Nonparametric regression with adaptive truncation via a convex hierarchical penalty

BY ASAD HARIS, ALI SHOJAIE AND NOAH SIMON

Department of Biostatistics, University of Washington, 1705 NE Pacific Street, Seattle, Washington 98195, U.S.A.

aharis@uw.edu ashojaie@uw.edu nrsimon@uw.edu

SUMMARY

We consider the problem of nonparametric regression with a potentially large number of covariates. We propose a convex, penalized estimation framework that is particularly well suited to high-dimensional sparse additive models and combines the appealing features of finite basis representation and smoothing penalties. In the case of additive models, a finite basis representation provides a parsimonious representation for fitted functions but is not adaptive when component functions possess different levels of complexity. In contrast, a smoothing spline-type penalty on the component functions is adaptive but does not provide a parsimonious representation. Our proposal simultaneously achieves parsimony and adaptivity in a computationally efficient way. We demonstrate these properties through empirical studies and show that our estimator converges at the minimax rate for functions within a hierarchical class. We further establish minimax rates for a large class of sparse additive models. We also develop an efficient algorithm that scales similarly to the lasso with the number of covariates and sample size.

Some key words: Additive model; High-dimensional data; Minimax estimation; Nonparametric regression; Sparsity.

1. INTRODUCTION AND MOTIVATION

Consider univariate nonparametric function estimation from observations $\{(x_i, y_i) \in \mathbb{R}^2 : i = 1, \dots, n\}$. Assume that $y_i = f(x_i) + \varepsilon_i$ ($i = 1, \dots, n$), where the ε_i are independent zero-mean sub-Gaussian random variables. There are many proposals for estimating f , including local polynomials (Stone, 1977), kernel smoothing (Nadaraya, 1964; Watson, 1964), and splines (Wahba, 1990). To begin, we focus on basis expansion estimators, also known as projection estimators (Čencov, 1962), which are widely used and arguably the simplest.

Let $y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ and $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ be the response and covariate vectors. For $v \in \mathbb{R}^n$, let $\|v\|_n^2 = n^{-1} \sum_{i=1}^n v_i^2$ be a modified ℓ_2 -norm, referred to as the empirical norm. Projection estimators are solutions to linear regression problems based on a set of basis functions $(\psi_k)_{k=1}^\infty$, with a truncation level K . More specifically, let $\Psi_K \in \mathbb{R}^{n \times K}$ be the $n \times K$ matrix with entries $\Psi_{K(i,k)} = \psi_k(x_i)$ ($k = 1, \dots, K; i = 1, \dots, n$). The basis expansion estimate of f is given by $\hat{f} = \sum_{k=1}^K \hat{\beta}_k^{\text{proj}} \psi_k$, where

$$\hat{\beta}^{\text{proj}} = \arg \min_{\beta \in \mathbb{R}^K} \frac{1}{2} \|y - \Psi_K \beta\|_n^2. \quad (1)$$

Fig. 1. (a) Plots of component functions f_1 (solid) and f_2 (dashed). (b) Minimum mean square errors as functions of n for our proposal (dashed) and (2) (solid). (c) Degrees of the fitted polynomial as functions of n : for our proposal, $\hat{f}_1$ (dashed) and $\hat{f}_2$ (dot-dashed) are shown; for (2) both component functions have the same degree (solid). All results are averaged over 100 replications.

To asymptotically balance bias and variance, $K \equiv K_n$ is allowed to vary with n . Unfortunately, choosing the truncation level K can be difficult in practice; it depends on the variance of ε_i , properties of f such as smoothness, and the choice of basis functions. Usually, K is chosen via split-sample validation. The challenge of tuning K becomes more evident in multivariate problems, which we describe next, and this is one of the main motivations for the present work.

Multivariate additive models (Hastie et al., 2009) easily follow from projection estimators, where each $x_i = (x_{i1}, \dots, x_{ip})^T$ is now a p -vector and the true underlying model is believed to be of the form $y_i = \sum_{j=1}^p f_j(x_{ij}) + \varepsilon_i$ ($i = 1, \dots, n$). The components f_j of this model can be estimated by using a basis expansion for each j and solving

$$\hat{\beta}_1^{\text{A-proj}}, \dots, \hat{\beta}_p^{\text{A-proj}} = \arg \min_{\beta_j \in \mathbb{R}^{K_j}} \frac{1}{2} \left\| y - \sum_{j=1}^p \Psi_{K_j}^j \beta_j \right\|_n^2, \quad (2)$$

where the $\Psi_{K_j}^j$ are K_j basis functions for feature j , and f_j is estimated as $\hat{f}_j = \sum_{k=1}^{K_j} \hat{\beta}_{jk}^{\text{A-proj}} \psi_k$.

In practice, the same truncation level is used for each feature, i.e., $K_j \equiv K$, to reduce the number of tuning parameters. When the f_j have widely different complexities, this leads to poor estimates. This limitation, which is illustrated by the simulation in Fig. 1, becomes more of a hindrance in higher dimensions, as p increases. In high-dimensional problems, when $p \gg n$, it is often assumed that $f_j \equiv 0$ for many components. A popular choice is then to add a sparsity-inducing penalty to the basis expansion framework (Ravikumar et al., 2009) and solve

$$\hat{\beta}_1^{\text{SPAM}}, \dots, \hat{\beta}_p^{\text{SPAM}} = \arg \min_{\beta_j \in \mathbb{R}^K} \frac{1}{2} \left\| y - \sum_{j=1}^p \Psi_K^j \beta_j \right\|_n^2 + \lambda \sum_{j=1}^p \left\| \Psi_K^j \beta_j \right\|_n. \quad (3)$$

In this paper, we propose a penalized estimation framework that penalizes function complexity, simultaneously selects the truncation level, and can be used to fit both univariate and multivariate additive models with or without sparsity. We present an extension to fully nonparametric multivariate settings, as well as a relaxed version, similar to the relaxed lasso (Meinshausen, 2007), which reduces the bias. While our univariate proposal is similar to (1), our additive proposal data-adaptively selects the truncation level for each feature f_j . This improves the prediction accuracy

and provides parsimonious estimates of f_j . We illustrate these advantages in a small simulation study with data $y_i = f_1(x_{i1}) + f_2(x_{i2}) + \varepsilon_i$ ($i = 1, \dots, n$), using the f_1 and f_2 shown in Fig. 1. We fit (2) using $K_j \equiv K$ selected to optimize mean square error, and compare it with our relaxed proposal. Tuning the parameters for our method also minimizes the mean square error. The results in Fig. 1 clearly demonstrate the superior performance of our method, which has lower mean square error while maintaining parsimony. In particular, we estimate the linear term, f_1 , by a linear function, whereas (2) uses an order-9 polynomial.

In addition to adaptability and parsimony, our proposal is computationally efficient and can work with thousands of observations and features. Moreover, its estimates attain minimax-optimal rates under standard smoothness assumptions, for univariate, multivariate, and sparse additive models. The univariate estimator converges at rate $n^{-2m/(2m+1)}$, where m is the degree of smoothness; similarly, the multivariate estimator attains the rate $n^{-2m/(2m+p)}$. For sparse additive models, under a suitable compatibility condition, our estimator converges at the rate $\max\{sn^{-2m/(2m+1)}, s \log p/n\}$, where s is the number of nonzero f_j ; even without the compatibility condition, consistency is achieved at the rate $\max\{sn^{-m/(2m+1)}, s(\log p/n)^{1/2}\}$.

2. METHODOLOGY

2.1. Motivation for adaptive truncation

Our proposal is motivated by the need to select the truncation levels in a data-driven manner. Let us first reconsider the simple projection estimator. The bias-variance trade-off and parsimony of the estimates $\hat{f}_j$ in (2) are controlled by the truncation levels K_j . While separately tuning K_j over each component function may be feasible in low dimensions, it quickly becomes infeasible for additive models, as the optimal truncation level requires searching over a p -dimensional space.

To bypass the tuning of multiple truncation levels K_j ($j = 1, \dots, p$), one can instead use n basis functions for each component and consider a penalized version of the truncation estimator,

$$\hat{\beta}_1^{\text{proj-pen}}, \dots, \hat{\beta}_p^{\text{proj-pen}} = \arg \min_{\beta_j \in \mathbb{R}^n} \frac{1}{2} \left\| y - \sum_{j=1}^p \Psi_n^j \beta_j \right\|_n^2 + \lambda \sum_{j=1}^p \max(k : \beta_{jk} \neq 0), \quad (4)$$

where the truncation level for each feature is determined using the penalty $\max(k : \beta_{jk} \neq 0)$. The estimator in (4) chooses the truncation level for each feature in a data-adaptive manner. However, the penalty in (4) is nonconvex, so solving (4) becomes infeasible in moderate- to high-dimensional problems. To mitigate this challenge, we formulate a convex problem using a novel penalty that can be seen as a convex relaxation of the penalty in (4).

Our approach is particularly suitable for basis functions that possess a natural hierarchy, that is, when $(\psi_k)_{k=1}^\infty$ become increasingly complex for higher values of k , as opposed to, say, natural splines, which rely on knot points. Examples of hierarchical basis functions include polynomial, trigonometric and wavelet basis functions; see the Supplementary Material.

2.2. The univariate proposal

Consider again the projection estimator (1). As noted in § 1, choosing the truncation level K is key here: too small a K will result in a large bias, while choosing K too large will over-inflate the variance. The bias and variance are balanced by taking $K = O\{n^{1/(2m+1)}\}$, where m relates to the smoothness of the underlying f and is unknown in practice. To circumvent this difficulty,

we use instead a complete basis with $K = n$ along with regularization to choose the truncation level. Our estimator is defined as

$$\hat{\beta}^{\text{hier}} = \arg \min_{\beta \in \mathbb{R}^n} \frac{1}{2} \|y - \Psi_n \beta\|_n^2 + \lambda \Omega(\beta), \quad \Omega(\beta) = \sum_{k=1}^n w_k \|\Psi_{k:n} \beta_{k:n}\|_n, \quad (5)$$

with $w_k = k^m - (k-1)^m$. Here, $\Psi_{k:n}$ denotes the submatrix of Ψ_n containing columns k to n , $\beta_{k:n}$ is the subvector of β containing entries k to n , and m and λ are tuning parameters. The choice of weights w_k is theoretically motivated and is detailed in § 5. Briefly, it defines a function class with desirable properties that allow us to establish convergence rates.

The hierarchical group lasso penalty, $\Omega(\beta)$, will result in solutions $\hat{\beta}^{\text{hier}}$ with hierarchical sparsity; that is, if $\hat{\beta}_k^{\text{hier}} = 0$ for some k , then $\hat{\beta}_{k'}^{\text{hier}} = 0$ for all $k' > k$. For sufficiently large λ , many entries of $\hat{\beta}^{\text{hier}}$ will be 0. For a given λ , we define the induced truncation level to be the minimal integer $K \leq n$ such that $\hat{\beta}_k^{\text{hier}} = 0$ for all integers $k > K$. Unlike the simple basis expansion estimator (1), this truncation level is data-adaptive, not prespecified.

Equation (5) involves two tuning parameters, m and λ . The parameter m is analogous to the smoothness parameter in smoothing splines (Wahba, 1990), or the number of bounded derivatives used in the simple projection estimator (Čencov, 1962). In practice, using $m = 2$ or 3 gives good results; this is similar to the use of cubic smoothing splines. On the other hand, λ determines the trade-off between goodness-of-fit and parsimony; a theoretically optimal λ -value is $\lambda \propto n^{-m/(2m+1)}$. Split-sample validation can be used to choose λ in practice.

As with the lasso, the regularization in (5) results in bias, which can reduce the overall mean square error. To reduce this bias, we can consider the relaxed version of our estimator in (5) as the simple basis expansion estimator with $K = \|\hat{\beta}^{\text{hier}}\|_0$, the truncation level selected by (5). This relaxed proposal is equivalent to using the penalty to select a truncation level. Therefore, in the univariate case, the relaxed estimator for a sequence of λ -values would match the simple basis expansion estimator (1) for a sequence of K -values. However, the advantages of our penalty become more clear in the case of multivariate additive models.

2.3. The additive proposal

Ideally, the additive projection estimator (2) is obtained by considering a different truncation level K_j for each feature. When p is small, this can be achieved by using split-sample validation and searching over all combinations of K_j ($j = 1, \dots, p$); however, the number of candidate models grows exponentially with p and quickly becomes unwieldy. Often, a single $K \equiv K_j$ is used in practice, which can lead to some f_j estimates having too many degrees of freedom. As illustrated in Fig. 1, using a single truncation level can lead to poor estimates.

Our proposal, which can be seen as a convex relaxation of (4), is designed to circumvent the above limitation of projection estimators in choosing the truncation level for models with multiple covariates. This differentiates our proposal from the projection estimator: in our framework, a single tuning parameter λ leads to different K_j for each fitted f_j .

Our additive framework is a direct extension of our univariate proposal (5). Specifically, we consider function estimates $\hat{f}_j = \sum_{k=1}^n \hat{\beta}_{jk}^{\text{A-hier}} \psi_k$, where

$$\hat{\beta}_1^{\text{A-hier}}, \dots, \hat{\beta}_p^{\text{A-hier}} = \arg \min_{\beta_j \in \mathbb{R}^n} \frac{1}{2} \left\| y - \sum_{j=1}^p \Psi_n^j \beta_j \right\|_n^2 + \lambda \sum_{j=1}^p \Omega_j(\beta_j), \quad (6)$$

with Ω_j being the hierarchical group lasso penalty with weights $w_k = k^m - (k - 1)^m$,

$$\Omega_j(\beta_j) = \sum_{k=1}^n w_k \|\Psi_{k:n}^j \beta_{j,k:n}\|_n. \quad (7)$$

The optimization problem (6) results in $\hat{\beta}_j$ estimates that are hierarchically sparse for each j . Specifically, for each j there is some minimal K_j such that $\hat{\beta}_{jk}^{\text{A-hier}} = 0$ for all integers $k > K_j$. Moreover, the major advantage of (6) is that the induced truncation level is featurewise adaptive, with a different K_j for each feature j . Additionally, as in the univariate setting, we can define a relaxed version of our estimator by fitting (2), where K_j is now determined by (6). As a result, our framework balances goodness-of-fit and parsimony for each feature individually, without requiring an exhaustive search. This is a major advantage over simple projection estimators.

The advantage of our method over simple projection estimators becomes more evident in high dimensions, when $p \gg n$. For instance, the popular estimator of [Ravikumar et al. \(2009\)](#), (3), is generally obtained using a single truncation level, which, as noted above, can result in poor estimates. Similar to their proposal, our sparse additive framework encourages featurewise sparsity using a group lasso penalty ([Yuan & Lin, 2006](#)), and is defined as

$$\hat{\beta}_1^{\text{S-hier}}, \dots, \hat{\beta}_p^{\text{S-hier}} = \arg \min_{\beta_j \in \mathbb{R}^n} \frac{1}{2} \left\| y - \sum_{j=1}^p \Psi_n^j \beta_j \right\|_2^2 + \lambda^2 \sum_{j=1}^p \Omega_j(\beta_j) + \lambda \sum_{j=1}^p \|\Psi_n^j \beta_j\|_n, \quad (8)$$

with $\Omega_j(\beta_j)$ as given in (7). We can again define a relaxed version which fits (2) with sparsity and K_j selected by (8). An important feature of the optimization problem (8) is that the tuning parameters for the two penalty terms λ and λ^2 are linked. This link is theoretically justified in § 4. Briefly, for an oracle λ , the choice of tuning parameters in (8) gives rate-optimal estimates. In practice, while this formulation gives good predictive performance in many cases, in other situations tuning sparsity and smoothness separately leads to strong predictive performance. Our numerical experiments in § 5 and 6 corroborate this finding.

As with $\hat{\beta}_j^{\text{SPAM}}$ in (3), for sufficiently large λ our proposal gives a sparse solution with most $\hat{\beta}_j^{\text{S-hier}} \equiv 0$. The two estimators differ, however, in their nonzero estimates: nonzero $\hat{\beta}_j^{\text{S-hier}}$ are hierarchically sparse, with a data-driven feature-specific induced truncation level, whereas nonzero $\hat{\beta}_j^{\text{SPAM}}$ in (3) all have the same complexity. This additional flexibility of our method proves critical in high dimensions, and is achieved without paying a price in computational or sample complexity. Moreover, with the tuning parameters in (8), this additional flexibility is in theory achieved with the same number of tuning parameters as the method of [Ravikumar et al. \(2009\)](#).

2.4. Relationship to existing methods

The univariate framework of § 2.2 builds upon existing penalized estimation methods. A popular choice is the smoothing spline estimator ([Wahba, 1990](#)), which sets $(\psi_k)_{k=1}^n$ to n natural splines with knots at the observed covariates $x_1, \dots, x_n$; this estimator is found by minimizing $\|y - \Psi\beta\|_n^2 + 2\lambda \|C^{1/2}\beta\|_n^2$ over $\beta \in \mathbb{R}^n$, using $C \in \mathbb{R}^{n \times n}$ and $C_{j,k} = \int \psi_j^{(m/2+1/2)}(t) \psi_k^{(m/2+1/2)}(t) dt$, where $\psi^{(k)}$ denotes the k th-order derivative of ψ . The smoothing spline eliminates the dependence on the truncation level and has an efficient-to-compute closed-form solution; but its estimated functions are piecewise-polynomial splines of degree m with n knots. As a result,

smoothing spline estimates are not parsimonious. To achieve more parsimonious estimates, [Mammen & van de Geer \(1997\)](#) used a data-driven approach to select the knots in spline functions. Their locally adaptive regression splines use the same natural spline basis and are found by minimizing $\|y - \Psi\beta\|_n^2 + 2\lambda(m!)^{-1}\|D\beta\|_1$ over $\beta \in \mathbb{R}^n$, using $D \in \mathbb{R}^{(n-m-1) \times n}$ with $D_{i,j} = \psi_j^{(m)}(t_i) - \psi_j^{(m)}(t_{i-1})$. This is closely related to the more computationally tractable trend filtering proposal of [Kim et al. \(2009\)](#) and [Tibshirani \(2014\)](#).

Despite their appealing properties in the univariate setting, locally adaptive regression splines and trend filtering are computationally difficult to extend to high-dimensional sparse additive models; even for a single feature, neither estimator has a closed-form solution. The estimator (3) of [Ravikumar et al. \(2009\)](#) overcomes this difficulty by using a fixed truncation level for all p components. As mentioned earlier, its main drawback is that all nonzero components of the additive model have the same complexity. The sparse partially linear additive model of [Lou et al. \(2016\)](#) partly mitigates this problem by setting some of the nonzero components to linear functions using a hierarchical penalty of the form $\sum_{j=1}^p \lambda_1 \|\beta_j\|_2 + \lambda_2 \|\beta_{j,-1}\|_2$; here $\beta_{j,1}$ is the coefficient of the linear term in the basis expansion, $\beta_{j,-1} = (\beta_{j,2}, \dots, \beta_{j,K})^\top \in \mathbb{R}^{K-1}$, and λ_1 and λ_2 are tuning parameters. The first term in the penalty sets all of the coefficients for the j th feature to zero, whereas the second term sets only the $K - 1$ coefficients corresponding to higher-order terms to zero.

Our additive and sparse additive proposals of § 2.3 generalize those of [Ravikumar et al. \(2009\)](#) and [Lou et al. \(2016\)](#). The first becomes a special case of (8) if the weights in (7) are set to $w_1 = 1$ and $w_k = 0$ for $k > 1$. Similarly, with an orthogonal design matrix, $(\Psi_K^j)^\top \Psi_K^j / n = I_K$ ($j = 1, \dots, p$), [Lou et al.](#)'s method is a special case of (8) with the weights in (7) set to $w_1 = w_2 = 1$ and $w_k = 0$ for $k > 2$. Our theoretical analysis in § 4.5 indicates that, in addition to the improved flexibility, our choice of weights (7) results in optimal rates of convergence.

There are other proposals for estimating sparse additive models, including extensions of trend filtering and smoothing splines for additive models. Extensions of trend filtering were either not shown to be rate-optimal ([Petersen et al., 2016](#)), or only shown to be so for low-dimensional additive models ([Sadhanala & Tibshirani, 2018](#)). These extensions are also computationally challenging beyond first-order trend filtering. Similarly, the extension of smoothing splines by [Meier et al. \(2009\)](#) is computationally inefficient and not rate-optimal.

Using properties of reproducing kernel Hilbert spaces, some authors have proposed estimators that offer minimax-optimal convergence rates for the prediction error over smooth additive classes ([Koltchinskii & Yuan, 2010](#); [Raskutti et al., 2012](#); [Yuan & Zhou, 2015](#)), similar to our results in § 4.5. However, these estimators are given as minimizers of (np) -dimensional second-order cone programs, for which the authors did not discuss efficient algorithms, at most mentioning generic convex solvers. The computation for general-purpose second-order convex cone program solvers scales roughly as $(np)^3$; thus, even for moderate p and n , these proposals quickly become intractable. We compare and contrast the strengths and weaknesses of existing proposals in Table 1.

3. COMPUTATIONAL CONSIDERATIONS AND EXTENSIONS

3.1. Conservative basis truncation

Our proposal (5) uses a basis expansion with n basis functions. In practice, for any reasonable choice of λ , $\hat{\beta}$ will never have n nonzero entries, and will generally have very few nonzero entries, $K_0 \ll n$. If we instead solve

Table 1. Comparison of existing methods for sparse additive models

	PE	SS/RKHS	TF
Scalability	✓	×	×
	Exact solution for univariate subproblem; scales as $O(npK)$	No exact solution for univariate subproblem; general convex solver scales as $O\{(np)^3\}$	Inefficient beyond first-order TF, particularly for high-dimensional and unequally spaced covariates
Adaptability	×	✓	✓
	Smoothness controlled by basis expansion order, K , fixed for all component functions	Smoothness controlled by smoothness norm, varied for component functions	Smoothness controlled by smoothness norm and number of knots, K , varied for component functions
Parsimony	✓	×	✓
	Component functions of expansions to order $K \ll n$	Component functions of expansions to order n	Component functions having sparsity in the number of knots

PE, projection estimators (Ravikumar et al., 2009; Lou et al., 2016); RKHS, reproducing kernel Hilbert spaces (Koltchinskii & Yuan, 2010; Raskutti et al., 2012; Yuan & Zhou, 2015); SS, smoothing splines (Meier et al., 2009); TF, trend filtering (Petersen et al., 2016; Sadhanala & Tibshirani, 2018).

$$\hat{\beta}^{\text{hier}(\tilde{K})} = \arg \min_{\beta \in \mathbb{R}^{\tilde{K}}} \frac{1}{2} \|y - \Psi_{\tilde{K}} \beta\|_2 + \lambda \sum_{k=1}^{\tilde{K}} w_k \|\Psi_{k:\tilde{K}} \beta_{k:\tilde{K}}\|_n \quad (9)$$

for $\tilde{K} < n$, then provided $\tilde{K} \geq K_0$, the solution will be identical to that of the original proposal (5). Even when not identical, so long as $\tilde{K}$ is sufficiently large, in particular $\tilde{K} \gtrsim n^{2m/(2m-1)(2m+1)}$ where $a_n \gtrsim b_n$ means $a_n \geq Cb_n$ for some constant C , the theoretical properties of (5) will be maintained. This bound relies on the smoothness of the underlying f ; choosing $\tilde{K} \gtrsim n^{2/3}$ gives a conservative upper bound which is independent of the underlying f . Our theoretical results do not establish tight bounds on function approximation, but we conjecture that they can be improved to obtain the usual $\tilde{K} \gtrsim n^{1/(2m+1)}$ truncation level. Additionally, as discussed in § 3.2, by using $\tilde{K}$ rather than n basis functions, the computational complexity decreases from $O(n^2)$ to $O(n\tilde{K})$. A similar result holds for the sparse additive framework with

$$\hat{\beta}_1^{\text{S-hier}(\tilde{K})}, \dots, \hat{\beta}_p^{\text{S-hier}(\tilde{K})} = \arg \min_{\beta_j \in \mathbb{R}^{\tilde{K}}} \frac{1}{2} \left\| y - \sum_{j=1}^p \Psi_{\tilde{K}}^j \beta_j \right\|_2 + \lambda^2 \sum_{j=1}^p \Omega_j(\beta_j) + \lambda \sum_{j=1}^p \|\Psi_{\tilde{K}}^j \beta_j\|_n, \quad (10)$$

where now $\Omega_j(\beta_j) = \sum_{k=1}^{\tilde{K}} w_k \|\Psi_{k:\tilde{K}}^j \beta_{j,k:\tilde{K}}\|_n$. Choosing the pretruncation level $\tilde{K}$ is easier than choosing the truncation level for the simple basis expansion estimator (Čencov, 1962): the latter requires an exact truncation level that is neither too large nor too small, whereas the former requires only a level that is not too small.

3.2. Algorithm for the univariate and sparse additive framework

An appealing feature of our framework is its computational efficiency. Problem (9) can be solved via a one-step coordinate descent algorithm. Using a QR decomposition $\Psi_{\tilde{K}} = UV$ with $U \in \mathbb{R}^{n \times \tilde{K}}$ and $U^T U/n = I_{\tilde{K}}$, we can rewrite (9) as

$$\underset{\tilde{\beta} \in \mathbb{R}^{\tilde{K}}}{\text{minimize}} \quad \frac{1}{2n} \|y - U\tilde{\beta}\|_2^2 + \lambda \sum_{k=1}^{\tilde{K}} w_k \|\tilde{\beta}_{k:\tilde{K}}\|_2, \quad (11)$$

where $\tilde{\beta} = V\beta$. Applying the results of [Jenatton et al. \(2010\)](#) gives us Algorithm 1.

Algorithm 1. One-step coordinate descent for the univariate problem.

Initialize $\beta^0 = \dots = \beta^{\tilde{K}} \leftarrow U^T y/n$

For $k = \tilde{K}, \dots, 1$

Update $\beta_{k:\tilde{K}}^{k-1} \leftarrow (1 - w_k \lambda / \|\beta_{k:\tilde{K}}^k\|_2) \beta_{k:\tilde{K}}^k$, where $(x)_+ = \max(x, 0)$

Return β^0

The reformulation in (11) can also be used to efficiently solve the sparse additive extension (10) via a block coordinate descent algorithm. Specifically, given a set of estimates $(\beta_j)_{j=1}^p$, we fix all but one of the vectors β_j and optimize over the non-fixed vector using Algorithm 1. Iterating until convergence yields the solution to problem (10); see the Supplementary Material.

Solving problem (9) requires a QR decomposition of the matrix $\Psi_{\tilde{K}}$ followed by the multiplication $U^T y$; these steps require $O(n\tilde{K}^2)$ and $O(n\tilde{K})$ operations, respectively. However, these steps are needed only once for a sequence of λ -values. For the additive proposal (10), p such QR decompositions are needed once for the entire λ sequence.

By Proposition 2 of [Jenatton et al. \(2010\)](#), for a given λ , problem (11) can be solved in $O(\tilde{K})$ operations. Each block update requires a matrix multiplication $U_j^T r_{-j}$ followed by solution of the proximal problem (11); see the Supplementary Material. This requires $O(n\tilde{K})$ operations. Thus, our sparse additive proposal requires $O(np\tilde{K})$ operations, which is the computational complexity of the lasso ([Friedman et al., 2010](#)) when $\tilde{K} = 1$.

The above computational complexity calculations indicate that our univariate and sparse additive estimates can be obtained very efficiently. In fact, using our R implementation ([R Development Core Team, 2019](#)), the median time for solving the univariate problem for an example with $\tilde{K} = n = 300$ is 0.17 seconds on an Intel® CORE™ i5-3337U 1.80 GHz processor. The median time for solving the sparse additive framework for the simulation setting of § 5.2 on a grid of 50 λ -values is 5.96 seconds.

3.3. Degrees of freedom

For regression with fixed design and $\varepsilon_i \sim N(0, \sigma^2)$, we consider the definition of degrees of freedom given by [Stein \(1981\)](#), $DF = \sum_{i=1}^n \text{cov}(y_i, \hat{y}_i)/\sigma^2$, where $\hat{y}_i$ are the fitted response values. We use Claim 3.2 of [Haris et al. \(2016\)](#) to derive an unbiased estimate of DF for the estimator (11), using the decomposition $\Psi_{\tilde{K}} = UV$ from § 3.2. Let $K_0 = \max(k : \hat{\beta}_k \neq 0)$, and let $U_{K_0} \in \mathbb{R}^{n \times K_0}$ denote the first K_0 columns of U . For a vector $v \in \mathbb{R}^{\tilde{K}}$, define $v_{k:K_0} \in \mathbb{R}^{K_0}$ as $v_{k:K_0} = (0, \dots, 0, v_k, v_{k+1}, \dots, v_{K_0})^T$. We arrive at the following lemma.

Fig. 2. Visual representation of the multivariate penalty with $p = 2$ and $\psi_j(x) \equiv x$.

LEMMA 1. *An unbiased estimator for the degrees of freedom of $\hat{\beta}$ in (9) is*

$$\hat{\text{DF}} = 1 + \text{tr} \left\{ U_{K_0} \left[I_{K_0} + \sum_{k=1}^{K_0} \lambda w_k \left\{ \frac{\text{diag}(1_{k:K_0})}{\|\hat{\beta}_{k:K_0}\|_2} - \frac{\hat{\beta}_{k:K_0} \hat{\beta}_{k:K_0}^T}{\|\hat{\beta}_{k:K_0}\|_2^3} \right\} \right]^{-1} \frac{U_{K_0}^T}{n} (I_n - 1_n 1_n^T / n) \right\},$$

where $\text{diag}(v) \in \mathbb{R}^{K_0 \times K_0}$ is a diagonal matrix with $v \in \mathbb{R}^{K_0}$ on the main diagonal.

3.4. Nonadditive multivariate regression

For vectors $x \in \mathbb{R}^p$ and $v_k \in \mathbb{Z}_+^p$, define $x^{v_k} = x_1^{v_{k1}} \times \dots \times x_p^{v_{kp}}$. Now, for functions $f^0 : \mathbb{R}^p \rightarrow \mathbb{R}$, consider the basis representation $f^0(x) = \sum_{k=1}^{\tilde{K}} \psi_k(x^{v_k}) \beta_k^0$ for univariate functions $(\psi_j)_{j=1}^\infty$ and $v_1, \dots, v_{\tilde{K}} \in \mathbb{Z}_+^p$, where

$$\|v_k\|_1 = 1 \quad (k = 1, \dots, p), \quad \|v_k\|_1 = 2 \quad \{k = p+1, \dots, (p+2)(p+1)/2 - 1\},$$

and so on. As in the univariate case, let $\Psi_{\tilde{K}} \in \mathbb{R}^{n \times \tilde{K}}$ be the matrix with entries $\Psi_{\tilde{K}(i,k)} = \psi_k(x_i^{v_k})$. Then our multivariate regression estimator is simply (9) with weights given by

$$w_{q_k} = k^m - (k-1)^m, \quad q_k = \binom{k+p-1}{p}, \quad (12)$$

and $w_k = 0$ for all other k . Figure 2 demonstrates the multivariate penalty for $p = 2$ and ψ_k the identity function, that is, for $z \in \mathbb{R}$ and $\psi_k(z) \equiv z$. It is clear from the figure how the multivariate penalty is a natural extension of the univariate one: when $\psi_k(z) = z$, the fitted model can be a multivariate polynomial of any degree. With this choice of basis functions, our multivariate proposal acts as a procedure for selecting the complexity level of interaction models. This problem can be solved using Algorithm 1 with a single pass over the basis elements.

3.5. Extension to classification

We can extend our method to the setting of binary classification via a logistic loss function. Let $y_i \in \{-1, 1\}$ ($i = 1, \dots, n$) be the observed response. We then fit

$$(\beta_0, \hat{\beta}) = \arg \min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^n} \frac{1}{2n} \sum_{i=1}^n \log \left(1 + \exp[-y_i \{\beta_0 + (\Psi_n \beta)_i\}] \right) + \lambda \Omega(\beta). \quad (13)$$

As with the least squares loss, (13) can be naturally extended to sparse additive models by using both penalties in (8). The problem can be solved efficiently via a proximal gradient descent algorithm (Combettes & Pesquet, 2011); see the Supplementary Material.

4. THEORETICAL RESULTS

4.1. Summary of theoretical contributions

To investigate finite-sample properties of our estimators, we combine previously developed ideas from empirical process theory and metric entropy with a number of novel results about convergence rates of sparse additive models, as well as the metric entropy of our hierarchical class.

Our new results in § 4.5 allow one to establish convergence rates for a broad class of penalized sparse additive model estimators. Under a compatibility condition on the features, these rates match the minimax lower bound for estimation of sparse additive models under independent component functions (Raskutti et al., 2009). Thus, our sparse additive estimators are rate-optimal. With no such assumptions, in Theorem 4 we obtain rates that are the additive analogues of assumption-free convergence rates for the lasso (Chatterjee, 2014). Such assumption-free convergence rates have not previously been derived for sparse additive models.

Finally, key to our theoretical analyses is the entropy of our hierarchical class; we calculate these entropies via matching upper and lower bounds in Lemmas 3 and 4. These new results enable us to show that our univariate and sparse additive estimators, (5) and (8), are minimax rate-optimal within the hierarchical univariate and hierarchical sparse additive classes, respectively.

4.2. Entropy-based rates

We begin by stating two well-known results. We then present our contributions in § 4.4 and 4.5. Firstly, Theorem 1 of Yang & Barron (1999) establishes a lower bound for the minimax rate subject to certain conditions. Secondly, a framework for establishing an upper bound on convergence rates is given by Theorem 10.2 of van de Geer (2000). Here, we require a slight generalization of this result, which we state below and prove in the Supplementary Material.

We first introduce some terminology and notation. For a set $\mathcal{F}$ equipped with some metric $d(\cdot, \cdot)$, the subset $\{f_1, \dots, f_N\} \subset \mathcal{F}$ is a δ -cover if for any $f \in \mathcal{F}$, $\min_{1 \leq i \leq N} d(f, f_i) \leq \delta$. The log-cardinality of the smallest δ -cover is the δ -entropy of $\mathcal{F}$ with respect to the metric $d(\cdot, \cdot)$. We denote by $H(\delta, \mathcal{F}, Q)$ the δ -entropy of a function class $\mathcal{F}$ with respect to the $\|\cdot\|_Q$ metric for a measure Q , where $\|f\|_Q^2 = \int \{f(x)\}^2 dQ(x)$. For a fixed sample $x_1, \dots, x_n$, we denote by Q_n the empirical measure $Q_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$ and use the shorthand $\|\cdot\|_n = \|\cdot\|_{Q_n}$.

THEOREM 1 (Yang & Barron, 1999, Theorem 1). *Consider the model $y_i = f^0(x_i) + \varepsilon_i$ ($i = 1, \dots, n$) with independent and identically distributed $\varepsilon_i \sim N(0, \sigma^2)$ and $x_i \sim Q$. Assume that the entropy condition $H(\delta, \mathcal{F}, Q) = A_0 \delta^{-\alpha}$ holds for some function class $\mathcal{F}$, for $\alpha \in (0, 2)$ and $A_0 > 0$. Then, for a constant A_1 that depends on A_0 , α and σ^2 ,*

$$\min_{\hat{f}} \max_{f^0 \in \mathcal{F}} E\left(\|\hat{f} - f^0\|_Q^2\right) \geq A_1 n^{-2/(2+\alpha)},$$

where the minimum is over the space of all measurable functions.

THEOREM 2 (van de Geer, 2000, Theorem 10.2). *Consider the model $y_i = f^0(x_i) + \varepsilon_i$ ($i = 1, \dots, n$) with independent sub-Gaussian noise ε_i . Let $\hat{f} = \arg \min_{f \in \mathcal{F}_n} \|y - f\|_n^2 + 2\lambda_n^2 \Omega(f | Q_n)$, for some function class $\mathcal{F}_n$ and seminorm $\Omega(\cdot | Q_n)$ on $\mathcal{F}_n$ which satisfy the entropy condition $H[\delta, \{f \in \mathcal{F}_n : \Omega(f | Q_n) \leq 1\}, Q_n] \leq A_0 \delta^{-\alpha}$ for $\alpha \in (0, 2)$. Then, for $\lambda_n^{-1} = n^{1/(2+\alpha)} \{\Omega(f_n^* | Q_n)\}^{(2-\alpha)/(2(2+\alpha))}$ and for any function $f_n^* \in \mathcal{F}_n$, there is a constant c such that for all $T \geq c$,*

with probability at least $1 - c \exp\{-(T/c)^2\}$,

$$\|\hat{f} - f^0\|_n^2 \leq 5 \max \left\{ 2\|f^0 - f_n^*\|_n^2, C_0 \lambda_n^2 \Omega(f_n^* | Q_n) \right\},$$

where C_0 is a constant that depends on α and T .

Before specializing Theorems 1 and 2 to our proposal, we briefly discuss their assumptions. The main assumption in Theorem 1 is an entropy condition for the function class $\mathcal{F}$, which contains f^0 . This is a common condition needed to quantify the size of an infinite-dimensional space $\mathcal{F}$. The entropy condition $H(\delta, \mathcal{F}, Q) = A_0 \delta^{-\alpha}$ is satisfied by many commonly used function classes; examples include bounded Lipschitz functions with $\alpha = 1$, bounded monotone functions with $\alpha = 1$, and m th-order Sobolev functions with $\alpha = 1/m$. Theorem 2 requires a similar entropy condition on a sequence of function spaces, but it does relax conditions on f^0 , allowing it to be arbitrary and not necessarily in a specific class. Requiring an entropy condition for a sequence $(\mathcal{F}_n)_{n=1}^\infty$ may seem restrictive; but often, as with our hierarchical function class defined below, for all n one has $\mathcal{F}_n \subseteq \mathcal{F}$ for some class $\mathcal{F}$. Thus it suffices to prove an entropy bound for $\mathcal{F}$. Finally, we also require the noise ε_i to be independent and sub-Gaussian in order to use standard results from the empirical processes literature. While the original theorem of Yang & Barron (1999) requires identically distributed Gaussian noise to prove a lower bound, the fact that Gaussian random variables are sub-Gaussian allows us to generalize the original result. In the following subsection, we define and establish entropy bounds for our hierarchical function class.

4.3. Entropy results for the proposed penalty

To set up the notation, we define the univariate function class

$$\mathcal{F}_n = \left\{ f_\beta(x) = \sum_{k=1}^{\tilde{K}_n} \psi_k(x) \beta_k : \int \psi_k \psi_l dQ = 0 \text{ for } k \neq l, \int \psi_k^2 dQ = 1 \right\}, \quad x \in \mathbb{R}, \quad (14)$$

and the multivariate function class

$$\mathcal{F}_{p,n} = \left\{ f_\beta(x) = \sum_{k=1}^{\tilde{K}_n} \psi_k(x^{v_k}) \beta_k : \int \psi_k(x^{v_k}) \psi_l(x^{v_l}) dQ = 0 \text{ for } k \neq l, \int \{\psi_k(x^{v_k})\}^2 dQ = 1 \right\}, \quad x \in \mathbb{R}^p, \quad (15)$$

where $v_k \in \mathbb{Z}_+^p$, x^{v_k} was defined in § 3.4, and Q is the probability measure associated with x . In (14) and (15), we allow for the limiting case of $n = \infty$ with $\tilde{K}_\infty = \infty$. With some abuse of notation, for $\beta \in \ell^2(\mathbb{R})$ we define $\|\beta_{k:\infty}\|_2^2 = \sum_{l=k}^\infty \beta_l^2$.

To specialize Theorems 1 and 2, we need to characterize $H(\delta, \mathcal{F}_\infty^M, Q)$ for $\mathcal{F}_\infty^M$ defined below in (16), and establish an upper bound for $H[\delta, \{f_\beta \in \mathcal{F}_n : \Omega(\beta) \leq 1\}, Q_n]$. In the next lemma, Lemma 2, we show that the calculation of $H(\delta, \mathcal{F}_\infty^M, Q)$ and $H[\delta, \{f_\beta \in \mathcal{F}_n : \Omega(\beta) \leq 1\}, Q_n]$ is equivalent to an entropy calculation for subsets of $\ell^2(\mathbb{R})$ and $\mathbb{R}^{\tilde{K}_n}$, respectively, with respect to the usual $\|\cdot\|_2$ norm. This reduction allows us to use simple volume arguments and existing results to establish the entropy conditions. The lemma considers our proposed penalty in full generality, that is, the penalty (5) with any set of nonnegative weights w_k . This lemma gives a similar reduction of entropy calculations for the multivariate case with little extra work.

LEMMA 2 (Reduction to $\ell^2(\mathbb{R})$ and $\mathbb{R}^{\tilde{K}_n}$). *Let $\mathcal{F}_n^M$ and $\mathcal{F}_{p,n}^M$ be, respectively, the univariate and multivariate hierarchical basis expansion classes with bounded penalty. Specifically,*

$$\mathcal{F}_n^M = \left\{ f_\beta \in \mathcal{F}_n : \sum_{k=1}^{\tilde{K}_n} w_k \|\beta_{k:\tilde{K}_n}\|_2 \leq M \right\}, \quad \mathcal{F}_{p,n}^M = \left\{ f_\beta \in \mathcal{F}_{p,n} : \sum_{k=1}^{\tilde{K}_n} w_k \|\beta_{k:\tilde{K}_n}\|_2 \leq M \right\}, \quad (16)$$

where we allow the limiting case of $n = \infty$. Then $H(\delta, \mathcal{F}_n^M, Q)$ or $H(\delta, \mathcal{F}_{p,n}^M, Q)$ is equal to $H(\delta, \mathcal{H}_{\tilde{K}_n}^{w/M})$, the entropy of $\mathcal{H}_{\tilde{K}_n}^{w/M}$ with respect to the $\|\cdot\|_2$ norm, where

$$\mathcal{H}_{\tilde{K}_n}^{w/M} = \left\{ \beta \in \mathbb{R}^{\tilde{K}_n} : \sum_{k=1}^{\tilde{K}_n} w_k / M \|\beta_{k:\tilde{K}_n}\|_2 \leq 1 \right\}.$$

Secondly, assume that the Gram matrix $\Psi_{\tilde{K}_n}^\top \Psi_{\tilde{K}_n} / n$ has a finite maximum eigenvalue, denoted by $\Lambda_{\max}$. Then, writing $\mathcal{H}_{\tilde{K}_n}^w = \mathcal{H}_{\tilde{K}_n}^{w/1}$, we have

$$H\left[\delta, \left\{ f_\beta \in \mathcal{F}_n : \sum_{k=1}^{\tilde{K}_n} w_k \|\Psi_{k:\tilde{K}_n} \beta_{k:\tilde{K}_n}\|_n \leq 1 \right\}, Q_n \right] \leq H(\delta \Lambda_{\max}^{-1/2}, \mathcal{H}_{\tilde{K}_n}^w).$$

The above inequality also holds with $\mathcal{F}_n$ replaced by $\mathcal{F}_{p,n}$.

Lemma 2 establishes the connections between the entropies of the function classes of interest and the set $\mathcal{H}_{\tilde{K}_n}^w$. It is easy to see that $H(\delta, \mathcal{H}_{\tilde{K}_n}^{w/M})$ and $H(\delta \Lambda_{\max}^{-1/2}, \mathcal{H}_{\tilde{K}_n}^w)$ are proportional to $H(\delta, \mathcal{H}_{\tilde{K}_n}^w)$, where the proportionality constants depend on M and $\Lambda_{\max}$, respectively. The next lemma establishes an upper bound on $H(\delta, \mathcal{H}_{\tilde{K}_n}^w)$ for the proposed choice of univariate and multivariate weights. This upper bound is all we need to specialize Theorem 2.

LEMMA 3 (An upper bound). *Suppose $\delta \geq 0$. For the region $\mathcal{H}_{\tilde{K}_n}^w$ with univariate weights $w_k = k^m - (k-1)^m$, $H(\delta, \mathcal{H}_{\tilde{K}_n}^w) \leq U_{E,1} \delta^{-1/m}$ for a constant $U_{E,1} > 0$. Moreover, for the multivariate weights (12) we have $H(\delta, \mathcal{H}_{\tilde{K}_n}^w) \leq U_{E,2} \delta^{-p/m}$ for a constant $U_{E,2} > 0$.*

While Lemma 3 is sufficient for applying Theorem 2, to invoke Theorem 1 we need an exact value for the entropy up to a proportionality constant. A natural way to obtain this is to find a lower bound for the entropy which matches the upper bound; we do this in the following lemma.

LEMMA 4 (A lower bound). *For $\delta \in \{(w_1 + \dots + w_{\tilde{K}_n+1})^{-m}, 1/2\}$, for the region $\mathcal{H}_{\tilde{K}_n}^w$ with univariate weights $w_k = k^m - (k-1)^m$ we have $H(\delta, \mathcal{H}_{\tilde{K}_n}^w) \geq L_{E,1} \delta^{-1/m}$, and for the multivariate weights (12) we have $H(\delta, \mathcal{H}_{\tilde{K}_n}^w) \geq L_{E,2} \delta^{-p/m}$, for constants $L_{E,1}, L_{E,2} > 0$ and where we assume, for simplicity, that $\tilde{K}_n = q_{\tilde{K}'} - 1$ for some $\tilde{K}'$ and $q_{\tilde{K}'}$ as defined in (12).*

Lemmas 2–4 demonstrate the motivation for our weights w_k ; they define a function class with the same entropy as an m th-order Sobolev class. In fact, our class $\mathcal{F}_\infty^M$ is a subset of the m th-order Sobolev class $\mathcal{G}_2^M$, where $\mathcal{G}_q^M = \{f_\beta \in \mathcal{F}_\infty : \sum_{k=1}^\infty (k^m |\beta_k|)^q \leq M^q\}$ is the weighted L_q

space (Rauhut & Ward, 2016). Furthermore, we prove that $\mathcal{G}_1^M \subseteq \mathcal{F}_\infty^M \subseteq \mathcal{G}_2^M$; see the Supplementary Material. While the Sobolev class $\mathcal{G}_2^M$ is common in the literature (Ravikumar et al., 2009; van de Geer, 2010), the class $\mathcal{G}_1^M$ has recently attracted increased attention in the function interpolation literature; see, for example, Rauhut & Ward (2016), Candes et al. (2008) and the references therein.

4.4. Specializing Theorems 1 and 2

The following proposition establishes a lower bound for the minimax rate of estimating f^0 , the true function which belongs to some function class $\mathcal{F}$. We consider three different choices for $\mathcal{F}$: the univariate class (14), the multivariate class (15), and the Sobolev class $\mathcal{G}_2^M$. To prove the result, we use the fact that if an upper bound for the convergence rate can be found that matches the lower bound, then we can conclude that our estimator is minimax.

PROPOSITION 1. *For the m th-order function class $\mathcal{F}_\infty^M = \{f \in \mathcal{F}_\infty : \sum_{k=1}^\infty w_k \|\beta_{k:\infty}\|_2 \leq M\}$ where $w_k = k^m - (k-1)^m$,*

$$\min_{\hat{f}} \max_{f^0 \in \mathcal{F}_\infty^M} E\left(\|\hat{f} - f^0\|_Q^2\right) \geq A_1 n^{-2m/(2m+1)}.$$

For the m th-order multivariate class $\mathcal{F}_{p,\infty}^M = \{f \in \mathcal{F}_{p,\infty} : \sum_{k=1}^\infty w_k \|\beta_{k:\infty}\|_2 \leq M\}$ where w_k are the weights defined in (12),

$$\min_{\hat{f}} \max_{f^0 \in \mathcal{F}_{p,\infty}^M} E\left(\|\hat{f} - f^0\|_Q^2\right) \geq A_2 n^{-2m/(2m+p)}.$$

Finally, for the m th-order Sobolev class $\mathcal{G}_2^M = \{f \in \mathcal{F}_\infty : \sum_{k=1}^\infty (k^m \beta_k)^2 \leq M^2\}$,

$$\min_{\hat{f}} \max_{f^0 \in \mathcal{G}_2^M} E\left(\|\hat{f} - f^0\|_Q^2\right) \geq A_3 n^{-2m/(2m+1)}.$$

We now specialize Theorem 2 to establish an upper bound for the convergence rate of the proposed univariate and multivariate estimators. The following proposition provides some interesting insights. Firstly, with respect to the empirical norm $\|\cdot\|_n$, our estimators achieve the minimax rate for the classes $\mathcal{F}_\infty^M$ and $\mathcal{F}_{p,\infty}^M$, as defined in (16). For the Sobolev class $\mathcal{G}_2^M$, if $\sum_{k=1}^\infty w_k \|\beta_{k:\infty}\|_2 \leq C(M)$ for all $f_\beta \in \mathcal{G}_2^M$, then our univariate estimator is minimax over the Sobolev class as well. This result also gives insight into the role of $\tilde{K}_n$.

PROPOSITION 2. *Consider the model $y_i = f^0(x_i) + \varepsilon_i$ ($i = 1, \dots, n$) for zero-mean, sub-Gaussian noise ε_i . Define the univariate and multivariate estimators as*

$$\hat{f}^{\text{uni}} = \arg \min_{f_\beta \in \mathcal{F}_n} \frac{1}{2} \|y - f_\beta\|_n^2 + \lambda_n^2 \Omega^{\text{uni}}(\beta), \quad \hat{f}^{\text{multi}} = \arg \min_{f_\beta \in \mathcal{F}_{p,n}} \frac{1}{2} \|y - f_\beta\|_n^2 + \lambda_n^2 \Omega^{\text{multi}}(\beta)$$

for $p = 1$ and $p > 1$, respectively, where Ω^{uni} is the penalty in (9) and Ω^{multi} is the penalty in (12). Assume that $\max_k \|\psi_k\|_\infty = \psi_{\max} < \infty$ and that the Gram matrix $\Psi_{\tilde{K}_n}^\top \Psi_{\tilde{K}_n} / n$ has a bounded maximum eigenvalue, denoted by $\Lambda_{\max}$.

Then, for $p = 1$ and $f^0 \in \mathcal{F}_\infty^M$ there is a constant $c > 0$ such that for all $T \geq c$, with probability at least $1 - c \exp\{-(T/c)^2\}$,

$$\|\hat{f}^{\text{uni}} - f^0\|_n^2 \leq 5 \max\{C_1 \tilde{K}_n^{-(2m-1)}, C_2 n^{-2m/(2m+1)}\},$$

where $C_1, C_2 > 0$ are constants that depend on $M, \psi_{\max}, \Lambda_{\max}, m$ and T .

For $p = 1$ and $f^0 \in \mathcal{G}_2^M$ there is a constant $c > 0$ such that for all $T \geq c$, with probability at least $1 - c \exp\{-(T/c)^2\}$,

$$\|\hat{f}^{\text{uni}} - f^0\|_n^2 \leq 5 \max\{C_1 \tilde{K}_n^{-(2m-1)}, C_2 C_3 n^{-2m/(2m+1)}\},$$

where $C_1, C_2 > 0$ are constants that depend on $M, \psi_{\max}, \Lambda_{\max}, m$ and T , and for $f^0 = \sum_{k=1}^\infty \psi_k \beta_k^0$ we have $C_3^{(2m+1)/2} = \sum_{k=1}^\infty w_k \|\beta_{k:\infty}^0\|_2$.

For $1 < p < 2m$ and $f^0 \in \mathcal{F}_{p,\infty}^M$, let $\tilde{K}'$ be such that $\tilde{K}_n = q_{\tilde{K}'} - 1$ for $q_{\tilde{K}'}$ as in (12); then there is a constant $c > 0$ such that for all $T \geq c$, with probability at least $1 - c \exp\{-(T/c)^2\}$,

$$\|\hat{f}^{\text{multi}} - f^0\|_n^2 \leq 5 \max\{C_1 \tilde{K}'^{-(2m-1)}, C_2 n^{-2m/(2m+p)}\},$$

where $C_1, C_2 > 0$ are constants that depend on $M, \psi_{\max}, \Lambda_{\max}, m$ and T .

The pre-truncation level $\tilde{K}_n$ in Proposition 2 is not the truncation order selected by our proposal; rather, it is the prespecified maximum order of our proposal with conservative truncation in (9). The above result demonstrates that we achieve usual nonparametric rates as long as the truncation level $\tilde{K}_n$ satisfies $\tilde{K}_n \gtrsim n^{2m/\{(2m+1)(2m-1)\}}$, justifying $n^{2/3}$ as a conservative choice. Furthermore, if a function belongs to the m th-order hierarchical class, then it also belongs to the m' th-order class for all $m' \leq m$, and Proposition 2 holds with m replaced by m' . This means we can misspecify the smoothness order in our estimator and still get nonparametric convergence rates.

4.5. Theoretical results for sparse additive models

We next establish convergence rates of high-dimensional sparse additive models in terms of a general entropy condition. Our first contribution is an oracle inequality for an upper bound on the prediction error of additive models, which establishes the consistency of estimators with slow convergence rates; these rates are $O(sv_n)$, where sv_n^2 is the minimax lower bound of Raskutti et al. (2009) for sparse additive models and s is the cardinality of the set S defined below. For completeness, in Theorem 3 we state the result of Raskutti et al. (2009), which assumes independent covariates. We then proceed to state a compatibility condition which leads to two propositions: firstly it establishes convergence rates of order $O(sv_n^2)$, and secondly it automatically establishes minimax rates for univariate regression as a special case of an additive model with $p = 1$. These results extend to a broad class of estimators; consequently, we can establish new results on convergence rates for some existing methods, including methods that extend smoothing splines and trend filtering (Meier et al., 2009; Sadhanala & Tibshirani, 2018).

Let f^0 be the true function such that $y_i = f^0(x_i) + \varepsilon_i$ ($i = 1, \dots, n$) for independent zero-mean noise ε_i and $x_i = (x_{i1}, \dots, x_{ip})^T \in \mathbb{R}^p$. Let f^* be a sparse additive approximation to f^0 ,

$$f^*(x_i) = c^0 + \sum_{j=1}^p f_j^*(x_{ij}) = c^0 + \sum_{j \in S} f_j^*(x_{ij}),$$

where $S = \{j : f_j^* \neq 0\}$, which we call the active set, is a subset of $\{1, \dots, p\}$ of size $s = |S|$ and $c^0 = E(\bar{y})$ with $\bar{y}$ denoting the sample mean. To ensure identifiability, we assume $\sum_{i=1}^n f_j^*(x_{ij}) = 0$ ($j = 1, \dots, p$). Consider the estimator $\hat{f} = \sum_{j=1}^p \hat{f}_j$ with

$$\hat{f}_1, \dots, \hat{f}_p = \arg \min_{(f_j)_{j=1}^p \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^n \left\{ y_i - \bar{y} - \sum_{j=1}^p f_j(x_{ij}) \right\}^2 + \lambda_n \sum_{j=1}^p I(f_j), \quad (17)$$

where $I(\cdot)$ is a penalty of the form $I(f_j) = \|f_j\|_n + \lambda_n \Omega(f_j)$ for a seminorm $\Omega(\cdot)$. We can think of $\Omega(f_j)$ as a smoothness penalty for function f_j .

THEOREM 3 (Raskutti et al., 2009, Theorem 1). *Consider n independent and identically distributed samples from the sparse additive model $y_i = \sum_{j \in S} f_j^0(x_{ij}) + \varepsilon_i$ ($i = 1, \dots, n$), where $|S| = s \leq p/4$, $x_i \sim Q$, $\varepsilon_i \sim N(0, \sigma^2)$ and $f_j^0 \in \mathcal{F}$, with $\mathcal{F}$ being a class that satisfies the entropy condition $H(\delta, \mathcal{F}, Q) = A_0 \delta^{-1/m}$ for $m > 1/2$. Further assume that the covariates are independent, so $Q = \bigotimes_{j=1}^p Q_j$. Then, for a constant $C > 0$,*

$$\min_{(\hat{f}_j)_{j=1}^p} \max_{(f_j^0)_{j=1}^p \in \mathcal{F}} E \left(\left\| \sum_{j=1}^p \hat{f}_j - f_j^0 \right\|_Q^2 \right) \geq \max \left\{ \frac{\sigma^2 s \log(p/s)}{32n}, C s \left(\frac{\sigma^2}{n} \right)^{2m/(2m+1)} \right\},$$

where the minimum is over the set of all measurable functions.

We now state the first key result of this section, which establishes an oracle inequality for additive models, as well as slow rates of convergence.

THEOREM 4. *Assume the model $y_i = f^0(x_i) + \varepsilon_i$ ($i = 1, \dots, n$), with zero-mean ε_i satisfying $\max_{i \in \{1, \dots, n\}} L^2(E[\exp\{(\varepsilon_i/L)^2\}]) - 1 \leq \sigma_0^2$ for constants L and σ_0 . Assume that the entropy condition $H[\delta, \{f \in \mathcal{F} : \Omega(f) \leq 1\}, Q_n] \leq A_0 \delta^{-1/m}$ holds for $m > 1/2$, some function class $\mathcal{F}$ and some constant A_0 . Let $\rho_n = \kappa \max\{n^{-m/(2m+1)}, (\log p/n)^{1/2}\}$, where $\kappa = \kappa(A_0, m, L, \sigma_0)$ is a sufficiently large positive constant. Then for $\lambda_n \geq 4\rho_n$, with probability at least $1 - 2 \exp(-c_1 n \rho_n^2) - c_2 \exp(-c_3 n \rho_n^2)$, the estimator (17) satisfies*

$$\begin{aligned} & \|\hat{f} - f^0\|_n^2 + \lambda_n \sum_{j \in S^c} \|\hat{f}_j - f_j^*\|_n + \frac{3\lambda_n^2}{2} \sum_{j \in S} \Omega(\hat{f}_j - f_j^*) \\ & \leq 3\lambda_n \sum_{j \in S} \|\hat{f}_j - f_j^*\|_n + 4\lambda_n^2 \sum_{j \in S} \Omega(f_j^*) + \|f^* - f^0\|_n^2, \end{aligned}$$

where $c_1 = c_1(A_0, \sigma_0)$, $c_2 = c_2(A_0, m, L, \sigma_0)$ and $c_3 \geq 1/c_2^2$ are positive constants. Furthermore, if the function class $\mathcal{F}$ satisfies $\sup_{f \in \mathcal{F}} \|f\|_n \leq R$, we have

$$\|\hat{f} - f^0\|_n^2 \leq 2C_s \max \left\{ s n^{-m/(2m+1)}, s \left(\frac{\log p}{n} \right)^{1/2} \right\} + \|f^* - f^0\|_n^2,$$

where $C_s \geq 0$ depends on κ , R and $\sum_{j \in S} \Omega(f_j^*)/s$.

Theorem 4 requires the same assumptions as Theorem 2, namely sub-Gaussian noise ε_i and an entropy condition on the univariate function class $\mathcal{F}$. Its second part assumes a bound on the univariate class $\mathcal{F}$ in order to control the term $\|\hat{f}_j - f^*\|_n$. In Proposition 3 we drop this bounded class assumption and establish fast rates of convergence using the compatibility condition stated next.

DEFINITION 1 (Compatibility condition). *We say that the compatibility condition is met for the set $\tilde{S}$ if for some constant $\phi(\tilde{S}) > 0$ and for all additive $f = \sum_{j=1}^p f_j$ satisfying $\sum_{j \in \tilde{S}^c} \|f_j\|_n \leq 4 \sum_{j \in \tilde{S}} \|f_j\|_n$ one has $\sum_{j \in \tilde{S}} \|f_j\|_n \leq |\tilde{S}|^{1/2} \|f\|_n / \phi(\tilde{S})$.*

The above condition is a functional analogue of the compatibility condition used to prove oracle inequalities for the lasso (van de Geer & Bühlmann, 2009). Such conditions are common in the high-dimensional literature: for example, Meier et al. (2009) and van de Geer (2010) used a similar condition; recently Raskutti et al. (2012) and Yuan & Zhou (2015) used a functional version of the restricted eigenvalue condition for proving fast rates in additive models.

PROPOSITION 3. *Assume the conditions of Theorem 4 and the compatibility condition for $S = \{j : f_j^* \neq 0\}$. Then, with probability at least $1 - 2 \exp(-c_1 n \rho_n^2) - c_2 \exp(-c_3 n \rho_n^2)$,*

$$\frac{1}{2} \|\hat{f} - f^0\|_n^2 \leq C_f \max \left\{ s n^{-2m/(2m+1)}, \frac{s \log p}{n} \right\} + 2 \|f^* - f^0\|_n^2,$$

where $C_f \geq 0$ is a constant that depends on $\phi(S)$ and $\sum_{j \in S} \Omega(f_j^*)/s$.

Misspecifying the order m is especially important here since f_j can have different orders of smoothness. Using the same argument as in the univariate case, let m_j ($j = 1, \dots, p$) denote the smoothness order of the j th component; then our results are valid for any $m \leq \min_j m_j$. We end this section by specializing Theorem 4 to univariate regression.

PROPOSITION 4. *Assuming the conditions of Theorem 4 with $p = 1$, the compatibility condition holds trivially with $\phi(S) = 1$. Moreover, for a constant $C_f \geq 0$ that depends on $\Omega(f^*)$,*

$$\frac{1}{2} \|\hat{f} - f^0\|_n^2 \leq C_f n^{-2m/(2m+1)} + 2 \|f^* - f^0\|_n^2$$

with probability at least $1 - 2 \exp\{-c_1 \kappa n^{1/(2m+1)}\} - c_2 \exp\{-c_3 \kappa n^{1/(2m+1)}\}$.

5. SIMULATION STUDIES

5.1. Simulation for univariate regression

We begin with a simulation to compare the performance of our univariate framework with that of smoothing splines (Wahba, 1990) and of trend filtering (Kim et al., 2009; Tibshirani, 2014). Smoothing splines and trend filtering are implemented in the R packages `splines` and `genlasso` (Arnold & Tibshirani, 2014).

We generate the data as $y_i = f^0(x_i) + \varepsilon_i$ ($i = 1, \dots, n$) for different choices of the function f^0 and errors $\varepsilon_i \sim N(0, \sigma^2)$ with σ^2 chosen to attain a fixed signal-to-noise ratio, $\text{SNR} = (n-1)^{-1} \sum_{i=1}^n \{f^0(x_i)\}^2 / \sigma^2$. For this simulation we consider a fixed design with $x_i = i/n$ ($i = 1, \dots, n$). This facilitates comparison with trend filtering, which can slow substantially for

Fig. 3. Average mean square error over 100 simulated datasets, plotted as a function of degrees of freedom for true models given by $g_1, \dots, g_4$ in (18). The lines represent results for our framework with $m = 3$ (—■—) and $m = 1$ (—●—), trend filtering of order 1 (—■—), 2 (—●—) and 3 (—▲—), and smoothing splines (···■···).

random x_i , particularly when the covariates are not uniformly distributed over a closed interval. We consider $n = 150$, $\text{SNR} \in \{2, 3\}$, and four different functions f^0 :

$$\begin{aligned}
 g_1(x) &= -0.43 + 4.83x - 14.65x^2 + 11.76x^3, \\
 g_2(x) &= 0.23 - 8.44x + 45.20x^2 - 81.41x^3 + 46.59x^4, \\
 g_3(x) &= \exp(-5x + 0.5) - 0.4 \sinh(2.5), \\
 g_4(x) &= -\sin(7x - 0.4).
 \end{aligned} \tag{18}$$

We apply our proposal using a sequence of 100 λ -values linear on the log scale from $\lambda_{\max}$, for which $\hat{\beta} = 0$, down to $10^{-4}\lambda_{\max}$. We apply smoothing splines on a grid of 100 values of degrees of freedom from 10 to 1. Trend filtering is applied on a sequence of λ -values automatically selected by its R implementation with an average length of 279, 448 or 612 for trend filtering of order 1, 2 or 3, respectively. Figure 3 displays the mean square prediction error, $\text{MSE} = \|f^0 - \hat{f}\|_n^2$, of our method with $m = 1$ and 3, smoothing splines, and trend filtering of orders 1, 2 and 3, as a function of the degrees of freedom. Our proposal appears to outperform the competitors in terms of mean square prediction error, especially for polynomials. We observe comparable performance for the exponential and sine functions. This also provides empirical evidence for the theoretical results, where our method was proved to converge with rates comparable to smoothing splines. Since the functions considered in this simulation are smooth, as expected our method with $m = 1$ does not converge as fast as the competing methods.

We also compare the methods using an optimal tuning parameter that minimizes their prediction error on an independent test set of size 75. The ratios of MSE for each method to the MSE of our proposal, along with the corresponding ratios of degrees of freedom, are shown in Table 2.

Table 2. Average mean square errors of existing approaches relative to our univariate proposal with $m = 3$; a value greater than 1 indicates a lower corresponding value for our method. The results presented are averages over 100 datasets, with standard errors ($\times 100$) in parentheses

	Degree-3 polynomial		Degree-4 polynomial		Exponential function		Sine function	
	DF	MSE	DF	MSE	DF	MSE	DF	MSE
SS	1.32 (03)	1.48 (07)	1.25 (30)	1.58 (09)	1.13 (03)	1.21 (04)	1.10 (03)	1.00 (04)
TF-1	1.88 (13)	2.40 (30)	1.92 (12)	2.46 (30)	1.67 (16)	1.76 (31)	1.85 (13)	1.88 (24)
TF-2	1.30 (09)	1.61 (24)	1.42 (09)	1.81 (26)	1.34 (13)	1.48 (29)	1.20 (10)	1.30 (21)
TF-3	1.06 (12)	1.37 (32)	1.27 (11)	1.66 (31)	1.47 (16)	1.51 (35)	1.34 (12)	1.48 (25)

MSE, mean square prediction error; DF, degrees of freedom; SS, smoothing splines; TF-1, first-order trend filter; TF-2, second-order trend filter; TF-3, third-order trend filter.

Fig. 4. Average mean square error over 100 simulated datasets, plotted as a function of degrees of freedom for our proposal with $m = 1$ ($\blacksquare$ — $\blacksquare$ —), 2 ($\bullet$ — $\bullet$ —) and 3 ($\blacktriangle$ — $\blacktriangle$ —), compared with the method of Ravikumar et al. (2009) with 3 ($- \blacksquare - \blacksquare -$), 5 ($- \bullet - \bullet -$), 8 ($- \blacktriangle - \blacktriangle -$), 10 ($\cdots \blacksquare \cdots$), 15 ($\cdots \bullet \cdots$) and 20 ($\cdots \blacktriangle \cdots$) basis functions: (a) polynomial basis functions; (b) natural spline basis functions.

5.2. Simulation for multivariate additive regression

Next, we compare the performance of our sparse additive estimator with the method of Ravikumar et al. (2009). Their method is implemented in the R package SAM (Zhao et al., 2014), which uses natural spline basis functions. For a fairer comparison, we also implement their method using a polynomial basis expansion. Due to a lack of R packages for the proposals of Meier et al. (2009) and Lou et al. (2016), we defer the comparison with these methods to future work.

We consider the simulation setting of Meier et al. (2009) with some modifications to incorporate high-dimensional data and a smaller signal-to-noise ratio. We generate $n = 200$ samples for $p = 500$ features. The data are generated as $y_i = 5f_1(x_{i1}) + 3f_2(x_{i2}) + 4f_3(x_{i3}) + 6f_4(x_{i4}) + \varepsilon_i$ ($i = 1, \dots, n$), where $\varepsilon_i \sim N(0, \sigma^2)$ with σ^2 such that $\text{SNR} = 3$, and

$$f_1(x) = x, \quad f_2(x) = (2x - 1)^2, \quad f_3 = 2 \sin(2\pi x) / \{2 - \sin(2\pi x)\},$$

$$f_4(x) = 0.1 \sin(2\pi x) + 0.2 \cos(2\pi x) + 0.3 \sin^2(2\pi x) + 0.4 \cos^3(2\pi x) + 0.5 \sin^3(2\pi x);$$

the covariates are independently drawn from $\text{Un}(0, 1)$. For $m = 1$, we use the parameterization (10). For $m = 2$ and 3, we use

$$\underset{\beta_j \in \mathbb{R}^{\tilde{K}}}{\text{minimize}} \quad \frac{1}{2} \left\| y - \sum_{j=1}^p \Psi_{\tilde{K}}^j \beta_j \right\|_n^2 + \gamma \lambda \sum_{j=1}^p \Omega_j(\beta_j) + (1 - \gamma) \lambda \sum_{j=1}^p \|\Psi_{\tilde{K}}^j \beta_j\|_n$$

Fig. 5. Analysis of the Parkinson's telemonitoring dataset: (a) low-dimensional results for our proposal (—), the simple truncation estimator (2) (·····), and our relaxed proposal (---); (b) high-dimensional results for our proposal (—■—), Ravikumar et al.'s proposal with 2 (—■—), 4 (—●—) and 8 (—▲—) basis functions, the lasso (··■··), and the elastic net (··●··); (c) results for the relaxed versions of all methods in panel (b).

with $\gamma = 0.01$ and 0.001 , respectively. All methods were fitted over a sequence of 50 λ -values, decreasing linearly on the log scale. We set the maximum number of basis functions $\tilde{K}$ to 20 for our estimator and use 3, 5, 8, 10, 15 and 20 basis functions for Ravikumar et al. (2009)'s proposal.

In terms of mean square prediction error, it is not surprising to observe in Fig. 4 superior performance of our method over that of Ravikumar et al. (2009)'s with polynomial basis functions. However, in the same figure, our method also seems to outperform the original proposal of Ravikumar et al. (2009) using natural splines. Overall, our proposal achieves the smallest mean square error for a substantial interval of values of degrees of freedom in both panels of Fig. 4.

6. ANALYSIS OF PARKINSON'S TELEMONITORING DATA

We apply our method to the Parkinson's telemonitoring dataset (Tsanas et al., 2010), obtained from the University of California Irvine Machine Learning Repository. The data consist of $n = 5875$ observations and $p = 18$ covariates, including 16 biomedical voice measurements, age, and time of reading. Our goal is to predict the motor Unified Parkinson's Disease Rating Scale score. Besides the analysis of the original dataset, we add 100 noise variables, uniformly generated from the unit interval, to study the sparsity properties of our proposal. We use two-thirds of the data as a training set and the remaining data as a test set, and average the results over 100 training-test splits.

We compare our additive framework to (2) with fixed truncation, $K_j = K$, in the low-dimensional setting and compare our sparse additive framework to that of Ravikumar et al. (2009), the lasso (Tibshirani, 1996), and the elastic net (Zou & Hastie, 2005) in the high-dimensional setting. For computational convenience, we fit our proposal using the conservative basis truncation (9) with $\tilde{K} = \lceil (2n/3)^{1/2} \rceil$, the square root of the number of training observations. In both settings, we fit our proposal with order $m = 3$. We fit Ravikumar et al. (2009)'s proposal with $K = 2, 4, 8$; for a fairer comparison, we use a polynomial basis expansion. Other values of K had comparable or worse performance and are not presented here. For each proposal, we also implement a relaxed version, refitting the selected nonzero coefficients via least squares. For a sequence of λ -values, or a sequence of K -values for (2), we calculate the mean square test error and the model parsimony, the total number of nonzero basis functions used. Lower values of model parsimony correspond to more sparsity in either the number of components or the truncation levels K_j . Figure 5 shows that our proposal outperforms competitors in the low- and high-dimensional settings in terms of mean square error for a sequence of fitted models. The relaxed version of Ravikumar et al. (2009)'s method achieves a lower test error but at the cost of reduced parsimony. The linear models, implemented by the lasso and elastic net, had a substantially higher test error than additive models.

ACKNOWLEDGEMENT

We thank the associate editor and two referees for insightful comments that have improved the paper. Shojaie was supported by the U.S. National Institutes of Health and the U.S. National Science Foundation, and Simon by the U.S. National Institutes of Health.

SUPPLEMENTARY MATERIAL

Supplementary material at *Biometrika* online includes additional figures, algorithm details, extension to binary responses, and proofs of the theoretical results in §4. The R package *HierBasis*, available at <https://github.com/asadharis/HierBasis>, implements the proposed methods.

REFERENCES

- ARNOLD, T. B. & TIBSHIRANI, R. J. (2014). *genlasso: Path Algorithm for Generalized Lasso Problems*. R package version 1.3.
- CANDÈS, E. J., WAKIN, M. B. & BOYD, S. P. (2008). Enhancing sparsity by reweighted ℓ_1 minimization. *J. Fourier Anal. Appl.* **14**, 877–905.
- ČENCOV, N. (1962). Evaluation of an unknown distribution density from observations. *Doklady* **3**, 1559–62.
- CHATTERJEE, S. (2014). Assumptionless consistency of the lasso. *arXiv*: 1303.5817v5.
- COMBETTES, P. L. & PESQUET, J. (2011). Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke & H. Wolkowicz, eds. New York: Springer, pp. 185–212.
- FRIEDMAN, J. H., HASTIE, T. J. & TIBSHIRANI, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* **33**, 1–22.
- HARIS, A., WITTEN, D. & SIMON, N. (2016). Convex modeling of interactions with strong heredity. *J. Comp. Graph. Statist.* **25**, 981–1004.
- HASTIE, T. J., TIBSHIRANI, R. J. & FRIEDMAN, J. H. (2009). *The Elements of Statistical Learning*. New York: Springer, 2nd ed.
- JENATTON, R., MAIRAL, J., BACH, F. R. & OBOZINSKI, G. R. (2010). Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. Association for Computing Machinery, pp. 487–94.
- KIM, S., KOH, K., BOYD, S. & GORINEVSKY, D. (2009). ℓ_1 trend filtering. *SIAM Rev.* **51**, 339–60.
- KOLTCHINSKII, V. & YUAN, M. (2010). Sparsity in multiple kernel learning. *Ann. Statist.* **38**, 3660–95.
- LOU, Y., BIEN, J., CARUANA, R. & GEHRKE, J. (2016). Sparse partially linear additive models. *J. Comp. Graph. Statist.* **25**, 1126–40.
- MAMMEN, E. & VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25**, 387–413.
- MEIER, L., VAN DE GEER, S. & BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37**, 3779–821.
- MEINSHAUSEN, N. (2007). Relaxed lasso. *Comp. Statist. Data Anal.* **52**, 374–93.
- NADARAYA, E. A. (1964). On estimating regression. *Theory Prob. Appl.* **9**, 141–2.
- PETERSEN, A., WITTEN, D. & SIMON, N. (2016). Fused lasso additive model. *J. Comp. Graph. Statist.* **25**, 1005–25.
- R DEVELOPMENT CORE TEAM (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, <http://www.R-project.org>.
- RASKUTTI, G., WAINWRIGHT, M. J. & YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13**, 389–427.
- RASKUTTI, G., YU, B. & WAINWRIGHT, M. J. (2009). Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*. Association for Computing Machinery, pp. 1563–70.
- RAUHUT, H. & WARD, R. (2016). Interpolation via weighted ℓ_1 minimization. *Appl. Comp. Harmon. Anal.* **40**, 321–51.
- RAVIKUMAR, P., LAFFERTY, J., LIU, H. & WASSERMAN, L. (2009). Sparse additive models. *J. R. Statist. Soc. B* **71**, 1009–30.
- SADHANALA, V. & TIBSHIRANI, R. J. (2018). Additive models with trend filtering. *arXiv*: 1702.05037v3.
- STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135–51.
- STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595–620.
- TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 267–88.
- TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42**, 285–323.
- TSANAS, A., LITTLE, M. A., MCSHARRY, P. E. & RAMIG, L. O. (2010). Accurate telemonitoring of Parkinson’s disease progression by noninvasive speech tests. *IEEE Trans. Biomed. Eng.* **57**, 884–93.
- VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge: Cambridge University Press.

- VAN DE GEER, S. (2010). The Lasso with within group structure. In *Institute of Mathematical Statistics Collections*, vol. 7. Beachwood, Ohio: Institute of Mathematical Statistics, pp. 235–44.
- VAN DE GEER, S. & BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.* **3**, 1360–92.
- WAHBA, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.
- WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā* **26**, 359–72.
- YANG, Y. & BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27**, 1564–99.
- YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B* **68**, 49–67.
- YUAN, M. & ZHOU, D.-X. (2015). Minimax optimal rates of estimation in high dimensional additive models: Universal phase transition. *arXiv*: 1503.02817.
- ZHAO, T., LI, X., LIU, H. & ROEDER, K. (2014). *SAM: Sparse Additive Modelling*. R package version 1.0.5.
- ZOU, H. & HASTIE, T. J. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67**, 301–20.

[Received on 17 April 2017. Editorial decision on 29 June 2018]