Towards More Flexible Models in High Dimensions

Asad Haris

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Ali Shojaie, Chair

Noah Simon

Johannes Lederer

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Towards More Flexible Models in High Dimensions

Asad Haris

Chair of the Supervisory Committee:
Associate Professor Ali Shojaie
Department of Biostatistics

Recently, technological advances have allowed us to gather large and high-dimensional data. In high-dimensional data, the number of variables measured on each subject is quite large, often larger than the number of subjects. Consequently, there is growing need for improved supervised learning methods. We consider the setting in which we have an outcome variable and p covariates measured for n subjects; our goal is to estimate the conditional relationship between covariates and outcome.

Fitting linear models to high-dimensional data has been extensively studied in the past two decades, and numerous methods have been proposed for this task, such as the lasso. On the other hand, more flexible or nonparametric modeling of high-dimensional data is relatively less studied. Desirable flexible models should be interpretable, computationally-efficient, and have theoretical guarantees. Existing literature fails to achieve these three properties simultaneously. In this dissertation, we extend the existing literature and address gaps within the literature. In Chapter 2, we present a general framework for fitting sparse interaction models. Our framework not only generalizes many existing methods, but allows us to build new estimators; we present two such novel estimators in Chapter 2. In Chapter 3, we develop a general framework for fitting sparse additive models; this framework encompasses state-of-the-art techniques for additive models. We develop an efficient algorithm for computation, and establish theoretical guarantees for our general framework. In Chapters

4 and 5, we develop two novel estimators for nonparametric regression and extend them to sparse additive models. The main appeal of these estimators is that the fitted models have a parsimonious representation; this facilitates interpretation of models. Using the general framework of Chapter 3, we derive efficient algorithms for the estimators of Chapters 4 and 5, and establish theoretical convergence rates.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Firstly, I would like to thank my dissertation advisor, Ali Shojaie. This dissertation would not be possible without his continuous guidance, encouragement and support. His mentorship not only molded our shared research work, but also facilitated professional growth as a statistcian. Ali challenged and pushed me to limits I never thought possible and for that I will always be grateful. I would like to thank Noah Simon who mentored me throughout my dissertation work. Noah was my dissertation co-advisor in all but name. His infectious enthusiam about statistical problems has been a tremendous positive force for me throughout my studies. I am extremely grateful for the advices and mentorship of other members of my dissertation committee members, Johannes Lederer and Carey Farquhar (Graduate School Reprsentative).

I would like to thank Daniela Witten, who offered me my first reseach assitant position. Daniela introduced me to the field of statistical learning, which led to a chapter of this dissertation and paved the way for future research in statistical learning. I would also like to thank Gary Chan, Ying Q. Chen and Susanne May. As their research assistant, I gained new perspectives and was able to explore problems beyond my research work.

Finally, I would like to thank Gitana Garofalo, cohort of the Biostatistics Entering Class of 2013 and members of Slab Lab, whose support and friendship carried me through these challenging, yet enriching five years.

I would also like to thank everyone, not mentioned by name, who have supported me throughout my studies.

DEDICATION

to my parents, Haris Aqil and Saira Haris

Chapter 1

INTRODUCTION

1.1 Motivation

Recent technological advances have substantially amplified our ability to collect *high-dimensional* data. In high-dimensional data the number of variables measured, p, is much larger than the number of samples, n. A canonical example is *gene expression profiling*: generally we measure the expression levels of thousands of genes for a small number of subjects. For example, the study by Burczynski et al. (2006) obtained expression levels of 22,000 genes (dimension, p) on 85 subjects (sample size, n). The ability to measure such large amounts of data has led to some exciting opportunities and results in many fields. For instance, Burczynski et al. (2006) used the data as a diagnostic tool for patients with irritable bowl disease (IBD). A diagnosis based on gene expression data can be a welcome alternative to existing, invasive procedures such as a biopsy.

While high-dimensional data offers new opportunities it also presents a new set of challenges due to the *curse-of-dimensionality* (Bellman, 1961). For conventional statistical techniques, the curse-of-dimensionality means that the information we can extract from a dataset decreases exponentially with dimension, p.

In this dissertaiton, we focus on the *supervised learning* problem. Our goal is to predict some outcome variable using a set of predictors or covariates; the set of predictors in high-dimensional data is quite large. To be precise, consider outcome data $y_1, \ldots, y_n \in \mathbb{R}$, and covariate data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$. For example, in Burczynski et al. (2006) y_i is a categorical variable of IBD type (ulcerative colitis vs. Crohn's disease), and $\boldsymbol{x}_i \in \mathbb{R}^{22{,}000}$ is the vector of

gene expression levels for subject i. The relationship between predictors and outcome is

$$\mathbb{E}[y_i|\boldsymbol{x}_i] = f(\boldsymbol{x}_i).$$

The goal of *supervised learning* is to estimate the unknown function, f. The *curse-of-dimensionality* states that accurately estimating f for large p requires an extremely large sample size, n. Additionally, with a large number of covariates computation can quickly become unwieldy and, interpreting models with thousands of variables can be unfeasible.

To address the curse-of-dimensionality, issues of computation, and interpretability, restrictive assumptions are commonly made in the literature. Two popular assumptions are *linearity* of f, and *sparsity* (i.e., f is a function of a small, but unknown, subset of covariates). To be precise, the assumption is that

$$f(\boldsymbol{x}) = x_1\beta_1 + \cdots + x_p\beta_p, \tag{1.1}$$

where most $\beta_j = 0$. Such assumptions not only address the curse-of-dimensionality, but also give a parsimonious model that is easy to interpret. In the past few decades, numerous methods have been proposed for fitting the restrictive linear model (1.1); these methods have been thoroughly studied. In contrast, there are far fewer proposals for fitting flexible models in high dimensions and few gaps in the literature of existing methods. In this dissertation work, we aim to bridge gaps in the literature and contribute to the growing body of work on flexible models for high-dimensional data.

In this dissertation, we consider two types of relaxations to the linear model (1.1):

1. Interaction models of the form

$$f(\boldsymbol{x}) = x_1\beta_1 + \cdots + x_p\beta_p + x_1x_2\beta_{1,2} + \cdots + x_{p-1}x_p\beta_{p-1,p}. \tag{1.2}$$

2. additive models of the form

$$f(\boldsymbol{x}) = f_1(x_1) + \cdots + f_p(x_p). \tag{1.3}$$

In Sections 1.2 and 1.3, we briefly motivate our work for fitting interaction and additive models in high dimensions, respectively. We also summarize the contributions of each chapter.

1.2 Flexibility by interaction models

In low dimensions, fitting the interaction model (1.2) is done by the usual least squares:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \tag{1.4}$$

where $\boldsymbol{y}$ is the response vector, and $\boldsymbol{X}$ is the design matrix with all main effects and interactions. Thus, the usual machinery for linear model easily extends to interaction models. However, in high dimensions, we know that least squares overfits the data; a popular solution for fitting linear models in high dimensions is the lasso which solves

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{\beta}\|_1, \tag{1.5}$$

where the ℓ_1 penalty induces sparsity, i.e., the fitted model fits many $\beta_j = 0$. Unlike the low-dimensional case, the linear model machinery does not extend to interaction models because for interaction models we usually wish to impose *strong heredity*. Strong heredity is a property whereby if an interaction term is included in the model, then both main effects terms must also be present. Similarly the so-called *weak heredity* is a property whereby if an interaction term is included in the model, then at least on of the corresponding main effects must also be present. Heredity constraints are desirable because they simplify interpretation (McCullagh, 1984) and experimental design (Bien et al., 2013). Recently, a number of penalized regression approaches have been proposed for fitting interaction models with

strong or weak heredity (Choi et al., 2010; Jenatton et al., 2011; Zhao et al., 2009; Bach et al., 2012; Radchenko and James, 2010; Lim and Hastie, 2013; Bien et al., 2013). However, there are some gaps in the literature. For instance, the proposed algorithm of Bien and Tibshirani (2014) is not computationally efficient, particularly for large p. On the other hand, the algorithm of Radchenko and James (2010) is not guaranteed to converge to the global optimum.

In Chapter 2 of this dissertation, we propose a general framework for fitting interaction models which obey strong heredity. Our general framework encompasses a wide class of estimators including the proposals of Bien and Tibshirani (2014) and Radchenko and James (2010). We develop an alternating directions method of multipliers (ADMM) algorithm for estimators within our framework. This algorithm has guaranteed convergence to the global optimum, can be easily specialized to any convex penalty function of interest, and allows for a straightforward extension to the setting of generalized linear models. We also present an estimator for the degrees of freedom and discuss extensions of our proposal to enforce weak heredity. Additionally, our framework facilitates the development of new methods for fitting interaction models with strong heredity. We present two novel proposals not previously studied in the literature and we compare our novel approaches to existing methods in the literature on simulated and HIV-1 data.

1.3 Flexibility by additive models

Additive models are fairly popular in the nonparametric literature and have been studied as early as 1980s. They offer a substantially more flexible alternative to linear models, while retaining some of the desirable interpretability of linear models. Additionally, additive models do not suffer from the curse of dimensionality. However, in high dimensions, traditional methods are no longer feasible due to over-fitting of the data. In the past decade, a number of solutions have been proposed for the high-dimensional case (Ravikumar et al., 2009; Meier et al., 2009; Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Yuan and Zhou, 2015; Lou et al., 2016; Petersen et al., 2016; Sadhanala and Tibshirani, 2017). Despite these

developments, some papers in the literature fail to establish fast convergence rates for the estimators whereas other papers fail to develop efficient algorithms for computation. Our work in Chapter 3 aims to bridge these gaps.

In Chapter 3, we present a general framework for fitting sparse additive models for a potentially large number of covariates. Our proposal not only encompasses the above mentioned proposals, but also includes parametric methods such as the lasso (Tibshirani, 1996). We present an efficient computational algorithm that easily scales to thousands of observations and features. We prove minimax optimal convergence bounds on these estimators under a weak *compatibility condition*. In addition, we characterize the rate of convergence when this compatibility condition is not met. We complement our theoretical results with empirical studies comparing some exiting methods.

In Chapters 4 and 5, we present two novel techniques for fitting additive models. The proposal of Chapter 4 uses a basis expansion for each component function, where the order of expansion for each component is selected data-adaptively. The resulting model is not only flexible but is easy to interpret and has a parsimonious representation. The proposal of Chapter 5 uses a *wavelet* based approach. Wavelets are a popular system of basis functions for representing functions and have been extensively used in the nonparametric regression literature (Antoniadis, 1997). However, extending wavelet methods to additive models is challenging. We present a novel technique for extension of wavelets to additive and sparse additive models for high-dimensional data. We establish convergence rates for the estimators of Chapters 4 and 5 and develop efficient algorithms. We compare the predictive performance of these proposals to existing methods in empirical studies.

1.4 Notations and conventions

In this section, we introduce some notation used throughout this dissertation.

Except where defined otherwise, we denote vectors in lower case bold font (e.g., $\boldsymbol{y}$, $\boldsymbol{\beta}$), and matrices in upper case bold font (e.g., $\boldsymbol{X}, \boldsymbol{\Psi}$). Scalar elements of a vector or matrix are denoted by regular font, e.g., the i-th element of vector $\boldsymbol{y}$ will be denote by y_i, similarly the

(i, j)-th element of matrix $\boldsymbol{X}$ will be denoted by $X_{i,j}$.

Except where noted otherwise, we consider the setting with data $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$, where $y_i \in \mathbb{R}$ is the response variable and $\boldsymbol{x}_i \in \mathbb{R}^p$ is the covariate vector. For an n-vector $\boldsymbol{r}$, we denote by $\|\boldsymbol{r}\|_n$ the *empirical norm* where $\|\boldsymbol{r}\|_n^2 = n^{-1} \sum_{i=1}^n r_i^2$. We similarly define the empirical norm for a function f given data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ by $\|f\|_n^2 = n^{-1} \sum_{i=1}^n (f(\boldsymbol{x}_i))^2$; we also use the short-hand notation

$$\|\boldsymbol{r} - f\|_n^2 = \frac{1}{n} \sum_{i=1}^n (r_i - f(\boldsymbol{x}_i))^2.$$

For a vector $\boldsymbol{x} \in \mathbb{R}^p$, we study additive functions of the form $f(\boldsymbol{x}) = f_1(x_1) + \cdots + f_p(x_p)$; for convenience we use the short-hand notation $f = \sum_{j=1}^p f_j$. We similarly define the $\|\cdot\|_n$ notation for components of additive models, i.e.,

$$\|f_j\|_n^2 = \frac{1}{n} \sum_{i=1}^n (f_j(x_{ij}))^2, \quad \|\boldsymbol{r} - f_j\|_n^2 = \frac{1}{n} \sum_{i=1}^n (r_i - f_j(x_{ij}))^2. \tag{1.6}$$

Finally, we use the following short-hand notation: $A \asymp B \Leftrightarrow A = cB$ and $A \lesssim B \Leftrightarrow A \leq cB$, for some constant c.

Chapter 2

CONVEX MODELING OF INTERACTIONS WITH STRONG HEREDITY

2.1 Introduction

2.1.1 Modeling interactions

In this chapter, we model a response variable with a set of main effects and second-order interactions. The problem can be formulated as follows: we are given a response vector $\boldsymbol{y}$ for n observations, an $n \times p_1$ matrix $\boldsymbol{X}$ of covariates and another $n \times p_2$ matrix $\boldsymbol{Z}$ of covariates. In what follows, the notation $\boldsymbol{X}_{\cdot,j}$ and $\boldsymbol{Z}_{\cdot,k}$ will denote the j-th column of $\boldsymbol{X}$ and k-th column of $\boldsymbol{Z}$, respectively. The goal is to fit the model

$$y_i = B_{0,0} + \sum_{j=1}^{p_1} B_{j,0} X_{i,j} + \sum_{k=1}^{p_2} B_{0,k} Z_{i,k} + \sum_{j=1}^{p_1} \sum_{k=1}^{p_2} B_{j,k} X_{i,j} Z_{i,k} + \varepsilon_i \quad (i = 1, \ldots, n), \tag{2.1}$$

where $\boldsymbol{B}$ is a $(p_1 + 1) \times (p_2 + 1)$ matrix of coefficients, of which the rows and columns are indexed from 0 to p_1 and 0 to p_2 for the variables $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively. In the special case where $\boldsymbol{X} = \boldsymbol{Z}$, the coefficient of the (j, k)-th interaction is $B_{j,k} + B_{k,j}$, and the coefficient of the j-th main effect is $B_{0,j} + B_{j,0}$.

For brevity, we re-write model (2.1) using array notation. We construct the $n \times (p_1 + 1) \times (p_2 + 1)$ array $\boldsymbol{W}$ as follows: for $i \in \{1, \ldots, n\}, j \in \{0, \ldots, p_1\}, k \in \{0, \ldots, p_2\}$,

$$W_{i,j,k} = \begin{cases} X_{i,j}Z_{i,k} & \text{for } j \neq 0 \text{ and } k \neq 0 \\ X_{i,j} & \text{for } k = 0 \text{ and } j \neq 0 \\ Z_{i,k} & \text{for } j = 0 \text{ and } k \neq 0 \\ 1 & \text{for } j = k = 0 \end{cases}. \tag{2.2}$$

Then (2.1) is equivalent to the model

$$\boldsymbol{y} = \boldsymbol{W} * \boldsymbol{B} + \boldsymbol{\varepsilon}, \tag{2.3}$$

where $\boldsymbol{B}$ is the matrix of coefficients as in (2.1), and $\boldsymbol{W} * \boldsymbol{B}$ denotes the n-vector whose i-th element takes the form $(\boldsymbol{W} * \boldsymbol{B})_i \equiv \sum_{j=0}^{p_1} \sum_{k=0}^{p_2} W_{i,j,k} B_{j,k}$. The model is displayed in the left panel of Figure 2.1.

In fitting models with interactions, we may wish to impose either *strong* or *weak* heredity (Hamada and Wu, 1992; Yates, 1978; Chipman, 1996; Joseph, 2006), defined as follows:

> *Strong Heredity:* If an interaction term is included in the model, then both of the corresponding main effects must be present. That is, if $B_{j,k} \neq 0$, then $B_{j,0} \neq 0$ *and* $B_{0,k} \neq 0$.

> *Weak Heredity:* If an interaction term is included in the model, then at least one of the corresponding main effects must be present. That is, if $B_{j,k} \neq 0$, then either $B_{j,0} \neq 0$ *or* $B_{0,k} \neq 0$.

Such constraints facilitate model interpretation (McCullagh, 1984), improve statistical power (Cox, 1984), and simplify experimental designs (Bien et al., 2013). In this chapter we propose a general convex regularized regression approach which naturally and efficiently enforces strong heredity.

2.1.2 Summary of previous work

A number of authors have considered the task of fitting interaction models under strong or weak heredity constraints. Constraints to enforce heredity (Peixoto, 1987; Friedman, 1991; Bickel et al., 2010; Park and Hastie, 2008; Wu et al., 2010) have been applied to conventional step-wise model selection techniques (Montgomery et al., 2012, chap. 10). Chipman (1996) and George and McCulloch (1993) proposed Bayesian methods. In more recent work, Hao and Zhang (2014) proposed `iFORM`, an approach that performs forward selection on the main effects, and allows interactions into the model once the main effects have already been selected. `iFORM` has a number of attractive properties, including suitability for the ultra-high-dimensional setting, computational efficiency, as well as proven theoretical guarantees.

In this chapter, we take a regularization approach to inducing strong heredity. A number of regularization approaches for this task have already been proposed in the literature; in fact, a strength of our proposal is that it provides a unified framework (and associated algorithm) of which several existing approaches can be seen as special cases. Choi et al. (2010) propose a non-convex approach, which amounts to a lasso (Tibshirani, 1996) problem with re-parametrized coefficients. Alternatively, some authors have enforced strong or weak heredity via convex penalties or constraints. Jenatton et al. (2011) and Zhao et al. (2009) describe a set of penalties that can be applied to a broad class of problems. As a special case they consider interaction models with strong or weak heredity; this has been further developed by Bach et al. (2012). Radchenko and James (2010), Lim and Hastie (2013) and Bien et al. (2013) propose penalties specifically designed for interaction models with sparsity and strong heredity. We now describe the latter two approaches in greater detail.

2.1.2.1 `hierNet` (Bien et al., 2013)

The `hierNet` approach of Bien et al. (2013) fits the model (2.1) with $\boldsymbol{X} = \boldsymbol{Z}$ and $p_1 = p_2 = p$. In the case of strong heredity, using the notation of (2.3), they consider the problem

$$\begin{aligned} \underset{\boldsymbol{B}\in\mathbb{R}^{(p+1)\times(p+1)},\ \boldsymbol{\beta}^{\pm}\in\mathbb{R}^p}{\text{minimize}} \quad & \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{W} * \boldsymbol{B}\right\|_2^2 + \lambda\sum_{j=1}^{p}(\beta_j^+ + \beta_j^-) + \frac{\lambda}{2}\|\boldsymbol{B}_{-0,-0}\|_1 \\ \text{subject to } & \boldsymbol{B} = \boldsymbol{B}^\top,\ \boldsymbol{B}_{0,-0} = \boldsymbol{\beta}^+ - \boldsymbol{\beta}^- \\ & \|\boldsymbol{B}_{j,-0}\|_1 \le \beta_j^+ + \beta_j^-,\ \beta_j^+ \ge 0, \beta_j^- \ge 0\ (j = 1,\ldots,p). \end{aligned} \tag{2.4}$$

Using this notation, the coefficient for the j-th main effect is $B_{0,j} + B_{j,0}$, and the coefficient for the (j,k)-th interaction is $B_{j,k} + B_{k,j}$. Strong heredity is imposed by the constraint $\|\boldsymbol{B}_{j,-0}\|_1 \le \beta_j^+ + \beta_j^-$.

2.1.2.2 `glinternet` (Lim and Hastie, 2013)

Like `hierNet`, the `glinternet` proposal of Lim and Hastie (2013) fits (2.1) with $\boldsymbol{X} = \boldsymbol{Z}$ and $p_1 = p_2 = p$. In order to describe this approach, we introduce some additional notation. Let α_k be the coefficient of the k-th main effect. We decompose α_k into p parameters, i.e. $\alpha_k = \alpha_k^{(0)} + \alpha_k^{(1)} + \ldots + \alpha_k^{(k-1)} + \alpha_k^{(k+1)} + \ldots + \alpha_k^{(p)}$. We let $\alpha_{jk} + \alpha_{kj}$ denote the coefficient for the interaction between X_j and X_k. Lim and Hastie (2013) propose to solve the optimization problem

$$\begin{aligned} \underset{\substack{\alpha_0,\,\{\alpha_{ij}\}_{i\neq j; i,j\neq 0},\\ \{\alpha_i^{(j)}\}_{j\neq i}\in\mathbb{R}}}{\text{minimize}} \quad & \frac{1}{2}\left\|\boldsymbol{y} - \alpha_0\boldsymbol{1} - \sum_{k=1}^{p}\sum_{j\neq k}\alpha_k^{(j)}\boldsymbol{X}_{\cdot,k} - \sum_{j\neq k}\alpha_{jk}\left(\boldsymbol{X}_{\cdot,j} * \boldsymbol{X}_{\cdot,k}\right)\right\|_2^2 \\ & + \lambda\left(\sum_{j=1}^{p}|\alpha_j^{(0)}| + \sum_{j\neq k}\sqrt{\left(\alpha_j^{(k)}\right)^2 + \left(\alpha_k^{(j)}\right)^2 + \alpha_{jk}^2}\right), \end{aligned} \tag{2.5}$$

where $\boldsymbol{X}_{\cdot,j} * \boldsymbol{X}_{\cdot,k}$ denotes element-wise multiplication. Strong heredity is enforced via the group lasso (Yuan and Lin, 2006) penalties: if either α_{jk} or α_{kj} is estimated as non-zero,

then $\alpha_j^{(k)}$ and $\alpha_k^{(j)}$ will be estimated to be non-zero, and hence so will α_j and α_k.

2.1.3 Organization of chapter

The rest of this chapter is organized as follows. In Section 2.2, we provide details of FAMILY, our proposed approach for modeling interactions. An unbiased estimator for its degrees of freedom is in Section 2.3, and an extension to weak heredity is in Section 2.4. We explore FAMILY's empirical performance in simulation in Section 2.5, and in an application to an HIV data set in Section 2.6. The Discussion is in Section 2.7.

2.2 Modeling interactions with FAMILY

We propose a *framework for modeling interactions with a convex penalty* (FAMILY). The FAMILY approach is the solution to a convex optimization problem, which (using the notation of Section 2.1.1) takes the form

$$\underset{\boldsymbol{B}\in\mathbb{R}^{(p_1+1)\times(p_2+1)}}{\text{minimize}} \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{W}*\boldsymbol{B}\right\|_n^2 + \lambda_1\sum_{j=1}^{p_1} P_r(\boldsymbol{B}_{j,\cdot}) + \lambda_2\sum_{k=1}^{p_2} P_c(\boldsymbol{B}_{\cdot,k}) + \lambda_3\|\boldsymbol{B}_{-0,-0}\|_1. \quad (2.6)$$

Here, λ_1, λ_2, and λ_3 are non-negative tuning parameters. P_r and P_c are convex penalty functions on the rows and columns of the coefficient matrix $\boldsymbol{B}$. The $\|\boldsymbol{B}_{-0,-0}\|_1$ term denotes the element-wise ℓ_1-norm on the interactions, which enforces sparsity on the interaction coefficients when λ_3 is large. The right panel of Figure 2.1 demonstrates the action of each penalty on the matrix $\boldsymbol{B}$.

As we will see, the choice of P_r and P_c will determine the type of structure (such as strong heredity) enforced on the fitted model. In the examples that follow, we take $P_r = P_c$; however, in principle, these two penalty functions need not be equal. For instance, if the features in Z are known to be of scientific importance, we might choose to perform feature selection on the main effects of X only. In this case, we might choose to use $P_r(\boldsymbol{b}) = \|\boldsymbol{b}\|_2$ and $P_c(\boldsymbol{b}) = 0$.

We suggest standardizing the columns of $\boldsymbol{X}$ and $\boldsymbol{Z}$ to have mean zero and variance one

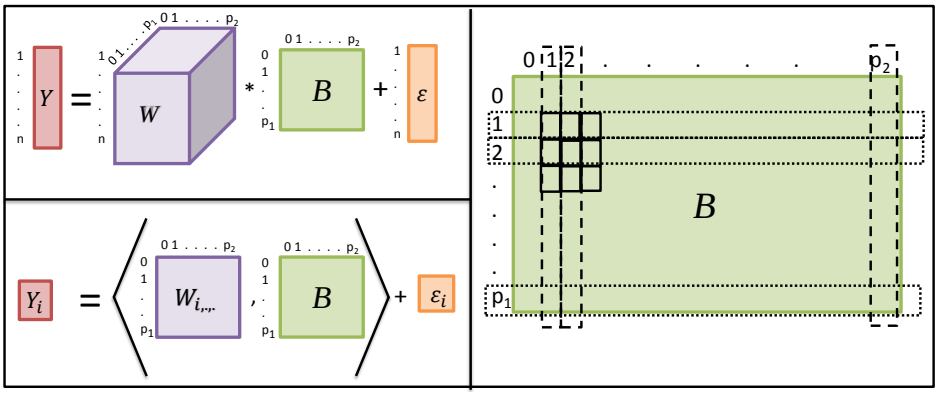


Figure 2.1: *Left:* The model (2.1), for all n observations *(top)* and for the i-th observation *(bottom)*. The notation $\langle \boldsymbol{W}_{i,\cdot,\cdot}, \boldsymbol{B} \rangle$ denotes the inner product, $\sum_{j,k} W_{i,j,k} B_{j,k}$. *Right:* In (2.6), the $(p_1+1) \times (p_2+1)$ coefficient matrix $\boldsymbol{B}$ is penalized by applying the P_r and P_c penalties to each of the p_1 rows (·····) and each of the p_2 columns (- - - - -), respectively. The ℓ_1 penalty is applied to each of the $p_1 p_2$ interactions (———).

before solving (2.6), in order to ensure that the main effects and interactions are on the same scale, as is standard practice for penalized regression estimators (Hastie et al., 2009). We take this approach in Sections 2.5 and 2.6.

2.2.1 Connections to lasso (Tibshirani, 1996)

The *main effects lasso* can be viewed as a special case of (2.6) where P_c and P_r are ℓ_1 penalties,

$$\underset{\boldsymbol{B} \in \mathbb{R}^{(p_1+1)\times(p_2+1)}}{\text{minimize}} \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{W} * \boldsymbol{B} \right\|_n^2 + \lambda_1 \sum_{j=1}^{p_1} \|\boldsymbol{B}_{j,\cdot}\|_1 + \lambda_2 \sum_{k=1}^{p_2} \|\boldsymbol{B}_{\cdot,k}\|_1 + \lambda_3 \|\boldsymbol{B}_{-0,-0}\|_1, \tag{2.7}$$

and where λ_3 is chosen sufficiently large as to shrink all of the interaction terms to 0. In this case, the lasso penalties on the rows and columns are applied only to the main effects.

In contrast, if we take $\lambda_3 = 0$, $\lambda_1 = \lambda_2 = \lambda$, and $P_c(\boldsymbol{b}) = P_r(\boldsymbol{b}) = |b_1| + 1/2 \|\boldsymbol{b}_{-1}\|_1$, where $\boldsymbol{b} = (b_1, \boldsymbol{b}_{-1}^\top)^\top$, then (2.6) yields the *all-pairs lasso*, which applies a lasso penalty to all main

effects and all interactions. In this case, (2.6) can be re-written more simply as

$$\underset{\boldsymbol{B}\in\mathbb{R}^{(p_1+1)\times(p_2+1)}}{\text{minimize}}\ \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{W}*\boldsymbol{B}\right\|_n^2+\lambda\|\boldsymbol{B}\|_1. \tag{2.8}$$

However, our main interest in this chapter is to develop a convex framework for modeling interactions that obeys strong heredity. Clearly, the all-pairs lasso does not satisfy strong heredity, and the main effects lasso does so only in a trivial way (by setting all interaction coefficient estimates to zero).

2.2.2 `FAMILY` *with strong heredity*

We now consider three choices of P_r and P_c in (2.6) that yield an estimator that obeys strong heredity. In Section 2.2.2.1, we consider the case where P_r and P_c are group lasso penalties. In Section 2.2.2.2, we consider the case where they are ℓ_∞ penalties. We consider a hybrid between an ℓ_1 and an ℓ_∞ norm in Section 2.2.2.3. The unit norm balls corresponding to these three penalties are displayed in Figure 2.2.

2.2.2.1 `FAMILY` *with an* ℓ_2 *penalty*

We first consider (2.6) in the case where $P_r(\boldsymbol{b}) = P_c(\boldsymbol{b}) = \|\boldsymbol{b}\|_2$, which we will refer to as `FAMILY.12`. The resulting optimization problem takes the form

$$\underset{\boldsymbol{B}\in\mathbb{R}^{(p_1+1)\times(p_2+1)}}{\text{minimize}}\ \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{W}*\boldsymbol{B}\right\|_n^2+\lambda_1\sum_{j=1}^{p_1}\|\boldsymbol{B}_{j,.}\|_2+\lambda_2\sum_{k=1}^{p_2}\|\boldsymbol{B}_{.,k}\|_2+\lambda_3\|\boldsymbol{B}_{-0,-0}\|_1. \tag{2.9}$$

This formulation will induce strong heredity, in the sense that an interaction between X_j and X_k can have a non-zero coefficient estimate only if both of the corresponding main effects are non-zero.

Problem 2.9 is closely related to `VANISH`, an approach for non-linear interaction modeling (Radchenko and James, 2010). In fact, if we take $\boldsymbol{X} = \boldsymbol{Z}$ and assume that all main effects and interactions are scaled to have norm one in (2.9), and consider the case of `VANISH` with

only linear main effects and interactions, then `VANISH` and (2.9) coincide exactly.

Radchenko and James (2010) attempt to solve the `VANISH` optimization problem via block coordinate descent. However, due to non-separability of the groups, their algorithm is not guaranteed convergence to the global optimum. In contrast, the algorithm in Section 2.2.3 is guaranteed convergence to the global optimum of (2.6) for any convex penalty, and can be extended to the case of generalized linear models.

2.2.2.2 `FAMILY` *with an* ℓ_∞ *penalty*

We now consider (2.6) in the case where $P_r(\boldsymbol{b}) = P_c(\boldsymbol{b}) = \|\boldsymbol{b}\|_\infty$; we refer to this in what follows as `FAMILY.linf`. We refer the reader to Duchi and Singer (2009) for a discussion of the properties of the ℓ_∞ norm, and its merits relative to the ℓ_2 norm in inducing group sparsity. In this case, (2.6) takes the form

$$\underset{\boldsymbol{B}\in\mathbb{R}^{(p_1+1)\times(p_2+1)}}{\text{minimize}} \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{W} * \boldsymbol{B}\right\|_n^2 + \lambda_1 \sum_{j=1}^{p_1} \|\boldsymbol{B}_{j,\cdot}\|_\infty + \lambda_2 \sum_{k=1}^{p_2} \|\boldsymbol{B}_{\cdot,k}\|_\infty + \lambda_3 \|\boldsymbol{B}_{-0,-0}\|_1. \quad (2.10)$$

This formulation also induces strong heredity.

2.2.2.3 `FAMILY` *with a hybrid* ℓ_1/ℓ_∞ *penalty*

Finally, we consider (2.6) with $P_r(\boldsymbol{b}) = P_c(\boldsymbol{b}) = \max(|b_1|, \|\boldsymbol{b}_{-1}\|_1)$. In this case, (2.6) takes the form

$$\begin{aligned}\underset{\boldsymbol{B}\in\mathbb{R}^{(p_1+1)\times(p_2+1)}}{\text{minimize}} \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{W} * \boldsymbol{B}\right\|_n^2 + \lambda_3 \|\boldsymbol{B}_{-0,-0}\|_1 + \\ \lambda_1 \sum_{j=1}^{p_1} \max(|B_{j,0}|, \|\boldsymbol{B}_{j,-0}\|_1) + \lambda_2 \sum_{k=1}^{p_2} \max(|B_{0,k}|, \|\boldsymbol{B}_{-0,k}\|_1).\end{aligned} \quad (2.11)$$

In the special case where $\boldsymbol{X} = \boldsymbol{Z}$, $\lambda_1 = \lambda_2 = \lambda$, and $\lambda_3 = \lambda/2$, (2.11) is in fact equivalent to the `hierNet` proposal of Bien et al. (2013). Details of this equivalence are given in Bien et al. (2013).

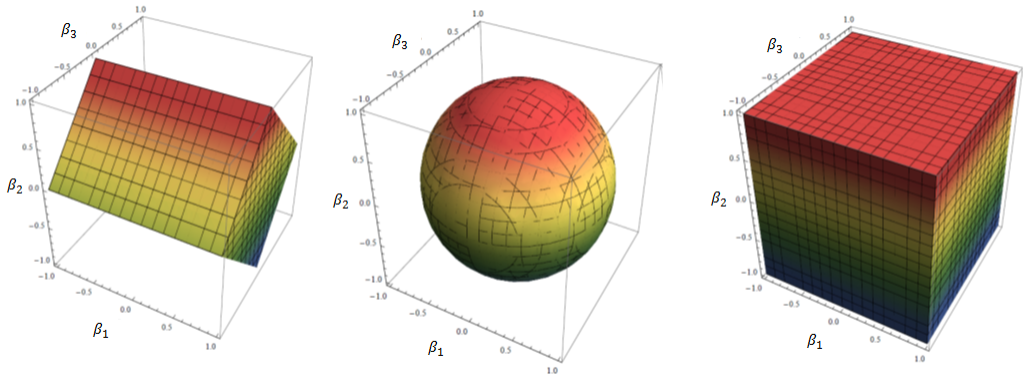


Figure 2.2: A graphical representation of the region $P(\boldsymbol{\beta}) \leq 1$, where $P(\boldsymbol{\beta}) = \max\left(|\beta_1|, |\beta_2| + |\beta_3|\right)$ *(left)*; $P(\boldsymbol{\beta}) = \sqrt{\beta_1^2 + \beta_2^2 + \beta_3^2}$ *(center)*; or $P(\boldsymbol{\beta}) = \max(|\beta_1|, |\beta_2|, |\beta_3|)$ *(right)*.

Bien et al. (2013) propose to solve `hierNet` via an ADMM algorithm which applies a generalized gradient descent loop within each update. This leads to computational inefficiency, especially for large p. In Section 2.2.3, we propose a simple, stand-alone ADMM algorithm for solving (2.6), which can be easily applied to solve (2.11), and consequently also the `hierNet` optimization problem.

Given its connection to Bien et al. (2013), we refer to (2.11) as `FAMILY.hierNet`.

2.2.2.4 Dual norms

Here we further consider the l_2, l_∞ and l_1/l_∞ hybrid penalties discussed in Sections 2.2.2.1-2.2.2.3. For an arbitrary penalty, the proximal operator is the solution to the optimization problem

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\beta}\|_2^2 + \lambda P(\boldsymbol{\beta}). \tag{2.12}$$

We begin by presenting a well-known lemma (see e.g. Proposition 1.1, Bach et al. (2011)).

Lemma 2.1. *Let $P(\boldsymbol{y})$ be a norm of $\boldsymbol{y}$ with dual norm $P_*(\boldsymbol{y}) \equiv \max_{\boldsymbol{z}} \{\boldsymbol{z}^\top \boldsymbol{y} : P(\boldsymbol{z}) \leq 1\}$. Then $\widehat{\boldsymbol{\beta}} = 0$ solves (2.12) if and only if $P_*(\boldsymbol{y}) \leq \lambda$.*

It is well-known that the ℓ_2 norm is its own dual norm, and that the ℓ_1 norm is dual to

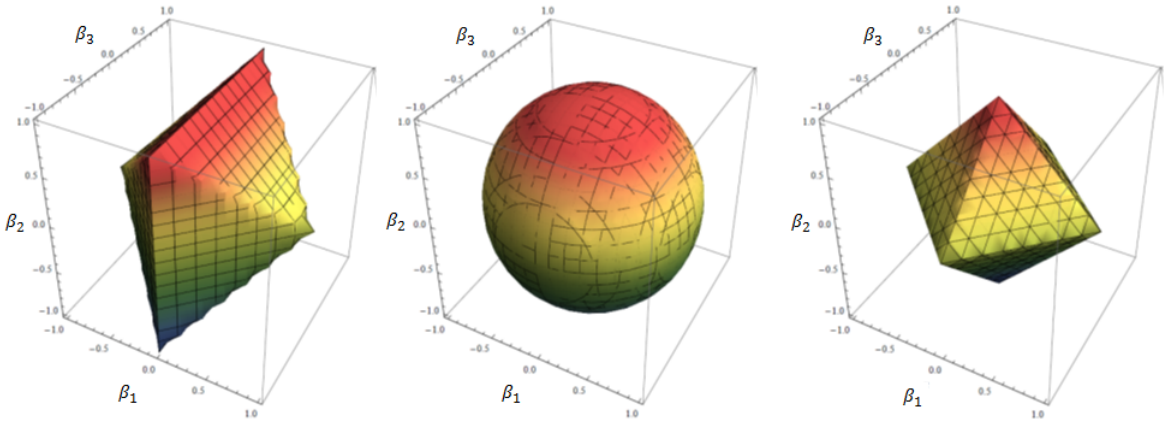


Figure 2.3: A graphical representation of the region $P_*(\beta) \leq 1$, where $P_*(\beta)$ is the dual norm for $P(\beta) = \max(|\beta_1|, |\beta_2| + |\beta_3|)$ *(left)*; $P(\beta) = \sqrt{\beta_1^2 + \beta_2^2 + \beta_3^2}$ *(center)*; or $P(\beta) = \max(|\beta_1|, |\beta_2|, |\beta_3|)$ *(right)*.

the ℓ_∞ norm. We now derive the dual norm for the `FAMILY.hierNet` penalty. This lemma is proven in Appendix A.2.

Lemma 2.2. *The dual norm of* $P(\boldsymbol{\beta}) = \max\{|\beta_1|, \|\boldsymbol{\beta}_{-1}\|_1\}$ *takes the form*

$$P_*(\boldsymbol{\beta}) = |\beta_1| + \|\boldsymbol{\beta}_{-1}\|_\infty. \tag{2.13}$$

Lemmas 2.1 and 2.2 provide insight into the values of $\boldsymbol{y}$ for which all variables are shrunken to zero in (2.12). The dual norm balls for the hybrid ℓ_1/ℓ_∞, ℓ_2, and ℓ_∞ norms are displayed in Figure 2.3. By Lemma 2.1, any y inside the dual norm ball leads to a zero solution of (2.12). For the hybrid ℓ_1/ℓ_∞ norm, the shape of the dual norm ball implies that the first element of $\boldsymbol{y}$ plays an outsize role in whether or not the coefficient vector is shrunken to zero. Consequently, the main effects play a larger role than the interactions in determining whether sparsity is induced. In contrast, for the ℓ_∞ and ℓ_2 norms, the main effect and interactions play an equal role in determining whether the coefficients are shrunken to zero.

2.2.3 *Algorithm for solving* `FAMILY`

A step-by-step ADMM algorithm for solving `FAMILY` is provided in Appendix A.1.2. Here, we present an overview of this algorithm. A gentle introduction to ADMM is provided in Appendix A.1.1.

2.2.3.1 *ADMM algorithm for solving* `FAMILY`

We now develop an ADMM algorithm to solve (2.6). We define the variable $\boldsymbol{\Theta} = (\boldsymbol{D}|\boldsymbol{E}|\boldsymbol{F})$, with $\boldsymbol{D}, \boldsymbol{E}, \boldsymbol{F} \in \mathbb{R}^{(p_1+1)\times(p_2+1)}$. That is, $\boldsymbol{\Theta}$ is a $(p_1+1)\times 3(p_2+1)$ matrix, which we partition into $\boldsymbol{D}$, $\boldsymbol{E}$, and $\boldsymbol{F}$ for convenience. Then (2.6) can be re-written as

$$\begin{aligned} &\underset{\boldsymbol{B}\in\mathbb{R}^{(p_1+1)\times(p_2+1)},}{\text{minimize}} \quad \left\{\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{W}*\boldsymbol{B}\|_n^2 + \lambda_1\sum_{j=1}^{p_1} P_r(\boldsymbol{D}_{j,.}) + \lambda_2\sum_{k=1}^{p_2} P_c(\boldsymbol{E}_{.,k}) + \lambda_3\|\boldsymbol{F}_{-0,-0}\|_1\right\} \\ &\boldsymbol{\Theta}\in\mathbb{R}^{(p_1+1)\times 3(p_2+1)} \\ &\text{subject to} \qquad \boldsymbol{B}(\boldsymbol{I}_{(p_2+1)\times(p_2+1)}|\boldsymbol{I}_{(p_2+1)\times(p_2+1)}|\boldsymbol{I}_{(p_2+1)\times(p_2+1)}) = \boldsymbol{\Theta}. \end{aligned} \tag{2.14}$$

The augmented Lagrangian corresponding to (2.14) takes the form

$$\begin{aligned} L_\rho(\boldsymbol{B},\boldsymbol{\Theta},\boldsymbol{\Gamma}) = &\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{W}*\boldsymbol{B}\|_n^2 + \lambda_1\sum_{j=1}^{p_1} P_r(\boldsymbol{D}_{j,.}) + \lambda_2\sum_{k=1}^{p_2} P_c(\boldsymbol{E}_{.,k}) + \lambda_3\|\boldsymbol{F}_{-0,-0}\|_1 \\ &+ \text{trace}\left(\boldsymbol{\Gamma}^\top(\boldsymbol{B}(\boldsymbol{I}|\boldsymbol{I}|\boldsymbol{I})-\boldsymbol{\Theta})\right) + \rho/2\|\boldsymbol{B}(\boldsymbol{I}|\boldsymbol{I}|\boldsymbol{I})-\boldsymbol{\Theta}\|_F^2, \end{aligned} \tag{2.15}$$

where $\boldsymbol{\Gamma}$ is a $(p_1+1)\times 3(p_2+1)$-dimensional dual variable. For convenience, we partition $\boldsymbol{\Gamma}$ as follows: $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1|\boldsymbol{\Gamma}_2|\boldsymbol{\Gamma}_3)$ where $\boldsymbol{\Gamma}_i$ $(i=1,2,3)$ is a $(p_1+1)\times(p_2+1)$ matrix.

The augmented Lagrangian (2.15) can be rewritten as

$$\begin{aligned} L_\rho(\boldsymbol{B},\boldsymbol{\Theta},\boldsymbol{\Gamma}) = &\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{W}*\boldsymbol{B}\|_n^2 + \lambda_1\sum_{j=1}^{p_1} P_r(\boldsymbol{D}_{j,.}) + \lambda_2\sum_{k=1}^{p_2} P_c(\boldsymbol{E}_{.,k}) + \lambda\|\boldsymbol{F}_{-0,-0}\|_1 \\ &+ \langle\boldsymbol{\Gamma_1},\boldsymbol{B}-\boldsymbol{D}\rangle + \langle\boldsymbol{\Gamma}_2,\boldsymbol{B}-\boldsymbol{E}\rangle + \langle\boldsymbol{\Gamma}_3,\boldsymbol{B}-\boldsymbol{F}\rangle \\ &+ \rho/2\|\boldsymbol{B}-\boldsymbol{D}\|_F^2 + \rho/2\|\boldsymbol{B}-\boldsymbol{E}\|_F^2 + \rho/2\|\boldsymbol{B}-\boldsymbol{F}\|_F^2. \end{aligned} \tag{2.16}$$

In order to develop an ADMM algorithm to solve (2.6), we must now simply figure out

how to minimize (2.16) with respect to $\boldsymbol{B}$ with $\boldsymbol{\Theta}$ held fixed, and how to minimize (2.16) with respect to $\boldsymbol{\Theta}$ with $\boldsymbol{B}$ held fixed. Minimizing (2.16) with respect to $\boldsymbol{B}$ amounts simply to a least squares problem. In order to minimize (2.16) with respect to $\boldsymbol{\Theta}$, we note that (2.16) can simply be minimized with respect to $\boldsymbol{D}$, $\boldsymbol{E}$, and $\boldsymbol{F}$ separately. Minimizing (2.16) with respect to $\boldsymbol{F}$ amounts simply to soft-thresholding (Friedman et al., 2007). Minimizing (2.16) with respect to $\boldsymbol{D}$ or with respect to $\boldsymbol{E}$ amounts to solving a problem that is equivalent to (2.12). We consider that problem next.

Details of the ADMM algorithm for solving (2.6) are given in Appendix A.1.2.

2.2.3.2 Solving (2.12) for ℓ_2, ℓ_∞, and hybrid ℓ_1/ℓ_∞ penalties

We saw in the previous section that the updates for $\boldsymbol{D}$ and $\boldsymbol{E}$ in the ADMM algorithm amount to solving the problem (2.12). For $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2$, (2.12) amounts to soft-shrinkage (Simon et al., 2013; Yuan and Lin, 2006), for which a closed-form solution is available. For $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_\infty$, an efficient algorithm was proposed by Duchi and Singer (2009). We now present an efficient algorithm for solving (2.12) for $P(\boldsymbol{\beta}) = \max\{|\beta_1|, \|\boldsymbol{\beta}_{-1}\|_1\}$.

Lemma 2.3. *Let $\widehat{\boldsymbol{\beta}}$ denote the solution to (2.12) with $P(\boldsymbol{\beta}) = \max\{|\beta_1|, \|\boldsymbol{\beta}_{-1}\|_1\}$. Then $\widehat{\boldsymbol{\beta}} = \boldsymbol{y} - \widehat{\boldsymbol{u}}$, where $\widehat{\boldsymbol{u}}$ is the solution to*

$$\begin{aligned} &\underset{\boldsymbol{u}\in\mathbb{R}^p,\ \lambda_1\in\mathbb{R}}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{u}\|^2 \\ &\text{subject to } |u_1| \le \lambda_1,\ \ \|\boldsymbol{u}_{-1}\|_\infty \le \lambda - \lambda_1,\ \ \ 0 \le \lambda_1 \le \lambda. \end{aligned} \tag{2.17}$$

We established in Section 2.2.2.4 that if $\lambda \ge |y_1| + \|\boldsymbol{y}_{-1}\|_\infty$, then the solution to (2.12) is zero. Therefore, we now restrict our attention to the case $\lambda < |y_1| + \|\boldsymbol{y}_{-1}\|_\infty$. For a fixed $\lambda_1 \in [0, \lambda]$, we can see by inspection that the solution to (2.17) is given by

$$u_1(\lambda_1) = \begin{cases} y_1 & |y_1| \le \lambda_1 \\ \lambda_1 \text{sgn}(y_1) & |y_1| > \lambda_1 \end{cases} \text{ and } u_i(\lambda_1) = \begin{cases} y_i & |y_i| \le \lambda - \lambda_1 \\ (\lambda - \lambda_1)\text{sgn}(y_i) & |y_i| > \lambda - \lambda_1 \end{cases}, \tag{2.18}$$

for $i = 2, \ldots, p$. Thus, (2.17) is equivalent to the problem

$$\underset{\lambda_1 \in [0,\lambda]}{\text{minimize}} \; \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{u}(\lambda_1)\|^2. \tag{2.19}$$

Theorem 2.1. *Let $\boldsymbol{z}$ denote the $(p-1)$-vector whose i-th element is $\lambda - |y_{i+1}|$. Then the solution to problem (2.19) is given by*

$$\widehat{\lambda}_1 = \begin{cases} \lambda & \text{if } \min_j \left\{ \frac{|y_1| + \sum_{i=1}^{j} z_{(i)}}{j+1} \right\} \geq \lambda \\ 0 & \text{if } \min_j \left\{ \frac{|y_1| + \sum_{i=1}^{j} z_{(i)}}{j+1} \right\} \leq 0 \\ \min_j \left\{ \frac{|y_1| + \sum_{i=1}^{j} z_{(i)}}{j+1} \right\} & \text{otherwise} \end{cases}. \tag{2.20}$$

Combining Theorem 2.1 and Lemma 2.3 gives us a solution for (2.12) with the hybrid ℓ_1/ℓ_∞ penalty. Proofs are given in Appendix A.2.

2.2.3.3 Convergence, computational complexity, and timing results

As mentioned in Appendix A.1.1, ADMM's convergence to the global optimum is guaranteed for the convex, closed and proper objective function (2.6) (Boyd et al., 2011). The computational complexity of the algorithm depends on the form of the penalty functions used.

The update for $\boldsymbol{B}$ is typically the most computationally-demanding step of the ADMM algorithm for (2.6). As pointed out in Appendix A.1.2, this can be done very efficiently. We perform the singular value decomposition for a $n \times (p_1+1)(p_2+1)$-dimensional matrix *once*, given the data matrix $\boldsymbol{W}$. Then, in each iteration of the ADMM algorithm, the update for $\boldsymbol{B}$ requires simply an efficient matrix inversion using the Woodbury matrix formula.

We now report timing results for our `R`-language implementation of `FAMILY`, available in the package `FAMILY` on `CRAN`, on an Intel® Xeon® E5-2620 processor. We considered an example with $n = 350$ and $p_1 = p_2 = 500$ (for a total of 251,000 features). Using the parametrization (2.33), running `FAMILY.12` with $\alpha = 0.7$ and a grid of 10 λ values takes a

median time of 330 seconds, and running `FAMILY.linf` takes a median time of 416 seconds.

2.2.4 Extension to generalized linear models

The `FAMILY` optimization problem (2.6) can be extended to the case of a general convex loss function $l(\cdot)$,

$$\underset{\boldsymbol{B}\in\mathbb{R}^{(p_1+1)\times(p_2+1)}}{\text{minimize}} \frac{1}{n} l(\boldsymbol{B}) + \lambda_1 \sum_{j=1}^{p_1} P_r(\boldsymbol{B}_{j,\cdot}) + \lambda_2 \sum_{k=1}^{p_2} P_c(\boldsymbol{B}_{\cdot,k}) + \lambda_3 \|\boldsymbol{B}_{-0,-0}\|_1. \tag{2.21}$$

For instance, in the case of a binary response variable $\boldsymbol{y}$, we could take $l(\cdot)$ to be the negative log likelihood under a binomial model. Then (2.21) corresponds to a penalized logistic regression problem with interactions. An ADMM algorithm for (2.21) can be derived just as in Section 2.2.3.1, with a modification to the update for $\boldsymbol{B}$. This is discussed in Appendix A.1.3.

2.2.5 Uniqueness of the `FAMILY` *solution*

The `FAMILY` optimization problem (2.6) is convex, and the algorithm presented in Section 2.2.3 is guaranteed to yield a solution that achieves the global minimum. But (2.6) is not strictly convex: this means that the solution might not be unique, in the sense that more than one value of $\boldsymbol{B}$ might achieve the global minimum. However, uniqueness of the *fitted values* resulting from (2.6) is straightforward. This is formalized in the following lemma. The proof is as in Lemma 1(ii) of Tibshirani (2013).

Lemma 2.4. *For a convex penalty function $P(\cdot)$, let $\widehat{\boldsymbol{B}}$ denote the solution to the problem*

$$\underset{\boldsymbol{B}\in\mathbb{R}^{(p_1+1)\times(p_2+1)}}{\textit{minimize}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{W} * \boldsymbol{B}\|_n^2 + P(\boldsymbol{B}). \tag{2.22}$$

The fitted values $\boldsymbol{W} * \widehat{\boldsymbol{B}}$ are unique.

2.3 Degrees of freedom

2.3.1 Review of degrees of freedom

Consider the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with fixed $\boldsymbol{X}$, and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_n)$. Then the degrees of freedom of a model-fitting procedure is defined as (Stein, 1981; Efron, 1986)

$$\mathrm{df} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(y_i, \hat{y}_i), \tag{2.23}$$

where $\hat{y}_i$ are the fitted response values. If certain conditions hold, then

$$\mathrm{df} = E\left[\sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i}\right]. \tag{2.24}$$

Therefore, $\sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i}$ is an unbiased estimator for the degrees of freedom of the model-fitting procedure.

Before presenting the main results of this section, we state a useful lemma.

Lemma 2.5. *Given a vector $\boldsymbol{x} \in \mathbb{R}^p$, and an even positive integer q,*

$$\frac{d^2 \|\boldsymbol{x}\|_q}{d\boldsymbol{x}^2} = (q-1)\mathrm{diag}\left[\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_q}\right)^{q-2}\right] \times \left[\frac{\boldsymbol{I}}{\|\boldsymbol{x}\|_q} - \frac{\boldsymbol{x}(\boldsymbol{x}^\top)^{q-1}}{\|\boldsymbol{x}\|_q^{q+1}}\right], \tag{2.25}$$

where $\mathrm{diag}(\boldsymbol{x})$ *is the diagonal matrix with $\boldsymbol{x}$ on the diagonal, and $(\boldsymbol{x})^q$ denotes the element-wise exponentiation of the vector $\boldsymbol{x}$.*

2.3.2 Degrees of freedom for a penalized regression problem

We now consider the degrees of freedom of the estimator that solves the problem

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \sum_d \lambda_d P_d(\boldsymbol{A}_d \boldsymbol{\beta}), \tag{2.26}$$

where $P_d(\cdot)$ is an ℓ_q norm for a positive q, and $\boldsymbol{A}_d$ is a $p \times p$ diagonal matrix with ones and zeros on the main diagonal. We define the active set to be $\mathcal{A} = \{j : \widehat{\beta}_j \neq 0\}$, the set of non-zero coefficient estimates. Let $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$ denote the coefficients of the active set, and let $\boldsymbol{X}_{\mathcal{A}}$ denote the matrix with columns corresponding to elements of the active set. Furthermore, we define $\boldsymbol{A}_d^{\mathcal{A}}$ to be the sub-matrix of $\boldsymbol{A}_d$ with rows and columns in $\mathcal{A}$.

Claim 2.1. *An unbiased estimator of the degrees of freedom of $\widehat{\boldsymbol{\beta}}$, the solution to (2.26), is given by*

$$\widehat{df} = \text{trace}\left(\boldsymbol{X}_{\mathcal{A}}\left[\boldsymbol{X}_{\mathcal{A}}^\top \boldsymbol{X}_{\mathcal{A}} + \sum_d \lambda_d \left(\boldsymbol{A}_d^{\mathcal{A}}\right)^\top \ddot{P}_d(\boldsymbol{A}_d^{\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}})\left(\boldsymbol{A}_d^{\mathcal{A}}\right)\right]^{-1}\boldsymbol{X}_{\mathcal{A}}^\top\right), \tag{2.27}$$

where $\ddot{P}_d(\cdot)$ is the Hessian of the function $P_d(\cdot)$, and where $\mathcal{A}$ is the active set.

The derivation for Claim 2.1 is outlined in Appendix A.3.

2.3.3 Degrees of freedom for `FAMILY`

In this section, we present estimates for the degrees of freedom of `FAMILY.l2` and `FAMILY.linf`. An estimate of the degrees of freedom of `FAMILY.hierNet` is given in Bien et al. (2013).

2.3.3.1 `FAMILY.l2`

We write `FAMILY.l2` in the form of (2.26),

$$\frac{1}{2}\|\boldsymbol{y} - \widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{B}}\|_2^2 + n\lambda_1 \sum_{j=1}^{p_1} \|\boldsymbol{A}_j\widetilde{\boldsymbol{B}}\|_2 + n\lambda_2 \sum_{k=p_1+1}^{p_1+p_2} \|\boldsymbol{A}_k\widetilde{\boldsymbol{B}}\|_2 + n\lambda_3\|\boldsymbol{A}_I\widetilde{\boldsymbol{B}}\|_1, \tag{2.28}$$

where $\widetilde{\boldsymbol{B}}$ is the vectorized version of $\boldsymbol{B}$, and $\widetilde{\boldsymbol{W}}$ is the $n \times (p_1+1)(p_2+1)$-dimensional matrix version of $\boldsymbol{W}$. We apply Claim 2.1 in order to obtain an unbiased estimate for `FAMILY.12`:

$$\begin{aligned}\widehat{\mathrm{df}}_{\ell_2} = \mathrm{trace}\Big(\widetilde{\boldsymbol{W}}_{\mathcal{A}}\Big[\widetilde{\boldsymbol{W}}_{\mathcal{A}}^\top\widetilde{\boldsymbol{W}}_{\mathcal{A}} + n\lambda_1\sum_{j=1}^{p_1}(\boldsymbol{A}_j^{\mathcal{A}})^\top\left[\ddot{P}(\boldsymbol{A}_j^{\mathcal{A}}\widehat{\widetilde{\boldsymbol{B}}}_{\mathcal{A}})\right](\boldsymbol{A}_j^{\mathcal{A}}) \\ + n\lambda_2\sum_{k=p_1+1}^{p_1+p_2}(\boldsymbol{A}_k^{\mathcal{A}})^\top\left[\ddot{P}(\boldsymbol{A}_k^{\mathcal{A}}\widehat{\widetilde{\boldsymbol{B}}}_{\mathcal{A}})\right](\boldsymbol{A}_k^{\mathcal{A}})\Big]^{-1}\widetilde{\boldsymbol{W}}_{\mathcal{A}}^\top\Big),\end{aligned} \tag{2.29}$$

where $\ddot{P}(\boldsymbol{v_0}) = \frac{d^2\|\boldsymbol{v}\|_2}{d\boldsymbol{v}^2}\Big|_{\boldsymbol{v}=\boldsymbol{v_0}}$ is of the form given in Lemma 2.5.

2.3.3.2 `FAMILY.linf`

The ℓ_∞ norm is not differentiable, and thus we cannot apply Claim 2.1 directly. Instead, we make use of the fact that $\lim_{q\to\infty}\|\boldsymbol{\beta}\|_q = \|\boldsymbol{\beta}\|_\infty$ in order to apply Claim 2.1 to a modified version of `FAMILY.linf` in which the ℓ_∞ norm is replaced with an ℓ_q norm for a very large value of q. This yields the estimator

$$\begin{aligned}\widehat{\mathrm{df}}_{\ell_\infty} = \mathrm{trace}\Big(\widetilde{\boldsymbol{W}}_{\mathcal{A}}\Big[\widetilde{\boldsymbol{W}}_{\mathcal{A}}^\top\widetilde{\boldsymbol{W}}_{\mathcal{A}} + n\lambda_1\sum_{j=1}^{p_1}(\boldsymbol{A}_j^{\mathcal{A}})^\top\left[\ddot{P}(\boldsymbol{A}_j^{\mathcal{A}}\widehat{\widetilde{\boldsymbol{B}}}_{\mathcal{A}})\right](\boldsymbol{A}_j^{\mathcal{A}}) \\ + n\lambda_2\sum_{k=p_1+1}^{p_1+p_2}(\boldsymbol{A}_k^{\mathcal{A}})^\top\left[\ddot{P}(\boldsymbol{A}_k^{\mathcal{A}}\widehat{\widetilde{\boldsymbol{B}}}_{\mathcal{A}})\right](\boldsymbol{A}_k^{\mathcal{A}})\Big]^{-1}\widetilde{\boldsymbol{W}}_{\mathcal{A}}^\top\Big),\end{aligned} \tag{2.30}$$

where $\ddot{P}(\boldsymbol{v}_0) = \frac{d^2\|\boldsymbol{v}\|_q}{d\boldsymbol{v}^2}\Big|_{\boldsymbol{v}=\boldsymbol{v}_0}$ is of the form given in Lemma 2.5. We use $q = 500$ in Section 2.3.4.

2.3.4 Numerical results

We now consider the numerical performance of our estimates of the degrees of freedom of `FAMILY` in a simple simulation setting. We use a fixed design matrix $\boldsymbol{X}$, with $n = 100$ rows and $p = 10$ main effects, and we let $\boldsymbol{X} = \boldsymbol{Z}$. We randomly selected 15 true interaction terms. We generated 100 different response vectors $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(100)}$ using independent Gaussian noise. We computed the true degrees of freedom as well as the estimated degrees of freedom from

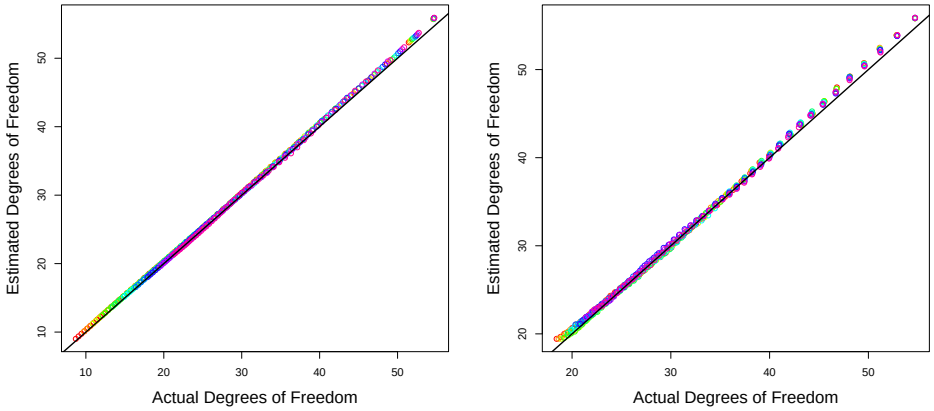


Figure 2.4: The estimated degrees of freedom as a function of the actual degrees of freedom, for *(Left:)* `FAMILY.l2` and *(Right:)* `FAMILY.linf`. To estimate the degrees of freedom for `FAMILY.linf`, we used $q = 500$ in (2.30). Several values of α in were used in the `FAMILY` optimization problem (using the reparametrization in (2.33)); each is shown in a different color. Each point corresponds to a different value of λ in the `FAMILY` optimization problem.

(2.29) and (2.30), averaged over the 100 simulated data sets. In Figure 2.4, we see almost perfect agreement between the true and estimated degrees of freedom.

2.4 Extension to weak heredity

We now consider a modification to the `FAMILY` optimization problem, (2.6), that imposes weak heredity. We assume that the main effects, interactions, and response have been centered to have mean zero.

In order to enforce weak heredity, we take an approach motivated by the latent overlap group lasso of Jacob et al. (2009). We let $\boldsymbol{W}^X$ denote the $n \times p_1 \times (p_2 + 1)$ array defined as follows: for $i \in \{1, \ldots, n\},\ j \in \{1, \ldots, p_1\},\ k \in \{0, \ldots, p_2\}$,

$$W^X_{i,j,k} = \begin{cases} X_{i,j} Z_{i,k} & \text{for } k \neq 0 \\ X_{i,j} & \text{for } k = 0 \end{cases}. \tag{2.31}$$

We let $\boldsymbol{W}^Z$ denote the $n \times (p_1+1) \times p_2$ array defined in an analogous way. We take $\boldsymbol{B}^X$ to be a $p_1 \times (p_2+1)$ matrix, and $\boldsymbol{B}^Z$ to be a $(p_1+1) \times p_2$ matrix.

We propose to solve the optimization problem

$$\begin{aligned} \underset{\substack{\boldsymbol{B}^X \in \mathbb{R}^{p_1 \times (p_2+1)} \\ \boldsymbol{B}^Z \in \mathbb{R}^{(p_1+1) \times p_2}}}{\text{minimize}} \quad & \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{W}^X * \boldsymbol{B}^X - \boldsymbol{W}^Z * \boldsymbol{B}^Z \right\|_n^2 \\ & + \lambda_1 \sum_{j=1}^{p_1} P_r(\boldsymbol{B}^X_{j,.}) + \lambda_2 \sum_{k=1}^{p_2} P_c(\boldsymbol{B}^Z_{.,k}) + \lambda_3 (\|\boldsymbol{B}^X_{.,-0}\|_1 + \|\boldsymbol{B}^Z_{-0,.}\|_1). \end{aligned} \tag{2.32}$$

Then the coefficient for the j-th main effect of X is $B^X_{j,0}$, the coefficient for the k-th main effect of Z is $B^Z_{0,k}$, and the coefficient for the (j,k) interaction is $B^X_{j,k} + B^Z_{j,k}$. If we take P_r and P_c to be either ℓ_2, ℓ_∞, or hybrid ℓ_1/ℓ_∞ penalties, then (2.32) imposes weak heredity: if the k-th column of $\boldsymbol{B}^Z$ has a zero estimate, then the (j,k)-th interaction coefficient estimate need not be zero. However, if the j-th row of $\boldsymbol{B}^X$ and the k-th column of $\boldsymbol{B}^Z$ have zero estimates, then the (j,k)-th interaction coefficient estimate is zero.

Problem (2.32) can be solved using an ADMM algorithm similar to that of Section 2.2.3. Since the focus of this chapter is on enforcing strong heredity, we leave the details of an algorithm for (2.32), as well as a careful numerical study, to future work.

2.5 Simulation study

We compare the performance of `FAMILY.l2` and `FAMILY.linf` to the all-pairs lasso (`APL`), the `hierNet` proposal of Bien et al. (2013), and the `glinternet` proposal of Lim and Hastie (2013). `APL` can be performed using the `glmnet R` package, and `hierNet` and `glinternet` are implemented in `R` packages available on `CRAN`. We also include the oracle model (Fan and Li, 2001) — an unpenalized model that uses only the main effects and interactions that are non-zero in the true model — in our comparisons.

The forward selection proposal of Hao and Zhang (2014), `iFORM`, is a fast screening approach for detecting interactions in ultra-high dimensional data. `iFORM` is intended for the

setting in which the true model is extremely sparse. In our simulation setting, we consider moderately sparse models, which fails to highlight the advantages of `iFORM`. Thus, we do not include results for `iFORM` in our simulation study.

To facilitate comparison with `hierNet` and `glinternet`, which require $\boldsymbol{X} = \boldsymbol{Z}$, we take $\boldsymbol{X} = \boldsymbol{Z}$ in our simulation study. Similar empirical results are obtained in simulations with $\boldsymbol{X} \neq \boldsymbol{Z}$; results are omitted due to space constraints.

We consider squared error loss in Section 2.5.1, and logistic regression loss in Section 2.5.2.

2.5.1 Squared error loss

2.5.1.1 Simulation set-up

We created a coefficient matrix $\boldsymbol{B}$, with $p = 30$ main effects and $\binom{p}{2} = 435$ interactions, for a total of 465 features. The first 10 main effects have non-zero coefficients, assigned uniformly from the set $\{-5, -4, \ldots, -1, 1, \ldots, 5\}$. The remaining main effects' coefficients equal zero. We consider three simulation settings, in which we randomly select 15, 30 or 45 non-zero interaction coefficients, chosen to obey strong heredity. The values for the non-zero coefficients were selected uniformly from the set $\{-10, -8, \ldots, -2, 2, \ldots, 8, 10\}$. Figure 2.5 displays $\boldsymbol{B}$ in each of the three simulation settings.

We generated a training set, a test set, and a validation set, each consisting of 300 observations. Each observation of $\boldsymbol{X} = \boldsymbol{Z}$ was generated independently from a $\mathcal{N}_p(0, I)$ distribution; $\boldsymbol{W}$ was then constructed according to (2.2). For each observation we generated an independent Gaussian noise term, with variance adjusted to maintain a signal-to-noise ratio of approximately 2.5 to 3.5. Finally, for each observation, a response was generated according to (2.3).

We applied `glinternet` and `hierNet` for 50 different values of the tuning parameters. For convenience, given that $\boldsymbol{X} = \boldsymbol{Z}$, we reparametrized the `FAMILY` optimization problem

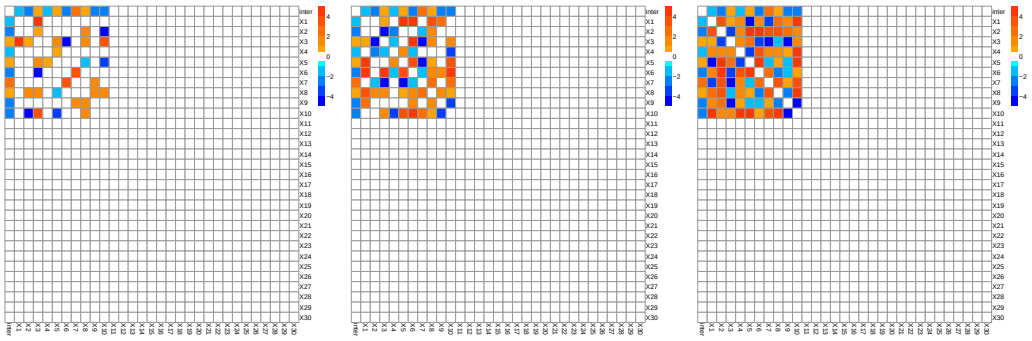


Figure 2.5: For the simulation study in Section 2.5, the heatmap of the matrix $\boldsymbol{B}$ is displayed in the case of 15 (*left*), 30 (*center*), and 45 (*right*) non-zero interactions. The first row and column of each heatmap represent the main effects.

(2.6) as

$$\begin{aligned}\underset{\boldsymbol{B}\in\mathbb{R}^{(p+1)\times(p+1)}}{\text{minimize}} \; & \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{W}*\boldsymbol{B}\|_n^2+\alpha\lambda\|\boldsymbol{B}_{-0,-0}\|_1 \\ & +(1-\alpha)\lambda\sqrt{p}\sum_{j=1}^{p}P_r(\boldsymbol{B}_{j,.})+(1-\alpha)\lambda\sqrt{p}\sum_{k=1}^{p}P_c(\boldsymbol{B}_{.,k}).\end{aligned} \tag{2.33}$$

We applied `FAMILY.l2` and `FAMILY.linf` over a 10×50 grid of (α,λ) values, with $\alpha\in(0,1)$ and λ chosen to give a suitable range of sparsity.

In principle, many methods are available for selecting the tuning parameters α and λ. These include Bayesian information criterion, generalized cross-validation, and others. Because we do not have an estimator for the degrees of freedom of the `glinternet` estimator, we opted to use a training/test/validation set approach. In greater detail, we fit each method to the training set, selected tuning parameters based on sum of squared residuals (SSR) on the test set, and then reported the SSR for that choice of tuning parameters on the validation set.

It is well-known that penalized regression techniques tend to yield models with over-shrunken coefficient estimates (Hastie et al., 2009; Fan and Li, 2001). To overcome this problem, we obtained *relaxed* versions of `FAMILY.l2`, `FAMILY.linf`, `hierNet`, and `glinternet`,

by refitting an unpenalized least squares model to the set of coefficients that are non-zero in the penalized fitted model (Meinshausen, 2007; Radchenko and James, 2010).

We also considered generating the observations of $\boldsymbol{X}$ from a $\mathcal{N}_p(0, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\Sigma}$ was an autoregressive or an exchangeable covariance matrix. We found that the choice of covariance matrix $\boldsymbol{\Sigma}$ led to little qualitative difference in the results. Therefore, we display only results for $\boldsymbol{\Sigma} = \boldsymbol{I}$ in Section 2.5.1.2.

2.5.1.2 Results

The left panel of Figure 2.6 displays ROC curves for `FAMILY.linf`, `FAMILY.l2`, `hierNet`, `glinternet`, and `APL`. These results indicate that `FAMILY.l2` outperforms all other methods in terms of variable selection, especially as the number of non-zero interaction coefficients increases. When there are 45 non-zero interactions, `FAMILY.linf` outperforms `glinternet`, `hierNet`, and `APL`.

The right panel of Figure 2.6 displays the test set SSR for all methods, as the tuning parameters are varied. We observe that relaxation leads to improvement for each method: it yields a much sparser model for a given value of the test error. This is not surprising, since the relaxation alleviates some of the over-shrinkage induced by the application of multiple convex penalties. The results further indicate that when relaxation is applied, `FAMILY.l2` performs the best, followed by `FAMILY.linf` and then the other competitors. We once again observe that the improvement of `FAMILY.l2` and `FAMILY.linf` over the competitors increases as the number of non-zero interaction coefficients increases.

Interestingly, the right-hand panel of Figure 2.6 indicates that though `FAMILY.l2` performs the best when relaxation is performed, it performs quite poorly when relaxation is not performed, in that the model with smallest test set SSR contains far too many non-zero interactions. This is consistent with the remark in Radchenko and James (2010) regarding over-shrinkage of coefficient estimates.

In Table 1, we present results on the validation set for the model that was fit on the training set using the tuning parameters selected on the test set, as described in Section 2.5.1.1.

We see that `FAMILY.l2` and `FAMILY.linf` outperform the competitors in terms of SSR, false discovery rate, and true positive rate, especially when relaxation is performed.

2.5.2 Logistic regression

2.5.2.1 Simulation set-up

We assume that each response y_i is a Bernoulli variable with probability p_i. We then model p_i as

$$\log\left(\frac{p_i}{1-p_i}\right) = (\boldsymbol{W} * \boldsymbol{B})_i \quad (i = 1, \ldots, n), \tag{2.34}$$

where $\boldsymbol{W} * \boldsymbol{B}$ is the n-vector defined in Section 2.1.1. The matrices $\boldsymbol{X}$ and $\boldsymbol{B}$ are generated in the exact same manner as in Section 2.5.1.1, but now with $n = 500$ observations in the training and test sets.

Once again, for convenience, we reparametrized `FAMILY.l2` and `FAMILY.linf` according to

$$\begin{aligned}\underset{\boldsymbol{B}\in\mathbb{R}^{(p+1)\times(p+1)}}{\text{minimize}} \; & -\frac{1}{n}\sum_{i=1}^{n}\left[y_i(\boldsymbol{W}*\boldsymbol{B})_i - \log\left(1+e^{(\boldsymbol{W}*\boldsymbol{B})_i}\right)\right] \\ & + \sqrt{p}(1-\alpha)\lambda\sum_{j=1}^{p} P_r(\boldsymbol{B}_{j,.}) + \sqrt{p}(1-\alpha)\lambda\sum_{k=1}^{p} P_c(\boldsymbol{B}_{.,k}) + \alpha\lambda\|\boldsymbol{B}_{-0,-0}\|_1.\end{aligned} \tag{2.35}$$

2.5.2.2 Results

The results for logistic regression are displayed in Figure 2.7. The ROC curves in the left-hand panel indicate that `FAMILY.linf` and `FAMILY.l2` outperform the competitors in terms of variable selection when there are 30 or 45 non-zero interactions. The SSR curves in the right-hand panel of Figure 2.7 indicate that the relaxed versions of `FAMILY.linf` and `FAMILY.l2` perform very well in terms of prediction error on the test set, especially as the number of non-zero interactions increases.

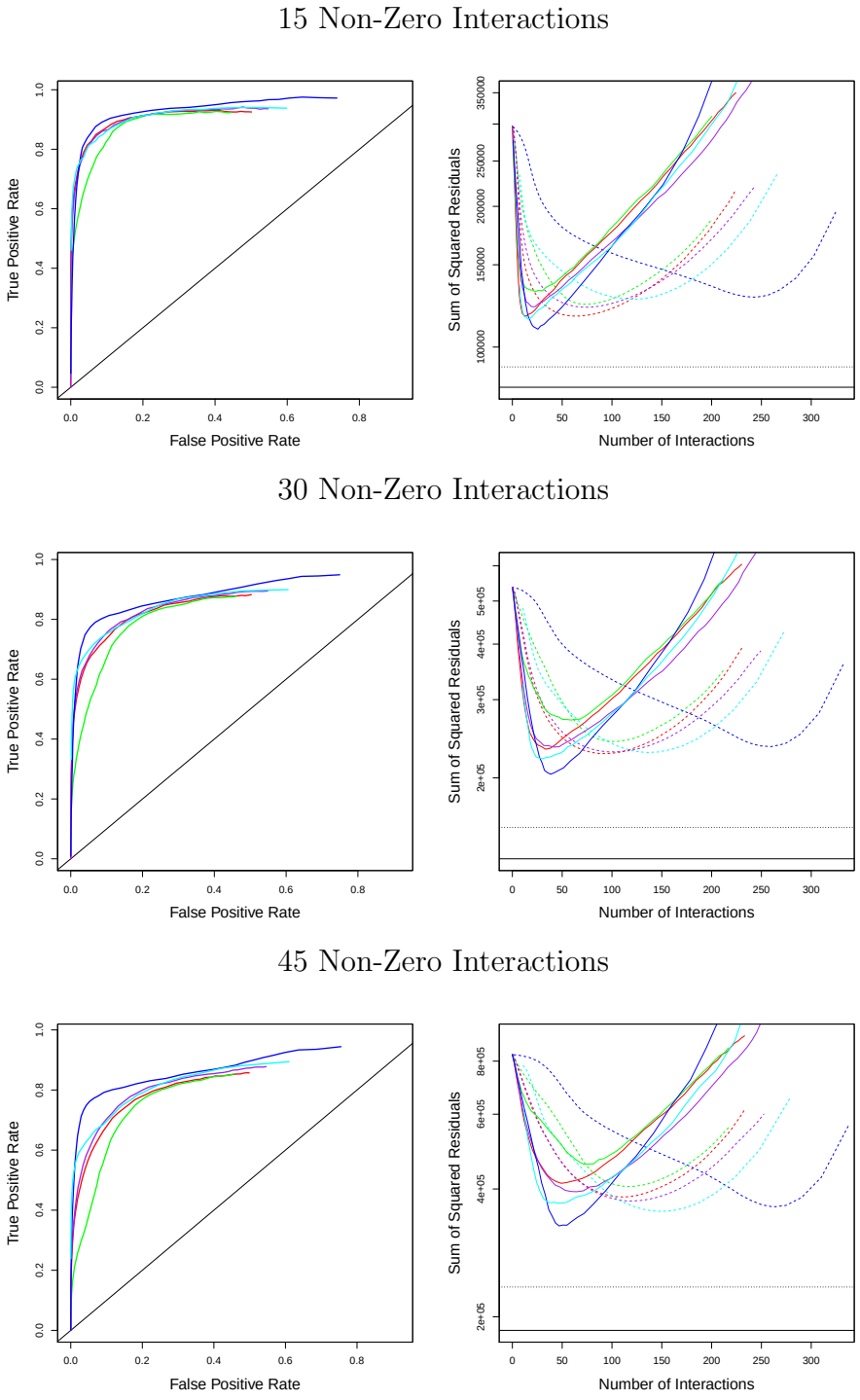


Figure 2.6: Results for the simulation study of Section 2.5.1, averaged over 100 simulated datasets. The colored lines indicate the results for `glinternet` (——), `hierNet` (——), `APL` (——), `FAMILY.l2` with $\alpha = 0.7$ (——), and `FAMILY.linf` with $\alpha = 0.83$ (——). *Left:* ROC curves for each proposal, along with the 45° line. *Right:* Sum of squared residuals (SSR), evaluated on the test set. Each method is shown with (——) and without (·····) relaxation. The two horizontal black lines indicate the test set SSR of the true model (——) and of the oracle model (·····).

Table 2.1: Simulation results, averaged over 100 simulated datasets, for the simulation set-up in Section 2.5.1. Tuning parameters were selected using a training/test/validation set approach, as described in Section 2.5.1.1. From left to right, the table's columns indicate the true number of non-zero interactions, the method used, whether or not relaxation was performed, the sum of squared residuals (SSR) on the validation set divided by the SSR of the oracle, the false discovery rate for the detection of non-zero interactions, the true positive rate for the detection of non-zero interactions, and the number of estimated non-zero interactions. Standard errors of the mean are reported in parentheses.

	Method	Relaxed	Relative SSR	FDR	TPR	Num. Inter.
15	`FAMILY.l2`	No	1.333 (0.012)	0.892 (0.002)	0.931 (0.006)	132.01 (2.3)
		Yes	1.133 (0.010)	0.399 (0.017)	0.837 (0.009)	22.94 (0.8)
	`FAMILY.linf`	No	1.348 (0.011)	0.855 (0.003)	0.915 (0.006)	97.85 (1.7)
		Yes	1.179 (0.011)	0.304 (0.017)	0.771 (0.010)	17.87 (0.6)
	`glinternet`	No	1.288 (0.011)	0.786 (0.004)	0.889 (0.007)	64.85 (1.4)
		Yes	1.230 (0.010)	0.209 (0.017)	0.691 (0.011)	14.23 (0.6)
	`hierNet`	No	1.359 (0.012)	0.816 (0.003)	0.881 (0.007)	73.12 (1.2)
		Yes	1.355 (0.013)	0.382 (0.023)	0.632 (0.013)	19.76 (1.4)
	`APL`	No	1.341 (0.011)	0.816 (0.004)	0.895 (0.007)	75.90 (1.6)
		Yes	1.308 (0.012)	0.375 (0.019)	0.749 (0.011)	20.65 (1.0)
30	`FAMILY.l2`	No	1.492 (0.016)	0.841 (0.003)	0.884 (0.006)	172.00 (3.3)
		Yes	1.218 (0.012)	0.352 (0.014)	0.800 (0.010)	39.09 (1.1)
	`FAMILY.linf`	No	1.476 (0.016)	0.790 (0.004)	0.846 (0.007)	124.00 (2.2)
		Yes	1.276 (0.013)	0.310 (0.016)	0.735 (0.008)	34.11 (1.0)
	`glinternet`	No	1.487 (0.015)	0.730 (0.005)	0.800 (0.007)	91.75 (1.8)
		Yes	1.446 (0.016)	0.328 (0.017)	0.627 (0.010)	31.07 (1.3)
	`hierNet`	No	1.567 (0.016)	0.754 (0.003)	0.797 (0.008)	98.95 (1.7)
		Yes	1.677 (0.019)	0.581 (0.013)	0.647 (0.012)	50.90 (1.8)
	`APL`	No	1.492 (0.016)	0.751 (0.004)	0.821 (0.007)	101.73 (1.8)
		Yes	1.484 (0.018)	0.411 (0.016)	0.676 (0.010)	37.78 (1.4)
45	`FAMILY.l2`	No	1.562 (0.020)	0.816 (0.003)	0.889 (0.005)	223.29 (4.0)
		Yes	1.219 (0.016)	0.203 (0.016)	0.833 (0.008)	49.09 (1.2)
	`FAMILY.linf`	No	1.531 (0.019)	0.754 (0.003)	0.841 (0.006)	156.59 (2.6)
		Yes	1.324 (0.023)	0.200 (0.019)	0.756 (0.009)	45.78 (1.5)
	`glinternet`	No	1.658 (0.021)	0.679 (0.004)	0.776 (0.005)	110.28 (1.4)
		Yes	1.689 (0.025)	0.415 (0.012)	0.610 (0.009)	50.07 (1.7)
	`hierNet`	No	1.746 (0.023)	0.699 (0.003)	0.772 (0.006)	116.46 (1.5)
		Yes	1.876 (0.027)	0.585 (0.006)	0.650 (0.008)	72.29 (1.5)
	`APL`	No	1.616 (0.021)	0.693 (0.004)	0.802 (0.005)	119.73 (1.8)
		Yes	1.633 (0.023)	0.456 (0.012)	0.674 (0.008)	59.40 (1.8)

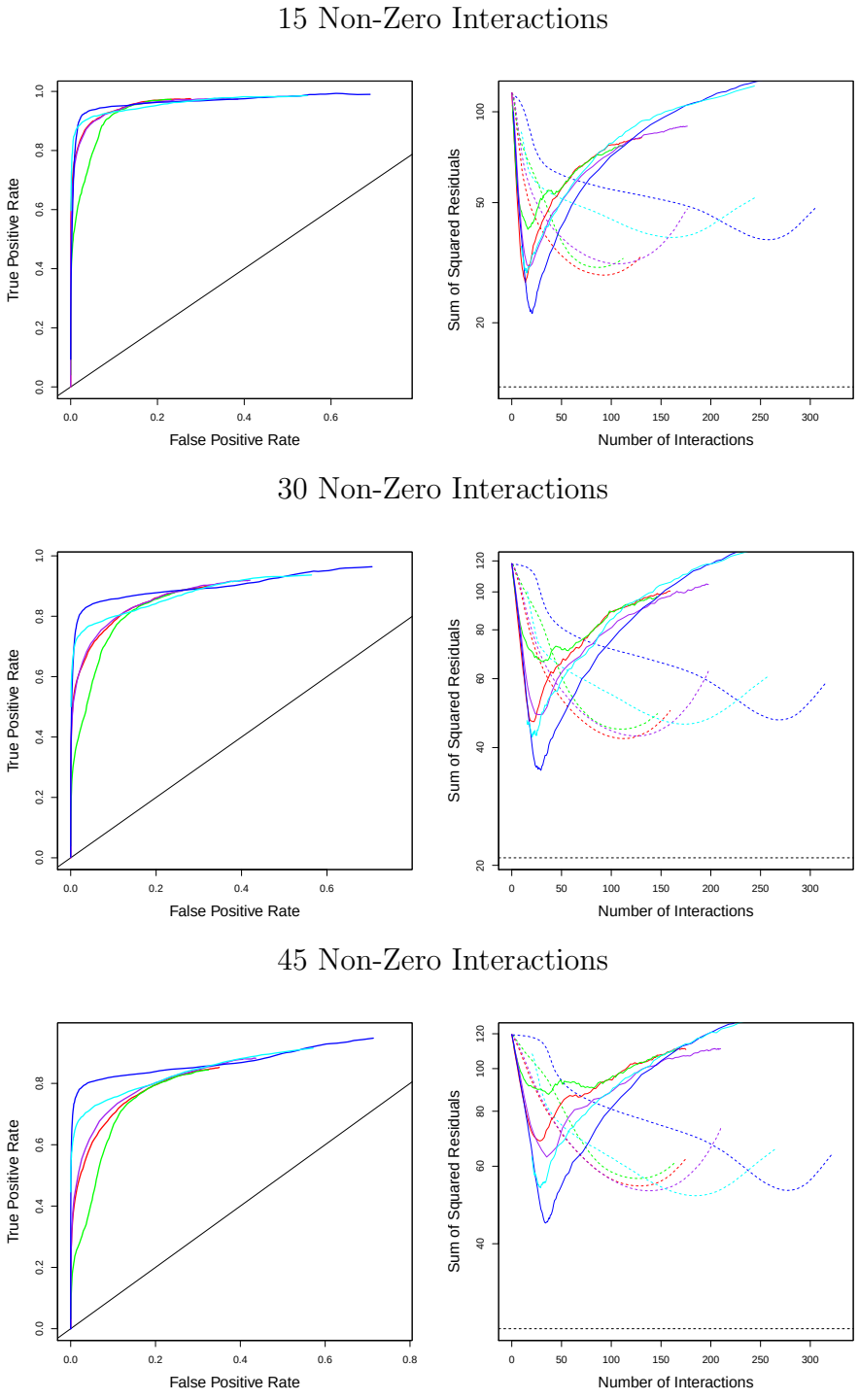


Figure 2.7: Results for the simulation study of Section 2.5.2, averaged over 100 simulated data sets. Details are as in Figure 2.6, but with $\alpha = 0.8$ for `FAMILY.linf` (——).

2.6 Application to HIV data

Rhee et al. (2006) study the susceptibility of the HIV-1 virus to 6 nucleoside reverse transcriptase inhibitors (NRTIs). The HIV-1 virus can become resistant to drugs via mutations in its genome sequence. Therefore, there is a need to model HIV-1's drug susceptibility as a function of mutation status. We consider one particular NRTI, 3TC. The data consists of a sparse binary matrix, with mutation status at each of 217 genomic locations for $n = 1057$ HIV-1 isolates. For each of the observations, there is a measure of susceptibility to 3TC. This data set was also studied by Bien et al. (2013).

Rather than working with all 217 genomic locations, we create bins of ten adjacent loci; this results in a design matrix with $p = 22$ features and $n = 1057$ observations. We perform the binning because the raw data contains mostly zeros, as most mutations occur in at most a few of the observations; by binning the observations, we obtain less sparse data. This binning is justified under the assumption that mutations in a particular region of the genome sequence result in a change to a binding site, in which case nearby mutations should have similar effects on a binding site, and hence similar associations with drug susceptibility. This binning is also needed for computational reasons, in order to allow for comparison to `hierNet` (specifically the version that enforces strong heredity) using the `R` package of Bien et al. (2013). (In Bien et al. (2013), all 217 genomic locations are analyzed using a much faster algorithm that enforces *weak* (rather than strong) heredity.)

We split the observations into equally-sized training and test sets. We fit `glinternet`, `hierNet`, `FAMILY.l2`, and `FAMILY.linf` on the training set for a range of tuning parameter values, and applied the fitted models to the test set. In Figure 2.8, the test set SSR is displayed as a function of the number of non-zero estimated interaction coefficients, averaged over 50 splits of the data into training and test sets. The figure reveals that all four methods give roughly similar results.

Figure 2.9 displays the estimated coefficient matrix, $\widehat{\boldsymbol{B}}$, that results from applying each of the four methods to all $n = 1057$ observations using the tuning parameter values that

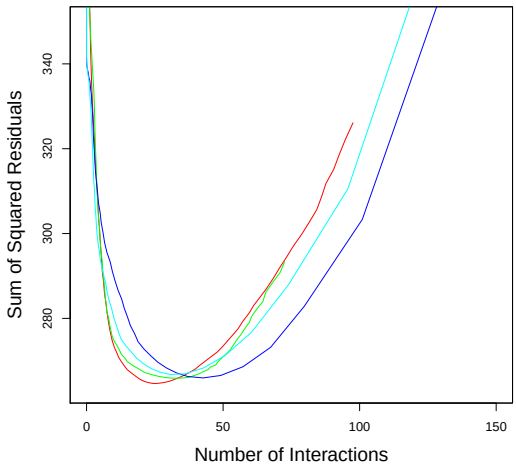


Figure 2.8: The test set SSR is displayed for the HIV-1 data of Section 2.6, as a function of the number of non-zero interaction terms. Results are averaged over 50 splits of the observations into a training set and a test set. The colored lines indicate the results for `glinternet` (——) , `hierNet` (——), `FAMILY.l2` with $\alpha = 0.944$ (——), and `FAMILY.linf` with $\alpha = 0.944$ (——).

minimized the average test set SSR. The estimated coefficients are qualitatively similar for all four methods. All four methods detect some non-zero interactions involving the 17th feature. `Glinternet` yields the sparsest model.

2.7 Discussion

In this chapter, we have introduced `FAMILY`, a framework that unifies a number of existing estimators for high-dimensional models with interactions. Special cases of `FAMILY` correspond to the all-pairs lasso, the main effects lasso, `VANISH`, and `hierNet`. Furthermore, we have explored the use of `FAMILY` with ℓ_2, ℓ_∞, and hybrid ℓ_1/ℓ_∞ penalties; these result in strong heredity and have good empirical performance.

The empirical results in Sections 2.5 and 2.6 indicate that the choice of penalty in `FAMILY` may be of little practical importance: for instance, `FAMILY.l2`, `FAMILY.linf`, and `FAMILY.hierNet` have similar performance. However, one could choose among penalties

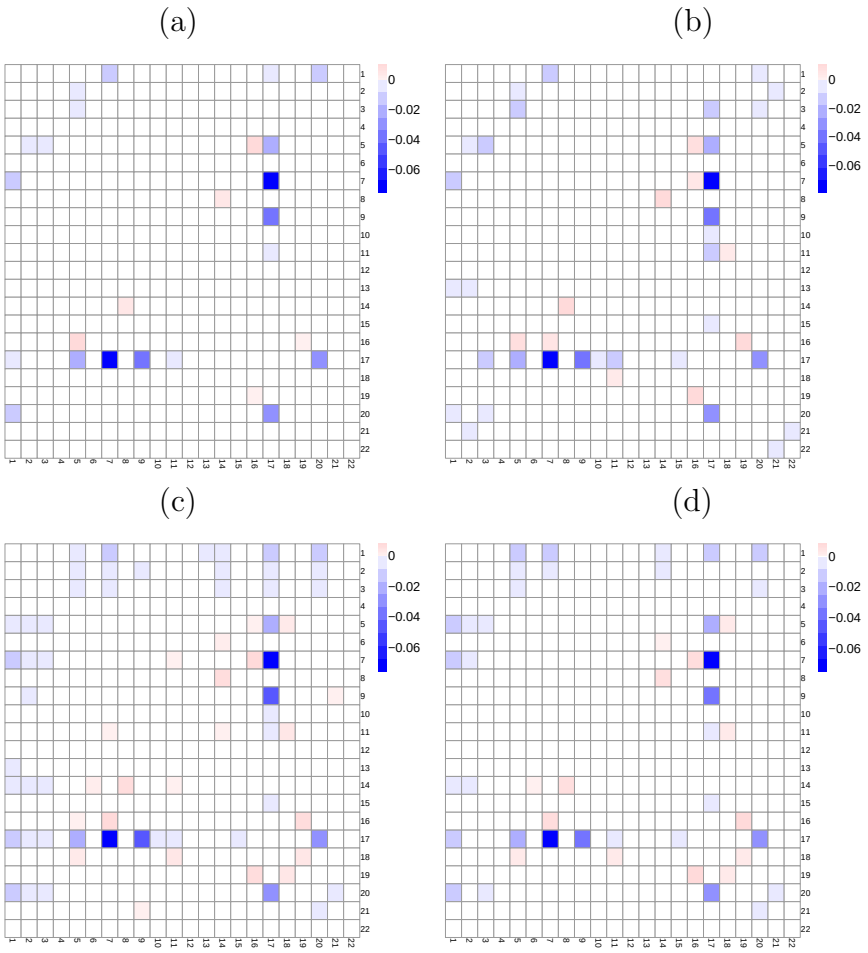


Figure 2.9: For the HIV-1 data of Section 2.6, the estimated coefficient matrix $\hat{B}_{-0,-0}$ is shown for *(a):* `glinternet`; *(b):* `hierNet`; *(c):* `FAMILY.l2` with $\alpha = 0.944$; and *(d):* `FAMILY.linf` with $\alpha = 0.944$. Main effects are not displayed.

using cross-validation or a related approach.

We have presented a simple ADMM algorithm that can be used to solve the `FAMILY` optimization problem for any convex penalty. It finds the global optimum for `VANISH` (unlike the proposal in Radchenko and James (2010)), and provides a simpler alternative to the original `hierNet` algorithm (Bien et al., 2013).

`FAMILY` could be easily extended to accommodate higher-order interaction models. For instance, to accommodate third-order interactions, we could take $\boldsymbol{B}$ to be a $(p+1) \times (p+1) \times (p+1)$ coefficient array. Instead of penalizing each row and each column of $\boldsymbol{B}$, we would instead penalize each 'slice' of the array.

In the simulation study in Section 2.5, we considered a setting with only $p_1 = p_2 = 30$ main effects. We did this in order to facilitate comparison to the `hierNet` proposal, which is very computationally intensive as implemented in the `R` package of Bien et al. (2013). However, our proposal can be applied for much larger values of p_1 and p_2, as discussed in Section 2.2.3.3.

The `R` package `FAMILY`, available on `CRAN`, implements the methods described in this chapter.

Chapter 3

GENERALIZED SPARSE ADDITIVE MODELS

3.1 Introduction

In this chapter, we model a response variable as an additive function of a potentially large number of covariates. The problem can be formulated as follows: we are given n observations with response $y_i \in \mathbb{R}$ and covariates $\boldsymbol{x}_i \in \mathbb{R}^p$ $(i = 1, \ldots, n)$. The goal is to fit the model

$$g\left(\mathbb{E}\left(y_i|\boldsymbol{x}_i\right)\right) = \beta_0 + \sum_{j=1}^{p} f_j\left(x_{ij}\right) \quad (i = 1, \ldots, n),$$

for a prespecified *link* function g, unknown intercept β_0 and, unknown component functions $f_1, \ldots, f_p$. The link function, g, is generally based on the outcome data-type, e.g., $g(x) = x$ or $g(x) = \log(x)$ for continuous or count response data, respectively. The estimands, $f_1, \ldots, f_p$, give the conditional relationships between each feature x_{ij} and the outcome y_i $(i = 1, \ldots, n;\ j = 1, \ldots, p)$. For identifiablility, we assume $\sum_{i=1}^{n} f_j(x_{ij}) = 0$ $(j = 1, \ldots, p)$. This is known as a generalized additive model (GAM) (Hastie and Tibshirani, 1990); it extends the generalized linear model (GLM) where each f_j is linear, and is a popular choice for modeling different types of response variables as a function of covariates. GAMs are popular because (a) they extend GLMs to model non-linear conditional relationships while (b) retaining some interpretability (we can examine the effect of each covariate x_{ij} individually on y_i while holding all other variables fixed) and, (c) they do not suffer from the *curse of dimensionality*.

While there are a number of proposals for estimating GAMs, a popular approach is to encode the estimation in the following convex optimization problem (Sadhanala and Tibshirani,

2017):

$$\widehat{\beta}, \widehat{f}_1, \ldots, \widehat{f}_p \leftarrow \underset{\beta \in \mathbb{R}, f_1, \ldots, f_p \in \mathcal{F}}{\operatorname{argmin}} \; -\frac{1}{n} \sum_{i=1}^{n} \ell\Big(y_i, \beta + \sum_{j=1}^{p} f_j\left(x_{ij}\right)\Big) + \lambda_{st} \sum_{j=1}^{p} P_{st}\left(f_j\right). \tag{3.1}$$

Here $\mathcal{F}$ is some suitable function class; $\ell(y_i, \theta)$ is the log-likelihood of y_i under parameter θ; $P_{st}(\cdot)$ is a structure-inducing penalty to control the wildness of the estimated functions, $\widehat{f}_j$; and $\lambda_{st} > 0$ is a penalty parameter which modulates the trade-off between goodness-of-fit and structure/smoothness of estimates. The class $\mathcal{F}$ is a general convex space, e.g., $\mathcal{F} = L^2[0, 1]$. Functions $-\ell(y_i, \theta)$ and $P_{st}(f_j)$ are convex in θ and f_j, respectively. The objective function in (3.1) is convex and for small dimension, p, can be solved via a general-purpose convex solver. However, many modern datasets are high-dimensional, often with more features than observations, i.e., $p > n$. Fitting even GLMs is challenging in such settings as conventional methods are known to overfit the data. A common assumption in this setting is *sparsity*, that is, only a small (but unknown) subset of features are informative for the outcome. In this case, it is desirable to apply feature selection: to build a model for which only a small subset of $\widehat{f}_j \not\equiv 0$.

A number of estimators have been proposed for fitting GAMs with sparsity. These estimators are generally solutions to a convex optimization problem. Though they differ in details, we show that most of these optimization problems can be written as:

$$\widehat{\beta}, \widehat{f}_1, \ldots, \widehat{f}_p \leftarrow \underset{\beta \in \mathbb{R}, f_1, \ldots, f_p \in \mathcal{F}}{\operatorname{argmin}} \; -\frac{1}{n} \sum_{i=1}^{n} \ell\Big(y_i, \beta + \sum_{j=1}^{p} f_j\left(x_{ij}\right)\Big) + \lambda_{st} \sum_{j=1}^{p} P_{st}\left(f_j\right) + \lambda_{sp} \sum_{j=1}^{p} \left\|f_j\right\|_n, \tag{3.2}$$

where $\|f_j\|_n = [n^{-1} \sum_{i=1}^{n} \{f_j(x_{ij})\}^2]^{1/2}$ is a group lasso-type penalty (Yuan and Lin, 2006) for feature-wise sparsity, and λ_{sp} a sparsity-related tuning parameter (Ravikumar et al., 2009; Lou et al., 2016; Petersen et al., 2016; Sadhanala and Tibshirani, 2017; Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Yuan and Zhou, 2015; Meier et al., 2009). For all of these previous proposals, gaps exist in the literature around efficient computation (Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Yuan and Zhou, 2015) and optimal statistical convergence properties (Ravikumar et al., 2009; Lou et al., 2016; Petersen et al., 2016; Sadhanala

and Tibshirani, 2017) of these estimators. Numerous authors suggest using general-purpose convex solvers for the problem (3.2) which roughly scale as $O(n^3p^3)$ (Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Yuan and Zhou, 2015). This chapter aims to bridge these gaps.

We present a general framework for sparse GAMs with two major contributions, a general algorithm for computing (3.2) and a theorem for establishing convergence rates. Briefly, our algorithm is based on accelerated proximal gradient descent. This reduces (3.2) to repeatedly solving a univariate penalized least squares problem. In many cases, this algorithm has a per iteration complexity of $O(np)$ — precisely that of state-of-the-art algorithms for the lasso (Friedman et al., 2010; Beck and Teboulle, 2009b). Our main theorem establishes fast convergence rates of the form $\max(s \log p/n, s\nu_n)$, where s is the number of signal variables and ν_n is the minimax rate of the univariate regression problem i.e., the problem (3.1) with $p = 1$. Nonparametric rates are established for a wide class of structural penalties P_{st} with $\nu_n = n^{-2m/(2m+1)}$, popular choices of P_{st} include m-th order Sobolev and Hölder norms, total variation norm of the m-th derivative and, norms of Reproducing Kernel Hilbert Spaces (RKHS). Parametric rates are also established with $\nu_n = T_n/n$ via a truncation-penalty; the number of parameters, T_n, can be fixed or allowed to grow with sample size. As a byproduct of our general theorem, our framework specifies $\lambda_{st} = \lambda_{sp}^2$ in (3.2), reducing the problem to a single tuning parameter. The highlight of this chapter is the generality of our framework: not only does it encompass many existing estimators for high-dimensional GAMs but also estimators for, low-dimensional GAMs, low-dimensional fully nonparametric models and, parametric models in low or high-dimensional settings.

The rest of the chapter is organized as follows. In Section 3.2 we detail our framework, and discuss various choices for P_{st}, illustrating that our framework to encompasses many existing proposals for high-dimensional GAMs. In Section 3.3 we present an algorithm for solving the optimization problem (3.2). Theoretical convergence rates are presented in Section 3.4 for a broad class of structural penalties. We explore the empirical performance of sparse GAMs for various choices of P_{st} in simulation in Section 3.5, and in an application to the Boston housing dataset in Section 3.6. Concluding remarks are in Section 3.7.

3.2 General framework for additive models

In this section, we present our general framework for estimating sparse GAMs, discuss its salient features, and review some existing methods as special cases of our framework. Before presenting our framework, we introduce some notation. For any function f and response/covariate pair, $(y, \boldsymbol{x})$, let $-\ell(f) \equiv -\ell(y, f(\boldsymbol{x}))$ denote a loss function; for given data $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$, let $\mathbb{P}_n \ell(f) \equiv n^{-1} \sum_{i=1}^n \ell(y_i, f(\boldsymbol{x}_i))$ denote an empirical average; and $\|f\|_n^2 \equiv n^{-1} \sum_{i=1}^n f(\boldsymbol{x}_i)^2$ denote the empirical norm. With some abuse of notation, we will use the shorthand f_j to denote the function $f_j \circ \pi_j$ where $\pi_j(\boldsymbol{x}) = x_j$ for $\boldsymbol{x} \in \mathbb{R}^p$.

We propose a general framework for obtaining a *Penalized Generalized Sparse Additive Model Estimator* (`PGSAME`), specifically a `PGSAME` is a solution to the following optimization problem:

$$\widehat{\beta}, \widehat{f}_1, \ldots, \widehat{f}_p \leftarrow \underset{\beta \in \mathbb{R}, f_1, \ldots, f_p \in \mathcal{F}}{\operatorname{argmin}} \underbrace{-\mathbb{P}_n \ell\Big(\beta + \sum_{j=1}^p f_j\Big)}_{\text{Goodness-of-fit}} + \underbrace{\lambda^2 \sum_{j=1}^p P_{st}\left(f_j\right)}_{\text{structure-inducing}} + \underbrace{\lambda \sum_{j=1}^p \|f_j\|_n}_{\text{sparsity-inducing}} . \tag{3.3}$$

This optimization problem balances 3 pieces which we discuss here. The first is a loss function based on goodness-of-fit to the observed data; the least squares loss, $-\ell(f) = (y - f(\boldsymbol{x}))^2$, is commonly used for continuous response. Our general framework requires only convexity and differentiability of $-\ell(y, \theta)$, with respect to θ. Later we consider loss functions given by the negative log-likelihood of exponential family distributions. The second piece is a penalty to induce smoothness/structure of the function estimates. Our framework requires P_{st} to be a *semi-norm* on $\mathcal{F}$. This choice is motivated by both statistical theory and computational efficiency; we discuss this along with possible choices of P_{st} in the following sub-sections. The final piece is a sparsity penalty $\|\cdot\|_n$, this estimates models with $\widehat{f}_j \equiv 0$ for many j. Surprisingly, P_{st} also plays an important role in obtaining an appropriate sparsity pattern. Briefly, if P_{st} is a squared semi-norm then either all $\widehat{f}_j \equiv 0$ or all $\widehat{f}_j \not\equiv 0$; to fit models where some $\widehat{f}_j \equiv 0$ and not others, the non-differentiablity of semi-norms at 0 is crucial, we detail

this in Section 3.2.2 below. Throughout this chapter, we require the function class $\mathcal{F}$ to be a convex cone, e.g., $L^2(\mathbb{R})$. Later for some specific results we will additionally require $\mathcal{F}$ to be a linear space.

Our framework has two additional desirable features. Firstly, the tuning parameters for structure (λ) and sparsity (λ^2) are coupled. The theoretical consequence of this is that, for properly chosen λ, we get rate-optimal estimates (shown in Section 3.4); the practical consequence is that we have a single tuning parameter. Secondly, our framework relaxes the usual distributional requirements of i.i.d. response from an exponential family; we require only y_i independent and $E(y_i - E[y_i])$ to be sub-Gaussian (or sub-Exponential). This demonstrates the generality of our framework and highlights our main innovation: the efficient algorithm of Section 3.3 and theoretical results of Section 3.4 apply to a very broad class of estimators, fill in the gaps of existing work and, can easily be applied for the development of future estimators.

3.2.1 Structure inducing penalties

We now present some possible choices of the structural penalty P_{st} followed by a discussion of the conditions we require on P_{st}. Recall the main requirement is that P_{st} is a semi-norm: a functional that obeys all the rules of a norm except one — for nonzero f we may have $P_{st}(f) = 0$. Some potential choices for smoothing semi-norms are:

1. k-th order Sobolev semi-norm $\quad P_{st} \leftarrow P_{sobolev}(f^{(k)}) = \sqrt{\int_x \left(f^{(k)}(x)\right)^2 dx}$;
2. k-th order total variation $\quad P_{st} \leftarrow TV(f^{(k)}) = \int_x \left|f^{(k+1)}(x)\right| dx$;
3. k-th order Hölder semi-norm $\quad P_{st} \leftarrow P_{holder}(f^{(k)}) = \sup_x \left|f^{(k)}(x)\right|$;
4. k-th order monotonicity $\quad P_{st} \leftarrow P_{mon}(f^{(k)}) \leftarrow \mathbb{I}\left(f;\ \{f : f^{(k+1)} \geq 0\}\right)$;
5. M-th dimensional linear subspace $\quad P_{st} \leftarrow P^M_{lin}(f) = \mathbb{I}\left(f;\ \text{span}\left\{g_1, \ldots, g_M\right\}\right)$;

here $\mathbb{I}$ is a convex indicator function defined as $\mathbb{I}(f;\mathcal{A}) = 0$ if $f \in \mathcal{A}$ and $\mathbb{I}(f;\mathcal{A}) = \infty$ if $f \notin \mathcal{A}$. As implied by the name, P_{st} imposes smoothness or structure on individual components $\widehat{f}_j$. For instance, $P_{sobolev}(f'')$ is a common measure of smoothness; small λ values leads to wiggly fitted functions $\widehat{f}_j$; on the other hand, sufficiently large λ values would lead to each component being a linear function. The convex indicator function, $\mathbb{I}(\cdot)$, can impose specific structural properties on $\widehat{f}_j$; e.g., $P_{mon}(f)$ fits a model with each $\widehat{f}_j$ a non-decreasing function.

The semi-norm requirement for P_{st} is important because: (a) it implies convexity leading to a convex objective function, (b) the first order absolute homogeneity ($P_{st}(\alpha f) = |\alpha| P_{st}(f)$) is needed for the algorithm of Section 3.3 and, (c) the triangle inequality is used throughout the proof of our theoretical results of Section 3.4. For non-sparse GAMs of the form (3.1), the existing literature does not use a semi-norm penalty; a common choice of smoothing penalty is $P_{st}(f) = P^2_{sobolev}(f'')$. In the following subsection, we discuss the issues with using squared semi-norm penalties in high dimensions, particularly their impact on the sparsity of estimated component functions.

3.2.2 Semi-norms vs squared semi-norms

For a semi-norm P_{semi}, using $P_{st} = P^2_{semi}$ in (3.3) can give poor theoretical performance (as noted in Meier et al. (2009) for $P_{semi} = P_{sobolev}$) and, can also be computationally expensive (as disscussed in Section 3.3). In this subsection, we show a surprising result: using a squared semi-norm penalty does not actually lead to a sparse solution.

To be precise, using $P_{st} = P^2_{semi}$ leads to an active set $\mathcal{S} = \{j : \widehat{f}_j \not\equiv 0\}$, for which either $|\mathcal{S}| = 0$ or $|\mathcal{S}| = p$; in contrast, using $P_{st} = P_{semi}$ can give active sets such that $0 < |\mathcal{S}| < p$. To demonstrate this phenomenon, we consider first the univariate problem

$$\widehat{f}_1 \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f\left(x_i\right)\right)^2 + \lambda_{st} P^2_{semi}\left(f\right) + \lambda_{sp} \left\|f\right\|_n , \tag{3.4}$$

and characterize conditions for which $\widehat{f}_1 \equiv 0$. Recall that $\widehat{f}_1$ minimizes the objective in (3.4) if for every direction h, the objective is minimized at $\varepsilon = 0$ along the path $\widehat{f}_1 + \varepsilon h$. In

the following lemma, we state necessary and sufficient conditions for $\widehat{f}_1$ to be 0 (proof in Appendix B.1).

Lemma 3.1. *For $\widehat{f}_1$ defined by (3.4), the following are equivalent: (a) $\widehat{f}_1 \equiv 0$, (b) for every direction $h \in \mathcal{F}$, $|n^{-1}\sum_i y_i h(x_i)/\|h\|_n| \leq \lambda_{sp}$, (c) $\|\boldsymbol{y}\|_n \leq \lambda_{sp}$.*

Condition (c) of Lemma 3.1 is problematic when we consider multiple features in our additive problem (3.3). For additive models, condition (c) implies that sparsity of component $\widehat{f}_j$, does not depend on covariate j. Thus if all smoothing penalties are squared semi-norms then for a given λ_{sp}, there exists a minimizer with either all $\widehat{f}_j \equiv 0$ or all $\widehat{f}_j \not\equiv 0$.

On the other hand, consider the optimization problem

$$\widehat{f}_2 \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f\left(x_i\right)\right)^2 + \lambda_{st} P_{semi}\left(f\right) + \lambda_{sp}\left\|f\right\|_n, \tag{3.5}$$

then we have the following lemma (proof in Appendix B.1).

Lemma 3.2. *For $\widehat{f}_2$ defined by (3.5), the following are equivalent: (a) $\widehat{f}_2 \equiv 0$, (b) for every direction h, there exists some $V \in [-1, 1]$ such that $\left|n^{-1}\sum_i y_i h(x_i)/\|h\|_n - \lambda_{st} V P_{semi}(h)/\|h\|_n\right| \leq \lambda_{sp}$.*

Additionally, if $\|\boldsymbol{y}\|_n \leq \lambda$ then $\widehat{f}_2 = \mathbf{0}$, but the converse is not necessarily true.

Unlike the squared semi-norm penalties, conditions for $\widehat{f}_2 \equiv 0$ involve the feature vector $\boldsymbol{x} = [x_1, \ldots, x_n]^\top$. Thus for an additive model the sparsity of component j depends on both the response vector $\boldsymbol{y}$, and j-th covariate $(x_{1j}, \ldots, x_{nj})$. Consequently, there are many $(\lambda_{sp}, \lambda_{st})$ pairs for which we will have some $\widehat{f}_j \equiv 0$ and some $\widehat{f}_j \not\equiv 0$. Additionally, Lemma 3.2 gives us a conservative value for $\lambda_{max} = \|\boldsymbol{y}\|_n$, i.e., the λ_{sp} value for which all $\widehat{f}_j \equiv 0$.

3.2.3 Relationship of existing methods to `PGSAME`

We now discuss some of the existing methods for sparse additive models in greater detail and, demonstrate that many existing proposals are special cases of our `PGSAME` framework.

One of the first proposals for sparse additive models, SpAM (Ravikumar et al., 2009), uses a basis expansion and solves

$$\underset{\beta_1,\ldots,\beta_j\in\mathbb{R}^M}{\operatorname{argmin}} \left\|\boldsymbol{y}-\sum_{j=1}^{p}\sum_{m=1}^{M}\beta_{jm}\boldsymbol{\psi}_{jm}\right\|_n^2+\lambda\sum_{j=1}^{p}\left\|\sum_{m=1}^{M}\beta_{jm}\boldsymbol{\psi}_{jm}\right\|_n, \tag{3.6}$$

where $\boldsymbol{\psi}_{jm}=[\psi_m(x_{1j}),\ldots,\psi_m(x_{nj})]^T\in\mathbb{R}^n$ for basis functions $\psi_1,\ldots,\psi_M$. This is a `PGSAME` with $P_{st}=\mathbb{I}\left(f;\operatorname{span}\left\{\psi_1,\ldots,\psi_M\right\}\right)$. The SpAM proposal is extended to partially linear models in SPLAM (Lou et al., 2016). There, a similar basis expansion is used, though with the particular choice $\psi_1(x)=x$. The SPLAM estimator solves

$$\underset{\beta_1,\ldots,\beta_j\in\mathbb{R}^M}{\operatorname{argmin}} \left\|\boldsymbol{y}-\sum_{j=1}^{p}\sum_{m=1}^{M}\beta_{jm}\boldsymbol{\psi}_{jm}\right\|_n^2+\lambda_1\sum_{j=1}^{p}\left\|\sum_{m=1}^{M}\beta_{jm}\boldsymbol{\psi}_{jm}\right\|_n+\lambda_2\sum_{j=1}^{p}\left\|\sum_{m=2}^{M}\beta_{jm}\boldsymbol{\psi}_{m}\right\|_n, \tag{3.7}$$

and is also a `PGSAME` with

$$P_{st}=\mathbb{I}\left(f;\operatorname{span}\left\{\psi_1,\ldots,\psi_M\right\}\right)+\sum_{j=1}^{p}\left\|\operatorname{Proj}_{\operatorname{span}(\psi_2,\ldots,\psi_M)}\left(f\right)\right\|_n,$$

where $\operatorname{Proj}_{\mathcal{A}}$ is the projection operator onto the set $\mathcal{A}$. The recently proposed extensions of trend filtering to additive models, is another example (Petersen et al., 2016; Sadhanala and Tibshirani, 2017): these methods can be written in our `PGSAME` framework with $P_{st}(f)=TV(f^{(k)})$.

Koltchinskii and Yuan (2010), Raskutti et al. (2012) and Yuan and Zhou (2015) discuss a similar framework to `PGSAME`s; however, they only consider structural penalties P_{st}, which are norms of Reproducing Kernel Hilbert Spaces (RKHS). Furthermore, they do not discuss efficient algorithms for solving the convex optimization problem. Using properties of RKHS, they note that their estimator is the minimum of a $d=np$ dimensional second order cone program (SOCP). The computation for general-purpose SOCP solvers scales roughly as d^3. Thus for even moderate p and n, these problems quickly become intractable.

Meier et al. (2009) give two proposals: the first solves the optimization problem

$$\underset{f_1,\ldots,f_p\in\mathcal{F}}{\operatorname{argmin}} \left\|\boldsymbol{y}-\sum_{j=1}^{p} f_j\right\|_n^2 + \sum_{j=1}^{p}\lambda_{sp}\sqrt{\|f_j\|_n^2+\lambda_{st}P_{st}^2\left(f_j\right)},$$

and is not a PGSAME; they note that this proposal gives a suboptimal rate. The second is a PGSAME of the form (3.3) with $P_{st}(f) = P_{sobolev}(f'')$. At the time Meier et al. (2009) focused on the first proposal as no computationally efficient method for solving the second was known to them. In a follow-up paper, van de Geer (2010) studied the theoretical properties of a PGSAME with an alternative, *diagonalized smoothness* structural penalty. The diagnolized smoothness penalty for a function with basis expansion $f_{\boldsymbol{\beta}}(x) = \sum_{j=1}^{n}\psi_j(x)\beta_j$, is defined as

$$P_{st}(f_{\boldsymbol{\beta}}) = \left(\sum_{j=1}^{n} j^{2m}\beta_j^2\right)^{1/2}, \tag{3.8}$$

for a smoothness parameter m. All of the above mentioned proposals either fail to provide an efficient computational algorithm or have sub-optimal convergence rates.

Various other proposals do not quite fall in this framework (Chouldechova and Hastie, 2015; Fan et al., 2012; Yin et al., 2012).

3.3 General purpose algorithm

Here we give a general algorithm for fitting PGSAMEs based on proximal gradient descent (Parikh and Boyd, 2014). We begin with some notation. We denote by $\dot{\ell}(y,\theta)$ and $\ddot{\ell}(y,\theta)$ the first and second derivatives of ℓ with respect to θ. For functions $f, g : \mathbb{R}^p \to \mathbb{R}$, let $\langle f, \dot{\ell}(g)\rangle_n \equiv n^{-1}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\{\dot{\ell}(y_i, g(\boldsymbol{x}_i))\}$, $\bar{\dot{\ell}}(g) \equiv n^{-1}\sum_{i=1}^{n}\dot{\ell}(y_i, g(\boldsymbol{x}_i))$ and, $\|f+\dot{\ell}(g)\|_n^2 \equiv n^{-1}\sum_{i=1}^{n}\{f(\boldsymbol{x}_i)+\dot{\ell}(y_i, g(\boldsymbol{x}_i))\}^2$.

We begin with a second order taylor expansion of the loss. For this, we first apply Taylor's theorem to $\ell(y_i, \beta+\theta_{i1}+\ldots+\theta_{ip})$ as a $(p+1)$ variate function of $(\beta, \theta_{i1}, \ldots, \theta_{ip})$. Note that for $|\ddot{\ell}(y,\theta)| \leq L$, the Hessian matrix, H_{p+1}, of $\ell(y_i, \beta+\theta_{i1}+\ldots+\theta_{ip})$ obeys the inequality

$\boldsymbol{a}^T H_{p+1}\boldsymbol{a} \leq (p+1)L\|a\|_2^2$ for all $\boldsymbol{a} \in \mathbb{R}^{p+1}$ (Zhan, 2005). This gives us the following bound:

$$\begin{aligned}
-\mathbb{P}_n\ell\Big(\beta + \sum_{j=1}^p f_j\Big) \leq & -\mathbb{P}_n\ell\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big) \\
& - (\beta - \beta^0)\bar{\dot{\ell}}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big) - \sum_{j=1}^p \Big\langle f_j - f_j^0, \dot{\ell}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)\Big\rangle_n \\
& + \frac{(p+1)L}{2}(\beta - \beta^0)^2 + \sum_{j=1}^p \frac{(p+1)L}{2}\left\|f_j - f_j^0\right\|_n^2,
\end{aligned}$$

which leads to the following majorizing inequality

$$\begin{aligned}
-\mathbb{P}_n\ell\Big(\beta + \sum_{j=1}^p f_j\Big) \leq & \frac{(p+1)L}{2}\Big[\beta - \Big\{\beta^0 + \frac{1}{(p+1)L}\bar{\dot{\ell}}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)\Big\}\Big]^2 \\
& + \sum_{j=1}^p \frac{(p+1)L}{2}\Big\|f_j - \Big\{f_j^0 + \frac{1}{(p+1)L}\dot{\ell}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)\Big\}\Big\|_n^2 + W,
\end{aligned} \tag{3.9}$$

where W is not a function of β or f_j for any j. Instead of minimizing the original problem (3.3), we minimize the majorizing surrogate

$$\begin{aligned}
& \frac{1}{2}\Big[\beta - \Big\{\beta^0 + t\bar{\dot{\ell}}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)\Big\}\Big]^2 + \frac{1}{2}\sum_{j=1}^p \Big\|f_j - \Big\{f_j^0 + t\dot{\ell}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)\Big\}\Big\|_n^2 \\
& \quad + t\lambda^2 \sum_{j=1}^p P_{st}\left(f_j\right) + t\lambda \sum_{j=1}^p \left\|f_j\right\|_n,
\end{aligned} \tag{3.10}$$

where $t = \{(p+1)L\}^{-1}$. Minimizing (3.10) and recentering our Taylor series at the new current iterate, is precisely the proximal gradient recipe. Updating the intercept β, is simply $\widehat{\beta} \leftarrow \beta^0 + t\bar{\dot{\ell}}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)$. Components $f_1, \ldots, f_p$, can be updated in parallel by solving the univariate problems:

$$\widehat{f}_j \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{2}\Big\|\Big\{f_j^0 + t\dot{\ell}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)\Big\} - f\Big\|_n^2 + t\lambda^2 P_{st}\left(f\right) + t\lambda\left\|f\right\|_n. \tag{3.11}$$

At first this problem still appears difficult due to the combination of structure and sparsity penalties. However, the following Lemma shows that things greatly simplify.

Lemma 3.3. *Suppose P_{st} is a semi-norm, and $\boldsymbol{r}$ is an n-vector. Consider the optimization problems*

$$\underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{r} - f\|_n^2 + \lambda_1 P_{st}(f) + \lambda_2 \|f\|_n, \tag{3.12}$$

$$\underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{r} - f\|_n^2 + \lambda_1 P_{st}(f). \tag{3.13}$$

If $\widetilde{f}$ is a solution to (3.13); then $\widehat{f}$ is a solution to (3.12) where $\widehat{f}$ is defined as

$$\widehat{f} = \left(1 - \lambda_2 / \|\widetilde{f}\|_n\right)_+ \widetilde{f}, \tag{3.14}$$

with $(z)_+ = \max(z, 0)$.

The proof is given in Appendix B.2. Thus we can get the solution to (3.11) by solving a problem in the form of (3.13), a classical univariate smoothing problem, and then applying (3.14), the simple soft-scaling operator. Putting things together, our proximal gradient algorithm for solving (3.3) is summarized in Algorithm 1.

Algorithm 1 is simple and can be quite fast: the time complexity is largely determined by the difficulty of solving the univariate smoothing problem of step 5. In many cases this takes $O(n)$ operations, allowing an iteration of proximal gradient descent to run in $O(np)$ operations. Complexity order $O(np)$ is the per-iteration time complexity of state-of-the-art algorithms for the lasso (Friedman et al., 2010; Beck and Teboulle, 2009a).

Any step-size t can be used in Algorithm 1 so long as inequality (3.9) still holds for $f_j^0 \equiv f_j^{k-1}$ and $f_j \equiv f_j^k$ when $(p+1)L$ is replaced by t^{-1}. Note that if $t \leq \{L(p+1)\}^{-1}$ this will always hold. However, often p_{active}^k, the number of j for which either of f_j^{k-1} or f_j^k is non-zero, will be small. In this case $t \leq \{L(p_{active}+1)\}^{-1}$ will satisfy the majorization condition. Since in practice we are interested in sparse models, generally $p_{active}^k \ll p$ and adaptive step-

Algorithm 1 General Proximal Gradient Algorithm for (3.3)

Initialize $f_1^0, \dots f_p^0 \leftarrow \mathbf{0}, \beta^0 \leftarrow 0$, $k \leftarrow 1$; choose a step-size t

while $k \leq max\_iter$ **and** not converged **do**

For each $i = 1, \dots, n$, set

$$\theta_i \leftarrow \beta^{k-1} + \sum_{j=1}^{p} f_j^{k-1}(x_{ij}),$$
$$r_i \leftarrow -\dot{\ell}(y_i, \theta_i).$$

Update

$$\beta^k \leftarrow \beta^{k-1} - t\sum_{i=1}^{n} r_i.$$

for $j = 1, \dots, p$ **do**

Set

$$f_j^{inter} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{2} \left\| \left(f_j^{k-1} - t\boldsymbol{r}\right) - f \right\|_n^2 + t\lambda^2 P_{st}(f). \tag{3.15}$$

Update

$$f_j^k \leftarrow \left(1 - t\lambda / \|f_j^{inter}\|_n\right)_+ f_j^{inter}.$$

end for

end while

return $\beta^k, f_1^k, \dots, f_p^k$

size optimization can be quite useful (Beck and Teboulle, 2009b) . This algorithm can also take advantage of Nesterov-style acceleration (Nesterov, 2007). Using acceleration changes the worst-case convergence rate after k steps from $O\left(k^{-1}\right)$ to $O\left(k^{-2}\right)$.

An important special case is the least squares loss $-\ell(y,\theta) = (y-\theta)^2$. In this case, we can use a block coordinate descent algorithm which can be more efficient than Algorithm 1, and does not require a step-size. In full detail, for a least squares loss we obtain Algorithm 2.

Algorithm 2 Block Coordinate Descent for Least Squares Loss

Initialize $f_1^0, \ldots f_p^0 \leftarrow \mathbf{0}, \beta^0 \leftarrow 0$, $\boldsymbol{r} \leftarrow \boldsymbol{y}$, $k \leftarrow 1$
while $k \leq max\_iter$ **and** not converged **do**
  Update

$$\beta^k \leftarrow n^{-1}\sum_{i=1}^{n} r_i, \quad \boldsymbol{r} \leftarrow \boldsymbol{r} - \beta^k \mathbf{1}.$$

  for $j = 1, \ldots, p$ **do**
    Set $\boldsymbol{r}_{-j}$ as

$$r_{-j,i} = r_i + f_j^{k-1}(x_{ij}).$$

    Update

$$f_j^k \leftarrow \left(1 - t\lambda/\|f_j^{inter}\|_n\right)_+ f_j^{inter},$$

    where

$$f_j^{inter} \leftarrow \underset{f\in\mathcal{F}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{r}_{-j} - f\right\|_n^2 + t\lambda^2 P_{st}\left(f\right). \tag{3.16}$$

    Update $\boldsymbol{r}$ to

$$r_i \leftarrow r_{-j,i} + f_j^k(x_{ij}).$$

  end for
end while
return $\beta^k, f_1^k, \ldots, f_p^k$

As noted above, the main computational hurdle is solving the univariate problem (3.13); we discuss this step in greater detail for various smoothness penalties in the following subsection.

3.3.1 Solving the univariate sub-problem

For many semi-norm smoothers there are already efficient solvers for solving (3.13): with the k-th order total variation penalty, (3.13) can be solved exactly in $2n$ operations for $k = 0$ (Johnson, 2013), or iteratively in roughly $O((k+1)n)$ operations for $k \geq 1$ (Ramdas and Tibshirani, 2015); with the convex indicator of an M dimensional linear subspace, (3.13) can be solved in $O(M^2 n)$ operations using linear regression; using a monotonicity indicator, (3.13) can be solved with the pool adjacent violators algorithm in $O(n)$ operations (Ayer et al., 1955).

For many other choices of P_{st}, we do not have efficient algorithms for solving (3.13); however, we might have fast algorithms for the slightly different optimization problem:

$$\widetilde{f}_{\widetilde{\lambda}} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{r} - f\|_n^2 + \widetilde{\lambda} P_{st}^{\nu}(f), \tag{3.17}$$

for $\nu > 1$. For example, the k-th order Sobolev penalty (Wahba, 1990) can be solved exactly in $O(kn)$ operations for $\nu = 2$. In the following Lemma, we show that the solution to (3.17) can be leveraged to solve the harder problem (3.13).

Lemma 3.4. *Given an n-vector $\boldsymbol{r}$, a convex linear space $\mathcal{F}$ over the field $\mathbb{R}$, and real $\nu > 1$, consider the optimization problems:*

$$\widehat{f}_{\lambda} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{r} - f\|_n^2 + \lambda P_{st}(f); \tag{3.18}$$

$$\widetilde{f}_{\lambda} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{r} - f\|_n^2 + \lambda P_{st}^{\nu}(f); \tag{3.19}$$

$$f_{null} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{r} - f\|_n^2 + \mathbb{I}(f \in \mathcal{F} : P_{st}(f) = 0); \tag{3.20}$$

$$f_{interp} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} P_{st}^{\nu}(f) + \mathbb{I}(r_i = f(x_i) \text{ for all } i), \tag{3.21}$$

where $P_{st}(\cdot)$ is a semi-norm on $\mathcal{F}$. Assume that the directional derivative

$$\nabla_h P_{st}^{\nu}(f) = \lim_{\varepsilon \to 0} \frac{P_{st}^{\nu}(f + \varepsilon h) - P_{st}^{\nu}(f)}{\varepsilon},$$

exists for all $h \in \mathcal{F}$. If $P_{st}(\widehat{f}_\lambda) \neq 0$ and $\nu\widetilde{\lambda} P_{st}^{\nu-1}(\widetilde{f}_{\widetilde{\lambda}}) = \lambda$, then $\widehat{f}_\lambda = \widetilde{f}_{\widetilde{\lambda}}$.

To determine if $P_{st}(\widehat{f}) = 0$, let $\mathcal{F} = \mathcal{F}_1 \oplus \mathcal{F}_2$, where $\oplus$ is such that, for all $f \in \mathcal{F}$ we have $f = f_0 + f_\perp$ where $\langle f_0, f_\perp \rangle_n = 0$ and $P_{st}(f) = P_{st}(f_\perp)$. Furthermore, let P_{st}^* be the dual norm over $\mathcal{F}_2$, given by

$$P_{st}^*(f_\perp) = \sup \quad \Big\{ |\langle f_\perp, f'_\perp \rangle_n| : P_{st}(f'_\perp) \leq 1, f'_\perp \in \mathcal{F}_2 \Big\}. \tag{3.22}$$

Then $f_{interp} - f_{null} \in \mathcal{F}_2$ and $\widehat{f}_\lambda = f_{null}$ if $\lambda \geq P_{st}^*(f_{interp} - f_{null})$.

The proof is given in Appendix B.2. This lemma allows us to first check if we should shrink entirely to a null fit with $P_{st}(\widehat{f}) = 0$ (usually a finite dimensional function), based on the dual semi-norm of the interpolating function f_{interp}. If we do not shrink to $P_{st}(\widehat{f}) = 0$, then there is an equivalence between $\widehat{f}$ and $\widetilde{f}$; and the problem is reduced to finding $\widetilde{\lambda}$ with $\nu\widetilde{\lambda} P_{st}^{\nu-1}(\widetilde{f}_{\widetilde{\lambda}}) = \lambda$ for the originally specified λ. This can be done in a number of ways; most simply by a combination of grid search and then local bisection noting that a) we need not try any $\widetilde{\lambda}$-values above $\lambda_{max} \equiv P_{st}(f_{interp})$, and b) $\widetilde{\lambda} P_{st}(\widetilde{f}_{\widetilde{\lambda}})$ is a smooth function of $\widetilde{\lambda}$. In fact, the grid search will often be unnecessary as we will generally have a good guess from the previous iterate of the proximal gradient algorithm, and can leverage the fact that $P_{st}(\widetilde{f}_{\widetilde{\lambda}})$ and $P_{st}(\widehat{f}_\lambda)$ are both smooth functions of $\boldsymbol{r}$.

We now discuss the dual norm (3.22), in greater detail. Consider the case where $P_{st}(f) = \|D\boldsymbol{f}\|_q$ for some matrix $D \in \mathbb{R}^{M \times n}$, vector $\boldsymbol{f} = [f(x_1), \ldots, f(x_n)]^\top \in \mathbb{R}^n$, and $q \geq 1$. Such penalties are common in the literature e.g., when P_{st} is the Sobolev semi-norm, total variation

norm, or any norm of a RKHS. Then the dual norm is given by

$$P_{st}^*(f) = \|D(D^\top D)^- \boldsymbol{f}\|_{\widetilde{q}}, \tag{3.23}$$

where $(D^\top D)^-$ is the Moore-Penrose pseudo inverse of $D^\top D$ and $\widetilde{q}$ satisfies $1/q + 1/\widetilde{q} = 1$.

3.4 Theoretical results

Here we prove rates of convergence for `PGSAME`s, estimators that fall within our framework (3.3). We first present the so-called *slow rates*, which require few assumptions, followed by *fast rates*, which require a compatibility and margin condition (defined and discussed below). Our fast rates match the minimax rates under Gaussian data with a least squares loss (Raskutti et al., 2009) and, our slow rates can be seen as an additive generalization of the lasso slow rates (Dalalyan et al., 2014). For both slow and fast rates, we first present a deterministic result; this result simply states that if we are within a special set, $\mathcal{T}$, then the convergence rates hold. In a following corollary we show that under suitable conditions (stated and discussed below) on the loss function, smoothness penalty, and data, we lie in $\mathcal{T}$ with high probability. Another feature of our results is that we allow for mean model misspecification with an additional *approximation error* term in the convergence rates; if the true mean model is additive then this term disappears.

To the best of our knowledge, the closest results to our work were established by Koltchinskii and Yuan (2010). However, they consider a more restrictive setting of Reproducing Kernel Hilbert Spaces (RKHS); where each additive component f_j belongs to a RKHS $\mathcal{H}_j$, and P_{st} is the norm on $\mathcal{H}_j$. Our work gives these rates for all semi-norm penalties and function classes $\mathcal{F}$, associated with certain non-restrictive entropy conditions. Before presenting the main results, we present some notation and definitions which will be used throughout the section.

3.4.1 Definitions and notation

We consider here properties of the solution to

$$\widehat{\beta}, \widehat{f}_1 \ldots, \widehat{f}_p \leftarrow \underset{\beta \in \mathcal{R}, \{f_j\}_{j=1}^p \in \mathcal{F}}{\arg\min} - \mathbb{P}_n \ell\Big(\beta + \sum_{j=1}^p f_j\Big) + \lambda \sum_{j=1}^p \left\{ \|f_j\|_n + \lambda P_{st}\left(f_j\right) \right\}, \tag{3.24}$$

where $\mathcal{R} \subseteq \mathbb{R}$ and $\mathcal{F}$ is some univariate function class. Note that in (3.24) we optimize β over $\mathcal{R}$; this is because we need $\mathcal{R}$ to be a bounded for proving the slow rates, the stronger *compatibility condition* allows us to take $\mathcal{R} = \mathbb{R}$ for proving fast rates.

For a function $f(\boldsymbol{x}) = \beta + \sum_{j=1}^p f_j(x_j)$ we use the shorthand notation

$$I(f) \equiv \sum_{j=1}^p \left[\|f_j\|_n + \lambda P_{st}\left(f_j\right) \right], \tag{3.25}$$

which defines a semi-norm on the function f. Furthermore, for any index set $\mathcal{S} \subset \{1, \ldots, p\}$ we define $I_{\mathcal{S}}(f)$ as $I_{\mathcal{S}}(f) = \sum_{j \in \mathcal{S}} \left[\|f_j\|_n + \lambda P_{st}(f_j) \right]$. We denote the target function by f^0 where

$$f^0 \leftarrow \underset{f \in \mathcal{F}^0}{\arg\min} - \mathbb{P}\ell\left(f\right), \tag{3.26}$$

for some function class $\mathcal{F}^0$ and, where $\mathbb{P}\ell(f) = n^{-1} \sum_{i=1}^n \mathbb{E}\left\{\ell(y_i, f(\boldsymbol{x}_i))\right\}$. We say the target function belongs to some class $\mathcal{F}^0$ to signify that f^0 does not need to belong to $\mathcal{F}$. We require no assumptions on the class $\mathcal{F}^0$ for the slow-rates of Theorem 3.1; we can take $\mathcal{F}^0$ to be the class of all measurable functions. For the fast rates we will require the margin condition on a subset of $\mathcal{F}^0$.

We define the *excess risk* for a function f as $\mathcal{E}(f) = \mathbb{P}\left(\ell(f^0) - \ell(f)\right)$, and we denote by $\nu_n(\cdot)$ the *empirical process term*, which is defined as

$$\nu_n(f) = \left(\mathbb{P}_n - \mathbb{P}\right)\left(-\ell(f)\right) = -\frac{1}{n} \sum_{i=1}^n \left\{ \ell(y_i, f(\boldsymbol{x}_i)) - \mathbb{E}\ell(y_i, f(\boldsymbol{x}_i)) \right\}. \tag{3.27}$$

Define the *δ-covering number*, $N(\delta, \mathcal{F}, \|\cdot\|_Q)$, as the size of the smallest δ-cover of $\mathcal{F}$

with respect to the norm $\|\cdot\|_Q$ induced by measure Q. We denote the *δ-entropy* of $\mathcal{F}$ by $H(\delta, \mathcal{F}, \|\cdot\|_Q) \equiv \log N(\delta, \mathcal{F}, \|\cdot\|_Q)$. Given fixed covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$, we denote the empirical measure by Q_n where $Q_n = n^{-1}\sum_{i=1}^n \delta_{\boldsymbol{x}_i}$, and for covariate j; we denote by $Q_{j,n}$ the empirical measure of $x_{1,j}, \ldots, x_{n,j}$. We define two different types of entropy bounds for a function class $\mathcal{F}$.

Definition 1 (Logarithmic Entropy)**.** *A univariate function class, $\mathcal{F}$, is said to have a logarithmic entropy bound if, for all $j = 1, \ldots, p$, and $\gamma > 0$, we have*

$$H(\delta, \{f_j \in \mathcal{F} : \|f_j\|_n + \gamma P_{st}(f_j) \leq 1\}, \|\cdot\|_{Q_{j,n}}) \leq A_0 T_n \log(1/\delta + 1), \tag{3.28}$$

for some constant A_0, and parameter T_n.

Definition 2 (Polynomial Entropy with Smoothness)**.** *A univariate function class, $\mathcal{F}$, is said to have a polynomial entropy bound with smoothness if, for all $j = 1, \ldots, p$ and $\gamma > 0$, we have*

$$H(\delta, \{f_j \in \mathcal{F} : \|f_j\|_n + \gamma P_{st}(f_j) \leq 1\}, \|\cdot\|_{Q_{j,n}}) \leq A_0(\delta\gamma)^{-\alpha}, \tag{3.29}$$

for some constant A_0, parameter $\alpha \in (0, 2)$.

The concept of entropy is commonly used in the literature, particularly in nonparametric statistics and empirical processes to quantify the size of function classes. The logarithmic entropy bound (3.28), holds for most finite dimensional classes of dimension T_n. For instance, it holds for $\mathcal{F} = L^2(\mathbb{R})$ with $P_{st}(f_j) = \mathbb{I}(f_j; \text{span}\{x, x^2, \ldots, x^{T_n}\})$. The bound (3.29) commonly holds for broader function classes, e.g., for $\mathcal{F} = L^2([0,1])$ with $P_{st}(f_j) = P_{sobolev}(f^{(k)})$ and $\alpha = 1/k$.

To simplify our presentation of bounds on the convergence rate, we use $A \precsim B$ to denote $A \leq cB$ for some constant $c > 0$. We write $A \asymp B$ if $A \precsim B$ and $B \precsim A$.

3.4.2 Main results

We now present our main results: upper bounds for the excess risk of PGSAMEs, i.e., bounds for $\mathcal{E}(\widehat{\beta}+\sum_{j=1}^{p}\widehat{f}_j)$. The following theorem shows that $\mathcal{E}(\widehat{\beta}+\sum_{j=1}^{p}\widehat{f}_j) \precsim \lambda$ over a special set $\mathcal{T}$. In the corollary that follows, we show that for appropriate λ values, and certain type of loss functions, we are within $\mathcal{T}$ with high probability.

Theorem 3.1 (Slow Rates for PGSAME)**.** *Let $\widehat{f}=\widehat{\beta}+\sum_{j=1}^{p}\widehat{f}_j$ be as defined in (3.24), and let $f^*=\beta^*+\sum_{j=1}^{p}f_j^*$ be an arbitrary additive function with $\sum_{i=1}^{n}f_j^*(x_{ij})=0$ and $\beta^*\in\mathcal{R}$. Assume that $-\ell(\cdot)$ and P_{st} are convex and that $\sup_{\beta\in\mathcal{R}}|\beta|<R$. Define M^* such that*

$$\rho M^* = \mathcal{E}(f^*) + 2\lambda I(f^*) + 2R\rho, \tag{3.30}$$

where $\lambda \geq 4\rho$. Furthermore, define the set $\mathcal{T}$ as follows

$$\mathcal{T} = \left\{Z_{M^*} \leq \rho(M^*+2R)\right\}, \text{ where } Z_{M^*} = \sup_{I(f-f^*)\leq M^*} \left|\nu_n(f)-\nu_n(f^*)\right|.$$

Then, on the set $\mathcal{T}$ we have

$$\mathcal{E}(\widehat{f}) + \lambda I(\widehat{f}-f^*) \leq \rho M^* + \rho(2R) + 2\lambda I(f^*) + \mathcal{E}(f^*).$$

Corollary 3.1.1. *Let $\widehat{f}$, f^* and $\mathcal{R}$ be as defined in Theorem 3.1. Assume that for any function f the loss $\ell(\cdot)$ is such that*

$$-\ell(f) = -\ell(y_i, f(\boldsymbol{x}_i)) = a y_i f(\boldsymbol{x}_i) + b(f(\boldsymbol{x}_i)), \tag{3.31}$$

for some $a\in\mathbb{R}\backslash\{0\}$ and function $b:\mathbb{R}\to\mathbb{R}$. Further assume that for $i=1,\ldots,n$, $y_i-\mathbb{E}[y_i]$ are uniformly sub-Gaussian, i.e.,

$$\max_{i=1,\ldots,n} K^2\left(\mathbb{E}e^{(y_i-E[y_i])^2/K^2}-1\right) \leq \sigma_0^2. \tag{3.32}$$

Finally, suppose $\mathcal{E}(f^*) = O(\lambda)$ *and* $I(f^*) = O(1)$. *Then with probability at-least* $1 - 2\exp(-C_1 n\rho^2) - C\exp(-C_2 n\rho^2)$ *we have the following cases:*

1. *If* $\mathcal{F}$ *has a logarithmic entropy bound, then for* $\lambda \asymp \rho \asymp \kappa \max\left(\sqrt{\frac{T_n}{n}}, \sqrt{\frac{\log p}{n}}\right)$,

$$\mathcal{E}(\widehat{f}) + \lambda I(\widehat{f} - f^*) \precsim \max\left(\sqrt{\frac{T_n}{n}}, \sqrt{\frac{\log p}{n}}\right), \tag{3.33}$$

with constants $\kappa = \kappa(a, K, \sigma_0, A_0)$, $C_1 = C_1(K, \sigma_0)$, $C = C(K, \sigma_0)$ *and* $C_2 = C_2(C, \kappa)$.

2. *If* $\mathcal{F}$ *has a polynomial entropy bound with smoothness, then for* $\lambda \asymp \rho \asymp \kappa \max\left(n^{-\frac{1}{2+\alpha}}, \sqrt{\frac{\log p}{n}}\right)$,

$$\mathcal{E}(\widehat{f}) + \lambda I(\widehat{f} - f^*) \precsim \max\left(n^{-\frac{1}{2+\alpha}}, \sqrt{\frac{\log p}{n}}\right), \tag{3.34}$$

with constants $\kappa = \kappa(a, K, \sigma_0, A_0, \alpha)$, $C_1 = C_1(K, \sigma_0)$, $C = C(K, \sigma_0)$ *and* $C_2 = C_2(C, \kappa)$.

We now proceed to show the fast rates of convergence. To establish these rates, we require *compatibility* and *margin* conditions. The compatibility condition, is based on the idea that $I(f)$ and $\|f\|$ are somehow compatible for some norm $\|\cdot\|$. This condition is common in the high-dimensional literature for proving fast rates (see van de Geer and Bühlmann (2009) for a discussion of compatibility and related conditions for the lasso). The margin condition, is based the idea that if $\mathcal{E}(f)$ is small then $\|f - f^0\|$ should also be small. This is another common condition in the literature for handling general convex loss functions (see e.g., Negahban et al., 2011; van de Geer, 2008).

Definition 3 (Compatibility Condition)**.** *The compatibility condition is said to hold for an index set* $\mathcal{S}_* \subset \{1, 2, \ldots, p\}$, *with compatibility constant* $\phi(\mathcal{S}_*) > 0$, *if for all* $\gamma > 0$ *and all*

functions f of the form $f(\boldsymbol{x}) = \beta + \sum_{j=1}^{p} f_j(x_j)$ that satisfy $\sum_{j \in \mathcal{S}_*^c} \|f_j\|_n + \gamma \sum_{j=1}^{p} P_{st}(f_j) \leq |\beta| + 3 \sum_{j \in \mathcal{S}_*} \|f_j\|_n$, it holds that

$$|\beta|/2 + \sum_{j \in \mathcal{S}_*} \|f_j\|_n \leq \|f\| \sqrt{|\mathcal{S}_*|}/\phi(\mathcal{S}_*), \tag{3.35}$$

for some norm $\|\cdot\|$.

Definition 4 (Margin Condition)**.** *The margin condition holds if there is strictly convex function G such that $G(0) = 0$ and for all $f \in \mathcal{F}^0_{local} \subset \mathcal{F}^0$ we have*

$$\mathcal{E}(f) \geq G(\|f - f^0\|), \tag{3.36}$$

for some norm on the function class $\mathcal{F}^0$; here $\mathcal{F}^0_{local}$ is a neighborhood of f^0 based on some norm (e.g., $\mathcal{F}^0_{local} = \{f : \|f - f^0\|_\infty \leq \eta\}$). In typical cases, the margin condition holds with $G(u) = cu^2$, for a positive constant c. We refer to this special case as the quadratic margin condition.

The following theorem establishes the bound $\mathcal{E}(\widehat{\beta} + \sum_{j=1}^{p} \widehat{f}_j) \precsim s\lambda^2$, where λ is the slow rate of Theorem 3.1, and s is the number of non-zero components of $f^* = \beta + \sum_{j=1}^{p} f_j^*$, a sparse additive approximation of f^0. As in Theorem 3.1, the bound holds over a set $\mathcal{T}$; the corollary following Theorem 3.1 shows that we lie in $\mathcal{T}$ with high probability.

Theorem 3.2 (Fast Rates for `PGSAME`)**.** *Suppose $-\ell(\cdot)$ and P_{st} are convex functions and with $\widehat{f}$ and f^* as defined in Theorem 3.1. Assume that f^* is sparse with $|\mathcal{S}_*| = s$ where $\mathcal{S}_* = \{j : f_j^* \neq 0\}$, and that the compatibility condition holds for $\mathcal{S}_*$. Further assume the quadratic margin condition holds with constant c, and that for a function $f(\boldsymbol{x}) = \beta + \sum_{j=1}^{p} f_j(x_j)$, $f \in \mathcal{F}^0_{local}$ iff $|\beta - \beta^*| + I(f - f^*) \leq M^*$. The constant M^* is defined as*

$$\rho M^* = \mathcal{E}(f^*) + \frac{16 s \lambda^2}{c \phi^2(\mathcal{S}_*)} + 2\lambda^2 \sum_{j \in \mathcal{S}_*} P_{st}(f_j^*),$$

and ρ is such that $\lambda \geq 8\rho$. Furthermore, define the set $\mathcal{T}$ as

$$\mathcal{T} = \{Z_{M^*} \leq \rho M^*\}, \text{ where } Z_{M^*} = \sup_{|\beta-\beta^*|+I(f-f^*)\leq M^*} |\nu_n(f) - \nu_n(f^*)|.$$

Then, on the set $\mathcal{T}$, we have

$$\mathcal{E}(\widehat{f}) + \lambda I(\widehat{f} - f^*) \leq 4\rho M^* = 4\mathcal{E}(f^*) + \frac{64s\lambda^2}{c\phi^2(\mathcal{S}_*)} + 8\lambda^2 \sum_{j\in\mathcal{S}_*} P_{st}(f_j^*). \tag{3.37}$$

Corollary 3.2.1. *Let $\widehat{f}$ and f^* be as defined in Theorem 3.1 and assume the conditions of Theorem 3.2. Furthermore, for any function f assume the loss $\ell(\cdot)$ is such that*

$$-\ell(f) = -\ell(y_i, f(\boldsymbol{x}_i)) = ay_i f(\boldsymbol{x}_i) + b(f(\boldsymbol{x}_i)), \tag{3.38}$$

for some $a \in \mathbb{R}\backslash\{0\}$ and function $b : \mathbb{R} \to \mathbb{R}$. Further assume that for $i = 1, \ldots, n$, $y_i - \mathbb{E}[y_i]$ are uniformly sub-Gaussian, i.e.

$$\max_{i=1,\ldots,n} K^2 \left[\mathbb{E}\exp\left\{(y_i - \mathbb{E}[y_i])^2/K^2\right\} - 1\right] \leq \sigma_0^2. \tag{3.39}$$

Finally suppose $\mathcal{E}(f^*) = O(s\lambda^2/\phi^2(\mathcal{S}_*))$ and $s^{-1}\sum_{j\in\mathcal{S}_*} P_{st}(f_j^*) = O(1)$. Then, with probability at-least $1 - 2\exp(-C_1 n\rho^2) - C\exp(-C_2 n\rho^2)$, we have the following cases:

1. *If $\mathcal{F}$ has a logarithmic entropy bound, for $\lambda \asymp \rho \asymp \kappa \max\left(\sqrt{\frac{T_n}{n}}, \sqrt{\frac{\log p}{n}}\right)$,*

$$\mathcal{E}(\widehat{f}) + \lambda I(\widehat{f} - f^*) \precsim \max\left(s\frac{T_n}{n}, s\frac{\log p}{n}\right), \tag{3.40}$$

with constants $\kappa = \kappa(a, K, \sigma_0, A_0)$, $C_1 = C_1(K, \sigma_0)$, $C = C(K, \sigma_0)$ and $C_2 = C_2(C, \kappa)$.

2. If $\mathcal{F}$ has a polynomial entropy bound with smoothness, then for $\lambda \asymp \rho \asymp \kappa \max\left(n^{-\frac{1}{2+\alpha}}, \sqrt{\frac{\log p}{n}}\right)$,

$$\mathcal{E}(\widehat{f}) + \lambda I(\widehat{f} - f^*) \precsim \max\left(sn^{-\frac{2}{2+\alpha}}, s\frac{\log p}{n}\right), \tag{3.41}$$

with constants $\kappa = \kappa(a, K, \sigma_0, A_0, \alpha)$, $C_1 = C_1(K, \sigma_0)$, $C = C(K, \sigma_0)$ *and* $C_2 = C_2(C, \kappa)$.

We will discuss the significance of our theoretical results in the next subsection by specializing them to some well studied special cases. Before discussing the specializations though, we conclude this section by a further generalization of Theorem 3.2. We will now assume the more general margin condition for which we need to define the additional notion of a *convex conjugate*.

Definition 5 (Convex Conjugate)**.** *Let G be a strictly convex function on $[0, \infty)$ with $G(0) = 0$. The convex conjugate of G, denoted by H, is defined as*

$$H(v) = \sup_u \{uv - G(u)\}, \ v \geq 0. \tag{3.42}$$

For the special case of $G(u) = cu^2$, one has $H(v) = v^2/(4c)$.

Theorem 3.3 (Fast Rates)**.** *Assume the conditions of Theorem 3.2 and define M^* as*

$$\rho M^* = \mathcal{E}(f^*) + H\left(\frac{8\lambda\sqrt{s}}{\phi(\mathcal{S}_*)}\right) + 2\lambda^2 \sum_{j \in \mathcal{S}_*} P_{st}(f_j^*), \tag{3.43}$$

where $H(\cdot)$ is the convex conjugate of G. Then on the set $\mathcal{T}$, we have

$$\mathcal{E}(\widehat{f}) + \lambda I(\widehat{f} - f^*) \leq 4\rho M^*. \tag{3.44}$$

Note on convex indicator penalties: The above results do not directly extend to some

convex indicator penalties. For some convex indicator penalties, such as $P_{st}(f) = \mathbb{I}(f; \{f : f' \geq 0\})$, we require a third type of entropy condition:

Definition 6 (Polynomial Entropy without Smoothness)**.** *The univariate function class, $\mathcal{F}$, is said to have a polynomial entropy with smoothness bound if for all $j = 1, \ldots, p$ we have*

$$H(\delta, \{f_j \in \mathcal{F} : \|f_j\|_n + \gamma P_{st}(f_j) \leq 1\}, \|\cdot\|_{Q_{j,n}}) \leq A_0 \delta^{-\alpha}, \tag{3.45}$$

for some constant A_0, parameter $\alpha \in (0, 2)$ and all $\gamma > 0$.

Our results do not extend to convex indicator penalties because our proof relies on the fact that $f_j - f_j^* \in \mathcal{F}$ for $f_j, f_j^* \in \mathcal{F}$; function classes with polynomial entropy without smoothness do not usually have this property. We defer the extension to convex indicator structural penalties to future work.

3.4.3 Special cases of `PGSAME`

In this subsection, we illustrate the main strength of our framework, namely its generalizability. We specialize our theoretical results to, various existing proposals for sparse additive models, low-dimensional additive models, and fully non-parametric regression problems. We also specialize our results to GLMs in low and high dimensions.

As discussed in Section 3.2.3, Meier et al. (2009) proposed a `PGSAME` with $P_{st}(f) = P_{sobolev}(f'')$. However, in their theoretical analysis they considered a larger class of structural penalties, namely penalties which satisfy the *polynomial entropy with smoothness* condition (3.29). Meier et al. (2009) establish a convergence rate of the order $s(\log p/n)^{2/(2+\alpha)}$ which is sub-optimal compared to our fast rate (3.41). Established rates for the diagnolized smoothness penalty of van de Geer (2010), were also sub-optimal and of the order $s(\log p)n^{-2/(2+\alpha)}$. Our work bridges the following gaps in the theoretical work of Meier et al. (2009) and van de Geer (2010): (a) we establish minimax rates under identical compatibility conditions, (b) we extend their result beyond least squares loss functions and, (c) we establish *slow rates* under

virtually no assumptions. Another special case is trend filtering additive models (Petersen et al., 2016; Sadhanala and Tibshirani, 2017). Theorem 3.1 improves upon the slow rates established by Petersen et al. (2016) of the order $\sqrt{\log(np)/n}$; Theorem 3.2 establishes fast rates solving the problem which Sadhanala and Tibshirani (2017) characterized as "... still an open problem".

Additive models in low dimensions can also be considered by simply setting $\mathcal{S}_* = \{1, \ldots, p\}$. In this case, the compatibility condition holds and we recover the usual convergence rates for generalized additive models of the form $pn^{-\frac{2}{2+\alpha}}$. With this, we recover the special case of univariate nonparametric regression, i.e., with $p = 1$. Another interesting case we recover is the multivariate nonparametric problem; to see this, say we have a single (but multivariate) component function $f_1 : \mathbb{R}^p \to \mathbb{R}$. For various choices of P_{st}, the bound (3.29) holds with $\alpha = p/m$ for some smoothness parameter m. Thus, we recover the usual nonparametric rate $n^{-2m/(2m+p)}$.

Finally, we also consider parametric regression models as special cases of `PGSAME`. Using a convex indicator for P_{st}, we can constrain each f_j to be a linear function leading to GLMs. For low-dimensional GLMs, Corollary 3.2.1 gives us the usual parametric rate p/n. For high-dimensional GLMs, not only does our theorem recover the lasso rates but our compatibility condition also matches the usual lasso compatibility condition (Bühlmann and van de Geer, 2011).

3.5 Simulation study

In this section we conduct a simulation study to compare estimators obtained by the following choices of smoothness penalty, $P_{st}(\cdot)$.

1. **SpAM (Ravikumar et al., 2009).** $P_{st}(f) = \mathbb{I}(f; \text{span}\{\psi_1, \ldots, \psi_M\})$ for $M \in \{3, 6, 10, 20, 30, 50, 80\}$. This is implemented in the `R` package `SAM` (Zhao et al., 2014).

2. **SSP (Meier et al., 2009).** $P_{st}(f) = \sqrt{\int_x (f^{(2)}(x))^2 \, dx}$, the Sobolev smoothness penalty (SSP). We implemented this using the algorithm and results of Section 3.3.

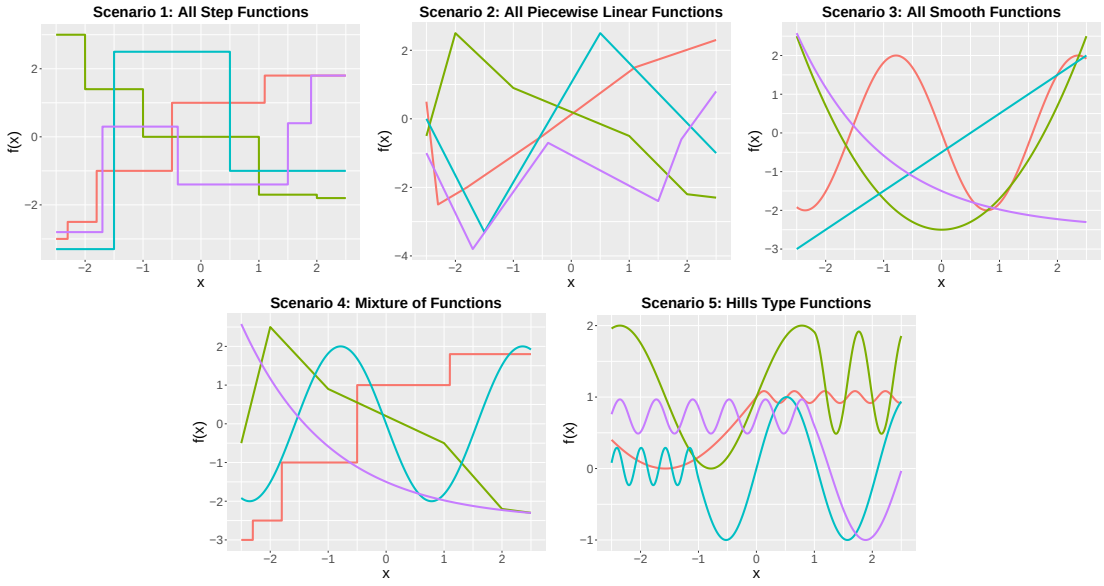


Figure 3.1: Plot of the 4 signal functions for each of the five simulation settings.

3. **TF (Sadhanala and Tibshirani, 2017).** $P_{st}(f) = \int_x |f^{(k+1)}(x)|\, dx$ for $k \in \{0, 1, 2\}$, trend filtering for additive models. We implemented this method using the algorithm of Section 3.3 where the univariate sub-problem (3.15) was solved using the `R` package `glmgen` (Arnold et al., 2014).

We simulate data for different simulation scenarios as follows: for a given sample size, n, and number of covariates, p, we draw 50 different $n \times p$ training design matrices $\boldsymbol{X}$ where each element is drawn from $\mathcal{U}(-2.5, 2.5)$. We replicate each of the 50 design matrices 10 times leading to a total of 500 design matrices. The response was generated as $y_i = f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) + \varepsilon_i$ $(i = 1, \ldots, n)$ where $\varepsilon_i \sim \mathcal{N}(0, 1)$. The remaining covariates are noise variables. We also generate an independent test set for each replicate with sample size $n/2$. We vary the sample size, $n \in \{100, 200, \ldots, 800\}$ and consider both, a low-dimensional $(p = 6)$ and high-dimensional $(p = 100)$ setting. We consider 5 different choices of the signal functions which we present graphically in Figure 3.1.

We implement each of the methods for a sequence of 50 λ values on the training set,

and select the tuning parameter λ^* which minimizes the test error ($\|\boldsymbol{y}_{test} - \widehat{\boldsymbol{y}}\|_n^2$). For the estimated model $\widehat{f}_{\lambda^*}$, we report the mean square error (MSE; $\|\widehat{f}_{\lambda^*} - f^0\|_n^2$) as a function of n.

In Figures 3.2 and 3.3, we plot the MSE as a function of n for the low and high-dimensional setting, respectively. For each simulation scenario, we plot the performance of SpAM for three different choices of M (low, moderate and high number of basis functions, M). Between the low and high-dimensional settings, we observe similar relative performances between the methods, with more variability of results in the high-dimensional setting. While there was no uniformly better method, we noted that for all, except Scenario 1, the Sobolev smoothness penalty and trend filtering of orders 1 and 2 had comparably good performance. Unsurprisingly, trend filtering of order 0 exhibited superior performance in Scenario 1 since each component is piecewise constant. In each scenario, we note the bias-variance trade-off of SpAM and its dependence on M: too small or high values of M led to high prediction error compared to other methods.

The dependence on M for SpAM, is further illustrated in Figure 3.4, where we plot functions estimated by SpAM for high-dimensional Scenario 4 with $n = 500$. We clearly see under fitted estimates for $M = 3$ (especially for the piecewise constant and linear functions) and high variance for $M = 30$. In the same figure, we also plot functions estimated by the SSP; SSP estimates exhibit a similar bias to that of SpAM with $M = 10$, but with a substantially smaller variance. Figure 3.5, similarly plots fitted example functions for trend filtering. Trend filtering with $k = 0$ estimates the piecewise constant function well, but estimating the other f_j's by piecewise constant functions incurs additional variance. Trend filtering with $k = 1$ and 2 estimates all other signal functions well.

3.6 Data analysis

We apply the methods from Section 3.5, to predict the value of owner-occupied homes in the suburbs of Boston using census data from 1970. The data consists of $n = 506$ measurements and 10 covariates, and has been studied in the additive models literature (Ravikumar et al.,

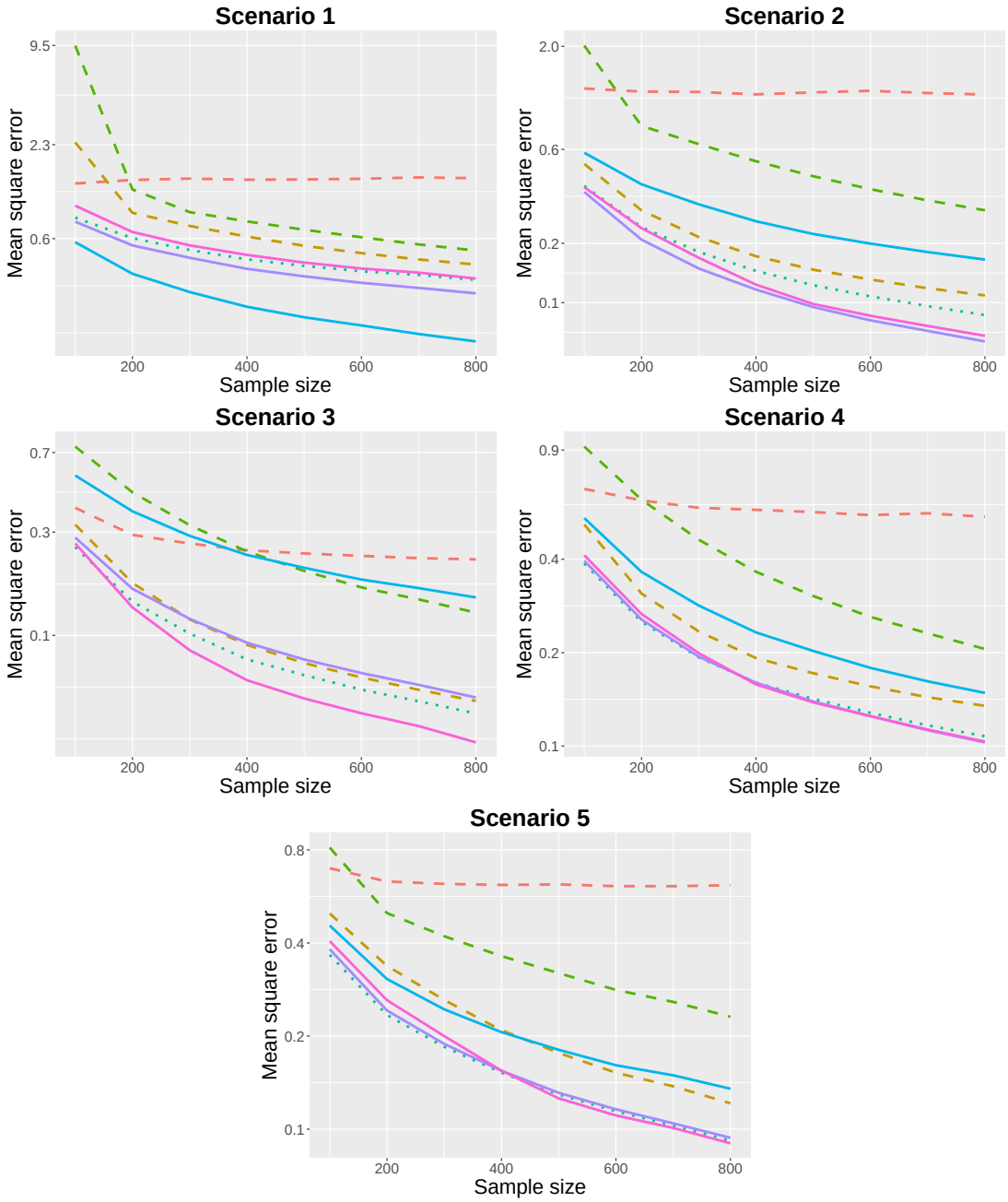


Figure 3.2: Plot of MSE as a function of sample size for each of the five scenarios for $p = 6$, averaged over 500 replications of the data. The dashed lines correspond to `SpAM` with small (- - -), moderate (- - -) and high (- - -) number of basis functions. The solid lines correspond to trend filtering of order $k = 0$ (——), 1 (——) and 2 (——). Finally, SSP is represented by the dotted line (·····).

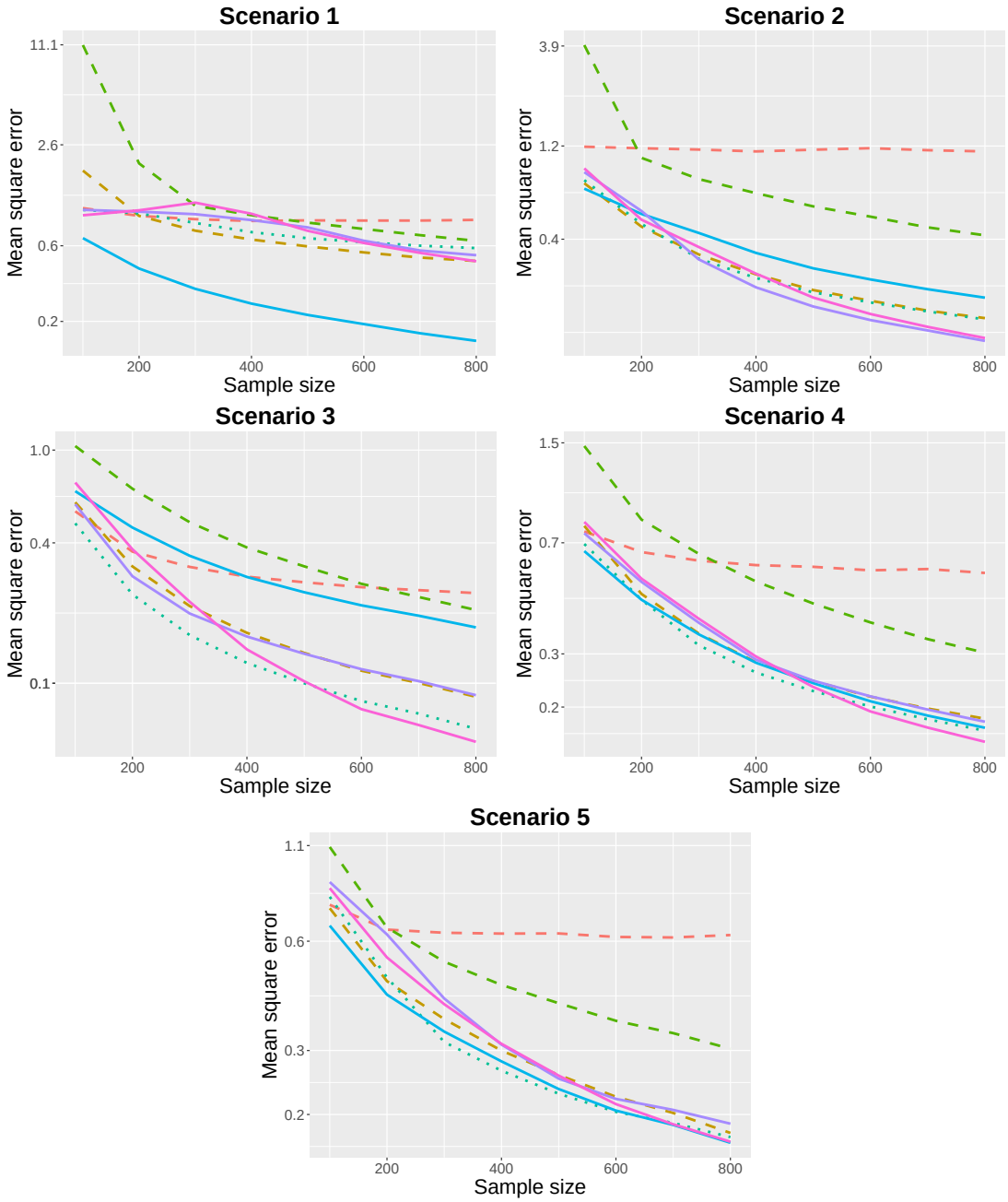
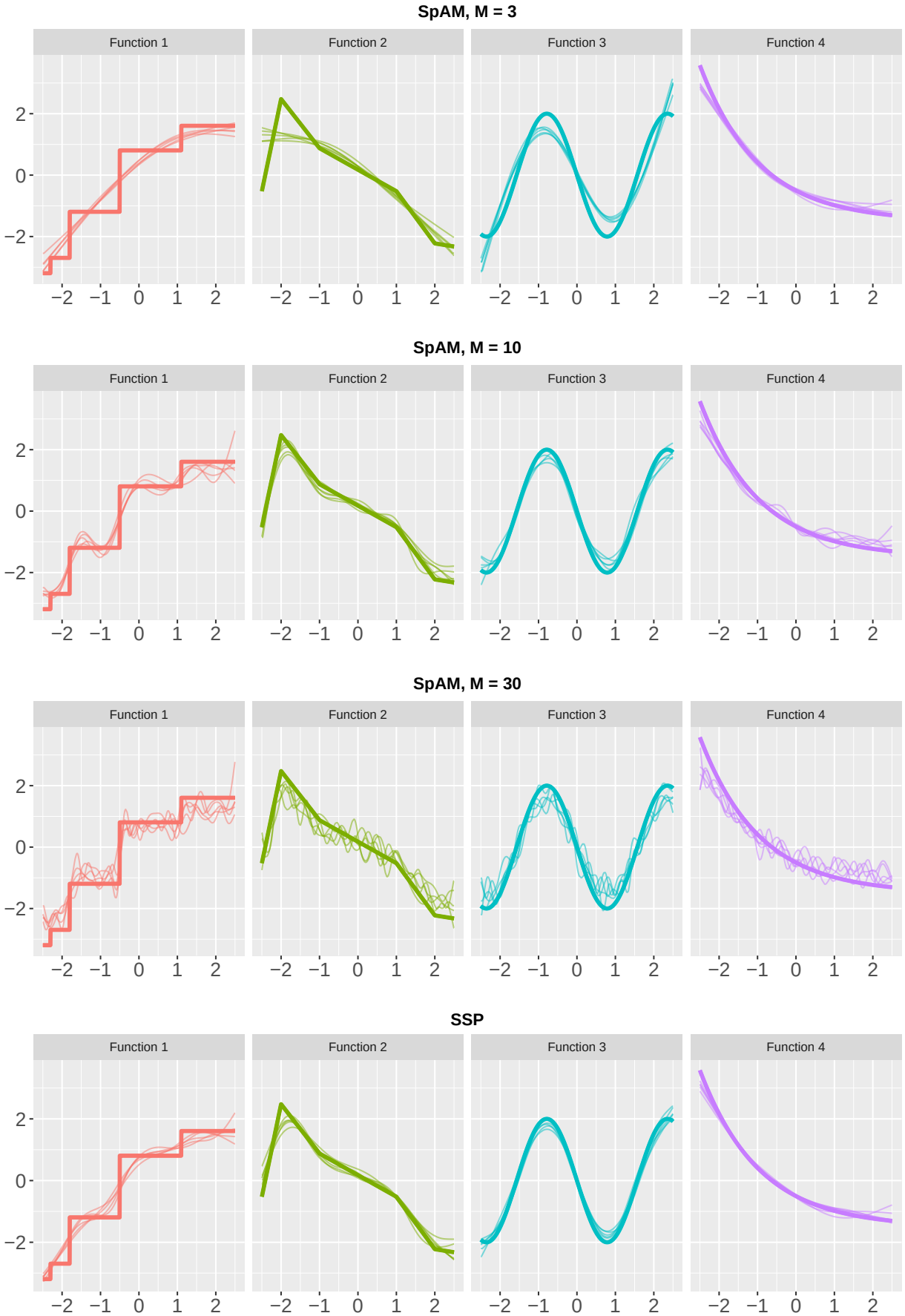


Figure 3.3: Plot of MSE as a function of sample size for each of the five scenarios for $p = 100$, averaged over 500 replications of the data. The line types and colors are the same as in Figure 3.2.


SpAM, M = 3
Function 1
Function 2
Function 3
Function 4
SpAM, M = 10
Function 1
Function 2
Function 3
Function 4
SpAM, M = 30
Function 1
Function 2
Function 3
Function 4
SSP
Function 1
Function 2
Function 3
Function 4

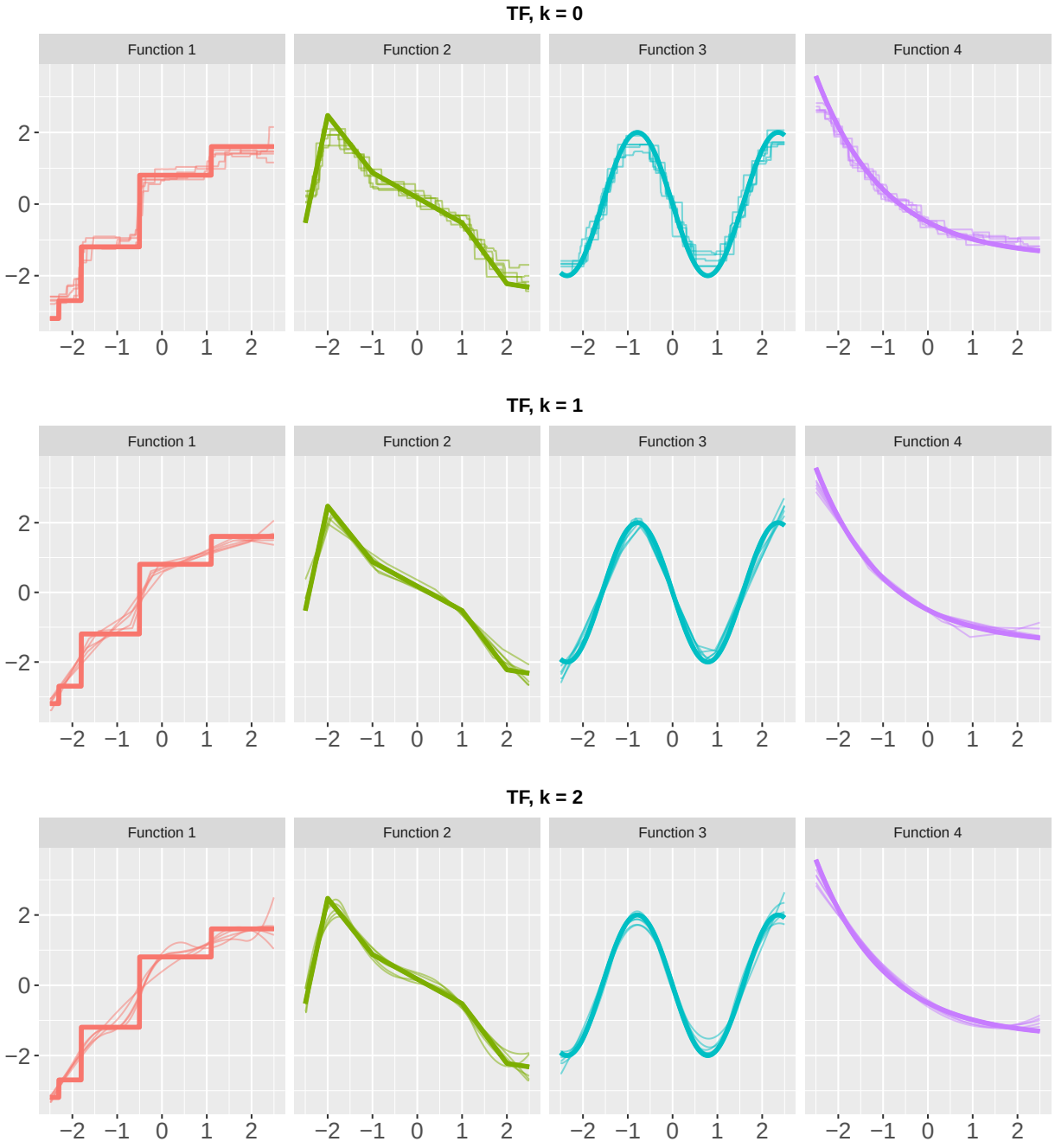


Figure 3.5: Examples of estimated signal functions by Trend Filtering (Sadhanala and Tibshirani, 2017) for Scenario 4.

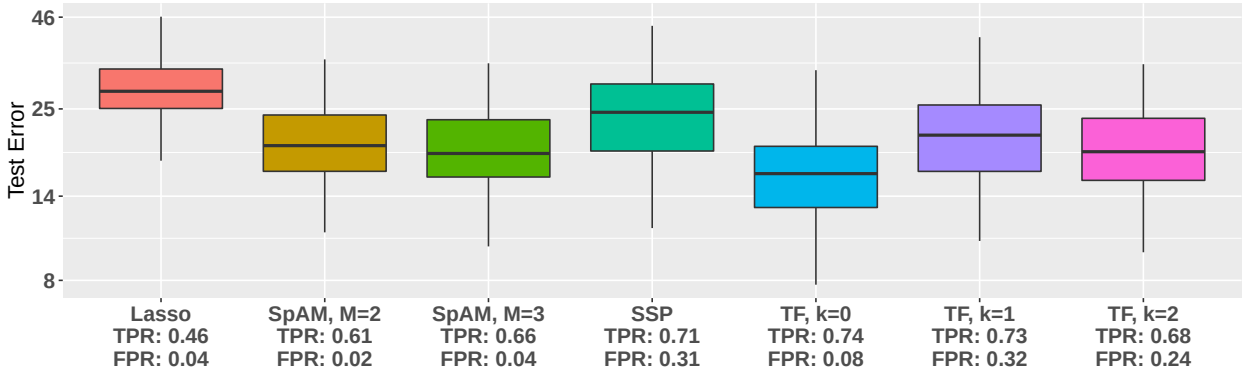


Figure 3.6: Box-plot of test errors for 100 different train/test splits of the data for each method. The average TPR and FPR was calculated using the original 10 covariates as 'signal' variables and remaining 20 as noise variables.

2009; Lin and Zhang, 2006). As done in the data analysis by Ravikumar et al. (2009), we add 10 noise covariates uniformly generated on the unit interval and 10 additional noise covariates obtained by randomly permuting the original covariates.

We fit SpAM with $M = 2$ and 3 basis functions, TF with orders $k = 0, 1, 2$, and SSP; we also fit the lasso Tibshirani (1996). Approximately 75% of the observations were used as training set and we reported the mean square prediction error on the test set. The final model was selected using 5-fold cross validation using the '1 standard error rule'. Results were presented for 100 splits of the data into training and test sets.

The box-plots of test error in the test set are shown in Figure 3.6. Since we added noise variables for the purpose of this analysis, we also state the average true positive rate (TPR) and false positive rate (FPR) in Figure 3.6. The box-plots demonstrate superior performance of trend filtering of order $k = 0$ over other methods in terms of lowest prediction error and highest TPR. The FPR of trend filtering with $k = 0$ was also low (under 10%). In Figure 3.7, we plot fitted functions for one split of the data for lasso, SpAM with $M = 3$, SSP and, TF with $k = 0$ for the 10 covariates of the original dataset. A striking feature of trend filtering fits is that many component functions are constant for extreme values of the covariates.

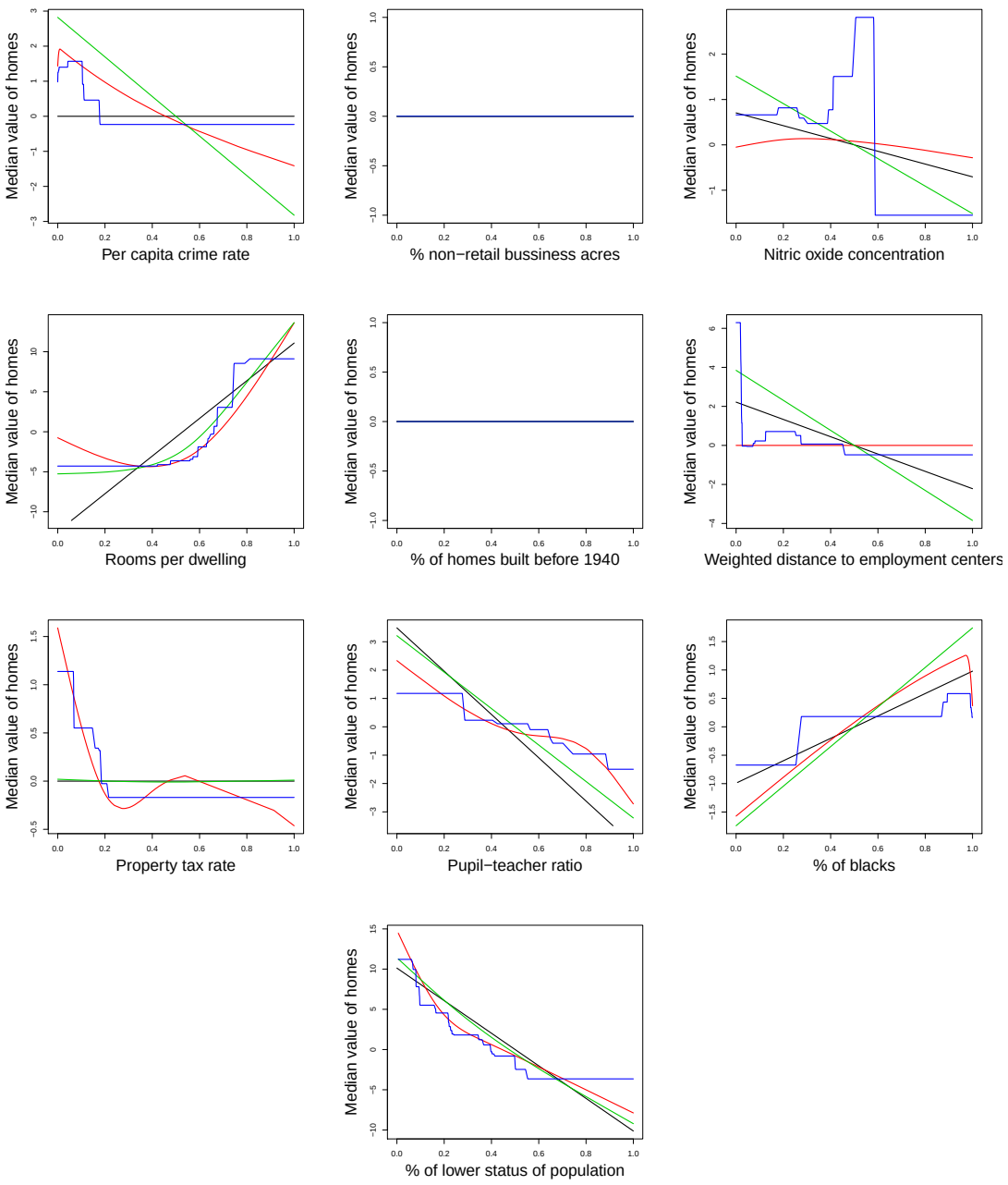


Figure 3.7: Plots of fitted functions for the original 10 covariates for a single split of the data into training and test sets for lasso (——), SpAM (——) with $M = 3$ basis functions, SSP (——) and, TF (——) of order $k = 0$.

3.7 Discussion

In this chapter, we introduced a general framework for non-parametric high-dimensional sparse additive models. We show that many existing proposals, such as SpAM (Ravikumar et al., 2009), SPLAM (Lou et al., 2016), Sobolev smoothness (Meier et al., 2009), and trend filtering additive models (Sadhanala and Tibshirani, 2017; Petersen et al., 2016), fall within our framework.

We established a proximal gradient descent algorithm which has a lasso-like per iteration complexity for certain choices of the structural penalty. Our theoretical analyses in Section 3.4 showed both fast rates, which match minimax rates under Gaussian noise, as well as slow rates, which only require a few weak assumptions.

The `R` package `PGSAME`, which will be made available on `github` soon, which implements the methods described in this chapter.

Chapter 4

NONPARAMETRIC REGRESSION WITH ADAPTIVE TRUNCATION VIA A CONVEX HIERARCHICAL PENALTY

4.1 Introduction

Consider first univariate nonparametric function estimation from observations $\{(x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, n\}$. Assume that $y_i = f(x_i) + \varepsilon_i$ $(i = 1, \ldots, n)$, where ε_i are independent, mean 0, sub-Gaussian random variables. There are many proposals for estimating f; local polynomials (Stone, 1977), kernel smoothing (Nadaraya, 1964; Watson, 1964), splines (Wahba, 1990), and others. To begin, we focus on basis expansions estimators, also known as projection estimators (Čencov, 1962), which are widely used, and arguably the simplest.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ and $\boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ be the response and covariate vectors. For $\boldsymbol{v} \in \mathbb{R}^n$, let $\|\boldsymbol{v}\|_n^2 = n^{-1} \sum_{i=1}^n v_i^2$ 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 $(\psi_k)_{k=1}^\infty$, along with a truncation level K. More specifically, let $\boldsymbol{\Psi}_K \in \mathbb{R}^{n \times K}$ be the $n \times K$ matrix with entries $\Psi_{K(i,k)} = \psi_k(x_i)$ $(k = 1, \ldots, K; i = 1, \ldots, n)$. The basis expansion estimate of f is then given by $\widehat{f} = \sum_{k=1}^K \widehat{\beta}_k^{proj} \psi_k$, where

$$\widehat{\boldsymbol{\beta}}^{proj} = \underset{\boldsymbol{\beta} \in \mathbb{R}^K}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\Psi}_K \beta\|_n^2 . \tag{4.1}$$

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 limitations of tuning K becomes clear in multivariate problems, which we describe next, and is one of our main motivations.

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

$$\widehat{\boldsymbol{\beta}}_1^{\text{A-proj}}, \ldots, \widehat{\boldsymbol{\beta}}_p^{\text{A-proj}} = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{K_j}}{\operatorname{argmin}} \frac{1}{2} \left\| \boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{\Psi}_{K_j}^j \boldsymbol{\beta}_j \right\|_n^2, \tag{4.2}$$

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

In practice, the same truncation level is used for each feature, $K_j \equiv K$, to reduce the number of tuning parameters. When f_j have widely different complexities, this strategy leads to poor estimates. This limitation, which is illustrated in the small simulation study summarized in Figure 4.1, becomes more hindering in higher dimensions, as p increases.

For high-dimensional problems, when $p \gg n$, it is often assumed that for many components $f_j \equiv 0$. A popular choice in this scenario is to add a sparsity-inducing penalty to the basis expansion framework (Ravikumar et al., 2009) and solve

$$\widehat{\boldsymbol{\beta}}_1^{\text{SPAM}}, \ldots, \widehat{\boldsymbol{\beta}}_p^{\text{SPAM}} = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{K}}{\operatorname{argmin}} \frac{1}{2} \left\| \boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{\Psi}_{K}^j \boldsymbol{\beta}_j \right\|_n^2 + \lambda \sum_{j=1}^{p} \left\| \boldsymbol{\Psi}_{K}^j \boldsymbol{\beta}_j \right\|_n. \tag{4.3}$$

In this chapter, we propose a penalized estimation framework motivated by the projection estimator: Our framework penalizes the function complexity and simultaneously selects the truncation level. This framework can be used to fit both univariate and multivariate additive models with or without sparsity. We also discuss an extension of our proposal for fully nonparametric multivariate settings. Finally, a relaxed version of our framework, similar to the relaxed lasso (Meinshausen, 2007), is presented which can further improve representational parsimony. For univariate problems, our method performs similarly to (4.1). However, for additive or sparse additive models, it automatically chooses a truncation level for each feature. These truncation levels will often differ between features based on the underlying complexity of the true f_j. This adaptability can vastly improve the prediction

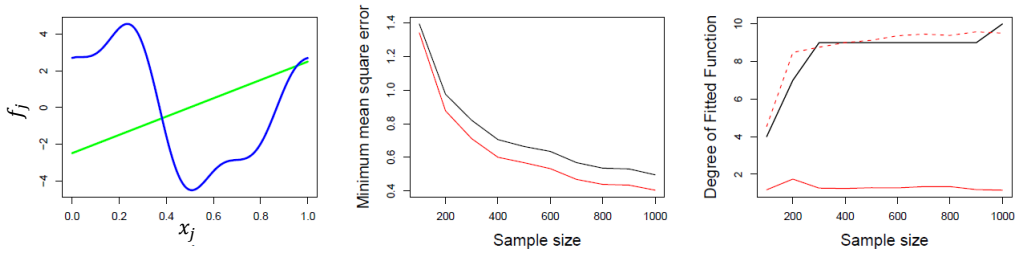


Figure 4.1: Left: Plots of component functions f_1 (——) and f_2 (——). Middle: Minimum mean square error as a function of n for our proposal (——) and (4.2) (——). Right: Degree of fitted polynomial as a function of n: For our proposal, $\widehat{f}_1$ (——) and $\widehat{f}_2$ (------) are shown; for (4.2) both component functions have the same degree (——). All results are averaged over 100 replications.

accuracy of our model; it additionally allows us to maintain as much parsimony as possible in estimating each f_j. The key innovation of our approach can thus be seen as obtaining a smooth and parsimoniously represented estimate of f_j due to our data-adaptive selection of the truncation level. We illustrate these advantages in a small simulation study using data $y_i = f_1(x_{i1}) + f_2(x_{i2}) + \varepsilon_i$ $(i = 1, \ldots, n)$ generated from f_1, f_2 shown in Figure 4.1. We fit (4.2) using $K_j \equiv K$, with K selected to optimize mean square error, and compare it to our relaxed proposal. Selected tuning parameters for each method, are those which minimize the mean square error. The results in Figure 4.1 clearly demonstrate the superior performance of our method, which exhibits a lower mean square error while maintaining parsimony. In particular, our proposal estimates the linear term, f_1, by a linear function, whereas (4.2) uses an order 9 polynomial for $\widehat{f}_1$.

In addition to adaptability and parsimony, our proposal also offers computational efficiency and theoretical guarantees. It can be applied to problems 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 the order of $O\left\{n^{-2m/(2m+1)}\right\}$ where m is the degree of smoothness; similarly, the multivariate estimator attains the rate $O\left\{n^{-2m/(2m+p)}\right\}$. For sparse additive models, under a suitable compatibility condition, our estimator converges at

$O\left[\max\left\{sn^{-2m/(2m+1)}, s\log p/n\right\}\right]$, where s is the number of non-zero f_j; even without the compatibility condition, it is consistent with convergence rate $O\left[\max\left\{sn^{-m/(2m+1)}, s(\log p/n)^{1/2}\right\}\right]$.

4.2 Methodology

4.2.1 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 tradeoff and parsimony of estimates $\widehat{f}_j$ in (4.2) are controlled by truncation levels K_j. While separately tuning K_j over each component function may be feasible in low dimensions, this becomes quickly 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, one can instead use n-basis functions for each component, and consider a penalized version of a truncation estimator,

$$\widehat{\boldsymbol{\beta}}_1^{\text{proj-pen}}, \ldots, \widehat{\boldsymbol{\beta}}_p^{\text{proj-pen}} = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\Big\|\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{\Psi}_n^j \boldsymbol{\beta}_j\Big\|_n^2 + \lambda \sum_{j=1}^{p} \max\{k \mid \beta_{jk} \neq 0\}, \tag{4.4}$$

where the truncation level for each feature is determined using the penalty $\max\{k \mid \beta_{jk} \neq 0\}$. The estimator in (4.4) chooses the truncation level for each feature data-adaptively. However, the penalty in (4.4) is non-convex. Therefore, solving (4.4) in moderate to high-dimensional problems becomes infeasible. We modify (4.4) to achieve this goal. Specifically, we formulate a convex problem using a novel penalty that can be seen as a convex relaxation of the penalty in (4.4).

Our approach is particularly suitable for basis functions which 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; we plot these examples in Figure 4.2.

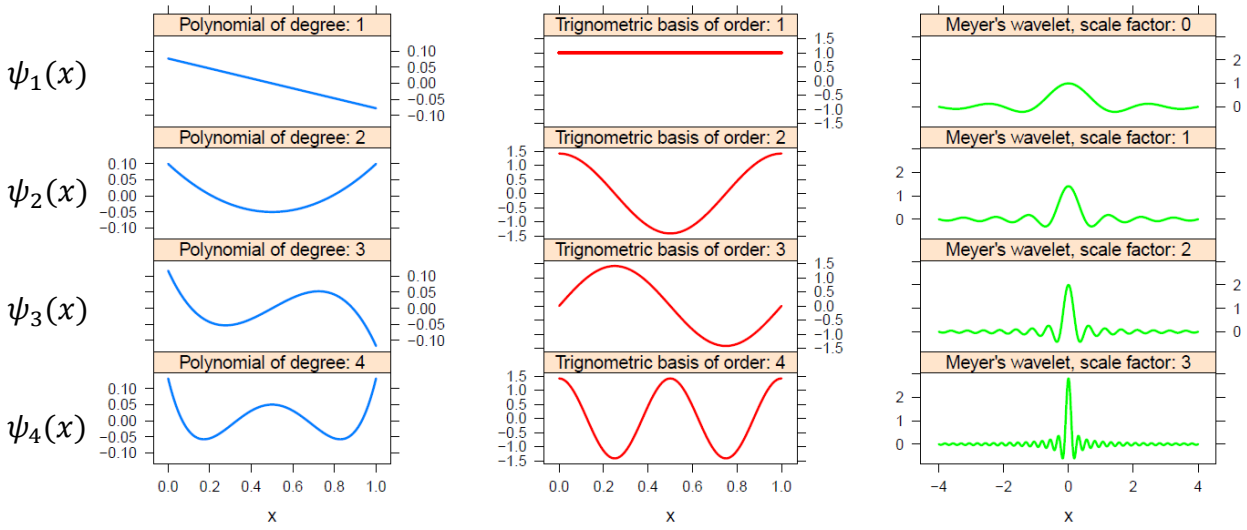


Figure 4.2: Examples of basis functions with natural hierarchical complexity; polynomial, trigonometric and wavelet basis functions are shown in the left, center and right panels, respectively.

4.2.2 The univariate proposal

Consider again the projection estimator (4.1). As noted in Section 4.1, choosing the truncation level K is key here: K too small will result in a large bias, while K too large will over-inflate the variance. The bias and variance are balanced by taking $K = O\left\{n^{1/(2m+1)}\right\}$, where m relates to the smoothness of the underlying f, and is unknown in practice. To circumvent this challenge, we use instead a complete basis with $K = n$ along with regularization to data-adaptively choose the truncation level. More specifically, our estimator is defined as

$$\widehat{\boldsymbol{\beta}}^{hier} = \underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{\Psi}_n\boldsymbol{\beta}\right\|_n^2 + \lambda\Omega\left(\boldsymbol{\beta}\right), \quad \text{where} \quad \Omega\left(\boldsymbol{\beta}\right) = \sum_{k=1}^{n} w_k \left\|\boldsymbol{\Psi}_{k:n}\boldsymbol{\beta}_{k:n}\right\|_n, \quad (4.5)$$

with $w_k = k^m - (k-1)^m$. Here, $\boldsymbol{\Psi}_{k:n}$ denotes the submatrix of $\boldsymbol{\Psi}_n$ containing columns k to n, $\boldsymbol{\beta}_{k:n}$ is the subvector of $\boldsymbol{\beta}$ containing entries k to n, and m and λ are tuning parameters. The choice of weights w_k is theoretically motivated, and detailed in Section 5. Briefly, it

defines a function class with desirable properties allowing us to establish convergence rates.

The hierarchical group lasso penalty $\Omega(\boldsymbol{\beta})$, will result in a solution $\widehat{\boldsymbol{\beta}}^{hier}$ with hierarchical sparsity: that is, if $\widehat{\beta}_k^{hier} = 0$ for some k, then $\widehat{\beta}_{k'}^{hier} = 0$ for all $k' > k$. For sufficiently large λ, many entries of $\widehat{\boldsymbol{\beta}}^{hier}$ will be 0. For a given λ, we define the induced truncation level to be the minimal integer $K \leq n$ such that $\widehat{\beta}_k^{hier} = 0$ for all integers $k > K$. Unlike the simple basis expansion estimator (4.1), this truncation level is data-adaptive, not prespecified.

Equation (4.5) 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 a 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/smoothness; 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 (4.5) 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 (4.5) as the simple basis expansion estimator with $K = \|\widehat{\boldsymbol{\beta}}^{hier}\|_0$, the truncation level selected by (4.5). This relaxed proposal is equivalent to using the penalty for selecting a truncation level. Therefore, in the univariate case, the relaxed estimator for a sequence of λ values would match the simple basis expansion estimator (4.1) for a sequence of K values. However, the advantages of our penalty become more clear in the case of multivariate additive models described next.

4.2.3 The additive proposal

Ideally, the additive projection estimator (4.2) 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 $K_j, (j = 1, \ldots, p)$; however, the number of candidate models grows exponentially in p and becomes quickly unwieldy. Often, a single $K \equiv K_j$ is used in practice, which can lead to some f_j estimates with too many degrees

of freedom. As illustrated in Figure 4.1, using a single truncation level can lead to poor estimates.

Our proposal, which can be seen as a convex relaxation of (4.4), is designed to circumvent the above limitation of projection estimators in choosing the truncation level for models with multiple covariates. This is what 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 (4.5). Specifically, we consider function estimates $\widehat{f}_j = \sum_{k=1}^{n} \widehat{\beta}_{jk}^{\text{A-hier}} \psi_k$, where

$$\widehat{\boldsymbol{\beta}}_1^{\text{A-hier}}, \ldots, \widehat{\boldsymbol{\beta}}_p^{\text{A-hier}} = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2} \left\| \boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{\Psi}_n^j \boldsymbol{\beta}_j \right\|_n^2 + \lambda \sum_{j=1}^{p} \Omega_j \left(\boldsymbol{\beta}_j \right), \tag{4.6}$$

and Ω_j is the hierarchical group lasso penalty with weights $w_k = k^m - (k-1)^m$:

$$\Omega_j \left(\boldsymbol{\beta}_j \right) = \sum_{k=1}^{n} w_k \left\| \boldsymbol{\Psi}_{k:n}^j \boldsymbol{\beta}_{j,k:n} \right\|_n. \tag{4.7}$$

The optimization problem (4.6) results in $\widehat{\boldsymbol{\beta}}_j$ estimates that are hierarchically sparse for each j. Specifically, for each j, there is some minimal K_j such that $\widehat{\beta}_{jk}^{A-hier} = 0$ for all integers $k > K_j$. Moreover, the major advantage of (4.6) is that the induced truncation level is feature-wise adaptive, with a different K_j for each feature j. Additionally, like the univariate setting, we can define a relaxed version of our estimator by fitting (4.2), where K_j is now determined by (4.6). 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 where K_j need to be searched exhaustively or set to the same level $K_j = K$.

The advantage of our method over simple projection estimators becomes even more significant in high dimensions, when $p \gg n$. For instance, the popular estimator of Ravikumar et al. (4.3), is generally obtained by using a single truncation level, which, as noted above, can result in poor estimators. Similar to their proposal, our sparse additive framework

encourages feature-wise sparsity using a group lasso penalty (Yuan and Lin, 2006), and is defined as

$$\widehat{\boldsymbol{\beta}}_1^{\text{S-hier}}, \ldots, \widehat{\boldsymbol{\beta}}_p^{\text{S-hier}} = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2} \left\| \boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{\Psi}_n^j \boldsymbol{\beta}_j \right\|_2 + \lambda^2 \sum_{j=1}^{p} \Omega_j \left(\boldsymbol{\beta}_j \right) + \lambda \sum_{j=1}^{p} \left\| \boldsymbol{\Psi}_n^j \boldsymbol{\beta}_j \right\|_n, \quad (4.8)$$

with $\Omega_j \left(\boldsymbol{\beta}_j \right)$ is defined in (4.7). We can again define a relaxed version which fits (4.2) with sparsity and K_j selected by (4.8). An important feature of the optimization problem (4.8) is that the tuning parameters for the two penalty terms λ and λ^2, are linked. This link is theoretically justified in Section 4.4. Briefly, for an oracle λ, the choice of tuning parameters in (4.8) gives rate-optimal estimates. In practice, while this formulation gives good predictive performance in many cases, in other cases tuning sparsity and smoothness separately leads to strong predictive performance. Our numerical experiments in Section 4.5 and 4.6 corroborate this finding.

As with $\widehat{\boldsymbol{\beta}}_j^{SPAM}$ in (4.3), for sufficiently large λ, our proposal gives a sparse solution with most $\widehat{\boldsymbol{\beta}}_j^{\text{S-hier}} \equiv 0$. The two estimators differ, however, in their nonzero estimates: non-zero $\widehat{\boldsymbol{\beta}}_j^{\text{S-hier}}$ are hierarchically sparse, with a data-driven feature-specific induced truncation level, whereas nonzero $\widehat{\boldsymbol{\beta}}_j^{SPAM}$ in (4.3) all have the same complexity. This additional flexibility of our methodology proves critical in high dimensions, and is achieved without paying a price in computational or sample complexity. Moreover, with the tuning parameters in (4.8), this additional flexibly is in theory achieved with the same number of tuning parameters as Ravikumar et al.'s method.

4.2.4 Relationship to existing methods

The univariate framework of Section 4.2.2 builds upon existing penalized methods for estimating regression functions. A popular penalized estimation method is the smoothing spline estimator (Wahba, 1990), which sets $(\psi_k)_{k=1}^n$ to n natural splines with knots at the observed

covariates $x_1, \ldots, x_n$; this estimator is found by minimizing

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Psi}\boldsymbol{\beta}\|_n^2 + \lambda\|\boldsymbol{C}^{1/2}\boldsymbol{\beta}\|_n^2$$

over $\beta \in \mathbb{R}^n$, using $\boldsymbol{C} \in \mathbb{R}^{n\times n}$ and $C_{j,k} = \int \psi_j^{(m/2+1/2)}(t)\psi_k^{(m/2+1/2)}(t)\,dt$ with $\psi^{(k)}$ denoting the k-th oder derivative of ψ. The smoothing spline eliminates the dependence on the truncation level and has an efficient-to-compute closed form solution; however, 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 and van de Geer (1997) use a data-driven approach to select the knots in spline functions. Their locally adaptive regression splines use the same natural spline basis and is found by minimizing

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Psi}\boldsymbol{\beta}\|_n^2 + \lambda(m!)^{-1}\|\boldsymbol{D}\boldsymbol{\beta}\|_1,$$

over $\boldsymbol{\beta} \in \mathbb{R}^n$, using $\boldsymbol{D} \in \mathbb{R}^{(n-m-1)\times n}$ with $D_{i,j} = \psi_j^{(m)}(t_i) - \psi_j^{(m)}(t_{i-1})$. The proposal of Mammen and van de Geer (1997) is closely related to the recent, more computationally tractable, trend filtering proposal (Kim et al., 2009; 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. Ravikumar et al.'s estimator (4.3) overcomes this difficulty by using a fixed truncation level for all p components. As mentioned earlier, the main drawback of (4.3) is that all nonzero components of the additive model have the same level of complexity. The recently proposed sparse partially linear additive model of Lou et al. (2016) partly mitigates this shortcoming by setting some of the nonzero components to linear functions using a hierarchical penalty of the form $\sum_{j=1}^p \lambda_1\|\boldsymbol{\beta}_j\|_2 + \lambda_2\|\boldsymbol{\beta}_{j,-1}\|_2$; here $\beta_{j,1}$ is the coefficient of the linear term in the basis expansion and $\boldsymbol{\beta}_{j,-1} = [\beta_{j,2}, \ldots, \beta_{j,K}]^\top \in \mathbb{R}^{K-1}$. Depending on the value of tuning parameters λ_1 and λ_2, the first term in the penalty sets all of the coefficients for the j-th feature to zero, whereas the second term only sets the $K-1$ coefficients corresponding to

Table 4.1: Comparison of existing methods for sparse additive models.

	PE	SS/RKHS	TF
Scalability	✓	×	×
	Exact solution for univariate sub-problem; scales as $O(npK)$.	No exact solution for univariate sub-problem; general convex solver scaling 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 order $K \ll n$ expansions.	Component functions of order n expansions.	Component functions have sparsity in number of knots.

PE, projection estimators (Ravikumar et al., 2009; Lou et al., 2016); RKHS, reproducing kernel Hilbert spaces (Raskutti et al., 2012; Koltchinskii and Yuan, 2010; Yuan and Zhou, 2015); SS, smoothing splines (Meier et al., 2009); TF, trend filtering (Petersen et al., 2016; Sadhanala and Tibshirani, 2017).

higher-oder terms to zero.

Our additive and sparse additive proposals of Section 4.2.3 can be seen as generalizations of the proposals by Ravikumar et al. (2009) and Lou et al. (2016). Specifically, Ravikumar et al.'s proposal becomes a special case of (4.8) if the weights in (4.7) are set to $w_1 = 1$ and $w_k = 0$ for $k > 1$. Similarly, with an orthogonal design matrix, $(\mathbf{\Psi}_K^j)^\top \mathbf{\Psi}_K^j / n = I_K$ $(j = 1, \ldots, p)$, Lou et al.'s method is a special case of (4.8) with weights in (4.7) set to $w_1 = w_2 = 1$ and $w_k = 0$ for $k > 2$. Our theoretical analysis in Section 4.4.5 indicates that, in addition to the improved flexibility, our choice of weights (4.7) results in optimal rates of convergence.

There are a number of 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 for low-dimensional additive models (Sadhanala and Tibshirani, 2017). 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.

There have also been some proposals which offer minimax-optimal convergence rates for the prediction error over smooth additive classes (Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Yuan and Zhou, 2015) similar to our results in Section 4.4.5. These proposals use properties of reproducing kernel Hilbert spaces, and their estimator is given as the minimizer of a (np)-dimensional second order cone program. However, they do not discuss efficient algorithms for solving the optimization problems, 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 become quickly intractable. We compare and contrast the strengths and weaknesses of existing proposals in Table 4.1.

4.3 Computational considerations and extensions

4.3.1 Conservative basis truncation

Our proposal (4.5) uses a basis expansion with n basis functions. In practice, for any reasonable choice of λ, $\widehat{\beta}$ will never have n nonzero entries, and will generally have very few non-zero entries, $K_0 \ll n$. If we instead solve

$$\widehat{\boldsymbol{\beta}}^{hier(\tilde{K})} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{\tilde{K}}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{\Psi}_{\tilde{K}}\boldsymbol{\beta}\right\|_2 + \lambda \sum_{k=1}^{\tilde{K}} w_k \left\|\boldsymbol{\Psi}_{k:\tilde{K}}\boldsymbol{\beta}_{k:\tilde{K}}\right\|_n, \tag{4.9}$$

for $\tilde{K} < n$, then so long as $\tilde{K} \geq K_0$, the solution will be identical to that of the original proposal (4.5). Even when not identical, so long as $\tilde{K}$ is sufficiently large, $\tilde{K} \gtrsim n^{2m/\{(2m-1)(2m+1)\}}$, where $a_n \gtrsim b_n$ means $a_n \geq Cb_n$ for some constant C, the theoretical properties of (4.5) 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 Section 4.3.2, 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

$$\widehat{\boldsymbol{\beta}}_1^{S-hier(\tilde{K})}, \ldots, \widehat{\boldsymbol{\beta}}_p^{S-hier(\tilde{K})} = \underset{\boldsymbol{\beta}_j\in\mathbb{R}^{\tilde{K}}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{\Psi}_{\tilde{K}}^j \boldsymbol{\beta}_j\right\|_2 + \lambda^2 \sum_{j=1}^{p} \Omega_j\left(\boldsymbol{\beta}_j\right) + \lambda \sum_{j=1}^{p} \left\|\boldsymbol{\Psi}_{\tilde{K}}^j \boldsymbol{\beta}_j\right\|_n, \tag{4.10}$$

where, now, $\Omega_j\left(\boldsymbol{\beta}_j\right) = \sum_{k=1}^{\tilde{K}} w_k \|\boldsymbol{\Psi}_{k:\tilde{K}}^j \boldsymbol{\beta}_{j,k:\tilde{K}}\|_n$. It is worth noting that choosing the pre-truncation level, $\tilde{K}$, is easier than 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; the former only requires a level that is not too small.

4.3.2 Algorithm for the univariate and sparse additive framework

An appealing feature of our estimator is its computational efficiency. Problem (4.9) can be solved via a one-step coordinate descent algorithm. Using a QR decomposition $\boldsymbol{\Psi} = \boldsymbol{U}\boldsymbol{V}$ with $\boldsymbol{U} \in \mathbb{R}^{n \times \tilde{K}}$ and $\boldsymbol{U}^\top \boldsymbol{U}/n = I_{\tilde{K}}$, we can re-writing (4.9) as

$$\underset{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{\tilde{K}}}{\text{minimize}} \; \frac{1}{2n} \left\| \boldsymbol{y} - \boldsymbol{U}\widetilde{\boldsymbol{\beta}} \right\|_2^2 + \lambda \sum_{k=1}^{\tilde{K}} w_k \left\| \widetilde{\boldsymbol{\beta}}_{k:\tilde{K}} \right\|_2, \tag{4.11}$$

where $\widetilde{\boldsymbol{\beta}} = \boldsymbol{V}\boldsymbol{\beta}$. Applying the results of Jenatton et al. (2010), gives us Algorithm 3 below.

Algorithm 3 One-Step coordinate descent for univariate setting

Initialize $\boldsymbol{\beta}^0 = \cdots = \boldsymbol{\beta}^{\tilde{K}} \leftarrow \boldsymbol{U}^\top \boldsymbol{y}/n$
For $k = \tilde{K}, \ldots, 1$
  Update $\boldsymbol{\beta}^{k-1}_{k:\tilde{K}} \leftarrow \left(1 - w_k \lambda / \|\boldsymbol{\beta}^k_{k:\tilde{K}}\|_2\right)_+ \boldsymbol{\beta}^k_{k:\tilde{K}}$, where $(x)_+ = \max(x, 0)$
Return $\boldsymbol{\beta}^0$

The reformulation in (4.11) can also be used to efficiently solve the sparse additive extension (4.10) via a block coordinate descent algorithm. Specifically, given a set of estimates $(\boldsymbol{\beta}_j)_{j=1}^p$, we can fix all but one of the vectors $\boldsymbol{\beta}_j$ and optimize over the non-fixed vector using Algorithm 3. Iterating until convergence yields the solution to problem (4.10), as described in Algorithm 4, Appendix C.1.

Solving problem (4.5) requires a QR decomposition of the matrix $\boldsymbol{\Psi}$ followed by the multiplication $\boldsymbol{U}^\top \boldsymbol{y}$; these steps require $O(n\tilde{K}^2)$ and $O(n\tilde{K})$ operations, respectively. However, these steps are only needed once for a sequence of λ values. For the additive proposal (4.9), p such QR decompositions are needed once for the entire sequence of λs.

By Proposition 2 of Jenatton et al. (2010), for a given λ, problem (4.11) can be solved in $O(\tilde{K})$ operations. Each block update requires a matrix multiplication $\boldsymbol{U}_j^\top \boldsymbol{r}_{-j}$ followed by solving the proximal problem (4.11), see Appendix C.1. This requires $O(n\tilde{K})$ operations. Thus, our sparse additive proposal requires $O(np\tilde{K})$ operations, which is equal to 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, the median time of 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 of solving the sparse additive framework for the simulation setting of Section 4.5.2 on a grid of 50 λ values is 5·96 seconds.

4.3.3 Degrees of freedom

For regression with fixed design and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, we consider the definition of degrees of freedom given by Stein (1981), $\mathrm{df} = \sum_{i=1}^{n} \mathrm{cov}(y_i, \widehat{y}_i)/\sigma^2$, where $\widehat{y}_i$ are the fitted response values. We apply Claim 2.1 of Chapter 2 to derive an unbiased estimate of df for the estimator (4.11), using the decomposition $\boldsymbol{\Psi} = \boldsymbol{U}\boldsymbol{V}$ from Section 4.3.2. Let $K_0 = \max\{k : \widehat{\beta}_k \neq 0\}$, and let $\boldsymbol{U}_{K_0} \in \mathbb{R}^{n \times K_0}$ denote the first K_0 columns of $\boldsymbol{U}$. Furthermore, for a vector $\boldsymbol{\nu} \in \mathbb{R}^n$, define $\boldsymbol{\nu}_{k:K_0} \in \mathbb{R}^{K_0}$ as $\boldsymbol{\nu}_{k:K_0} = [0, 0, \ldots, 0, \nu_k, \nu_{k+1}, \ldots, \nu_{K_0}]^\top$. We arrive at the following lemma.

Lemma 4.1. *An unbiased estimator for the degrees of freedom of $\widehat{\boldsymbol{\beta}}$ in (4.5) is given by*

$$\widehat{df} = 1 + \mathrm{tr}\left\{ \boldsymbol{U}_{K_0} \left[\boldsymbol{I}_{K_0} + \sum_{k=1}^{K_0} \lambda w_k \left\{ \frac{\mathrm{diag}(\mathbf{1}_{k:K_0})}{\|\widehat{\boldsymbol{\beta}}_{k:K_0}\|_2} - \frac{\widehat{\boldsymbol{\beta}}_{k:K_0} \widehat{\boldsymbol{\beta}}_{k:K_0}^\top}{\|\widehat{\boldsymbol{\beta}}_{k:K_0}\|_2^3} \right\} \right]^{-1} \frac{\boldsymbol{U}_{K_0}^\top}{n} \left(\boldsymbol{I}_n - \mathbf{1}_n \mathbf{1}_n^\top / n \right) \right\},$$

where $\mathrm{diag}(\boldsymbol{\nu}) \in \mathbb{R}^{n \times n}$ *is a diagonal matrix with* $\boldsymbol{\nu} \in \mathbb{R}^n$ *on the main diagonal.*

4.3.4 Non-additive multivariate regression

For vectors $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{\nu}_k \in \mathbb{Z}_+^p$, define $\boldsymbol{x}^{\boldsymbol{\nu}_k} = x_1^{\nu_{11}} \times \cdots \times x_p^{\nu_{kp}}$. Now for functions $f^0 : \mathbb{R}^p \to \mathbb{R}$, consider the basis representation $f^0(\boldsymbol{x}) = \sum_{k=1}^{\tilde{K}} \psi_k(\boldsymbol{x}^{\boldsymbol{\nu}_k}) \beta_k^0$, for univariate

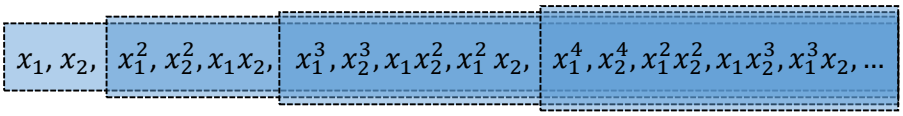


Figure 4.3: Visual representation of the multivariate penalty with $p = 2$ and $\psi_j(x) \equiv x$.

functions $(\psi_j)_{j=1}^{\infty}$, and $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{\tilde{K}} \in \mathbb{Z}_+^p$, where

$$\|\boldsymbol{\nu}_k\|_1 = 1 \;\; (k = 1, \ldots, p), \quad \|\boldsymbol{\nu}_k\|_1 = 2 \;\; \left(k = p+1, \ldots, \binom{p+2}{p} - 1\right),$$

and so on. As in the univariate case, let $\boldsymbol{\Psi}_{\tilde{K}} \in \mathbb{R}^{n \times \tilde{K}}$ be the matrix with entries $\Psi_{\tilde{K}(i,k)} = \psi_k(\boldsymbol{x_i}^{\boldsymbol{\nu}_k})$. Then, our multivariate regression estimator is simply (4.9) with weights given by

$$w_{q_k} = k^m - (k-1)^m, \quad q_k = \binom{k+p-1}{p}, \tag{4.12}$$

and $w_k = 0$ for all other k. Figure 4.3 demonstrates the multivariate penalty for $p = 2$ and ψ_k the identity function; that is, for $z \in \mathbb{R}$, $\psi_k(z) \equiv z$. 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 3 with a single pass over the basis elements.

4.3.5 Extension to classification

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

$$(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\beta}_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^n}{\arg\min} \; \frac{1}{2n} \sum_{i=1}^{n} \log\left(1 + \exp\left[-y_i\{\beta_0 + (\boldsymbol{\Psi}_n \boldsymbol{\beta})_i\}\right]\right) + \lambda \Omega(\boldsymbol{\beta}). \tag{4.13}$$

As with linear regression, (4.13) can be naturally extended to sparse additive models, by using both penalties in (4.8). The problem can be efficiently solved via a proximal gradient descent algorithm (Combettes and Pesquet, 2011); see Appendix C.1.

4.4 Theoretical results

4.4.1 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, and the metric entropy of our hierarchical class.

Similar to our theoretical results of Chapter 3, the results in Section 4.4.5 allow one to establish convergence rates for a broad class of penalized sparse additive model estimators. The results presented here utilize a different proof technique and are specialized to the least squares loss. Results specific to the least squares loss allow us to relax two conditions from the results of Chapter 3, namely, we do not assume a bounded intercept for proving slow rates and we do not require a margin condition for proving fast rates. Under a compatibility condition on the component features, these rates match the minimax lower bound for estimation of sparse additive models under independent component functions, established previously by Raskutti et al. (2009), see Proposition 4.3. Thus, our additive and sparse additive estimators are rate-optimal.

Finally, key for our theoretical analyses is the entropy of our hierarchical class; we calculate these with matching upper and lower bounds in Lemmas 4.3 and 4.4. These new results allows us to show that our univariate and sparse additive estimators, (4.5) and (4.8), are minimax rate-optimal within the hierarchical univariate and hierarchical sparse additive classes, respectively.

4.4.2 Entropy-based rates

We begin by stating two well-known results from the literature. We then present our contributions in Sections 4.4.4 and 4.4.5. Firstly, Theorem 1 of Yang and 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 Appendix C.7.

We first introduce some terminology and notation for the entropy of a set. For a set $\mathcal{F}$ equipped with some metric $d(\cdot, \cdot)$, the subset $\{f_1, \ldots, f_N\} \subset \mathcal{F}$ 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 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 $x_1, \ldots, x_n$, we denote by Q_n the empirical measure $Q_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$ and use the short-hand notation $\|\cdot\|_n = \|\cdot\|_{Q_n}$.

Theorem 4.1 (Theorem 1, Yang and Barron (1999))**.** *Consider the model $y_i = f^0(x_i) + \varepsilon_i$ $(i = 1, \ldots, n)$, with independent and identically distributed $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $x_i \sim Q$. Assume 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 depending on A_0, α and σ^2,*

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

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

Theorem 4.2 (Theorem 10·2, van de Geer (2000))**.** *Consider the model $y_i = f^0(x_i) + \varepsilon_i$ $(i = 1, \ldots, n)$, with independent sub-Gaussian noise ε_i. Let*

$$\widehat{f} = \arg\min_{f \in \mathcal{F}_n} \frac{1}{2} \|\boldsymbol{y} - f\|_n^2 + \lambda_n^2 \Omega(f | Q_n),$$

for some function class $\mathcal{F}_n$ and semi-norm $\Omega(\cdot | Q_n)$ 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 any function $f_n^* \in \mathcal{F}_n$, and

$$\lambda_n^{-1} = n^{1/(2+\alpha)} \left\{\Omega(f_n^*|Q_n)\right\}^{(2-\alpha)/\{2(2+\alpha)\}},$$

there is a constant c such that for all $T \geq c$, with probability at least $1 - c\exp\{-(T/c)^2\}$,

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

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

Before specializing Theorems 4.1 and 4.2 to our proposal, we briefly discuss their assumptions. The main assumption for Theorem 4.1 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 commonly used function classes. Examples include, bounded Lipscitz functions with $\alpha = 1$, bounded monotone functions with $\alpha = 1$, and m-th order Sobolev functions with $\alpha = 1/m$. Theorem 4.2 requires a similar entropy condition on a sequence of function spaces, but it does relax conditions on f_0 allowing it to be arbitrary, 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, $\mathcal{F}_n \subseteq \mathcal{F}$ 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 to use standard results from the empirical processes literature. While the original theorem of Yang and 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 section, we define and establish entropy bounds for our hierarchical function class.

4.4.3 Entropy results for the proposed penalty

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

$$\mathcal{F}_n = \left\{ f_{\boldsymbol{\beta}}(x) = \sum_{k=1}^{\tilde{K}_n} \psi_k(x)\beta_k \,:\, \int \psi_k \psi_l \, dQ = 0 \text{ for } k \neq l, \int \psi_k^2 \, dQ = 1 \right\}, \tag{4.14}$$

for $x \in \mathbb{R}$, and the multivariate function class

$$\mathcal{F}_{p,n} = \left\{ f_{\boldsymbol{\beta}}(\boldsymbol{x}) = \sum_{k=1}^{\tilde{K}_n} \psi_k(\boldsymbol{x}^{\boldsymbol{\nu}_k})\boldsymbol{\beta}_k \,:\, \int \psi_k(\boldsymbol{x}^{\boldsymbol{\nu}_k})\psi_l(\boldsymbol{x}^{\boldsymbol{\nu}_l}) \, dQ = 0 \text{ for } k \neq l, \int \{\psi_k(\boldsymbol{x}^{\boldsymbol{\nu}_k})\}^2 \, dQ = 1 \right\}, \tag{4.15}$$

for $\boldsymbol{x} \in \mathbb{R}^p$, where $\boldsymbol{\nu}_k \in \mathbb{Z}_+^p$, $\boldsymbol{x}^{\boldsymbol{\nu}_k}$ was defined in Section 4.3.4, and Q is the probability measure associated with x. In (4.14) and (4.15), we allow for the limiting case of $n = \infty$ with $\tilde{K}_\infty = \infty$. With some abuse of notation, for $\boldsymbol{\beta} \in \ell^2(\mathbb{R})$, we define $\|\boldsymbol{\beta}_{k:\infty}\|_2^2 = \sum_{l=k}^{\infty} \beta_l^2$.

To specialize Theorems 4.1 and 4.2, we need to characterize $H(\delta, \mathcal{F}_\infty^M, Q)$, for $\mathcal{F}_\infty^M$ defined below in (4.16), 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 4.2, we show that the calculation of $H(\delta, \mathcal{F}_\infty^M, Q)$ and $H(\delta, \{f_{\boldsymbol{\beta}} \in \mathcal{F}_n : \Omega(\boldsymbol{\beta}) \leq 1\}, Q_n)$ 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 for establishing the entropy conditions. The lemma considers our proposed penalty in full generality, that is the penalty (4.5) with any set of non-negative weights w_k. This lemma gives a similar reduction of entropy calculations for the multivariate case with little extra work.

Lemma 4.2 (Reduction to $\ell^2(\mathbb{R})$ and $\mathbb{R}^{\tilde{K}_n}$)**.** *Let $\mathcal{F}_n^M$ and $\mathcal{F}_{p,n}^M$ be the univariate and multivariate hierarchical basis expansion class with bounded penalty, respectively. Specifically,*

$$\mathcal{F}_n^M = \{f_{\boldsymbol{\beta}} \in \mathcal{F}_n : \sum_{k=1}^{\tilde{K}_n} w_k \|\boldsymbol{\beta}_{k:\tilde{K}_n}\|_2 \leq M\}, \quad \mathcal{F}_{p,n}^M = \{f_{\boldsymbol{\beta}} \in \mathcal{F}_{p,n} : \sum_{k=1}^{\tilde{K}_n} w_k \|\boldsymbol{\beta}_{k:\tilde{K}_n}\|_2 \leq M\}, \tag{4.16}$$

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

$H(\delta, \mathcal{H}^{w/M}_{\tilde{K}_n})$, the entropy of $\mathcal{H}^{w/M}_{\tilde{K}_n}$ with respect to the $\|\cdot\|_2$ norm, where

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

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

$$H\Big(\delta, \Big\{f_{\boldsymbol{\beta}} \in \mathcal{F}_n : \sum_{k=1}^{\tilde{K}_n} w_k \|\boldsymbol{\Psi}_{k:\tilde{K}_n}\boldsymbol{\beta}_{k:\tilde{K}_n}\|_n \leq 1\Big\}, Q_n\Big) \leq H\Big(\delta\Lambda^{-1/2}_{\max}, \mathcal{H}^{w}_{\tilde{K}_n}\Big).$$

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

Lemma 4.2 establishes the connections between entropy of the function classes of interest and the set $\mathcal{H}^{w}_{\tilde{K}_n}$. It is easy to see that $H(\delta, \mathcal{H}^{w/M}_{\tilde{K}_n})$ and $H(\delta\Lambda^{-1/2}_{\max}, \mathcal{H}^{w}_{\tilde{K}_n})$ are proportional to $H(\delta, \mathcal{H}^{w}_{\tilde{K}_n})$ where the proportionality constants depend on M and $\Lambda_{\max}$, respectively. The next lemma establishes an upper bound for $H(\delta, \mathcal{H}^{w}_{\tilde{K}_n})$ for the proposed choice of univariate and multivariate weights. This upper bound is all we need to specialize Theorem 4.2.

Lemma 4.3 (An upper bound)**.** *Suppose $\delta \geq 0$. For the region $\mathcal{H}^{w}_{\tilde{K}_n}$ with univariate weights $w_k = k^m - (k-1)^m$, $H(\delta, \mathcal{H}^{w}_{\tilde{K}_n}) \leq U_{E,1}\delta^{-1/m}$, for constant $U_{E,1} > 0$. Moreover, for the multivariate weights (4.12), we have $H(\delta, \mathcal{H}^{w}_{\tilde{K}_n}) \leq U_{E,2}\delta^{-p/m}$, for constant $U_{E,2} > 0$.*

While Lemma 4.3 is sufficient for applying Theorem 4.2, to invoke Theorem 4.1 we need an exact value for the entropy up to a proportionality constant. A natural way to achieve this is to find a lower bound for the entropy which matches the upper bound; we do this in the following lemma.

Lemma 4.4 (A lower bound)**.** *For $\delta \in ((w_1 + \cdots + w_{\tilde{K}_n+1})^{-m}, 1/2)$, for the region $\mathcal{H}^{w}_{\tilde{K}_n}$ with univariate weights, $w_k = k^m - (k-1)^m$, we have $H(\delta, \mathcal{H}^{w}_{\tilde{K}_n}) \geq L_{E,1}\delta^{-1/m}$, and for the multivariate weights (4.12) we have $H(\delta, \mathcal{H}^{w}_{\tilde{K}_n}) \geq L_{E,2}\delta^{-p/m}$, 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* (4.12).

The lemmas above 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 $\mathcal{G}_q^M = \{f_{\boldsymbol{\beta}} \in \mathcal{F}_\infty : \sum_{k=1}^\infty (k^m|\beta_k|)^q \leq M^q\}$ is the weighted L_q space (Rauhut and Ward, 2016). Furthermore, in Appendix C.5 we prove that $\mathcal{G}_1^M \subseteq \mathcal{F}_\infty^M \subseteq \mathcal{G}_2^M$. 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 gained attention in the function interpolation literature; see, for example, Rauhut and Ward (2016); Candes et al. (2008) and references therein.

4.4.4 Specializing Theorems 4.1 and 4.2

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 $\mathcal{F}$: 1) the univariate class (4.14); 2) the multivariate class (4.15); and 3) the Sobolev class $\mathcal{G}_2^M$. To prove the result, we use the fact that if an upper bound for the convergence rates can be found that matches the lower bound, then we can conclude that our estimator is minimax.

Proposition 4.1. *For the m-th order function class* $\mathcal{F}_\infty^M \equiv \{f \in \mathcal{F}_\infty : \sum_{k=1}^\infty w_k \|\boldsymbol{\beta}_{k:\infty}\|_2 \leq M\}$, *where* $w_k = k^m - (k-1)^m$, *we have*

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

For the m-th order multivariate class $\mathcal{F}_{p,\infty}^M \equiv \{f \in \mathcal{F}_{p,\infty} : \sum_{k=1}^n w_k \|\boldsymbol{\beta}_{k:\infty}\|_2 \leq M\}$, *where w_k are the weights defined in* (4.12), *we have*

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

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

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

As the last step in our analysis, we next specialize Theorem 4.2 to establish an upper bound for the convergence rate of the proposed univariate and multivariate estimators. The following proposition reveals 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 (4.16). For the Sobolev class, $\mathcal{G}_2^M$, if $\sum_{k=1}^\infty w_k \|\boldsymbol{\beta}_{k:\infty}\|_2 \leq C(M)$ for all $f_{\boldsymbol{\beta}} \in \mathcal{G}_2^M$, 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 4.2. *Consider the model* $y_i = f^0(x_i) + \varepsilon_i$ $(i = 1, \ldots, n)$ *for mean zero, sub-Gaussian noise* ε_i. *Define the univariate and multivariate estimators as*

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

for $p = 1$ *and* $p > 1$, *respectively, where* Ω^{uni} *is the penalty in* (4.9) *and* Ω^{multi} *is the penalty in* (4.12). *Assume that* $\max_k \|\psi_k\|_\infty = \psi_{\max} < \infty$ *and that the Gram matrix* $\boldsymbol{\Psi}_{\tilde{K}_n}^\top \boldsymbol{\Psi}_{\tilde{K}_n}/n$ *has a bounded maximum eigenvalue denoted by* $\Lambda_{\max}$. *Then we have the following:*

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* $1 - c\exp\{-(T/c)^2\}$,

$$\|\widehat{f}^{uni} - f^0\|_n^2 \leq 5 \max\left\{C_1 \tilde{K}_n^{-(2m-1)},\ C_2 n^{-\frac{2m}{2m+1}}\right\},$$

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

For $p = 1$, $f^0 \in \mathcal{G}_2^M$ *there is a constant* $c > 0$ *such that for all* $T \geq c$, *with probability at*

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

$$\|\widehat{f}^{uni} - f^0\|_n^2 \leq 5\max\left\{C_1\tilde{K}_n^{-(2m-1)},\ C_2C_3n^{-\frac{2m}{2m+1}}\right\},$$

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

For $p > 1$, $f^0 \in \mathcal{F}_{p,\infty}^M$, *assume that* $p < 2m$ *and define the integer* $\tilde{K}'$ *such that* $\tilde{K}_n = q_{\tilde{K}'} - 1$ *for* $q_{\tilde{K}'}$ *as defined in* (4.12). *Then there is a constant* $c > 0$ *such that for all* $T \geq c$, *with probability at least* $1 - c\exp\{-(T/c)^2\}$,

$$\|\widehat{f}^{multi} - f^0\|_n^2 \leq 5\max\left\{C_1\tilde{K}'^{-(2m-1)},\ C_2n^{-\frac{2m}{2m+p}}\right\},$$

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 4.2 is not the truncation order selected by our proposal; rather, it is the pre-specified maximum order of our proposal with conservative truncation, (4.9). The above result demonstrates that we achieve usual non-parametric 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 mth order hierarchical class, then it also belongs to the m'th order class for all $m' \leq m$ and Proposition 4.2 holds with m replaced by m'. This means we can misspecify the smoothness order in our estimator and still get nonparametric rates.

4.4.5 Theoretical results for sparse additive models

In this section, we establish the 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. This inequality establishes the consistency of estimators with slow convergence rates; specifically, these rates are $O(s\nu_n)$ where $s\nu_n^2$ is the minimax lower bound of Raskutti et al. (2009) for sparse additive model

and, s is the carnality of the set $\mathcal{S}_*$ defined below; for completeness, in Theorem 4.3 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 the order of $O(s\nu_n^2)$ and, secondly, it automatically establishes minimax rates for univariate regression as a special case of an additive model with $p = 1$. This section's contributions extend to a broad class of estimators; consequently, we can establish new results on convergence rates for some existing methods, particularly methods that extend smoothing splines and trend filtering (Meier et al., 2009; Sadhanala and Tibshirani, 2017).

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

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

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

$$\widehat{f}_1, \ldots, \widehat{f}_p = \underset{(f_j)_{j=1}^p \in \mathcal{F}}{\arg\min} \frac{1}{2n} \sum_{i=1}^{n} \left\{ Y_i - \bar{Y} - \sum_{j=1}^{p} f_j(x_{ij}) \right\}^2 + \lambda_n \sum_{j=1}^{p} I(f_j) \ , \tag{4.17}$$

where $I(\cdot)$ is a penalty of the form $I(f_j) = \|f_j\|_n + \lambda_n \Upsilon(f_j)$, for a semi-norm $\Upsilon(\cdot)$. We can think of $\Upsilon(f_j)$ as a smoothness penalty for function f_j.

Theorem 4.3 (Theorem 1, Raskutti et al. (2009))**.** *Consider n independent identically distributed samples from the sparse additive model $y_i = \sum_{j \in \mathcal{S}_*} f_j^0(x_{ij}) + \varepsilon_i$ $(i = 1, \ldots, n)$, where $|\mathcal{S}_*| = s \leq p/4$, $x_i \sim Q$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and, $f_j^0 \in \mathcal{F}$ where $\mathcal{F}$ is a class satisfying the entropy condition $H(\delta, \mathcal{F}, Q) = A_0 \delta^{-1/m}$, with $m > 1/2$. Further assume the covariates are*

independent, so, $Q = \bigotimes_{j=1}^{p} Q_j$*. Then for a constant* $C > 0$*,*

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

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

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

Theorem 4.4. *Assume the model* $y_i = f^0(x_i) + \varepsilon_i$ $(i = 1, \ldots, n)$*, with mean-zero* ε_i *satisfies* $\max_{i \in \{1,\ldots,n\}} L^2 \left(E\left[\exp\left\{ (\varepsilon_i / L)^2 \right\} \right] - 1 \right) \leq \sigma_0^2$ *, for constants* L *and* σ_0*. Assume the entropy condition* $H(\delta, \{f \in \mathcal{F} : \Upsilon(f) \leq 1\}, Q_n) \leq A_0 \delta^{-1/m}$*, holds for* $m > 1/2$*, for 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 (4.17) satisfies*

$$\|\widehat{f} - f^0\|_n^2 + \lambda_n \sum_{j \in \mathcal{S}_*^c} \|\widehat{f}_j - f_j^*\|_n + \frac{3\lambda_n^2}{2} \sum_{j \in \mathcal{S}_*} \Upsilon(\widehat{f}_j - f_j^*) \leq 3\lambda_n \sum_{j \in \mathcal{S}_*} \|\widehat{f}_j - f_j^*\|_n + 4\lambda_n^2 \sum_{j \in \mathcal{S}_*} \Upsilon(f_j^*) + \|f^* - f^0\|_n^2,$$

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*

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

where $C_s \geq 0$ *depends on* κ*,* R *and* $\sum_j \Upsilon(f_j^*)/s$*.*

The above theorem makes the same assumptions as Theorem 4.2, namely sub-Gaussian noise ε_i, and an entropy condition on the univariate function class $\mathcal{F}$. The second part of the theorem, assumes a bound on the univariate class $\mathcal{F}$ to control the term $\|\widehat{f}_j - f^*\|_n$. In the following proposition we drop this bounded class assumption and establish fast rates of convergence using the compatibility condition stated next.

Definition 7 (Compatibility Condition)**.** *We say that the compatibility condition is met for the set $\mathcal{S}_*$, if for some constant $\phi(\mathcal{S}_*) > 0$, and for all $f \in \mathcal{F} = \{f : f = \sum_{j=1}^p f_j\}$, satisfying $\sum_{j \in \mathcal{S}_*^c} \|f_j\|_n \leq 4 \sum_{j \in \mathcal{S}_*} \|f_j\|_n$, it holds that $\sum_{j \in \widetilde{\mathcal{S}_*}} \|f_j\|_n \leq |\mathcal{S}_*|^{1/2} \|f\|_n / \phi(\mathcal{S}_*)$.*

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

Proposition 4.3. *Assume the conditions of Theorem 4.4 and the compatibility condition for $\mathcal{S}_* = \{j : f_j^* \not\equiv 0\}$ hold. Then, with probability at least $1 - 2\exp(-c_1 n \rho_n^2) - c_2 \exp(-c_3 n \rho_n^2)$,*

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

where $C_f \geq 0$ is a constant that depends on $\phi(\mathcal{S}_*)$ and $\sum_j \Upsilon(f_j^*)/s$.

Misspecifying the order m is especially important here since f_j can have different orders of smoothness. Using the same argument as the univariate case, let m_j $(j = 1, \ldots, p)$ denote the smoothness order of the jth component, then our results are valid for any $m \leq \min_j m_j$. We end this section by specializing Theorem 4.4 to case of univariate regression.

Proposition 4.4. *Assuming the conditions of Theorem 4.4 with $p = 1$, the compatibility condition holds trivially with $\phi(\mathcal{S}_*) = 1$, and we have*

$$\frac{1}{2}\|\widehat{f} - f^0\|_n^2 \leq C_f n^{-\frac{2m}{2m+1}} + 2\|f^* - f^0\|_n^2,$$

with probability at least $1 - 2\exp\left\{-c_1 \kappa n^{1/(2m+1)}\right\} - c_2 \exp\left\{-c_3 \kappa n^{1/(2m+1)}\right\}$ for a constant $C_f \geq 0$ that depends on $\Upsilon(f^*)$.

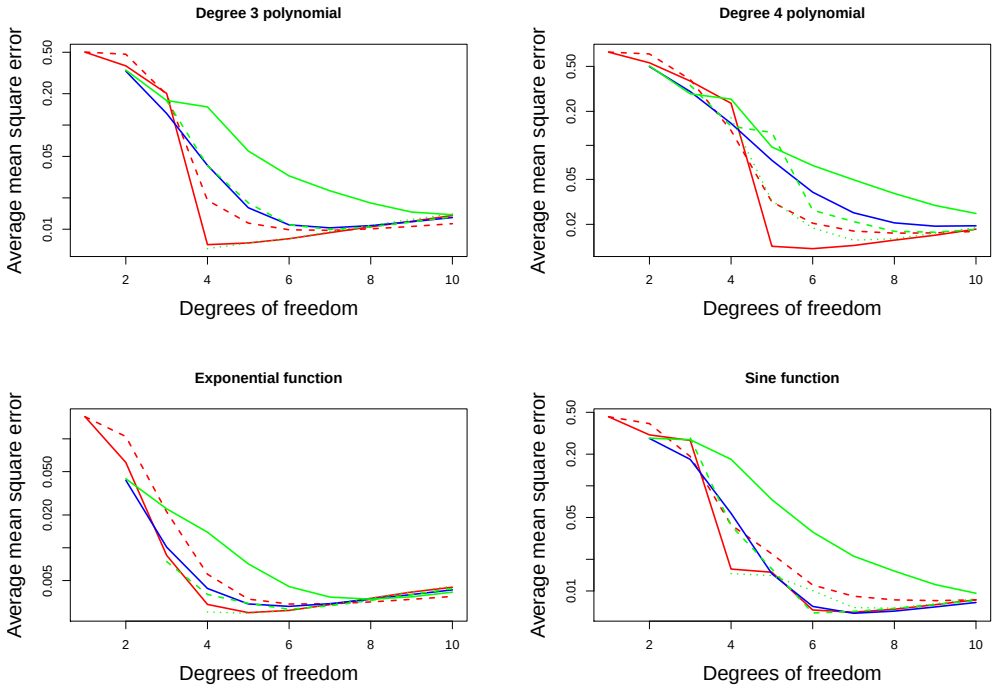


Figure 4.4: Average mean square error, over 100 simulated datasets, as a function of degrees of freedom for true models given by $g_1, \ldots, g_4$ in (4.18). The colored lines indicate results for our framework with $m = 3$ (——) and 1 (- - - - -), Trend Filtering of order 1 (——), 2 (- - - - -) and 3 (·····), and Smoothing Splines (——).

4.5 Simulation studies

4.5.1 Simulation for univariate regression

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

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

$1, \ldots, n$). This facilitates comparison to trend filtering, which can become substantially slow for random x_i, particularly when the covariates are not uniformly distributed over a closed interval. We consider $n = 150$, SNR $\in \{2, 3\}$, and four different f^0 functions, denoted g_t $(t = 1, \ldots, 4)$, defined as

$$\begin{aligned} g_1(x) &= -0{\cdot}43 + 4{\cdot}83x - 14{\cdot}65x^2 + 11{\cdot}76x^3, \\ g_2(x) &= 0{\cdot}23 - 8{\cdot}44x + 45{\cdot}20x^2 - 81{\cdot}41x^3 + 46{\cdot}59x^4, \\ g_3(x) &= \exp(-5x + 0{\cdot}5) - 0{\cdot}4\sinh(2{\cdot}5), \quad g_4(x) = -\sin(7x - 0{\cdot}4). \end{aligned} \tag{4.18}$$

We applied our proposal using a sequence of 100 λ values linear on the log scale from $\lambda_{\max}$, for which $\widehat{\boldsymbol{\beta}} = 0$, down to $10^{-4}\lambda_{\max}$. We applied smoothing splines on a grid of 100 values of degrees of freedom from 10 to 1. Trend filtering is applied to a sequence of λ values automatically selected by its `R` implementation with an average length of 279, 448 and 612 for trend filtering of order 1, 2 and 3, respectively.

Figure 4.4 displays the mean square prediction error, MSE $= \|f^0 - \widehat{f}\|_n^2$, of our method with $m = 1$ and 3, smoothing splines and trend filtering of orders 1, 2 and 3, as a function of 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 we proved our method to converge with rates comparable to smoothing splines. Since the functions considered in this simulation are smooth, as expected, we see that our method with $m = 1$ does not converge as fast as competing methods.

Figure 4.5, shows examples of some fitted models for a fixed value of degrees of freedom. Our method seems to perform very well and is mostly robust to changes in the value of m. The smoothing splines estimates are unable to do as well for the same effective degrees of freedom. The plots in the bottom panel of Figure 4.5 also suggest that first order trend filter can perform poorly in presence of model misspecification.

To compare the performances for an optimal value of λ, for each method we find a λ^*

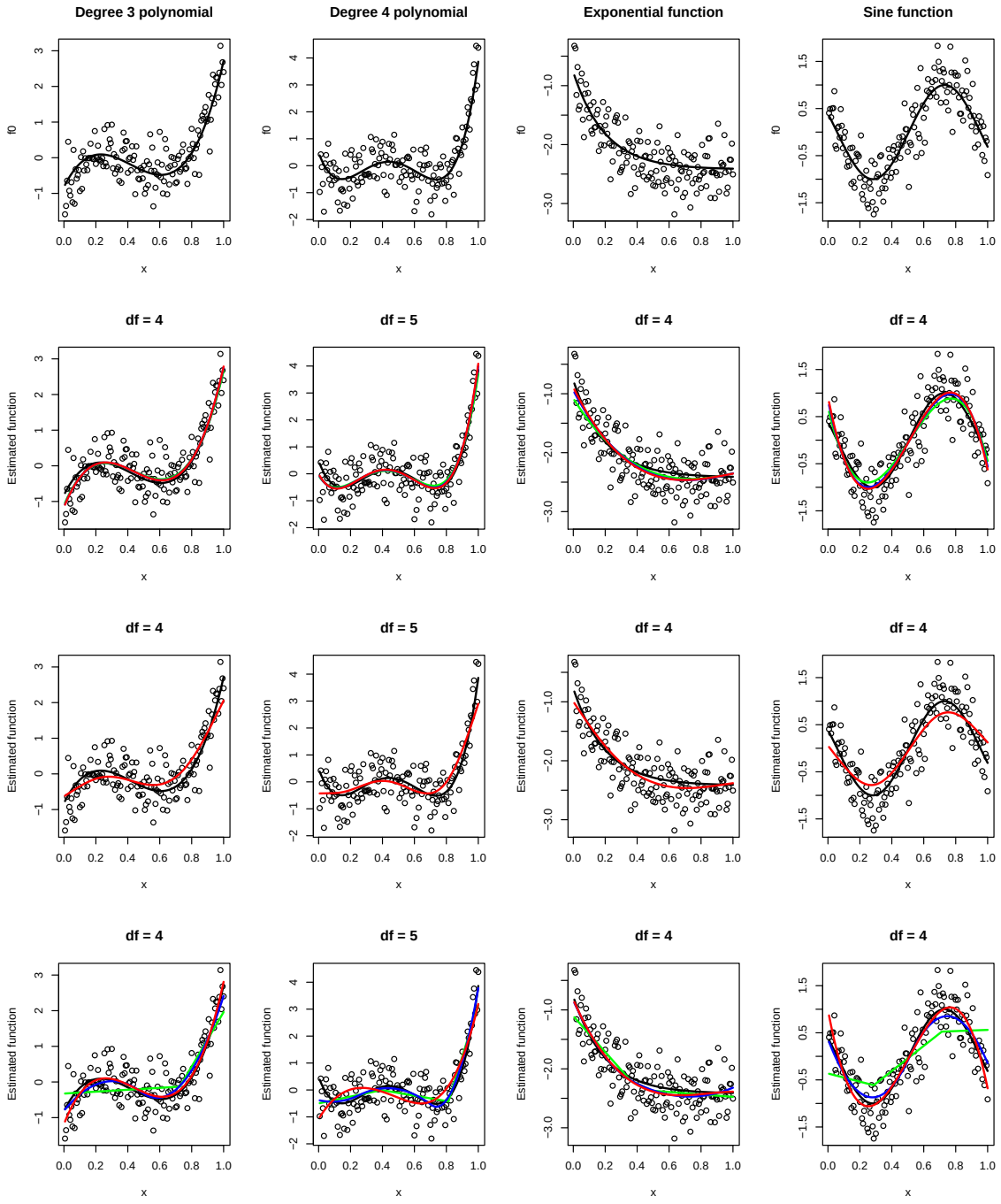


Figure 4.5: Scatterplots of simulated data along with true and estimated functions. The top row includes plots of the simulated data along with the true function used for generating the data. The other three rows show the fitted functions for each method and the degrees of freedom corresponding to the fitted model, our univariate proposal, smoothing splines and trend filtering are shown in rows 2-4, respectively. For trend filtering and our proposal with $m = 1$, 2 and 3 are shown in green (——), blue (——) and red (——). Only the available R implementation with order $m = 3$ is shown for smoothing splines.

Table 4.2: 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 along with 100× standard errors 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.

that minimizes the prediction error on an independent test set of size $n_{test} = 75$. For this λ^*, we report the ratio of MSE for each method and the MSE of our proposal in Table 4.2. The table also includes similar ratios for degrees of freedom.

4.5.2 Simulation for multivariate additive regression

We proceed with a simulation study to illustrate the performance of our sparse additive framework compared to the performance of Ravikumar et al.'s method. Ravikumar et al.'s method is implemented in the `R` package `SAM` (Zhao et al., 2014) which uses natural spline basis functions. To facilitate 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 to these methods to future work.

We consider the simulation setting of Meier et al. (2009) with some modifications to have high-dimensional data and smaller signal-to-noise ratio. We generate $n = 200$ samples for $p = 500$ features. The data is generated as $y_i = 5f_1(x_{i1}) + 3f_2(x_{i2}) + 4f_3(x_{i3}) + 6f_4(x_{i4}) + \varepsilon_i$

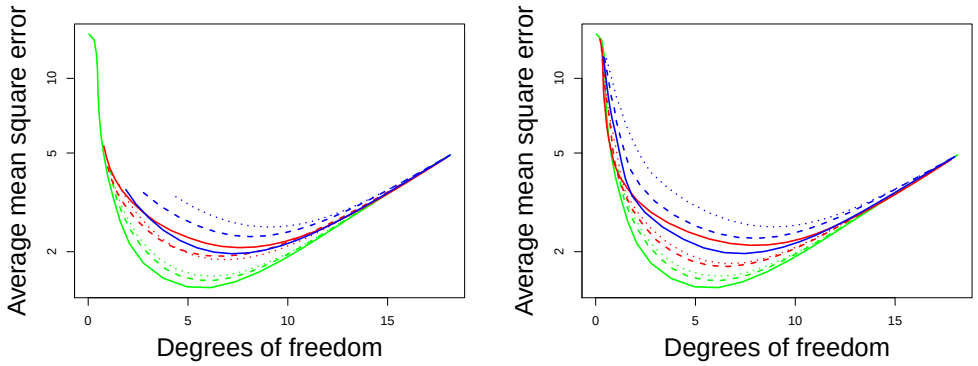


Figure 4.6: Average mean square error, over 100 simulated datasets, as a function of degrees of freedom for our proposal with $m = 1$ (——), $m = 2$ (- - - - -) and $m = 3$ (·····), compared to Ravikumar et al. (2009) with 3 (——), 5 (- - - - -), 8 (·····), 10 (——), 15 (- - - - -) and 20 (·····) basis functions. Left: natural spline basis functions. Right: polynomial basis functions.

$(i = 1, \ldots, n)$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with σ^2 such that SNR $= 3$ and

$$f_1(x) = x, \quad f_2(x) = (2x-1)^2, \quad f_3 = 2\sin(2\pi x)/\left\{2 - \sin(2\pi x)\right\},$$
$$f_4(x) = 0{\cdot}1\sin(2\pi x) + 0{\cdot}2\cos(2\pi x) + 0{\cdot}3\sin^2(2\pi x) + 0{\cdot}4\cos^3(2\pi x) + 0{\cdot}5\sin^3(2\pi x)\ ;$$

the covariates are independently drawn from Uniform$(0, 1)$. We implemented the parametrization (4.10) for $m = 1$,

$$\underset{\beta_j \in \mathbb{R}^{\tilde{K}}}{\text{minimize}} \frac{1}{2}\left\| y - \sum_{j=1}^{p} \Psi_{\tilde{K}}^{j} \beta_j \right\|_n^2 + \gamma\lambda \sum_{j=1}^{p} \Omega_j(\beta_j) + (1-\gamma)\lambda \sum_{j=1}^{p} \|\Psi_{\tilde{K}}^{j} \beta_j\|_n \,, \tag{4.19}$$

for $m = 2$ and 3 with $\gamma = 0{\cdot}01$ and $0{\cdot}001$, respectively. All methods were fit for a sequence of 50 λ values, decreasing linearly on the log-scale. We fix the maximum number of basis functions $\tilde{K} = 20$ for our sparse additive framework and implement Ravikumar et al.'s proposal with 3, 5, 8, 10, 15 and, 20 basis functions.

In terms of mean square prediction error, it is not surprising to observe superior perfor-

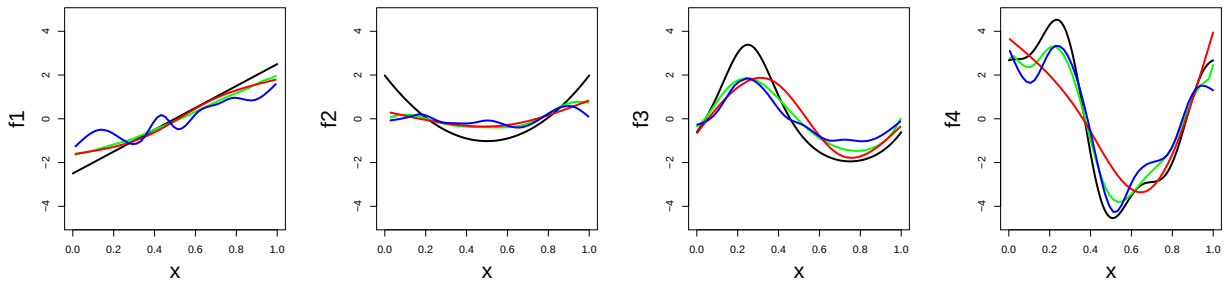


Figure 4.7: The first 4 component functions of the simulation study from Section 4.5.2. The estimates of our proposal are shown in green (——), whereas that of Ravikumar et al. (2009) fitted with 20 and 3 basis functions are shown in blue (——) and red (——), respectively. In each case, the tuning parameter leading to the smallest mean square error was used.

mance of our methodology over that of Ravikumar et al.'s proposal with polynomial basis in Figure 4.6. 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 degrees of freedom values in both panels of Fig. 4.6.

In Figure 4.7, we show some of the fitted functions by our proposal and that of Ravikumar et al. (2009) using the λ value which minimizes the test set error for Ravikumar et al.'s proposal with 3 and 15 basis functions.

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 consists 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. Apart from 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 2/3 of the data as training and remaining as test, and average the results over 100 training-test splits.

We compare the performance of our additive framework to (4.2) with fixed truncation, $K_j = K$, in the low-dimensional setting and compare our sparse additive framework to that of Ravikumar et al. (2009), lasso (Tibshirani, 1996), and elastic net (Zou and Hastie, 2005) in the high-dimensional setting. For computational convenience, we fit our proposal with conservative basis truncation (4.9) 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.'s proposal with $K = 2, 4, 8$; to facilitate 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; re-fitting the selected non-zero coefficients via least squares. For a sequence of λ values, or sequence of K for (4.2), we calculate the mean square test error and the model parsimony, the number of total non-zero basis functions used. Lower values of model parsimony correspond to more sparsity in either the number of components or truncation levels K_j. Figure 4.8, shows that our proposal outperforms competitors in the low and high-dimensional setting in terms of mean square error for a sequence of fitted models. The relaxed version of Ravikumar et al.'s method is able to achieve a lower test error but at the cost of added parsimony. The linear models, implemented by lasso and elastic net, had a substantially higher test error compared to additive models.

4.7 Discussion

In this chapter we introduced a novel approach to non-parametric regression and high-dimensional additive models. Recall the original motivation: for non-parametric regression, especially additive models, we require an estimator that can adapt to function complexity in a data-adaptive way. We showed that state-of-the-art methods like those of Ravikumar et al. (2009) and Lou et al. (2016) are unable to do that effectively. More data adaptive proposals, such as the sparsity smoothness penalty of Meier et al. (2009), come at a cost of highly complex fitted models even for simple underlying surfaces. The use of hierarchical penalty allows us to adaptively fit simple models for simple functions as shown in Sections 4.5. Our

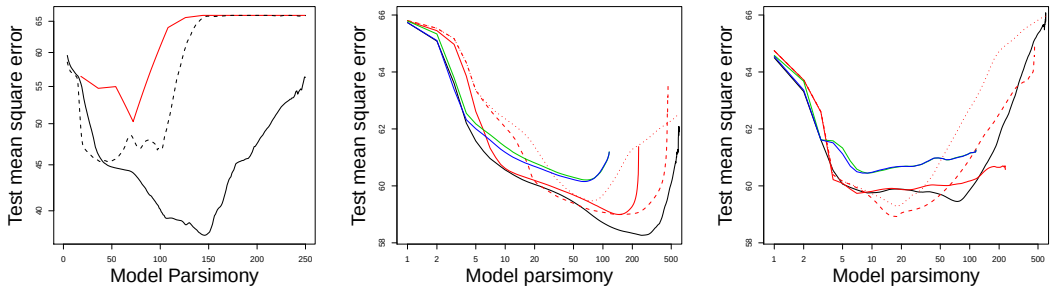


Figure 4.8: Results of analysis of parkinsons telemonitoring dataset. Left: Results for low dimensional case for our proposal (——), The simple truncation estimator (4.2) (——) and our relaxed proposal (------). Centre: Results for high-dimensional case for our proposal (——), Ravikumar et al.'s proposal with 2 (——), 4 (------), and 8 (.....) basis functions, lasso (——) and elastic net (——). Right: Results for the relaxed versions of all methods in the centre panel.

theoretical analyses in Section 4.4 establishes fast convergence rates for a broad class sparse additive estimators.

The `R` package `HierBasis`, available on `https://github.com/asadharis/HierBasis`, implements the methods described in this chapter.

Chapter 5

WAVELET REGRESSION AND ADDITIVE MODELS FOR IRREGULARLY SPACED DATA

5.1 Introduction

We consider the canonical task of estimating a regression function, f, from observations $\{(x_i, y_i) : i = 1, \ldots, n\}$, with $\boldsymbol{x}_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ and $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$ $(i = 1, \ldots, n)$, where ε_i are independent, mean 0, sub-Gaussian random variables. A popular approach for estimating f is to use linear combinations of a pre-specified set of *basis functions*, e.g., polynomials, splines (Wahba, 1990), wavelets Daubechies (1992), or other systems (Čencov, 1962). The weights, or coefficients, in such a linear combination are often determined using some form of penalized regression. In this chapter, we focus on estimators that use a *wavelet* basis. Wavelet-based estimators have compelling theoretical properties; however, a number of issues, discussed in the reminder of this section, have limited their adaptation in many non-parametric applications. The approach proposed in this chapter overcomes these issues. Throughout the chapter, we assume basic knowledge of wavelet methods though some key points will be reviewed; for a detailed introduction to wavelets see books by Daubechies (1992); Percival and Walden (2006); Vidakovic (2009); Nason (2010); Ogden (2012).

Wavelets are a system of orthonormal basis functions for $L^2([0, 1])$. Wavelets are popular for representing functions because they allow *time and frequency localization* (Daubechies, 1990) as opposed to, say, Fourier bases, which allow only frequency localization. Additionally, wavelet-based methods are computationally efficient; the main ingredient of wavelet regression is the discrete wavelet transform (DWT) and its inverse (IDWT) which can be computed in $O(n)$ operations (Mallat, 1989). Unfortunately, traditional wavelet methods require stringent conditions on the data, specifically that $x_i = i/n$ with $n = 2^J$ for some integer

J. This is not a problem in many signal processing applications with regularly sampled signals; however, in general non-parametric regression, this condition will rarely be satisfied. A simple solution for general data types is to ignore irregular spacing of data (Cai and Brown, 1999; Sardy et al., 1999) and/or artificially extend the signal such that $n = 2^J$ (Strang and Nguyen, 1996, Ch. 8). Other solutions include transformations (Cai and Brown, 1998; Pensky and Vidakovic, 2001) or interpolation (Hall et al., 1997; Kovac and Silverman, 2000; Antoniadis and Fan, 2001) of the data to a regular grid of size 2^J. The literature on univariate wavelet methods is quite extensive and cannot be adequately discussed within this chapter. In contrast, the literature on wavelet methods for multiple covariates is rather limited, particularly when the number of covariates is large.

For the multivariate settings with $\boldsymbol{x}_i \in \mathbb{R}^p$ for $p \geq 2$, we consider estimating an additive model, that is $\widehat{f}(\boldsymbol{x}_i) = \sum_j \widehat{f}_j(x_{ij})$. Additive models naturally extend linear models to capture non-linear conditional relationships, while retaining some interpretability; they also do not suffer from the *curse of dimensionality*. Despite these benefits, wavelet-based additive models have received limited attention. This is most likely because data with multiple covariates are rarely available on a regular grid of size $n = 2^J$. Sardy and Tseng (2004) fit additive models by treating data as if regularly spaced; however, they do not discuss the case when n is not a power of 2. A number of proposals transform the data to a regular grid (Amato and Antoniadis, 2001; Zhang et al., 2003; Grez and Vidakovic, 2018). However, to do this, the density of the covariates must be estimated, which unnecessarily invokes the curse of dimensionality. In addition, to the best of our knowledge, there are no wavelet-based methods for fitting additive models in high dimensions (when $p > n$) that induce sparsity, i.e., for many j, give a solution with $\widehat{f}_j \equiv 0$.

In this chapter, we give a simple proposal that effectively extends wavelet-based methods to non-parametric modeling with a potentially large number of covariates. More specifically, we present an interpolation-based approach for dealing with irregularly spaced data when n is not necessarily a power of 2. However, unlike existing interpolation methods, we do not transform the raw data $(\boldsymbol{x}_i, y_i)$. As a result, our method naturally extends to additive

and sparse additive models. We also propose a penalized estimation framework to induce sparsity in high dimensions. We develop a proximal gradient descent method for computation of our estimator, which leverages fast algorithms for DWT and sparse matrix multiplication. Furthermore, we establish adaptive minimax convergence rates (up to a $\log n$ factor) similar to that of existing wavelet methods for regularly spaced data. We also establish convergence rates for our (sparse) additive proposal for a potentially large number of covariates. We discuss some extensions of our proposal to generalized additive models and an adaptive version of our proposal which exhibits improved performance.

The remainder of this chapter is organized as follows. In Section 5.2 we present our univariate, additive and sparse additive proposals. The univariate case ($p = 1$) is mainly presented to motivate our proposal. We also present our main algorithm for computing the estimator. In Section 5.3 we establish the convergence rates of our estimators, and present empirical studies in Section 5.4. Concluding remarks are presented in Section 5.5.

5.2 Methodology

5.2.1 Short background on wavelets

We begin with a quick review of wavelet methods for nonparametric regression covering 3 main ingredients: (1) wavelet basis functions, (2) the discrete wavelet transform (DWT) and, (3) shrinkage.

First, *wavelets* are a system of orthonormal basis functions for $L^2([0,1])$ or $L^2(\mathbb{R})$. The bases are generated by translations and dilations of special functions $\phi(\cdot)$ and $\psi(\cdot)$ called the *father* and *mother* wavelet, respectively. In greater detail, for any $j_0 \geq 0$, a function $f \in L^2([0,1])$ can be written as

$$f(x) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k}\phi_{j_0 k}(x) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}(x), \tag{5.1}$$

where

$$\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k), \quad \psi_{jk}(x) = 2^{j/2}\psi(2^j x - k).$$

The coefficients $\alpha_{j_0 k}$ and β_{jk} are called the father and mother wavelet coefficients, respectively. The index j is called the *resolution level* and j_0 is the *minimum resolution level*. Different choices of ϕ and ψ generate various wavelet families; popular choices are Daubechies (Daubechies, 1988), Coiflets (Daubechies, 1993), Meyer wavelets (Meyer, 1985), and Spline wavelets (Chui, 1992); for an overview of wavelet families, see Ogden (2012). Here, we can consider functions with a truncated basis expansion, i.e., functions of the form $f(x) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k}\phi_{j_0 k}(x) + \sum_{j=j_0}^{J}\sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}(x)$, for some J. For regular data with $x_i = i/n$ $(i = 1, \ldots, n)$ and $n = 2^J$ for some J, we can calculate the vector $\boldsymbol{f} = [f(1/n), f(2/n), \ldots, f(n/n)]^\top$ efficiently via our second ingredient described next.

Any vector $\boldsymbol{f} = [f(1/n), f(2/n), \ldots, f(n/n)]^\top$, for function f with truncated wavelet basis expansion of order J, can be written as a linear combination of that truncated wavelet basis. In particular $\boldsymbol{f} = \boldsymbol{W}^\top \boldsymbol{d}$, where $\boldsymbol{d} = \left(\alpha_{j_0 0}, \ldots, \alpha_{j_0 2^{j_0}-1}, \beta_{j_0 0}, \beta_{j_0 1}, \ldots, \beta_{J 2^J-1}\right)^\top$ is the vector of wavelet coefficients, and the rows of $\boldsymbol{W}$ contain the corresponding wavelet basis functions evaluated at $x_i = i/n$. Specifically, $\boldsymbol{W}$ is an orthogonal matrix with $W_{li} \approx \sqrt{n}\psi_{jk}(i/n)$, or $W_{li} \approx \sqrt{n}\phi_{jk}(i/n)$, for some l; the $\sqrt{n}$ factor is due to convention in the literature and software implementation. By orthogonality, $\boldsymbol{d} = \boldsymbol{W}\boldsymbol{f}$; this transformation from $\boldsymbol{f}$ to its wavelet coefficients via multiplication by W is known as the discrete wavelet transform (DWT). The transformation from wavelet coefficients to fitted values, via multiplication by $\boldsymbol{W}^\top$ is known as the inverse discrete wavelet transform (IDWT). The DWT and IDWT can be computed in $O(n)$ operations via Mallat's pyramid algorithm (Mallat, 1989). However, this is only possible for $n = 2^J$.

Finally, shrinkage is employed to give the estimates of the form $\widehat{\boldsymbol{f}} = \boldsymbol{W}^\top \widehat{\boldsymbol{d}}$; for ease of exposition, we will assume $j_0 = 0$; i.e., all except the first element of $\boldsymbol{d}$ correspond to mother wavelet coefficients. Our methodology and theoretical results do not depend on the choice

of j_0. The wavelet shrinkage estimator is given by

$$\widehat{\boldsymbol{d}} \leftarrow \underset{\boldsymbol{d}\in\mathbb{R}^n}{\arg\min} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{W}^\top \boldsymbol{d}\|_2^2 + \lambda \sum_{i=2}^{n} |d_i|, \tag{5.2}$$

for a positive tuning parameter λ, and given data $\{(i/n, y_i) \in \mathbb{R}^2 : i = 1, \ldots, n\}$. The ℓ_1 penalty shrinks the mother coefficients and also induces sparsity. The sparsity-inducing penalty is motivated by the desirable *parsimony* property of wavelets: many functions in $L^2([0,1])$ are a sparse linear combination of wavelet bases. The optimization problem (5.2) can be solved exactly as follows: define $\widetilde{\boldsymbol{d}} = \boldsymbol{W}\boldsymbol{y}$, the DWT of $\boldsymbol{y}$. Then, $\widehat{d}_1 = \widetilde{d}_1$ and $\widehat{d}_i = sgn(\widetilde{d}_i)(|\widetilde{d}_i| - 2\lambda)_+$ $(i = 2, \ldots, n)$ where $(x)_+ = \max(x, 0)$. Thus, for regularly spaced data with $n = 2^J$, wavelet bases provide an efficient nonparametric estimator. In the following subsection, we discuss some existing methods for dealing with irregularly spaced data and present our novel proposal, `waveMesh`.

5.2.2 A novel interpolation scheme

The common approach to dealing with irregularly spaced data is to map outcomes at the data $\{(x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, n\}$ to approximate outcomes on the regular grid $\{(i/n, y_i') \in \mathbb{R}^2 : i = 1, \ldots, 2^J\}$ via either interpolation or transformation of the data. The novelty of our approach is a reversal of the direction of interpolation, i.e., interpolation from fitted values on the regular grid $i/2^J$ $(i = 1, \ldots, 2^J)$ to approximated fits on the raw data x_i $(i = 1, \ldots, n)$. In greater detail, given $\boldsymbol{f} = [f(1/K), \ldots, f(K/K)]$, a vector of fitted values on the regular grid at i/K $(i = 1, \ldots, K = 2^J)$, we define $\widetilde{f}(\cdot)$ as the function obtained by interpolation of $\boldsymbol{f}$. Specifically, we are interested in interpolation schemes defined as linear transformations of $\boldsymbol{f}$. For example, linear interpolation gives $\widetilde{f}(x) = \boldsymbol{r}(x)^\top \boldsymbol{f}$, where $\boldsymbol{r}(x) = t(x)\boldsymbol{e}_{\lfloor Kx \rfloor} + (1 - t(x))\boldsymbol{e}_{\lceil Kx \rceil}$, with $t(z) = (Kz - \lfloor Kz \rfloor)/(\lceil Kz \rceil - \lfloor Kz \rfloor)$, where $\boldsymbol{e}_j$ is the j-th coordinate vector. Given such an interpolation scheme, let $\boldsymbol{R} \in \mathbb{R}^{n \times 2^J} = (\boldsymbol{r}(x_1), \ldots, \boldsymbol{r}(x_n))^\top$ denote our interpolation matrix; note that $\boldsymbol{R}\boldsymbol{f} \equiv [\widetilde{f}(x_1), \ldots, \widetilde{f}(x_n)]^\top$. Our proposal, `waveMesh`, solves

the following convex optimization problem

$$\widehat{\boldsymbol{d}} \leftarrow \underset{\boldsymbol{d}\in\mathbb{R}^K}{\arg\min}\ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}\|_2^2 + \lambda\|\boldsymbol{d}_{-1}\|_1, \tag{5.3}$$

where $K = 2^{\lceil \log_2 n \rceil}$, $\boldsymbol{d}_{-1} = [d_2, \ldots, d_n]^\top \in \mathbb{R}^{K-1}$, and $\boldsymbol{W} \in \mathbb{R}^{K\times K}$ is the usual DWT matrix. To evaluate the `waveMesh` estimate at a new point $x \in \mathbb{R}$, one can use $\boldsymbol{r}(x)^\top \boldsymbol{W}^\top \widehat{\boldsymbol{d}}$, with $\boldsymbol{r}(\cdot)$ as given by the chosen interpolation scheme. The advantage of `waveMesh`, over existing methods, is that it can naturally be extended to additive models. Given data $\{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{p+1} : i = 1, \ldots, n\}$, let $\boldsymbol{R}_j \in \mathbb{R}^{n\times K}$ be the interpolation matrix corresponding to covariate j, that is, $\boldsymbol{R}_j \boldsymbol{f} = [\widetilde{f}(x_{1j}), \ldots, \widetilde{f}(x_{nj})]^\top$. Then, `waveMesh` can be extended to fitting additive models by the following optimization problem:

$$\widehat{\boldsymbol{d}}_1, \ldots, \widehat{\boldsymbol{d}}_p \leftarrow \underset{\boldsymbol{d}_1,\ldots,\boldsymbol{d}_p\in\mathbb{R}^K}{\arg\min} \frac{1}{2}\Big\|\boldsymbol{y} - \sum_{j=1}^p \boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}_j\Big\|_2^2 + \lambda \sum_{j=1}^p \|\boldsymbol{d}_{j,-1}\|_1, \tag{5.4}$$

and $\widehat{\boldsymbol{f}} = [\widehat{f}(\boldsymbol{x}_1), \ldots, \widehat{f}(\boldsymbol{x}_n)]^\top = \sum_{j=1}^p \widehat{\boldsymbol{f}}_j = \sum_{j=1}^p \boldsymbol{R}_j \boldsymbol{W}^\top \widehat{\boldsymbol{d}}_j$. Finally, we can extend additive `waveMesh` to fitting sparse additive models for a potentially large number of covariates. This can be achieved by adding a sparsity inducing penalty for each component f_j as follows:

$$\widehat{\boldsymbol{d}}_1, \ldots, \widehat{\boldsymbol{d}}_p \leftarrow \underset{\boldsymbol{d}_1,\ldots,\boldsymbol{d}_p\in\mathbb{R}^K}{\arg\min} \frac{1}{2}\Big\|\boldsymbol{y} - \sum_{j=1}^p \boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}_j\Big\|_2^2 + \sum_{j=1}^p \left[\lambda_1\|\boldsymbol{d}_{j,-1}\|_1 + \lambda_2\|\boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}_j\|_2\right]. \tag{5.5}$$

5.2.3 Algorithm for `waveMesh` and sparse additive `waveMesh`

We now present a proximal gradient descent algorithm (Parikh and Boyd, 2014) for solving the optimization problem (5.3). For convex loss ℓ and penalty P, the proximal gradient descent algorithm iteratively finds the minimizer of $\{\ell(\boldsymbol{d}) + P(\boldsymbol{d})\}$ via the iteration:

$$\boldsymbol{d}^{(l+1)} \leftarrow \underset{\boldsymbol{d}\in\mathbb{R}^K}{\arg\min}\ \frac{1}{2}\Big\|\left(\boldsymbol{d}^{(l)} - t_l \nabla \ell(\boldsymbol{d}^{(l)})\right) - \boldsymbol{d}\Big\|_2^2 + t_l P(\boldsymbol{d}),$$

for a step-size $t_l > 0$. The algorithm is guaranteed to converge as long as $t_l \leq L^{-1}$ where L is the Lipschitz constant of $\nabla \ell(\cdot)$. The step-size can be fixed or selected via a line search algorithm. For (5.3), we obtain the following iterative scheme:

$$\boldsymbol{d}^{(l+1)} \leftarrow \underset{\boldsymbol{d} \in \mathbb{R}^K}{\arg\min} \frac{1}{2} \left\| \left\{ (\boldsymbol{I}_K - t_l \boldsymbol{R}^\top \boldsymbol{R}) \boldsymbol{W}^\top \boldsymbol{d}^{(l)} + t_l \boldsymbol{R}^\top \boldsymbol{y} \right\} - \boldsymbol{W}^\top \boldsymbol{d} \right\|_2^2 + t_l \lambda \|\boldsymbol{d}_{-1}^{(l)}\|_1. \tag{5.6}$$

Our algorithm has a number of desirable features which make it computationally efficient. Firstly, (5.6) is the traditional wavelet problem for regularly spaced data (5.2), with response vector $\boldsymbol{r} = \{(\boldsymbol{I}_K - t_l \boldsymbol{R}^\top \boldsymbol{R}) \boldsymbol{W}^\top \boldsymbol{d}^{(l)} + t_l \boldsymbol{R}^\top \boldsymbol{y}\}$. The vector $\boldsymbol{r}$ can efficiently be calculated via the sparsity of R and Mallets algorithm for DWT. Secondly, we can use a fixed step size with $t_l = L_{\max}^{-1}$ where $L_{\max}$ is the maximum eigenvalue of $\boldsymbol{R}^\top \boldsymbol{R}$. Again, the maximum eigenvalue can be efficiently computed for sparse matrices, e.g., if $\boldsymbol{R}$ is the linear interpolation matrix then $\boldsymbol{R}^\top \boldsymbol{R}$ is tridiagonal, and its eigenvalues can be calculated in $O(K \log k)$ operations.

The procedure (5.6) can also be used to solve the additive (5.4) and sparse additive (5.5) extensions via a block coordinate descent algorithm. Specifically, given a set of estimates $\boldsymbol{d}_j$ $(j = 1, \ldots, p)$ we can fix all but one of the vectors $\boldsymbol{d}_j$ and optimize over the non-fixed vector, by solving

$$\underset{\boldsymbol{d} \in \mathbb{R}^K}{\text{minimize}} \frac{1}{2} \|\boldsymbol{r}_j - \boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}\|_2^2 + \lambda_1 \|\boldsymbol{d}_{-1}\|_1 + \lambda_2 \|\boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}\|_2, \tag{5.7}$$

for some vector $\boldsymbol{r}_j \in \mathbb{R}^n$. For additive `waveMesh` ($\lambda_2 = 0$), this reduces to the univariate problem which can be solved via the algorithm (5.6). For sparse additive `waveMesh` ($\lambda_2 \neq 0$), the problem can be solved by solving (5.7) with $\lambda_2 = 0$ following by a soft-scaling operation (Lemma 3.3). We detail our algorithm for sparse additive `waveMesh` in Appendix D.1.

5.2.4 Some extensions and variations

In this subsection, we discuss some variations and extensions of `waveMesh`, namely (1) using a conservative order for the wavelet basis expansion, (2) extending `waveMesh` for more general loss functions and, (3) using a weighted ℓ_1 penalty for shrinkage of wavelet coefficients.

While in (5.3) we set $K = 2^{\lceil \log_2 n \rceil}$, we could, instead, set K to be any power of 2. Since the main computational step in our algorithm is the DWT and IDWT which requires $O(K)$ operations, a smaller value of K can greatly reduce the computation time. Furthermore, using a smaller K can lead to superior predictive performance in some settings; this is formalized in our theoretical results of Section 5.3 and observed in the simulation studies of Section 5.4.

Secondly, `waveMesh` can be extended to other loss functions appropriate for various data types. For example, we can extend our methodology to the setting of binary classification via a logistic loss function. Let $y_i \in \{-1, 1\}$ $(i = 1, \ldots, n)$ be the observed response. For the univariate case, we get

$$\widehat{\boldsymbol{d}} \leftarrow \underset{\boldsymbol{d} \in \mathbb{R}^K}{\arg\min} \frac{1}{2} \sum_{i=1}^{n} \log\left(1 + \exp\left[-y_i (\boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d})_i\right]\right) + \lambda \|\boldsymbol{d}_{-1}\|_1. \tag{5.8}$$

Like the least squares loss, this naturally extends to (sparse) additive models. The problem can be efficiently solved via a proximal gradient descent algorithm described in the supplementary material.

Finally, we consider a variation of our ℓ_1 penalty motivated by the `SURESHRINK` procedure of Donoho and Johnstone (1995). For a vector $\boldsymbol{d} \in \mathbb{R}^K$ of discrete father and mother wavelet coefficients, denote by $\boldsymbol{d}_{[j]}$ the discrete mother wavelet coefficients at resolution level j. For this particular extensions, we require that the minimum resolution level $j_0 > 1$. We then propose to solve

$$\widehat{\boldsymbol{d}} \leftarrow \underset{\boldsymbol{d} \in \mathbb{R}^K}{\arg\min} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}\|_2^2 + \lambda \sum_{j=j_0}^{\log_2 K} \sqrt{2 \log(j)} \|\boldsymbol{d}_{[j]}\|_1. \tag{5.9}$$

In Section 5.4 we show that the above estimator outperforms the usual `waveMesh` estimator (5.3) in terms of prediction error.

5.3 Theoretical results

In this section, we study finite sample properties of our univariate estimator (5.3), and sparse additive estimator (5.5). We begin with a quick introduction to Besov spaces and their connection to wavelet bases. We establish minimax convergence rates (up to a $\log n$ factor) for our univariate proposal. We note that our estimator (5.3) ca be seen as a lasso estimator Tibshirani (1996) with design matrix $\boldsymbol{RW}^\top$; this allows us to use well-known results for the lasso estimator to easily establish minimax rates. Finally, we also establish rates for the sparse additive `waveMesh` proposal for a specific penalty.

Besov spaces on the unit interval, $B^s_{q_1,q_2}$, are function spaces with specific degrees of smoothness in their derivative, i.e., $B^s_{q_1,q_2} = \left\{ g \in L^2([0,1]) : \|g\|_{B^s_{q_1,q_2}} < C \right\}$, for the Besov norm $\|\cdot\|_{B^s_{q_1,q_2}}$. The constants (s, q_1, q_2) are the parameters of Besov spaces; for a function $g \in L^2([0,1])$ with the wavelet bases expansion (5.1), the Besov norm is defined as

$$\|g\|_{B^s_{q_1,q_2}} = \|\boldsymbol{\alpha}_{j_0}\|_{q_1} + \left[\sum_{j=j_0}^{\infty} \left\{ 2^{j(s+1/2-1/q_1)} \|\boldsymbol{\beta}_j\|_{q_1} \right\}^{q_2} \right]^{1/q_2}, \tag{5.10}$$

where $\boldsymbol{\alpha}_{j_0} \in \mathbb{R}^{2^{j_0}}$ is the vector of father wavelet coefficients with minimum resolution level j_0 and $\boldsymbol{\beta}_j \in \mathbb{R}^{2^j}$ is the vector of mother wavelet coefficients at resolution level j. For completeness, we also define $\|g\|_{B^s_{q_1,\infty}} = \|\boldsymbol{\alpha}_{j_0}\|_{q_1} + \sup_{j \geq j_0} \left\{ 2^{j(s+1/2-1/q_1)} \|\boldsymbol{\beta}_j\|_{q_1} \right\}$. We consider Besov spaces because they generalize well-known classes such as the Sobolev ($B^s_{2,2},\ s = 1, 2, \ldots$), and Hölder ($B^s_{\infty,\infty},\ s > 0$) spaces and the class of bounded total variation functions (sandwiched between $B^1_{1,1}$ and $B^1_{1,\infty}$). The theorem below establishes near minimax convergence rates for the prediction error of our estimator. An attractive feature of our estimator is that it achieves this rate without any information about the parameters $(s,\ q_1,\ q_2)$. We recover the usual wavelet rates of Donoho (1995) under the special case when $x_i = i/n$ and $R = I_n$. Additionally, the theorem justifies the use of $K < n$ basis functions: if the true function is sufficiently smooth, we recover the usual rates with an additional $\log K$ factor instead of $\log n$.

Theorem 5.1. *Consider the model $y_i = f^0(x_i) + \varepsilon_i$ $(i = 1, \ldots, n)$ for mean zero, sub-Gaussian noise ε_i. Define the estimator $\widehat{\boldsymbol{f}} = \boldsymbol{R}\boldsymbol{W}^\top \widehat{\boldsymbol{d}} = [\widehat{f}(x_1), \ldots, \widehat{f}(x_n)]^T$ for linear interpolation matrix $\boldsymbol{R}$ where*

$$\widehat{\boldsymbol{d}} \leftarrow \underset{\boldsymbol{d} \in \mathbb{R}^K}{\arg\min} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}\|_2^2 + \lambda \|\boldsymbol{d}_{-1}\|_1,$$

for the usual DWT transform matrix $\boldsymbol{W} \in \mathbb{R}^{K \times K}$ associated with some orthogonal wavelet family. Further define, $\boldsymbol{f}^0 = [f^0(x_1), \ldots, f^0(x_n)]^\top$ and $\widetilde{\boldsymbol{f}}^0 = [f^0(1/K), \ldots, f^0(K/K)]^\top$. Assume that $f^0 \in B^s_{q_1,q_2}$ and the mother wavelet ψ, has r null moments and r continuous derivatives where $r > \max\{1, s\}$. Suppose $\lambda \geq c_1\sqrt{t^2 + 2\log K}$ for some $t > 0$. Then, for sufficiently large K (specifically $K \geq c_1 n^{1/(2s+1)}$ for some constant c_1), we have with probability at-least $1 - 2\exp(-t^2/2)$ the following inequality:

$$\frac{1}{n}\left\|\boldsymbol{f}^0 - \widehat{\boldsymbol{f}}\right\|_2^2 \leq C\left(\frac{\log K}{n}\right)^{\frac{2s}{2s+1}} + \frac{2}{n}\|\boldsymbol{f}^0 - \boldsymbol{R}\widetilde{\boldsymbol{f}}^0\|_2^2,$$

where the constant c_1 depends on R and the distribution of ε_i, and the constant C depends on $\boldsymbol{R}$.

The above theorem includes an interpolation error term $\|\boldsymbol{f}^0 - \boldsymbol{R}\widetilde{\boldsymbol{f}}^0\|_2^2$. For a sufficiently large K (particularly $K = n$), with any reasonable interpolator, the approximation error will disappear.

For the sparse additive model, we consider a different model motivated by the Besov norm (5.10). The theorem below provides convergence rates for the estimated function at the data $\widehat{\boldsymbol{f}} = \sum_{j=1}^p \widehat{\boldsymbol{f}}_j = \sum_{j=1}^p \boldsymbol{R}_j \boldsymbol{W}^\top \widehat{\boldsymbol{d}}_j$ where

$$\widehat{\boldsymbol{d}}_1, \ldots, \widehat{\boldsymbol{d}}_p \leftarrow \underset{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p \in \mathbb{R}^K}{\arg\min} \frac{1}{2}\left\|\boldsymbol{y} - \sum_{j=1}^p \boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}_j\right\|_2^2 + \sum_{j=1}^p \left[\lambda_1 P_s(\boldsymbol{d}_j) + \lambda_2 \|\boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}_j\|_2\right], \quad (5.11)$$

where the penalty P_s is the discrete version of the Besov norm for $B^s_{1,1}$. Specifically, for $\boldsymbol{d}$ a vector of father coefficients, $\alpha_{j_0 k}$ $(k = 0, \ldots, 2^{j_0} - 1)$, and mother wavelet coefficients

β_{jk} $(j = j_0, \ldots, J; k = 0, \ldots, 2^j - 1)$ the penalty is

$$P_s(\boldsymbol{d}) = \sum_{k=0}^{2^{j_0}-1} |\alpha_{j_0 k}| + \sum_{j=j_0}^{J} \left(2^{j(s-1/2)} \sum_{k=0}^{2^j-1} |\beta_{jk}| \right). \tag{5.12}$$

Convergence rates for our proposal `waveMesh` can easily be derived using the general theoretical framework of Section 4.4.5. With a similar compatibility condition (stated below) we can prove minimax rates for `waveMesh` without any additional work. As a special case, we also establish minimax convergence rates for the low-dimensional problem for which the compatibility condition trivially holds. Furthermore, the compatibility condition can be relaxed at the cost of proving a slower rate, similar to the slow rates of Theorem 4.4. The final ingredient is entropy bounds for the Besov space $B^s_{q_1,q_2}$; such bounds are well established in the literature, see for example Nickl and Pötscher (2007).

Definition 8. *The compatibility condition is said to hold for an index set $\mathcal{S}_* \subset \{1, 2, \ldots, p\}$, with compatibility constant $\vartheta(\mathcal{S}_*) > 0$, if for all $\gamma > 0$ and any set of discrete wavelet coefficients vector $(\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p)$, that satisfy*

$$\sum_{j \in \mathcal{S}_*^c} n^{-1} \|\boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}_j\|_2 + \gamma \sum_{j=1}^{p} P_s(\boldsymbol{d}_j) \leq 3 \sum_{j \in \mathcal{S}_*} \|\boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}_j\|,$$

it holds that

$$\sum_{j \in \mathcal{S}_*} \|\boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}_j\|_2 \leq \Big\| \sum_{j=1}^{p} \boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}_j \Big\|_2 \sqrt{|\mathcal{S}_*|} / \vartheta(\mathcal{S}_*).$$

Theorem 5.2. *Assume the model $y_i = f^0(\boldsymbol{x}_i) + \varepsilon_i$ $(i = 1, \ldots, n)$ with mean zero, sub-Gaussian ε_i. Let $\widehat{f} = \sum_{j=1}^{p} \widehat{\boldsymbol{f}}_j$ be as defined in (5.11), and let*

$$\boldsymbol{f}^* = \sum_{j \in \mathcal{S}_*} \boldsymbol{f}_j^* = \sum_{j \in \mathcal{S}_*} \boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}_j^*,$$

be an arbitrary sparse additive function with $\mathcal{S}_* \subset \{1, 2, \ldots, p\}$. *Let*

$$\rho = \kappa \max\{n^{-2s/(2s+1)}, (\log p/n)^{1/2}\},$$

for a constant κ *that depends on the distribution of* ε_i *and* s. *Suppose* $\lambda \geq 4\rho$. *Then, with probability at-least* $1 - 2\exp(-c_1 n\rho^2) - c_2 \exp(-c_3 n\rho^2)$, *we have*

$$\frac{1}{n}\left\|\boldsymbol{f}^0 - \widehat{\boldsymbol{f}}\right\|_2^2 \leq C_1 \max\left\{|\mathcal{S}_*| n^{-\frac{s}{2s+1}}, |\mathcal{S}_*|\Big(\frac{\log p}{n}\Big)^{1/2}\right\} + \frac{1}{n}\left\|\boldsymbol{f}^0 - \boldsymbol{f}^*\right\|_2^2,$$

where constants c_1, c_2 *depend on the distribution of* ε_i *and* s, *and* C_1 *depends on* κ *and* $|\mathcal{S}_*|^{-1}\sum_{j\in\mathcal{S}_*} P_s(\boldsymbol{d}_j^*)$. *Furthermore, if the compatibility condition holds for* $\mathcal{S}_*$ *with constant* $\vartheta(\mathcal{S}_*)$ *we have*

$$\frac{1}{n}\left\|\boldsymbol{f}^0 - \widehat{\boldsymbol{f}}\right\|_2^2 \leq C_2 \max\left\{|\mathcal{S}_*| n^{-\frac{2s}{2s+1}}, |\mathcal{S}_*|\frac{\log p}{n}\right\} + \frac{4}{n}\left\|\boldsymbol{f}^0 - \boldsymbol{f}^*\right\|_2^2,$$

where the constant C_2 *depends on* $\vartheta(\mathcal{S}_*)$ *and* $|\mathcal{S}_*|^{-1}\sum_{j\in\mathcal{S}_*} P_s(\boldsymbol{d}_j^*)$.

5.4 Simulation studies

5.4.1 Simulation for univariate regression

We begin with a simulation to compare the performance of univariate `waveMesh` to the traditional interpolation method of Kovac and Silverman (2000), isometric wavelet method of Sardy et al. (1999)—which treats the data as if it were regularly spaced—and adaptive lifting method of Nunes et al. (2006). The former two methods are implemented in the `R` package `wavethres` (Nason, 2016) and the latter is implemented in the `adlift` (Nunes and Knight, 2017) package.

We generate the data as $y_i = f^0(x_i) + \varepsilon_i$ $(i = 1, \ldots, n)$ for different choices of function f^0 and n. To facilitate comparison with isometric wavelets, we consider only samples sizes that were a power of 2. The errors are distributed as $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with σ^2 chosen such

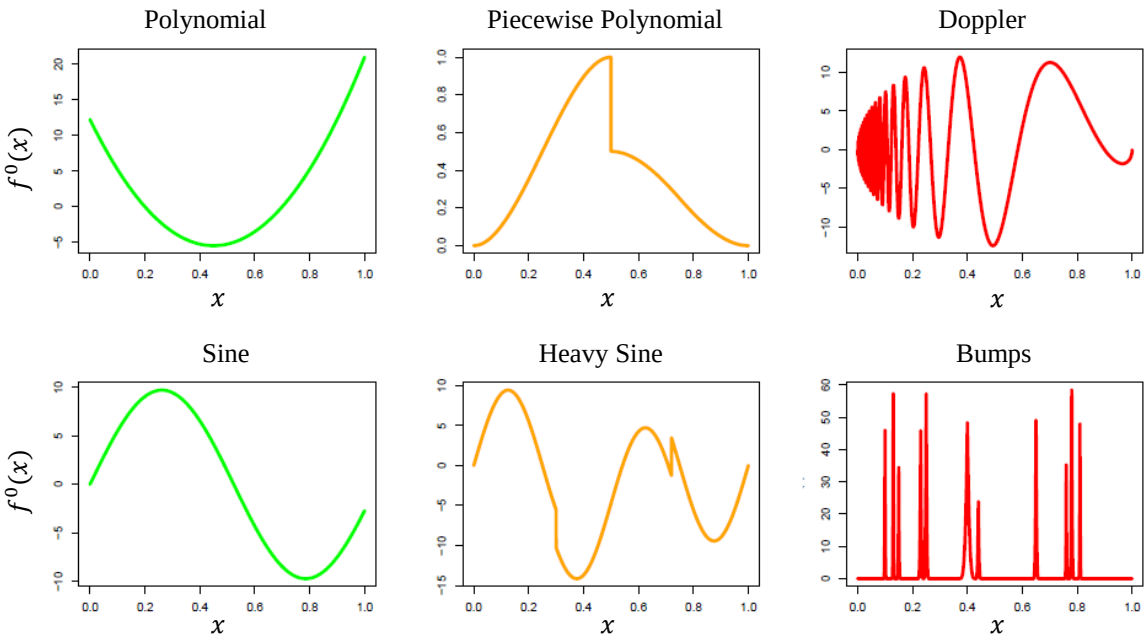


Figure 5.1: Plots of functions f^0 for the simulation study of Section 5.4. Functions in green are the most smooth and well-behaved followed by functions with moderate smoothness in orange. Finally, functions in red are highly irregular functions, e.g., functions with unbounded total variation.

that SNR $= 5$, where SNR $= \text{var}(\boldsymbol{f}^0)/\sigma^2$. We consider two different choices of the covariate, $x_i \sim \mathcal{U}[0,1]$ and $x_i \sim \mathcal{N}(0,1)$ scaled to lie in $[0,1]$. We consider 6 different choices for the function f^0 shown in Figure 5.1. We apply our proposal, `waveMesh`, the interpolation proposal of Kovac and Silverman (2000) and isometric wavelet proposal of Sardy et al. (1999), for a sequence of 50 λ values linear on the log scale and select the λ value which minimizes the mean square error, MSE $= n^{-1}\left\|\boldsymbol{f}^0 - \widehat{\boldsymbol{f}}\right\|_2^2$. For adaptive lifting, the `R` implementation automatically selects a tuning parameter. We also implement `waveMesh` using a small grid, i.e., we fit 5.3 with $K = 2^5$ and 2^6.

Table 5.1 shows the ratio of MSE between our proposal with $K = 2^{\lceil \log_2 n \rceil}$ and other proposals for uniformly distributed x_i. We observe that our proposal has the smallest MSE

for all functions except the Bumps function. Even for the Bumps function, `waveMesh` exhibits superior prediction performance over other methods for $n = 512$. We also observe that `waveMesh` with smaller values of K often outperforms the full `waveMesh` ($K = 2^{\lceil \log_2 n \rceil}$) method in terms of MSE. Results for normally distributed x_i are presented in Table 5.2. In that case, we again observe that our method outperforms existing methods for a number of simulation scenarios, except for a few cases with polynomial and Bumps functions.

5.4.2 Simulation for multivariate additive regression

We proceed with a simulation study to illustrate the performance of additive `waveMesh` compared to the the proposal of Sardy and Tseng (2004), AMlet. We use the author-provided `R` implementation for the Amlet proposal; due to a lack of `R` packages for other proposals we defer the comparison to future work. We consider the following simulation setting: we generate data with $y_i = f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) + \varepsilon_i$ $(i = 1, \ldots, 2^{10})$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $x_i \sim \mathcal{U}[0, 1]$, ans σ^2 such that SNR = 10. The four functions $f_1, \ldots, f_4$ are the polynomial, sine, piecewise polynomial and heavy sine functions presented in Figure 5.1. We consider sample sizes $n = 2^6, 2^7, 2^8$ and 2^9 and results were averaged over 100 data sets. Sample sizes were powers of 2 to facilitate comparison to AMlet; existing `R` implementation of AMlet requires sample size to be a power of 2. The universal threshold rule was used for AMlet as detailed in Sardy and Tseng (2004), 5-fold cross validation was used for additive `waveMesh` for selection of λ.

Table 5.3 shows the MSE of both proposals for various choices of n. The results clearly indicate that additive `waveMesh` offers substantial improvement over AMlet, especially for smaller values of n. These results support our theoretical analysis and underscore the advantages of `waveMesh` in sparse high-dimensional additive models.

5.4.3 Effect of truncation level K

In this subsection, we present simulation results which study the effects of using different truncation levels K. In Figures 5.2 to 5.7 we plot the results for each of the 6 functions

Table 5.1: Table of results for $x_i \sim \mathcal{U}[0,1]$ averaged over 100 replications of the data. The table presents the ratio MSE / MSE_{FG} along with $100\times$ the standard error, where MSE_{FG} is the MSE of `waveMesh` with $K = 2^{\lceil \log_2 n \rceil}$. Boldface values represent the method with the smallest MSE within each row of the table.

		`waveMesh` $K = 2^5$	`waveMesh` $K = 2^6$	Interpolation	Isometric	Adaptive Lifting
Polynomial	n = 64	1.19 (5.51)	**1.00** (0.00)	1.24 (4.11)	1.78 (7.56)	4.28 (29.86)
	n = 128	0.92 (5.57)	**0.77** (3.07)	1.12 (6.00)	1.33 (7.18)	3.57 (31.27)
	n = 256	1.00 (6.20)	**0.85** (3.15)	1.61 (9.04)	1.50 (7.67)	4.29 (31.29)
	n = 512	0.78 (3.18)	**0.72** (2.58)	1.76 (6.11)	1.13 (2.64)	3.61 (26.47)
Sine	n = 64	**0.97** (3.14)	1.00 (0.00)	1.47 (5.81)	1.59 (6.72)	3.62 (33.65)
	n = 128	**0.76** (3.18)	0.76 (1.96)	1.29 (6.08)	1.46 (5.24)	2.98 (19.78)
	n = 256	**0.66** (2.50)	0.70 (2.22)	1.93 (9.49)	1.34 (4.23)	3.41 (18.80)
	n = 512	0.57 (2.34)	**0.56** (2.22)	2.13 (7.78)	1.24 (3.66)	3.63 (28.42)
Piecewise	n = 64	**0.85** (1.97)	1.00 (0.00)	1.18 (3.12)	1.31 (3.62)	1.63 (9.07)
Polynomial	n = 128	**0.77** (2.00)	0.82 (1.52)	1.26 (2.75)	1.22 (2.61)	1.40 (7.36)
	n = 256	0.82 (1.92)	**0.79** (1.59)	1.42 (3.18)	1.14 (2.11)	1.15 (6.04)
	n = 512	1.01 (2.43)	**0.86** (1.70)	1.71 (3.56)	1.15 (1.99)	1.25 (7.24)
Heavy Sine	n = 64	**0.84** (2.44)	1.00 (0.00)	1.12 (3.04)	1.41 (3.17)	1.70 (8.35)
	n = 128	**0.75** (2.66)	0.82 (1.16)	1.17 (3.32)	1.50 (4.75)	1.56 (8.26)
	n = 256	**0.66** (1.64)	0.72 (1.14)	1.37 (2.98)	1.33 (2.58)	1.53 (6.74)
	n = 512	**0.58** (1.59)	0.60 (1.18)	1.58 (3.05)	1.29 (1.60)	1.50 (9.21)
Bumps	n = 64	2.11 (2.30)	1.00 (0.00)	1.70 (1.75)	**0.72** (1.34)	1.07 (5.12)
	n = 128	2.86 (2.77)	2.11 (1.62)	1.40 (1.59)	**0.63** (0.83)	0.85 (2.43)
	n = 256	4.81 (6.82)	3.47 (4.39)	1.43 (1.89)	**0.88** (0.99)	0.97 (2.00)
	n = 512	7.45 (9.13)	5.69 (6.77)	1.32 (1.35)	**1.19** (1.03)	1.23 (2.34)
Doppler	n = 64	**0.98** (1.69)	1.00 (0.00)	1.15 (3.45)	1.33 (3.20)	1.30 (3.65)
	n = 128	1.24 (2.02)	**0.89** (1.04)	1.07 (2.13)	1.44 (2.57)	1.18 (3.22)
	n = 256	1.71 (3.92)	**0.94** (1.38)	1.20 (2.11)	1.29 (1.99)	1.30 (3.44)
	n = 512	2.58 (4.85)	1.26 (2.01)	1.21 (1.48)	**1.10** (1.31)	1.23 (3.36)

Table 5.2: Table of results for $x_i \sim \mathcal{N}(0,1)$ averaged over 100 replications of the data. The table presents the ratio MSE / MSE_{FG} along with $100\times$ the standard error, where MSE_{FG} is the MSE of `waveMesh` with $K = 2^{\lceil \log_2 n \rceil}$. Boldface values represent the method with the smallest MSE within each row of the table.

		`waveMesh` $K = 2^5$	`waveMesh` $K = 2^6$	Interpolation	Isometric	Adaptive Lifting
Polynomial	n = 64	1.47 (13.17)	1.41 (11.32)	**0.51** (3.04)	1.45 (10.11)	1.59 (9.30)
	n = 128	0.78 (5.25)	0.77 (4.95)	**0.40** (2.96)	0.87 (4.69)	0.88 (4.84)
	n = 256	**0.39** (3.75)	0.51 (3.89)	0.43 (2.38)	0.64 (2.81)	0.76 (5.97)
	n = 512	0.90 (4.57)	0.77 (4.03)	**0.29** (0.98)	0.43 (1.59)	0.33 (2.36)
Sine	n = 64	0.92 (9.55)	0.99 (1.49)	1.48 (11.85)	2.22 (21.67)	3.61 (35.07)
	n = 128	**0.89** (8.74)	0.91 (3.77)	1.71 (10.85)	1.83 (15.18)	3.53 (33.07)
	n = 256	**0.48** (2.39)	0.73 (1.53)	1.48 (8.74)	1.51 (8.18)	2.73 (22.25)
	n = 512	**0.36** (1.22)	0.64 (1.63)	1.03 (5.54)	0.74 (2.77)	1.21 (7.62)
Piecewise	n = 64	**0.78** (1.92)	0.99 (1.01)	1.50 (6.50)	1.64 (7.54)	2.18 (14.06)
Polynomial	n = 128	0.86 (2.29)	**0.83** (2.04)	1.89 (7.42)	1.59 (4.60)	1.65 (8.86)
	n = 256	1.25 (3.80)	**0.90** (2.22)	1.64 (5.21)	1.09 (3.24)	1.15 (6.94)
	n = 512	1.79 (2.71)	1.27 (2.34)	1.76 (3.24)	**0.96** (1.54)	1.01 (4.29)
Heavy Sine	n = 64	**0.73** (1.81)	1.00 (0.65)	1.23 (4.40)	1.26 (4.03)	1.54 (6.83)
	n = 128	**0.54** (1.70)	0.78 (1.40)	1.30 (5.02)	1.14 (2.78)	1.12 (6.04)
	n = 256	**0.47** (0.93)	0.65 (0.98)	1.17 (3.08)	0.89 (1.99)	0.93 (5.45)
	n = 512	**0.38** (0.87)	0.54 (1.08)	1.40 (2.91)	0.77 (1.24)	0.84 (3.94)
Bumps	n = 64	1.27 (0.62)	1.00 (0.06)	0.85 (1.19)	**0.36** (0.79)	0.53 (2.24)
	n = 128	3.40 (4.69)	2.25 (2.81)	1.35 (2.28)	**0.69** (1.50)	0.76 (1.64)
	n = 256	6.49 (10.88)	3.71 (5.58)	1.31 (2.03)	1.18 (1.41)	**1.10**† (2.52)
	n = 512	8.83 (10.06)	5.43 (6.03)	1.29 (1.82)	1.28 (1.37)	**1.11**† (1.90)
Doppler	n = 64	**0.75** (1.84)	1.00 (0.67)	1.36 (4.74)	1.53 (4.32)	1.56 (6.01)
	n = 128	0.99 (1.87)	**0.81** (1.44)	1.43 (4.75)	1.49 (3.81)	1.40 (4.35)
	n = 256	0.58 (1.11)	**0.52** (1.06)	1.26 (3.25)	1.15 (1.86)	0.98 (3.77)
	n = 512	0.98 (1.52)	**0.58** (1.05)	1.24 (2.38)	0.98 (1.48)	0.85 (2.21)

considered in the simulation of Section 5.4.1.

In the left panel of each figure we plot the MSE as a function of sample size, n. This is done for the full grid method where we take $K = 2^{\log_2 n}$, and for `waveMesh` with $K = 2^4, 2^5$ and 2^6 which we refer to as 4 Grid, 5 Grid and 6 Grid, respectively. In the right panel of each figure we present the computation time as a function of sample size n for `waveMesh` with $K = 2^4, 2^5, 2^6$ and $2^{\log_2 n}$.

We see in Figures 5.6 and 5.7, that using a small order K leads to substantially high MSE. This is most likely due to the nature of the underlying functions. The Doppler function is an example of function which does not have a bounded variation, estimating such functions by interpolation is extremely difficult and in general we need a full grid, i.e. $K = n$. On the other hand for all other functions, i.e. polynomial, sine etc, we see a clear advantage of using $K = 2^7$ basis functions. We also see in some figures that while using $K = 2^6$ leads to substantially smaller MSE using too small a value of K can be lead to poor prediction performance. We see this even in the simple cases of estimating a polynomial or sine function.

We notice on the right panels the clear computational advantage of using fewer than n basis functions. We observe the computation time for fixed K generally does not vary too much with increasing sample size. This is because the main computational step is the DWT and IDWT via Mallets algorithm. The other matrix multiplications are sparse and can be computed efficiently.

Table 5.3: MSE and standard error of additive `waveMesh` and AMlet over 100 data sets of size $n = 64, 128, 256, 512$.

	$n = 64$	$n = 128$	$n = 256$	$n = 512$
Additive `waveMesh`	**10.95** (0.28)	**8.34** (0.17)	**5.32** (0.10)	**3.76** (0.06)
AMlet	100.48 (1.83)	45.49 (1.09)	19.57 (0.33)	8.90 (0.11)

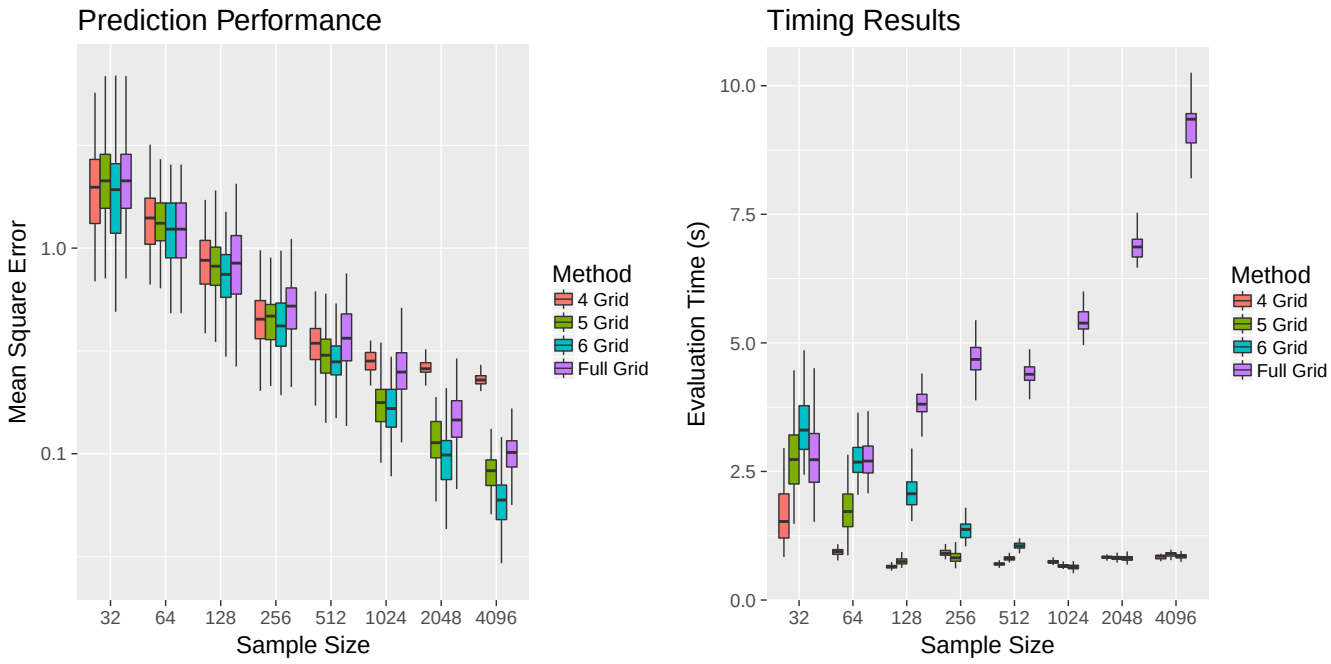


Figure 5.2: Results of simulation study of Section 5.4.3: Polynomial function.

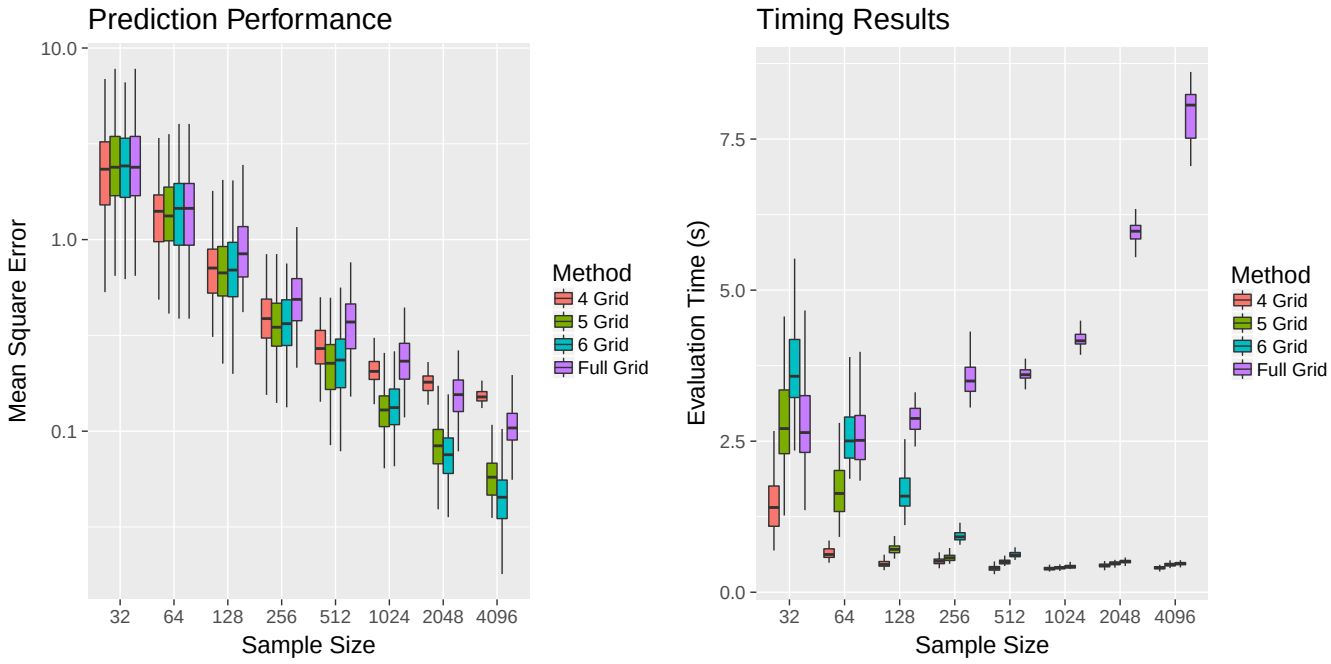


Figure 5.3: Results of simulation study of Section 5.4.3: Sine function.

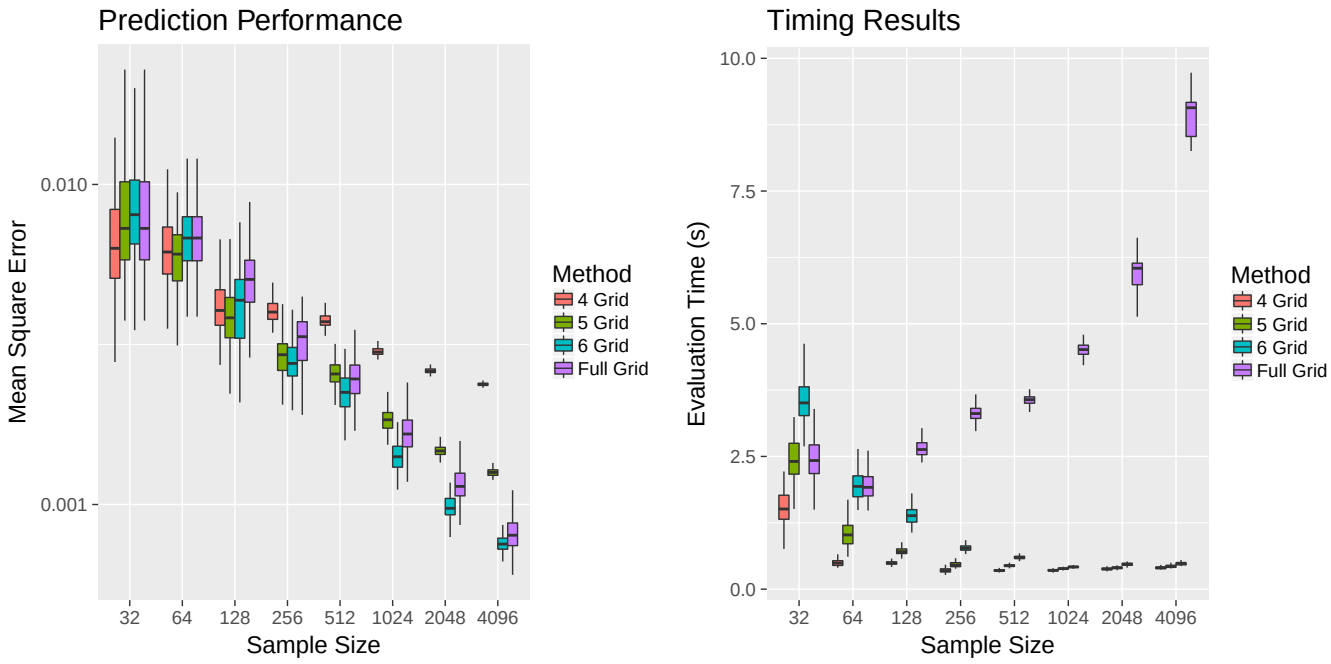


Figure 5.4: Results of simulation study of Section 5.4.3: Piecewise Polynomial function.

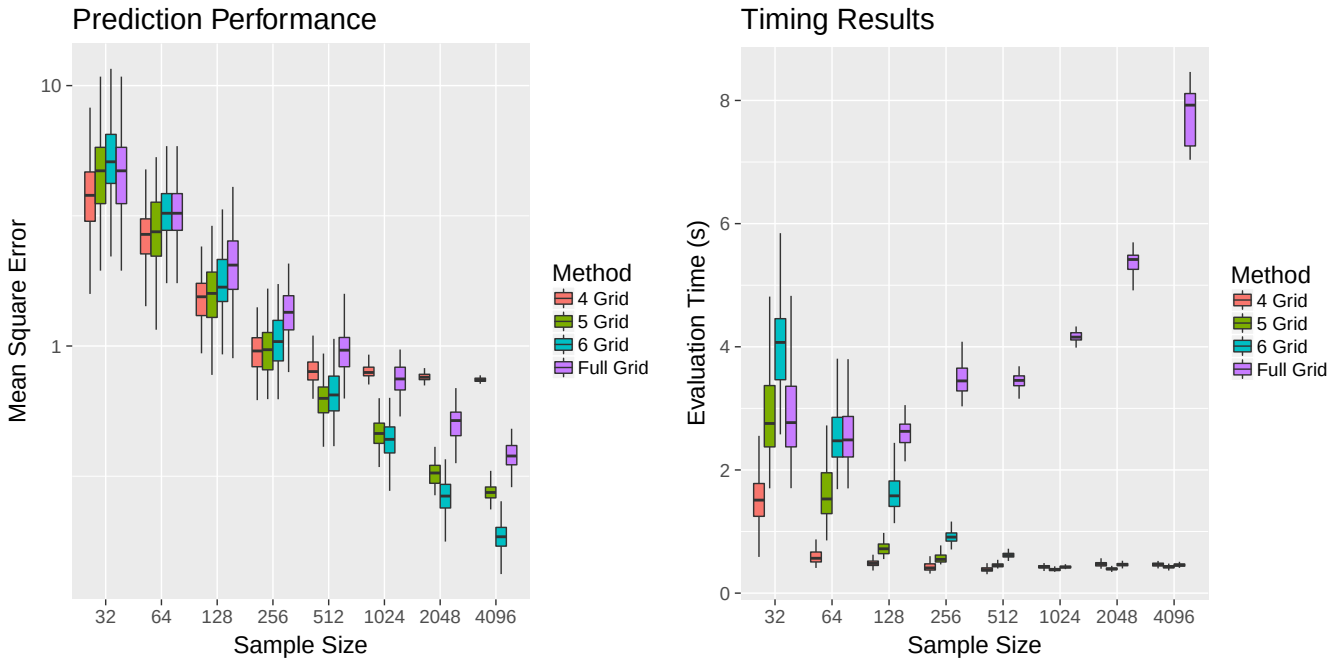


Figure 5.5: Results of simulation study of Section 5.4.3: Heavy sine function.

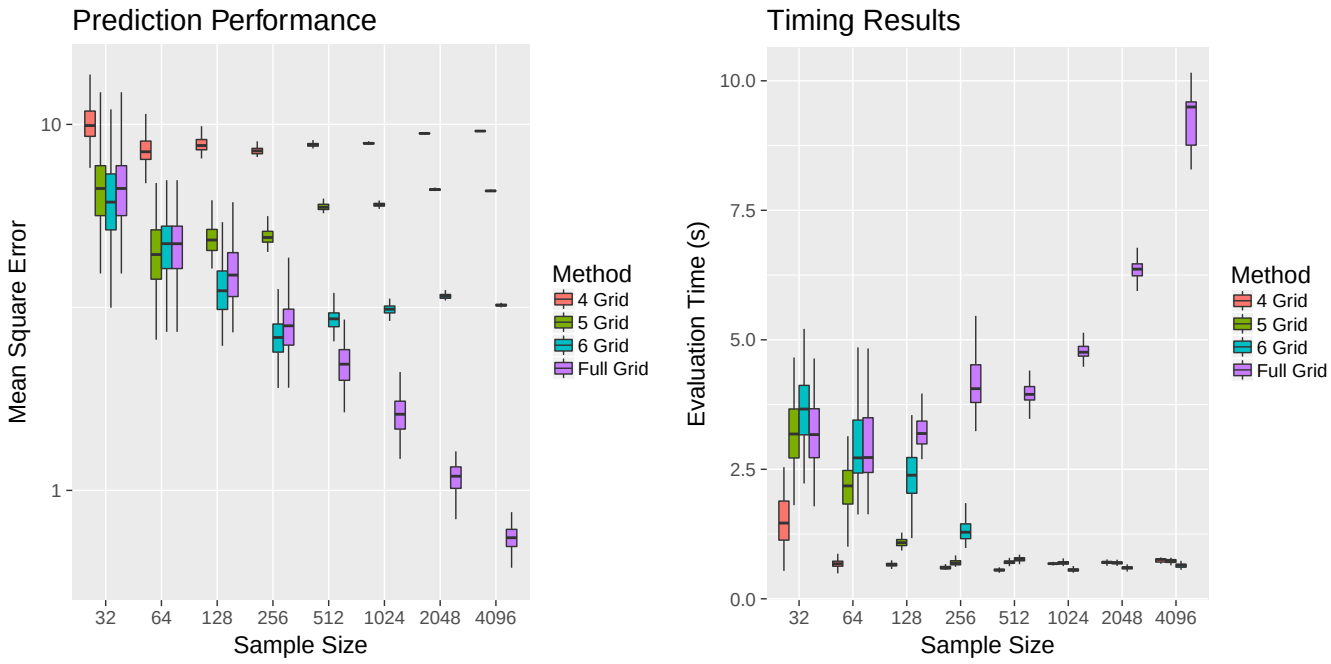


Figure 5.6: Results of simulation study of Section 5.4.3: Doppler function.

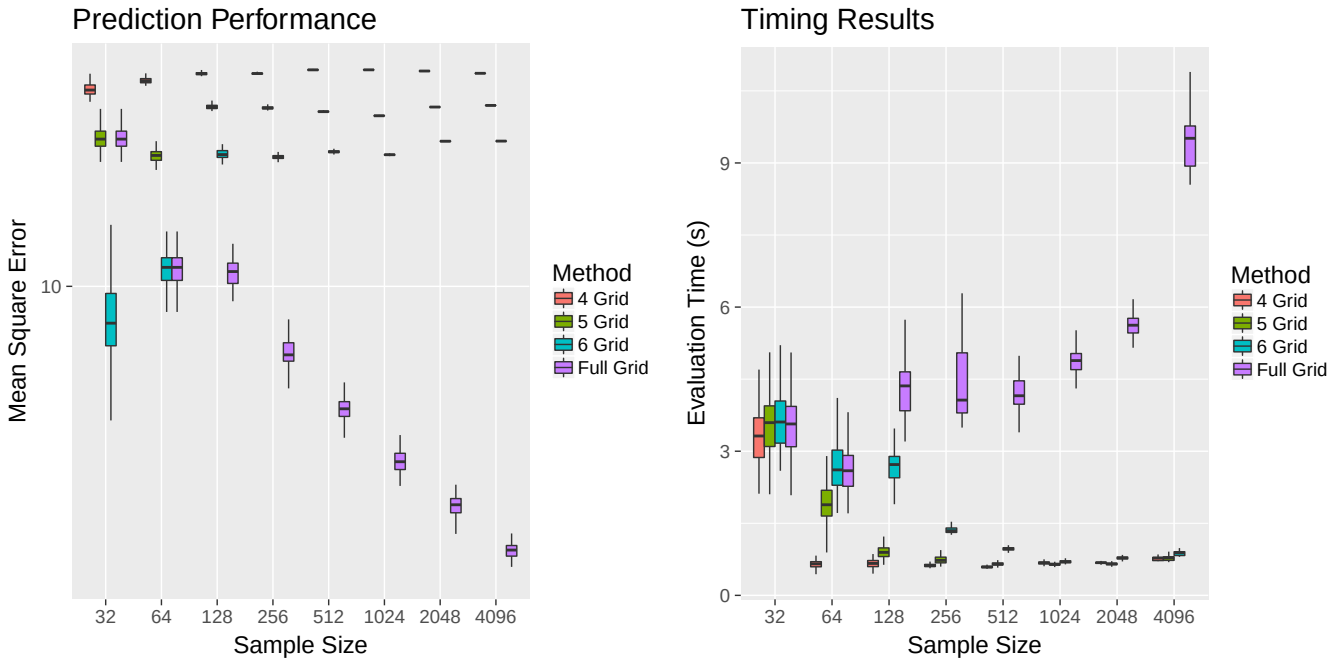


Figure 5.7: Results of simulation study of Section 5.4.3: Bumps function.

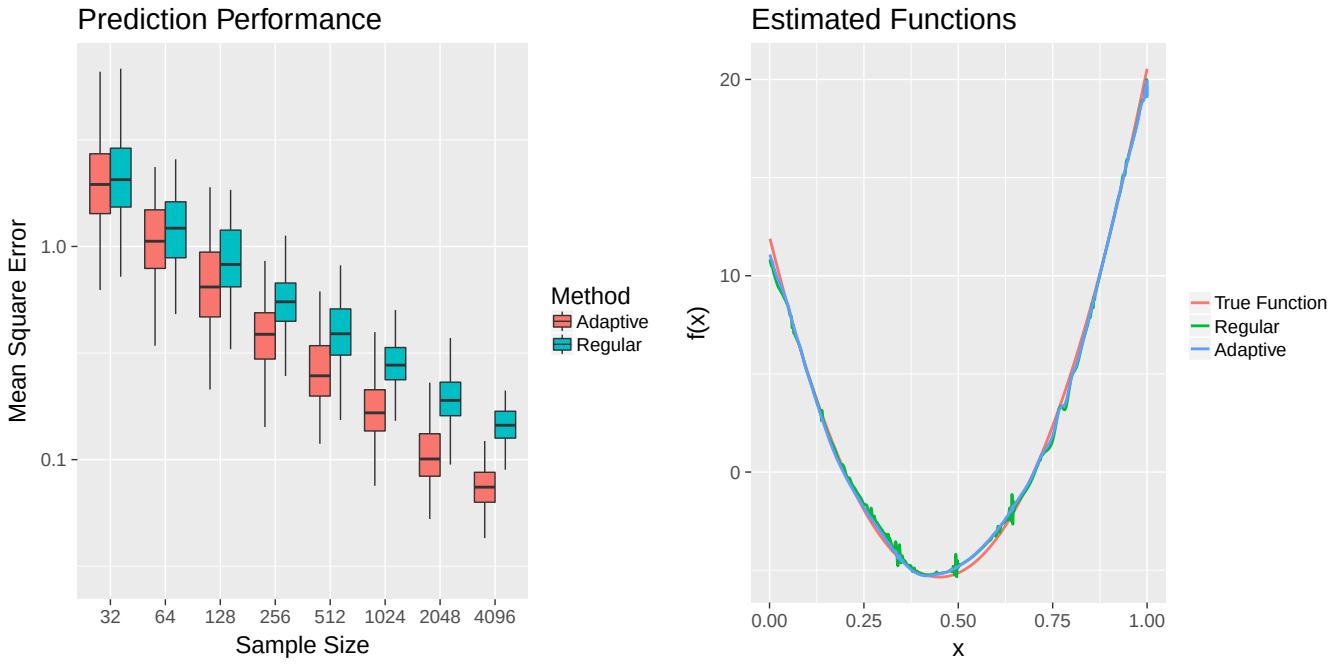


Figure 5.8: Results of simulation study of Section 5.4.4: Polynomial function.

5.4.4 Simulation study for adaptive `waveMesh`

In this subsection, we present some simulation results regarding the adaptive `waveMesh` estimator (5.9). In the left panel of Figure 5.8 to 5.13 we present the MSE as a function of sample size for regular `waveMesh` with $K = n$ and adaptive `waveMesh`. We present the minimum MSE over a sequence of 50 λ values. We see that our adaptive estimator uniformly outperforms the regular estimator in terms of prediction error. The results indicates that if we have a good procedure for selecting the tuning parameter, i.e., if we pick close to the theoretically ideal tuning parameter then adaptive `waveMesh` will have a lower MSE.

5.5 Discussion

In this chapter, we introduced `waveMesh`, a novel method for non-parametric regression using wavelets. Unlike traditional methods, `waveMesh` does not require the independent variable

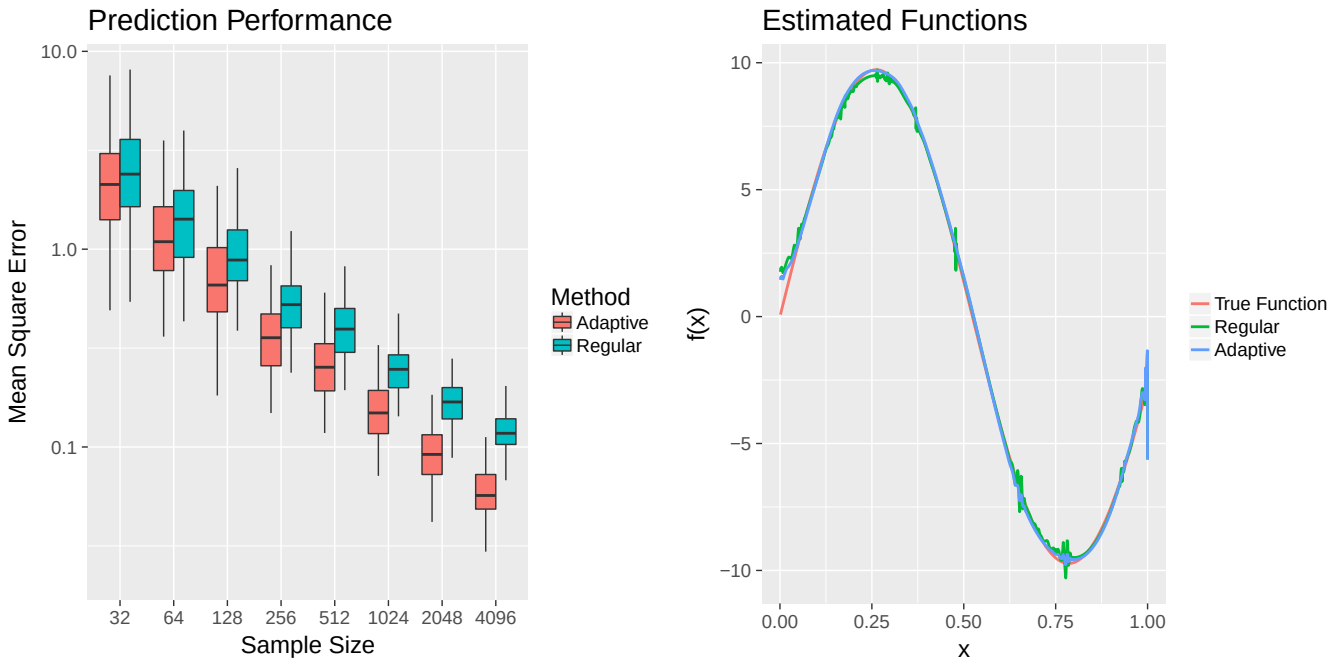


Figure 5.9: Results of simulation study of Section 5.4.4: Sine function.

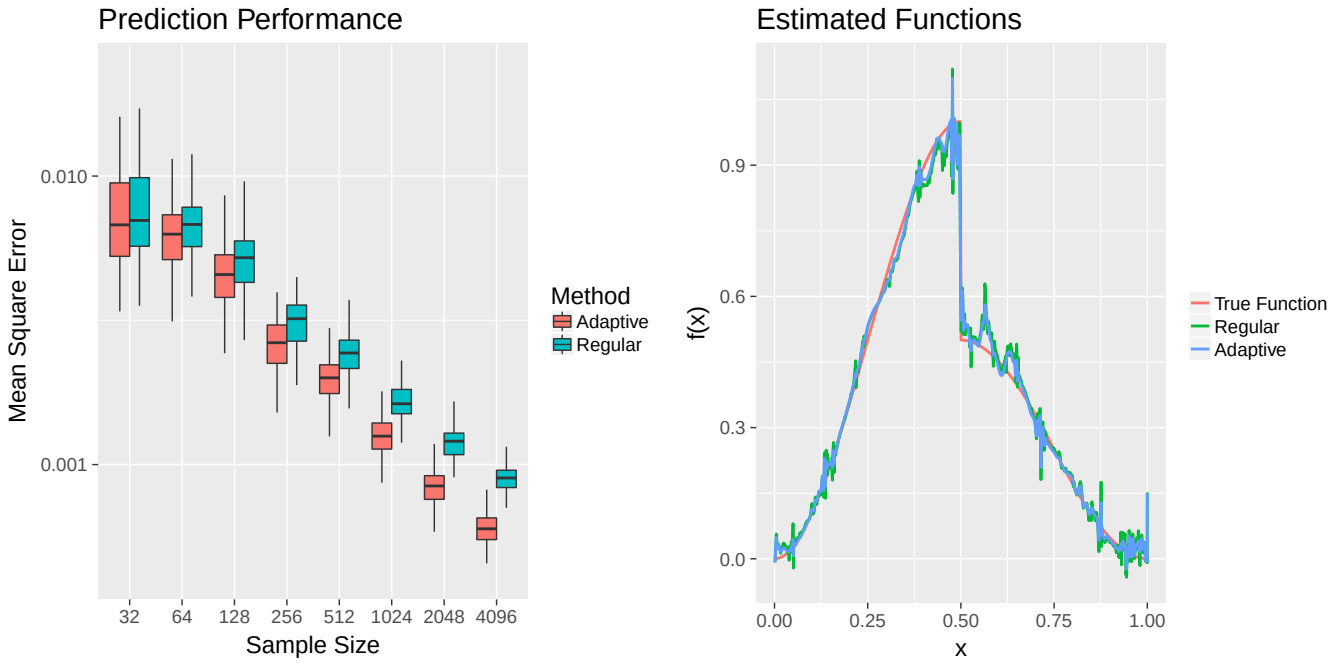


Figure 5.10: Results of simulation study of Section 5.4.4: Piecewise Polynomial function.

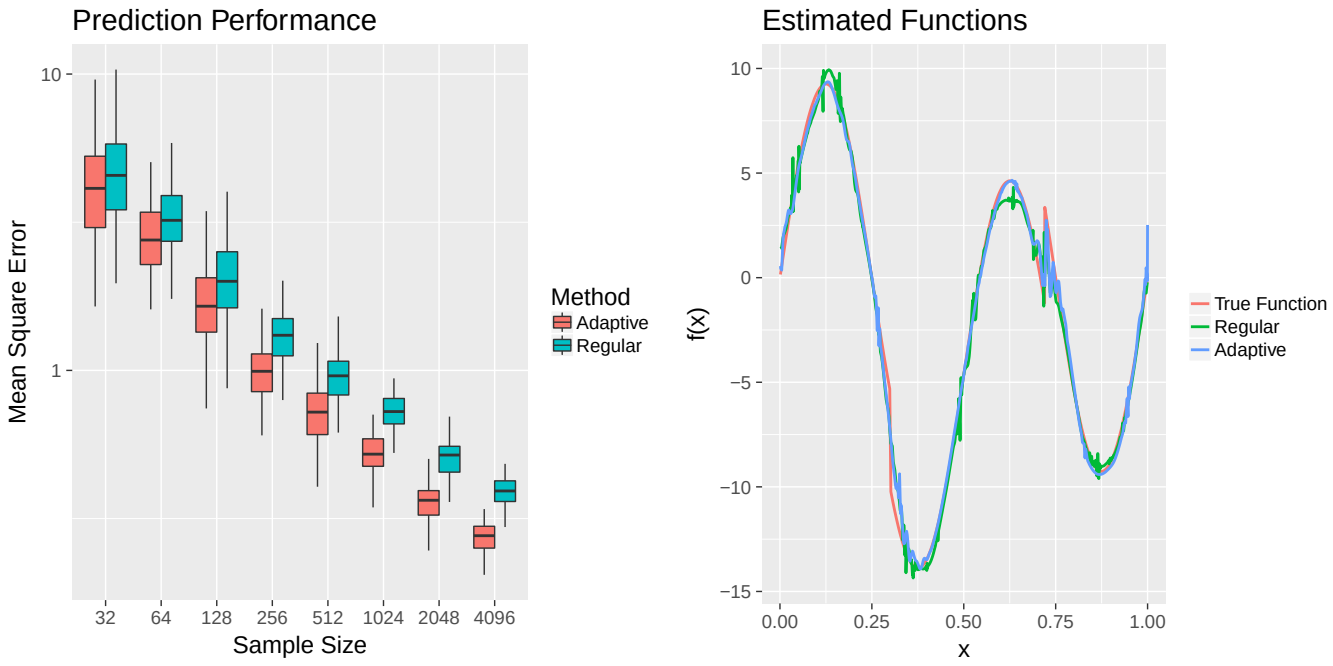


Figure 5.11: Results of simulation study of Section 5.4.4: Heavy sine function.

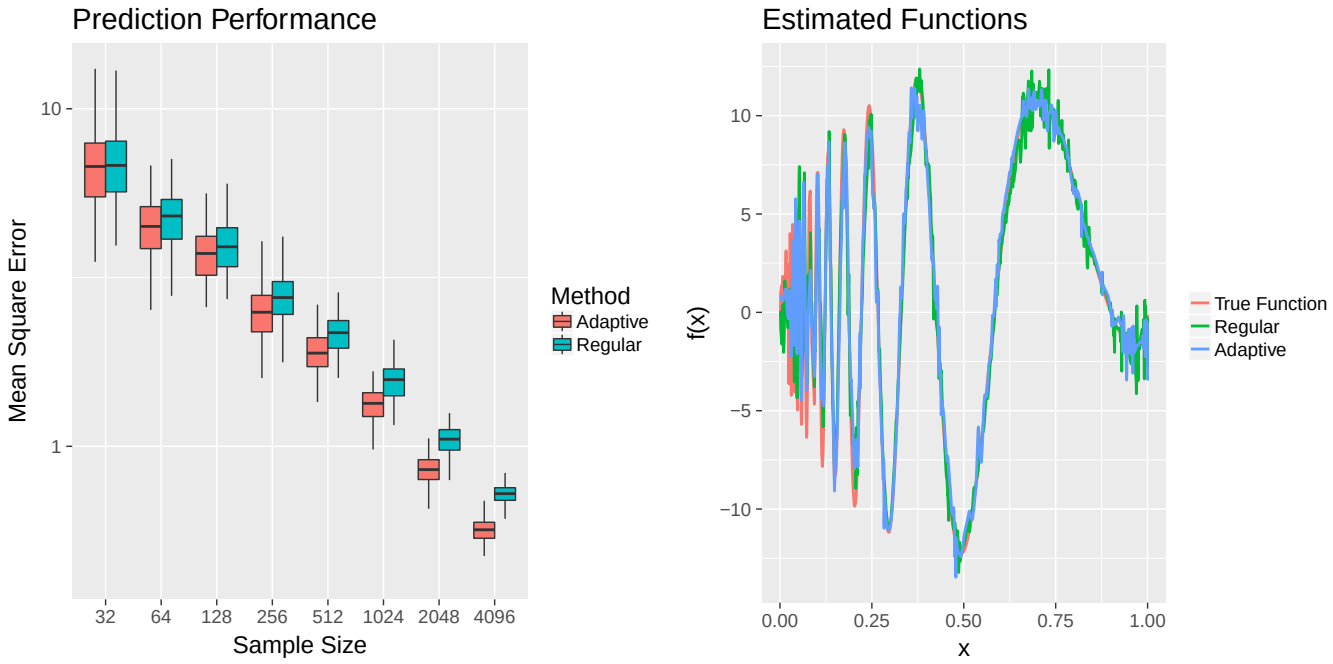


Figure 5.12: Results of simulation study of Section 5.4.4: Doppler function.

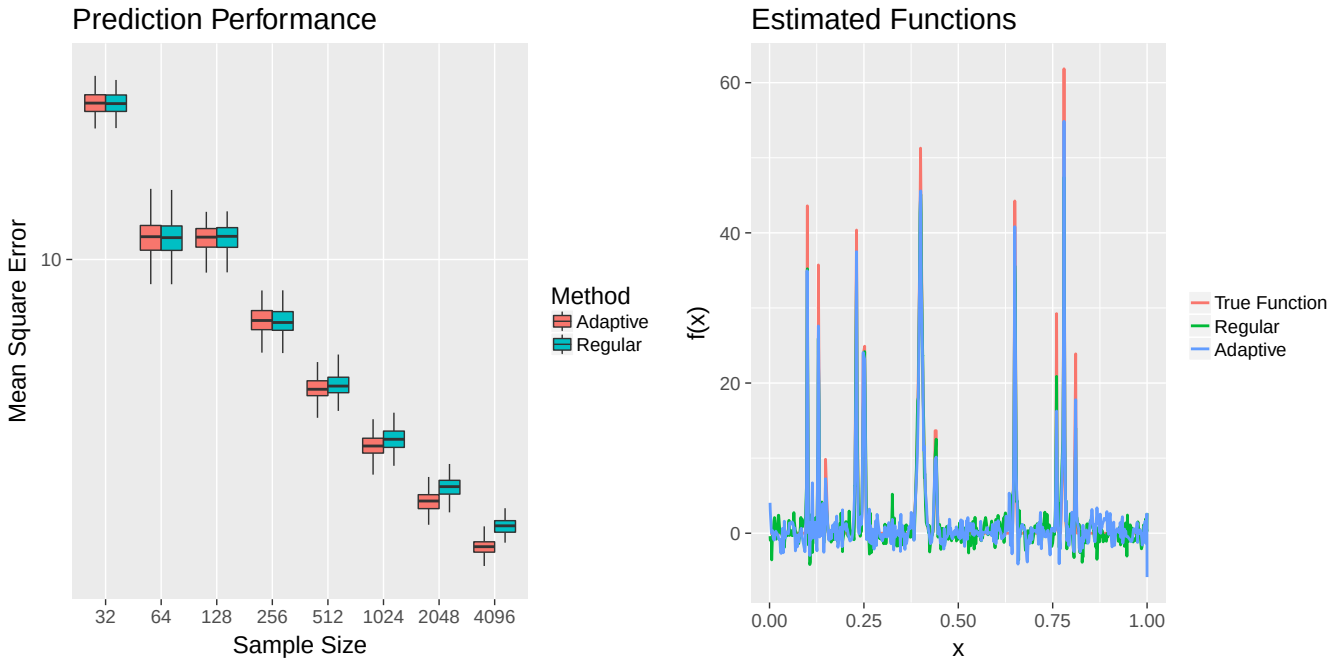


Figure 5.13: Results of simulation study of Section 5.4.4: Bumps function.

to be uniformly spaced on the unit interval nor does it require the sample size to be a power of 2. We achieve this using a novel interpolation approach for wavelets. The main appeal of our proposal is that it naturally extends to multivariate additive models for a potentially large number of covariates.

We proposed an efficient proximal gradient descent algorithm, which leverages exiting techniques for fast computation of the DWT. We established minimax convergence rates for our univariate proposal over a large class of Besov spaces. For a particular Besov space, we also establish minimax convergence rates for our (sparse) additive framework.

Chapter 6

DISCUSSION

6.1 Summary

In this dissertation we presented a number of techniques for moving beyond linearity when modelling data in high dimensions. We presented two general frameworks for fitting (1) interaction models with strong heredity and (2) additive models, in high dimensions. The strength of our general frameworks is that they bridge gaps in the existing literature and facilitate the development of future estimators.

In Chapter 2, we proposed `FAMILY`, for fitting interaction models with strong heredity. `FAMILY` leveraged the sparsity-inducing properties of group penalties and resulting sparsity pattern by using overlapping groups. Our proposal is a generalization of several existing methods, such as `VANISH`, `hierNet`, the all-pairs lasso, and the lasso using only main effects. `FAMILY` can be formulated as the solution to a convex optimization problem, which we solved using an efficient alternating directions method of multipliers (ADMM) algorithm. The algorithm has guaranteed convergence to the global optimum, can be easily specialized to any convex penalty function of interest, and allows for a straightforward extension to the setting of generalized linear models. We derived an unbiased estimator of the degrees of freedom of `FAMILY`, and explored its performance in a simulation study and on an HIV sequence dataset.

In Chapter 3, we presented a unified framework for estimation of generalized additive models in high dimensions. We discussed how our framework, `PGSAME`, defines a large class of penalized regression estimators, encompassing many existing methods. We presented an efficient computational algorithm for `PGSAME`s that easily scales to thousands of observations and features. We proved minimax optimal convergence bounds on these estimators under a

weak compatibility condition. In addition, we characterized the rate of convergence when this compatibility condition is not met. Finally, we also showed that the optimal penalty parameters for our structure and sparsity penalties are linked, allowing cross-validation to be conducted over only a single tuning parameter. We complemented our theoretical results with empirical studies comparing some existing methods.

In Chapter 4, we proposed a convex, penalized estimator for nonparametric regression and extended it to sparse additive models for a potentially large number of covariates. We proposed a hierarchical group lasso penalty which allowed automatic selection of truncation level of a basis expansion. Thus our proposal simultaneously achieves parsimony and adaptivity in a computationally efficient framework. We demonstrated these properties through empirical studies on both real and simulated datasets. We developed an efficient algorithm for our proposal that scales similarly to the lasso with the number of covariates and sample size. We showed that our estimator converges at the minimax rate for functions within a hierarchical class. We further established minimax rates for a large class of sparse additive models.

In Chapter 5, we presented `waveMesh`, a novel approach for nonparametric regression using wavelet basis functions. Our proposal can be applied to non-equispaced data with sample size not necessarily a power of 2. We developed an efficient proximal gradient descent algorithm for computing the estimator and established adaptive minimax convergence rates. The main appeal of `waveMesh` was demonstrated by the natural extensions to additive and sparse additive models for a potentially large number of covariates. We discussed how `waveMesh` can be extended to general, convex loss functions, and proposed *adaptive* `waveMesh`, a variation which led to improved prediction performance. We proved minimax optimal convergence rates under a weak compatibility condition for sparse additive models. We specialized our results to the low-dimensional case where the compatibility condition holds; in addition, we established convergence rates for when the condition is not met. Our theoretical results were complemented with an empirical study comparing `waveMesh` to existing methods.

6.2 Limitations of dissertation work

In this section, we highlight some issues and limitations of our work. Addressing these limitations is a promising direction for future research.

- *The choice of tuning parameters is an open problem.* Most of our theoretical results established convergence rates under the so-called *prediction optimal* tuning parameter. Prediction optimal tuning parameters were presented up to a constant. Accurate estimation of the optimal tuning parameters is an interesting challenge. This is a particular limitation in the case of our univariate `waveMesh` proposal, because many existing wavelet methods offer a theoretically motivated estimator for the tuning parameters.

- *Results regarding convergence rates ignored constants.* The convergence rates presented throughout this section did not optimize or explicitly specify constants. Many convergence rates for the estimation of f^0 were of the form

$$\|\widehat{f} - f^0\|_n^2 \leq constant \times rate,$$

  e.g., for nonparametric univariate regression we had $rate = n^{-\frac{2m}{2m+1}}$. While ignoring the constants is common in the nonparametric literature, it would be desirable to get some bounds for the *constant* term. One limitation of our work is that we cannot guarantee that the *constant* is not too large.

- *The compatibility condition for sparse additive models was not verified to hold.* Our theoretical results for (generalized) sparse additive models relied on the compatibility condition. A few authors in the literature prove the compatibility condition holds with high probability (Meier et al., 2009; van de Geer, 2010) while others impose conditions, such as independence of covariates, to prove a similar restricted eigenvalue condition or restricted strong convexity condition (Raskutti et al., 2012). In the spirit of our

general framework, it would be desirable to derive sufficient conditions under which compatibility holds with high probability.

6.3 Future research

The work presented in this dissertation can be extended in many ways. As discussed in the final section of Chapter 2, `FAMILY` can easily be extended to higher order interactions terms. An efficient implementation and empirical comparison of such a technique for high dimensional data is a promising avenue for future research. Furthermore, the recent work on studying the finite sample properties of general sparsity inducing penalties (van de Geer, 2016), can be leveraged to establish convergence rates for `FAMILY`. We conjecture that it can be theoretically shown that using our proposal can achieve better rates than the all-pairs lasso over the class of models that obey strong heredity.

The work on additive models from Chapters 3 to 5 can also be extended in various ways. One possible extension is adding interaction terms to the additive model, i.e., a model of the form:

$$f(\boldsymbol{x}) = f_1(x_1) + \cdots + f_p(x_p) + f_{1,2}(x_1, x_2) + \cdots + f_{p-1,p}(x_{p-1}, x_p). \tag{6.1}$$

This problem raises the same challenges as Chapter 2 regarding interaction terms and imposing strong or weak heredity constraints. The nonparametric nature of the model does not allow a simple extension of our framework, `FAMILY`. However, it also offers exciting new opportunities for building more flexible models by allowing the order of interactions to grow; enforcing sparsity can allow us to build such models in high-dimensions.

Another direction of future research is developing *smoothness adaptive* estimators for additive models. We believe that our wavelet proposal, `waveMesh`, is well suited for this task. We discuss this in some detail here. Firstly, recall that by *smoothness adaptive*, we mean estimators which can adapt to the smoothness level of the underlying function class.

Recall the univariate smoothing spline estimator

$$\widehat{\boldsymbol{d}} \leftarrow \underset{\boldsymbol{d}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{N}\boldsymbol{d}\|_n^2 + \lambda\|\boldsymbol{D}_m\boldsymbol{d}\|_n^2, \tag{6.2}$$

where $\boldsymbol{N}_{i,j} = \psi_j(x_i)$ and $[\boldsymbol{D}_m^\top \boldsymbol{D}_m]_{j,k} = \int \psi_j^{(m)}(t)\psi_k^{(m)}(t)\,dt$. For the estimator $\widehat{f} = \boldsymbol{N}\widehat{\boldsymbol{d}}$ we obtain convergence rates of the form:

$$\|\widehat{f} - f^0\|_n^2 \lesssim n^{-\frac{2m}{2m+1}},$$

if f^0 belongs to the m-th order Sobolev class. The smoothness order, m, needs to be specified correctly in (6.2) to obtain the correct rate; thus, smoothing splines are *not* smoothness adaptive. On the other hand the `waveMesh` estimator

$$\widehat{\boldsymbol{d}} \leftarrow \underset{\boldsymbol{d}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{R}\boldsymbol{W}^\top\boldsymbol{d}\|_n^2 + \lambda\|\boldsymbol{d}_{-1}\|_1^2, \tag{6.3}$$

achieves the same rate (6.3) without requiring any knowledge of the smoothness order of f^0. Thus we showed in Theorem 5.1 that `waveMesh` *is* smoothness adaptive. In fact, the smoothness adaptivity of wavelets is well known in the literature, see, e.g., Donoho and Johnstone (1995).

A more challenging problem, however, is building smoothness adaptive estimators for additive models. This is challenging because now different components can have different smoothness orders. Ideally we would like to achieve rates of the form

$$\|\sum_j (\widehat{f}_j - f_j^0)\|_n^2 \lesssim \sum_j n^{-\frac{2m_j}{2m_j+1}}. \tag{6.4}$$

Establishing such rates generally requires either correct specification of a smoothness penalty or simultaneous selection of multiple tuning parameters for each component function. We believe that rates like (6.4) can be achieved by a single tuning parameter of our additive `waveMesh` proposal. We present a sketch of the proof here. For convenience, we will ignore

the interpolation error term. Note that our (non-sparse) additive proposal can be written as

$$\widehat{\boldsymbol{d}} \leftarrow \underset{\boldsymbol{d}\in\mathbb{R}^{pK}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{d}\|_n^2 + \lambda\|\boldsymbol{d}_M\|_1, \tag{6.5}$$

where $\boldsymbol{X} = [\boldsymbol{R}_1\boldsymbol{W}^\top|\boldsymbol{R}_2\boldsymbol{W}^\top|\cdots|\boldsymbol{R}_{p-1}\boldsymbol{W}^\top|\boldsymbol{R}_p\boldsymbol{W}^\top] \in \mathbb{R}^{n\times pK}$ and $\boldsymbol{d}_M$ is the vector of all mother wavelet coefficients. Just as we proceeded in the univariate case, we note that the problem (6.5) is a lasso problem and (under suitable assumptions) we can establish rates of the form

$$\|\sum_j(\widehat{f}_j - f_j^0)\|_n^2 \lesssim \underbrace{s\frac{\log(pK)}{n}}_{(I)} + \underbrace{\sum_j \|f_j^* - f_j^0\|_n^2}_{(II)}, \tag{6.6}$$

where $s = \|\boldsymbol{d}^*\|_0$ for the oracle $f_j^* = \boldsymbol{R}_j\boldsymbol{W}^\top\boldsymbol{d}_j^*$. In the univariate case, we selected s such that it balanced the terms (I) and (II) and achieved the minimax rate. We can do this again by defining $s_j = \|\boldsymbol{d}_j^*\|_0$ and noting that

$$\begin{aligned}\|\sum_j(\widehat{f}_j - f_j^0)\|_n^2 &\lesssim s\frac{\log(pK)}{n} + \sum_j \|f_j^* - f_j^0\|_n^2 \\ &= \sum_j \Big(s_j\frac{\log(pK)}{n} + \|f_j^* - f_j^0\|_n^2\Big).\end{aligned}$$

As in the univariate case, we believe the optimal s_j (obtained by selecting an appropriate f_j^*) will balance the two terms for each component leading to the rate (6.4) up to a $\log(pK)$ factor. The additional log term which appears here and also in our univariate proposal is unfortunate. However, we believe this term can be eliminated by appropriately weighting the wavelet penalty or using *adaptive* rates as done in the *adaptive lasso* (Zou, 2006). The main challenge of the above approach will be careful treatment of the various details, such as interpolation error, validation of the compatibility condition, etc.

Finally, moving on to the goal of sparse additive `waveMesh`, we consider a slight variation

of our proposal

$$\widehat{\boldsymbol{d}}_1, \ldots, \widehat{\boldsymbol{d}}_p \leftarrow \underset{\boldsymbol{d}_j \in \mathbb{R}^K}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}_j\|_n^2 + \sum_j \lambda_1 \|\boldsymbol{d}_{j,-1}\|_1 + \lambda_2 \|\boldsymbol{d}_j\|_2. \tag{6.7}$$

The problem (6.7) is simply the *sparse group lasso* problem of Simon et al. (2013). We may be able to establish smoothness adaptive rates if we can establish lasso-like rates for the sparse group lasso. We believe this can be achieved using the recently proposed framework of general sparsity inducing penalties (van de Geer, 2016). In fact, a generalization of this problem to establishing rates of the hierarchical group lasso is of independent interest. We conjecture that this might allow us to prove smoothness adaptability of our proposal in Chapter 4.

BIBLIOGRAPHY

Amato, U. and A. Antoniadis

2001. Adaptive wavelet series estimation in separable nonparametric regression models. *Statistics and Computing*, 11(4):373–394.

Antoniadis, A.

1997. Wavelets in statistics: a review. *Journal of the Italian Statistical Society*, 6(2):97.

Antoniadis, A. and J. Fan

2001. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–967.

Arnold, T., V. Sadhanala, and R. Tibshirani

2014. *glmgen: Fast algorithms for generalized lasso problems*. R package version 0.0.3.

Arnold, T. B. and R. J. Tibshirani

2014. *genlasso: Path algorithm for generalized lasso problems*. R package version 1.3.

Ayer, M., H. D. Brunk, G. M. Ewing, W. Reid, E. Silverman, et al.

1955. An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, 26(4):641–647.

Bach, F., R. Jenatton, J. Mairal, and G. Obozinski

2011. Convex optimization with sparsity-inducing norms. In *Optimization for Machine Learning*, S. Sra, S. Nowozin, and S. J. Wright, eds., chapter 2, Pp. 19–53. Cambridge, MA, USA: MIT Press.

Bach, F., R. Jenatton, J. Mairal, and G. Obozinski

2012. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468.

Beck, A. and M. Teboulle

2009a. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202.

Beck, A. and M. Teboulle

2009b. Gradient-based algorithms with applications to signal recovery. *Convex optimization in signal processing and communications*, Pp. 42–88.

Bellman, R. E.

1961. *Adaptive control processes: a guided tour*. Princeton University Press.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov

2010. Hierarchical selection of variables in sparse high-dimensional regression. In *Borrowing strength: theory powering applications–a Festschrift for Lawrence D. Brown*, J. O. Berger, T. T. Cai, and I. M. Johnstone, eds., Pp. 56–69. Beachwood, Ohio, USA: Institute of Mathematical Statistics.

Bien, J., J. Taylor, and R. Tibshirani

2013. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141.

Bien, J. and R. Tibshirani

2014. *hierNet: A Lasso for Hierarchical Interactions*. R package version 1.6.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein

2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Bühlmann, P. and S. van de Geer

2011. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Burczynski, M. E., R. L. Peterson, N. C. Twine, K. A. Zuberek, B. J. Brodeur, L. Casciotti, V. Maganti, P. S. Reddy, A. Strahs, F. Immermann, W. Spinelli, U. Schwertschlag, A. M.

Slager, M. M. Cotreau, and A. J. Dorner

2006. Molecular classification of crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *The Journal of Molecular Diagnostics*, 8(1):51–61.

Cai, T. and L. D. Brown

1998. Wavelet shrinkage for nonequispaced samples. *The Annals of Statistics*, 26(5):1783–1799.

Cai, T. T. and L. D. Brown

1999. Wavelet estimation for samples with random uniform design. *Statistics & probability letters*, 42(3):313–321.

Candes, E. J., M. B. Wakin, and S. P. Boyd

2008. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, 14(5):877–905.

Chipman, H.

1996. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36.

Choi, N. H., W. Li, and J. Zhu

2010. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364.

Chouldechova, A. and T. Hastie

2015. Generalized additive model selection. *arXiv preprint arXiv:1506.03850*.

Chui, C. K.

1992. An introduction to wavelets. *Philadelphia, SIAM*, 38.

Combettes, P. L. and J. Pesquet

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, and H. Wolkowicz, eds., Pp. 185–212. Springer New York.

Cox, D. R.

1984. Interaction. *International Statistical Review/Revue Internationale de Statistique*, 52:1–24.

Dalalyan, A. S., M. Hebiri, and J. Lederer

2014. On the prediction performance of the lasso. *Preprint. Available at arXiv:1402.1700*.

Daubechies, I.

1988. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 41(7):909–996.

Daubechies, I.

1990. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, 36(5):961–1005.

Daubechies, I.

1992. *Ten lectures on wavelets*, volume 61. Siam.

Daubechies, I.

1993. Orthonormal bases of compactly supported wavelets ii. variations on a theme. *SIAM Journal on Mathematical Analysis*, 24(2):499–519.

Donoho, D. L.

1995. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3):613–627.

Donoho, D. L. and I. M. Johnstone

1995. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, 90(432):1200–1224.

Duchi, J. and Y. Singer

2009. Efficient online and batch learning using forward backward splitting. *The Journal of Machine Learning Research*, 10:2899–2934.

Dumer, I.

2006. Covering an ellipsoid with equal balls. *Journal of Combinatorial Theory, Series A*, 113(8):1667–1676.

Efron, B.

1986. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470.

Fan, J., Y. Feng, and R. Song

2012. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*.

Fan, J. and R. Li

2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani

2007. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., T. Hastie, and R. Tibshirani

2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Friedman, J. H.

1991. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

George, E. I. and R. E. McCulloch

1993. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Grez, G. A. S. and B. Vidakovic

2018. Empirical wavelet-based estimation for non-linear additive regression models. *arXiv preprint arXiv:1803.04558*.

Hall, P., B. A. Turlach, et al.

1997. Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *The Annals of Statistics*, 25(5):1912–1925.

Hamada, M. and C. J. Wu

1992. Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24(3):130–137.

Hao, N. and H. H. Zhang

2014. Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301.

Hastie, T., R. Tibshirani, and J. Friedman

2009. *The Elements of Statistical Learning*. Springer New York.

Hastie, T. J. and R. J. Tibshirani

1990. *Generalized additive models*, volume 43. CRC Press.

Jacob, L., G. Obozinski, and J.-P. Vert

2009. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, Pp. 433–440. ACM.

Jenatton, R., J.-Y. Audibert, and F. Bach

2011. Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, 12:2777–2824.

Jenatton, R., J. Mairal, F. R. Bach, and G. R. Obozinski

2010. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Pp. 487–494.

Johnson, N. A.

2013. A dynamic programming algorithm for the fused lasso and l 0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260.

Joseph, V. R.

2006. A Bayesian approach to the design and analysis of fractionated experiments. *Technometrics*, 48(2):219–229.

Kim, S., K. Koh, S. Boyd, and D. Gorinevsky

2009. ℓ_1 trend filtering. *SIAM review*, 51(2):339–360.

Koltchinskii, V. and M. Yuan

2010. Sparsity in multiple kernel learning. *Ann. Statist.*, 38(6):3660–3695.

Kovac, A. and B. W. Silverman

2000. Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of the American Statistical Association*, 95(449):172–183.

Lim, M. and T. Hastie

2013. Learning interactions through hierarchical group-lasso regularization. *Preprint. Available at arXiv:1308.2719*.

Lin, Y. and H. H. Zhang

2006. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297.

Lou, Y., J. Bien, R. Caruana, and J. Gehrke

2016. Sparse partially linear additive models. *Journal of Computational and Graphical Statistics*, 25(4):1126–1140.

Mallat, S. G.

1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693.

Mammen, E. and S. van de Geer

1997. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413.

McCullagh, P.

1984. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292.

Meier, L., S. van de Geer, and P. Bühlmann

2009. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821.

Meinshausen, N.

2007. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393.

Meyer, Y.

1985. Principe d'incertitude, bases hilbertiennes et algebres d'operateurs. *Séminaire Bourbaki*, 662:1985–1986.

Montgomery, D. C., E. A. Peck, and G. G. Vining

2012. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons.

Nadaraya, E. A.

1964. On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142.

Nason, G.

2010. *Wavelet methods in statistics with R*. Springer Science & Business Media.

Nason, G.

2016. *wavethresh: Wavelets Statistics and Transforms*. R package version 4.6.8.

Negahban, S., P. Ravikumar, M. J. Wainwright, and B. Yu

2011. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Manuscript, University of California, Berkeley, Dept. of Statistics and EECS*.

Nesterov, Y.

2007. Gradient methods for minimizing composite objective function. Technical report, UCL.

Nickl, R. and B. M. Pötscher

2007. Bracketing metric entropy rates and empirical central limit theorems for function classes of besov-and sobolev-type. *Journal of Theoretical Probability*, 20(2):177–199.

Nunes, M. and M. Knight

2017. *adlift: An Adaptive Lifting Scheme Algorithm*. R package version 1.3-3.

Nunes, M. A., M. I. Knight, and G. P. Nason

2006. Adaptive lifting for nonparametric regression. *Statistics and Computing*, 16(2):143–159.

Ogden, T.

2012. *Essential wavelets for statistical applications and data analysis*. Springer Science & Business Media.

Parikh, N. and S. Boyd

2014. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239.

Park, M. Y. and T. Hastie

2008. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50.

Peixoto, J. L.

1987. Hierarchical variable selection in polynomial regression models. *The American Statistician*, 41(4):311–313.

Pensky, M. and B. Vidakovic

2001. On non-equally spaced wavelet regression. *Annals of the Institute of Statistical Mathematics*, 53(4):681–690.

Percival, D. B. and A. T. Walden

2006. *Wavelet methods for time series analysis*, volume 4. Cambridge university press.

Petersen, A., D. Witten, and N. Simon

2016. Fused lasso additive model. *Journal of Computational and Graphical Statistics*, 25(4):1005–1025.

R Core Team

2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radchenko, P. and G. M. James

2010. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105(492):1541–1553.

Ramdas, A. and R. J. Tibshirani

2015. Fast and flexible admm algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, 25(3):839–858.

Raskutti, G., M. J. Wainwright, and B. Yu

2012. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *The Journal of Machine Learning Research*, 13(1):389–427.

Raskutti, G., B. Yu, and M. J. Wainwright

2009. Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness. In *Advances in Neural Information Processing Systems*, Pp. 1563–1570.

Rauhut, H. and R. Ward

2016. Interpolation via weighted ℓ_1 minimization. *Applied and Computational Harmonic Analysis*, 40(2):321 – 351.

Ravikumar, P., J. Lafferty, H. Liu, and L. Wasserman

2009. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030.

Rhee, S.-Y., J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer

2006. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360.

Sadhanala, V. and R. J. Tibshirani

2017. Additive Models with Trend Filtering. *ArXiv e-prints*.

Sardy, S., D. B. Percival, A. G. Bruce, H.-Y. Gao, and W. Stuetzle

1999. Wavelet shrinkage for unequally spaced data. *Statistics and Computing*, 9(1):65–75.

Sardy, S. and P. Tseng

2004. Amlet, ramlet, and gamlet: automatic nonlinear fitting of additive models, robust and generalized, with wavelets. *Journal of Computational and Graphical Statistics*, 13(2):283–309.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani

2013. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Stein, C. M.

1981. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Stone, C. J.

1977. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620.

Strang, G. and T. Nguyen

1996. *Wavelets and filter banks*. SIAM.

Tibshirani, R.

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

Tibshirani, R. J.

2013. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490.

Tibshirani, R. J.

2014. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323.

Tibshirani, R. J., J. Taylor, et al.

2012. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232.

Tsanas, A., M. A. Little, P. E. McSharry, and L. O. Ramig

2010. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE transactions on Biomedical Engineering*, 57(4):884–893.

van de Geer, S.

2000. *Empirical Processes in M-estimation*, volume 6. Cambridge University Press.

van de Geer, S.

2008. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, Pp. 614–645.

van de Geer, S.

2010. The lasso with within group structure. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*, Pp. 235–244. Institute of Mathematical Statistics.

van de Geer, S.

2016. *Estimation and Testing Under Sparsity: École d'Été de Probabilités De Saint-Flour XLV - 2015*, 1st edition. Springer Publishing Company, Incorporated.

van de Geer, S. and P. Bühlmann

2009. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Čencov, N.

1962. Evaluation of an unknown distribution density from observations. *Doklady*, 3:1559–1562.

Vidakovic, B.

2009. *Statistical modeling by wavelets*, volume 503. John Wiley & Sons.

Wahba, G.

1990. *Spline models for observational data*. SIAM.

Watson, G. S.

1964. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372.

Wu, J., B. Devlin, S. Ringquist, M. Trucco, and K. Roeder

2010. Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic epidemiology*, 34(3):275–285.

Yang, Y. and A. Barron

1999. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599.

Yates, F.

1978. *The design and analysis of factorial experiments*. Imperial Bureau of Soil Science.

Yin, J., X. Chen, and E. Xing

2012. Group sparse additive models. *arXiv preprint arXiv:1206.4673*.

Yuan, M. and Y. Lin

2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Yuan, M. and D.-X. Zhou

2015. Minimax optimal rates of estimation in high dimensional additive models: Universal phase transition. *arXiv preprint arXiv:1503.02817*.

Zhan, X.

2005. Extremal eigenvalues of real symmetric matrices with entries in an interval. *SIAM journal on matrix analysis and applications*, 27(3):851–860.

Zhang, S., M.-Y. Wong, et al.

2003. Wavelet threshold estimation for additive regression models. *The Annals of Statistics*, 31(1):152–173.

Zhao, P., G. Rocha, and B. Yu

2009. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497.

Zhao, T., X. Li, H. Liu, and K. Roeder

2014. *SAM: Sparse Additive Modelling*. R package version 1.0.5.

Zou, H.

2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and T. Hastie

2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Appendix A

TECHNICAL DETAILS FOR "CONVEX MODELING OF INTERACTIONS WITH STRONG HEREDITY"

A.1 Alternating directions method of multipliers

A.1.1 Overview of ADMM

We will solve (2.6) using the *alternating directions method of multipliers* (ADMM) algorithm, which we briefly review here. We refer the reader to Boyd et al. (2011) for a detailed discussion.

ADMM provides a simple, general, and efficient approach for solving a problem of the form

$$\underset{\boldsymbol{x}}{\text{minimize}}\ f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}), \tag{A.1}$$

where f_1 and f_2 are convex, closed and proper. The key insight behind ADMM is that (A.1) can be re-written as

$$\underset{\boldsymbol{x},\boldsymbol{y}}{\text{minimize}}\ \{f_1(\boldsymbol{x}) + f_2(\boldsymbol{y})\} \text{ subject to } \boldsymbol{x} = \boldsymbol{y}. \tag{A.2}$$

The augmented Lagrangian corresponding to (A.2) takes the form

$$L_\rho(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\gamma}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{y}) + \boldsymbol{\gamma}^\top(\boldsymbol{x} - \boldsymbol{y}) + (\rho/2)\|\boldsymbol{x} - \boldsymbol{y}\|_2^2,$$

where γ is a dual variable and $\rho \in \mathbb{R}$ is a positive constant. The resulting ADMM algorithm

involves iterating the following steps until convergence,

$$\begin{aligned}\boldsymbol{x}^{k+1} &= \underset{\boldsymbol{x}}{\operatorname{argmin}}\ L_\rho(\boldsymbol{x}, \boldsymbol{y}^k, \boldsymbol{\gamma}^k)\\ \boldsymbol{y}^{k+1} &= \underset{\boldsymbol{y}}{\operatorname{argmin}}\ L_\rho(\boldsymbol{x}^{k+1}, \boldsymbol{y}, \boldsymbol{\gamma}^k)\\ \boldsymbol{\gamma}^{k+1} &= \boldsymbol{\gamma}^k + \rho(\boldsymbol{x}^{k+1} - \boldsymbol{y}^{k+1})\ ,\end{aligned}$$

where k indexes the iterations. Under a few simple conditions, the ADMM algorithm converges to the global optimum (Boyd et al., 2011).

A.1.2 `FAMILY` *with squared error loss*

A.1.2.1 The ADMM algorithm

The augmented Lagrangian corresponding to (2.6) was given in (2.16). The complete ADMM algorithm is as follows:

1. Initialize ρ^0, $\boldsymbol{B}^0$, $\boldsymbol{\Theta}^0$ and $\boldsymbol{\Gamma}^0$.

2. Choose $\varepsilon^{pri} > 0$, $\varepsilon^{dual} > 0$.

3. Repeat for $i = 1, 2, 3, ...$ until $r^i < \varepsilon^{pri}$ and $s^i < \varepsilon^{dual}$, where r^i and s^i are the primal and dual residuals, respectively, defined as

$$\begin{aligned}s^i &= \rho^i \|(\boldsymbol{D}^i|\boldsymbol{E}^i|\boldsymbol{F}^i) - (\boldsymbol{D}^{i-1}|\boldsymbol{E}^{i-1}|\boldsymbol{F}^{i-1})\|_F\\ r^i &= \|(\boldsymbol{B}^i|\boldsymbol{B}^i|\boldsymbol{B}^i) - (\boldsymbol{D}^i|\boldsymbol{E}^i|\boldsymbol{F}^i)\|_F.\end{aligned}$$

   (a) Update ρ^i as described in Boyd et al. (2011):

$$\rho^i = \begin{cases} 2\rho^{i-1} & \text{if } r^{i-1} > 10s^{i-1} \\ \rho^{i-1}/2 & \text{if } 10r^{i-1} < s^{i-1} \\ \rho^{i-1} & \text{otherwise} \end{cases}.$$

(b) Update $\boldsymbol{B}^i$ as the solution to the least squares problem:

$$\begin{aligned}\boldsymbol{B}^i = \underset{\boldsymbol{B}}{\operatorname{argmin}} \ & \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{W} * \boldsymbol{B}\|_n^2 \\ & + \frac{3\rho^i}{2}\left\|\frac{1}{3\rho^i}\left[\rho^i(\boldsymbol{D}^{i-1} + \boldsymbol{E}^{i-1} + \boldsymbol{F}^{i-1}) - (\boldsymbol{\Gamma}_1^{i-1} + \boldsymbol{\Gamma}_2^{i-1} + \boldsymbol{\Gamma}_3^{i-1})\right] - \boldsymbol{B}\right\|_F^2.\end{aligned}$$

(c) Update $\boldsymbol{D}^i$ and $\boldsymbol{E}^i$ using the proximal operators discussed in Section 2.2.3.2:

$$\begin{aligned}\boldsymbol{D}^i &= \underset{\boldsymbol{D}}{\operatorname{argmin}} \ \tfrac{\rho^i}{2}\left\|\boldsymbol{D} - \left(\boldsymbol{B}^i + \tfrac{\boldsymbol{\Gamma}_1^{i-1}}{\rho^i}\right)\right\|_F^2 + \lambda_1 \textstyle\sum_{j=1}^{p_1} P_r(\boldsymbol{D}_{j,.}), \\ \boldsymbol{E}^i &= \underset{\boldsymbol{E}}{\operatorname{argmin}} \ \tfrac{\rho^i}{2}\left\|\boldsymbol{E} - \left(\boldsymbol{B}^i + \tfrac{\boldsymbol{\Gamma}_2^{i-1}}{\rho^i}\right)\right\|_F^2 + \lambda_2 \textstyle\sum_{j=1}^{p_2} P_c(\boldsymbol{E}_{.,k}).\end{aligned}$$

(d) Update $\boldsymbol{F}^i$ as follows:

$$\begin{aligned}\boldsymbol{F}^i_{0,.} &= \boldsymbol{B}^i_{0,.} + \frac{\boldsymbol{\Gamma}_{3}{}^{i-1}_{0,.}}{\rho^i}, \\ \boldsymbol{F}^i_{.,0} &= \boldsymbol{B}^i_{.,0} + \frac{\boldsymbol{\Gamma}_{3}{}^{i-1}_{.,0}}{\rho^i}, \\ \boldsymbol{F}^i_{j,k} &= \operatorname{sign}\left(\boldsymbol{B}^i_{j,k} + \frac{\boldsymbol{\Gamma}_{3}{}^{i-1}_{j,k}}{\rho^i}\right)\left(\left|\boldsymbol{B}^i_{j,k} + \frac{\boldsymbol{\Gamma}_{3}{}^{i}_{j,k}}{\rho^i}\right| - \frac{\lambda_3}{\rho^i}\right)_+ \quad \text{for } j \neq 0, k \neq 0.\end{aligned}$$

(e) Update $\boldsymbol{\Gamma}^i$ as follows:

$$\begin{aligned}\boldsymbol{\Gamma}_1^i &= \boldsymbol{\Gamma}_1{}^{i-1} + \rho^i\left(\boldsymbol{B}^i - \boldsymbol{D}^i\right), \\ \boldsymbol{\Gamma}_2^i &= \boldsymbol{\Gamma}_2{}^{i-1} + \rho^i\left(\boldsymbol{B}^i - \boldsymbol{E}^i\right), \\ \boldsymbol{\Gamma}_3^i &= \boldsymbol{\Gamma}_3{}^{i-1} + \rho^i\left(\boldsymbol{B}^i - \boldsymbol{F}^i\right).\end{aligned}$$

A.1.2.2 Update for $\boldsymbol{B}$ in step 3(b)

The update for $\boldsymbol{B}$ in Step 3(b) is a least squares problem with a $n \times (p_1+1)(p_2+1)$ design matrix. Here we show that clever matrix algebra can be applied in order to avoid solving

this least squares problem in each iteration. For convenience, we omit the superscripts in Step 3(b).

Let $\widetilde{\boldsymbol{B}}, \widetilde{\boldsymbol{D}}, \widetilde{\boldsymbol{E}}, \widetilde{\boldsymbol{F}}, \widetilde{\boldsymbol{\Gamma}}_1, \widetilde{\boldsymbol{\Gamma}}_2$, and $\widetilde{\boldsymbol{\Gamma}}_3$ denote the vectorized versions of $\boldsymbol{B}, \boldsymbol{D}, \boldsymbol{E}, \boldsymbol{F}, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2$, and $\boldsymbol{\Gamma}_3$. And let $\widetilde{\boldsymbol{W}}$ denote the $n \times (p_1+1)(p_2+1)$-dimensional matrix version of $\boldsymbol{W}$. Then the objective of Step 3(b) can be rewritten as

$$\frac{1}{2}\left\| \begin{bmatrix} \frac{1}{\sqrt{n}}\boldsymbol{y} \\ \frac{\rho(\widetilde{\boldsymbol{D}}+\widetilde{\boldsymbol{E}}+\widetilde{\boldsymbol{F}})-(\widetilde{\boldsymbol{\Gamma}}_1+\widetilde{\boldsymbol{\Gamma}}_2+\widetilde{\boldsymbol{\Gamma}}_3)}{\sqrt{3\rho}} \end{bmatrix} - \begin{bmatrix} \frac{1}{\sqrt{n}}\widetilde{\boldsymbol{W}} \\ \sqrt{3\rho}\boldsymbol{I}_{(1+p_1)(1+p_2)} \end{bmatrix} \widetilde{\boldsymbol{B}} \right\|_F^2. \tag{A.3}$$

Therefore, before performing the ADMM algorithm described in Section A.1.2, we compute the SVD of $\widetilde{\boldsymbol{W}}$. Then for each iteration of Step 3(b), the Woodbury matrix identity can be very quickly applied in order to minimize (A.3).

A.1.3 `FAMILY` *for generalized linear models*

We now consider the extension of `FAMILY` to GLMs (Section 2.2.4). The resulting ADMM algorithm is as in Section A.1.2, except that the update for B in Step 3(b) now takes the form

$$\underset{\boldsymbol{B}\in\mathbb{R}^{(p_1+1)\times(p_2+1)}}{\operatorname{argmin}} \frac{1}{n} l(\boldsymbol{W} * \boldsymbol{B}) + \frac{3\rho^i}{2}\left\| \frac{1}{3\rho^i}\left[\rho^i(\boldsymbol{D}^{i-1}+\boldsymbol{E}^{i-1}+\boldsymbol{F}^{i-1}) - (\boldsymbol{\Gamma}_1^{i-1}+\boldsymbol{\Gamma}_2^{i-1}+\boldsymbol{\Gamma}_3^{i-1})\right] - \boldsymbol{B}\right\|_F^2. \tag{A.4}$$

To solve this problem, we perform a second-order Taylor expansion of (A.4), in which we approximate the Hessian using a multiple of the identity (e.g., for logistic regression, we use the upper bound of $(1/4)\boldsymbol{I}$). Details are omitted in the interest of brevity.

A.2 Proofs of results in Section 2.2

Proof of Lemma 2.2. The result follows from the definition of the dual norm.

$$\begin{aligned}
P_*(\boldsymbol{z}) &= \sup\{\boldsymbol{z}^\top\boldsymbol{\beta} : P(\boldsymbol{\beta}) \leq 1\} \\
&= \sup\{\boldsymbol{z}^\top\boldsymbol{\beta} : \max(|\beta_1|, \|\boldsymbol{\beta}_{-1}\|_1) \leq 1\} \\
&= \sup\{\boldsymbol{z}^\top\boldsymbol{\beta} : |\beta_1| \leq 1 \text{ and } \|\boldsymbol{\beta}_{-1}\|_1 \leq 1\} \\
&= \sup\{z_1\beta_1 + \boldsymbol{z}_{-1}^\top\boldsymbol{\beta}_{-1} : |\beta_1| \leq 1 \text{ and } \|\boldsymbol{\beta}_{-1}\|_1 \leq 1\} \\
&= \sup\{z_1\beta_1 : |\beta_1| \leq 1\} + \sup\{\boldsymbol{z}_{-1}^\top\boldsymbol{\beta}_{-1} : \|\boldsymbol{\beta}_{-1}\|_1 \leq 1\} \\
&= |z_1| + \|\boldsymbol{z}_{-1}\|_\infty.
\end{aligned}$$

□

Proof of Lemma 2.3. Consider the series of equalities:

$$\begin{aligned}
\min_{\boldsymbol{\beta}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\beta}\|^2 + \lambda P(\boldsymbol{\beta}) &= \min_{\boldsymbol{\beta}} \max_{P_*(\boldsymbol{u}) \leq \lambda} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\beta}\|^2 + \boldsymbol{\beta}^\top\boldsymbol{u} \\
&= \max_{P_*(\boldsymbol{u}) \leq \lambda} \min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\beta}\|^2 + \beta^T u \\
&= \max_{P_*(\boldsymbol{u}) \leq \lambda} \ \frac{1}{2}\|\boldsymbol{y} - (\boldsymbol{y} - \boldsymbol{u})\|^2 + (\boldsymbol{y} - \boldsymbol{u})^T\boldsymbol{u} \\
&= \max_{P_*(\boldsymbol{u}) \leq \lambda} \ \boldsymbol{y}^\top\boldsymbol{u} - \frac{1}{2}\|\boldsymbol{u}\|^2 \\
&= \max_{P_*(\boldsymbol{u}) \leq \lambda} \ -\frac{1}{2}\|\boldsymbol{u} - \boldsymbol{y}\|^2 + \textit{constant}.
\end{aligned}$$

This is equivalent to the problem

$$\begin{aligned}
&\underset{\boldsymbol{u} \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{u}\|^2 \\
&\text{subject to} \ \ |u_1| + \|\boldsymbol{u}_{-1}\|_\infty \leq \lambda,
\end{aligned}$$

which, in turn, is equivalent to (2.17). □

A.2.1 Proof of Theorem 2.1

We consider the function

$$f(\lambda_1) = \frac{1}{2}\|\boldsymbol{u}(\lambda_1) - \boldsymbol{y}\|^2, \tag{A.5}$$

where $\boldsymbol{u}(\lambda_1)$ is a vector-valued function of λ_1, as defined in (2.18). We wish to minimize this function over the interval $[0, \lambda]$. We will prove this theorem using a series of claims.

Claim A.1. *The function $f(\lambda_1)$ is convex on $\mathbb{R}$.*

Proof. Note that

$$\begin{aligned}(y_1 - u_1(\lambda_1))^2 &= (y_1 - y_1)^2\mathbf{1}(|y_1| \leq \lambda_1) + (y_1 - \lambda_1\text{sign}(y_1))^2\mathbf{1}(|y_1| > \lambda_1) \\ &= (y_1 - \lambda_1\text{sign}(y_1))^2\mathbf{1}(|y_1| > \lambda_1)\end{aligned} \tag{A.6}$$

and

$$(y_i - u_i(\lambda_1))^2 = (y_i - (\lambda - \lambda_1)\text{sign}(y_i))^2\mathbf{1}(\lambda_1 > \lambda - |y_i|). \tag{A.7}$$

By inspection, both (A.6) and (A.7) are convex. The result follows from the fact that the sum of convex functions is convex. □

Claim A.2. *The derivative of $f(\lambda_1)$ is given by*

$$\frac{d}{d\lambda_1}f(\lambda_1) = [\lambda_1 - |y_1|]\mathbf{1}(|y_1| > \lambda_1) + \sum_{i=1}^{p-1}\left[\lambda_1 - z_{(i)}\right]\mathbf{1}(\lambda_1 > z_{(i)}), \tag{A.8}$$

where z is as defined in Theorem 2.1.

Proof. Note that $f(\lambda_1)$ can be rewritten as

$$f(\lambda_1) = (y_1 - \lambda_1\text{sign}(y_1))^2\mathbf{1}(|y_1| > \lambda_1) + \sum_{i=2}^{p}(y_i - (\lambda - \lambda_1)\text{sign}(y_i))^2\mathbf{1}(\lambda_1 > \lambda - |y_i|).$$

The result follows by inspection.

□

Claim A.3. *Define*

$$\lambda_1(m) = \frac{|y_1| + \sum_{j=1}^{m} z_{(j)}}{m+1}. \tag{A.9}$$

Then

$$\underset{\lambda_1 \in \mathbb{R}}{\operatorname{argmin}}\ f(\lambda_1) = \min_m \lambda_1(m). \tag{A.10}$$

Proof. Let $z_{(p)} \equiv \infty$, and define $\lambda_1(m) \equiv \frac{|y_1|+\sum_{j=1}^{m} z_{(j)}}{m+1}$. The optimality conditions for $f(\lambda_1)$ guarantee that if $\lambda_1(m) \in (z_{(m)}, z_{(m+1)}]$, then $\hat{\lambda}_1 = \lambda_1(m)$.

If the set $\arg\min_m \lambda_1(m)$ contains a single element, then define $k \equiv \arg\min_m \lambda_1(m)$; otherwise, let k be the smallest element of the set. To complete the proof, it suffices to show that $\lambda_1(k) \in (z_{(k)}, z_{(k+1)}]$.

First, we will show that $\lambda_1(k) > z_{(k)}$. By definition of $\lambda_1(k)$, we know that $\lambda_1(k) < \lambda_1(k-1)$. In other words,

$$\frac{|y_1| + \sum_{j=1}^{k} z_{(j)}}{k+1} < \frac{|y_1| + \sum_{j=1}^{k-1} z_{(j)}}{k}.$$

Rearranging terms, we find that

$$\left(|y_1| + \sum_{j=1}^{k} z_{(j)}\right)\left(1 - \frac{1}{k+1}\right) < |y_1| + \sum_{j=1}^{k-1} z_{(j)}.$$

Consequently,

$$z_{(k)} - \frac{|y_1| + \sum_{j=1}^{k} z_{(j)}}{k+1} < 0.$$

This means that $z_{(k)} < \lambda_1(k)$.

We now use a similar argument to show that $\lambda_1(k) \leq z_{(k+1)}$. By definition of $\lambda_1(k)$, we

know that $\lambda_1(k) \leq \lambda_1(k+1)$. In other words,

$$\frac{|y_1| + \sum_{j=1}^{k} z_{(j)}}{k+1} \leq \frac{|y_1| + \sum_{j=1}^{k+1} z_{(j)}}{k+2}.$$

Rearranging terms, we find that

$$\left(|y_1| + \sum_{j=1}^{k} z_{(j)}\right)\left(1 + \frac{1}{k+1}\right) \leq |y_1| + \sum_{j=1}^{k+1} z_{(j)} = \left(|y_1| + \sum_{j=1}^{k} z_{(j)}\right) + z_{(k+1)}.$$

This implies that $\lambda_1(k) \leq z_{(k+1)}$. □

Since $f(\lambda_1)$ is convex, its minimizer in the interval $[0, \lambda]$ is simply the projection of its minimizer on $\mathbb{R}$ (given in Claim A.3) into the interval. This completes the proof of Theorem 2.1.

□

A.3 Proofs of results in Section 2.3

Derivation of Claim 2.1. As mentioned in the main text, an unbiased estimate for the degrees of freedom of (2.26) is given by

$$\widehat{\mathrm{df}} = \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i} = \mathrm{trace}\left(\frac{d\hat{\boldsymbol{y}}}{d\boldsymbol{y}}\right), \tag{A.11}$$

provided that $\hat{\boldsymbol{y}}(\boldsymbol{y})$ is almost differentiable. The proof that $\hat{\boldsymbol{y}}(\boldsymbol{y})$ is almost differentiable follows from arguments similar to those in Tibshirani et al. (2012).

We now derive an explicit form for (A.11). To evaluate $\frac{d\hat{\boldsymbol{y}}}{d\boldsymbol{y}}$, we first note that $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$, the solution of (2.26) restricted to the active set, takes the form

$$\widehat{\boldsymbol{\beta}}_{\mathcal{A}} = \underset{\boldsymbol{\beta}_{\mathcal{A}}}{\mathrm{argmin}} \left\{\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}\|_2^2 + \sum_{d} \lambda_d P_d(\boldsymbol{A}_d^{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}})\right\}. \tag{A.12}$$

Therefore, $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$ must satisfy

$$-\boldsymbol{X}_{\mathcal{A}}^{\top}(\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}})+\sum_{d}\lambda_{d}(\boldsymbol{A}_{d}^{\mathcal{A}})^{\top}\dot{P}_{d}(\boldsymbol{A}_{d}^{\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}})=0. \tag{A.13}$$

We then differentiate with respect to $\boldsymbol{y}$ and apply the chain rule, to obtain

$$-\boldsymbol{X}_{\mathcal{A}}^{\top}+\boldsymbol{X}_{\mathcal{A}}^{\top}\boldsymbol{X}_{\mathcal{A}}\frac{d\widehat{\boldsymbol{\beta}}_{\mathcal{A}}}{d\boldsymbol{y}}+\sum_{d}\lambda_{d}(\boldsymbol{A}_{d}^{\mathcal{A}})^{\top}\ddot{P}_{d}(\boldsymbol{A}_{d}^{\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}})\left(\boldsymbol{A}_{d}^{\mathcal{A}}\right)\frac{d\widehat{\boldsymbol{\beta}}_{\mathcal{A}}}{d\boldsymbol{y}}=0. \tag{A.14}$$

Solving for $\frac{d\widehat{\boldsymbol{\beta}}_{\mathcal{A}}}{d\boldsymbol{y}}$ gives us

$$\frac{d\widehat{\boldsymbol{\beta}}_{\mathcal{A}}}{d\boldsymbol{y}}=\left[\boldsymbol{X}_{\mathcal{A}}^{\top}\boldsymbol{X}_{\mathcal{A}}+\sum_{d}\lambda_{d}(\boldsymbol{A}_{d}^{\mathcal{A}})^{\top}\ddot{P}_{d}(\boldsymbol{A}_{d}^{\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}})(\boldsymbol{A}_{d}^{\mathcal{A}})\right]^{-1}\boldsymbol{X}_{\mathcal{A}}^{\top}. \tag{A.15}$$

Form the definition of $\widehat{\boldsymbol{y}}=\boldsymbol{X}_{\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$, we get

$$\frac{d\widehat{\boldsymbol{y}}}{d\boldsymbol{y}}=\boldsymbol{X}_{\mathcal{A}}\frac{d\widehat{\boldsymbol{\beta}}_{\mathcal{A}}}{d\boldsymbol{y}}=\boldsymbol{X}_{\mathcal{A}}\left[\boldsymbol{X}_{\mathcal{A}}^{\top}\boldsymbol{X}_{\mathcal{A}}+\sum_{d}\lambda_{d}(\boldsymbol{A}_{d}^{\mathcal{A}})^{\top}\ddot{P}_{d}(\boldsymbol{A}_{d}^{\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}})(\boldsymbol{A}_{d}^{\mathcal{A}})\right]^{-1}\boldsymbol{X}_{\mathcal{A}}^{\top}. \tag{A.16}$$

In order to make this derivation entirely rigorous, we would need to show that $\widehat{\boldsymbol{\beta}}$ is unique, and that with probability one, within some neighbourhood of $\boldsymbol{y}$, the active set $\mathcal{A}$ does not change as a function of $\boldsymbol{y}$.

□

Appendix B

TECHNICAL DETAILS FOR "GENERALIZED SPARSE ADDITIVE MODELS"

B.1 Proof of results in Section 3.2.2

Proof of Lemma 3.1. (a) $\Leftrightarrow$ (b). By a simple calculation we see that P_{semi}^2 is pathwise differentiable at 0, and its derivative is given by $\frac{\partial}{\partial\varepsilon}P_{semi}^2(\varepsilon h)\big|_{\varepsilon=0} = \frac{\partial}{\partial\varepsilon}\varepsilon^2 P_{semi}^2(h)\big|_{\varepsilon=0} = 0$ for any function h. Now for any direction h, we can calculate a subdifferential of the objective in (3.4) evaluated at $f \equiv 0$:

$$\frac{\partial}{\partial\varepsilon}\left[n^{-1}\sum_{i=1}^{n}\left(y_i - \varepsilon h\left(x_i\right)\right)^2 + \lambda^2 P_{semi}^2\left(\varepsilon h\right) + \lambda\left\|\varepsilon h\right\|_n\right]_{\varepsilon=0} \ni -n^{-1}\sum_i y_i h(x_i) + \lambda U\left\|h\right\|_n \equiv \delta_{quad}(U),$$

for any $U \in [-1,1]$. By the sub-gradient conditions, $\widehat{f}_1 \equiv 0$ if and only if for every h, $\delta_{quad}(U) = 0$ for some $U \in [-1,1]$. For any given h, we have that

$$\delta_{quad}(U) = 0 \Leftrightarrow -n^{-1}\sum_i y_i h(x_i) + \lambda U\left\|h\right\|_n = 0 \Leftrightarrow \left|n^{-1}\sum_i y_i \frac{h(x_i)}{\left\|h\right\|_n}\right| = \lambda|U|.$$

From this we see that for a given h, $\delta_{quad}(U) = 0$ for some $U \in [-1,1]$ if and only if $|n^{-1}\sum_i y_i h(x_i)/\|h\|_n| \le \lambda$. It follows that, $\widehat{f}_1 \equiv 0$ if and only if $|n^{-1}\sum_i y_i h(x_i)/\|h\|_n| \le \lambda$ for every direction h.

(b) $\Rightarrow$ (c). Since $|n^{-1}\sum_i y_i h(x_i)/\|h\|_n| \le \lambda$ holds for every direction h, we simply consider the special direction such that $h(x_i) = y_i$. This implies (c).

(b) $\Leftarrow$ (c). Assuming that $\|\boldsymbol{y}\|_n \le \lambda$ then, for every h

$$\left|n^{-1}\sum_i y_i \frac{h(x_i)}{\left\|h\right\|_n}\right| \le \|\boldsymbol{y}\|_n \frac{\|h\|_n}{\|h\|_n} \le \|\boldsymbol{y}\|_n \le \lambda.$$

□

Proof of Lemma 3.2. (a) ⇔ (b). By the absolute homogeneity of semi-norms,

$$\frac{\partial}{\partial\varepsilon}P_{semi}(\varepsilon h)\big|_{\varepsilon=0} = \frac{\partial}{\partial\varepsilon}|\varepsilon|P_{semi}(h)\big|_{\varepsilon=0}.$$

Hence the subdifferential evaluated at $f \equiv 0$, in the direction of h, is

$$\frac{\partial}{\partial\varepsilon}\left[n^{-1}\sum_{i=1}^{n}\left(y_i - \varepsilon h\left(x_i\right)\right)^2 + \lambda^2 P_{semi}\left(\varepsilon h\right) + \lambda\left\|\varepsilon h\right\|_n\right]_{\varepsilon=0} \ni -n^{-1}\sum_i y_i h(x_i) + \lambda^2 V P_{semi}(h) + \lambda U\left\|h\right\|_n$$
$$\equiv \delta_{semi}(U,V),$$

for any $(U,V) \in [-1,1]^2$. As in the previous lemma, $\widehat{f}_2 \equiv 0$ if and only if for every h, there exists $(U,V) \in [-1,1]^2$ such that $\delta_{semi}(U,V) = 0$. For a given h, we have that

$$\delta_{semi}(U,V) = 0 \Leftrightarrow n^{-1}\sum_i y_i h(x_i) - \lambda^2 V P_{semi}(h) = \lambda U\left\|h\right\|_n$$
$$\Leftrightarrow \left|n^{-1}\sum_i y_i \frac{h(x_i)}{\|h\|_n} - \lambda^2 V\frac{P_{semi}(h)}{\|h\|_n}\right| = \lambda|U|.$$

Thus for a given h, $\delta_{semi}(U,V) = 0$ for some $(U,V) \in [-1,1]^2$ if and only if

$$\left|n^{-1}\sum_i y_i \frac{h(x_i)}{\|h\|_n} - \lambda^2 V\frac{P_{semi}(h)}{\|h\|_n}\right| \le \lambda,$$

for some $V \in [-1,1]$. This proves the first part.

Now we will show that $\|\boldsymbol{y}\|_n \le \lambda$ implies (b). That is, for every h, there exists some $V \in [-1,1]$ such that

$$\left|n^{-1}\sum_i y_i \frac{h(x_i)}{\|h\|_n} - \lambda^2 V\frac{P_{semi}(h)}{\|h\|_n}\right| \le \lambda.$$

We can simply take $V = 0$ for every h, which reduces the above inequality to the one seen in the proof of Lemma 3.1. Thus $\|\boldsymbol{y}\|_n \le \lambda \Rightarrow \widehat{f}_2 \equiv 0$. □

B.2 Proof of results in Section 3.3

Proof of Lemma 3.3. If $\widetilde{f} \equiv 0$, then $\widehat{f} \equiv 0$ is trivially the solution to (3.12). Thus, throughout this proof, we consider $\widetilde{f} \not\equiv 0$.

Case 1: $\|\widetilde{f}\|_n \geq \lambda_2$. In this case $c\widehat{f} = \widetilde{f}$ where $c = (1 - \lambda_2/\|\widetilde{f}\|_n)^{-1}$. Let $f_T \in \mathcal{F}$ be some arbitrary function and define the function $h = f_T - \widehat{f}$. We will show that along the path $\widehat{f} + \varepsilon h$ for all $\varepsilon \in [0, 1]$, the objective

$$\frac{1}{2}\left\|r - (\widehat{f} + \varepsilon h)\right\|_n^2 + \lambda_1 P_{st}\left(\widehat{f} + \varepsilon h\right) + \lambda_2 \|\widehat{f} + \varepsilon h\|_n \tag{B.1}$$

is minimized at $\varepsilon = 0$. We begin by noting that

$$\frac{1}{2}\left\|r - (\widetilde{f} + \varepsilon c h)\right\|_n^2 + \lambda_1 P_{st}\left(\widetilde{f} + \varepsilon c h\right),$$

is minimized at $\varepsilon = 0$ because

$$\widetilde{f} + \varepsilon c h = \widetilde{f} + \varepsilon c f_T - \varepsilon c \widehat{f} = (1 - \varepsilon)\widetilde{f} + \varepsilon c f_T \in \mathcal{F},$$

for all $\varepsilon \in [0, 1]$ since $\mathcal{F}$ is a convex cone. By the sub-gradient condition, we have

$$-\langle r - \widetilde{f}, ch\rangle_n + \lambda_1 \vartheta_1 = 0,$$

for some $\vartheta_1 \in \partial\, P_{st}\left(\widetilde{f} + \varepsilon c h\right)\Big|_{\varepsilon=0}$, or equivalently

$$c\left[-\langle r - c\widehat{f}, h\rangle_n + \lambda_1 \vartheta_2\right] = 0,$$

for some $\vartheta_2 \in \partial\, P_{st}\left(\widehat{f} + \varepsilon h\right)\Big|_{\varepsilon=0}$.

At $\widehat{f} + \varepsilon h$, one possible sub-gradient of the objective (B.1) is

$$-\langle r - \widehat{f}, h\rangle_n + \lambda_1 \vartheta_2 + \lambda_2 \frac{\langle \widehat{f}, h\rangle_n}{\|\widehat{f}\|_n}.$$

By the definition of c, we have that $\lambda_2/\|\widehat{f}\|_n = c\lambda_2/\|\widetilde{f}\|_n = c(1-1/c) = c-1$, and thus the above sub-gradient is

$$-\langle r - \widehat{f}, h\rangle_n + \lambda_1 \vartheta_2 + (c-1)\langle \widehat{f}, h\rangle_n = -\langle r - c\widehat{f}, h\rangle_n + \lambda_1 \vartheta_2 = 0.$$

Thus we have shown that the objective function (B.1) is minimized at $\varepsilon = 0$. Since f_T was an arbitrary function, we conclude that $\widehat{f}$ is the solution of (3.12).

Case 2: $\|\widetilde{f}\|_n < \lambda_2$**.** In this case we will show that $\widehat{f} \equiv 0$. For this, we consider the path εf_T for $\varepsilon \in [0,1]$ for an arbitrary $f_T \in \mathcal{F}$. We will show that the function

$$\frac{1}{2}\|r - \varepsilon f_T\|_n^2 + \lambda_1 P_{st}(\varepsilon f_T) + \lambda_2 \|\varepsilon f_T\|_n, \tag{B.2}$$

is minimized at $\varepsilon = 0$ and since f_T is arbitrary that will complete the proof.

As in the previous case, we begin by looking at the sub-gradient conditions for $\widetilde{f}$. The expression

$$\frac{1}{2}\left\|r - (\widetilde{f} + \varepsilon f_T)\right\|_n^2 + \lambda_1 P_{st}\left(\widetilde{f} + \varepsilon f_T\right),$$

is minimized at $\varepsilon = 0$ by definition of $\widetilde{f}$. This leads us to the sub-gradient condition

$$-\langle r - \widetilde{f}, f_T\rangle_n + \lambda_1 \vartheta_1 = 0 \Leftrightarrow \vartheta_1 = \frac{\langle r - \widetilde{f}, f_T\rangle_n}{\lambda_1}.$$

Now we describe the the sub-gradient conditions for (B.2). All sub-gradients of (B.2) at

$\varepsilon = 0$ are given by

$$-\langle r, f_T\rangle_n + \nu_1\lambda_1 P_{st}(f_T) + \nu_2\lambda_2\|f_T\|_n, \tag{B.3}$$

for real values $(\nu_1, \nu_2) \in [-1, 1]^2$. To complete the proof we need to find ν_1 and ν_2 such that (B.3) is 0 and, $(\nu_1, \nu_2) \in [-1, 1]^2$. Setting $\nu_1 = \vartheta_1/P_{st}(f_T)$ and $\nu_2 = \langle\widetilde{f}, f_T\rangle/(\lambda_2\|f_T\|_n)$ clearly makes (B.3) 0 and so we need only prove that our choice of ν_1 and ν_2 lie within the interval $[-1, 1]$.

Showing $|\nu_1| \leq 1$ is equivalent to $|\vartheta_1| \leq P_{st}(f_T)$. ϑ_1 is a member of the sub-gradient set

$$\partial\, P_{st}\left(\widetilde{f} + \varepsilon f_T\right)\Big|_{\varepsilon=0} = \left\{u \geq 0 : P_{st}(\widetilde{f} + \eta f_T) - P(\widetilde{f}) \geq u \times \eta \quad \forall\, \eta \geq 0\right\}.$$

Thus ϑ_1 must satisfy the inequality

$$\begin{aligned}\vartheta_1\eta &\leq P_{st}(\widetilde{f} + \eta f_T) - P_{st}(\widetilde{f}) \\ &\leq P_{st}(\widetilde{f}) + \eta P_{st}(f_T) - P_{st}(\widetilde{f}) = \eta P_{st}(f_T),\end{aligned}$$

where the second inequality holds because P_{st} is a semi-norm. This proves that $|\vartheta_1| \leq P_{st}(f_T)$.

Showing $|\nu_2| \leq 1$ is easier and follows by definition:

$$|\nu_2| = \frac{|\langle\widetilde{f}, f_T\rangle|}{\lambda_2\|f_T\|_n} \leq \frac{\|\widetilde{f}\|_n\|f_T\|_n}{\lambda_2\|f_T\|_n} = \frac{\|\widetilde{f}\|_n}{\lambda_2},$$

which is less than 1 since $\|\widetilde{f}\|_n < \lambda_2$.

□

Prof of Lemma 3.4. Consider an arbitrary direction $\widehat{f}_\lambda + \varepsilon h$ for some function h and ε in an open interval. We first consider the case $P_{st}(\widehat{f}_\lambda) \neq 0$. In this case if the directional derivative

$\nabla_h P^\nu_{st}(\widehat{f})$ exists then so does the directional derivative of $P_{st}(\widehat{f})$ and is given by

$$\nabla_h P_{st}(\widehat{f}_\lambda) = \frac{\nabla_h P^\nu_{st}(\widehat{f}_\lambda)}{\nu P^{\nu-1}_{st}(\widehat{f}_\lambda)}.$$

This follows from standard arguments for derivative of power functions. Here we present the simple case of integer valued $\nu > 1$.

$$\begin{aligned}
\nabla_h P_{st}(\widehat{f}_\lambda) &= \lim_{\varepsilon\to 0} \frac{P_{st}(\widehat{f}_\lambda + \varepsilon h) - P_{st}(\widehat{f}_\lambda)}{\varepsilon} \\
&= \lim_{\varepsilon\to 0} \frac{P_{st}(\widehat{f}_\lambda + \varepsilon h) - P_{st}(\widehat{f}_\lambda)}{\varepsilon} \times \frac{\sum_{l=1}^{\nu}\{P_{st}(\widehat{f}_\lambda + \varepsilon h)\}^{\nu-l}\{P_{st}(\widehat{f}_\lambda)\}^{l-1}}{\sum_{l=1}^{\nu}\{P_{st}(\widehat{f}_\lambda + \varepsilon h)\}^{\nu-l}\{P_{st}(\widehat{f}_\lambda)\}^{l-1}} \\
&= \lim_{\varepsilon\to 0} \frac{P^\nu_{st}(\widehat{f}_\lambda + \varepsilon h) - P^\nu_{st}(\widehat{f}_\lambda)}{\varepsilon} \times \lim_{\varepsilon\to 0} \frac{1}{\sum_{l=1}^{\nu}\{P_{st}(\widehat{f}_\lambda + \varepsilon h)\}^{\nu-l}\{P_{st}(\widehat{f}_\lambda)\}^{l-1}} \\
&= \nabla_h P^\nu_{st}(\widehat{f}_\lambda) \times \frac{1}{\sum_{l=1}^{\nu}\{P_{st}(\widehat{f}_\lambda)\}^{\nu-1}} = \frac{\nabla_h P^\nu_{st}(\widehat{f}_\lambda)}{\nu P^{\nu-1}_{st}(\widehat{f}_\lambda)}.
\end{aligned}$$

Now, by the gradient condition, for $f_\varepsilon = \widehat{f}_\lambda + \varepsilon h$,

$$\frac{\partial}{\partial \epsilon}\left[\frac{1}{2}\left\|r - f_\varepsilon\right\|_n^2 + \lambda P_{st}\left(f_\varepsilon\right)\right]_{\epsilon=0} = -\langle r - \widehat{f}_\lambda, h\rangle_n + \lambda \frac{\nabla_h P^\nu_{st}(\widehat{f}_\lambda)}{\nu P^{\nu-1}_{st}(\widehat{f}_\lambda)} = 0, \tag{B.4}$$

Similarly, for the path $f_{\widetilde{\varepsilon}} = \widetilde{f}_{\widetilde{\lambda}} + \widetilde{\varepsilon} h$, by the gradient condition,

$$\frac{\partial}{\partial \widetilde{\varepsilon}}\left[\frac{1}{2}\left\|r - f_{\widetilde{\varepsilon}}\right\|_n^2 + \widetilde{\lambda} P^\nu_{st}\left(f_{\widetilde{\varepsilon}}\right)\right]_{\widetilde{\varepsilon}=0} = -\langle r - \widetilde{f}_{\widetilde{\lambda}}, h\rangle_n + \widetilde{\lambda}\nabla_h P^\nu_{st}\left(\widetilde{f}_{\widetilde{\lambda}}\right) = 0.$$

This is exactly the optimality condition (B.4) with $\widetilde{\lambda} = \lambda(\nu P^{\nu-1}_{st}(\widehat{f}_\lambda))^{-1}$. Thus, if $\nu\widetilde{\lambda}P^{\nu-1}_{st}(\widetilde{f}_{\widetilde{\lambda}}) = \lambda$ then $\widehat{f}_\lambda = \widetilde{f}_{\widetilde{\lambda}}$.

Now to show the case $P_{st}(\widehat{f}) = 0$, we need to find conditions for which the objective

$$\frac{1}{2}\|f_{interp} - f\|_n^2 + \lambda P_{st}(f) \tag{B.5}$$

is minimized by $\widehat{f} = f_{null}$. We consider the functions $f_{h,\varepsilon} = (1-\varepsilon)f_{null} + \varepsilon h$ for $\varepsilon \in [0,1]$

and show that for all $h \in \mathcal{F}$, the objective

$$\frac{1}{2}\|f_{interp} - f_{h,\varepsilon}\|_n^2 + \lambda P_{st}(f_{h,\varepsilon}) \tag{B.6}$$

is minimized at $\varepsilon = 0$, if and only if $\lambda \geq P_{st}^*(f_{interp} - f_{null})$.

To see this, note that all subgradients of (B.6) at $\varepsilon = 0$, are of the form

$$\langle f_{interp} - f_{null}, f_{null} - h\rangle_n + \lambda\kappa P_{st}(h),$$

for $\kappa \in [-1, 1]$. For 0 to be a sub-gradient of (B.6) we need to have

$$\lambda\kappa P_{st}(h) = \langle f_{interp} - f_{null}, h - f_{null}\rangle_n. \tag{B.7}$$

Consequently, (B.6) is minimized at $\varepsilon = 0$ if and only if for all $h \in \mathcal{F}$

$$\langle f_{interp} - f_{null}, h - f_{null}\rangle_n \leq \lambda P_{st}(h). \tag{B.8}$$

Using the decomposition $h = h_1 + h_2 \in \mathcal{F}_1 \oplus \mathcal{F}_2$, the above condition becomes

$$\frac{\langle f_{interp} - f_{null}, h_1 - f_{null}\rangle_n}{P_{st}(h_2)} + \frac{\langle f_{interp} - f_{null}, h_2\rangle_n}{P_{st}(h_2)} \leq \lambda. \tag{B.9}$$

Now if $f_{interp} - f_{null} \in \mathcal{F}_2$, then the first part of the LHS is 0 and the second part is bounded above by $P_{st}^*(f_{interp} - f_{null})$. To complete the proof we show that $f_{interp} - f_{null}$ is, infact, a member of $\mathcal{F}_2$. For this, it suffices to show that $\langle f_{interp} - f_{null}, f_{null} - h_{null}\rangle_n = 0$ for all $h_{null} \in \mathcal{F}_1$. We know that f_{null} is the solution to the problem

$$\underset{f \in \mathcal{F}_1}{\text{minimize}}\ \frac{1}{2}\|f_{interp} - f\|_n^2;$$

in other words, for all $h_{null} \in \mathcal{F}_1$, the expression

$$\frac{1}{2}\|f_{interp} - (1-\varepsilon)f_{null} - \varepsilon h_{null}\|_n^2, \tag{B.10}$$

is minimized by $\varepsilon = 0$. Equivalently (by the gradient condition), for all $h_{null} \in \mathcal{F}_1$,

$$\langle f_{interp} - f_{null}, f_{null} - h_{null}\rangle_n = 0. \tag{B.11}$$

□

B.3 Proofs of results in Section 3.4.2

In this section we present the proof Theorem 3.1 and 3.3 for the sake of completeness. The arguments presented here are only a slight modification to those of Bühlmann and van de Geer (2011) for proving LASSO rates. One notable difference is that we explicitly handle an unpenalized intercept term; another is our handling of the structural penalty, $P_{st}(\cdot)$. Throughout the proofs, we will utilize the so-called *basic inequalities*. Hence, for the sake of convenience, we state and prove these basic inequalities as a separate lemma.

Lemma B.1 (Basic Inequality)**.** *Let $\widehat{f}(\boldsymbol{x}) = \widehat{\beta} + \sum_{j=1}^p \widehat{f}_j(x_j)$ be as defined in (3.24), and let $f^*(\boldsymbol{x}) = \beta^* + \sum_{j=1}^p f_j^*(x_j)$ be an arbitrary additive function with $\beta^* \in \mathcal{R}$ and $f_j^* \in \mathcal{F}$. Then we have the following basic inequality*

$$\mathcal{E}(\widehat{f}) + \lambda I(\widehat{f}) \leq -\left[\nu_n(\widehat{f}) - \nu_n(f^*)\right] + \lambda I(f^*) + \mathcal{E}(f^*). \tag{B.12}$$

If we further assume that $-\ell(\cdot)$ and $P_{st}(\cdot)$ are convex, then for all $t \in (0,1)$ and $\widetilde{f} = t\widehat{f} + (1-t)f^*$ we have the following basic inequality

$$\mathcal{E}(\widetilde{f}) + \lambda I(\widetilde{f}) \leq -\left[\nu_n(\widetilde{f}) - \nu_n(f^*)\right] + \lambda I(f^*) + \mathcal{E}(f^*). \tag{B.13}$$

Proof. For the first inequality, note that

$$-\mathbb{P}_n \ell\left(\widehat{f}\right) + \lambda I(\widehat{f}) \leq -\mathbb{P}_n \ell\left(f^*\right) + \lambda I(f^*),$$

which is equivalent to

$$\begin{aligned}
&\lambda I(\widehat{f}) \leq \mathbb{P}_n \ell\left(\widehat{f}\right) - \mathbb{P}_n \ell\left(f^*\right) + \lambda I(f^*) \\
\Leftrightarrow\ &\mathbb{P}(\ell(f^*)) - \mathbb{P}(\ell(\widehat{f})) + \lambda I(\widehat{f}) \leq \mathbb{P}_n \ell\left(\widehat{f}\right) - \mathbb{P}(\ell(\widehat{f})) - \mathbb{P}_n \ell\left(f^*\right) + \mathbb{P}(\ell(f^*)) + \lambda I(f^*) \\
\Leftrightarrow\ &\mathbb{P}(\ell(f^*)) - \mathbb{P}(\ell(\widehat{f})) + \lambda I(\widehat{f}) \leq -\left[(\mathbb{P}_n - \mathbb{P})(-\ell(\widehat{f})) - (\mathbb{P}_n - \mathbb{P})(-\ell(f^*))\right] + \lambda I(f^*) \\
\Leftrightarrow\ &\mathbb{P}(\ell(f^*)) - \mathbb{P}(\ell(\widehat{f})) + \lambda I(\widehat{f}) \leq -\left[\nu_n(\widehat{f}) - \nu_n(f^*)\right] + \lambda I(f^*) \\
\Leftrightarrow\ &\mathbb{P}(\ell(f^*)) - \mathbb{P}(\ell(f^0)) + \mathbb{P}(\ell(f^0)) - \mathbb{P}(\ell(\widehat{f})) + \lambda I(\widehat{f}) \leq -\left[\nu_n(\widehat{f}) - \nu_n(f^*)\right] + \lambda I(f^*) \\
\Leftrightarrow\ &-\mathcal{E}(f^*) + \mathcal{E}(\widehat{f}) + \lambda I(\widehat{f}) \leq -\left[\nu_n(\widehat{f}) - \nu_n(f^*)\right] + \lambda I(f^*).
\end{aligned}$$

For the second inequality we have by convexity

$$\begin{aligned}
-\mathbb{P}_n \ell\left(\widetilde{f}\right) + \lambda I(\widetilde{f}) &\leq t\left[-\mathbb{P}_n \ell\left(\widehat{f}\right) + \lambda I(\widehat{f})\right] + (1-t)\left[-\mathbb{P}_n \ell\left(f^*\right) + \lambda I(f^*)\right] \\
&\leq -\mathbb{P}_n \ell\left(f^*\right) + \lambda I(f^*),
\end{aligned}$$

after which we simply need to repeat the arguments for the previous basic inequality with $\widehat{f}$ replaced by $\widetilde{f}$. □

Proof of Theorem 3.1. Define

$$t = \frac{M^*}{M^* + I(\widehat{f} - f^*)}, \tag{B.14}$$

and $\widetilde{f} = t\widehat{f} + (1-t)f^*$. Then (B.14) implies $I(\widetilde{f} - f^*) \leq M^*$ and by Lemma B.1, we obtain

$$\mathcal{E}(\widetilde{f}) + \lambda I(\widetilde{f}) \leq Z_{M^*} + \lambda I(f^*) + \mathcal{E}(f^*).$$

By applying the triangle inequality $I(\widetilde{f} - f^* + f^*) \geq I(\widetilde{f} - f^*) - I(f^*)$, to the left hand side we obtain on the set $\mathcal{T}$ (where $Z_{M^*} \leq \rho M^* + 2R\rho$)

$$\mathcal{E}(\widetilde{f}) + \lambda I(\widetilde{f} - f^*) \leq \rho M^* + 2R\rho + 2\lambda I(f^*) + \mathcal{E}(f^*).$$

Recall the definition of M^*, given by

$$\rho M^* = \mathcal{E}(f^*) + 2\lambda I(f^*) + 2R\rho, \tag{B.15}$$

from which we obtain

$$\mathcal{E}(\widetilde{f}) + \lambda I(\widetilde{f} - f^*) \leq 2\rho M^* \leq 2\frac{\lambda}{4}M^* \Rightarrow I(\widetilde{f} - f^*) \leq \frac{M^*}{2}.$$

Now by the definition of $\widetilde{f}$ we have

$$I(\widehat{f} - f^*) = \frac{I(\widetilde{f} - f^*)}{t} = I(\widetilde{f} - f^*)\left[1 + \frac{I(\widehat{f} - f^*)}{M^*}\right] \leq \frac{M^*}{2} + \frac{I(\widehat{f} - f^*)}{2},$$

which implies that $I(\widehat{f} - f^*) \leq M^*$. Now we can repeat the above arguments with $\widetilde{f}$ replaced by $\widehat{f}$ which gives us

$$\mathcal{E}(\widehat{f}) + \lambda I(\widehat{f} - f^*) \leq \rho M^* + \rho(2R) + 2\lambda I(f^*) + \mathcal{E}(f^*).$$

□

Proof of Theorem 3.3. As in the proof of Theorem 3.1, we begin by defining t as

$$t = \frac{M^*}{M^* + |\widehat{\beta} - \beta^*| + I(\widehat{f} - f^*)},$$

and $\widetilde{f} = t\widehat{f} + (1-t)f^*$ which gives us $|\widetilde{\beta} - \beta^*| + I(\widetilde{f} - f^*) \leq M^*$, i.e. $\widetilde{f} \in \mathcal{F}^0_{local}$. Lemma B.1

implies that on the set $\mathcal{T}$ (where $Z_{M^*} \leq \rho M^*$) we have

$$\mathcal{E}(\widetilde{f}) + \lambda I(\widetilde{f}) \leq Z_{M^*} + \mathcal{E}^* + \lambda I(f^*) \leq \rho M^* + \mathcal{E}(f^*) + \lambda I(f^*). \tag{B.16}$$

Be definition of $\mathcal{S}_*$ and $I(\cdot)$, we have by the triangle inequality

$$\lambda I(f^*) = \lambda \sum_{j \in \mathcal{S}_*} \Big\{ \|f_j^*\|_n + \lambda P_{st}(f_j^*) \Big\} \leq \lambda \sum_{j \in \mathcal{S}_*} \Big\{ \|f_j^* - \widetilde{f}_j\|_n + \|\widetilde{f}_j\|_n + \lambda P_{st}(f_j^*) \Big\},$$

and by the reverse triangle inequality we have

$$\begin{aligned}
\lambda I(\widetilde{f}) &= \lambda \sum_{j \in \mathcal{S}_*} \Big\{ \|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j) \Big\} + \lambda \sum_{j \in \mathcal{S}_*^c} \Big\{ \|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j) \Big\} \\
&\geq \lambda \sum_{j \in \mathcal{S}_*} \Big\{ \|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j - f_j^*) - \lambda P_{st}(f_j^*) \Big\} + \lambda \sum_{j \in \mathcal{S}_*^c} \Big\{ \|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j) \Big\} \\
&= \lambda \sum_{j \in \mathcal{S}_*} \Big\{ \|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j - f_j^*) - \lambda P_{st}(f_j^*) \Big\} + \lambda \sum_{j \in \mathcal{S}_*^c} \Big\{ \|\widetilde{f}_j - f_j^*\|_n + \lambda P_{st}(\widetilde{f}_j - f_j^*) \Big\},
\end{aligned}$$

where the last equality follows from the fact that $f_j^* = 0$ for all $j \in \mathcal{S}_*^c$. With the above two inequalities combined with (B.16) we get

$$\begin{aligned}
\mathcal{E}(\widetilde{f}) + \lambda &\sum_{j \in \mathcal{S}_*} \Big\{ \|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j - f_j^*) - \lambda P_{st}(f_j^*) \Big\} + \lambda \sum_{j \in \mathcal{S}_*^c} \Big\{ \|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j) \Big\} \\
&\leq \lambda \sum_{j \in \mathcal{S}_*} \Big\{ \|f_j^* - \widetilde{f}_j\|_n + \|\widetilde{f}_j\|_n + \lambda P_{st}(f_j^*) \Big\} + \rho M^* + \mathcal{E}(f^*),
\end{aligned}$$

which simplifies to

$$\mathcal{E}(\widetilde{f}) + \lambda \sum_{j \in S_*^c} \|\widetilde{f}_j - f_j^*\|_n + \lambda \sum_{j=1}^{p} \lambda P_{st}(\widetilde{f}_j - f_j^*) \leq \lambda \sum_{j \in \mathcal{S}_*} \|f_j^* - \widetilde{f}_j\|_n + 2\lambda^2 \sum_{j \in \mathcal{S}_*} P_{st}(f_j^*) + \rho M^* + \mathcal{E}(f^*). \tag{B.17}$$

Now we add $\lambda|\widetilde{\beta}-\beta^*|+\lambda\sum_{j\in\mathcal{S}_*}\|\widetilde{f}_j-f_j^*\|_n$ to both sides of (B.17) to obtain

$$\mathcal{E}(\widetilde{f})+\lambda\Big\{|\widetilde{\beta}-\beta^*|+I(\widetilde{f}-f^*)\Big\}\leq 2\lambda\sum_{j\in\mathcal{S}_*}\|f_j^*-\widetilde{f}_j\|_n+\lambda|\widetilde{\beta}-\beta^*|+\rho M^*+\mathcal{E}(f^*)+2\lambda^2\sum_{j\in\mathcal{S}_*}P_{st}(f_j^*). \tag{B.18}$$

Case I. If

$$2\lambda\sum_{j\in\mathcal{S}_*}\|f_j^*-\widetilde{f}_j\|_n+\lambda|\widetilde{\beta}-\beta^*|\leq\rho M^*+\mathcal{E}(f^*)+2\lambda^2\sum_{j\in\mathcal{S}_*}P_{st}(f_j^*),$$

then (B.18) simplifies to

$$\begin{aligned}\mathcal{E}(\widetilde{f})+\lambda\Big\{|\widetilde{\beta}-\beta^*|+I(\widetilde{f}-f^*)\Big\}&\leq 2\rho M^*+2\mathcal{E}(f^*)+4\lambda^2\sum_{j\in\mathcal{S}_*}P_{st}(f_j^*)\\&\leq 4\rho M^*\leq 4\frac{\lambda}{8}M^*=\lambda M^*/2,\end{aligned}$$

which indicates that $|\widetilde{\beta}-\beta^*|+I(\widetilde{f}-f^*)\leq M^*/2$ which implies that $|\widehat{\beta}-\beta^*|+I(\widehat{f}-f^*)\leq M^*$ and hence we can redo the above arguments and replace $\widetilde{f}$ by $\widehat{f}$.

Case II. If instead

$$2\lambda\sum_{j\in\mathcal{S}_*}\|f_j^*-\widetilde{f}_j\|_n+\lambda|\widetilde{\beta}-\beta^*|\geq\rho M^*+\mathcal{E}(f^*)+2\lambda^2\sum_{j\in\mathcal{S}_*}P_{st}(f_j^*),$$

then we have

$$\mathcal{E}(\widetilde{f})+\lambda\Big\{|\widetilde{\beta}-\beta^*|+I(\widetilde{f}-f^*)\Big\}\leq 4\lambda\sum_{j\in\mathcal{S}_*}\|f_j^*-\widetilde{f}_j\|_n+2\lambda|\widetilde{\beta}-\beta^*|. \tag{B.19}$$

This is equivalent to

$$\mathcal{E}(\widetilde{f})+\lambda\Big\{\sum_{j\in\mathcal{S}_*^c}\|\widetilde{f}_j-f_j^*\|_n+\lambda\sum_{j=1}^{p}P_{st}(\widetilde{f}_j-f_j^*)\Big\}\leq 3\lambda\sum_{j\in\mathcal{S}_*}\|f_j^*-\widetilde{f}_j\|_n+\lambda|\widetilde{\beta}-\beta^*|,$$

which means we have that

$$\sum_{j\in\mathcal{S}_*^c}\|\widetilde{f}_j - f_j^*\|_n + \lambda\sum_{j=1}^{p} P_{st}(\widetilde{f}_j - f_j^*) \leq 3\sum_{j\in\mathcal{S}_*}\|f_j^* - \widetilde{f}_j\|_n + |\widetilde{\beta}-\beta^*|,$$

and hence by the compatibility condition (B.19) reduces to

$$\mathcal{E}(\widetilde{f}) + \lambda\Big\{|\widetilde{\beta}-\beta^*| + I(\widetilde{f}-f^*)\Big\} \leq 4\lambda\|\widetilde{f}-f^*\|\sqrt{s}/\phi(\mathcal{S}_*). \tag{B.20}$$

Since $\widetilde{f}$ and f^* are in $\mathcal{F}^0_{local}$, we invoke the inequality $uv \leq H(v) + G(u)$ to obtain

$$\begin{aligned}\frac{4\lambda\sqrt{s}\|\widetilde{f}-f^*\|}{\phi(\mathcal{S}_*)} &\leq \frac{8\lambda\sqrt{s}}{\phi(\mathcal{S}_*)}\left[\frac{\|\widetilde{f}-f^0\|}{2} + \frac{\|f^*-f^0\|}{2}\right] \\ &\leq H\left(\frac{8\lambda\sqrt{s}}{\phi(\mathcal{S}_*)}\right) + G\left(\frac{\|\widetilde{f}-f^0\|}{2} + \frac{\|f^*-f^0\|}{2}\right).\end{aligned}$$

By the convexity of G and the margin condition we obtain

$$\frac{4\lambda\sqrt{s}\|\widetilde{f}-f^*\|}{\phi(\mathcal{S}_*)} \leq H\left(\frac{8\lambda\sqrt{s}}{\phi(\mathcal{S}_*)}\right) + \frac{\mathcal{E}(\widetilde{f})}{2} + \frac{\mathcal{E}(f^*)}{2}.$$

Hence we have

$$\frac{\mathcal{E}(\widetilde{f})}{2} + \lambda\Big\{|\widetilde{\beta}-\beta^*| + I(\widetilde{f}-f^*)\Big\} \leq H\left(\frac{8\lambda\sqrt{s}}{\phi(\mathcal{S}_*)}\right) + \frac{\mathcal{E}(f^*)}{2} \leq \rho M^* \leq \lambda M^*/8, \tag{B.21}$$

which implies that $|\widetilde{\beta}-\beta^*|+I(\widetilde{f}-f^*) \leq M^*/2$ which in turn implies $|\widehat{\beta}-\beta^*|+I(\widehat{f}-f^*) \leq M^*$ and hence we can redo the above arguments and replace $\widetilde{f}$ by $\widehat{f}$.

Thus we have shown that

$$\mathcal{E}(\widehat{f}) + \lambda\Big\{|\widehat{\beta}-\beta^*| + I(\widehat{f}-f^*)\Big\} \leq 4\rho M^*. \tag{B.22}$$

□

B.4 The set $\mathcal{T}$

Theorems 3.1 and 3.2 show inequalities holding over the set $\mathcal{T}$. In this section we will show that $\mathcal{T}$ occurs with high probability. This will be shown for the two different types of entropy bounds considered in Section 3.4. We consider the special case of loss functions linear in Y_i as in Corollaries 3.1.1 and 3.2.1 and bound the term $\nu_n(f) - \nu_n(f^*)$ in the following theorem.

Theorem B.1. *Let $\boldsymbol{x}_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ denote the fixed covariates and response, respectively, for $i = 1, \ldots, n$. Assume that for any function f the loss $\ell(\cdot)$ is such that*

$$-\ell(f) = -\ell(f, \boldsymbol{x}_i, Y_i) = aY_i f(\boldsymbol{x}_i) + b(f(\boldsymbol{x}_i)), \tag{B.23}$$

for some $a \in \mathbb{R}\backslash\{0\}$ and function $b : \mathbb{R} \to \mathbb{R}$. Further assume that $Y_i - \mathbb{E}Y_i = Y_i - \mu_i$ are uniformly sub-Gaussian, i.e.

$$\max_{i=1,\ldots,n} K^2 \left(\mathbb{E}e^{(Y_i - \mu_i)^2 / K^2} - 1 \right) \leq \sigma_0^2, \tag{B.24}$$

then with probability at-least $1 - 2\exp\left[-n\rho^2 C_1\right] - C\exp\left[-n\rho^2 C_2\right]$ the following inequality holds

$$\nu_n(f) - \nu_n(f^*) \leq \rho \left[|\beta - \beta^*| + \sum_{j=1}^{p} \|f_j - f_j^*\|_n + \lambda P_{st}(f_j - f_j^*) \right], \tag{B.25}$$

for variables ρ, λ and positive constants C, C_1, C_2 which we specify in the following 3 cases.

Case 1. *If $\mathcal{F}$ has a logarithmic entropy bound, then the inequality (B.25) holds with $\rho = \kappa \max\left(\sqrt{\frac{T_n}{n}}, \sqrt{\frac{\log p}{n}} \right)$ and $\lambda = 1$ for constants $\kappa = \kappa(a, K, \sigma_0, A_0)$, $C_1 = C_1(K, \sigma_0)$, $C = C(K, \sigma_0)$ and $C_2 = C_2(C, \kappa)$.*

Case 2. *If $\mathcal{F}$ has a polynomial entropy bound with smoothness, then the inequality (B.25) holds with $\rho = \kappa \max\left(n^{-\frac{1}{2+\alpha}}, \sqrt{\frac{\log p}{n}} \right)$ for constants $\kappa = \kappa(a, K, \sigma_0, A_0, \alpha)$, $C_1 = C_1(K, \sigma_0)$, $C = C(K, \sigma_0)$ and $C_2 = C_2(C, \kappa)$. The parameter satisfies $\lambda \asymp \rho$ and $\lambda \geq 8\rho$.*

In light of the above theorem, for the case of Theorem 3.1 where

$$Z_{M^*} = \sup_{I(f-f^*)\leq M^*} |\nu_n(f) - \nu_n(f^*)|,$$

and $\beta, \beta^* \in \mathcal{R}$ where $\mathcal{R}$ is uniformly bounded by R then we have with probability at-least $1 - 2\exp\left[-n\rho^2 C_1\right] - C\exp\left[-n\rho^2 C_2\right]$,

$$Z_{M^*} \leq \rho\left(M^* + 2R\right).$$

In the case of Theorem 3.3 where

$$Z_{M^*} = \sup_{|\beta-\beta^*|+I(f-f^*)\leq M^*} |\nu_n(f) - \nu_n(f^*)|,$$

we have with probability at-least $1 - 2\exp\left[-n\rho^2 C_1\right] - C\exp\left[-n\rho^2 C_2\right]$,

$$Z_{M^*} \leq \rho M^*.$$

To prove Theorem B.1, we use a few technical lemmas from van de Geer (2000) namely Lemma 8.2 and Corollary 8.3; for the sake of completeness we state these lemmas in Appendix B.5.

Proof of Theorem B.1. We begin by noting that for any arbitrary function we have

$$\nu_n(f) = (\mathbb{P}_n - \mathbb{P})(-\ell(f)) = \frac{1}{n}\sum_{i=1}^{n}\left[aY_i f(\boldsymbol{x}_i) + b(f(\boldsymbol{x}_i))\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} aY_i f(\boldsymbol{x}_i) + b(f(\boldsymbol{x}_i))\right].$$

Since we assume the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are fixed we obtain

$$\nu_n(f) = \frac{1}{n}\sum_{i=1}^{n} a(Y_i - \mu_i) f(\boldsymbol{x}_i) \equiv a\langle \boldsymbol{Y} - \boldsymbol{\mu}, f\rangle_n,$$

where $\mu_i = \mathbb{E}Y_i$. Thus for additive functions f and f^* we obtain

$$\begin{aligned}\nu_n(f) - \nu_n(f^*) = a\langle \boldsymbol{Y} - \boldsymbol{\mu}, f - f^* \rangle_n &= \frac{1}{n}\sum_{i=1}^n a(Y_i - \mu_i)\left[\beta - \beta^* + \sum_{j=1}^p f_j(x_{ij}) - f_j^*(x_{ij})\right] \\ &= a(\beta - \beta^*)\frac{1}{n}\sum_{i=1}^n (Y_i - \mu_i) + \sum_{j=1}^p \frac{1}{n}\sum_{i=1}^n a(Y_i - \mu_i)(f_j(x_{ij}) - f_j^*(x_{ij})) \\ &= a(\beta - \beta^*)(\overline{\boldsymbol{Y}} - \overline{\boldsymbol{\mu}}) + \sum_{j=1}^p a\langle \boldsymbol{Y} - \boldsymbol{\mu}, f_j - f_j^* \rangle_n.\end{aligned}$$

From now on we will assume, without loss of generality, that $|a| = 1$ since this constant is absorbed into a constant κ which we define later.

To control the first term, $(\beta - \beta^*)(\overline{\boldsymbol{Y}} - \overline{\boldsymbol{\mu}})$, we simply apply Lemma B.2. For the second part, we consider 2 cases.

Case 1: Logarithmic Entropy. We first note that if the entropy bound holds, then the same bound holds (upto a constant) for the class

$$\left\{ \frac{f_j - f_j^*}{\|f_j - f_j^*\|_n + \lambda P_{st}(f_j - f_j^*)} : f_j \in \mathcal{F} \right\}, \tag{B.26}$$

for some $f_j^* \in \mathcal{F}$ for all $j = 1, \ldots, p$. Now we apply Lemma B.3 to the above class by first noting that $R \leq 1$ and then using the bound for Dudley's integral

$$A_0^{1/2} T_n^{1/2} \int_0^1 \log^{1/2}\left(\frac{1}{u} + 1\right) du \leq \widetilde{A}_0 T_n^{1/2}, \tag{B.27}$$

we have for all δ that satisfy

$$\delta \geq 2C\widetilde{A}_0 \sqrt{\frac{T_n}{n}}, \tag{B.28}$$

where the constant C depends only on K and σ_0, we have

$$\mathbb{P}\left(\sup_{f_j \in \mathcal{F}} \left| \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \mu_i)\left(f_j(x_{ij}) - f_j^*(x_{ij})\right)}{\|f_j - f_j^*\|_n + \lambda P_{st}(f_j - f_j^*)} \right| \geq \delta \right) \leq C \exp\left[-\frac{n\delta^2}{4C^2} \right].$$

We can now take $\delta = \rho \geq 2C\widetilde{A}_0 \max\left\{\sqrt{\frac{T_n}{n}}, \sqrt{\frac{\log p}{n}}\right\} \geq 2C\widetilde{A}_0\sqrt{\frac{T_n}{n}}$ which holds for all $\kappa \geq 2C\widetilde{A}_0$. Applying the above result with a union bound gives us

$$\begin{aligned}
&\mathbb{P}\left(\max_{j=1,\ldots,p}\sup_{f_j\in\mathcal{F}}\frac{\left|\langle \boldsymbol{Y}-\boldsymbol{\mu}, f_j - f_j^*\rangle_n\right|}{\|f_j-f_j^*\|_n + \lambda P_{st}(f_j-f_j^*)} \geq \rho\right) \\
&\leq pC\exp\left[-\frac{n\rho^2}{4C^2}\right] = C\exp\left[-\frac{n\rho^2}{4C^2} + \log p\right] \\
&= C\exp\left[-n\rho^2\left(\frac{1}{4C^2} - \frac{\log p}{n\rho^2}\right)\right] \leq C\exp\left[-n\rho^2 C_2\right],
\end{aligned}$$

for a constant $C_2 > 0$ that depends on C and $\widetilde{A}_0$. To see this, note that

$$\frac{1}{4C^2} - \frac{\log p}{n\rho^2} = \frac{1}{4C^2} - \frac{1}{\kappa\max\left\{\frac{T_n}{\log p}, 1\right\}} \geq \frac{1}{4C^2} - \frac{1}{\kappa},$$

which is positive if $\kappa > 4C^2$. Thus we can take the constant κ such that $\kappa > \max\left\{4C^2, 2C\widetilde{A}_0\right\}$. Hence κ depends on $C(K,\sigma_0)$ and $\widetilde{A}_0(A_0)$.

Case 2: Polynomial Entropy with Smoothness. Now we note that same entropy bound holds for the class

$$\widetilde{\mathcal{F}} = \left\{\frac{f_j - f_j^*}{\|f_j - f_j^*\|_n + \lambda P_{st}(f_j - f_j^*)} : f_j \in \mathcal{F}\right\}, \tag{B.29}$$

and we can now apply Lemma B.3 by noting that

$$\int_0^1 H^{1/2}(u, \widetilde{\mathcal{F}}, Q_n)\, du \leq \widetilde{A}_0\lambda^{-\alpha/2},$$

for some constant $\widetilde{A}_0 = \widetilde{A}_0(A_0)$. For some $C = C(K,\sigma_0)$ and all $\delta \geq 2C\widetilde{A}_0\lambda^{-\alpha/2}n^{-1/2}$ we have

$$\mathbb{P}\left(\sup_{f_j\in\mathcal{F}}\frac{\left|\langle \boldsymbol{Y}-\boldsymbol{\mu}, f_j - f_j^*\rangle_n\right|}{\|f_j-f_j^*\|_n + \lambda P_{st}(f_j-f_j^*)} \geq \delta\right) \leq C\exp\left[-\frac{n\delta^2}{4C^2}\right]. \tag{B.30}$$

Since $\lambda \geq \rho$ we note that $2C\widetilde{A}_0\lambda^{-\alpha/2}n^{-1/2} \leq 2C\widetilde{A}_0\rho^{-\alpha/2}n^{-1/2}$ and that

$$2C\widetilde{A}_0\rho^{-\alpha/2}n^{-1/2} \leq \rho \Leftrightarrow \rho \geq \left(2C\widetilde{A}_0\right)^{\frac{2}{2+\alpha}} n^{-\frac{1}{2+\alpha}}.$$

Which holds by definition since $\rho = \kappa \max\left(\sqrt{\frac{\log p}{n}}, n^{-\frac{1}{2+\alpha}}\right) \geq \kappa n^{-\frac{1}{2+\alpha}}$ and κ is sufficiently large (any $\kappa \geq \left(2C\widetilde{A}_0\right)^{\frac{2}{2+\alpha}}$ would suffice). Therefore, we can take $\delta = \rho$ in (B.30) along with a union bound to obtain

$$\begin{aligned}
\mathbb{P}\left(\max_{j=1,\ldots,p} \sup_{f_j \in \mathcal{F}} \frac{\left|\langle \boldsymbol{Y} - \boldsymbol{\mu}, f_j - f_j^* \rangle_n\right|}{\|f_j - f_j^*\|_n + \lambda P_{st}(f_j - f_j^*)} \geq \rho\right) &\leq pC \exp\left[-\frac{n\rho^2}{4C^2}\right] \\
&= C \exp\left[-n\rho^2\left(\frac{1}{4C^2} - \frac{\log p}{n\rho^2}\right)\right] \\
&\leq C \exp\left[-n\rho^2 C_2\right],
\end{aligned}$$

for some positive constant $C_2 = C_2(C, \widetilde{A}_0)$ exactly as in Case 1.

□

B.5 Some results from van de Geer (2000)

Lemma B.2 (Lemma 8.2 of van de Geer (2000))**.** *Suppose that $Y_1 - \mu_1, \ldots, Y_n - \mu_n$ are mean zero sub-Gaussian random variables, i.e., they satisfy (B.24). Then for all $\gamma \in \mathbb{R}^n$ and $\rho > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n (Y_i - \mu_i)\gamma_i\right| \geq \rho\right) \leq 2\exp\left[-\frac{\rho^2}{8(K^2 + \sigma_0^2)\sum_{i=1}^n \gamma_i^2}\right], \tag{B.31}$$

in particular if $\gamma_i = 1/n$ then we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i - \mu_i\right| \geq \rho\right) \leq 2\exp\left[-\frac{n\rho^2}{8(K^2 + \sigma_0^2)}\right]. \tag{B.32}$$

Lemma B.3 (Corollary 8.3 of van de Geer (2000))**.** *Suppose that $\sup_{f_j \in \mathcal{F}} \|f_j\|_n \leq R$ for a univariate function class $\mathcal{F}$ and that $Y_1 - \mu_1, \ldots, Y_n - \mu_n$ are mean zero sub-Gaussian random*

variables, i.e. they satisfy (B.24)*. Then for some constant* $C = C(K, \sigma_0)$*, and for all* $\delta > 0$ *satisfying*

$$\sqrt{n}\delta \geq 2C \left(\int_0^R H^{1/2}(u, \mathcal{F}, Q_n)\, du \vee R \right), \tag{B.33}$$

we have

$$\mathbb{P}\left(\sup_{f_j \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu_i) f_j(x_{ij}) \right| \geq \delta \right) \leq C \exp\left[-\frac{n\delta^2}{4C^2R^2} \right]. \tag{B.34}$$

Appendix C

TECHNICAL DETAILS FOR "NONPARAMETRIC REGRESSION WITH ADAPTIVE TRUNCATION VIA A CONVEX HIERARCHICAL PENALTY"

C.1 Algorithms for additive framework and extension to classification

Here we give an algorithm for our additive and sparse-additive framework as well as an algorithm for the extension of our proposal to classification. We use a block-wise coordinate descent algorithm for solving the additive and sparse additive proposal. This algorithm cyclically iterates through features, and for each feature applies the univariate solution detailed in Algorithm 3. The exact details are given in Algorithm 4 below.

Algorithm 4 Block coordinate descent for the additive and sparse additive framework

Initialize $\boldsymbol{\beta}_j \leftarrow 0$ for $j = 1, \ldots, p$
While $l \leq max\_iter$ **and** not converged
  For $j = 1, \ldots, p$
    Set $\boldsymbol{r}_{-j} \leftarrow \boldsymbol{y} - \sum_{j' \neq j} \boldsymbol{\Psi}^{j'} \boldsymbol{\beta}_{j'}$
    Update $\boldsymbol{\beta}_j \leftarrow \underset{\boldsymbol{\beta} \in \mathbb{R}^{\tilde{K}}}{\arg\min} \frac{1}{2} \left\| \boldsymbol{r}_{-j} - \boldsymbol{\Psi}^j \boldsymbol{\beta} \right\|_n^2 + \lambda^2 \sum_{k=1}^{\tilde{K}} \widetilde{w}_k \left\| \boldsymbol{\Psi}^j_{k:\tilde{K}} \boldsymbol{\beta}_{j,k:\tilde{K}} \right\|_n,$
      where $\widetilde{w}_1 = w_1 + \lambda^{-1}$ and $\widetilde{w}_k = w_k$ for $k = 2, \ldots, \tilde{K}$.
Return $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$

We also give an algorithm for the extension of our method to classification based on proximal gradient descent. To begin let $L(\beta_0, \boldsymbol{\beta}) = 1/(2n) \sum_{i=1}^{n} \log\left(1 + \exp\left[-y_i \left\{\beta_0 + (\boldsymbol{\Psi}\boldsymbol{\beta})_i\right\}\right]\right)$. We denote by $\nabla L(\beta_0, \boldsymbol{\beta})$, the derivative of L at the point $(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{\tilde{K}+1}$. Algorithm 5 presents the steps for solving (4.13). The algorithm for extension of additive models to classification can be similarly derived and is omitted in the interest of brevity.

Algorithm 5 Proximal gradient descent for extension to classification

Initialize $(\beta_0^0, \boldsymbol{\beta}^0)$
For $l = 1, 2, \ldots$ until convergence
Select a step size t_l via line search
Update

$$(\beta_0^l,\, \boldsymbol{\beta}^l) \leftarrow \underset{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{\tilde{K}+1}}{\arg\min} \frac{1}{2} \left\|(\beta_0, \boldsymbol{\beta}) - \left\{(\beta_0^{l-1}, \boldsymbol{\beta}^{l-1}) - t_l \nabla L(\beta_0^{l-1}, \boldsymbol{\beta}^{l-1})\right\}\right\|_2^2 + \lambda\Omega(\boldsymbol{\beta}).$$

Return $(\beta_0^l, \boldsymbol{\beta}^l)$

C.2 Proofs for Section 4.4.3

of Lemma 4.2. Firstly, we have for $f_1(x) = \sum_{k=1}^{\tilde{K}_n} \psi(x)\beta_k^{[1]}$, $f_2(x) = \sum_{k=1}^{\tilde{K}_n} \psi(x)\beta_k^{[2]} \in \mathcal{F}_n$

$$\begin{aligned}
\|f_1 - f_2\|_Q^2 &= \int (f_1 - f_2) dQ = \int \left\{ \sum_{k=1}^{\tilde{K}_n} \psi_k(x)(\beta_k^{[1]} - \beta_k^{[2]}) \right\}^2 dQ \\
&= \int \left\{ \sum_{k=1}^{\tilde{K}_n} \psi_k^2(x) \left(\beta_k^{[1]} - \beta_k^{[2]}\right)^2 + \sum_{k\neq l} \psi_k(x)\psi_l(x) \left(\beta_k^{[1]} - \beta_k^{[2]}\right)\left(\beta_l^{[1]} - \beta_l^{[2]}\right) \right\} dQ \\
&= \left\|\beta^{[1]} - \beta^{[2]}\right\|_2^2,
\end{aligned}$$

where the final equality follows due to the orthonormality of ψ_k. Similarly for $f_1, f_2 \in \mathcal{F}_{p,n}$ we can show that $\|f_1 - f_2\|_Q^2 = \|\beta^{[1]} - \beta^{[2]}\|_2^2$. Thus if $\{\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^N\}$ is the smallest δ-cover of $H_{\tilde{K}_n}^{w/M}$ then the functions $f_{\boldsymbol{\beta}}$ associated with $\{\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^N\}$ form the smallest δ-cover with respect to the L_Q norm. This can be extended to the case $n = \infty$. This proves the first part.

Secondly, note that for $f_1, f_2 \in \mathcal{F}_n$ (or $\mathcal{F}_{p,n}$) we have

$$\|f_1 - f_2\|_n^2 = (\boldsymbol{\beta}^{[1]} - \boldsymbol{\beta}^{[2]})^\top \frac{\boldsymbol{\Psi}^\top \boldsymbol{\Psi}}{n} (\boldsymbol{\beta}^{[1]} - \boldsymbol{\beta}^{[2]}) \leq \Lambda_{\max} \|\boldsymbol{\beta}^{[1]} - \boldsymbol{\beta}^{[2]}\|_2^2,$$

thus if $\{\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^N\}$ is the smallest δ-cover for $\mathcal{H}_{\tilde{K}_n}^w$, then the associated functions

$\{f^1, \ldots, f^N\}$ is a $\Lambda_{\max}^{1/2}\delta$ cover of $\{f_{\boldsymbol{\beta}} \in \mathcal{F}_n(\mathcal{F}_{p,n}) : \sum_{k=1}^{\tilde{K}_n} w_k \|\boldsymbol{\Psi}_{k:\tilde{K}_n}\boldsymbol{\beta}_{k:\tilde{K}_n}\| \leq 1\}$ with respect to the Q_n metric. Since this is a cover and not the smallest cover, we have

$$H(\Lambda_{\max}^{1/2}\delta, \{f_{\boldsymbol{\beta}} \in \mathcal{F}_n(\text{or } \mathcal{F}_{p,n}) : \sum_{k=1}^{\tilde{K}_n} w_k \|\boldsymbol{\Psi}_{k:\tilde{K}_n}\boldsymbol{\beta}_{k:\tilde{K}_n}\| \leq 1\}, Q_n) \leq H(\delta, \mathcal{H}^w_{\tilde{K}_n}),$$

and since the inequality holds for all $\delta > 0$, we can select $\delta = \delta'\Lambda_{\max}^{-1/2}$ giving us the result. □

Proof of Lemma 4.3. For the Ellipsoid $E^w_{\tilde{K}_n}$ where

$$E^w_{\tilde{K}_n} = \left\{ \boldsymbol{\beta} \in \mathbb{R}^{\tilde{K}_n} : \sum_{k=1}^{\tilde{K}_n} \beta_k^2 \left(w_1 + \cdots + w_k\right)^2 \leq 1 \right\}, \tag{C.1}$$

we show that $\mathcal{H}^w_{\tilde{K}_n} \subset E^w_{\tilde{K}_n}$ in Lemma C.4. Dumer (2006) proved an upper bound for ellipsoids which we state in Appendix C.6.1. For the special case of $w_k = k^m - (k-1)^m$, this theorem yields the desired upper bound as shown in Corollary C.1.1. Therefore we have $H(\delta, \mathcal{H}^w_{\tilde{K}_n}) \leq H(\delta, E^w_{\tilde{K}_n}) \leq U_{E,1}\delta^{-1/m}$.

Similarly, we can consider the special case of our multivariate framework weights in Corollary C.1.2, which gives us the result $H(\delta, \mathcal{H}^w_{\tilde{K}_n}) \leq H(\delta, E^w_{\tilde{K}_n}) \leq U_{E,2}\delta^{-p/m}$. □

of Lemma 4.4. Let d be the integer such that $(w_1 + \cdots + w_{d+1})^{-1} \leq \delta \leq (w_1 + \cdots + w_d)^{-1}$ for $\delta \in ((w_1 + \cdots + w_{\tilde{K}_n+1})^{-1}, 1)$. Note that since $\delta \geq (w_1 + \cdots + w_{\tilde{K}_n+1})^{-1}$, $d \leq \tilde{K}_n$. We define the truncated region as

$$\widetilde{H}^w_d = \left\{ \boldsymbol{\beta} \in \mathcal{H}^w_{\tilde{K}_n} : \beta_j = 0 \ \forall \ j \geq d+1 \right\}.$$

Then we have that $\mathcal{H}^w_d \subset \widetilde{\mathcal{H}}^w_d \subseteq \mathcal{H}^w_{\tilde{K}_n}$ where $\mathcal{H}^w_d$ is simply viewing $\widetilde{\mathcal{H}}^w_d$ as a subset of $\mathbb{R}^d$. Let $\mathbb{B}_n(r)$ be the n-ball of radius r. By Lemma C.5, we have $\mathbb{B}_d\left((w_1 + \cdots + w_d)^{-1}\right) \subset \mathcal{H}^w_d$. The lower bound of the entropy of a ball can be obtained by a simple volume argument. Since

$(w_1 + \cdots + w_d)^{-1} \geq \delta$ then $\mathbb{B}_d(\delta) \subseteq \mathbb{B}_d((w_1 + \cdots + w_d)^{-1})$ and hence

$$H(\delta/2, \mathcal{H}_d^w) \geq H(\delta/2, \mathbb{B}_d(\delta)) \geq \log \frac{Vol(\mathbb{B}_d(\delta))}{Vol(\mathbb{B}_d(\delta/2))} = d \log(2).$$

Since the above inequality holds for $\delta \leq 1$, for $\delta \in ((w_1 + \cdots + w_{\tilde{K}_n+1})^{-1}, 1/2)$ we have $H(\delta, \mathcal{H}_d^w) \geq d \log 2$.

Now for the univariate case we have $(w_1 + \cdots + w_{d+1})^{-1} = (d+1)^{-m} \leq \delta \Rightarrow (d+1) \geq \delta^{-1/m}$ and hence we have

$$H(\delta, \mathcal{H}_d^w) \geq d \log 2 \geq (\delta^{-\frac{1}{m}} - 1) \log 2 = \delta^{-\frac{1}{m}} \left(1 - \delta^{1/m}\right) \log 2 \geq \delta^{-\frac{1}{m}} \left(1 - 2^{-1/m}\right) \log 2.$$

Now for the multivariate case, the argument is slightly different due to presence of zero weights. As before, there is some d' such that $(w_1 + \cdots + w_{q_{d'}-1})^{-1} \leq \delta \leq (w_1 + \cdots + w_{q_{d'}})^{-1}$ and hence $d = q_{d'} - 1$. Note that by assumption we have $\tilde{K}_n = q_{\tilde{K}'} - 1$ and hence $\delta \geq (w_1 + \cdots + w_{q_{\tilde{K}'}})^{-1}$ which implies that $d' \leq \tilde{K}'$ and hence $d \leq \tilde{K}_n$. Finally we have that since $w_1 + \cdots + w_{q_{d'}-1} = w_1 + \cdots + w_{q_{d'-1}} = (d'-1)^m$, therefore $d' - 1 \geq \delta^{-1/m}$. Now we have that

$$\begin{aligned} H(\delta, \mathcal{H}_d^w) &\geq d \log(2) = (q_{d'} - 1) \log(2) = \left\{ \binom{d' + p - 1}{p} - 1 \right\} \log(2) \\ &\geq \left\{ \frac{(d' + p - 1)^p}{p^p} - 1 \right\} \log(2) \geq \left\{ \frac{(\delta^{-\frac{1}{m}} + p)^p}{p^p} - 1 \right\} \log(2) \\ &= \delta^{-\frac{p}{m}} \underbrace{\left\{ \frac{(1 + p\delta^{1/m})^p}{p^p} - \delta^{\frac{p}{m}} \right\}}_{g(\delta)} \log(2) \geq \delta^{-\frac{p}{m}} A \log(2), \end{aligned}$$

where the last inequality follows from the fact that $g(\delta) > 0$ for all $\delta \in (0, 1)$. □

C.3 Details for Proposition 4.2

C.3.1 Univariate case

Firstly, if $f^0(x) = \sum_{k=1}^{\infty} \psi_k(x)\beta_k^0$ then we select $f_n^*(x) = \sum_{k=1}^{\tilde{K}_n} \psi_k(x)\beta_k^0 \in \mathcal{F}_n$. Secondly, we note that for the univariate estimator we have $\Omega(f_n^*|Q_n) = \Omega^{uni}(\boldsymbol{\beta}^0_{1:\tilde{K}_n})$. For brevity we will drop the dependence on $\boldsymbol{\beta}^0$ and denote $\Omega^{uni}(\boldsymbol{\beta}^0_{1:\tilde{K}_n})$ by Ω. Thus we have

$$\lambda_n^2 \Omega(f_n^*|Q_n) = n^{-\frac{2}{2+\alpha}} \Omega^{-\frac{2-\alpha}{2+\alpha}} \Omega = n^{-\frac{2}{2+\alpha}} \Omega^{\frac{2\alpha}{2+\alpha}} = n^{-\frac{2m}{2m+1}} \Omega^{\frac{2}{2m+1}},$$

where we use the fact that for our class $\alpha = 1/m$. For the term $\Omega(\beta^0_{1:\tilde{K}_n})$ we have

$$\Omega(\beta^0_{1:\tilde{K}_n}) = \sum_{k=1}^{\tilde{K}_n} w_k \Big\{ \left(\boldsymbol{\beta}^0_{1:\tilde{K}_n}\right)^\top \frac{\boldsymbol{\Psi}^\top_{k:\tilde{K}_n} \boldsymbol{\Psi}_{k:\tilde{K}_n}}{n} \boldsymbol{\beta}^0_{1:\tilde{K}_n} \Big\}^{1/2} \leq \sqrt{\Lambda_{\max}} \sum_{k=1}^{\tilde{K}_n} w_k \|\boldsymbol{\beta}^0_{1:\tilde{K}_n}\|_2 \leq \sqrt{\Lambda_{\max}} M,$$

for $f^0 \in \mathcal{F}_\infty^M$. For $\mathcal{G}_2^M$, we do not have the above bound and hence we keep the Ω term in the inequality.

For the truncation error we note that

$$\begin{aligned}
\|f^0 - f_n^*\|_n^2 &= \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{k=\tilde{K}_n+1}^{\infty} \psi_k(x_i)\beta_k^0 \right\}^2 \\
&\leq \psi_{\max}^2 \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=\tilde{K}_n+1}^{\infty} \beta_k^0 \right)^2 = \psi_{\max}^2 \left(\sum_{k=\tilde{K}_n+1}^{\infty} \beta_k^0 \right)^2 \\
&= \psi_{\max}^2 \left(\sum_{k=\tilde{K}_n+1}^{\infty} \frac{k^m}{k^m} |\beta_k^0| \right)^2 \\
&\leq \psi_{\max}^2 \left(\sum_{k=\tilde{K}_n+1}^{\infty} k^{2m} (\beta_k^0)^2 \right) \left(\sum_{k=\tilde{K}_n+1}^{\infty} \frac{1}{k^{2m}} \right), \\
&\leq \psi_{\max}^2 \frac{1}{n} \sum_{i=1}^{n} \left(M^2 \sum_{k=\tilde{K}_n+1}^{\infty} \frac{1}{k^{2m}} \right) = \psi_{\max}^2 M^2 \sum_{k=\tilde{K}_n+1}^{\infty} \frac{1}{k^{2m}},
\end{aligned}$$

where the last inequality follows from the proof of Lemma C.4. The result now follows since

$$\sum_{k=\tilde{K}_n+1}^{\infty} k^{-2m} \leq \left\{(2m-1)(\tilde{K}_n+1)^{2m-1}\right\}^{-1} \leq \frac{1}{2m-1}\frac{1}{\tilde{K}_n^{2m-1}}.$$

C.3.2 Multivariate case

Now we assume that $f^0(x) = \sum_{k=1}^{\infty} \psi_k(x^{\nu_k})\beta_k^0$ for $x \in \mathbb{R}^p$ and $\boldsymbol{\nu}_k \in \mathbb{Z}_+^p$. Then we take $f_n^*(\boldsymbol{x}) = \sum_{k=1}^{\tilde{K}_n} \psi_k(\boldsymbol{x}^{\boldsymbol{\nu}_k})\beta_k^0$. Now by the same calculations as in the univariate case, we have

$$\lambda_n^2\Omega(f_n^*|Q_n) = n^{-\frac{2m}{2m+p}}\Omega^{\frac{2p}{2m+p}} \leq n^{-\frac{2m}{2m+p}}[\sqrt{\Lambda_{\max}}M]^{\frac{2p}{2m+p}},$$

since in this case $\alpha = p/m$.

For the truncation error we note that $\tilde{K}_n = q_{\tilde{K}'} - 1$ and hence

$$\begin{aligned}
\|f^0 - f_n^*\|_n^2 &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{k=\tilde{K}_n+1}^{\infty}\psi_k({x_i}^{\nu_k})\beta_k^0\right\}^2 \leq \psi_{\max}^2\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=q_{\tilde{K}'}}^{\infty}\beta_k^0\right)^2 \\
&= \psi_{\max}^2\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{k:\|\boldsymbol{\nu}_k\|_1=\tilde{K}'}\frac{\tilde{K}'^m}{\tilde{K}'^m}|\beta_k^0| + \sum_{k:\|\boldsymbol{\nu}_k\|_1=\tilde{K}'+1}\frac{(\tilde{K}'+1)^m}{(\tilde{K}'+1)^m}|\beta_k^0| + \cdots\right\}^2 \\
&= \psi_{\max}^2\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{R=\tilde{K}'}^{\infty}\frac{R^m}{R^m}\sum_{k:\|\boldsymbol{\nu}_k\|_1=R}|\beta_k^0|\right)^2 \\
&= \psi_{\max}^2\frac{1}{n}\sum_{i=1}^{n}\left\{\left(\sum_{R=\tilde{K}'}^{\infty}\frac{1}{R^{2m}}\right)^{1/2}\left(\sum_{R=K'}^{\infty}R^m\sum_{k:\|\boldsymbol{\nu}_k\|_1=R}|\beta_k^0|\right)^{1/2}\right\}^2 \\
&\leq \psi_{\max}^2 M^2\sum_{R=\tilde{K}'}^{\infty}\frac{1}{R^{2m}} \leq \frac{M^2\psi_{\max}^2}{2m-1}\frac{1}{(\tilde{K}')^{2m-1}}.
\end{aligned}$$

C.4 Proof of Theorem 4.4

C.4.1 Initial results

Recall that $(\widehat{f}_j)_{j=1}^p \in \mathcal{F}$ where $\mathcal{F}$ is some arbitrary univariate function class . We denote the functions $\widehat{f}(\boldsymbol{x}) = \sum_{j=1}^p \widehat{f}_j(x_j)$ and $f^0(\boldsymbol{x}) = \sum_{j=1}^p f_j^0(x_j)$ for $\boldsymbol{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$. For the proof of Theorem 4.4, λ_n and ρ_n are functions of n but for convenience we will simply write λ, ρ. Throughout this proof, instead of the smoothness level m, we will use $\alpha = 1/m$. Thus the entropy condition is $H(\delta, \{f \in \mathcal{F} : \Upsilon(f) \leq 1\}, Q_n) \leq A_0 \delta^{-\alpha}$, for $\alpha \in (0, 2)$, and so forth.

We begin the proof of Theorem 4.4 with a basic inequality.

Lemma C.1 (Basic inequality)**.** *For any function $f^* = \sum_{j=1}^p f_j^*$, where $f_j^* \in \mathcal{F}$ and, the solution $\widehat{f}$ of (4.17), we have the following basic inequality*

$$\frac{1}{2}\|\widehat{f} - f^0\|_n^2 + \lambda I_p(\widehat{f}) \leq |\langle \boldsymbol{\varepsilon}, \widehat{f} - f^* \rangle_n| + \lambda I_p(f^*) + |\bar{\varepsilon}| \sum_{j=1}^p \|\widehat{f}_j - f_j^*\|_n + \frac{1}{2}\|f^* - f^0\|_n^2,$$

where $\langle \boldsymbol{\varepsilon}, f \rangle_n = \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)$, $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^n \varepsilon_i$ and $I_p(f) = \sum_{j=1}^p I(f_j) = \sum_{j=1}^p \|f_j\|_n + \lambda \Upsilon(f_j)$ for an additive function f.

Proof. We have

$$\frac{1}{2n}\sum_{i=1}^n \left\{Y_i - \bar{Y} - \widehat{f}(\boldsymbol{x}_i)\right\}^2 + \lambda I_p(\widehat{f}) \leq \frac{1}{2n}\sum_{i=1}^n \left\{Y_i - \bar{Y} - f^*(\boldsymbol{x}_i)\right\}^2 + \lambda I_p(f_j^*),$$

$$\Leftrightarrow \frac{1}{2n}\sum_{i=1}^n \left\{\varepsilon_i + c^0 - \bar{Y} - (\widehat{f} - f^0)(\boldsymbol{x}_i)\right\}^2 + \lambda I_p(\widehat{f}) \leq \frac{1}{2n}\sum_{i=1}^n \left\{\varepsilon_i + c^0 - \bar{Y} - (f^* - f^0)(\boldsymbol{x}_i)\right\}^2 + \lambda I_p(f_j^*)$$

$$
\begin{aligned}
&\Rightarrow \frac{1}{2n}\sum_{i=1}^{n}\left(\varepsilon_i + c^0 - \bar{Y}\right)^2 + (\widehat{f} - f^0)^2(\boldsymbol{x_i}) - 2(\varepsilon_i + c^0 - \bar{Y})(\widehat{f} - f^0)(\boldsymbol{x_i}) + \lambda I_p(\widehat{f}) \\
&\leq \frac{1}{2n}\sum_{i=1}^{n}\left(\varepsilon_i + c^0 - \bar{Y}\right)^2 + (f^* - f^0)^2(\boldsymbol{x_i}) - 2(\varepsilon_i + c^0 - \bar{Y})(f^* - f^0)(\boldsymbol{x_i}) + \lambda I_p(f^*) \\
&\Rightarrow \frac{1}{2}\|\widehat{f} - f^0\|_n^2 - \langle \boldsymbol{\varepsilon} + c^0 - \bar{Y},\, \widehat{f} - f^0\rangle_n + \lambda I_p(\widehat{f}) \\
&\leq \frac{1}{2}\|f^* - f^0\|_n^2 - \langle \boldsymbol{\varepsilon} + c^0 - \bar{Y},\, f^* - \widehat{f} + \widehat{f} - f^0\rangle_n + \lambda I_p(f^*) \\
&\Rightarrow \frac{1}{2}\|\widehat{f} - f^0\|_n^2 - \langle \boldsymbol{\varepsilon} + c^0 - \bar{Y},\, \widehat{f} - f^0\rangle_n + \lambda I_p(\widehat{f}) \\
&\leq \frac{1}{2}\|f^* - f^0\|_n^2 - \langle \boldsymbol{\varepsilon} + c^0 - \bar{Y},\, f^* - \widehat{f}\rangle_n - \langle \boldsymbol{\varepsilon} + c^0 - \bar{Y},\, \widehat{f} - f^0\rangle_n + \lambda I_p(f^*),
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\frac{1}{2}\|\widehat{f} - f^0\|_n^2 + \lambda I_p(\widehat{f}) \leq \frac{1}{2}\|f^* - f^0\|_n^2 - \langle \boldsymbol{\varepsilon} + c^0 - \bar{Y},\, f^* - \widehat{f}\rangle_n + \lambda I_p(f^*) \\
\Rightarrow &\frac{1}{2}\|\widehat{f} - f^0\|_n^2 + \lambda I_p(\widehat{f}) \leq |\langle \boldsymbol{\varepsilon},\, \widehat{f} - f^*\rangle_n| + \sum_{j=1}^{p}\langle c^0 - \bar{Y}, \widehat{f}_j - f_j^*\rangle_n + \lambda I_p(f^*) + \frac{1}{2}\|f^* - f^0\|_n^2 \\
\Rightarrow &\frac{1}{2}\|\widehat{f} - f^0\|_n^2 + \lambda I_p(\widehat{f}) \leq |\langle \boldsymbol{\varepsilon},\, \widehat{f} - f^*\rangle_n| + |c^0 - \bar{Y}|\sum_{j=1}^{p}\|\widehat{f}_j - f_j^*\|_n + \lambda I_p(f^*) + \frac{1}{2}\|f^* - f^0\|_n^2.
\end{aligned}
$$

Now for the second term note that:

$$
|c^0 - \bar{Y}| = \left|\frac{1}{n}\sum_{i=1}^{n}(c^0 - Y_i)\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\left\{c^0 - c^0 - \sum_{j=1}^{p} f_j^0(x_{i,j}) - \varepsilon_i\right\}\right| = |\bar{\varepsilon}|.
$$

Which leads us to

$$
\frac{1}{2}\|\widehat{f} - f^0\|_n^2 + \lambda I_p(\widehat{f}) \leq |\langle \boldsymbol{\varepsilon},\, \widehat{f} - f^*\rangle_n| + \lambda I_p(f^*) + |\bar{\varepsilon}|\sum_{j=1}^{p}\|\widehat{f}_j - f_j^*\|_n + \frac{1}{2}\|f^* - f^0\|_n^2.
$$

□

Lemma C.2 (Bounding the term $|\bar{\varepsilon}|$)**.** *For* $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots \varepsilon_n)^\top$ *such that* $\mathbb{E}(\varepsilon_i) = 0$ *and*

$$L^2 \left\{ \mathbb{E}\left(e^{\varepsilon_i^2 / L^2} \right) - 1 \right\} \leq \sigma_0^2 \ ,$$

for all $\kappa > 0$ *and*

$$\rho = \kappa \max \left\{ n^{-\frac{1}{2+\alpha}}, \left(\frac{\log p}{n} \right)^{1/2} \right\},$$

we have that with probability at least $1 - 2\exp\left(-n\rho^2/c_1\right)$,

$$|\bar{\varepsilon}| \leq \rho,$$

for a constant c_1 *that depends on* L *and* σ_0.

Proof. By Lemma 8·2 of van de Geer (2000) (with $\gamma_n = 1_n/n$) we have for all $t > 0$

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \right| \geq t \right) \leq 2\exp\left\{ -\frac{nt^2}{8(L^2 + \sigma_0^2)} \right\}.$$

The result follows by setting $t = \rho$. □

Lemma C.3 (Bounding the term $|\langle \boldsymbol{\varepsilon}, \widehat{f} - f^* \rangle_n|$)**.** *For* $\lambda \geq 4\rho$ *where*

$$\rho = \kappa \max \left\{ n^{-\frac{1}{2+\alpha}}, \left(\frac{\log p}{n} \right)^{1/2} \right\},$$

for some constant κ*, if*

$$H(\delta, \{f \in \mathcal{F} : \Upsilon(f) \leq 1\}, Q_n) \leq A_0 \delta^{-\alpha},$$

we then have with probability at least $1 - c_2 \exp\left(-c_3 n\rho^2\right)$

$$|\langle \boldsymbol{\varepsilon}, \widehat{f}_j - f_j^* \rangle_n| \leq \rho \|\widehat{f}_j - f_j^*\|_n + \rho\lambda\Upsilon(\widehat{f}_j - f_j^*),$$

for all $j = 1, \ldots, p$ and positive constants c_2 and c_3.

Proof. Firstly, for $\mathcal{F}_0 = \{f \in \mathcal{F} : \Upsilon(f) \leq 1\}$ we have by assumption a δ cover $f_1, \ldots, f_N$ such that for all $f \in \mathcal{F}_0$ we have $\min_{j\in\{1,\ldots,N\}} \|f_j - f\|_n \leq \delta$. Now we are interested in the set $\mathcal{F}_{0,\lambda} = \{f \in \mathcal{F} : \lambda\Upsilon(f) \leq 1\}$. Firstly, for a function $f \in \mathcal{F}_{0,\lambda}$,

$$\min_{j\in\{1,\ldots,N\}} \|f - f_j/\lambda\|_n = \min_{j\in\{1,\ldots,N\}} \frac{1}{\lambda}\|\lambda f - f_j\|_n \leq \frac{\delta}{\lambda},$$

because $\Upsilon(\lambda f) = \lambda\Upsilon(f) \leq 1 \Rightarrow \lambda f \in \mathcal{F}_0$. This means that the set $\{f_1/\lambda, \ldots, f_N/\lambda\}$ is a δ/λ cover of the set $\mathcal{F}_{0,\lambda}$.

This implies that $H(\delta, \mathcal{F}_0, Q_n) \leq A_0\delta^{-\alpha} \Rightarrow H(\delta/\lambda, \mathcal{F}_{0,\lambda}, Q_n) \leq A_0\delta^{-\alpha}$ or equivalently $H(\delta, \mathcal{F}_{0,\lambda}, Q_n) \leq A_0(\delta\lambda)^{-\alpha}$. Finally, since $\{f \in \mathcal{F} : I(f) \leq 1\} \subset \{f \in \mathcal{F} : \Upsilon(f) \leq \lambda^{-1}\}$ we have

$$H(\delta, \{f \in \mathcal{F} : I(f) \leq 1\}, Q_n) \leq A_0(\delta\lambda)^{-\alpha}.$$

The same entropy bound holds for the class

$$\widetilde{\mathcal{F}} = \left\{ \frac{f_j - f_j^*}{\|f_j - f_j^*\|_n + \lambda\Omega(f_j - f_j^*)} : f_j \in \mathcal{F} \right\}, \tag{C.2}$$

and we can now apply Corollary 8.3 of van de Geer (2000) by noting that

$$\int_0^1 H^{1/2}(u, \widetilde{\mathcal{F}}, Q_n)\, du \leq \widetilde{A}_0\lambda^{-\alpha/2},$$

for some constant $\widetilde{A}_0 = \widetilde{A}_0(A_0)$. For some $c_2 = c_2(L, \sigma_0)$ and all $\delta \geq 2c_2\widetilde{A}_0\lambda^{-\alpha/2}n^{-1/2}$ we have

$$\text{pr}\left(\sup_{f_j\in\mathcal{F}} \frac{\left|\langle \boldsymbol{\varepsilon}, f_j - f_j^*\rangle_n\right|}{\|f_j - f_j^*\|_n + \lambda\Omega(f_j - f_j^*)} \geq \delta \right) \leq c_2 \exp\left(-\frac{n\delta^2}{4c_2^2}\right). \tag{C.3}$$

Since $\lambda \geq \rho$ we note that $2c_2\widetilde{A}_0\lambda^{-\alpha/2}n^{-1/2} \leq 2c_2\widetilde{A}_0\rho^{-\alpha/2}n^{-1/2}$ and that

$$2c_2\widetilde{A}_0\rho^{-\alpha/2}n^{-1/2} \leq \rho \Leftrightarrow \rho \geq \left(2c_2\widetilde{A}_0\right)^{\frac{2}{2+\alpha}} n^{-\frac{1}{2+\alpha}}.$$

Which holds by definition since $\rho = \kappa \max\left\{(\log p/n)^{1/2}, n^{-1/(2+\alpha)}\right\} \geq \kappa n^{-1/(2+\alpha)}$ and κ is sufficiently large (any $\kappa \geq \left(2c_2\widetilde{A}_0\right)^{2/(2+\alpha)}$ would suffice). Therefore, we can take $\delta = \rho$ in (C.3) along with a union bound to obtain

$$\begin{aligned}
\mathbb{P}\left(\max_{j=1,\ldots,p}\sup_{f_j\in\mathcal{F}}\frac{\left|\langle \boldsymbol{\varepsilon}, f_j - f_j^*\rangle_n\right|}{\|f_j - f_j^*\|_n + \lambda\Omega(f_j - f_j^*)} \geq \rho\right) &\leq pc_2\exp\left(-\frac{n\rho^2}{4c_2^2}\right) \\
&= c_2\exp\left\{-n\rho^2\left(\frac{1}{4c_2^2} - \frac{\log p}{n\rho^2}\right)\right\} \\
&\leq c_2\exp\left(-n\rho^2 c_3\right),
\end{aligned}$$

for some positive constant $c_3 = c_3(c_2, \widetilde{A}_0)$.

Finally, we show that $c_3 > 0$. This follows from the fact that $1/(4c_2^2) - \log p/(n\rho^2) > 0 \Leftrightarrow n\rho^2 > 4c_2^2\log p$. This holds since $n\rho^2 \geq \kappa^2\log p$ for κ sufficiently large. Thus, we have with probability at least $1 - c_2\exp\left(c_3 n\rho^2\right)$ for all $j = 1, \ldots, p$

$$|\langle \boldsymbol{\varepsilon}, \widehat{f}_j - f_j^*\rangle_n| \equiv |\langle \boldsymbol{\varepsilon}, \widehat{\Delta}_j\rangle_n| \leq \rho\|\widehat{\Delta}_j\|_n + \rho\lambda\Upsilon(\widehat{\Delta}_j)\ .$$

□

C.4.2 Using the active set

So far we have shown that, for $\lambda \geq 4\rho$, with probability at least $1 - 2\exp(-n\rho/c_1) - c_2\exp(-c_3 n\rho^2)$, the following inequality holds

$$\begin{aligned}
\|\widehat{f} - f^0\|_n^2 + 2\lambda\sum_{j=1}^p I(\widehat{f}_j) &\leq 2|\langle\varepsilon, \widehat{f} - f^*\rangle_n| + 2|\bar{\varepsilon}|\sum_{j=1}^p \|\widehat{\Delta}_j\|_n + 2\lambda\sum_{j=1}^p I(f_j^*) + \|f^* - f^0\|_n^2 \\
&\leq \left\{\sum_{j=1}^p 2\rho\|\widehat{\Delta}_j\|_n + 2\rho\lambda\Upsilon(\widehat{\Delta}_j)\right\} + \left(2\rho\sum_{j=1}^p \|\widehat{\Delta}_j\|_n\right) \\
&\quad + \left\{2\lambda\sum_{j=1}^p I(f_j^*)\right\} + \|f^* - f^0\|_n^2 \\
\Rightarrow \|\widehat{f} - f^0\|_n^2 + 2\lambda\sum_{j=1}^p I(\widehat{f}_j) &\leq \sum_{j=1}^p \left\{\lambda\|\widehat{\Delta}_j\|_n + \frac{\lambda^2}{2}\Upsilon(\widehat{\Delta}_j) + 2\lambda\|f_j^*\|_n + 2\lambda^2\Upsilon(f_j^*)\right\} + \|f^* - f^0\|_n^2.
\end{aligned}$$

For notational convenience we will exclude the $\|f^* - f^0\|_n^2$ term in the following manipulations. If $\mathcal{S}_*$ is the active set then we have on the right hand side,

$$\begin{aligned}
\text{RHS} &= \lambda\sum_{j\in\mathcal{S}_*}\left\{\|\widehat{\Delta}_j\|_n + \frac{\lambda}{2}\Upsilon(\widehat{\Delta}_j) + 2\|f_j^*\|_n + 2\lambda\Upsilon(f_j^*)\right\} + \lambda\sum_{j\in\mathcal{S}_*^c}\left\{\|\widehat{f}_j\|_n + \frac{\lambda}{2}\Upsilon(\widehat{f}_j)\right\} \\
&\leq \lambda\sum_{j\in\mathcal{S}_*}\left\{\|\widehat{\Delta}_j\|_n + \frac{\lambda}{2}\Upsilon(\widehat{\Delta}_j) + 2\|\widehat{\Delta}_j\|_n + 2\|\widehat{f}_j\|_n + 2\lambda\Upsilon(f_j^*)\right\} + \lambda\sum_{j\in\mathcal{S}_*^c}\left\{\|\widehat{f}_j\|_n + \frac{\lambda}{2}\Upsilon(\widehat{f}_j)\right\} \\
&= 3\sum_{j\in\mathcal{S}_*}\lambda\|\widehat{\Delta}_j\|_n + 2\sum_{j\in\mathcal{S}_*}\lambda^2\Upsilon(f_j^*) + \sum_{j\in\mathcal{S}_*^c}\lambda\|\widehat{f}_j\| + \frac{1}{2}\sum_{j\in\mathcal{S}_*^c}\lambda^2\Upsilon(\widehat{f}_j) + 2\sum_{j\in\mathcal{S}_*}\lambda\|\widehat{f}_j\|_n + \frac{1}{2}\sum_{j\in\mathcal{S}_*}\lambda^2\Upsilon(\widehat{\Delta}_j),
\end{aligned}$$

where the inequality holds by the decomposition $\|f_j^*\|_n = \|f_j^* - \widehat{f}_j + \widehat{f}_j\|_n \leq \|\widehat{\Delta}_j\|_n + \|\widehat{f}_j\|_n$. On the left hand side we have

$$\begin{aligned}
\text{LHS} &= \|\widehat{f} - f^0\|_n^2 + 2\lambda\sum_{j\in\mathcal{S}_*}\left\{\|\widehat{f}_j\|_n + \lambda\Upsilon(\widehat{f}_j)\right\} + 2\lambda\sum_{j\in\mathcal{S}_*^c}\left\{\|\widehat{f}_j\|_n + \lambda\Upsilon(\widehat{f}_j)\right\} \\
&\geq \|\widehat{f} - f^0\|_n^2 + 2\lambda\sum_{j\in\mathcal{S}_*}\left\{\|\widehat{f}_j\|_n + \lambda\Upsilon(\widehat{\Delta}_j) - \lambda\Upsilon(f_j^*)\right\} + 2\lambda\sum_{j\in\mathcal{S}_*^c}\left\{\|\widehat{f}_j\|_n + \lambda\Upsilon(\widehat{f}_j)\right\},
\end{aligned}$$

where the inequality follows from the triangle inequality $\Upsilon(\widehat{f}_j) + \Upsilon(f_j^*) \geq \Upsilon(\widehat{\Delta}_j)$ since $\Upsilon(\cdot)$ is a semi-norm. By re-arranging the terms we obtain the inequality

$$\|\widehat{f}-f^0\|_n^2+\lambda \sum_{j\in\mathcal{S}_*^c}\left\{\|\widehat{f}_j\|_n + \frac{3\lambda}{2}\Upsilon(\widehat{f}_j)\right\}+\frac{3\lambda^2}{2}\sum_{j\in\mathcal{S}_*}\Upsilon(\widehat{\Delta}_j) \leq 3\lambda\sum_{j\in\mathcal{S}_*}\|\widehat{\Delta}_j\|_n+4\lambda^2\sum_{j\in\mathcal{S}_*}\Upsilon(f_j^*)+\|f^*-f^0\|_n^2$$

which implies that

$$\|\widehat{f}-f^0\|_n^2+\lambda\sum_{j\in\mathcal{S}_*^c}\|\widehat{\Delta}_j\|_n+\frac{3\lambda^2}{2}\sum_{j=1}^{p}\Upsilon(\widehat{\Delta}_j) \leq 3\lambda\sum_{j\in\mathcal{S}_*}\|\widehat{\Delta}_j\|_n+4\lambda^2\sum_{j\in\mathcal{S}_*}\Upsilon(f_j^*)+\|f^*-f^0\|_n^2.$$

This implies the slow rates for convergence for $\lambda \geq 4\rho$ and $s = |\mathcal{S}_*|$

$$\frac{1}{2}\|\widehat{f}-f^0\|_n^2+ \leq s\lambda\left\{3R+2\lambda\sum_{j\in\mathcal{S}_*}\Upsilon(f_j^*)/s\right\}+\frac{1}{2}\|f^*-f^0\|_n^2.$$

This completes the proof of Theorem 4.4. In the next section we prove the oracle inequality with fast rates via the compatibility condition.

C.4.3 Using the compatibility condition

Recall the compatibility condition for $f = \sum_{j=1}^p f_j$, whenever

$$\sum_{j\in\mathcal{S}_*^c}\|f_j\|_n \leq 4\sum_{j\in\mathcal{S}_*}\|f_j\|_n, \tag{C.4}$$

then we have

$$\sum_{j\in\mathcal{S}_*}\|f_j\|_n \leq s^{1/2}\|f\|_n/\phi(\mathcal{S}_*).$$

Once we assume the compatibility condition we can prove Corollary 4.3 by considering the following two cases.

Case 1: $\lambda \sum_{j \in \mathcal{S}_*} \|\widehat{\Delta}_j\|_n \geq 4\lambda^2 \sum_{j \in \mathcal{S}_*} \Upsilon(f_j^*)$ in which case we have

$$\|\widehat{f} - f^0\|_n^2 + \lambda \sum_{j \in \mathcal{S}_*^c} \|\widehat{\Delta}_j\|_n + \frac{3\lambda^2}{2} \sum_{j=1}^{p} \Upsilon(\widehat{\Delta}_j) \leq 4\lambda \sum_{j \in \mathcal{S}_*} \|\widehat{\Delta}_j\|_n + \|f^* - f^0\|_n^2 \ ,$$

hence for the function $\widehat{f} - f^* = \sum_{j=1}^{p} \widehat{\Delta}_j$ (C.4) holds and hence by the compatibility condition we have

$$\begin{aligned}
\|\widehat{f} - f^0\|_n^2 + \lambda \sum_{j \in \mathcal{S}_*^c} \|\widehat{\Delta}_j\|_n + \frac{3\lambda^2}{2} \sum_{j=1}^{p} \Upsilon(\widehat{\Delta}_j) &\leq \frac{4\lambda s^{1/2}}{\phi(\mathcal{S}_*)} \|\widehat{f} - f^*\|_n + \|f^* - f^0\|_n^2 \\
&\leq \frac{4\lambda s^{1/2}}{\phi(\mathcal{S}_*)} \|\widehat{f} - f^0\|_n + \frac{4\lambda s^{1/2}}{\phi(\mathcal{S}_*)} \|f^* - f^0\|_n + \|f^* - f^0\|_n^2 \\
&\leq 2 \left\{ \frac{2\lambda (2s)^{1/2}}{\phi(\mathcal{S}_*)} \right\} \left(\frac{\|\widehat{f} - f^0\|_n}{2^{1/2}} \right) + 2 \left\{ \frac{2\lambda s^{1/2}}{\phi(\mathcal{S}_*)} \right\} \left(\|f^* - f^0\|_n \right) + \|f^* - f^0\|_n^2 \\
&\leq \frac{4\lambda^2 (2s)}{\phi^2(\mathcal{S}_*)} + \frac{\|\widehat{f} - f^0\|_n^2}{2} + \frac{4\lambda^2 s}{\phi^2(\mathcal{S}_*)} + \|f^* - f^0\|_n^2 + \|f^* - f^0\|_n^2 \\
&\leq \frac{12\lambda^2 s}{\phi^2(\mathcal{S}_*)} + \frac{\|\widehat{f} - f^0\|_n^2}{2} + 2\|f^* - f^0\|_n^2,
\end{aligned}$$

where we use the inequality $2ab \leq a^2 + b^2$ and this implies that

$$\frac{1}{2} \|\widehat{f} - f^0\|_n^2 + \lambda \sum_{j \in \mathcal{S}_*^c} \|\widehat{\Delta}_j\|_n + \frac{3\lambda^2}{2} \sum_{j=1}^{p} \Upsilon(\widehat{\Delta}_j) \leq \frac{12 s \lambda^2}{\phi^2(\mathcal{S}_*)} + 2\|f^* - f^0\|_n^2.$$

Case 2: $\lambda \sum_{j \in \mathcal{S}_*} \|\widehat{\Delta}_j\|_n \leq 4\lambda^2 \sum_{j \in \mathcal{S}_*} \Upsilon(f_j^*)$ in which case we have

$$\begin{aligned}
\|\widehat{f} - f^0\|_n^2 + \lambda \sum_{j \in \mathcal{S}_*^c} \|\widehat{\Delta}_j\|_n + \frac{3\lambda^2}{2} \sum_{j=1}^{p} \Upsilon(\widehat{\Delta}_j) &\leq 16\lambda^2 \sum_{j \in \mathcal{S}_*} \Upsilon(f_j^*) + \|f^* - f^0\|_n^2 \\
&\leq 16 s \lambda^2 \sum_{j \in \mathcal{S}_*} \Upsilon(f_j^*)/s + \|f^* - f^0\|_n^2,
\end{aligned}$$

which implies

$$\frac{1}{2}\|\widehat{f}-f^0\|_n^2+\lambda\sum_{j\in\mathcal{S}_*^c}\|\widehat{\Delta}_j\|_n+\frac{3\lambda^2}{2}\sum_{j\in\mathcal{S}_*}\Upsilon(\widehat{\Delta}_j)\leq 16s\lambda^2\sum_{j\in\mathcal{S}_*}\Upsilon(f_j^*)/s+2\|f^*-f^0\|_n^2.$$

C.5 Constraining the proposed penalty region

Recall the following definitions

$$\mathcal{H}_{K_n}^w=\left\{\boldsymbol{\beta}\in\mathbb{R}^{K_n}:\sum_{k=1}^{K_n}w_k\|\boldsymbol{\beta}_{k:K_n}\|_2\leq 1\right\},\tag{C.5}$$

$$E_{K_n}^w=\left\{\boldsymbol{\beta}\in\mathbb{R}^{K_n}:\sum_{k=1}^{K_n}\beta_k^2\left(w_1+\ldots+w_k\right)^2\leq 1\right\}.\tag{C.6}$$

Lemma C.4. *For the regions $\mathcal{H}_{K_n}^w$ and $E_{K_n}^w$ as defined in (C.5) and (C.6), respectively, we have $\mathcal{H}_{K_n}^w\subseteq E_{K_n}^w$ for all $n\geq 1$ and non-negative weights.*

Proof. It is sufficient to show $\sum_{k=1}^{K_n}\beta_k^2\left(w_1+\ldots+w_k\right)^2\leq\left(\sum_{k=1}^{K_n}w_k\|\boldsymbol{\beta}_{k:K_n}\|_2\right)^2$. We now have

$$\begin{aligned}
\left(\sum_{k=1}^{K_n}w_k\|\boldsymbol{\beta}_{k:K_n}\|_2\right)^2&=\sum_{m=1}^{K_n}w_m^2\|\boldsymbol{\beta}_{m:K_n}\|_2^2+2\sum_{m<k}w_kw_m\|\boldsymbol{\beta}_{m:K_n}\|_2\|\boldsymbol{\beta}_{k:K_n}\|_2\\
&=\sum_{m=1}^{K_n}w_m^2\sum_{l=m}^{K_n}\beta_l^2+2\sum_{m<k}w_kw_m\|\boldsymbol{\beta}_{k:K_n}\|_2^2\underbrace{\frac{\|\boldsymbol{\beta}_{m:K_n}\|}{\|\boldsymbol{\beta}_{k:K_n}\|}}_{\geq 1}\\
&\geq\sum_{l=1}^{K_n}\sum_{m=1}^{K_n}w_m^2\beta_l^2\mathbf{1}(l\geq m)+2\sum_{k=2}^{K_n}\sum_{m=1}^{k-1}w_kw_m\sum_{l=1}^{K_n}\beta_l^2\mathbf{1}(l\geq k)\\
&=\sum_{l=1}^{K_n}\beta_l^2\sum_{m=1}^{l}w_m^2+2\sum_{l=1}^{K_n}\beta_l^2\sum_{k=2}^{K_n}\sum_{m=1}^{k-1}w_kw_m\mathbf{1}(l\geq k)\\
&=\sum_{l=1}^{K_n}\beta_l^2\left(\sum_{m=1}^{l}w_m^2+2\sum_{k=2}^{l}\sum_{m=1}^{k-1}w_kw_m\right)=\sum_{l=1}^{K_n}\beta_l^2\left(\sum_{m=1}^{l}w_m\right)^2.
\end{aligned}$$

□

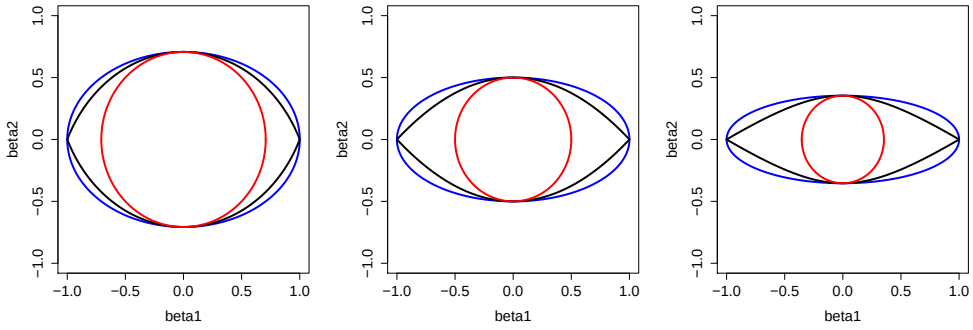


Figure C.1: Demonstration of Lemma C.4 and Lemma C.5 for the special case of $w_k = a_{j,m} = k^m - (k-1)^m$ and $K_n = 2$. We have in blue (——) the region E_2^w, C_2^w in red (——) and H_2^w in black (——). From left to right we have the plots for $m = 0.5$, 1 and 1.5.

Lemma C.5. *For the region $\mathcal{H}_{K_n}^w$ as defined in (C.5), we have the inclusion $\mathbb{B}_{K_n}^w \subset \mathcal{H}_{K_n}^w$ where*

$$\mathbb{B}_{K_n}^w = \left\{ \boldsymbol{\beta} \in \mathbb{R}^{K_n} : \sum_{k=1}^{K_n} \beta_k^2 \leq (w_1 + \ldots + a_{K_n})^{-2} \right\}. \tag{C.7}$$

Proof. Let $\beta \in \mathbb{B}_{K_n}^w$ and for brevity we denote $\|\cdot\| = \|\cdot\|_2$. Then for $\beta \in \mathbb{B}_{K_n}^w$

$$\begin{aligned} 1 &\geq \|\boldsymbol{\beta}\| \left(w_1 + w_2 + \ldots + w_{K_n}\right) \\ &\geq \|\boldsymbol{\beta}\| \left(w_1 \frac{\|\boldsymbol{\beta}_{1:K_n}\|}{\|\boldsymbol{\beta}_{1:K_n}\|} + w_2 \frac{\|\boldsymbol{\beta}_{2:K_n}\|}{\|\boldsymbol{\beta}_{1:K_n}\|} + \ldots + w_{K_n} \frac{\|\boldsymbol{\beta}_{K_n:K_n}\|}{\|\boldsymbol{\beta}_{1:K_n}\|} \right)^2 \\ &= w_1 \|\boldsymbol{\beta}_{1:K_n}\| + w_2 \|\boldsymbol{\beta}_{2:K_n}\| + \ldots + w_{K_n} \|\boldsymbol{\beta}_{K_n:K_n}\|, \end{aligned}$$

which implies that $\beta \in \mathcal{H}_{K_n}^w$. □

In Figure C.1, we demonstrate the above two lemma's for $K_n = 2$ for the special case of $w_k = k^m - (k-1)^m$.

We now present the proof of the claim made in 4.4.3 regarding the relationship of our

function class to weighted L_p spaces. Recall the definition of our class $\mathcal{F}_\infty$:

$$\mathcal{F}_\infty = \left\{ f_{\boldsymbol{\beta}}(x) = \sum_{k=1}^{\infty} \psi_k(x)\beta_k \ : \int \psi_k \psi_l \, dQ = 0 \text{ for } k \neq l, \int \psi_k^2 \, dQ = 1 \right\}.$$

Lemma C.6. *For the hierarchical function class,*

$$\mathcal{F}_\infty^M = \{ f_{\boldsymbol{\beta}} \in \mathcal{F}_\infty : \sum_{k=1}^{\infty} \{k^m - (k-1)^m\} \|\boldsymbol{\beta}_{k:\infty}\|_2 \leq M \},$$

and the weighted L_p class

$$\mathcal{G}_q^M = \{ f_{\boldsymbol{\beta}} \in \mathcal{F}_\infty : \sum_{k=1}^{\infty} (k^m |\beta_k|)^q \leq M^q \},$$

we have the following relationship:

$$\mathcal{G}_1^M \subseteq \mathcal{F}_\infty^M \subseteq \mathcal{G}_2^M. \tag{C.8}$$

Proof. The inclusion $\mathcal{F}_\infty^M \subseteq \mathcal{G}_2^M$ follows from the proof of Lemma C.4 above. For the first inclusion it suffices to show that $\sum_{k=1}^{\infty} \{k^m - (k-1)^m\} \|\boldsymbol{\beta}_{k:\infty}\|_2 \leq \sum_{k=1}^{\infty} k^m |\beta_k|$. This follows

from the fact that the ℓ_q norm is decreasing in q.

$$\begin{aligned}
\sum_{k=1}^{\infty}\{k^m-(k-1)^m\}\|\boldsymbol{\beta}_{k:\infty}\|_2 &\leq \sum_{k=1}^{\infty}\{k^m-(k-1)^m\}\|\boldsymbol{\beta}_{k:\infty}\|_1 \\
&= \sum_{k=1}^{\infty}\{k^m-(k-1)^m\}\sum_{j=k}^{\infty}|\beta_j| \\
&= \sum_{k=1}^{\infty}\sum_{j=1}^{\infty}\{k^m-(k-1)^m\}|\beta_j|\mathbf{1}(j>k) \\
&= \sum_{j=1}^{\infty}|\beta_j|\sum_{k=1}^{j}\{k^m-(k-1)^m\} \\
&= \sum_{j=1}^{\infty}|\beta_j|j^m.
\end{aligned}$$

□

C.6 Some entropy results for ellipsoids

C.6.1 An upper bound

In this section we establish some entropy results for the ellipsoid (C.6) and the circle (C.7) which will allow us to establish entropy rates for the penalty region $\mathcal{H}^w_{K_n}$.

Since K_n can potentially be ∞, or arbitrarily large, we need a way to handle this dimension. It turns out that this can be done using a simple argument which we demonstrate in the following theorem.

Theorem C.1. *(Dumer, 2006) For any $\theta \in (0, 1/2)$, the δ-entropy of the ellipsoid $E^w_{K_n}$ satisfies the following inequality*

$$H(\delta,\ E^w_{K_n}) \leq \sum_{k=1}^{d-1}\log\left(\frac{1}{\delta\sum_{l=1}^{k}w_l}\right)+\mu_\theta\log(3/\theta)\ ,$$

where $\mu_\theta \leq K_n$ is the largest integer such that $w_1+\cdots+w_{\mu_\theta} < \left\{(1-\theta)^{1/2}\delta\right\}^{-1}$ and $d \leq K_n+1$ is the largest integer such that $w_1+\cdots+w_{d-1} \leq \delta^{-1}$. If $\delta^{-1} \leq w_1$ then $H(\delta, E^w_{K_n}) = 0$ holds

trivially.

Corollary C.1.1 (Sobolev Ellipsoids)**.** *For Theorem C.1, let* $w_k = k^m - (k-1)^m$*. Then we have the following upper bound:*

$$H(\delta,\, E^w_{K_n}) \leq U_E \delta^{-1/m},$$

for some constant U_E *which only depends on* m *and* θ*.*

Proof. Firstly, we note that with this definition of w_k, we can let $K_n = \infty$. Thus if we can show that $H(\delta, E^w_\infty) \leq U\delta^{-1/m}$ then the result follows since $E^w_{K_n} \subset E^w_\infty$ for all $K_n < \infty$.

Now we have $w_1 + \cdots + w_{\mu_\theta} = \mu_\theta^m$, hence

$$\mu_\theta^m < \frac{\delta^{-1}}{(1-\theta)^{1/2}} \Rightarrow \mu_\theta \log(3/\theta) < \log(3/\theta) \left\{ \frac{\delta^{-1}}{(1-\theta)^{1/2}} \right\}^{1/m} = U_1 \delta^{-\frac{1}{m}}.$$

Now for the second part we use the fact that $w_1 + \cdots + w_{d-1} = (d-1)^m \leq \delta^{-1} < d^m$ and we obtain

$$\begin{aligned}
\sum_{j=1}^{d-1} \log\left(\frac{1}{\delta j^m}\right) &= (d-1)\log(\delta^{-1}) + \log\left[\frac{1}{\{(d-1)!\}^m}\right] \leq \delta^{-1/m}\log(\delta^{-1}) - m\log\{(d-1)!\} \\
&\leq \delta^{-1/m}\left[\log(\delta^{-1}) - m\delta^{1/m}\log\{(d-1)!\}\right] \leq \delta^{-1/m}\left[\log(d^m) - m\delta^{1/m}\log\{(d-1)!\}\right] \\
&\leq \delta^{-1/m} m\left[\log(d) - d^{-1}\log\{(d-1)!\}\right].
\end{aligned}$$

Now by sterling's inequality we have for all $d \in \{1, 2, \ldots\}$

$$\begin{aligned}
\log(d+1) - \frac{\log(d!)}{d+1} &\leq \log(d+1) - \frac{\log\left(2^{1/2}\pi^{1/2}d^{d+1/2}e^{-d}\right)}{d+1} \\
&= \log(d+1) + \frac{d}{d+1} - \frac{\log(2^{1/2}\pi^{1/2})}{d+1} - \frac{d+1/2}{d+1}\log d \\
&\leq \log(d+1) + 1 - \frac{d+1-1+1/2}{d+1}\log d \\
&= 1 + \log\left(\frac{d+1}{d}\right) + (1/2)\frac{\log d}{d+1} \\
&\leq 1 + \log\left(1+\frac{1}{d}\right) + (1/2)\frac{\log d}{d+1} \leq 1 + \log 2 + 1.
\end{aligned}$$

This implies that

$$\sum_{j=1}^{d-1} \log\left(\frac{1}{\delta j^m}\right) \leq \delta^{-1/m} m \left[\log(d) - d^{-1}\log\{(d-1)!\}\right] \leq U_2 \delta^{-1/m}.$$

□

Corollary C.1.2 (Multivariate framework)**.** *For Theorem C.1, let* $w_{q_k} = k^m - (k-1)^m$ *where for a fixed dimension* p *we define*

$$q_k = \sum_{l=1}^{k} \binom{l+p-2}{l-1} = \binom{k+p-1}{p},$$

and all other $w_k = 0$*. Then we have the following upper bound:*

$$H(\delta,\, E^w_{K_n}) \leq U_E \delta^{-p/m},$$

for some constant U_E *which only depends on* m *and* θ*.*

Proof. Firstly, since $w_1 = 1$ the entropy is 0 for $\delta \geq 1$ and hence we will restrict ourselves to $\delta \in (0, 1)$. We note that we must have $\mu_\theta = q_{k_1} - 1$ for some integer k_1. This is because all

weights after q_{k_1-1} are zero until w_{q_k}. Now we have by definition

$$w_1 + \cdots + w_{q_{k_1}-1} = (k_1 - 1)^m \leq \left\{\delta(1-\theta)^{1/2}\right\}^{-1},$$

and we have

$$\begin{aligned}
\mu_\theta = q_{k_1} - 1 &= \binom{k_1 + p - 1}{p} - 1 < \frac{(k_1 + p - 1)^p}{p!} \\
&\leq \frac{\left[\left\{\delta(1-\theta)^{1/2}\right\}^{-1/m} + p\right]^p}{p!} = \frac{\delta^{-\frac{p}{m}}\left\{(1-\theta)^{-1/(2m)} + p\delta^{1/m}\right\}^p}{p!} \\
&\leq \delta^{-\frac{p}{m}} \frac{\left\{(1-\theta)^{-1/(2m)} + p\right\}^p}{p!},
\end{aligned}$$

where the second line follows from the inequality

$$\binom{n}{k} \leq n^k/k!.$$

This implies that for $\delta \in (0,\, 1)$

$$\mu_\theta \log(3/\theta) \leq U_1 \delta^{-\frac{p}{m}}.$$

Similarly, there is an integer k_2 such that $d - 1 = q_{k_2} - 1$. Which means that $(k_2 - 1)^m \leq \delta^{-1} \leq k_2^m$. For the other term we have

$$\begin{aligned}
&\sum_{k=1}^{d-1} \log\left(\frac{1}{\delta \sum_{l=1}^k w_l}\right) = (d-1)\left\{\log(\delta^{-1}) - \frac{\sum_{k=1}^{d-1} \log(\sum_{l=1}^k w_l)}{d-1}\right\} \\
&= (d-1)\left\{\log(\delta^{-1}) - \frac{(q_2 - 1)\log(1^m) + (q_3 - q_2)\log(2^m) + \cdots + (q_{k_2} - 1 - q_{k_2-1})\log(q_{k_2-1}^m)}{d-1}\right\} \\
&= (d-1)\left[\log(\delta^{-1}) - \frac{m\{f(k_2) - \log(q_{k_2-1})\}}{d-1}\right],
\end{aligned}$$

where $f(k_2) = (q_2 - 1)\log(1) + (q_3 - q_2)\log(2) + \cdots + (q_{k_2} - q_{k_2-1})\log(q_{k_2-1}) = \sum_{l=1}^{k_2}(q_l -$

$q_{l-1})\log(q_{l-1})$. Hence we have

$$\sum_{j=1}^{d-1}\log\left(\frac{1}{\delta\sum_{l=1}^{k}a_l}\right) \leq (d-1)\left[\log(k_2^m) - \frac{m\{f(k_2)-\log(q_{k_2-1})\}}{d-1}\right]$$
$$= m(d-1)\left\{\log(k_2) - \frac{f(k_2)-\log(q_{k_2-1})}{q_{k_2}-1}\right\}.$$

Now by induction we can show that $\frac{f(k_2)-\log(q_{k_2-1})}{q_{k_2}-1} \geq \frac{\log\{(k_2-1)!\}}{k_2}$ which implies that

$$m(d-1)\left\{\log(k_2) - \frac{f(k_2)-\log(q_{k_2-1})}{q_{k_2}-1}\right\} \leq m(d-1)\left[\log(k_2) - \frac{\log\{(k_2-1)!\}}{k_2}\right]$$
$$\leq (d-1)m\left\{2+\log(2)\right\}.$$

Finally, we note that

$$d-1 = q_{k_2}-1 = \binom{k_2+p-1}{p} - 1 < \binom{k_2+p-1}{p}$$
$$\leq \frac{(k_2+p-1)^p}{p!} \leq \frac{(\delta^{-1/m}+p)^p}{p!} = \delta^{-\frac{p}{m}}\frac{\left(1+p\delta^{1/m}\right)^p}{p!} \leq \delta^{-\frac{p}{m}}\frac{(1+p)^p}{p!}.$$

□

C.7 Proof of Theorem 4.2

Proof. By definition

$$\frac{1}{2}\|\widehat{f}-\boldsymbol{y}\|_n^2 + \lambda_n^2\Omega(\widehat{f}|Q_n) \leq \frac{1}{2}\|f_n^*-\boldsymbol{y}\|_n^2 + \lambda_n^2\Omega(f_n^*|Q_n),$$

which leads to the following inequality

$$\frac{1}{2}\|\widehat{f}-f^0\|_n^2 + \lambda_n^2\Omega(\widehat{f}|Q_n) \leq |\langle\boldsymbol{\varepsilon},\widehat{f}-f_n^*\rangle_n| + \frac{1}{2}\|f_n^*-f^0\|_n^2 + \lambda_n^2\Omega(f_n^*|Q_n)\ ,$$

where $\langle \boldsymbol{\varepsilon}, f\rangle_n = \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)$. Via the simple decomposition $\|\widehat{f} - f_n^*\|_n^2 \leq 2\|\widehat{f} - f^0\|_n^2 + 2\|f^0 - f_n^*\|_n^2$ we obtain

$$\begin{aligned}
\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 + \lambda_n^2\Omega(\widehat{f}|Q_n) &\leq \|\widehat{f} - f^0\|_n^2 + \|f^0 - f_n^*\|_n^2 + 2\lambda_n^2\Omega(\widehat{f}|Q_n) \\
&= \|f^0 - f_n^*\|_n^2 + 2\left\{\frac{1}{2}\|\widehat{f} - f^0\|_n^2 + \lambda_n^2\Omega(\widehat{f}|Q_n)\right\} \\
&\leq \|f^0 - f_n^*\|_n^2 + 2\left\{|\langle \varepsilon, \widehat{f} - f_n^*\rangle_n| + \frac{1}{2}\|f_n^* - f^0\|_n^2 + \lambda_n^2\Omega(f_n^*|Q_n)\right\} \\
&= 2|\langle \boldsymbol{\varepsilon}, \widehat{f} - f_n^*\rangle_n| + 2\lambda_n^2\Omega(f_n^*|Q_n) + \|f^0 - f_n^*\|_n^2 \\
&\leq \max\left\{4|\langle \boldsymbol{\varepsilon}, \widehat{f} - f_n^*\rangle_n| + 4\lambda_n^2\Omega(f_n^*|Q_n),\ 2\|f^0 - f_n^*\|_n^2\right\}.
\end{aligned}$$

Thus our basic inequality is given by

$$\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 + \lambda_n^2\Omega(\widehat{f}|Q_n) \leq 2\max\left\{2|\langle \boldsymbol{\varepsilon}, \widehat{f} - f_n^*\rangle_n| + 2\lambda_n^2\Omega(f_n^*|Q_n),\ \|f^0 - f_n^*\|_n^2\right\}.$$

Hence from the basic inequality either $\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 + \lambda_n^2\Omega(\widehat{f}|Q_n) \leq 2\|f^0 - f_n^*\|_n^2$ which implies the result or

$$\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 + \lambda_n^2\Omega(\widehat{f}|Q_n) \leq 4\langle \boldsymbol{\varepsilon}, \widehat{f} - f_n^*\rangle_n + 4\lambda_n^2\Omega(f_n^*|Q_n)\ .$$

Now note that $H(\delta, \{f \in \mathcal{F}_n : \Omega(f|Q_n) \leq 1\}, Q_n) \leq A_1\delta^{-\alpha}$ implies

$$H\left(\delta, \left\{\frac{f - f_n^*}{\Omega(f|Q_n) + \Omega(f_n^*|Q_n)} : f \in \mathcal{F}_n\right\}, Q_n\right) \leq \widetilde{A}_1\delta^{-\alpha}.$$

Thus we invoke Lemma 8·4 of van de Geer (2000) and conclude that with probability at least $1 - c\exp\{-(T/c)^2\}$ for a constant $c > 0$ and all $T \geq c$, we have

$$|\langle \boldsymbol{\varepsilon}, \widehat{f} - f_n^*\rangle_n| \leq Tn^{-1/2}\|\widehat{f} - f_n^*\|_n^{1-\frac{\alpha}{2}}\left\{\Omega(\widehat{f}|Q_n) + \Omega(f_n^*|Q_n)\right\}^{\frac{\alpha}{2}}.$$

Define the set $\mathcal{T}$ as

$$\mathcal{T} = \left\{ \sup_{f \in \mathcal{F}_n} \left| \langle \boldsymbol{\varepsilon},\, f - f_n^* \rangle_n \right| \leq T n^{-1/2} \|f - f_n^*\|_n^{1-\frac{\alpha}{2}} \left\{ \Omega(f|Q_n) + \Omega(f_n^*|Q_n) \right\}^{\frac{\alpha}{2}} \right\} ,$$

then on the set $\mathcal{T}$ we have

$$\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 + \lambda_n^2 \Omega(\widehat{f}|Q_n) \leq 4 T n^{-1/2} \|\widehat{f} - f_n^*\|_n^{1-\frac{\alpha}{2}} \left\{ \Omega(\widehat{f}|Q_n) + \Omega(f_n^*|Q_n) \right\}^{\frac{\alpha}{2}} + 4\lambda_n^2 \Omega(f_n^*|Q_n) \ .$$

Which means we have either

$$\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 + \lambda_n^2 \Omega(\widehat{f}|Q_n) \leq 8\lambda_n^2 \Omega(f_n^*|Q_n),$$

which is of the desired form or

$$\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 + \lambda_n^2 \Omega(\widehat{f}|Q_n) \leq 8 T n^{-1/2} \|\widehat{f} - f_n^*\|_n^{1-\frac{\alpha}{2}} \left\{ \Omega(\widehat{f}|Q_n) + \Omega(f_n^*|Q_n) \right\}^{\frac{\alpha}{2}} \ . \tag{C.9}$$

We now consider (C.9) only.

C.7.1 Case 1: $\Omega(\widehat{f}|Q_n) \geq \Omega(f_n^*|Q_n)$

In this case we have

$$\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 + \lambda_n^2 \Omega(\widehat{f}|Q_n) \leq 8 T n^{-1/2} \|\widehat{f} - f_n^*\|_n^{1-\frac{\alpha}{2}} \left\{ 2\Omega(\widehat{f}|Q_n) \right\}^{\frac{\alpha}{2}} .$$

which gives us

$$\begin{aligned} \lambda_n^2 \Omega(\widehat{f}|Q_n) &\leq 8 T n^{-1/2} \|\widehat{f} - f_n^*\|_n^{1-\frac{\alpha}{2}} \left\{ 2\Omega(\widehat{f}|Q_n) \right\}^{\frac{\alpha}{2}} \\ \Leftrightarrow \left\{ \Omega(\widehat{f}|Q_n) \right\}^{1-\frac{\alpha}{2}} &\leq 2^{3+\frac{\alpha}{2}} T n^{-\frac{1}{2}} \lambda_n^{-2} \|\widehat{f} - f_n^*\|_n^{1-\frac{\alpha}{2}} \\ \Leftrightarrow \Omega(\widehat{f}|Q_n) &\leq \left(2^{3+\frac{\alpha}{2}} T n^{-\frac{1}{2}} \lambda_n^{-2} \right)^{\frac{2}{2-\alpha}} \|\widehat{f} - f_n^*\|_n. \end{aligned}$$

Plugging this into the right hand side of (C.7.1) and solving for $\|\widehat{f} - f_n^*\|_n$ we obtain

$$\begin{aligned}
&\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 \leq T 2^{3+\frac{\alpha}{2}} n^{-\frac{1}{2}} \|\widehat{f} - f_n^*\|_n^{1-\frac{\alpha}{2}} \left(2^{3+\frac{\alpha}{2}} T n^{-\frac{1}{2}} \lambda_n^{-2}\right)^{\frac{2}{2-\alpha}\frac{\alpha}{2}} \|\widehat{f} - f_n^*\|_n^{\alpha/2} \Rightarrow \\
&\frac{1}{2}\|\widehat{f} - f_n^*\|_n \leq T^{\frac{2}{2-\alpha}} 2^{\frac{6+\alpha}{2-\alpha}} n^{-\frac{1}{2-\alpha}} \lambda_n^{-\frac{2\alpha}{2-\alpha}} \Rightarrow \\
&\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 \leq C_1 n^{-\frac{2}{2-\alpha}} \lambda_n^{-\frac{4\alpha}{2-\alpha}} = C_1 \lambda_n^2 \Omega(f_n^*|Q_n) \ ,
\end{aligned}$$

where $C_1 = T^{\frac{4}{2-\alpha}} 2^{\frac{14+\alpha}{2-\alpha}}$ and recall the definition $\lambda_n^{-1} = n^{\frac{1}{2+\alpha}} \left\{\Omega(f_n^*|Q_n)\right\}^{\frac{2-\alpha}{2(2+\alpha)}}$.

C.7.2 Case 2: $\Omega(\widehat{f}|Q_n) \leq \Omega(f_n^*|Q_n)$

In this case we have

$$\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 + \lambda_n^2 \Omega(\widehat{f}|Q_n) \leq 8 T n^{-1/2} \|\widehat{f} - f_n^*\|_n^{1-\frac{\alpha}{2}} \left\{2\Omega(f_n^*|Q_n)\right\}^{\frac{\alpha}{2}} .$$

From which we directly get

$$\begin{aligned}
\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 &\leq 8 T n^{-\frac{1}{2}} \|\widehat{f} - f_n^*\|_n^{1-\frac{\alpha}{2}} \left\{2\Omega(f_n^*|Q_n)\right\}^{\frac{\alpha}{2}} \Rightarrow \\
\frac{1}{2}\|\widehat{f} - f_n^*\|_n^{1+\frac{\alpha}{2}} &\leq 2^{3+\frac{\alpha}{2}} T n^{-\frac{1}{2}} \left\{\Omega(f_n^*|Q_n)\right\}^{\frac{\alpha}{2}} \Rightarrow \\
\|\widehat{f} - f_n^*\|_n &\leq 2^{\frac{8+\alpha}{2+\alpha}} T^{\frac{2}{2+\alpha}} n^{-\frac{1}{2+\alpha}} \left\{\Omega(f_n^*|Q_n)\right\}^{\frac{\alpha}{2+\alpha}} \Rightarrow \\
\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 &\leq C_2 n^{-\frac{2}{2+\alpha}} \left\{\Omega(f_n^*|Q_n)\right\}^{\frac{2\alpha}{2+\alpha}} = C_2 \lambda_n^2 \Omega(f_n^*|Q_n),
\end{aligned}$$

where $C_2 = T^{4/(2+\alpha)} 2^{(14+\alpha)/(2+\alpha)}$. Thus we have shown that on the set $\mathcal{T}$ we have

$$\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 \leq \max(8, C_1, C_2) \lambda_n^2 \Omega(f_n^*|Q_n) = C_0 \lambda_n^2 \Omega(f_n^*|Q_n).$$

We have shown that with probability at least $1 - c\exp\left\{-(T/c)^2\right\}$ we have the inequality

$$\frac{1}{2}\|\widehat{f} - f_n^*\|_n^2 \leq \max\left\{2\|f^0 - f_n^*\|_n^2,\ C_0 \lambda_n^2 \Omega(f_n^*|Q_n)\right\}$$

To complete the proof we note that

$$\begin{aligned}\frac{1}{2}\|\widehat{f} - f^0\|_n^2 &\leq \|\widehat{f} - f_n^*\|_n^2 + \|f_n^* - f^0\|_n^2 \\ &\leq 2\max\left\{2\|f^0 - f_n^*\|_n^2,\ C_0\lambda_n^2\Omega(f_n^*|Q_n)\right\} + \|f_n^* - f^0\|_n^2 \\ &\leq \frac{5}{2}\max\left\{2\|f^0 - f_n^*\|_n^2,\ C_0\lambda_n^2\Omega(f_n^*|Q_n)\right\}.\end{aligned}$$

□

Appendix D

TECHNICAL DETAILS FOR "WAVELET REGRESSION AND ADDITIVE MODELS FOR IRREGULARLY SPACED DATA"

D.1 Details of algorithms

Here we give an algorithm for our additive and sparse-additive framework as well as an algorithm for the extension of our proposal to classification. We use a block-wise coordinate descent algorithm for solving the additive and sparse additive proposal. This algorithm cyclically iterates through features, and for each feature applies the univariate solution detailed in the main chapter. The exact details are given in Algorithm 6 below.

Algorithm 6 Block coordinate descent for the additive and sparse additive framework

Initialize $\boldsymbol{d}_j \leftarrow 0$ for $j = 1, \ldots, p$
While $l \leq max\_iter$ **and** not converged
  For $j = 1, \ldots, p$
    Set $\boldsymbol{r}_{-j} \leftarrow \boldsymbol{y} - \sum_{j' \neq j} \boldsymbol{R}_{j'} \boldsymbol{W}^\top \boldsymbol{d}_{j'}$
    Update $\boldsymbol{d}_j \leftarrow \underset{\boldsymbol{d} \in \mathbb{R}^{\tilde{K}}}{\arg\min} \frac{1}{2} \left\| \boldsymbol{r}_{-j} - \boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d} \right\|_2^2 + \lambda_1 \|\boldsymbol{d}_{-1}\|_1 + \lambda_2 \|\boldsymbol{R}_j \boldsymbol{W}^\top \boldsymbol{d}\|_2,$
Return $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p$

We also give an algorithm for the extension of our method to classification based on proximal gradient descent. To begin let $L(\boldsymbol{d}) = 1/(2n) \sum_{i=1}^n \log\left(1 + \exp\left[-y_i \left\{\left(RW^\top \boldsymbol{d}\right)_i\right\}\right]\right)$, or more generally let it be some differentiable convex loss function. We denote by $\nabla L(\boldsymbol{d})$, the derivative of L at the point $\boldsymbol{d} \in \mathbb{R}^{\tilde{K}}$. Algorithm 7 presents the steps for solving the univariate `waveMesh` problem with general loss. The algorithm for extension of additive models to classification (or other loss functions) can be similarly derived and is omitted in the interest of brevity.

Algorithm 7 Proximal gradient descent for extension to classification

Initialize $\boldsymbol{d}^0$
For $l = 1, 2, \ldots$ until convergence
Select a step size t_l via line search
Update

$$\boldsymbol{d}^l \leftarrow \underset{\boldsymbol{d}\in\mathbb{R}^{\tilde{K}}}{\arg\min} \frac{1}{2}\left\|\boldsymbol{d} - \left\{\boldsymbol{d}^{l-1} - t_l \nabla L(\boldsymbol{d}^{l-1})\right\}\right\|_2^2 + t_l\lambda\|\boldsymbol{d}_{-1}\|_1.$$

Return $\boldsymbol{d}^l$

D.2 Proofs for univariate results

Here we present the proof for Theorem 5.1. We consider the estimator

$$\widehat{\boldsymbol{d}} \leftarrow \underset{\boldsymbol{d}\in\mathbb{R}^{K}}{\arg\min} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{R}\boldsymbol{W}^\top\boldsymbol{d}\|_2^2 + \lambda\|\boldsymbol{d}_M\|_1, \tag{D.1}$$

where $\boldsymbol{d}_M$ denotes the sub-vector corresponding to the mother wavelet coefficients. We use this notation to generalize the case of $j_0 = 0$ where j_0 denotes the minimum resolution level. One nice feature about (D.1) is that it is exactly the lasso problem (Tibshirani, 1996) with design matrix $\boldsymbol{R}\boldsymbol{W}^\top$.

Proof of Theorem 1. We can divide the proof into three parts, (1) the deterministic part, (2) the stochastic part and (3) the approximation error part. The first 2 parts are standard in the lasso literature, for this reason we will use the results from the book by van de Geer (2016).

Deterministic Part

As per Theorem 2.1 of van de Geer (2016) let λ_ε satisfy

$$\lambda_\varepsilon \geq \|\boldsymbol{W}\boldsymbol{R}^\top\boldsymbol{\varepsilon}\|_\infty/n,$$

where $\boldsymbol{\varepsilon}$ is the noise vector. Define for $\lambda > \lambda_\varepsilon$

$$\overline{\lambda} = \lambda + \lambda_\varepsilon, \quad \underline{\lambda} = \lambda - \lambda_\varepsilon,$$

and stretching factor $L = \overline{\lambda}/\underline{\lambda}$. Further more, for an index set $\mathcal{S} \subset \{1, \ldots, K\}$ and stretching factor L define the *compatibility constant* as

$$\widehat{\vartheta}^2(L, \mathcal{S}) = \min \left\{ n^{-1}|\mathcal{S}| \|\boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}\|_2^2 : \|\boldsymbol{d}_{\mathcal{S}}\|_1 = 1, \|\boldsymbol{d}_{-\mathcal{S}}\|_1 \leq L \right\}, \tag{D.2}$$

where $\boldsymbol{d}_S$ is the vector $\boldsymbol{d}$ with values equal to 0 for indices in $\mathcal{S}$. Similarly $\boldsymbol{d}_{-\mathcal{S}}$ is the vector $\boldsymbol{d}$ with values equal to 0 for indices in $\mathcal{S}^c$. Then we have for any set $\mathcal{S}$, and vector $\boldsymbol{d}^*$ we have

$$n^{-1}\|\widehat{\boldsymbol{f}} - \boldsymbol{f}^0\|_2^2 \leq n^{-1}\|\boldsymbol{f}^0 - \boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}^*\|_2^2 + \frac{|\mathcal{S}|\overline{\lambda}^2}{\widehat{\vartheta}^2(L, \mathcal{S})}. \tag{D.3}$$

For simplicity we take the $\lambda = 2\lambda_\varepsilon$ giving us $\overline{\lambda} = 3\lambda_\varepsilon$, $\underline{\lambda} = \lambda_\varepsilon$ and $L = 3$. □

We consider a quick calculation of the compatibility constant $\widehat{\vartheta}(L, \mathcal{S})$. Let $\Lambda_{\min}(\boldsymbol{R})$ be the minimum eigenvalue of $\boldsymbol{R}$, this will normally be greater than 0 if $K < n$. We then note that:

$$\begin{aligned} n^{-1}|\mathcal{S}| \|\boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}\|_2^2 &\geq \Lambda_{\min}(\boldsymbol{R}) n^{-1}|\mathcal{S}| \|\boldsymbol{d}\|_2^2 \\ &= \Lambda_{\min}(\boldsymbol{R}) n^{-1}|\mathcal{S}| \left\{ \|\boldsymbol{d}_S\|_2^2 + \|\boldsymbol{d}_{-\mathcal{S}}\|_2^2 \right\} \\ &\geq \Lambda_{\min}(\boldsymbol{R}) n^{-1}|\mathcal{S}| \left\{ \frac{\|\boldsymbol{d}_S\|_1^2}{|\mathcal{S}|} + \frac{\|\boldsymbol{d}_{-\mathcal{S}}\|_1^2}{K - |\mathcal{S}|} \right\}, \end{aligned}$$

and minimizing the right hand side under the constraints $\|\boldsymbol{d}_{\mathcal{S}}\|_1 = 1$ and $\|\boldsymbol{d}_{-\mathcal{S}}\|_1 \leq L$ we can get that it is bounded below by $\Lambda_{\min}(\boldsymbol{R}) n^{-1}$. This gives us one possible value for the compatibility constant $\widehat{\vartheta}^2(L, \mathcal{S})$, notice that this includes the special case of traditional wavelet regression with $\boldsymbol{R} = \boldsymbol{I}$ and $\Lambda_{\min}(\boldsymbol{R}) = 1$.

Thus we have that

$$n^{-1}\|\widehat{\boldsymbol{f}} - \boldsymbol{f}^0\|_2^2 \leq n^{-1}\|\boldsymbol{f}^0 - \boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}^*\|_2^2 + \frac{9n|\mathcal{S}|\lambda_\varepsilon^2}{\Lambda_{\min}(\boldsymbol{R})}. \tag{D.4}$$

Stochastic part

We focus on obtaining a possible values for λ_ε. We start with the simple case where $\boldsymbol{R} = \boldsymbol{I}$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$, i.e. the traditional wavelet approach with regularly spaced data. In this case we need to find a λ_ε such that

$$\lambda_\varepsilon \geq \|\boldsymbol{W}\boldsymbol{\varepsilon}\|_\infty / n. \tag{D.5}$$

First note that $\boldsymbol{\varepsilon}' = \boldsymbol{W}\boldsymbol{\varepsilon}/\sigma \sim \mathcal{N}(0, I)$ by orthogonality of $\boldsymbol{W}$. Hence we have

$$\mathbb{P}\left(\|\boldsymbol{\varepsilon}'\|_\infty > \sqrt{t^2 + 2\log n}\right) \leq 2p\exp\left[-\frac{t^2 + 2\log p}{2}\right] = 2\exp(-t^2/2). \tag{D.6}$$

Thus with probability at-least $1 - 2\exp(-t^2/2)$ we have $\sigma\sqrt{t^2 + 2\log n} \geq \|\boldsymbol{W}\boldsymbol{\varepsilon}\|_\infty$. Thus in this case we can take $\lambda_\varepsilon = n^{-1}\sigma\sqrt{t^2 + 2\log n}$. In the general case we would have the mean zero, sub-Gaussian K-vector $\boldsymbol{W}\boldsymbol{R}^\top\boldsymbol{\varepsilon}$. By a slightly more involved argument we can show that we can take $\lambda_\varepsilon = n^{-1}c_1\sqrt{t^2 + 2\log K}$ where c_1 depends on the distribution of $\boldsymbol{\varepsilon}$ (i.e., the parameters of the sub-gaussian distribution) and matrix $\boldsymbol{R}$.

Thus we have shown so far that with probability at-least $1 - 2\exp(-t^2/2)$ we have

$$n^{-1}\|\widehat{\boldsymbol{f}} - \boldsymbol{f}^0\|_2^2 \leq n^{-1}\|\boldsymbol{f}^0 - \boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}^*\|_2^2 + \frac{9c_1^2}{\Lambda_{\min}(\boldsymbol{R})}\frac{|\mathcal{S}|(t^2 + 2\log K)}{n}, \tag{D.7}$$

or without worrying about optimal constants we get the rate

$$n^{-1}\|\widehat{\boldsymbol{f}} - \boldsymbol{f}^0\|_2^2 \leq n^{-1}\|\boldsymbol{f}^0 - \boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}^*\|_2^2 + C\frac{|\mathcal{S}|\log K}{n}. \tag{D.8}$$

To obtain our result we just need the final step: approximation error.

Approximation error part

Now we will bound the term $n^{-1}\|\boldsymbol{f}^0 - \boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}^*\|_2^2$. We will define specific types of vectors $\boldsymbol{d}^*$ which leads to specific sparse indes sets $\mathcal{S}$. We begin with the decomposition:

$$n^{-1}\|\boldsymbol{f}^0 - \boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}^*\|_2^2 \leq 2n^{-1}\|\boldsymbol{f}^0 - \boldsymbol{R}\widetilde{\boldsymbol{f}}^0\|_2^2 + 2n^{-1}\|\boldsymbol{R}\widetilde{\boldsymbol{f}}^0 - \boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}^*\|_2^2, \tag{D.9}$$

where $\widetilde{f}^0$ is the function obtained by interpolating f^0 from the data $(i/K, f^0(i/K))$ for $i = 1, \ldots, K$ and $\widetilde{f}^0 = [\widetilde{f}^0(1/K), \ldots, \widetilde{f}^0(K/K)]^\top$.

For the second term, define $\Lambda_{\max}(\boldsymbol{R})$ as the maximum eigenvalue of $\boldsymbol{R}^\top\boldsymbol{R}$ then

$$n^{-1}\|\boldsymbol{R}\widetilde{\boldsymbol{f}}^0 - \boldsymbol{R}\boldsymbol{W}^\top \boldsymbol{d}^*\|_2^2 \leq \Lambda_{\max}(\boldsymbol{R})n^{-1}\|\widetilde{\boldsymbol{f}}^0 - \boldsymbol{W}^\top \boldsymbol{d}^*\|_2^2 \leq \Lambda_{\max}(\boldsymbol{R})\|\widetilde{\boldsymbol{f}}^0 - \boldsymbol{W}^\top \boldsymbol{d}^*\|_\infty^2.$$

For the last part we now define $\boldsymbol{d}^*$, the vector of wavelet coefficients such that it defines a function f^* as a linear combination of wavelet basis functions. To be precise we have that

$$f^*(x) = \sum_{k=0}^{2^{j_0}-1} \phi_{j_0k}(x)\alpha_{j_0k}^0 + \sum_{j=j_0}^{J^*-1}\sum_{k=0}^{2^j-1} \psi_{jk}(x)\beta_{jk}^0, \tag{D.10}$$

for some integer J^*, and where α_{j0k}^0 and β_{jk}^0 are the wavelet coefficients of the true function f^0. Now we obtain:

$$\begin{aligned}
\max_x |f^*(x) - f^0(x)| &= \max_x \Big| \sum_{j=J^*}^{\infty}\sum_{k=0}^{2^j-1} \psi_{jk}(x)\beta_{jk}^0 \Big| \\
&\leq \max_x \max_{j\geq J^*,k} |\psi_{jk}(x)| \sum_{j=J^*}^{\infty}\sum_{k=0}^{2^j-1} |\beta_{jk}^0| \\
&= \max_x \max_{j\geq J^*,k} |\psi_{jk}(x)| \sum_{j=J^*}^{\infty} \|\boldsymbol{\beta}_j^0\|_1,
\end{aligned}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^{2^j}$ is the mother wavelet coefficient vector at level j. Now assuming that

$f^0 \in B^s_{q_1,q_2}$

$$\sum_{j=J^*}^{\infty} \|\boldsymbol{\beta}_j^0\|_1 = \sum_{j=J^*}^{\infty} \frac{2^{js'}}{2^{js'}} \|\boldsymbol{\beta}_j^0\|_1, \qquad (s' = s - 1/2)$$
$$\leq \left[\sum_{j=J^*}^{\infty} \left(2^{js'} \|\boldsymbol{\beta}_j^0\|_1\right)^{q_2}\right]^{1/q_2} \left[\sum_{j=J^*}^{\infty} 2^{-js'q_2'}\right]^{1/q_2'},$$

where q_2' is such that $1/q_2 + 1/q_2' = 1$. Using the inequality $\|\boldsymbol{\beta}_j^0\|_1 \leq 2^{j(1-1/q_1)} \|\boldsymbol{\beta}_j^0\|_{q_1}$ we get

$$\sum_{j=J^*}^{\infty} \|\boldsymbol{\beta}_j^0\|_1 \leq \left[\sum_{j=J^*}^{\infty} \left(2^{j(s+1/2-1/q_1)} \|\boldsymbol{\beta}_j^0\|_{q_1}\right)^{q_2}\right]^{1/q_2} \left[\sum_{j=J^*}^{\infty} 2^{-js'q_2'}\right]^{1/q_2'}$$
$$= \left[\sum_{j=J^*}^{\infty} \left(2^{j(s+1/2-1/q_1)} \|\boldsymbol{\beta}_j^0\|_{q_1}\right)^{q_2}\right]^{1/q_2} \times C_2 2^{-J^* s},$$

where the second term can be obtained by looking at $S_\infty - S_{J^*-1}$ where $S_n = \sum_{j=0}^{n} 2^{-js'q_2'}$. The first term is bounded because $f^0 \in B^s_{q_1,q_2}$.

Putting the pieces together

Thus we have shown so far, by taking $\boldsymbol{d}^*$ as defined above and $\mathcal{S}$ being the active set of $\boldsymbol{d}^*$ (i.e. $|\mathcal{S}| = 2^{J^*}$), that the rate is of the form (upto constants)

$$n^{-1} \|\widehat{\boldsymbol{f}} - \boldsymbol{f}^0\|_2^2 \leq 2n^{-1} \|\boldsymbol{f}^0 - \boldsymbol{R}\widetilde{\boldsymbol{f}}^0\|_2^2 + C_2 2^{-(2s)J^*} + C_3 2^{J^*} \frac{\log K}{n}.$$

Treating the above as a function of J^* and minimizing we obtain the approximate truncation order $|\mathcal{S}| = \mathcal{O}(n^{1/(2s+1)})$ which minimizes the right hand side. Finally, putting all the different pieces together we obtain the bound:

$$n^{-1} \|\widehat{\boldsymbol{f}} - \boldsymbol{f}^0\|_2^2 \leq C_4 \left(\frac{\log K}{n}\right)^{\frac{2s}{2s+1}} + 2n^{-1} \|\boldsymbol{f}^0 - \boldsymbol{R}\widetilde{\boldsymbol{f}}^0\|_2^2.$$