Journal of Machine Learning Research 23 (2022) 1-56 Submitted 2/20; Revised 4/21; Published 05/22

Generalized Sparse Additive Models

Asad Haris ASAD.HARIS@UBC.CA
Department of Earth, Ocean and Atmospheric Sciences
University of British Columbia
2020 – 2207 Main Mall
Vancouver, BC, Canada V6T 1Z4

Noah Simon NRSIMON@UW.EDU
Ali Shojaie ASHOJAIE@UW.EDU
Department of Biostatistics
University of Washington
Seattle, WA 98195-7232, USA

Editor: Garvesh Raskutti

Abstract

We present a unified framework for estimation and analysis of generalized additive models in high dimensions. The framework defines a large class of penalized regression estimators, encompassing many existing methods. An efficient computational algorithm for this class is presented that easily scales to thousands of observations and features. We prove minimax optimal convergence bounds for this class under a weak compatibility condition. In addition, we characterize the rate of convergence when this compatibility condition is not met. Finally, we also show that the optimal penalty parameters for structure and sparsity penalties in our framework are linked, allowing cross-validation to be conducted over only a single tuning parameter. We complement our theoretical results with empirical studies comparing some existing methods within this framework.

Keywords: Generalized Additive Models, Sparsity, Minimax, High-Dimensional, Penalized Regression

1. Introduction

In this paper, 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 $\boldsymbol{x}_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$. The goal is to fit the model

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

for a prespecified *link* function g, unknown intercept β and, unknown component functions $f_1, \ldots, f_p$. The link function, g, is generally based on the outcome data-type, for example, $g(x) = x$ 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 for all i and j. For identifiability, we assume $\sum_{i=1}^{n} f_j(x_{ij}) = 0$ for all $j = 1, \ldots, p$. This model 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 they extend GLMs to model non-linear conditional relationships while retaining some interpretability (we can examine the effect of each covariate x_{ij} individually on y_i while holding all other variables fixed); they also 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:

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

Here $\mathcal{F}$ is some suitable function class; $\ell(y_i, \theta)$ is the log-likelihood of y_i under parameter θ; P_{st} 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, for example, $\mathcal{F} = L^2[0, 1]$. Functions $-\ell(y_i, \theta)$ and $P_{st}(f_j)$ are convex in θ and f_j, respectively. The objective function in (1) is convex and for small dimension, p, can be solved via a general-purpose convex solver. However, many modern data sets are high-dimensional, often with more features than observations, that is, $p > n$. Fitting even GLMs is challenging in such settings as conventional methods are known to overfit the data. A common assumption in the high-dimensional setting is *sparsity*, which states that, only a small (but unknown) subset of features is 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:

$$\widehat{\beta}, \widehat{f}_1, \ldots, \widehat{f}_p \leftarrow \underset{\beta \in \mathbb{R}, f_1, \ldots, f_p \in \mathcal{F}}{\operatorname{argmin}} \; - n^{-1} \sum_{i=1}^{n} \ell\Big(y_i, \beta + \sum_{j=1}^{p} f_j\left(x_{ij}\right)\Big) + \lambda_{st} \sum_{j=1}^{p} P_{st}\left(f_j\right) + \lambda_{sp} \sum_{j=1}^{p} \left\|f_j\right\|_n, \tag{2}$$

where $\|f_j\|_n^2 = n^{-1} \sum_{i=1}^{n} \{f_j(x_{ij})\}^2$ 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; Meier et al., 2009; Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Yuan and Zhou, 2016; Lou et al., 2016; Petersen et al., 2016; Sadhanala and Tibshirani, 2019; Tan and Zhang, 2019). However, previous proposals consists of gaps around efficient computation (Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Yuan and Zhou, 2016; Tan and Zhang, 2019) and/or optimal statistical convergence properties (Ravikumar et al., 2009; Lou et al., 2016; Petersen et al., 2016; Sadhanala and Tibshirani, 2019). General-purpose convex solvers have also been suggested (Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Yuan and Zhou, 2016) as an alternative for solving problem (2), but they roughly scale as $O(n^3 p^3)$ and are hence inefficient. This manuscript aims to bridge these gaps.

We present a general framework for sparse GAMs with two major contributions, a general algorithm for computing (2) and a theorem for establishing convergence rates. Briefly, our algorithm is based on accelerated proximal gradient descent. This reduces (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 $\max(s \log p/n, s\xi_n)$, where s is the number of signal variables and ξ_n is the minimax rate of the univariate regression problem, that is, problem (1) with $p = 1$. Nonparametric rates are established for a wide class of structural penalties P_{st} with $\xi_n = n^{-2m/(2m+1)}$, 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 $\xi_n = T_n/n$ via a truncation-penalty; the number of parameters, T_n, can be fixed or allowed to grow with sample size.

The highlight of this paper is the generality of the proposed framework: not only does it encompass many existing estimators for high-dimensional GAMs, but also estimators for low-dimensional fully nonparametric models and, parametric models in low or high dimensions. Brief examples of our framework's generalizability include: establishing minimax convergence rates (under well-studied assumptions) where only sub-optimal rates existed (Lou et al., 2016; Ravikumar et al., 2009; Meier et al., 2009; van de Geer, 2010); recovering minimax rates for a general class of loss functions as opposed to only least squares (Raskutti et al., 2012; Yuan and Zhou, 2016; Tan and Zhang, 2019); establishing consistency (albeit at a slower rate) while relaxing strong assumptions on the design matrix and function class (Raskutti et al., 2012; Yuan and Zhou, 2016; Tan and Zhang, 2019).

Extending GAMs (1), to `GSAM`s (2), appears to simply be a matter of adding a sparsity penalty. However, our manuscript proves a surprising result: sparsity in GAMs can only be achieved if P_{st} is a semi-norm penalty, as opposed to a squared semi-norm. Thus, the originally proposed GAMs (Hastie and Tibshirani, 1990) cannot be extended to high-dimensions by simply adding a sparsity penalty. Finally, as a byproduct of our general theorem, we also determine that $\lambda_{st} = \lambda_{sp}^2$ in (2), results in optimal convergence rates, reducing the problem to a single tuning parameter. Empirical studies showed a single tuning parameter to yield comparable or better performance compared to finding two tuning parameters over a grid.

The rest of the paper is organized as follows. In Section 2, we detail our framework and discuss various choices of structural penalties, P_{st}, illustrating that our framework encompasses many existing proposals. In Section 3 we present an algorithm for solving the optimization problem (2) for a broad class of P_{st} penalties, and establish their theoretical convergence rates in Section 4. We explore the empirical performance of various choices of P_{st} in simulation in Section 5, and in an application to the Boston housing data set and gene expression data sets in Section 6. Concluding remarks are in Section 7.

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. 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; 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 $\boldsymbol{x} \in \mathbb{R}^p$.

Our general framework for obtaining a *Generalized Sparse Additive Model* (GSAM) encompasses estimators that can be obtained by solving the following problem:

$$\widehat{\beta}, \widehat{f}_1, \ldots, \widehat{f}_p \leftarrow \underset{\beta \in \mathbb{R}, f_1, \ldots, f_p \in \mathcal{F}}{\operatorname{argmin}} \underbrace{-\mathbb{P}_n \ell\left(\beta + \sum_{j=1}^{p} f_j\right)}_{\text{Goodness-of-fit}} + \underbrace{\lambda^2 \sum_{j=1}^{p} P_{st}\left(f_j\right)}_{\text{structure-inducing}} + \underbrace{\lambda \sum_{j=1}^{p} \|f_j\|_n}_{\text{sparsity-inducing}} . \quad (3)$$

This optimization problem balances three terms. 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 $\|\cdot\|_n$, which encourages models with $\widehat{f}_j \equiv 0$ for many j. For some choices of P_{st} the smoothness and sparsity pattern of f_j are intrinsically linked (for example, Ravikumar et al., 2009; Lou et al., 2016); for other structure-inducing penalties the formulation (3) appears to decouple structure and sparsity. However, this manuscript highlights the surprising role of P_{st} in obtaining an appropriate sparsity pattern. Briefly, if P_{st} is a squared norm then either all $\widehat{f}_j \equiv 0$ or all $\widehat{f}_j \not\equiv 0$. We detail this result and its extension to semi-norms in Section 2.2. Throughout this manuscript, we require the function class $\mathcal{F}$ to be a convex cone, for example, $L^2(\mathbb{R})$. Later for some specific results, we will additionally require $\mathcal{F}$ to be a linear space.

As noted before, the tuning parameters for structure (λ^2) and sparsity (λ) are coupled in our framework. The theoretical consequence of this is that, for properly chosen λ, we get rate-optimal estimates, up to a constant (details in Section 4). The practical consequence is that we have a single tuning parameter. While fine tuning an estimator with two tuning parameters can lead to improved prediction performance, a single tuning parameter is adequate for most choices of P_{st} as seen in our empirical experiments of Section 5.

Furthermore, 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 and theoretical results of Section 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.

2.1 Structure Inducing Penalties

We now present some possible choices of the structural penalty P_{st} followed by a discussion of the conditions on P_{st} that lead to desirable estimation and computation. 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 $\qquad P_{st} \leftarrow P_{sobolev}(f^{(k)}) = \sqrt{\int_x \left\{f^{(k)}(x)\right\}^2 dx}$;

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

here $TV(\cdot)$ is the total variation norm, $TV(f) = \sup\{\sum_{i=1}^{o} |f(z_{i+1}) - f(z_i)| : z_1 < \ldots < z_o$ is a partition of $[0,1]\}$, and $\mathbb{I}$ 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 $\widehat{f}_j$; for example, $P_{mon}(f)$ 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 and, (c) the triangle inequality is used throughout the proof of our theoretical results of Section 4. For our context, we consider convex indicators of cones as a semi-norm because, the first order homogeneity condition can be relaxed. For our algorithm, we only require $P_{st}(\alpha f) = \alpha P_{st}(f)$ for $\alpha > 0$; for our theoretical results we treat convex indicators of cones as a special case and discuss them at the end of Section 4.2. For non-sparse GAMs of the form (1), the existing literature does not necessarily 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.

2.2 Using a Squared Smoothness Penalty

Given a semi-norm P_{semi}, using $P_{st} = P^2_{semi}$ in (3) may 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). In this subsection, we show a surprising result: using a squared semi-norm penalty results in fitting models that are either not sparse (all $\widehat{f}_j \neq 0$) or not flexible (all f_j belong to some restrictive parametric class). In other words, the original GAMs (Hastie and Tibshirani, 1990) cannot be extended to high dimensions/sparsity by simply adding a sparsity penalty. This highlights a key contribution of this manuscript: not only do we present a framework for fitting `GSAM` but also, prove that naïve GAM extensions are not feasible.

In greater detail, using $P_{st} = P^2_{norm}$ where P_{norm} is a norm, leads to an active set, $S = \{j : \widehat{f}_j \not\equiv 0\}$, for which either $|S| = 0$ or $|S| = p$. If $P_{st} = P^2_{semi}$, for a semi-norm P_{semi}, we can get $0 < |S| < p$; however, now all $\widehat{f}_j \in \mathcal{F}_0$, where $\mathcal{F}_0$ is a *finite-dimensional* function class. In contrast, using $P_{st} = P_{semi}$ can give active sets such that $0 < |S| < p$ and each $\widehat{f}_j$ can be modeled nonparametrically.

Before presenting our main results, we note that throughout this section we deal with finite valued semi-norms, this excludes convex indicator penalties. It should be noted that many convex indicator penalties impose a parametric structure, thus excluding them from

the discussion of this section. Other convex indicator penalties are challenging to deal with in this context and other settings (see for example, Remark 18). We now present our result in Lemma 1 for a squared norm penalty followed by the extension to squared semi-norms in Corollary 2 (proofs in Appendix D).

Lemma 1 *Let $\mathcal{F}$ be a nonparametric function class in the following sense: for any covariate-response pair $(\boldsymbol{y}, \boldsymbol{x}) \in \mathbb{R}^{n\times 2}$ there exists some $f \in \mathcal{F}$ such that $f(x_i) = y_i$ for all i. Consider the optimization problem*

$$\widehat{f}_1, \ldots, \widehat{f}_p \leftarrow \underset{f_1,\ldots,f_p\in\mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} f_j(x_{ij})\right)^2 + \sum_{j=1}^{p}\left\{\lambda_{st} P_{st}^2(f_j) + \lambda_{sp}\|f_j\|_n\right\}, \quad (4)$$

for a norm P_{st} on $\mathcal{F}$. Then, for any λ_{sp}, either $\widehat{f}_j \equiv 0$ for all j or $\widehat{f}_j \not\equiv 0$ for all j.

Corollary 2 *For a semi-norm P_{st} in (4), we define its null set as $\mathcal{F}_0 \subset \mathcal{F}$ such that $P_{st}(f_0) = 0$ for all $f_0 \in \mathcal{F}_0$ (note that $\mathcal{F}_0$ contains the zero function). Consider an arbitrary, non-empty, index subset $I \subset \{1, \ldots, p\}$. If $\widehat{f}_{j'} = 0$ for all $j' \in I$, then for all $j \in I^c$, $\widehat{f}_j \in \mathcal{F}_0$.*

The above results imply that using a squared semi-norm means sacrificing either sparsity or flexibility in our modeling approach. In most cases the set $\mathcal{F}_0$ is parametric, e.g, a commonly used penalty $P_{st} = P_{sobolev}(f'')$, leads to $\mathcal{F}_0$ as the set of linear functions. Thus we can either fit sparse, parametric models or non-sparse, nonparametric models but *not both*. In other words, for parametric regression we can simply add a sparsity-inducing penalty; for example, the elastic net (Zou and Hastie, 2005), which adds a sparsity penalty to ridge regression, leading to sparse linear models. In contrast, simply adding a sparsity penalty to traditional GAMs (Hastie and Tibshirani, 1990) is not sufficient because, fitted models will not be sparse GAMs.

2.3 Relationship of Existing Methods to `GSAM`

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

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

where $\boldsymbol{\psi}_{jm} = [\psi_m(x_{1j}), \ldots, \psi_m(x_{nj})]^\top \in \mathbb{R}^n$ for basis functions $\psi_1, \ldots, \psi_M$. This is a `GSAM` with $P_{st} = \mathbb{I}(f; \operatorname{span}\{\psi_1, \ldots, \psi_M\})$. 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

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

and is also a GSAM 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 Proj_A is the projection operator onto the set A. The recently proposed extensions of trend filtering to additive models are other examples (Petersen et al., 2016; Sadhanala and Tibshirani, 2019); these methods can be written in our GSAM framework with $P_{st}(f) = TV(f)$.

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 GSAM; they note that this proposal gives a suboptimal rate of convergence. The second is a GSAM of the form (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 one was known to them. In a follow-up paper, van de Geer (2010) studied the theoretical properties of a GSAM 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

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

for a smoothness parameter m.

Koltchinskii and Yuan (2010), Raskutti et al. (2012) and Yuan and Zhou (2016) discuss a similar framework to GSAMs; 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. Recently, Tan and Zhang (2019) studied GSAMs beyond the RKHS framework similar to this manuscript, however, there are important differences. Firstly, Tan and Zhang (2019) consider only the least squares loss. Secondly, the authors do not present an algorithm for estimation; instead they prove that under a least squares loss and special smoothness penalties the solution to the optimization problem is finite dimensional. Thirdly, their results (like the minimax rates proved by Yuan and Zhou, 2016) include the more general notion of *weak sparsity* (van de Geer, 2016); however, extending this notion beyond the least squares loss is left for future research. All of the above mentioned proposals either fail to provide an efficient computational algorithm or have sub-optimal convergence rates. There are also a number of other proposals that do not quite fall in the GSAM framework (Chouldechova and Hastie, 2015; Fan et al., 2012; Yin et al., 2012).

3. General-Purpose Algorithm

Here we give a general algorithm for fitting `GSAM`s 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 $\langle f, \dot{\ell}(g)\rangle_n \equiv n^{-1}\sum_{i=1}^n f(\boldsymbol{x}_i)\{\dot{\ell}(y_i, g(\boldsymbol{x}_i))\}$, $\mathbb{P}_n\dot{\ell}(g) \equiv n^{-1}\sum_{i=1}^n \dot{\ell}(y_i, g(\boldsymbol{x}_i))$ 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 at some arbitrary point $\beta^0 + \sum_{j=1}^p f_j^0$. 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:

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

which leads to the following majorizing inequality

$$
\begin{aligned}
-\mathbb{P}_n\ell\Big(\beta + \sum_{j=1}^p f_j\Big) \leq\ & \frac{(p+1)L}{2}\Big[\beta - \Big\{\beta^0 + \frac{1}{(p+1)L}\mathbb{P}_n\dot{\ell}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)\Big\}\Big]^2 \\
& + \sum_{j=1}^p \frac{(p+1)L}{2}\Big\|f_j - \Big\{f_j^0 + \frac{1}{(p+1)L}\dot{\ell}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)\Big\}\Big\|_n^2 + W,
\end{aligned} \tag{5}
$$

where W is not a function of β or f_j for any j. Instead of minimizing the original problem (3), we minimize the majorizing surrogate

$$
\begin{aligned}
\frac{1}{2}\Big[\beta - \Big\{\beta^0 + t\mathbb{P}_n\dot{\ell}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)\Big\}\Big]^2 + \frac{1}{2}\sum_{j=1}^p \Big\|f_j - \Big\{f_j^0 + t\dot{\ell}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)\Big\}\Big\|_n^2 & \\
+ t\lambda^2\sum_{j=1}^p P_{st}\left(f_j\right) + t\lambda\sum_{j=1}^p \left\|f_j\right\|_n, &
\end{aligned} \tag{6}
$$

where $t = \{(p+1)L\}^{-1}$. Minimizing (6) and re-centering our Taylor series at the current iteration, is precisely the proximal gradient recipe. Updating the intercept β, is simply $\widehat{\beta} \leftarrow \beta^0 + t\mathbb{P}_n\dot{\ell}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)$. Components $f_1, \ldots, f_p$, can be updated in parallel by solving the univariate problems:

$$
\widehat{f}_j \leftarrow \underset{f\in\mathcal{F}}{\operatorname{argmin}}\ \frac{1}{2}\Big\|\Big\{f_j^0 + t\dot{\ell}\Big(\beta^0 + \sum_{j=1}^p f_j^0\Big)\Big\} - f\Big\|_n^2 + t\lambda^2 P_{st}\left(f\right) + t\lambda\left\|f\right\|_n. \tag{7}
$$

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

$$\underset{f\in\mathcal{F}}{\operatorname{argmin}}\ \frac{1}{2}\|\boldsymbol{r}-f\|_n^2+\lambda_1 P_{st}(f)+\lambda_2\|f\|_n, \tag{8}$$

$$\underset{f\in\mathcal{F}}{\operatorname{argmin}}\ \frac{1}{2}\|\boldsymbol{r}-f\|_n^2+\lambda_1 P_{st}(f). \tag{9}$$

If $\widetilde{f}$ is a solution to (9); then $\widehat{f}$ is a solution to (8) where $\widehat{f}$ is defined as

$$\widehat{f}=\Big(1-\lambda_2/\|\widetilde{f}\|_n\Big)_+\widetilde{f}, \tag{10}$$

with $(z)_+=\max(z,0)$. Additionally, if $\lambda_2\geq\|\boldsymbol{r}\|_n$, then $\widehat{f}\equiv 0$.

The proof is given in Appendix E. Using Lemma 3, we can get the solution to (7) by solving a problem in the form of (9), a classical univariate smoothing problem, and then applying (10), the simple soft-scaling operator. Putting this together, leads to Algorithm 1 below for solving (3). The general recipe used to derive Algorithm 1, is the well known proximal gradient descent algorithm. Thus, well established convergence results in the literature can be used (see for example, Beck and Teboulle, 2009b), under mild conditions: we require a convex ℓ with Lipschitz first derivative. These conditions hold for many loss functions particularly the negative log-likelihood of exponential family distributions. Convergence of the infinite-dimensional optimization over $\mathcal{F}$ follows from the finite-dimensional analog of problem (3); we prove this in Appendix E.

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 (5) 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. While a fixed step size t ensures theoretical convergence, in practice we can achieve a substantial speedup by adaptively selecting t for each iteration. We could use $t_k=\{L(p^k_{active}+1)\}^{-1}$, where p^k_{active} is the number of j for which either of f_j^{k-1} or f_j^k is non-zero. Since we are interested in sparse models, generally $p^k_{active}\ll p$, leading to a substantial efficiency gain. If we do not have a suitable bound for L, we could use a data-dependent scheme such as the backtracking line search (Beck and Teboulle, 2009b). The algorithm can also take advantage of Nesterov-style acceleration (Nesterov, 2007), which improves 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 calculation. We present the full details of the algorithm in Appendix A.

Algorithm 1 General Proximal Gradient Algorithm for (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

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

  Update $\beta^k \leftarrow \beta^{k-1} - t\sum_{i=1}^n r_i$.
  for $j = 1, \dots, p$ **do**
    Set

$$f_j^{inter} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{2} \left\| \left(f_j^{k-1} - t\boldsymbol{r} \right) - f \right\|_n^2 + t\lambda^2 P_{st}(f). \tag{11}$$

    Update

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

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

Algorithm 1 is developed for a given λ value. Alternatively, we recommend using a decreasing sequence of λ values, linear on the log scale starting at $\lambda_{\max} = \|\boldsymbol{y}\|_n$. Another computational consideration is the method of tuning parameter selection: our numerical experiments suggest K-fold cross validation as a suitable choice; however, we note that other tuning parameter selection techniques such as generalized cross validation, AIC and BIC can be used. Finding a theoretically optimal tuning parameter method, or proving the estimator selected by cross validation obtains the same fast convergence rate, is a challenging problem which we defer to future work.

As noted above, the main computational hurdle in Algorithm 1 is solving the univariate problem (9). In the following subsection, we discuss this step in greater detail for various smoothness penalties.

3.1 Solving the Univariate Subproblem

For many semi-norm smoothers there are already efficient solvers for solving (9): with the k-th order total variation penalty, (9) 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, (9) can be solved in $O(M^2 n)$ operations using linear regression; using a monotonicity indicator, (9) 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 (9); however, we might have fast algorithms for the slightly different optimization problem:

$$\widetilde{f}_{\widetilde{\lambda}} \leftarrow \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{2} \|\boldsymbol{r} - f\|_n^2 + \widetilde{\lambda} P_{st}^{\tau}(f), \tag{12}$$

for $\tau > 1$. For example, the k-th order Sobolev penalty (Wahba, 1990) can be solved exactly in $O(kn)$ operations for $\tau = 2$. In the following Lemma, we show that the solution to (12) can be leveraged to solve the harder problem (9).

Lemma 4 *Given an n-vector $\boldsymbol{r}$, a convex linear space $\mathcal{F}$ over the field $\mathbb{R}$, and real $\tau > 1$, consider the optimization problems:*

$$\begin{aligned}
\widehat{f}_{\lambda} &\leftarrow \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{2} \|\boldsymbol{r} - f\|_n^2 + \lambda P_{st}(f); \\
\widetilde{f}_{\lambda} &\leftarrow \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{2} \|\boldsymbol{r} - f\|_n^2 + \lambda P_{st}^{\tau}(f); \\
f_{null} &\leftarrow \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{2} \|\boldsymbol{r} - f\|_n^2 + \mathbb{I}(f \in \mathcal{F} : P_{st}(f) = 0); \\
f_{interp} &\leftarrow \operatorname*{argmin}_{f \in \mathcal{F}} P_{st}^{\tau}(f) + \mathbb{I}(r_i = f(x_i) \text{ for all } i),
\end{aligned}$$

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

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

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

$$P_{st}^*(f_\perp) = \sup \left\{ |\langle f_\perp, f'_\perp \rangle_n| : P_{st}(f'_\perp) \leq 1, f'_\perp \in \mathcal{F}_2 \right\}. \tag{13}$$

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 E. 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 $\tau \widetilde{\lambda} P_{st}^{\tau-1}(\widetilde{f}_{\widetilde{\lambda}}) = \lambda$ 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 $\lambda_{max} \equiv \|f_{interp}\|_n$ (by Lemma 3), 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 $\boldsymbol{r}$. To assess the computational

impact of an added grid search, we looked at the run-time for the proximal problem with $P_{st} = P_{sobolev}$ (which requires a grid search) and with $P_{st} \in \{TV(f^{(0)}), TV(f^{(1)}), TV(f^{(2)})\}$ (which does not use Lemma 4). For 100 replications of the proximal problem on a quad-core Intel®Core™, i7-10510U CPU @1.80GHz, the median run-time with $n = 500$ for $P_{st} = P_{sobolev}$ was 693.20 μs. In contrast, the median run-time for $P_{st}(f) = TV(f^{(k)})$ for $k = 0, 1, 2$ was 514.15, 2968.30, 4884.90 μs, respectively. These median run-times were calculated via a small simulation study; details of this experiment along with detailed timing results are presented in Appendix B.

To complete the discussion, we give the explicit form of the dual norm (13) for the case where $P_{st}(f) = \|D\vec{f}\|_q$ for a matrix $D \in \mathbb{R}^{M\times n}$, a vector $\vec{f} = [f(x_1), \ldots, f(x_n)]^\top \in \mathbb{R}^n$, and $q \geq 1$. Such penalties are common in the literature, for example, when P_{st} is the Sobolev semi-norm, total variation norm, or any RKHS norm. For $P_{st}(f) = \|D\vec{f}\|_q$, the dual norm is given by

$$P^*_{st}(f) = \|D(D^\top D)^- \vec{f}\|_{\widetilde{q}},$$

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

4. Theoretical Results

Here we prove rates of convergence for `GSAM`s, estimators that fall within our framework (3). We first present the so-called *slow rates*, which require few assumptions, followed by *fast rates*, which require compatibility and margin conditions (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., 2017). 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. We then 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. Throughout, we also 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.

4.1 Definitions and Notation

We consider here properties of the solution to

$$\widehat{\beta}, \widehat{f}_1 \ldots, \widehat{f}_p \leftarrow \underset{\beta\in\mathcal{R}, \{f_j\}_{j=1}^p \in \mathcal{F}}{\arg\min} - \mathbb{P}_n \ell\Big(\beta + \sum_{j=1}^p f_j\Big) + \lambda \sum_{j=1}^p \left\{\|f_j\|_n + \lambda P_{st}\left(f_j\right)\right\}, \tag{14}$$

where $\mathcal{R} \subseteq \mathbb{R}$ and $\mathcal{F}$ is some univariate function class. Note that in (14) 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

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

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

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

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 7; 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

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

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 5 (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*

$$H(\delta, \{f_j \in \mathcal{F} : \|f_j\|_n + \gamma P_{st}(f_j) \leq 1\}, \|\cdot\|_{Q_{j,n}}) \leq A_0 T_n \log\left(1/\delta + 1\right), \tag{15}$$

for some constant A_0, and parameter T_n.

Definition 6 (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*

$$H(\delta, \{f_j \in \mathcal{F} : \|f_j\|_n + \gamma P_{st}(f_j) \leq 1\}, \|\cdot\|_{Q_{j,n}}) \leq A_0 (\delta\gamma)^{-\alpha}, \tag{16}$$

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 (15) 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 (16) commonly holds for broader function classes, for example, 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 $B \precsim A$.

4.2 Main Results

We now present our main results: upper bounds for the excess risk of `GSAM`s, specifically, 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 7 (Slow Rates for `GSAM`) *Let $\widehat{f} = \widehat{\beta} + \sum_{j=1}^p \widehat{f}_j$ be as defined in (14), 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*

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

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

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

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

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

Corollary 8 *Let $\widehat{f}$, f^* and $\mathcal{R}$ be as defined in Theorem 7. Assume that for any function f the loss $\ell(\cdot)$ is such that*

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

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 independent, uniformly sub-Gaussian:

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

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

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

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

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 with smoothness, then for* $\lambda \asymp \rho \asymp \kappa \max\left(n^{-\frac{1}{2+\alpha}}, \sqrt{\frac{\log p}{n}}\right)$,

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

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

In the above corollary, the assumption $I(f^*) = O(1)$ is often reasonable in high-dimensions; if omitted, with the same high probability, the above rates will be multiplied by the term $\sum_{j=1}^p \|f_j^*\|_n$. Now for high-dimensions, we commonly assume sparsity, $f^* = \sum_{j \in S} f_j^*$, where $|S|$ is small. The dependence of the rate on sparsity can be directly expressed by the inequality $\sum_{j \in S} \|f_j^*\|_n \leq |S| \max_{j \in S} \|f_j^*\|_n$. Another possible assumption for high dimensions is *weak sparsity*, which states that, the effect size of most component functions is very small. In this case, the preceding inequality would not be tight but we essentially have $\sum_{j=1}^p \|f_j^*\|_n = O(1)$.

We now proceed to show the fast rates of convergence. To establish these rates, we require the *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 for example, Negahban et al., 2011; van de Geer, 2008).

Definition 9 (Compatibility Condition) *The compatibility condition is said to hold for an index set $S \subset \{1, 2, \ldots, p\}$, with compatibility constant $\phi(S) > 0$, 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 $\sum_{j \in S^c} \|f_j\|_n + \gamma \sum_{j=1}^p P_{st}(f_j) \leq |\beta| + 3 \sum_{j \in S} \|f_j\|_n$, it holds that*

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

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

Definition 10 (Margin Condition) *The margin condition holds if there is strictly convex function G such that $G(0) = 0$ and for all $f \in \mathcal{F}_{local}^0 \subset \mathcal{F}^0$ we have*

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

for some norm $\|\cdot\|$ on $\mathcal{F}^0$; here $\mathcal{F}_{local}^0$ is a neighborhood of f^0 (for example, $\mathcal{F}_{local}^0 = \{f : \|f - f^0\|_\infty \leq \eta\}$). 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 norm $\|\cdot\|$ used in the definitions above is most often the empirical norm, $\|\cdot\|_n$. Our proof is the same for any norm $\|\cdot\|$, as long as the same norm is used for both conditions. Note that the margin condition is strictly a condition on the loss function $\ell(\cdot)$, implying that it is not dependent on the class, $\mathcal{F}$, or dimension, p. While the margin condition is established for well-known choices of $\ell(\cdot)$ (see for example, van de Geer, 2016), in Appendix H, we present a framework for verifying the quadratic margin condition for loss functions of the form: $-\ell(f) = ay_i f(\boldsymbol{x}_i) + b(f(\boldsymbol{x}_i))$. While the compatibility condition is difficult to prove, the *theoretical compatibility condition* (defined below) can be verified under suitable conditions. In Appendix H, we prove that (under mild conditions), the theoretical compatibility condition implies the original compatibility condition with high probability.

Definition 11 (Theoretical Compatibility Condition) *The theoretical compatibility condition is said to hold for an index set $S \subset \{1, 2, \ldots, p\}$, for a compatibility constant $\widetilde{\phi}(S)$, if for some $\eta \in (0, 1/5)$, all $\lambda > 0$, and all functions of the form $f(\boldsymbol{x}) = \beta + \sum_{j=1}^{p} f_j(x_j)$ that satisfy $\sum_{j \in S^c} \|f_j\| + \frac{1-5\eta}{1-\eta} \sum_{j=1}^{p} \lambda P_{st}(f_j) \leq |\beta| + \frac{3(1+\eta)}{1-\eta} \sum_{j \in S} \|f_j\|$, it holds that*

$$|\beta| + \sum_{j \in S} \|f_j\| \leq \frac{\sqrt{|S|}\|f\|}{\widetilde{\phi}(S)},$$

where $\|f\|^2 = \int [f(\boldsymbol{x})]^2 dQ(\boldsymbol{x})$ is the population level L_2 norm.

The theoretical compatibility condition holds trivially when we have independent covariates. In general, establishing verifying it depends on the smoothness penalty P_{st}; for example, for the Sobolev norm, Meier et al. (2009) established sufficient conditions for the compatibility condition to hold. An important special case, is when $P_{st}(f)$ projects component functions to a finite dimensional space (for example, Ravikumar et al., 2009; Lou et al., 2016). In this case, our condition reduces to the well-known, group lasso compatibility condition, for which sufficient conditions are well established in the literature (for example, Bühlmann and van de Geer, 2011).

We now present our second theorem which 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 7, 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 7, the bound holds over a set $\mathcal{T}$; Corollary 13 following the theorem shows that we lie in $\mathcal{T}$ with high probability.

Theorem 12 (Fast Rates for `GSAM`) *Suppose $-\ell(\cdot)$ and P_{st} are convex functions and with $\widehat{f}$ and f^* as defined in Theorem 7. Assume that f^* is sparse with $|S_*| = s$ where $S_* = \{j : f_j^* \neq 0\}$, and that the compatibility condition holds for S_*. Further assume the quadratic margin condition holds with constant c, and that for a function $f(\boldsymbol{x}) = \beta + \sum_{j=1}^{p} f_j(x_j)$, $f \in \mathcal{F}_{local}^0$ if and only if $|\beta - \beta^*| + I(f - f^*) \leq M^*$. The constant M^* is defined as*

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

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 $\mathcal{T}$,

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

Corollary 13 *Let $\widehat{f}$ and f^* be as defined in Theorem 7 and assume the conditions of Theorem 12. Furthermore, for any function f assume the loss $\ell(\cdot)$ is such that*

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

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 independent, uniformly sub-Gaussian:

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

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

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

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

   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 $\lambda \asymp \rho \asymp \kappa \max\left(n^{-\frac{1}{2+\alpha}}, \sqrt{\frac{\log p}{n}}\right)$,*

$$\mathcal{E}(\widehat{f}) + \lambda I(\widehat{f} - f^*) \precsim \max\left(s n^{-\frac{2}{2+\alpha}}, s\frac{\log p}{n}\right), \tag{17}$$

   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 these specializations, we conclude this section by further generalizing Theorem 12. We will now assume a more general margin condition, for which we need to define the additional notion of a *convex conjugate*.

Definition 14 (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*

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

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

Theorem 15 (Fast Rates) *Assume the conditions of Theorem 12 and define M^* as*

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

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

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

Remark 16 (Additional tuning parameters) *Note that our convergence rates include the term $\sum_{j \in S_*} P_{st}(f_j^*)$, or constants which depend on it. For some choices of P_{st} this can lead to poor finite sample performance. In such cases, prediction performance can be improved by solving instead*

$$\widehat{\beta}, \widehat{f}_1, \ldots, \widehat{f}_p \leftarrow \underset{\beta \in \mathbb{R}, f_1, \ldots, f_p \in \mathcal{F}}{\operatorname{argmin}} - \mathbb{P}_n \ell\left(\beta + \sum_{j=1}^p f_j\right) + (1-\zeta)\lambda \sum_{j=1}^p P_{st}(f_j) + \zeta\lambda \sum_{j=1}^p \|f_j\|_n, \qquad (18)$$

where $\zeta \in [0, 1]$ is an additional tuning parameter. In theory, using two tuning parameters should lead to improved prediction, however in practice, tuning parameter selection over a discrete grid can become computationally cumbersome. A moderately-sized search grid might not yield a lower MSE and in fact, can lead to substantially higher MSE, particularly for large n. We illustrate this phenomenon via a small simulation study in Appendix C: to the simulation study of Section 5, specifically Scenarios 3 and 4, we additionally fit `GSAM`*s by solving* (18)*. Using a grid of ten ζ values, we observe improved prediction in some cases however, a ten-grid is too coarse to exhibit uniformly lower MSE.*

Remark 17 (Constants in convergence rates) *Our convergence rates are presented* up to constants*. To illustrate this fact, we consider the problem*

$$\widehat{\beta}, \widehat{f}_1, \ldots, \widehat{f}_p \leftarrow \underset{\beta \in \mathbb{R}, f_1, \ldots, f_p \in \mathcal{F}}{\operatorname{argmin}} - \mathbb{P}_n \ell\left(\beta + \sum_{j=1}^p f_j\right) + \Theta\lambda^2 \sum_{j=1}^p P_{st}(f_j) + \lambda \sum_{j=1}^p \|f_j\|_n, \qquad (19)$$

for a constant Θ that does not depend on n or p. While (19) *will have a different convergence rate than those presented in Theorems 7—15, the two rates will only differ by a constant that depends on Θ. Optimizing these constants is an interesting open problem that is beyond the scope of this manuscript.*

Remark 18 (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 19 (Polynomial Entropy without Smoothness) *The univariate function class, $\mathcal{F}$, is said to have a polynomial entropy without smoothness bound if for all $j = 1, \ldots, p$ we have*

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

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.

Remark 20 (Sub-Exponential residuals) *In Corollaries 8 and 13 we can replace the requirement of uniformly sub-Gaussian residuals by the weaker condition of* uniformly sub-Exponential residuals. *To be precise, we would require*

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

However, sub-Exponential residuals would firstly require bounds for the δ-entropy with bracketing. *The δ-entropy with bracketing is a stronger notion than δ-entropy without bracketing (the δ-entropy with bracketing is always larger than the δ-entropy without bracketing). Secondly, we would also require uniform bounds for each univariate function, specifically, we need*

$$\max_{j=1,\ldots,p} \sup_x |f_j(x)/\|f_j\|_n| \leq R.$$

Remark 21 (Comparison of rates to existing work) *Here, we highlight the key differences between our theoretical result and existing work. Koltchinskii and Yuan (2010) and Raskutti et al. (2012) establish same rates of convergence as those in Corollaries 8 and 13. However, their work requires stronger assumptions. Both papers are restricted to the setting reproducing of kernel hilbert spaces (RKHS) and only allow an RKHS norm as the choice of smoothness penalty. They also assume known bounds on the additive functions or individual components. Additionally, Raskutti et al. (2012) assumes independence of covariates as opposed to a more general compatibility condition. The work of Yuan and Zhou (2016), extends the RKHS framework with results capturing the notion of* weak sparsity*; assuming bounds on the term $\sum_{j=1}^p \|f_j^*\|_n^q$ for $0 \leq q < 1$ they present rates (up to a constant) of the form*

$$n^{-\frac{2}{2+\alpha}} + \left(\frac{\log p}{n}\right)^{1-q/2}. \tag{20}$$

Similarly, Tan and Zhang (2019) present rates of the form

$$\left(n^{-\frac{1}{2+\alpha(1-q)}} + \sqrt{\frac{\log p}{n}}\right)^{2-q}. \tag{21}$$

Both (20) *and* (21) *match our established rates for the limiting case of $q = 0$. While Tan and Zhang (2019) relax some of the strong assumptions of Yuan and Zhou (2016), both*

papers deal exclusively with the least squares loss function. It is not clear if results like (20) *and* (21) *can be established for general loss functions.*

Tan and Zhang (2019) also present convergence rates in greater generality, namely, in terms of the integral of the entropy number. However the only special case they consider are function classes with polynomial entropy with smoothness. For this special case, their convergence rates match ours under a least squares loss; it is not clear if their results extend to GAMs nor is it clear what their rates are for other commonly used function classes.

4.3 Special Cases of `GSAM`

In this subsection, we illustrate the main strength of our framework, namely its generalizability. We specialize our theoretical results to, various existing `GSAM`s, fully non-parametric regression, and also to (sparse-)GLMs. As per a reviewer's suggestion, in Table A.1 in Appendix A, we summarize existing `GSAM`s and their limitations which our framework overcomes.

Firstly, the proposals of Ravikumar et al. (2009) and Lou et al. (2016), established convergence rates substantially slower than the minimax rates and only for the least squares loss. Our general framework, establishes the following convergence rates for both methods: $\mathcal{E}(\widehat{f}) \precsim \max(sM/n, s\log p/n) + \mathcal{E}(f^*)$, where M is the order of the basis expansion used for each $\widehat{f}_j$. For an additive f^0, $\mathcal{E}(f^*)$ is decreasing in M; and we require a value of M which balances the two terms in the rate. For function classes with polynomial entropy with smoothness, we recover rates (17) with $M \asymp n^{\frac{\alpha}{2+\alpha}}$. As noted in Section 4.2, the margin condition holds for a large class of loss functions; for both methods, the compatibility condition reduces to the well-studied, group lasso compatibility condition.

Next we consider the proposal of Meier et al. (2009) (see also Bühlmann and van de Geer, 2011; van de Geer, 2010); their theoretical results were limited to the least squares loss and the resulting convergence rate takes the form $\mathcal{E}(\widehat{f}) \precsim (s\log p/n)^{2/(2+\alpha)}$. This rate is sub-optimal compared to our fast rate (17). 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. As another example, we consider our own previous work (Haris et al., 2018), a `GSAM` which uses wavelet basis functions. Once the univariate problem ($p = 1$) was solved for the wavelet bases, extending it to `GSAM` was trivially achieved using the results of this manuscript. The above examples demonstrate that not only do our theoretical results and proximal gradient descent algorithm improve existing results in the literature, but also, can be applied to fully develop any `GSAM` as long as we can solve the uni-variate problem of the form (12).

Next we show how some seemingly unrelated problems can also be treated as special cases of our framework. Firstly, we recover the special case of univariate nonparametric regression, that is, with $p = 1$: the compatibility condition trivially holds leading to the usual rates $\mathcal{E}(\widehat{f}) \precsim n^{-2/(2+\alpha)}$. Next, we recover the multivariate nonparametric regression problem: suppose we have a single (but multivariate) component function $f_1 : \mathbb{R}^p \to \mathbb{R}$. For various choices of P_{st}, the bound (16) holds with $\alpha = p/m$ for some smoothness parameter

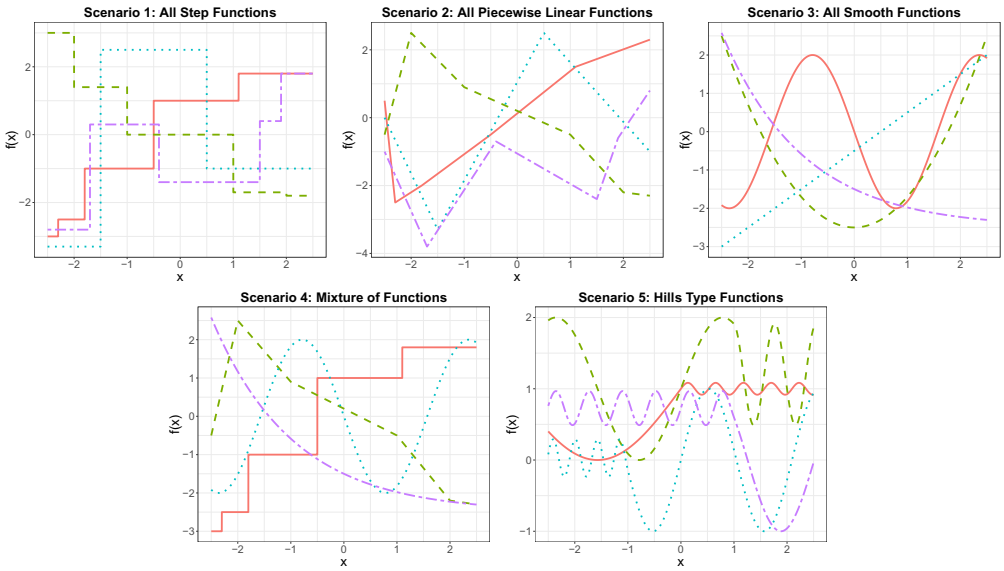

Figure 1: Plot of the 4 signal functions for each of the five simulation settings.

m. Again, the compatibility condition holds trivially, leading to the usual nonparametric rate $n^{-2m/(2m+p)}$. Finally, parametric regression models are also special cases of `GSAM`. 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 13 gives the usual parametric rate, p/n. For high-dimensional GLMs, not only does our theorem recover the lasso rate, but our compatibility condition also matches that of lasso (Bühlmann and van de Geer, 2011).

5. Simulation Study

In this section, to complement our theoretical results, we conduct a simulation study to study the finite sample performance of various `GSAM`s as a function of n. The `GSAM`s we study are existing techniques in the literature obtained by various choices of the smoothness penalty, $P_{st}(\cdot)$. Our aim is to study the convergence of various methods with increasing n. For a more detailed simulation study we refer the reader to the original papers for each method (cited below). We consider the following choices for $P_{st}(\cdot)$:

1. **SpAM (Ravikumar et al., 2009).** $P_{st}(f) = \mathbb{I}(f; \text{span}\{\psi_1, \ldots, \psi_M\})$ for $M \in \{3, 6, 10, 20, 30, 50, 80\}$. We use the `SAM` `R`-package (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). Given the lack of efficient software for this method, we implemented it using the algorithm and results of Section 3.

3. **TF (Sadhanala and Tibshirani, 2019).** $P_{st}(f) = \int_x |f^{(k+1)}(x)| \, dx$ for $k \in \{0, 1, 2\}$, trend filtering for additive models. We implemented this method using the algorithm of Section 3 where the univariate sub-problem (11) was solved using the `R` package `glmgen` (Arnold et al., 2014).

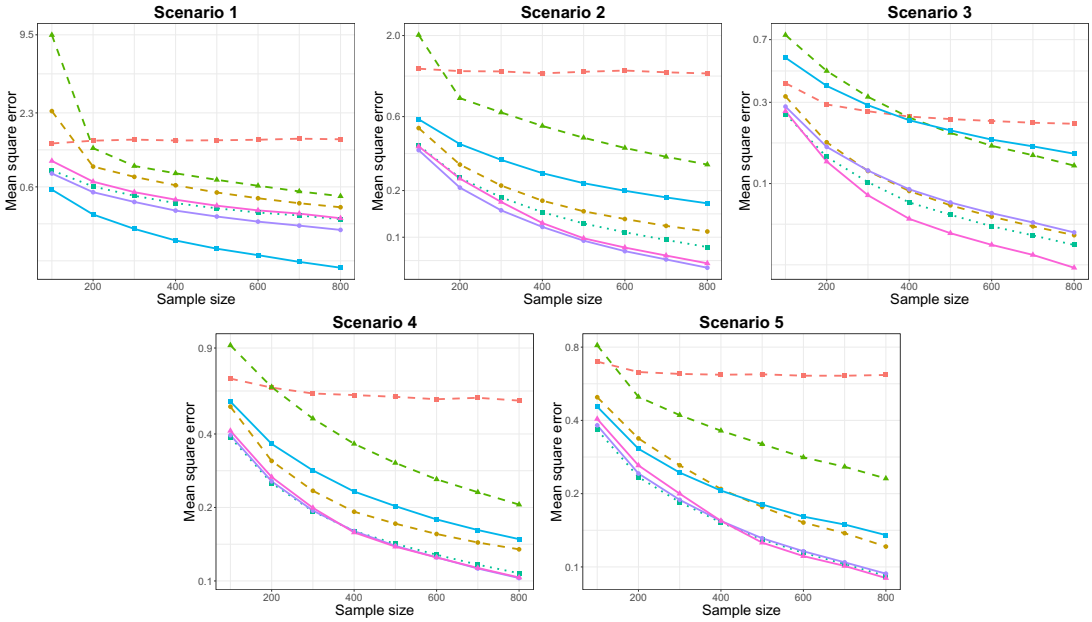

Figure 2: Plot of MSEs versus sample size for each of five scenarios for $p = 6$, averaged over 500 replications. The dashed lines correspond to `SpAM` with small (– ■ – ■ –), moderate (– ● – ● –) and high (– ▲ – ▲ –) number of basis functions. The solid lines correspond to trend filtering of order $k = 0$ (—■—■—), 1 (—●—●—) and 2 (—▲—▲—). SSP is represented by the dotted line (··■··■··).

We simulate data for each of five simulation scenarios as follows: Given a sample size n and a number of covariates p, we draw 50 different $n \times p$ training design matrices 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 is generated as $y_i = f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) + \varepsilon_i$ 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 ($p = 100$) settings. We consider five different choices of the signal functions as shown in Figure 1.

We fit each method over a sequence of 50 λ values on the training set, and select the tuning parameter λ^* which minimizes the test error ($\|y_{test} - \widehat{y}\|_n^2$). 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 2 and 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. The exact value of M presented varies by scenario, for example, in Scenario 4, low, moderate and high values of M correspond to $M = 3, 10$ and 30, respectively. In both low- and high-dimensional settings, we observe similar relative performances between the methods, with more variability in results for the high-dimensional setting. While there is no uniformly

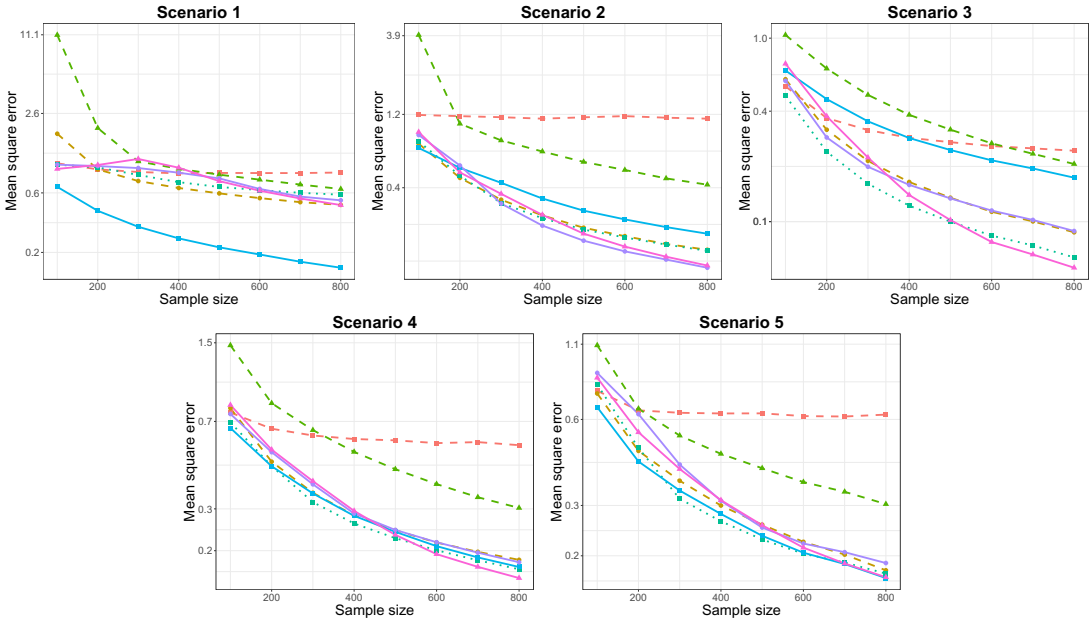

Figure 3: Plots of MSE versus sample size for each of five scenarios for $p = 100$, averaged over 500 replications. The line types and colors are the same as in Figure 2.

superior method, for all, except Scenario 1, the Sobolev smoothness penalty and trend filtering of orders 1 and 2 had comparably good performances. Unsurprisingly, trend filtering of order 0 exhibits superior performance in Scenario 1, where each component is piecewise constant. In each scenario, the bias-variance trade-off of SpAM depends on the choice of M: too small or large values of M lead to high prediction error compared to other methods.

In Appendix A, we plot examples of fitted functions for the various methods. The dependence on M for SpAM, is further illustrated in Figure A.1, where we plot functions estimated by SpAM for high-dimensional Scenario 4 with $n = 500$. We observe large bias 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 A.2 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.

6. Data Analysis

6.1 Boston Housing Data

We use the methods of Section 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., 2009;

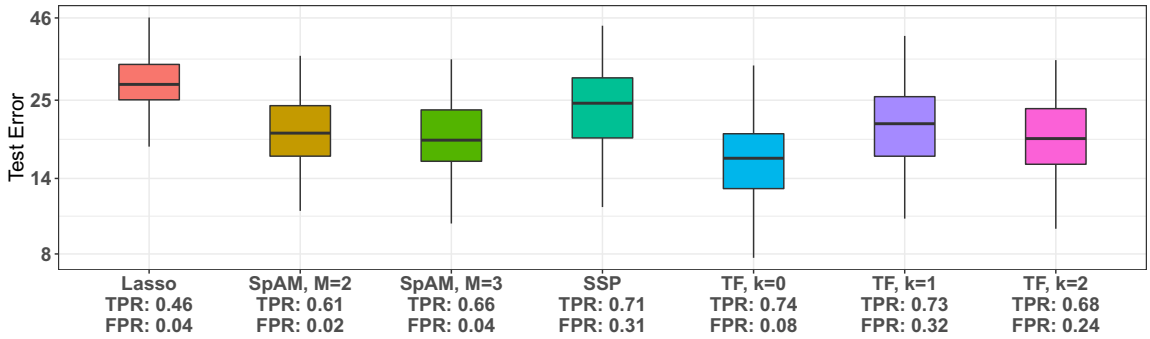

Figure 4: Box-plot of test errors for 100 different train/test splits of the data for each method. The average TPR and FPR was calculated using the original 10 covariates as 'signal' variables and remaining 20 as noise variables.

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 SSP, SpAM with $M = 2$ and 3 basis functions, and TF with orders $k = 0, 1, 2$; we also fit the lasso (Tibshirani, 1996). Approximately 75% of the observations are used as training set, and the mean square prediction error on the test set is reported. The final model is selected using 5-fold cross validation using the '1 standard error rule'. Results are 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 4. 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 4. The box-plots demonstrate superior performance of TF of order $k = 0$ over other methods in terms of lowest prediction error and highest TPR. The FPR of TF with $k = 0$ is also low (under 10%). In Figure A.3 of Appendix A, 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 data set. A striking feature of TF fits is that many component functions are constant for extreme values of the covariates.

6.2 Gene Expression Data

We now fit `GSAM`s for classification of gene expression data. We used the Curated Microarray Database (CuMiDa) (Feltes et al., 2019): a repository of gene-expression data sets curated from the Gene Expression Omnibus (GEO). Using gene expression measurements, we aimed to classify observations as either cancer cells or normal cells. We consider the following data sets/cancer types:

1. *Lung*: 54,675 gene expression measurements from 114 lung tissue samples from non-smoking women with non-small cell lung carcinoma; data set consists of 56 tumor, and 58 normal tissue samples. Available on CuMiDa with accession number GSE19804.

Method	Cancer type			
	Lung	*Prostate*	*Breast*	*Oral cavity*
	$n = 114; p = 54,675$	$n = 124; p = 12,620$	$n = 289; p = 35,980$	$n = 103; p = 54,675$
Lasso	0.985 (3.46)	0.713 (14.07)	0.951 (3.93)	0.930 (10.43)
SpAM, $M = 2$	0.982 (3.49)	0.726 (13.56)	0.948 (3.89)	0.923 (14.14)
SpAM, $M = 3$	0.984 (3.29)	0.763 (12.94)	**0.955** (3.50)	0.918 (13.06)
SpAM, $M = 10$	0.970 (5.48)	0.727 (13.46)	0.946 (4.08)	0.940 (8.17)
SSP	0.987 (2.72)	0.765 (11.71)	0.934 (4.69)	0.950 (7.08)
TF, $k = 0$	0.980 (4.07)	0.761 (13.04)	0.935 (4.66)	**0.953** (7.32)
TF, $k = 1$	**0.988** (2.65)	**0.771** (12.50)	0.936 (4.22)	0.947 (9.14)

Table 1: Table results for the analysis of gene expression data. For 100 different splits of the data into a training, testing and validation set, we report mean AUC along with $10^3\times$ mean SE, on the validation set. The method with the highest mean AUC is highlighted for each cancer type.

2. *Prostate*: 12,620 gene expression measurements from 124 prostate tissue samples; data set consists of 64 primary prostate tumor, and 60 normal tissue samples. Available on CuMiDa with accession number GSE6919_U95B.

3. *Breast*: 35,980 gene expression measurements from 289 breast tissue samples; data set consists of 143 breast adenocarcinoma, and 143 normal tissue samples. Available on CuMiDa with accession number GSE70947.

4. *Oral cavity*: 54,675 gene expression measurements from 103 mucosa cell samples; data set consists of 74 samples with oral cavity cancer, and 29 normal cell samples. Available on CuMiDa with accession number GSE42743.

Our goal is to correctly classify samples as either normal or cancer samples. We split the data as follows: 60% as training, 20% as validation and 20% as test data. On the training data we fit the lasso, SpAM with $M \in \{2, 3, 10\}$, SSP and, TF with $k \in \{0, 1\}$. TF with $k = 2$ was excluded due to numerical instability of the current implementation of the algorithm in `glmgen`; SpAM with other values of M yielded similar performance and thus the results are omitted here. All methods were fit for a sequence of λ values, using $(\lambda_{sp}, \lambda_{st}) = (\lambda, \lambda^2)$ for `GSAM`s. The λ value with the smallest area under the curve (AUC) for the ROC curve on the validation set was selected, and the corresponding model was used to classify samples in the test set. The experiment was repeated for 50 splits of the data into training, validation and testing.

In Table 1, we report the mean AUC on the test set and the estimated standard error based on 50 replications of the experiment. For the lung and breast cancer data sets, we observe similar performance between the lasso and other additive models; this suggests a low signal in the data to detect non-linearities. However, for prostate and oral cavity cancer, we observe a substantial gain (mean AUC beyond one SE) when using a `GSAM` instead of a linear model. In summary, this data analysis validates our intuition and theoretical results: using a `GSAM` will lead to comparable or better performance than using a linear model.

7. Conclusion

In this paper, 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, 2019; 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. The computational framework presented in this paper, effectively reduce the problem of fitting high-dimensional GSAMs to fitting a univariate regression model with the relevant smoothness penalty. While algorithms and theoretical results for specific GSAMs, as well as some theoretical results for certain types of general GSAMs, exists, to the best of our knowledge, the general algorithm for GSAMs is a key novel contribution in this paper. Our theoretical analyses in Section 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 `GSAM`, available on `https://github.com/asadharis/GSAM`, implements the methods described in this paper.

Acknowledgments

The authors gratefully acknowledge the support for this project from the National Science Foundation (NSF grant DMS-1915855) and the National Institute of Health (NIH grant R01GM114029). The authors would like to thank the anonymous reviewers for their insightful and helpful comments.

Appendix A. Additional Figures, Table and Algorithm

Algorithm A.1 is the block coordinate descent algorithm that can be used to estimate `GSAM`s.

Figure A.1 shows estimated functions for SpAM with 3,10 and 30 basis functions, and for SSP.

Figure A.2 shows estimated functions for trend filtering of orders $k = 0, 1, 2$.

Figure A.3 shows estimated functions for the various methods for the analysis of Boston housing data.

Table A.1 summary of existing `GSAM`s in the literature and their shortcomings (lack of efficient algorithms and/or limited theoretical results).

Algorithm A.1 Block Coordinate Descent for Least Squares Loss

Initialize $f_1^0, \dots f_p^0 \leftarrow \mathbf{0}, \beta^0 \leftarrow 0$, $\boldsymbol{r} \leftarrow \boldsymbol{y}$, $k \leftarrow 1$
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 $j = 1, \dots, p$ **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

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

    Update $\boldsymbol{r}$ to

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

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

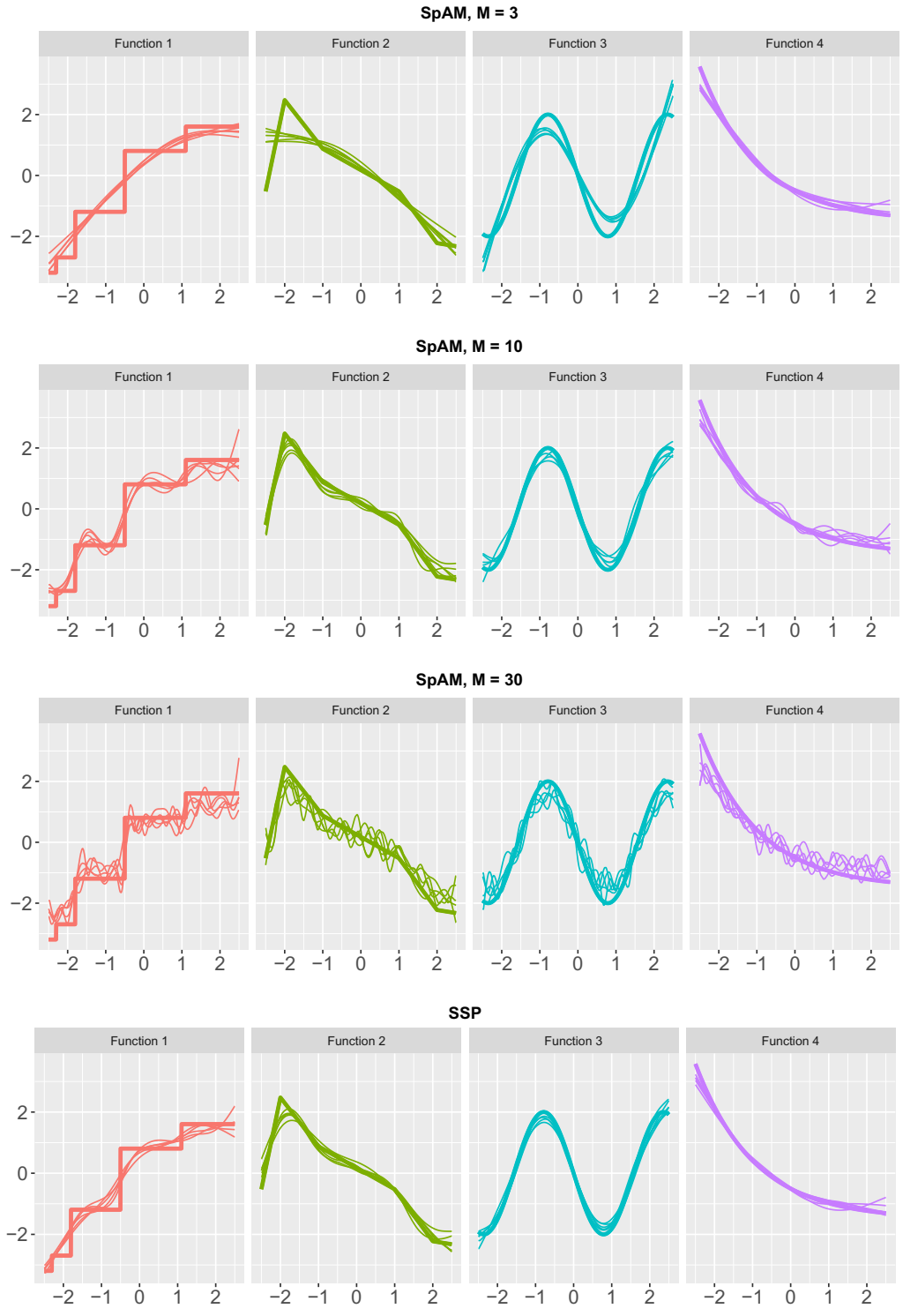

Figure A.1: Examples of estimated signal functions by SpAM (Ravikumar et al., 2009) and Sobolev Smoothness Penalty (Meier et al., 2009) for Scenario 4.

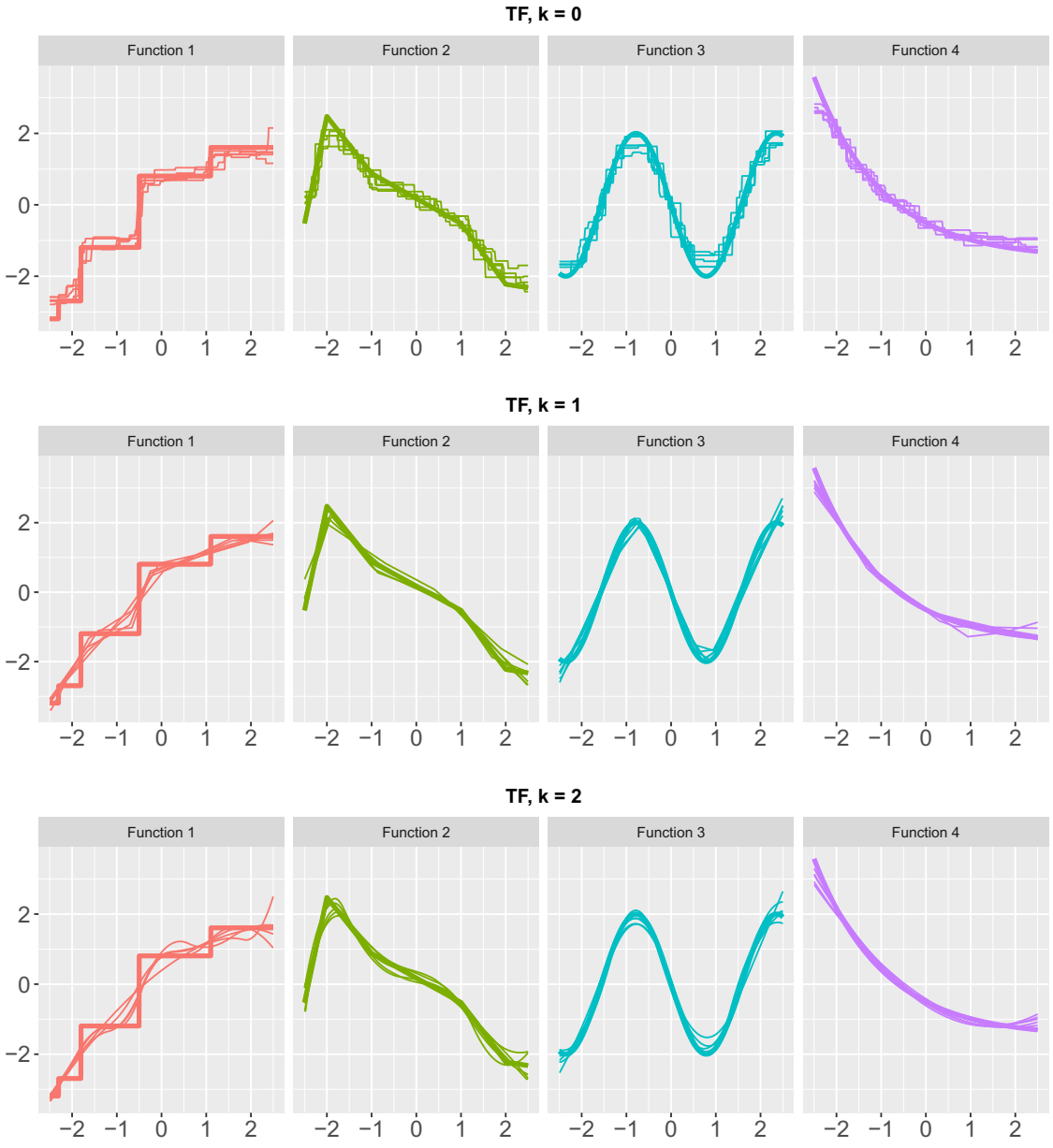

Figure A.2: Examples of estimated signal functions by Trend Filtering (Sadhanala and Tibshirani, 2019) for Scenario 4.

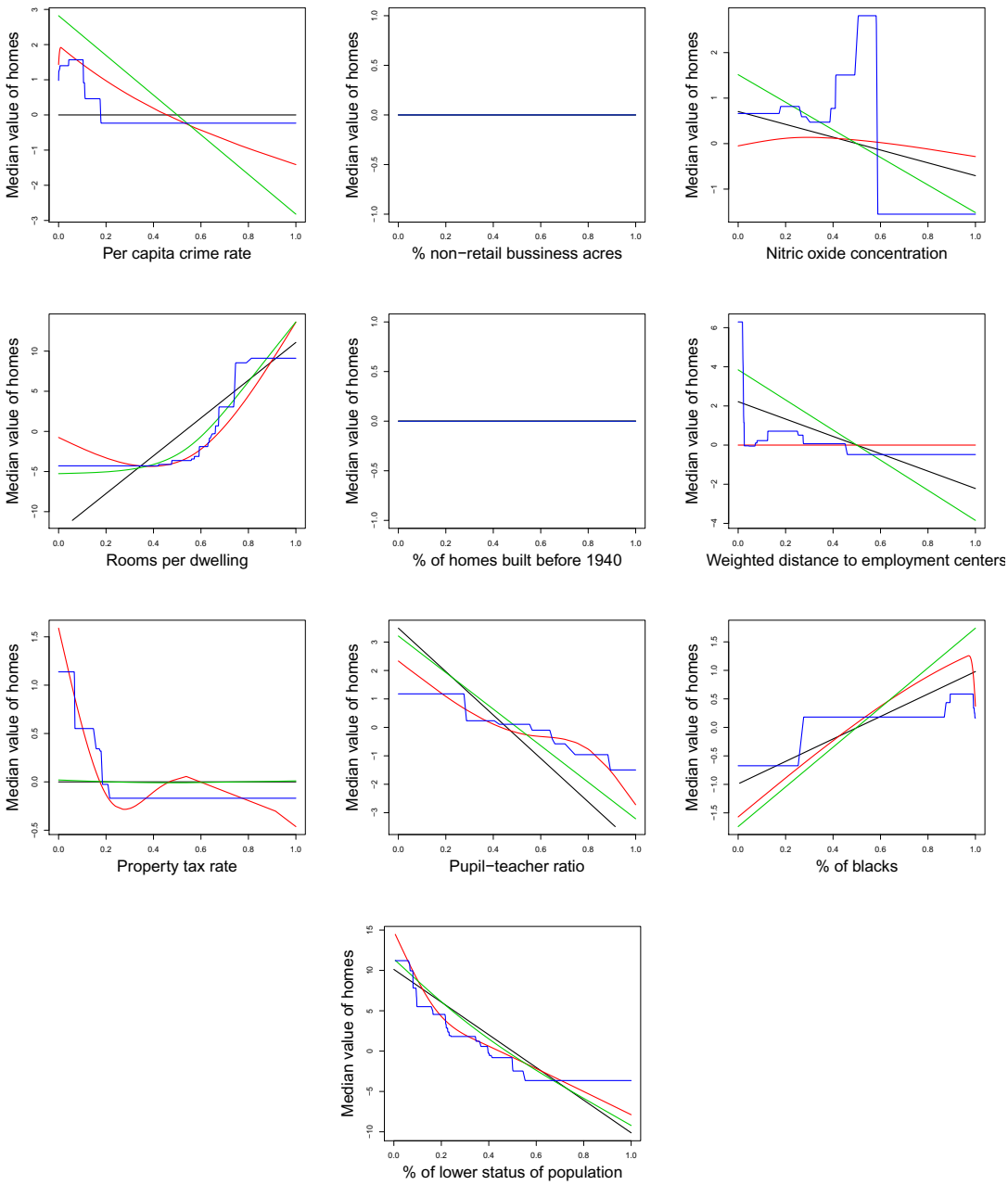

Figure A.3: Plots of fitted functions for the original 10 covariates for a single split of the data into training and test sets for lasso (——), SpAM (——) with $M = 3$ basis functions, SSP (——) and, TF (——) of order $k = 0$.

		Method	Function class $\mathcal{F}$	Loss/link‡ $\ell(\cdot)$	Smoothness Penalty $P_{st}(\cdot)$
No Efficient Algorithm	*Limited Theoretical Results*	Ravikumar et al. (2009)	Sobolev space	Least squares and logistic†	$\mathbb{I}(f; \text{span}\{\psi_1, \ldots, \psi_M\})$
		Lou et al. (2016)	General class of functions	Convex loss†	$\mathbb{I}(f; \text{span}\{\psi_1, \ldots, \psi_M\}) + \sum_{j=1}^{p} \| \text{Proj}_{\text{span}(\psi_2,\ldots,\psi_M)}(f) \|_n$
		Petersen et al. (2016)	General class of functions	Convex loss†	$\text{TV}(f)$
		Sadhanala and Tibshirani (2019)	Functions with finite k-th order total variation	Least squares	$\text{TV}(f^{(k)})$
		Meier et al. (2009)	General class of smooth functions	Least squares and logistic†	$\sqrt{\int_x \{f^{(2)}(x)\}^2 dx}$
		Raskutti et al. (2012)	RKHS	Least squares	Norm of RKHS space
		Yuan and Zhou (2016)	RKHS	Least squares	Quasi-norm of RKHS space
		Tan and Zhang (2019)	General class of functions	Least squares	General smoothness semi-norms
		Koltchinskii and Yuan (2010)	RKHS	Convex loss	Norm of RKHS space

†: Theoretical results only available for least squares loss/identity link function.

‡: The link function is generally incorporated into the loss function, for example, least squares loss corresponds to the identity link and logistic loss to the logit link.

Table A.1: Summary of methods generalized by our `GSAM` framework. The various methods in the literature are grouped according to (a) limited theoretical results (only for least squares or sub-optimal rates) and (b) absence of an efficient algorithm.

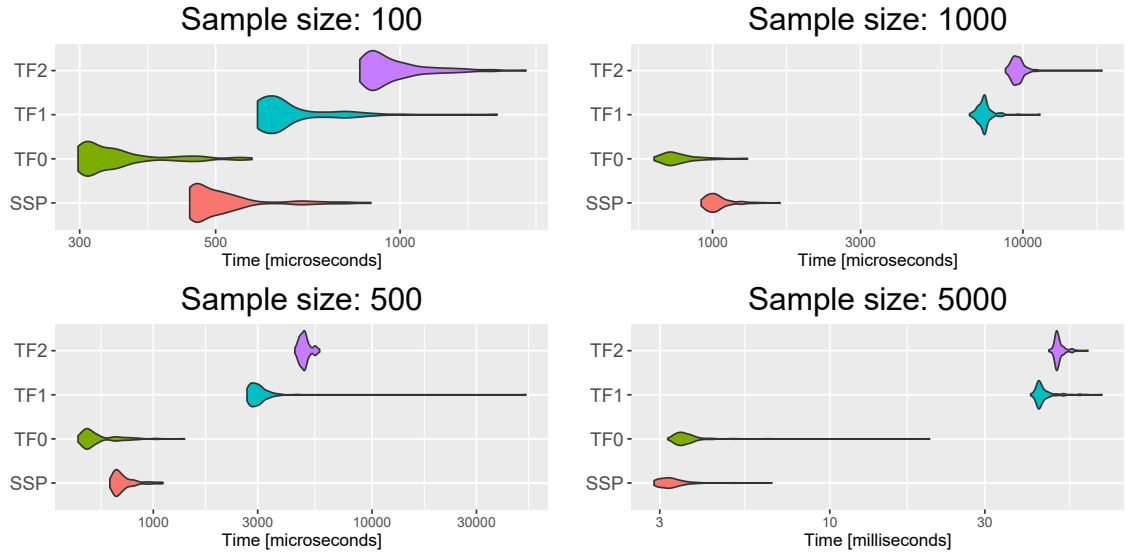

Figure B.1: Timing results for 100 implementations of the proximal problem for SSP and trend filtering.

Appendix B. Numerical Experiments for Comparing Run-Times

In this appendix, we present a detailed comparison of the run-times for solving the univariate proximal problem for various choices of the smoothness penalty, P_{st}. In greater detail, we study the proximal problem:

$$\min_{f \in \mathcal{F}} \frac{1}{2}\|\boldsymbol{r} - f\|_n^2 + \lambda \|f\|_n + \lambda^2 P_{st}(f).$$

We generated data as $x_i \sim Unif[-1, 1]$ and $r_i = \sin(1.5\pi x_i)/2 + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, 0.5^2)$ for $i = 1, \ldots, n$. We considered sample sizes $n = 100, 500, 1000$ and 5000.

In Table B.1 we present the timing results for repeatedly implementing the proximal solver 100 times for $n = 500$. Unsurprisingly, SpAM is the fastest method as it can be viewed as a standardized group lasso problem (Simon and Tibshirani, 2012) where each proximal problem has a closed-form solution. The SSP penalty is slightly slower than trend filtering for order $k = 0$ but still orders of magnitude faster than trend filtering for $k = 1, 2$.

In Figure B.1, we present violin plots for the timing results comparing SSP to trend filtering. This figure validates our previous observations: despite the added grid search, SSP is still much faster than trend filtering of order $k = 1, 2$. Another interesting feature is that SSP is the fastest method for $n = 5000$.

While it is encouraging to see competitive computation time despite an added grid search, we must highlight one limitation of this experiment: each method was implemented based on existing code in R packages, and other publicly available sources. Therefore, there are bound to be differences in the efficacy of the code written including the choice of programming language used. For instance the SSP proximal problem is solved using

	Time (μs)					
	Min	Lower Quantile	Mean	Median	Upper Quantile	Max
SpAM, $M = 3$	23.70	26.10	36.61	30.30	42.95	138.20
SpAM, $M = 5$	25.30	28.90	40.15	32.90	42.20	143.60
SpAM, $M = 10$	31.80	35.05	44.20	39.15	49.25	102.60
SpAM, $M = 15$	38.00	42.95	55.48	46.85	56.90	202.00
SpAM, $M = 20$	42.50	47.55	57.31	51.60	61.75	163.70
SpAM, $M = 30$	55.00	61.00	75.79	69.25	77.10	248.20
SpAM, $M = 50$	77.00	85.10	100.75	92.35	102.60	225.50
SSP	633.00	670.20	733.72	693.20	761.10	1107.80
TF, $k = 0$	452.80	489.80	571.64	514.15	579.05	1390.20
TF, $k = 1$	2681.40	2815.95	3495.12	2968.30	3134.55	50508.80
TF, $k = 2$	4436.40	4710.05	4921.01	4884.90	5008.40	5759.70

Table B.1: Summary of timing results for 100 implementations of proximal solver for $n = 500$.

`FORTRAN`, whereas others use `C++`; additionally, speed is impacted by other functions which might be written in `R`, for example, using `for` loops to construct the matrix of basis functions.

Appendix C. Numerical Experiments for Additional Tuning Parameters

In this appendix, we empirically investigate the impact of decoulping the tuning parameters for sparsity and smoothness. Recall our decoupled `GSAM`:

$$\widehat{\beta}, \widehat{f}_1, \ldots, \widehat{f}_p \leftarrow \underset{\beta \in \mathbb{R}, f_1, \ldots, f_p \in \mathcal{F}}{\operatorname{argmin}} - \mathbb{P}_n \ell \left(\beta + \sum_{j=1}^{p} f_j \right) + (1-\zeta)\lambda \sum_{j=1}^{p} P_{st}\left(f_j\right) + \zeta\lambda \sum_{j=1}^{p} \|f_j\|_n , \tag{C.1}$$

for a second tuning parameter $\zeta \in [0, 1]$. To demonstrate the performance of decoupling tuning parameters, we implemented (C.1) using some of the data from our simulation study in Section 5. In greater detail, we consider Scenario 3 (all smooth functions) and Scenario 4 (mixture of functions) from the simulation study in Section 5. We generated data with $n \in \{100, 300, 500, 700, 900, 1000\}$ and $p \in \{6, 100\}$. We implement both versions of `GSAM`, with the smoothness penalties, SSP and TF with $k = 0, 1, 2$. For the originally proposed `GSAM` (with coupled tuning parameters), we use a sequence of fifty λ values and, for (C.1), we additionally used a sequence of ten $\zeta \in [10^{-3}, 1 - 10^{-5}]$ values (resulting in a 10×50 grid of tuning parameters). We report the *oracle* mean square error (MSE):

$$\text{MSE}_{\text{coupled}} = \min_{\lambda} \|\widehat{f}_\lambda - f^0\|_n^2,$$
$$\text{MSE}_{\text{decoupled}} = \min_{\zeta,\lambda} \|\widehat{f}_{\zeta,\lambda} - f^0\|_n^2.$$

All results were averaged over 100 replications of the data.

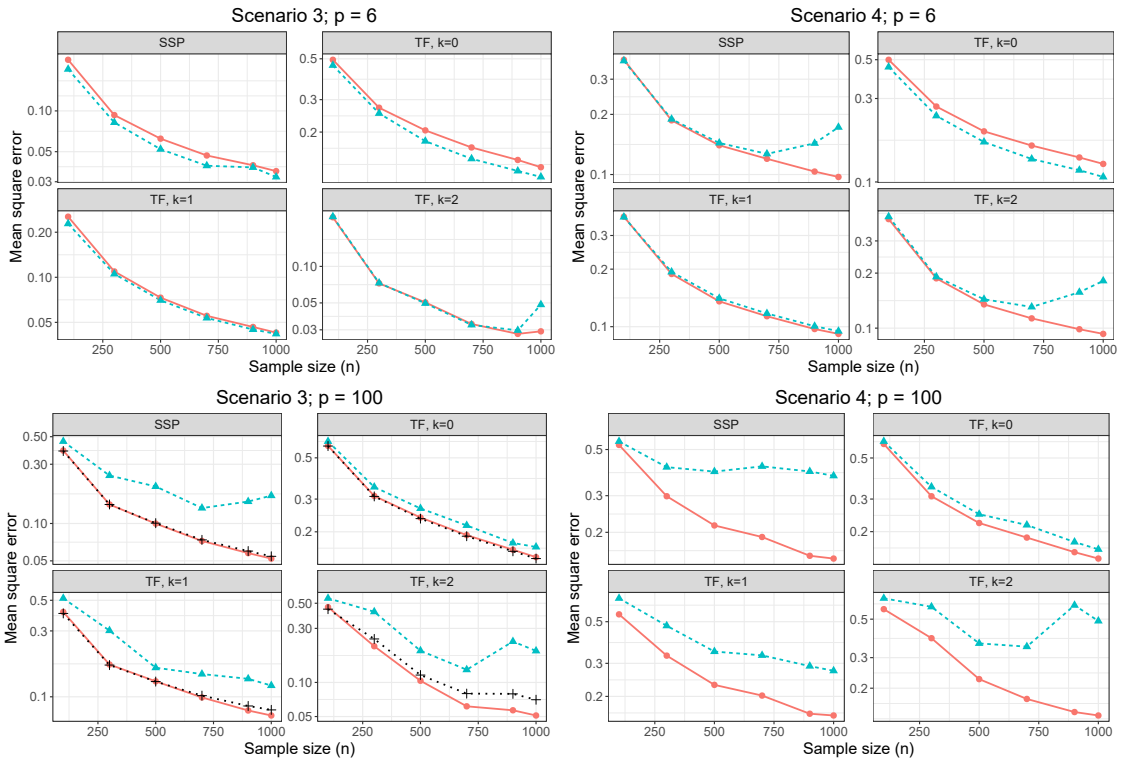

Figure C.1: Plot of oracle MSEs versus sample size for each of scenarios 3 and 4 for $p = 6$ and 100, averaged over 100 replications. The lines correspond to `GSAM` with coupled (—●—) and decoupled (- -▲- -) tuning parameters. For Scenario 3 with $p = 100$, we also consider a finer grid of ζ values for decoupled tuning parameters (···+···).

In Figure C.1 we plot the oracle MSE as a function of the sample size. For low-dimensional data we observe that decoupling the tuning parameters can lead to a lower MSE in some cases. For example, in Scenario 3, with $p = 6$ we observe the decoupled `GSAM` to have lower MSE for all methods for almost all sample sizes. However, even with $p = 6$ we note decoupled tuning parameters lead to a high MSE for large n whereas coupled tuning parameters have a monotone decreasing MSE curve. This phenomenon is exacerbated for $p = 100$, where we observe the coupled tuning parameters `GSAM` to uniformly beat the decoupled version in terms of oracle MSE. This behavior is likely due to the precision of our tuning parameters' grid: using a finer grid of ζ values could lead to a lower MSE but at a high computational cost. We study the use of a finer grid for Scenario 3 with $p = 100$: the third line in the panel is $\text{MSE}_{\text{decoupled}}$ for a sequence of thirty $\zeta \in [0.700, 0.999]$ values. We clearly observe a substantial reduction of oracle MSE from the original decoupled `GSAM`; additionally, for most methods, for some n, the decoupled `GSAM` outperforms the coupled `GSAM`.

Appendix D. Proof of Results in Section 2.2

Proof [Proof of Lemma 1] For brevity, we write our optimization problem as

$$\widehat{f}_1, \ldots, \widehat{f}_p \leftarrow \underset{f_1,\ldots,f_p \in \mathcal{F}}{\operatorname{argmin}} \mathcal{L}(f_1, \ldots, f_p),$$

where $\mathcal{L}(\cdot)$ is the objective function. The proof follows by contradiction: we assume some $\widehat{f}_j \equiv 0$ while others are non-zero, and looking at the sub-gradient conditions we arrive at a contradiction.

In greater detail, assume without loss of generality, that for some $k < p$, $\widehat{f}_1, \ldots, \widehat{f}_k \neq 0$ and $\widehat{f}_{k+1}, \ldots, \widehat{f}_p = 0$. Define the paths $\widehat{f}_{j,\varepsilon_j} = \widehat{f}_j + \varepsilon_j h_j$ for any direction $h_j \in \mathcal{F}$ for $j = 1, \ldots, p$. The sub-gradient conditions state, that for any direction $h_j \in \mathcal{F}$, we must have

$$0 \in \partial_{\varepsilon_j} \mathcal{L}(\widehat{f}_{1,\varepsilon_1}, \ldots, \widehat{f}_{p,\varepsilon_p})\Big|_{\varepsilon_1=\ldots=\varepsilon_p=0},$$

for all $j \in \{1, \ldots, p\}$, where ∂_{ε_j} denotes the sub-gradient set with respect to ε_j. For the non-zero functions, $j \in \{1, \ldots, k\}$, the sub-gradient conditions are:

$$-\langle y - \widehat{f}_1 - \widehat{f}_2 - \ldots - \widehat{f}_k, h_j\rangle_n + \lambda_1 \partial_{\varepsilon_j} P_{st}^2(\widehat{f}_{j,\varepsilon_j}) + \lambda_2 \frac{\langle \widehat{f}_j, h_j\rangle_n}{\|\widehat{f}_j\|_n} \ni 0,$$

where $\langle a, b\rangle_n = n^{-1}\sum_{i=1}^n a(\boldsymbol{x}_i)b(\boldsymbol{x}_i)$. Since the sub-gradient conditions must be met *for all* $h_j \in \mathcal{F}$, we set $h_j = \widehat{f}_j$. This implies that

$$\begin{aligned}
\partial_{\varepsilon_j} P_{st}^2(\widehat{f}_{j,\varepsilon_j})\Big|_{\varepsilon_j=0} &= \partial_{\varepsilon_j} P_{st}^2\{(1+\varepsilon_j)\widehat{f}_j\}\Big|_{\varepsilon_j=0} \\
&= \partial_{\varepsilon_j}(1+\varepsilon_j)^2\big|_{\varepsilon_j=0} P_{st}^2\{\widehat{f}_j\} \\
&= 2P_{st}^2(\widehat{f}_j),
\end{aligned}$$

where the second equality follows from properties of a norm. Thus, for $j \in \{1, \ldots, k\}$, we must have

$$\begin{aligned}
&-\langle y - \widehat{f}_1 - \widehat{f}_2 - \ldots - \widehat{f}_k, \widehat{f}_j\rangle_n + 2\lambda_1 P^2(\widehat{f}_j) + \lambda_2 \|\widehat{f}_j\|_n = 0, \\
\Leftrightarrow &\frac{\langle y - \widehat{f}_1 - \widehat{f}_2 - \ldots - \widehat{f}_k, \widehat{f}_j\rangle_n}{\lambda_2 \|\widehat{f}_j\|_n} = 1 + \frac{2\lambda_1 P^2(\widehat{f}_j)}{\lambda_2 \|\widehat{f}_j\|_n}.
\end{aligned} \tag{D.1}$$

On the other hand for $j' \in \{k+1, \ldots, p\}$,

$$\begin{aligned}
\partial_{\varepsilon_{j'}} P_{st}^2(\widehat{f}_{j',\varepsilon_{j'}})\Big|_{\varepsilon_{j'}=0} &= \partial_{\varepsilon_{j'}} P_{st}^2(\varepsilon_{j'} h_{j'})\Big|_{\varepsilon_{j'}=0} \\
&= \partial_{\varepsilon_{j'}} \varepsilon_{j'}^2\Big|_{\varepsilon_{j'}=0} P_{st}^2(h_{j'}) = 0. \\
\partial_{\varepsilon_{j'}} \|\widehat{f}_{j',\varepsilon_{j'}}\|_n\Big|_{\varepsilon_{j'}=0} &= \partial_{\varepsilon_{j'}} \|\varepsilon_{j'} h_{j'}\|_n\Big|_{\varepsilon_{j'}=0} \\
&= \partial_{\varepsilon_{j'}} |\varepsilon_{j'}|\Big|_{\varepsilon_{j'}=0} \|h_{j'}\|_n,
\end{aligned}$$

By definition of a sub-gradient we know that $\partial_{\varepsilon_{j'}}|\varepsilon_{j'}|\Big|_{\varepsilon_{j'}=0} = [-1,1]$. Therefore, the sub-gradient conditions for $j' \in \{k+1,\ldots,p\}$ imply that

$$-\langle y - \widehat{f}_1 - \widehat{f}_2 - \ldots - \widehat{f}_k, h_{j'}\rangle_n + \lambda_2 U_{j'}\|h_{j'}\|_n = 0, \tag{D.2}$$

for some $U_{j'} \in [-1,1]$. Now with (D.1) and (D.2), and a clever choice of $h_{j'}$ we arrive at a contradiction: setting $h_{j'} = h_j = \widehat{f}_j \neq 0$ for any $j \leq k$,

$$\begin{aligned} U_{j'} &= \frac{\langle y - \widehat{f}_1 - \widehat{f}_2 - \ldots - \widehat{f}_k, \widehat{f}_j\rangle_n}{\lambda_2\|\widehat{f}_j\|_n} && \text{by (D.2)} \\ &= 1 + \frac{2\lambda_1 P^2(\widehat{f}_j)}{\lambda_2\|\widehat{f}_j\|_n} && \text{by (D.1)} \\ &> 1, \end{aligned}$$

because $\widehat{f}_j \neq 0$, which leads to a contradiction.

■

Proof [Proof of Corollary 2] The proof is essentially identical to that of Lemma 1. Assume without loss of generality that $I = \{k+1,\ldots,p\}$. Assume for contradiction that, there is some $j \in \{1,\ldots,k\}$ such that $\widehat{f}_j \in \mathcal{F}\backslash\mathcal{F}_0$. Then as in the proof of Lemma 1, we can arrive at a contradiction showing $U_{j'} > 1$ for all $j' \in I$. ■

Appendix E. Proof of Results in Section 3

Proof [Proof of Lemma 3] If $\widetilde{f} = 0$, then $\widehat{f} = 0$ is trivially the solution to (8). Thus, throughout this proof, we consider $\widetilde{f} \neq 0$.
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

$$\frac{1}{2}\left\|r - (\widehat{f} + \varepsilon h)\right\|_n^2 + \lambda_1 P_{st}\left(\widehat{f} + \varepsilon h\right) + \lambda_2\|\widehat{f} + \varepsilon h\|_n \tag{E.1}$$

is minimized at $\varepsilon = 0$. We begin by noting that

$$\frac{1}{2}\left\|r - (\widetilde{f} + \varepsilon ch)\right\|_n^2 + \lambda_1 P_{st}\left(\widetilde{f} + \varepsilon ch\right),$$

is minimized at $\varepsilon = 0$ because

$$\widetilde{f} + \varepsilon ch = \widetilde{f} + \varepsilon c f_T - \varepsilon c\widehat{f} = (1-\varepsilon)\widetilde{f} + \varepsilon c f_T \in \mathcal{F},$$

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 $\vartheta_1 \in \partial\, P_{st}\left(\widetilde{f} + \varepsilon ch\right)\Big|_{\varepsilon=0}$, 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 (E.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 (E.1) is minimized at $\varepsilon = 0$. Since f_T was an arbitrary function, we conclude that $\widehat{f}$ is the solution of (8).

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

$$\frac{1}{2}\|r - \varepsilon f_T\|_n^2 + \lambda_1 P_{st}\left(\varepsilon f_T\right) + \lambda_2 \|\varepsilon f_T\|_n, \tag{E.2}$$

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 (E.2). All sub-gradients of (E.2) at $\varepsilon = 0$ are given by

$$-\langle r, f_T\rangle_n + \nu_1 \lambda_1 P_{st}(f_T) + \nu_2 \lambda_2 \|f_T\|_n, \tag{E.3}$$

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 (E.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 (E.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 $\|\widetilde{f}\|_n < \lambda_2$.

Sufficient conditions for $\widehat{f}_j \equiv 0$: for the second part of this Lemma, we proceed from the sub-gradient condition (E.3). If we set $\nu_1 = 0$, then $\widehat{f} \equiv 0$ if for every direction f_T there exists $\nu_2 \in [-1, 1]$ such that

$$\nu_2 \lambda_2 \|f_T\|_n = \langle r, f_T\rangle_n.$$

Which is equivalent to

$$\left|\frac{\langle r, f_T\rangle_n}{\|f_T\|_n}\right| \leq \lambda_2. \tag{E.4}$$

If $\lambda_2 \geq \|r\|_n$, then (E.4) is satisfied for all f_T because by the Cauchy-Schwarz inequality

$$\left|\frac{\langle r, f_T\rangle_n}{\|f_T\|_n}\right| \leq \left|\frac{\|r\|_n \|f_T\|_n}{\|f_T\|_n}\right| = \|r\|_n \leq \lambda_2.$$

■

Proof [Prof of Lemma 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 $f_\varepsilon = \widehat{f}_\lambda + \varepsilon h$,

$$\frac{\partial}{\partial \epsilon}\left[\frac{1}{2}\left\|r - f_\varepsilon\right\|_n^2 + \lambda P_{st}\left(f_\varepsilon\right)\right]_{\epsilon=0} = -\langle r - \widehat{f}_\lambda, h\rangle_n + \lambda \frac{\nabla_h P_{st}^\nu(\widehat{f}_\lambda)}{\nu P_{st}^{\nu-1}(\widehat{f}_\lambda)} = 0, \tag{E.5}$$

Similarly, for the path $f_{\widetilde{\varepsilon}} = \widetilde{f}_{\widetilde{\lambda}} + \widetilde{\varepsilon} h$, by the gradient condition,

$$\frac{\partial}{\partial \widetilde{\varepsilon}}\left[\frac{1}{2}\left\|r - f_{\widetilde{\varepsilon}}\right\|_n^2 + \widetilde{\lambda} P_{st}^\nu\left(f_{\widetilde{\varepsilon}}\right)\right]_{\widetilde{\varepsilon}=0} = -\langle r - \widetilde{f}_{\widetilde{\lambda}}, h\rangle_n + \widetilde{\lambda} \nabla_h P_{st}^\nu\left(\widetilde{f}_{\widetilde{\lambda}}\right) = 0.$$

This is exactly the optimality condition (E.5) with $\widetilde{\lambda} = \lambda(\nu P_{st}^{\nu-1}(\widehat{f}_\lambda))^{-1}$. Thus, if $\nu\widetilde{\lambda} P_{st}^{\nu-1}(\widetilde{f}_{\widetilde{\lambda}}) = \lambda$ 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

$$\frac{1}{2}\|f_{interp} - f\|_n^2 + \lambda P_{st}(f)$$

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

$$\frac{1}{2}\|f_{interp} - f_{h,\varepsilon}\|_n^2 + \lambda P_{st}(f_{h,\varepsilon}) \tag{E.6}$$

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 (E.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 (E.6) we need to have

$$\lambda\kappa P_{st}(h) = \langle f_{interp} - f_{null}, h - f_{null}\rangle_n.$$

Consequently, (E.6) is minimized at $\varepsilon = 0$ if and only if for all $h \in \mathcal{F}$

$$\langle f_{interp} - f_{null}, h - f_{null}\rangle_n \leq \lambda P_{st}(h).$$

Using the decomposition $h = h_0 + h_\perp \in \mathcal{F}_1 \oplus \mathcal{F}_2$, the above condition becomes

$$\frac{\langle f_{interp} - f_{null}, h_0 - f_{null}\rangle_n}{P_{st}(h_\perp)} + \frac{\langle f_{interp} - f_{null}, h_\perp\rangle_n}{P_{st}(h_\perp)} \leq \lambda.$$

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

$$\frac{1}{2}\|f_{interp} - (1-\varepsilon)f_{null} - \varepsilon h_{null}\|_n^2,$$

is minimized by $\varepsilon = 0$. Equivalently (by the gradient condition), for all $h_{null} \in \mathcal{F}_1$,

$$\langle f_{interp} - f_{null}, f_{null} - h_{null}\rangle_n = 0.$$

■

Proof [Proof of Convergence of Infinite-dimensional problem] We begin by re-writing our optimization problem as follows:

$$\widehat{\beta}, \widehat{\eta}_1, \ldots, \widehat{\eta}_p \leftarrow \underset{\beta\in\mathbb{R}, \eta_1,\ldots,\eta_p \in \mathbb{R}^n}{\operatorname{argmin}} -\mathbb{P}_n \ell\left(\beta + \sum_{j=1}^p \eta_j\right) + \lambda^2 \sum_{j=1}^p P_{st}\left(\eta_j\right) + \lambda \sum_{j=1}^p \|\eta_j\|_n, \tag{E.7}$$
$$P_{st}(\eta_j) = \min_{f\in\mathcal{F}} P_{st}(f) + \mathbb{I}(\eta_{ji} = f(x_{ji}) \text{ for all } i),$$

in which case, the term $P_{st}(\eta_j)$, in (E.7), is a semi-norm over $\mathbb{R}^n$. We recover our estimate by

$$\widehat{f}_j \leftarrow \underset{f\in\mathcal{F}}{\operatorname{argmin}}\ P_{st}(f) + \mathbb{I}\Big(\widehat{\eta}_{ji} = f(x_{ji}) \text{ for all } i\Big). \tag{E.8}$$

Thus a proximal gradient descent algorithm will solve the finite-dimensional problem (E.7) by standard convergence guarantees. We then just need to solve (E.8) for a given $\widehat{\eta}_j$. Our Algorithm 1, does exactly that. To see this, re-write our main proximal problem, (11):

$$\eta_j^{inter} \leftarrow \underset{\eta\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\left\|\left(\eta_j^{k-1} - t\boldsymbol{r}\right) - \eta\right\|_n^2 + t\lambda^2 P_{st}\left(\eta\right), \tag{E.9}$$
$$f_j^{inter} \leftarrow \underset{f\in\mathcal{F}}{\operatorname{argmin}}\ P_{st}(f) + \mathbb{I}(\eta_{ji}^{inter} = f(x_{ji}) \text{ for all } i). \tag{E.10}$$

Thus, we see that (E.9) generates a proximal gradient descent algorithm which solves (E.7) and (E.10) is exactly the problem (E.8). This completes the proof.
■

Appendix F. Proofs of Results in Section 4.2

In this section we present the proof Theorem 7 and 15 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 use the so-called *basic inequalities*. Hence, for the sake of convenience, we state and prove these basic inequalities as a separate lemma.

Lemma F.1 (Basic Inequality) *Let $\widehat{f}(\boldsymbol{x}) = \widehat{\beta} + \sum_{j=1}^{p} \widehat{f}_j(x_j)$ be as defined in (14), 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*

$$\mathcal{E}(\widehat{f}) + \lambda I(\widehat{f}) \leq -\left[\nu_n(\widehat{f}) - \nu_n(f^*)\right] + \lambda I(f^*) + \mathcal{E}(f^*).$$

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

$$\mathcal{E}(\widetilde{f}) + \lambda I(\widetilde{f}) \leq -\left[\nu_n(\widetilde{f}) - \nu_n(f^*)\right] + \lambda I(f^*) + \mathcal{E}(f^*).$$

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 [Proof of Theorem 7] Define

$$t = \frac{M^*}{M^* + I(\widehat{f} - f^*)}, \tag{F.1}$$

and $\widetilde{f} = t\widehat{f} + (1-t)f^*$. Then (F.1) implies $I(\widetilde{f} - f^*) \leq M^*$ and by Lemma F.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

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

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 [Proof of Theorem 15] As in the proof of Theorem 7, 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 $|\widetilde{\beta} - \beta^*| + I(\widetilde{f} - f^*) \leq M^*$ implying that, $\widetilde{f} \in \mathcal{F}^0_{local}$. Lemma F.1 implies that on the set $\mathcal{T}$ (where $Z_{M^*} \leq \rho M^*$) we have

$$\mathcal{E}(\widetilde{f}) + \lambda I(\widetilde{f}) \leq Z_{M^*} + \mathcal{E}^* + \lambda I(f^*) \leq \rho M^* + \mathcal{E}(f^*) + \lambda I(f^*). \tag{F.2}$$

Be definition of S_* and $I(\cdot)$, we have by the triangle inequality

$$\lambda I(f^*) = \lambda \sum_{j \in S_*} \Big\{ \|f_j^*\|_n + \lambda P_{st}(f_j^*) \Big\} \leq \lambda \sum_{j \in S_*} \Big\{ \|f_j^* - \widetilde{f}_j\|_n + \|\widetilde{f}_j\|_n + \lambda P_{st}(f_j^*) \Big\},$$

and by the reverse triangle inequality we have

$$\begin{aligned}
\lambda I(\widetilde{f}) &= \lambda \sum_{j \in S_*} \Big\{ \|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j) \Big\} + \lambda \sum_{j \in S_*^c} \Big\{ \|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j) \Big\} \\
&\geq \lambda \sum_{j \in S_*} \Big\{ \|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j - f_j^*) - \lambda P_{st}(f_j^*) \Big\} + \lambda \sum_{j \in S_*^c} \Big\{ \|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j) \Big\} \\
&= \lambda \sum_{j \in S_*} \Big\{ \|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j - f_j^*) - \lambda P_{st}(f_j^*) \Big\} + \lambda \sum_{j \in S_*^c} \Big\{ \|\widetilde{f}_j - f_j^*\|_n + \lambda P_{st}(\widetilde{f}_j - f_j^*) \Big\},
\end{aligned}$$

where the last equality follows from the fact that $f_j^* = 0$ for all $j \in S_*^c$. With the above two inequalities combined with (F.2) we get

$$\begin{aligned}\mathcal{E}(\widetilde{f})+\lambda \sum_{j\in S_*} \Big\{\|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j - f_j^*) - \lambda P_{st}(f_j^*)\Big\} + \lambda \sum_{j\in S_*^c} \Big\{\|\widetilde{f}_j\|_n + \lambda P_{st}(\widetilde{f}_j)\Big\} \\ \leq \lambda \sum_{j\in S_*} \Big\{\|f_j^* - \widetilde{f}_j\|_n + \|\widetilde{f}_j\|_n + \lambda P_{st}(f_j^*)\Big\} + \rho M^* + \mathcal{E}(f^*),\end{aligned}$$

which simplifies to

$$\begin{aligned}\mathcal{E}(\widetilde{f}) + \lambda \sum_{j\in S_*^c} \|\widetilde{f}_j - f_j^*\|_n + \lambda \sum_{j=1}^{p} \lambda P_{st}(\widetilde{f}_j - f_j^*) \leq \\ \lambda \sum_{j\in S_*} \|f_j^* - \widetilde{f}_j\|_n + 2\lambda^2 \sum_{j\in S_*} P_{st}(f_j^*) + \rho M^* + \mathcal{E}(f^*).\end{aligned} \tag{F.3}$$

Now we add $\lambda|\widetilde{\beta} - \beta^*| + \lambda \sum_{j\in S_*} \|\widetilde{f}_j - f_j^*\|_n$ to both sides of (F.3) to obtain

$$\begin{aligned}\mathcal{E}(\widetilde{f}) + \lambda\Big\{|\widetilde{\beta} - \beta^*| + I(\widetilde{f} - f^*)\Big\} \leq 2\lambda \sum_{j\in S_*} \|f_j^* - \widetilde{f}_j\|_n + \lambda|\widetilde{\beta} - \beta^*| + \\ \rho M^* + \mathcal{E}(f^*) + 2\lambda^2 \sum_{j\in S_*} P_{st}(f_j^*).\end{aligned} \tag{F.4}$$

Case I. If

$$2\lambda \sum_{j\in S_*} \|f_j^* - \widetilde{f}_j\|_n + \lambda|\widetilde{\beta} - \beta^*| \leq \rho M^* + \mathcal{E}(f^*) + 2\lambda^2 \sum_{j\in S_*} P_{st}(f_j^*),$$

then (F.4) simplifies to

$$\begin{aligned}\mathcal{E}(\widetilde{f}) + \lambda\Big\{|\widetilde{\beta} - \beta^*| + I(\widetilde{f} - f^*)\Big\} &\leq 2\rho M^* + 2\mathcal{E}(f^*) + 4\lambda^2 \sum_{j\in S_*} P_{st}(f_j^*) \\ &\leq 4\rho M^* \leq 4\frac{\lambda}{8} M^* = \lambda M^*/2,\end{aligned}$$

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

$$2\lambda \sum_{j\in S_*} \|f_j^* - \widetilde{f}_j\|_n + \lambda|\widetilde{\beta} - \beta^*| \geq \rho M^* + \mathcal{E}(f^*) + 2\lambda^2 \sum_{j\in S_*} P_{st}(f_j^*),$$

then we have

$$\mathcal{E}(\widetilde{f}) + \lambda\Big\{|\widetilde{\beta} - \beta^*| + I(\widetilde{f} - f^*)\Big\} \leq 4\lambda \sum_{j\in S_*} \|f_j^* - \widetilde{f}_j\|_n + 2\lambda|\widetilde{\beta} - \beta^*|. \tag{F.5}$$

This is equivalent to

$$\mathcal{E}(\widetilde{f}) + \lambda\Big\{ \sum_{j\in S_*^c} \|\widetilde{f}_j - f_j^*\|_n + \lambda \sum_{j=1}^{p} P_{st}(\widetilde{f}_j - f_j^*)\Big\} \leq 3\lambda \sum_{j\in S_*} \|f_j^* - \widetilde{f}_j\|_n + \lambda|\widetilde{\beta} - \beta^*|,$$

which means we have that

$$\sum_{j\in S_*^c} \|\widetilde{f}_j - f_j^*\|_n + \lambda \sum_{j=1}^{p} P_{st}(\widetilde{f}_j - f_j^*) \leq 3 \sum_{j\in S_*} \|f_j^* - \widetilde{f}_j\|_n + |\widetilde{\beta} - \beta^*|,$$

and hence by the compatibility condition (F.5) reduces to

$$\mathcal{E}(\widetilde{f}) + \lambda\Big\{|\widetilde{\beta} - \beta^*| + I(\widetilde{f} - f^*)\Big\} \leq 4\lambda\|\widetilde{f} - f^*\|\sqrt{s_*}/\phi(S_*).$$

Since $\widetilde{f}$ and f^* are in $\mathcal{F}^0_{local}$, we invoke the inequality $uv \leq H(v) + G(u)$ to obtain

$$\begin{aligned}\frac{4\lambda\sqrt{s_*}\|\widetilde{f} - f^*\|}{\phi(S_*)} &\leq \frac{8\lambda\sqrt{s_*}}{\phi(S_*)}\left[\frac{\|\widetilde{f} - f^0\|}{2} + \frac{\|f^* - f^0\|}{2}\right] \\ &\leq H\left(\frac{8\lambda\sqrt{s_*}}{\phi(S_*)}\right) + G\left(\frac{\|\widetilde{f} - f^0\|}{2} + \frac{\|f^* - f^0\|}{2}\right).\end{aligned}$$

By the convexity of G and the margin condition we obtain

$$\frac{4\lambda\sqrt{s_*}\|\widetilde{f} - f^*\|}{\phi(S_*)} \leq H\left(\frac{8\lambda\sqrt{s_*}}{\phi(S_*)}\right) + \frac{\mathcal{E}(\widetilde{f})}{2} + \frac{\mathcal{E}(f^*)}{2}.$$

Hence we have

$$\frac{\mathcal{E}(\widetilde{f})}{2} + \lambda\Big\{|\widetilde{\beta} - \beta^*| + I(\widetilde{f} - f^*)\Big\} \leq H\left(\frac{8\lambda\sqrt{s_*}}{\phi(S_*)}\right) + \frac{\mathcal{E}(f^*)}{2} \leq \rho M^* \leq \lambda M^*/8,$$

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

$$\mathcal{E}(\widehat{f}) + \lambda\Big\{|\widehat{\beta} - \beta^*| + I(\widehat{f} - f^*)\Big\} \leq 4\rho M^*. \tag{F.6}$$

■

Appendix G. The Set $\mathcal{T}$

Theorems 7 and 12 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 4. We consider the special case of loss functions linear in Y_i as in Corollaries 8 and 13 and bound the term $\nu_n(f) - \nu_n(f^*)$ in the following theorem.

Theorem G.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*

$$-\ell(f) = -\ell(f, \boldsymbol{x}_i, Y_i) = aY_i f(\boldsymbol{x}_i) + b(f(\boldsymbol{x}_i)),$$

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:

$$\max_{i=1,\ldots,n} K^2 \left(\mathbb{E} e^{(Y_i - \mu_i)^2 / K^2} - 1 \right) \leq \sigma_0^2, \tag{G.1}$$

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

$$\nu_n(f) - \nu_n(f^*) \leq \rho \left[|\beta - \beta^*| + \sum_{j=1}^{p} \|f_j - f_j^*\|_n + \lambda P_{st}(f_j - f_j^*) \right], \tag{G.2}$$

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* (G.2) *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* (G.2) *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 7 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 15 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 G.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 I.

Proof [Proof of Theorem G.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 I.1. 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

$$\left\{\frac{f_j - f_j^*}{\|f_j - f_j^*\|_n + \lambda P_{st}(f_j - f_j^*)} : f_j \in \mathcal{F}\right\},$$

for some $f_j^* \in \mathcal{F}$ for all $j = 1, \ldots, p$. Now we apply Lemma I.2 to the above class by first noting that $R \leq 1$ and then using the bound for Dudley's integral

$$A_0^{1/2} T_n^{1/2} \int_0^1 \log^{1/2}\left(\frac{1}{u} + 1\right) du \leq \widetilde{A}_0 T_n^{1/2},$$

we have for all δ that satisfy

$$\delta \geq 2C\widetilde{A}_0\sqrt{\frac{T_n}{n}},$$

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

$$\begin{aligned}
&\mathbb{P}\left(\max_{j=1,\dots,p}\sup_{f_j\in\mathcal{F}}\frac{\left|\langle \boldsymbol{Y}-\boldsymbol{\mu}, f_j - f_j^*\rangle_n\right|}{\|f_j-f_j^*\|_n + \lambda P_{st}(f_j-f_j^*)} \geq \rho\right)\\
&\leq pC\exp\left[-\frac{n\rho^2}{4C^2}\right] = C\exp\left[-\frac{n\rho^2}{4C^2}+\log p\right]\\
&= C\exp\left[-n\rho^2\left(\frac{1}{4C^2}-\frac{\log p}{n\rho^2}\right)\right] \leq C\exp\left[-n\rho^2 C_2\right],
\end{aligned}$$

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

$$\widetilde{\mathcal{F}} = \left\{\frac{f_j - f_j^*}{\|f_j-f_j^*\|_n + \lambda P_{st}(f_j-f_j^*)} : f_j \in \mathcal{F}\right\},$$

and we can now apply Lemma I.2 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

$$\mathbb{P}\left(\sup_{f_j\in\mathcal{F}}\frac{\left|\langle \boldsymbol{Y}-\boldsymbol{\mu}, f_j - f_j^*\rangle_n\right|}{\|f_j-f_j^*\|_n + \lambda P_{st}(f_j-f_j^*)} \geq \delta\right) \leq C\exp\left[-\frac{n\delta^2}{4C^2}\right]. \tag{G.3}$$

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 (G.3) along

with a union bound to obtain

$$\begin{aligned}\mathbb{P}\left(\max_{j=1,\ldots,p}\sup_{f_j\in\mathcal{F}}\frac{\left|\langle \boldsymbol{Y}-\boldsymbol{\mu}, f_j-f_j^*\rangle_n\right|}{\|f_j-f_j^*\|_n+\lambda P_{st}(f_j-f_j^*)}\geq\rho\right) &\leq pC\exp\left[-\frac{n\rho^2}{4C^2}\right]\\ &= C\exp\left[-n\rho^2\left(\frac{1}{4C^2}-\frac{\log p}{n\rho^2}\right)\right]\\ &\leq C\exp\left[-n\rho^2 C_2\right],\end{aligned}$$

for some positive constant $C_2 = C_2(C, \widetilde{A}_0)$ exactly as in Case 1.

■

Appendix H. On the Margin and Compatibility Coniditions

In this appendix, we present a detailed discussion of the margin and compatibility conditions. These conditions are the main assumptions made for our theoretical results and in this appendix we discuss suitable conditions under which the conditions hold.

H.1 Margin Condition

We now present a general framework for proving the quadratic margin condition for loss functions given by the negative log-likelihood of exponential family distributions. Specifically, for loss functions of the type

$$-\ell(f) = -\ell(y_i; f(\boldsymbol{x}_i)) = ay_i f(\boldsymbol{x}_i) + b(f(\boldsymbol{x}_i)). \tag{H.1}$$

The excess risk is

$$\begin{aligned}\mathcal{E}(f) &= \mathbb{P}\{\ell(f^0)-\ell(f)\}\\ &= \frac{1}{n}\sum_{i=1}^n\{-a\mathbb{E}(y_i)\}\{f^0(\boldsymbol{x}_i)-f(\boldsymbol{x}_i)\}-[b\{f^0(\boldsymbol{x}_i)\}-b\{f(\boldsymbol{x}_i)\}]\\ &= \frac{1}{n}\sum_{i=1}^n[b'\{f^0(\boldsymbol{x}_i)\}]\{f^0(\boldsymbol{x}_i)-f(\boldsymbol{x}_i)\}-[b\{f^0(\boldsymbol{x}_i)\}-b\{f(\boldsymbol{x}_i)\}],\end{aligned}$$

where the last equality holds because the expectation of the score function is zero. Now by Taylor's theorem for $b(\cdot)$ and the mean value theorem for $b'(\cdot)$, the above simplifies to

$$\begin{aligned}\mathcal{E}(f) &= n^{-1}\sum_{i=1}^n b''(\zeta_{1i})\{f^0(\boldsymbol{x}_i)-f_i(\boldsymbol{x}_i)\}^2-\frac{b''(\zeta_{2i})}{2}\{f^0(\boldsymbol{x}_i)-f(\boldsymbol{x}_i)\}^2\\ &\geq \underbrace{\min_i\left\{b''(\zeta_{1i})-\frac{b''(\zeta_{2i})}{2}\right\}}_{C}\|f^0-f\|_n^2,\end{aligned}$$

where ζ_{1i}, ζ_{2i} lie between $f(\boldsymbol{x}_i), f^0(\boldsymbol{x}_i)$. Thus for loss functions of the type (H.1), we only need to find a constant C in a neighborhood of f^0. As an example, consider the least squares loss where

$$b(\theta) = \theta^2/2, \quad b''(\theta) = 1 \Rightarrow C = 1/2.$$

Similarly, for logistic regression we have

$$b(\theta) = \log\{\exp(\theta) + 1\}, \quad b''(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)}.$$

Now on the set $\mathcal{F}^0_{local} = \{f : \|f - f^0\|_\infty \leq \eta\}$, we need only look at the interval $(f^0(\boldsymbol{x}_i) - \eta, f^0(\boldsymbol{x}_i) + \eta)$. As a simple case, say $f^0(\boldsymbol{x}_i) = 0$, then

$$b''(\zeta_{1i}) - \frac{b''(\zeta_{2i})}{2} \geq \frac{\exp\{\eta\}}{[1 + \exp\{\eta\}]^2} - \frac{\exp\{0\}}{2[1 + \exp\{0\}]^2} > 0,$$

for $\eta < \log(3 + 2\sqrt{2})$. While tedious, for any $f^0(\boldsymbol{x}_i)$, we can find an η such that $b''(\zeta_{1i}) - b''(\zeta_{2i})/2 > 0$ for all ζ_{1i}, ζ_{2i}. The above examples indicate that the margin condition is satisfied in most of the relevant problems.

H.2 Compatibility Condition

We now show that under suitable conditions, the theoretical compatibility condition implies the compatibility condition. Recall first our notation, $\|f\|^2 = \int [f(\boldsymbol{x})]^2 dQ(\boldsymbol{x})$ where Q is the probability measure of our observed data. We denote by Q_n, the empirical measure associated with Q, and $\|f\|_n^2 = n^{-1} \sum_{i=1}^n f^2(\boldsymbol{x}_i)$.

Lemma H.1 *On the set*

$$\mathcal{S} = \left\{ \sup_{f \in \mathcal{F}} \frac{\Big| \|f\|_n - \|f\| \Big|}{\sum_{j=1}^p \|f_j\| + \lambda P(f_j)} \leq \eta \right\}, \tag{H.2}$$

with $\eta \in (0, 1/5)$, suppose we have

$$c_1 = \eta \left\{ \frac{2(2+\eta)}{1-\eta} + \frac{3(1+\eta)}{1-5\eta} \right\} \frac{\sqrt{|S|}}{\widetilde{\phi}(S)} < 1, \tag{H.3}$$

if the theoretical compatibility condition holds with η as in (H.2)*. Then, the empirical compatibility condition holds with constant*

$$\{\phi(S)\}^{-1} = \left\{ (1+\eta) + \frac{3\eta(1+\eta)}{1-5\eta} \right\} \frac{1}{(1-c_1)\widetilde{\phi}(S)}.$$

Remark H.2 *The event $\mathcal{S}$, ensures that our empirical norm is relatively close to the theoretical norm; existing literature shows that $\mathcal{S}$ holds with high probability under suitable entropy conditions for small η (see for example, Section 5 of Tan and Zhang, 2019). The condition* (H.3)*, simply forces our index set S to not be too large; thus, highlighting the role of sparsity in high dimensions.*

To prove Lemma H.1, we will state and prove a sequence of smaller lemmas, the proof will then follow immediately.

For ease of exposition/reference, we restate our theoretical and (empirical) compatibility conditions. For simplicity, we do not include the intercept term; explicitly including an intercept term does not change the proof.

Definition H.3 (Empirical Compatibility Condition) *The compatibility condition is said to hold for an index set $S \subset \{1, 2, \ldots, p\}$ with $s = |S|$, and with compatibility constant $\phi(S) > 0$, if for all $\lambda > 0$ and all functions f of the form $f(\boldsymbol{x}) = \sum_{j=1}^{p} f_j(x_j)$ that satisfy*

$$\sum_{j \in S^c} \|f_j\|_n + \sum_{j=1}^{p} \lambda P_{st}(f_j) \leq 3 \sum_{j \in S} \|f_j\|_n, \tag{Empirical-A}$$

it holds that

$$\sum_{j \in S} \|f_j\|_n \leq \|f\|_n \sqrt{s}/\phi(S). \tag{Empirical-B}$$

Definition H.4 (Theoretical Compatibility Condition) *The theoretical compatibility condition is said to hold for an index set $S \subset \{1, 2, \ldots, p\}$ with $s = |S|$, and for a compatibility constant $\widetilde{\phi}(S) > 0$, if for some $\eta \in (0, 1/5)$, all $\lambda > 0$, and all functions of the form $f(\boldsymbol{x}) = \sum_{j=1}^{p} f_j(x_j)$ that satisfy*

$$\sum_{j \in S^c} \|f_j\| + \frac{1 - 5\eta}{1 - \eta} \sum_{j=1}^{p} \lambda P_{st}(f_j) \leq \frac{3(1 + \eta)}{1 - \eta} \sum_{j \in S} \|f_j\|, , \tag{Theoretical-A}$$

it holds that

$$\sum_{j \in S} \|f_j\| \leq \|f\| \sqrt{s}/\widetilde{\phi}(S). \tag{Theoretical-B}$$

Proof [Proof of Lemma H.1] We will show that on the set $\mathcal{S}$ given in (H.2), we have the following sequence of implications.

$$\text{(Empirical-A)} \overset{1}{\Rightarrow} \text{(Theoretical-A)} \overset{2}{\Rightarrow} \text{(Theoretical-B)} \overset{3}{\Rightarrow} \text{(Empirical-B)}$$

1. The result of Lemma H.5.

2. By assuming the theoretical compatility condition.

3. Follows immediately from Lemmas H.6 and H.7.

■

Lemma H.5 *On the set $\mathcal{S}$,* (Empirical-A) $\Rightarrow$ (Theoretical-A).

Proof We have

$$\left| \|f_j\|_n - \|f_j\| \right| \leq \eta \left(\|f_j\| + \lambda P(f_j) \right),$$

which is equivalent to

$$(1-\eta)\|f_j\| - \eta\lambda P(f_j) \leq \|f_j\|_n \leq (1+\eta)\|f_j\| + \eta\lambda P(f_j). \tag{H.4}$$

Thus we have the following:

$$\begin{aligned}
&\sum_{j\in S^c} \|f_j\| + \sum_{j=1}^{n} \lambda P(f_j) \\
&\leq \frac{1}{1-\eta} \sum_{j\in S^c} \|f_j\|_n + \frac{\eta}{1-\eta} \sum_{j\in S^c} \lambda P(f_j) + \sum_{j=1}^{p} \lambda P(f_j) && \text{by (H.4)} \\
&= \frac{1}{1-\eta} \sum_{j\in S^c} \|f_j\|_n + \frac{1}{1-\eta} \sum_{j\in S^c} \lambda P(f_j) + \sum_{j\in S} \lambda P(f_j) \\
&= \frac{1}{1-\eta} \left\{ \sum_{j\in S^c} \|f_j\|_n + \sum_{j=1}^{p} \lambda P(f_j) \right\} + \frac{\eta}{1-\eta} \sum_{j\in S} \lambda P(f_j) \\
&\leq \frac{3}{1-\eta} \sum_{j\in S} \|f_j\|_n + \frac{\eta}{1-\eta} \sum_{j\in S} \lambda P(f_j) && \text{by (Empirical-A)} \\
&\leq \frac{3}{1-\eta} \times \left\{ \sum_{j\in S} (1+\eta)\|f_j\| + \eta\lambda P(f_j) \right\} + \frac{\eta}{1-\eta} \sum_{j\in S} \lambda P(f_j) && \text{by (H.4)} \\
&= \frac{3(1+\eta)}{1-\eta} \sum_{j\in S} \|f_j\| + \frac{4\eta}{1-\eta} \sum_{j\in S} \lambda P(f_j).
\end{aligned}$$

Taking the smoothness norm term to the left hand side completes the proof. ■

Lemma H.6 *On the set $\mathcal{S}$, assuming the theoretical compatibility condition, if*

$$c_1 = \eta \left\{ \frac{2(2+\eta)}{1-\eta} + \frac{3(1+\eta)}{1-5\eta} \right\} \frac{\sqrt{s}}{\widetilde{\phi}(S)} < 1,$$

then

$$\|f\| \leq \frac{1}{1-c_1} \|f\|_n.$$

Proof On the set $\mathcal{S}$ we have

$$\|f\| \leq \|f\|_n + \eta \left\{ \sum_{j=1}^{p} \|f_j\| + \lambda P(f_j) \right\} \qquad \text{by definition}$$

For terms inside the brackets we have:

$$\begin{aligned}\sum_{j=1}^{p}\|f_j\| &= \sum_{j\in S}\|f_j\| + \sum_{j\in S^c}\|f_j\| \\ &\leq \sum_{j\in S}\|f_j\| + \frac{3(1+\eta)}{1-\eta}\sum_{j\in S}\|f_j\| \qquad \text{by (Theoretical-A)} \\ &= \frac{2(2+\eta)}{1-\eta}\sum_{j\in S}\|f_j\|,\end{aligned}$$

and

$$\sum_{j=1}^{p}\lambda P(f_j) \leq \frac{3(1+\eta)}{1-5\eta}\sum_{j\in S}\|f_j\| \qquad \text{by (Theoretical-A)}.$$

So now we have:

$$\begin{aligned}\|f\| &\leq \|f\|_n + \eta\left\{\frac{2(2+\eta)}{1-\eta}\sum_{j\in S}\|f_j\| + \frac{3(1+\eta)}{1-5\eta}\sum_{j\in S}\|f_j\|\right\} \\ &\leq \|f\|_n + \eta\left\{\frac{2(2+\eta)}{1-\eta} + \frac{3(1+\eta)}{1-5\eta}\right\}\frac{\sqrt{s}\|f\|}{\widetilde{\phi}(S)}. \qquad \text{by (Theoretical-B)}\end{aligned}$$

By definition of c_1 we get

$$(1-c_1)\|f\| \leq \|f\|_n,$$

which completes the proof. ■

Lemma H.7 *Under the conditions of Lemma H.6,* (Theoretical-B) ⇒ (Empirical-B) *with*

$$\{\phi(S)\}^{-1} = \left\{(1+\eta) + \frac{3\eta(1+\eta)}{1-5\eta}\right\}\frac{1}{(1-c_1)\widetilde{\phi}(S)}.$$

Proof We have that

$$\begin{aligned}\sum_{j\in S}\|f_j\|_n &\leq (1+\eta)\sum_{j\in S}\|f_j\| + \eta\sum_{j\in S}\lambda P(f_j) \qquad \text{by (H.4)} \\ &\leq (1+\eta)\sum_{j\in S}\|f_j\| + \frac{3\eta(1+\eta)}{1-5\eta}\sum_{j\in S}\|f_j\| \qquad \text{by (Theoretical-A)} \\ &\leq \left\{(1+\eta) + \frac{3\eta(1+\eta)}{1-5\eta}\right\}\frac{\sqrt{s}\|f\|}{\widetilde{\phi}(S)} \qquad \text{by (Theoretical-B)} \\ &\leq \left\{(1+\eta) + \frac{3\eta(1+\eta)}{1-5\eta}\right\}\frac{\sqrt{s}}{\widetilde{\phi}(S)}\frac{\|f\|_n}{1-c_1}, \qquad \text{by Lemma H.6} \\ &\equiv \|f\|_n\sqrt{s}/\phi(S).\end{aligned}$$

■

Appendix I. Some Results from van de Geer (2000)

Lemma I.1 (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, satisfying* (G.1)*. Then for all $\gamma \in \mathbb{R}^n$ and $\rho > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(Y_i-\mu_i)\gamma_i\right| \geq \rho\right) \leq 2\exp\left[-\frac{\rho^2}{8(K^2+\sigma_0^2)\sum_{i=1}^{n}\gamma_i^2}\right],$$

in particular if $\gamma_i = 1/n$ then we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}Y_i-\mu_i\right| \geq \rho\right) \leq 2\exp\left[-\frac{n\rho^2}{8(K^2+\sigma_0^2)}\right].$$

Lemma I.2 (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, satisfying* (G.1)*. Then for some constant $C = C(K,\sigma_0)$, and for all $\delta > 0$ satisfying*

$$\sqrt{n}\delta \geq 2C\left(\int_0^R H^{1/2}(u,\mathcal{F},Q_n)\,du \vee R\right),$$

we have

$$\mathbb{P}\left(\sup_{f_j\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_i)f_j(x_{ij})\right| \geq \delta\right) \leq C\exp\left[-\frac{n\delta^2}{4C^2R^2}\right].$$

References

Taylor Arnold, Veeranjaneyulu Sadhanala, and Ryan Tibshirani. *glmgen: Fast algorithms for generalized lasso problems*, 2014. R package version 0.0.3.

Miriam Ayer, Hugh D. Brunk, George M. Ewing, William T. Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647, 1955.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009a.

Amir Beck and Marc Teboulle. Gradient-based algorithms with applications to signal recovery. *Convex Optimization in Signal Processing and Communications*, pages 42–88, 2009b.

Peter Bühlmann and Sara van de Geer. *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

Alexandra Chouldechova and Trevor J. Hastie. Generalized additive model selection. *ArXiv Preprint arXiv:1506.03850*, 2015.

Arnak S. Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 2017.

Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 2012.

Bruno C. Feltes, Eduardo B. Chandelier, Bruno I. Grisci, and Márcio Dorn. CuMiDa: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *Journal of Computational Biology*, 26(4):376–386, 2019. doi: 10.1089/cmb.2018.0238. URL `https://doi.org/10.1089/cmb.2018.0238`. PMID: 30789283.

Jerome H. Friedman, Trevor J. Hastie, and Robert J. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1–22, 2010.

Asad Haris, Ali Shojaie, and Noah Simon. Wavelet regression and additive models for irregularly spaced data. In *Advances in Neural Information Processing Systems*, pages 8973–8983, 2018.

Trevor J. Hastie and Robert J. Tibshirani. *Generalized Additive Models*, volume 43. CRC Press, 1990.

Nicholas A. Johnson. A dynamic programming algorithm for the fused lasso and l_0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260, 2013.

Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 12 2010. doi: 10.1214/10-AOS825. URL `http://dx.doi.org/10.1214/10-AOS825`.

Yi Lin and Hao H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006.

Yin Lou, Jacob Bien, Rich Caruana, and Johannes Gehrke. Sparse partially linear additive models. *Journal of Computational and Graphical Statistics*, 25(4):1126–1140, 2016. doi: 10.1080/10618600.2015.1089775. URL `http://dx.doi.org/10.1080/10618600.2015.1089775`.

Lukas Meier, Sara van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

Sahand Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Manuscript, University of California, Berkeley, Dept. of Statistics and EECS*, 2011.

Yurii Nesterov. Gradient methods for minimizing composite objective function. Technical report, UCL, 2007.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

Ashley Petersen, Daniela Witten, and Noah Simon. Fused lasso additive model. *Journal of Computational and Graphical Statistics*, 25(4):1005–1025, 2016.

Aaditya Ramdas and Ryan J. Tibshirani. Fast and flexible ADMM algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, 25(3):839–858, 2015.

Garvesh Raskutti, Bin Yu, and Martin J. Wainwright. Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness. In *Advances in Neural Information Processing Systems*, pages 1563–1570, 2009.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *The Journal of Machine Learning Research*, 13(1):389–427, 2012.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

Veeranjaneyulu Sadhanala and Ryan J. Tibshirani. Additive models with trend filtering. *The Annals of Statistics*, 47(6):3032–3068, 2019.

Noah Simon and Robert Tibshirani. Standardization and the group lasso penalty. *Statistica Sinica*, 22(3):983, 2012.

Zhiqiang Tan and Cun-Hui Zhang. Doubly penalized estimation in additive regression with high-dimensional data. *The Annals of Statistics*, 47(5):2567 – 2600, 2019. doi: 10.1214/18-AOS1757. URL `https://doi.org/10.1214/18-AOS1757`.

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

Sara van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

Sara van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, pages 614–645, 2008.

Sara van de Geer. The lasso with within group structure. In J. Antoch, M. Hušková, and P. K. Sen, editors, *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*, volume 7 of *IMS Collections*, pages 235–244. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2010. doi: 10.1214/10-IMSCOLL723. URL `https://doi.org/10.1214/10-IMSCOLL723`.

Sara van de Geer. *Estimation and Testing Under Sparsity: École d'Été de Probabilités De Saint-Flour XLV - 2015*. Springer Publishing Company, Incorporated, 1st edition, 2016. ISBN 3319327739, 9783319327730.

Sara van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Junming Yin, Xi Chen, and Eric P. Xing. Group sparse additive models. In *Proceedings of the International Conference on Machine Learning*, volume 2012, page 871. NIH Public Access, 2012.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Ming Yuan and Ding-Xuan Zhou. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564 – 2593, 2016. doi: 10.1214/15-AOS1422. URL `https://doi.org/10.1214/15-AOS1422`.

Xingzhi Zhan. Extremal eigenvalues of real symmetric matrices with entries in an interval. *SIAM Journal on Matrix Analysis and Applications*, 27(3):851–860, 2005.

Tuo Zhao, Xingguo Li, Han Liu, and Kathryn Roeder. *SAM: Sparse Additive Modelling*, 2014. URL `http://CRAN.R-project.org/package=SAM`. R package version 1.0.5.

Hui Zou and Trevor J. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.