Faithful Variable Screening for High-Dimensional Convex Regression

Min Xu[1], Minhua Chen[2,3], and John Lafferty[2,3]

[1]Machine Learning Department, Carnegie Mellon University
[2]Department of Statistics, University of Chicago
[3]Department of Computer Science, University of Chicago

July 24, 2014

Abstract

We consider the problem of estimating a convex function of several variables from noisy values of the function at a finite sample of input points. Convex function estimation is subject to the curse of dimensionality where the sample size necessary for consistent estimation increases exponentially with the dimensionality of the observed variables p. However, if the function is sparse, with $s << p$ relevant variables, then one could achieve consistency in the high-dimensional setting by first identifying the s variables. We develop a faithful screening procedure to compute a set S that contains the s relevant variables. Our approach is a two-stage method that estimates a sum of p one-dimensional convex functions, followed by one-dimensional concave regression fits on the residuals. The method is based on quadratic programming, and in contrast to standard sparse additive models, requires no tuning parameters for smoothness. Under appropriate assumptions, we prove that the procedure is faithful in the population setting, yielding no false negatives, and we give a finite sample statistical analysis. In addition, we introduce algorithms for efficiently carrying out the required quadratic programs. The approach leads to significant computational and statistical advantages over fitting a full model, and provides an effective, practical approach to variable screening in convex regression.

1 Introduction

Shape restrictions such as monotonicity, convexity, and concavity provide a natural way of limiting the complexity of many statistical estimation problems. Shape-constrained estimation is not as well understood as more traditional nonparametric estimation involving smoothness constraints. For instance, the minimax rate of convergence for multivariate convex regression has yet to be rigorously established in full generality. Even the one-dimensional case is challenging, and has been of recent interest [**?**].

In this paper we study the problem of variable selection in multivariate convex regression. Assuming that the regression function is convex and sparse, our goal is to identify the relevant variables. We show that it suffices to estimate a sum of one-dimensional convex functions, leading to significant computational and statistical advantages. This is in contrast to general nonparametric regression, where fitting an additive model can result in false negatives. Our approach is based on a two-stage quadratic programming procedure. In the first stage, we fit an convex additive model, imposing a sparsity penalty. In the second stage, we fit a concave function on the residual for each variable. As we show, this non-intuitive second stage is in general necessary. Our first result is that this procedure is faithful in the population setting, meaning that it results in no false negatives,

under mild assumptions on the density of the covariates. Our second result is a finite sample statistical analysis of the procedure, where we upper bound the statistical rate of convergence. An additional contribution is to show how the required quadratic programs can be formulated to be more scalable. We give simulations to illustrate our method, showing that it performs in a manner that is consistent with our analysis.

Estimation of convex functions arises naturally in several applications. Examples include geometric programming [Boyd and Vandenberghe, 2004], computed tomography [Prince and Willsky, 1990], target reconstruction [Lele et al., 1992], image analysis [Goldenshluger and Zeevi, 2006] and circuit design [Hannah and Dunson, 2012]. Other applications include queuing theory [Chen and Yao, 2001] and economics, where it is of interest to estimate concave utility functions [Meyer and Pratt, 1968]. See Lim and Glynn [2012] for other applications. Beyond cases where the assumption of convexity is natural, the convexity assumption can be attractive as a tractable, nonparamametric relaxation of the linear model.

Variable selection in general nonparametric regression or function estimation is a notoriously difficult problem. Lafferty and Wasserman [2008] develop a greedy procedure for adjusting bandwidths in a local linear regression estimator, and show that the procedure achieves the minimax rate as if the relevant variables were isolated in advance. But the method only provably scales to dimensions p that grow logarithmically in the sample size n, so that $p = O(\log n)$. This is in contrast to the high dimensional scaling behavior known to hold for sparsity selection in linear models using ℓ_1 penalization, where the sample size n is logarithmic in the dimension p. Bertin and Lecué [2008] develop an optimization-based approach in the nonparametric setting, applying the lasso in a local linear model at each test point. Here again, however, the method only scales as $p = O(\log n)$, the low-dimensional regime. An approximation theory approach to the same problem is presented in DeVore et al. [2011], using techniques based on hierarchical hashing schemes, similar to those used for "junta" problems [Mossel et al., 2004]. Here it is shown that the sample complexity scales as $n > \log p$ if one adaptively selects the points on which the high-dimensional function is evaluated.

Comminges and Dalalyan [2012] show that the exponential scaling $n = O(\log p)$ is achievable if the underlying function is assumed to be smooth with respect to a Fourier basis. They also give support for the intrinsic difficulty of variable selection in nonparametric regression, giving lower bounds showing that consistent variable selection is not possible if $n < \log p$ or if $n < \exp s$, where s is the number of relevant variables. Variable selection over kernel classes is studied by Koltchinskii and Yuan [2010].

Perhaps most closely related to the present work is the framework studied by Raskutti et al. [2012] for sparse additive models, where sparse regression is considered under an additive assumption, with each component function belonging to an RKHS. An advantage of working over an RKHS is that nonparametric regression with a sparsity-inducing regularization penalty can be formulated as a finite dimensional convex cone optimization. On the other hand, smoothing parameters for the component Hilbert spaces must be chosen, leading to extra tuning parameters that are difficult to select in practice. In addition, the additive model must be assumed to be correct for consistent variable selection.

While nonparametric, the convex regression problem is naturally formulated using finite dimensional convex optimization, with no additional tuning parameters. The convex additive model can be used for convenience, without assuming it to actually hold, for the purpose of variable selection. As we show, our method scales to high dimensions, with a dependence on the intrinsic dimension s that scales polynomially, rather than exponentially as in the general case analyzed in Comminges and Dalalyan [2012].

In the following section we give a high-level summary of our technical results, including additive faithfulness, variable selection consistency, and high dimensional scaling. In Section 4 we give a detailed account of our method and the conditions under which we can guarantee consistent variable selection. In Section 6 we show how the required quadratic programs can be reformulated to be more efficient and scalable. In Section 7 we give the details of our finite sample analysis, showing that a sample size growing as $n = O\big(\text{poly}(s) \log p\big)$ is sufficient for variable selection. In Section **??** we report the results of simulations that illustrate our methods and theory. The full proofs are

given in a technical appendix.

2 Related Work

Variable selection in general nonparametric regression or function estimation is a notoriously difficult problem. Lafferty and Wasserman [2008] develop a greedy procedure for adjusting bandwidths in a local linear regression estimator, and show that the procedure achieves the minimax rate as if the relevant variables were isolated in advance. But the method only provably scales to dimensions p that grow logarithmically in the sample size n, so that $p = O(\log n)$. Note that this is the opposite of the high dimensional scaling behavior known to hold for sparsity selection in linear models using ℓ_1 penalization, where the sample size n is logarithmic in the dimension p. Bertin and Lecué [2008] develop an optimization-based approach in the nonparametric setting, applying the lasso in a local linear model at each test point. Here again, however, the method only scales as $p = O(\log n)$, the low-dimensional regime. An approximation theory approach to the same problem is presented in DeVore et al. [2011], using techniques based on hierarchical hashing schemes, similar to those used for "junta" problems Mossel et al. [2004]. Here it is shown that the sample complexity scales as $n > \log p$ if one adaptively selects the points on which the high-dimensional function is evaluated.

Comminges and Dalalyan [2012] show that the exponential scaling $p = O(\log n)$ is achievable if the underlying function is assumed to be smooth with respect to a Fourier basis. They also give support for the intrinsic difficulty of variable selection in nonparametric regression, giving lower bounds showing that sparsistency is not possible if $n < \log p$ or if $n < \exp s$, where s is the number of relevant variables. Variable selection over kernel classes is studied by Koltchinskii and Yuan [2010].

Perhaps most closely related to the present work, is the framework studied by Raskutti et al. [2012] for sparse additive models, where sparse regression is considered under an additive assumption, with each component function belonging to an RKHS. An advantage of working over an RKHS, in contrast to the other papers mentioned above, is that nonparametric regression with a sparsity-inducing regularization penalty can be formulated as a finite dimensional convex cone optimization. On the other hand, smoothing parameters for the component Hilbert spaces must be chosen, leading to extra tuning parameters that are difficult to select in practice. In addition, the additive model must be assumed to be correct for sparsistent variable selection.

An attraction of the convex function estimation framework we consider in this paper is that the additive model can be used for convenience, without assuming it to actually hold. While nonparametric, the problem is naturally formulated using finite dimensional convex optimization, but with no additional tuning parameters. As we show below, our method scales to high dimensions, with a dependence on the intrinsic dimension s that scales polynomially, rather than exponentially as in the general case analyzed in Comminges and Dalalyan [2012].

Notation. If $\boldsymbol{x}$ is a vector, we use $\boldsymbol{x}_{-k}$ to denote the vector with the k-th coordinate removed. If $\boldsymbol{v} \in \mathbb{R}^n$, then $v_{(1)}$ denotes the smallest coordinate of $\boldsymbol{v}$ in magnitude, and $v_{(j)}$ denotes the j-th smallest; $\mathbf{1}_n \in \mathbb{R}^n$ is the all ones vector. If $X \in \mathbb{R}^p$ and $S \subset \{1, ..., p\}$, then X_S is the subvector of X restricted to the coordinates in S. Given n samples $X^{(1)}, ..., X^{(n)}$, we use $\overline{X}$ to denote the empirical average. Given a random variable X_k and a scalar x_k, we use $\mathbb{E}[\cdot \,|\, x_k]$ as a shorthand for $\mathbb{E}[\cdot \,|\, X_k = x_k]$.

3 Overview of Results

In this section we provide a high-level description of our technical results. The full technical details, the precise statement of the results, and their detailed proofs are provided in following sections.

Our main contribution is an analysis of an additive approximation for identifying relevant variables in convex regression. We prove a result that shows when and how the additive approximation can be used without introducing false negatives in the population setting. In addition, we develop algorithms for the efficient implementation of the quadratic programs required by the procedure.

We first establish some notation, to be used throughout the paper. If $\boldsymbol{x}$ is a vector, we use $\boldsymbol{x}_{-k}$ to denote the vector with the k-th coordinate removed. If $\boldsymbol{v} \in \mathbb{R}^n$, then $v_{(1)}$ denotes the smallest coordinate of $\boldsymbol{v}$ in magnitude, and $v_{(j)}$ denotes the j-th smallest; $\mathbf{1}_n \in \mathbb{R}^n$ is the all ones vector. If $X \in \mathbb{R}^p$ is a random variable and $S \subset \{1, ..., p\}$, then X_S is the subvector of X restricted to the coordinates in S. Given n samples $X^{(1)}, ..., X^{(n)}$, we use $\overline{X}$ to denote the sample mean. Given a random variable X_k and a scalar x_k, we use $\mathbb{E}[\,\cdot \,|\, x_k]$ as a shorthand for $\mathbb{E}[\,\cdot \,|\, X_k = x_k]$.

3.1 Faithful screening

The starting point for our approach is the observation that least squares nonparametric estimation under convexity constraints is equivalent to a finite dimensional quadratic program. Specifically, the infinite dimensional optimization

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} (Y_i - f(\boldsymbol{x}_i))^2 \\ \text{subject to} \quad & f : \mathbb{R}^p \to \mathbb{R} \text{ is convex} \end{aligned} \tag{3.1}$$

is equivalent to the finite dimensional quadratic program

$$\begin{aligned} \text{minimize}_{f,\beta} \quad & \sum_{i=1}^{n} (Y_i - f_i)^2 \\ \text{subject to} \quad & f_j \geq f_i + \beta_i^T (\boldsymbol{x}_j - \boldsymbol{x}_i), \text{ for all } i, j. \end{aligned} \tag{3.2}$$

See Boyd and Vandenberghe [2004], Section 6.5.5. Here f_i is the estimated function value $f(\boldsymbol{x}_i)$, and the vectors $\beta_i \in \mathbb{R}^d$ represent supporting hyperplanes to the epigraph of f. Importantly, this finite dimensional quadratic program does not have tuning parameters for smoothing the function.

For general regression, using an additive approximation for variable selection may make errors. In particular, the nonlinearities in the regression function may result in an additive component being wrongly zeroed out. We show that this will not happen for convex regression under appropriate conditions.

We say that a differentiable function f depends on variable x_k if $\partial_{x_k} f \neq 0$ with probability greater than zero. An additive approximation is given by

$$\{f_k^*\}, \mu^* := \underset{f_1, ..., f_p, \mu}{\arg\min} \Big\{ \mathbb{E}\Big(f(X) - \mu - \sum_{k=1}^{p} f_k(X_k) \Big)^2 \,:\, \mathbb{E} f_k(X_k) = 0 \Big\}. \tag{3.3}$$

We say that f is *additively faithful* in case $f_k^* = 0$ implies that f does not depend on coordinate k. Additive faithfulness is a desireable property since it implies that an additive approximation may allow us to screen out irrelevant variables.

Our first result shows that convex multivariate functions are additively faithful under the following assumption on the distribution of the data.

Definition 3.1. Let $p(\boldsymbol{x})$ be a density supported on $[0,1]^p$. Then p satisfies the *boundary points condition* if for all j, and for all $\boldsymbol{x}_{-j}$,

$$\frac{\partial p(\boldsymbol{x}_{-j} \,|\, x_j)}{\partial x_j} = \frac{\partial^2 p(\boldsymbol{x}_{-j} \,|\, x_j)}{\partial x_j^2} = 0 \quad \text{at } x_j = 0 \text{ and } x_j = 1.$$

As discussed in Section 4, this is a relatively weak condition. Our first result is that this condition suffices in the population setting of convex regression.

Let p be a positive density supported on $C = [0,1]^p$ that satisfies the boundary points property. If f is convex and twice differentiable, then f is additively faithful under p.

Intuitively, an additive approximation zeroes out variable k when, fixing x_k, every "slice" of f integrates to zero. We prove this result by showing that "slices" of convex functions that integrate to zero cannot be "glued together" while still maintaining convexity.

While this shows that convex functions are additively faithful, it is difficult to estimate the optimal additive functions. The difficulty is that f_k^* need not be a convex function, as we show through a counterexample in Section 4. Since the true regression function f is convex, it is natural to ask when it is sufficient to estimate a convex additive model. Unfortunately, a convex additive approximation is not generally faithful. In other words, it could be that $f_k^* \equiv 0$ even for a relevant variable x_k. But our next result shows that this type of error can be detected by fitting a *concave* function to the residual. If this concave function is zero, we can then safely mark x_k as an irrelevant variable.

Suppose $p(\boldsymbol{x})$ is a positive density on $C = [0,1]^p$ that satisfies the boundary points condition. Suppose that f is convex and twice-differentiable. and that $\partial_{x_k} f$, $\partial_{x_k} p(\boldsymbol{x}_{-k} \,|\, x_k)$, and $\partial^2_{x_k} p(\boldsymbol{x}_{-k} \,|\, x_k)$ are all continuous as functions on C. Define

$$\{f_k^*\}_{k=1}^p, \mu^* = \arg\min_{\{f_k\},\mu} \Big\{ \mathbb{E}\Big(f(X) - \mu - \sum_{k=1}^s f_k(X_k) \Big)^2 \,:\, f_k \in \mathcal{C}^1,\, \mathbb{E} f_k(X_k) = 0 \Big\} \tag{3.4}$$

where $\mathcal{C}^1$ is the set of univariate convex functions, and

$$g_k^* = \arg\min_{g_k} \Big\{ \mathbb{E}\Big(f(X) - \mu^* - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k \Big)^2 \,:\, g_k \in \text{-}\mathcal{C}^1, \mathbb{E} g_k(X_k) = 0 \Big\}, \tag{3.5}$$

with -$\mathcal{C}^1$ denoting the set of univariate concave functions. Then $f_k^* = 0$ and $g_k^* = 0$ implies that f does not depend on x_k, i.e., $\partial_{x_k} f(\boldsymbol{x}) = 0$ with probability one.

This result naturally suggests a two-stage screening procedure for variable selection. In the first stage we fit a sparse convex additive model $\{\widehat{f}_k\}$. In the second stage we fit a concave function $\widehat{g}_k$ to the residual for each variable having a zero convex component $\widehat{f}_k$. If both $\widehat{f}_k = 0$ and $\widehat{g}_k = 0$, we can safely discard variable x_k. As a shorthand, we refer to this two-stage procedure as AC/DC. In the AC stage we fit an additive convex model. In the DC stage we fit decoupled concave functions on the residuals. The decoupled nature of the DC stage allows all of the fits to be carried out in parallel. Our next results concern the required optimizations, and their finite sample statistical performance.

3.2 Optimization

In Section 6 we present optimization algorithms for the additive convex regression stage. The convex constraints for the additive functions, analogous to the multivariate constraints (3.2), are that each component $f_k(\cdot)$ can be represented by its supporting hyperplanes, i.e.,

$$f_{ki'} \geq f_{ki} + \beta_{ki}(x_{ki'} - x_{ki}) \quad \text{for all } i, i' \tag{3.6}$$

where $f_{ki} := f_k(x_{ki})$ and β_{ki} is the subgradient at point x_{ki}. While this apparently requires $O(n^2 p)$ equations to impose the supporting hyperplane constraints, in fact, only $O(np)$ constraints suffice. This is because univariate convex functions are characterized by the condition that the subgradient, which is a scalar, must increase monotonically. This observation leads to a reduced quadratic program with $O(np)$ variables and $O(np)$ constraints.

Directly applying a QP solver to this optimization is still computationally expensive for relatively large n and p. We thus develop a block coordinate descent method, where in each step we solve a sparse quadratic program involving $O(n)$ variables and $O(n)$ constraints. This is efficiently solved using optimization packages such as MOSEK. The details of these optimizations are given in Section 6.

3.3 Finite sample analysis

In Section 7 we analyze the finite sample variable selection consistency of convex additive modeling, without making the assumption that the true regression function f_0 is additive. Our analysis first establishes a sufficient deterministic condition for variable selection consistency, and then considers a stochastic setting. Our proof technique decomposes the KKT conditions for the optimization in a manner that is similar to the now standard *primal-dual witness* method [Wainwright, 2009].

We prove separate results that allow us to analyze false negative rates and false positive rates. To control false positives, we analyze scaling conditions on the regularization parameter λ_n for group sparsity needed to zero out irrelevant variables $k \in S^c$, where $S \subset \{1, \ldots, p\}$ is the set of relevant variables. To control false negatives, we analyze the restricted regression where the variables in S^c are zeroed out, following the primal-dual strategy.

Each of our theorems uses a subset of the following assumptions:

A1: X_S, X_{S^c} are independent.

A2: f_0 is convex and twice-differentiable.

A3: $|\partial_{x_k} f_0| \leq L$ for all k

A4: The noise is mean-zero sub-Gaussian, independent of X.

Our analysis involves parameters α_f and α_g, which are measures of the signal strength of the weakest variable:

$$\alpha_f = \inf_{f \in \mathcal{C}^p \,:\, \exists k,\, f_k^* \neq 0 \,\wedge\, f_k = 0} \Big\{ \mathbb{E}\big(f_0(X) - f(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X)\big)^2 \Big\}$$

$$\alpha_g = \min_{k \in S \,:\, g_k^* \neq 0} \Big\{ \mathbb{E}\big(f_0(X) - f^*(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X) - g_k^*(X_k)\big)^2 \Big\}.$$

Intuitively, if α_f is small, then it is easier to make a false omission in the additive convex stage of the procedure. If α_g is small, then it is easier to make a false omission in the decoupled concave stage of the procedure.

We make strong assumptions on the covariates in A1 in order to make very weak assumptions on the true regression function f_0 in A2; in particular, we do not assume that f_0 is additive. Relaxing this condition is an important direction for future work. We also include an extra Lipschitz constraint so that we can use existing covering number results [Bronshtein, 1976]. Recent work Guntuboyina and Sen [2013] shows that the Lipschitz constraint is not required with more advanced empirical process theory techniques; we leave the incorporation of this development as future work.

Our main result is the following. Suppose assumptions A1-A4 hold. Let $\{\widehat{f}_i\}$ be any AC solution and let $\{\widehat{g}_k\}$ be any DC solution, both estimated with regularization parameter λ scaling as $\lambda = \Theta\Big(sLb\sqrt{\frac{1}{n}\log^2 np}\Big)$. Suppose in addition that

$$\frac{\alpha_f}{\sigma} \geq c(Lb)^3 \sqrt{\frac{s^5}{n^{4/5}} \log^2 np} \tag{3.7}$$

$$\frac{\alpha_g^2}{\sigma} \geq c(Lb)^3 \sqrt{\frac{s^5}{n^{4/5}} \log^2 2np}. \tag{3.8}$$

Then, for sufficiently large n, with probability at least $1 - \frac{1}{n}$:

$$\widehat{f}_k \neq 0 \text{ or } \widehat{g}_k \neq 0 \text{ for all } k \in S$$

$$\widehat{f}_k = 0 \text{ and } \widehat{g}_k = 0 \text{ for all } k \notin S.$$

This shows that variable selection consistency is achievable under exponential scaling of the ambient dimension, $p = O(\exp(n^c)) for c < 1$, as for linear models. The cost of nonparametric estimation is reflected in the scaling with respect to $s = |S|$, which can grow only as $o(n^{4/25})$.

We remark that Comminges and Dalalyan [2012] show that under traditional smoothness constraints, variable selection is achievable only if $n > O(e^s)$. Here we demonstrate that convexity yields the scaling $n = O(\text{poly}(s))$.

4 Additive Faithfulness

For general regression, an additive approximation may result in a relevant variable being incorrectly marked as irrelevant. Such mistakes are inherent to the approximation and may persist even in the population setting. In this section we give examples of this phenomenon, and then show how the convexity assumption changes the behavior of the additive approximation. We begin with a lemma that characterizes the components of the additive approximation under mild conditions.

Lemma 4.1. *Let F be a distribution on $C = [0,1]^p$ with a positive density function p. Let $f : C \to \mathbb{R}$ be an integrable function, and define*

$$f_1^*, ..., f_p^*, \mu^* := \arg\min \left\{ \mathbb{E}\Big(f(X) - \mu - \sum_{k=1}^{p} f_k(X_k)\Big)^2 \,:\, \mathbb{E} f_k(X_k) = 0,\ \forall k = 1, \dots, p \right\}. \tag{4.1}$$

Then $\mu^* = \mathbb{E}f(X)$,

$$f_k^*(x_k) = \mathbb{E}\Big[f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) \,|\, x_k\Big] - \mathbb{E}f(X) \tag{4.2}$$

and this solution is unique.

Lemma 4.1 follows from the stationarity conditions of the optimal solution. This result is known, and criterion (**??**) is used in the backfitting algorithm for fitting additive models. We include a proof as our results build on it.

Proof. Let $f_1^*, ..., f_p^*, \mu^*$ be the minimizers as defined. We first show that the optimal μ is $\mu^* = \mathbb{E}f(X)$ for any $f_1, ..., f_k$ such that $\mathbb{E}f_k(X_k) = 0$. This follows from the stationarity condition, which states that $\mu^* = \mathbb{E}[f(X) - \sum_k f_k(X_k)] = \mathbb{E}[f(X)]$. Uniqueness is apparent because the second derivative is strictly larger than zero and strong convexity is guaranteed.

We now turn our attention toward the f_k^*s. It must be that f_k^* minimizes

$$\mathbb{E}\Big[\big(f(X) - \mu^* - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - f_k(X_k)\big)^2\Big] \tag{4.3}$$

subject to $\mathbb{E}f_k(X_k) = 0$. Fixing x_k, we will show that the value

$$\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}(X_{k'}) \,|\, x_k] - \mu^* \tag{4.4}$$

uniquely minimizes

$$\min_{f_k(x_k)} \int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x})\Big(f(\boldsymbol{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'}) - f_k(x_k) - \mu^*\Big)^2 d\boldsymbol{x}_{-k}. \tag{4.5}$$

The first-order optimality condition gives us

$$\int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}) f_k(x_k) d\boldsymbol{x}_{-k} = \int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x})(f(\boldsymbol{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'}) - \mu^*) d\boldsymbol{x}_{-k} \tag{4.6}$$

$$p(x_k) f_k(x_k) = \int_{\boldsymbol{x}_{-k}} p(x_k) p(\boldsymbol{x}_{-k} \,|\, x_k)(f(\boldsymbol{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'}) - \mu^*) d\boldsymbol{x}_{-k} \tag{4.7}$$

$$f_k(x_k) = \int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k} \,|\, x_k)(f(\boldsymbol{x}) - \sum_{k' \neq k} f_{k'}(x_{k'}) - \mu^*) d\boldsymbol{x}_{-k} \tag{4.8}$$

The square error objective is strongly convex, and the second derivative with respect to $f_k(x_k)$ is $2p(x_k)$, which is always positive under the assumption that p is positive. Therefore, the solution $f_k^*(x_k) = \mathbb{E}[f(X) \,|\, x_k] - \mathbb{E}f(X)$ is unique. Noting that $\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}(X_{k'})|x_k] - \mathbb{E}f(X)$ has mean zero as a function of x_k completes the proof. □

In the case that the distribution in Lemma 4.1 is a product distribution, the additive components take on a simple form.

Corollary 4.1. *Let F be a product distribution on $C = [0,1]^p$ with density function p which is positive on C. Let $\mu^*, f_k^*(x_k)$ be defined as in Lemma 4.1. Then $\mu^* = \mathbb{E}f(X)$ and $f_k^*(x_k) = \mathbb{E}[f(X) \,|\, x_k] - \mathbb{E}f(X)$ and this solution is unique.*

In particular, if F is the uniform distribution, then $f_k^*(x_k) = \int f(x_k, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k}$.

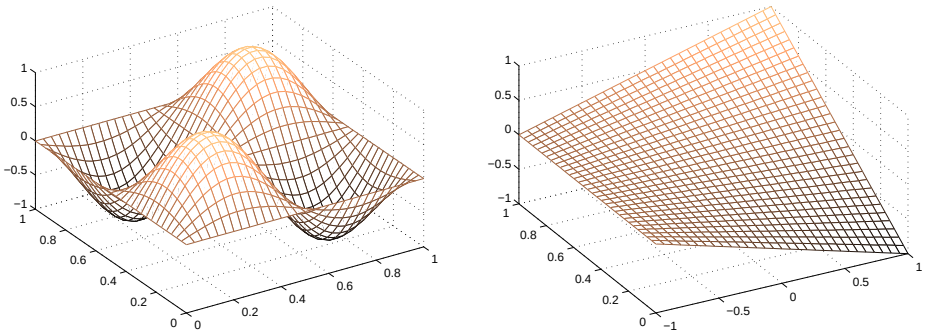
Example 4.1. Using Corollary 4.1, we give two examples of *additively unfaithfulness* under the uniform distribution—where relevant variables are erroneously marked as irrelevant under an additive approximation. First, consider the following function:

$$f(x_1, x_2) = \sin(2\pi x_1)\sin(2\pi x_2) \quad \text{(egg carton)} \tag{4.9}$$

defined for $(x_1, x_2) \in [0,1]^2$. Then $\int_{x_2} f(x_1, x_2) dx_2 = 0$ and $\int_{x_1} f(x_1, x_2) dx_1 = 0$ for each x_1 and x_2. An additive approximation would set $f_1 = 0$ and $f_2 = 0$. Next, consider the function

$$f(x_1, x_2) = x_1 x_2 \quad \text{(tilting slope)} \tag{4.10}$$

defined for $x_1 \in [-1, 1],\ x_2 \in [0, 1]$. In this case $\int_{x_1} f(x_1, x_2) dx_1 = 0$ for each x_2; therefore, we expect $f_2 = 0$ under the additive approximation. This function, for every fixed x_2, is a zero-intercept linear function of x_1 with slope x_2.



(a) egg carton (b) tilting slope

Figure 1: Two additively unfaithful functions. Relevant variables are zeroed out under an additive approximation because every "slice" of the function integrates to zero.

In order to exploit additive models in variable selection, it is important to understand when the additive approximation accurately captures all of the relevant variables. We call this property *additive faithfulness*. We first formalize the intuitive notion that a multivariate function f *depends on* a coordinate x_k.

Definition 4.1. Let F be a distribution on $C = [0,1]^p$, and $f : C \to \mathbb{R}$. We say that f *depends on coordinate* k if, for all $x_k \in [0,1]$, the set $\left\{x_k' \in [0,1] \,:\, f(x_k, \boldsymbol{x}_{-k}) = f(x_k', \boldsymbol{x}_{-k}) \text{ for almost all } \boldsymbol{x}_{-k}\right\}$ has probability strictly less than one. If f is differentiable, then f depends on k if $\partial_{x_k} f \neq 0$ with probability greater than zero.

Suppose we have the additive approximation

$$f_k^*, \mu^* := \operatorname*{arg\,min}_{f_1,\ldots,f_p,\mu} \Big\{ \mathbb{E}\Big[\Big(f(X) - \sum_{k=1}^{p} f_k(X_k) - \mu \Big)^2 \Big] \,:\, \mathbb{E} f_k(X_k) = 0 \Big\}. \tag{4.11}$$

We say that f is *additively faithful* under F in case $f_k^* = 0$ implies that f does not depend on coordinate k.

Additive faithfulness is an attractive property because it implies that, in the population setting, the additive approximation yields consistent variable selection.

4.1 Additive Faithfulness of Convex Functions

We now show that under a general class of distributions which we characterize below, convex multivariate functions are additively faithful.

Definition 4.2. A density $p(\boldsymbol{x})$ be a density supported on $[0,1]^p$ satisfies the *boundary points condition* if, for all j, and for all $\boldsymbol{x}_{-j}$:

$$\frac{\partial p(\boldsymbol{x}_{-j} \,|\, x_j)}{\partial x_j} = \frac{\partial^2 p(\boldsymbol{x}_{-j} \,|\, x_j)}{\partial x_j^2} = 0 \quad \text{at } x_k = 0, x_k = 1 \tag{4.12}$$

The boundary points condition is a weak condition. For instance, it is satisfied when the density is flat at the boundary of support—more precisely, when the joint density satisfies the property that $\frac{\partial p(x_j, \boldsymbol{x}_{-j})}{\partial x_j} = \frac{\partial^2 p(x_j, \boldsymbol{x}_{-j})}{\partial x_j^2} = 0$ at points $x_j = 0, x_j = 1$. The boundary points property is also trivially satisfied when p is a product density.

The following theorem is the main result of this section.

Theorem 4.1. *Let p be a positive density supported on $C = [0,1]^p$ that satisfies the boundary points property. If f is convex and twice differentiable, then f is additively faithful under p.*

We pause to give some intuition before we presenting the full proof. Suppose that the the underlying distribution has a product density. Then we know from Lemma 4.1 that the additive approximation zeroes out k when, fixing x_k, every "slice" of f integrates to zero. We prove Theorem 4.1 by showing that "slices" of convex functions that integrate to zero cannot be "glued" together while still maintaining convexity.

Proof. Fixing k and using the result of Lemma 4.1, we need only show that for all x_k, $\mathbb{E}[f(X) - \sum_{k'} f_{k'}(X_{k'}) \,|\, x_k] - \mathbb{E} f(X) = 0$ implies that f does not depend on coordinate k, i.e., $\partial_{x_k} f(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$.

Let us use the shorthand notation that $r(\boldsymbol{x}_{-k}) = \sum_{k' \neq k} f_{k'}(x_{k'})$ and assume without loss of generality that $\mu^* = E[f(X)] = 0$. We then assume that for all x_k,

$$\mathbb{E}[f(X) - r(X_{-k}) \,|\, x_k] \equiv \int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k} \,|\, x_k) \big(f(\boldsymbol{x}) - r(\boldsymbol{x}_{-k}) \big) = 0. \tag{4.13}$$

We let $p'(\boldsymbol{x}_{-k} \,|\, x_k)$ denote $\frac{\partial p(\boldsymbol{x}_{-k} \,|\, x_k)}{\partial x_k}$ and $p''(\boldsymbol{x}_{-k} \,|\, x_k)$ denote $\frac{\partial^2 p(\boldsymbol{x}_{-k} \,|\, x_k)}{\partial x_k^2}$ and likewise for $f'(x_k, \boldsymbol{x}_{-k})$ and $f''(x_k, \boldsymbol{x}_{-k})$. We then differentiate under the integral, valid because all functions are bounded, and obtain

$$\int_{\boldsymbol{x}_{-k}} p'(\boldsymbol{x}_{-k} \,|\, x_k) \big(f(\boldsymbol{x}) - r(\boldsymbol{x}_{-k}) \big) + p(\boldsymbol{x}_{-k} \,|\, x_k) f'(x_k, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0 \tag{4.14}$$

$$\int_{\boldsymbol{x}_{-k}} p''(\boldsymbol{x}_{-k} \,|\, x_k) \big(f(\boldsymbol{x}) - r(\boldsymbol{x}_{-k}) \big) + 2p'(\boldsymbol{x}_{-k} \,|\, x_k) f'(x_k, \boldsymbol{x}_{-k}) + p(\boldsymbol{x}_{-k} \,|\, x_k) f''(x_k, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0. \tag{4.15}$$

By the boundary points condition, we have that $p''(\boldsymbol{x}_{-k} \mid x_k)$ and $p'(\boldsymbol{x}_{-k} \mid x_k)$ are zero at $x_k = x_k^0 \equiv 0$. The integral equations then reduce to the following:

$$\int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k} \mid x_k^0) f'(x_k^0, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0 \tag{4.16}$$

$$\int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k} \mid x_k^0) f''(x_k^0, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0. \tag{4.17}$$

Because f is convex, $f(x_k, \boldsymbol{x}_{-k})$ must be a convex function of x_k for all $\boldsymbol{x}_{-k}$. Therefore, for all $\boldsymbol{x}_{-k}$, $f''(x_k^0, \boldsymbol{x}_{-k}) \geq 0$. Since $p(\boldsymbol{x}_{-k} \mid x_k^0) > 0$ by the assumption that p is a positive density, we have that $\forall \boldsymbol{x}_{-k}, f''(x_k^0, \boldsymbol{x}_{-k}) = 0$ necessarily.

The Hessian of f at $(x_k^0, \boldsymbol{x}_{-k})$ then has a zero at the k-th main diagonal entry. A positive semidefinite matrix with a zero on the k-th main diagonal entry must have only zeros on the k-th row and column; see proposition 7.1.10 of Horn and Johnson [1990]. Thus, at all $\boldsymbol{x}_{-k}$, the gradient of $f'(x_k^0, \boldsymbol{x}_{-k})$ with respect to $\boldsymbol{x}_{-k}$ must be zero. Therefore, $f'(x_k^0, \boldsymbol{x}_{-k})$ must be constant for all $\boldsymbol{x}_{-k}$. By equation 4.4, we conclude that $f'(x_k^0, \boldsymbol{x}_{-k}) = 0$ for all $\boldsymbol{x}_{-k}$. We can use the same reasoning for the case where $x_k = x_k^1$ and deduce that $f'(x_k^1, \boldsymbol{x}_{-k}) = 0$ for all $\boldsymbol{x}_{-k}$.

Because $f(x_k, \boldsymbol{x}_{-k})$ as a function of x_k is convex, it must be that, for all $x_k \in (0, 1)$ and for all $\boldsymbol{x}_{-k}$,

$$0 = f'(x_k^0, \boldsymbol{x}_{-k}) \leq f'(x_k, \boldsymbol{x}_{-k}) \leq f'(x_k^1, \boldsymbol{x}_{-k}) = 0 \tag{4.18}$$

Therefore f does not depend on x_k.

□

Theorem 4.1 plays an important role in our finite sample analysis, where we show that the additive approximation is variable selection consistent (or "sparsistent"), even when the true function is not additive.

Remark 4.1. We assume twice differentiability in Theorems 4.1 to simplify the proof. We expect, however, that this this smoothness condition is not necessary—every convex function can be approximated arbitrarily well by a smooth convex function.

Remark 4.2. We have not found natural conditions under which the opposite direction of additive faithfulness holds—conditions implying that if f does not depend on coordinate k, then f_k^* will be zero in the additive approximation. Suppose, for example, that f is only a function of X_1, X_2, and that (X_1, X_2, X_3) follows a degenerate 3-dimensional distribution where $X_3 = f(X_1, X_2) - f^*(X_1) - f_2^*(X_2)$. In this case X_3 exactly captures the additive approximation error. The best additive approximation of f would have a component $f_3^*(x_3) = x_3$ even though f does not depend on x_3.

Remark 4.3. It is possible to get a similar result for distributions with unbounded support, by using a limit condition $\lim_{|x_k| \to \infty} \frac{\partial p(\boldsymbol{x}_{-k} \mid x_k)}{\partial x_k} = 0$. Such a limit condition is too strong however, as it is not obeyed by the multivariate Gaussian distribution.

The following example shows that not all convex functions are additively faithful under multivariate Gaussian distributions.

Example 4.2. Consider a quadratic form $f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T} H \boldsymbol{x} + c^\mathsf{T} \boldsymbol{x}$ and a Gaussian distribution $X \sim N(0, \Sigma)$ where $\Sigma_{jj} = 1$ for all j. As we show in the appendix, the additive approximation has the following closed form.

1. If f does not depend on x_j, then $f_j^*(x_j) = 0$.
2. If f depends on x_j, then
$$f_j^*(x_j) = H_j^\mathsf{T} \Sigma_j x_j^2 + c_j x_j \tag{4.19}$$
where H_j and Σ_j are the jth rows of H_j and Σ_j, respectively.

Let $H = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$ and let $\Sigma = \begin{pmatrix} 1 & -\alpha \\ -\alpha & 1 \end{pmatrix}$. It can be checked that if $\alpha = \frac{1}{2}$, then $f_1^* = 0$ and additive faithfulness is violated. Moreover, if $\alpha > \frac{1}{2}$, then f_1^* is a concave function.

4.2 Convex Additive Models

Although convex functions are additively faithful—under appropriate conditions—it is difficult to estimate the optimal additive functions f_k^*s as defined in equation (4.1). The reason is that f_k^* need not be a convex function, as example 4.2 shows. Since the true regression function f is convex, it is natural to ask when it is sufficient to estimate an convex additive model

$$\{f_k^*\}_{k=1}^p = \arg\min \Big\{ \mathbb{E}\Big(f(X) - \sum_{k=1}^p f_k(X_k) \Big)^2 \,:\, f_k \in \mathcal{C}^1,\, \mathbb{E} f_k(X_k) = 0 \Big\} \tag{4.20}$$

where $\mathcal{C}^1$ is the set of univariate convex functions. While the convex functions $\{f_k^*\}$ are not additively faithful by themselves, faithfulness can be restored by coupling the f_k^*s with a set of concave functions:

$$g_k^* = \arg\min \Big\{ \mathbb{E}\Big(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k \Big)^2 \,:\, g_k \in \text{-}\mathcal{C}^1, \mathbb{E} g_k(X_k) = 0 \Big\}. \tag{4.21}$$

Theorem 4.2. *Suppose $p(\boldsymbol{x})$ is a positive density on $C = [0,1]^p$ that satisfies the boundary points condition. Suppose that f is convex and twice-differentiable, and that that $\partial_{x_k} f$, $\partial_{x_k} p(\boldsymbol{x}_{-k} \,|\, x_k)$, and $\partial_{x_k}^2 p(\boldsymbol{x}_{-k} \,|\, x_k)$ are all continuous as functions on C. Let f_k^* and g_k^* be as defined in equations (4.6) and (4.7). Then $f_k^* = 0$ and $g_k^* = 0$ implies that $\partial_{x_k} f(\boldsymbol{x}) = 0$ with probability one, so that f does not depend on x_k.*

Lemma 4.2. *Suppose $p(x)$ is a positive density on $[0,1]^p$ satisfying the boundary points condition. For any function $\phi(X_{-k})$ that does not depend on X_k,*

$$f_k^* = \arg\min_{f_k} \mathbb{E}\Big[\big(f(X) - \phi(X_{-k}) - f_k(X_k) \big)^2 \Big] \tag{4.22}$$

is given by $f_k^*(x_k) = \mathbb{E}\big[f(X) - \phi(X_{-k}) \,|\, x_k \big]$, and $f_k^* = 0$ implies that $\partial_{x_k} f(x) = 0$.

Proof. In the proof of Theorem 4.1, the only property of $r(\boldsymbol{x}_{-k})$ we used was the fact that $\partial_{x_k} r(\boldsymbol{x}_{-k}) = 0$. Therefore, the proof here is identical to that of Theorem 4.1 except that we let $\phi(\boldsymbol{x}_{-k}) = r(\boldsymbol{x}_{-k})$. □

Proof of theorem 4.2. Fix k. Let f_k^* and g_k^* be defined as in equation 4.6 and equation 4.7. Then we have that

$$f_k^* = \arg\min_{f_k} \Big\{ \mathbb{E}\big(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - f_k \big)^2 \,:\, f_k \in \mathcal{C}^1,\, \mathbb{E} f_k(X_k) = 0 \Big\} \tag{4.23}$$

$$g_k^* = \arg\min_{g_k} \Big\{ \mathbb{E}\big(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k \big)^2 \,:\, g_k \in \text{-}\mathcal{C}^1,\, \mathbb{E} g_k(X_k) = 0 \Big\} \tag{4.24}$$

Therefore, $f_k^* = 0$ and $g_k^* = 0$ implies that $\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) \,|\, x_k] = 0$. Now we use Lemma 4.2 with $\phi(\boldsymbol{x}_{-k}) = f(\boldsymbol{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'})$ and conclude that $f_k^* = 0$ and $g_k^* = 0$ together imply that f does not depend on x_k. □

4.3 Estimation Procedure

Theorem 4.2 naturally suggests a two-stage screening procedure for variable selection. In the first stage we fit a sparse convex additive model. In the second stage, for every variable marked as irrelevant in the first stage, we fit a univariate *concave* function separately on the residual for that variable. We refer to this procedure as AC/DC (additive convex/decoupled concave).

More precisely, given samples $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$, we perform the following steps.

Input: $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$, regularization parameter λ.

AC Stage: Estimate a sparse additive convex model:

$$\widehat{f}_1, ..., \widehat{f}_p, \widehat{\mu} = \underset{f_1,...,f_p \in \mathcal{C}_0^1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \mu - \sum_{k=1}^{p} f_k(x_{ik}) \Big)^2 + \lambda \sum_{k=1}^{p} \|f_k\|_\infty \tag{4.27}$$

DC Stage: Estimate concave functions for each k such that $\|\widehat{f}_k\|_\infty = 0$:

$$\widehat{g}_k = \underset{g_k \in \text{-}\mathcal{C}_0^1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \widehat{\mu} - \sum_{k'} \widehat{f}_{k'}(x_{ik'}) - g_k(x_{ik}) \Big)^2 + \lambda \|g_k\|_\infty \tag{4.28}$$

Output: Component functions $\{\widehat{f}_k\}$ and relevant variables $\widehat{S}$ where

$$\widehat{S}^c = \big\{ k : \|\widehat{f}_k\| = 0 \text{ and } \|\widehat{g}_k\| = 0 \big\}. \tag{4.29}$$

Figure 2: The AC/DC algorithm for variable selection in convex regression. The AC stage fits a sparse additive convex regression model, using a quadratic program that imposes an group sparsity penalty for each component function. The DC stage fits decoupled concave functions on the residuals, for each component that is zeroed out in the AC stage.

1. *AC Stage*: Estimate an additive convex model

$$\widehat{f}_1, ..., \widehat{f}_p, \widehat{\mu} = \underset{f_1,...,f_p \in \mathcal{C}_0^1,\, \mu \in \mathbb{R}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \mu - \sum_{k=1}^{p} f_k(x_{ik}) \Big)^2 + \lambda \sum_{k=1}^{p} \|f_k\|_\infty. \tag{4.25}$$

2. *DC Stage*: For each k such that $\|\widehat{f}_k\|_\infty = 0$, estimate a concave function on the residual:

$$\widehat{g}_k = \underset{g_k \in \text{-}\mathcal{C}_0^1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \widehat{\mu} - \sum_{k'} \widehat{f}_{k'}(x_{ik'}) - g_k(x_{ik}) \Big)^2 + \lambda \|g_k\|_\infty. \tag{4.26}$$

3. Output as the set of relevant variables $\widehat{S} = \{k : \|\widehat{f}_k\|_\infty > 0 \text{ or } \|\widehat{g}_k\|_\infty > 0\}$.

For identifiability, we impose the constraint $\sum_{i=1}^n f_k(x_{ik}) = 0$ for each k. The set of one-dimensional convex (concave) functions satisfying this mean zero constraint is denoted $\mathcal{C}_0^1$ ($\text{-}\mathcal{C}_0^1$). We use an ℓ_∞/ℓ_1 penalty in equation (5.1) and an ℓ_∞ penalty in equation (5.2) to encourage sparsity. Other penalties can also produce sparse estimates, such as a penalty on the derivative of each of the component functions. The $\|\cdot\|_\infty$ norm is convenient for both theoretical analysis and implementation.

The optimization in (5.1) appears to be infinite dimensional, but it is equivalent to a finite dimensional quadratic program. In the following section, we give the details of this optimization, and show how it can be reformulated to be more computationally efficient.

5 Optimization

We now describe in detail the optimization algorithm for the additive convex regression stage. The second decoupled concave regression stage follows a very similar procedure.

Let $\boldsymbol{x}_i \in \mathbb{R}^p$ be the covariate, let y_i be the response and let ϵ_i be the mean zero noise. The regression function $f(\cdot)$ we estimate is the summation of functions $f_k(\cdot)$ in each variable dimension and a scalar offset μ. We impose an additional constraint that each $f_k(\cdot)$ is an univariate convex function, which can be represented by its supporting hyperplanes, i.e.,

$$f_{i'k} \geq f_{ik} + \beta_{ik}(x_{i'k} - x_{ik}) \quad \text{for all } i, i' = 1, \ldots, n, \tag{5.1}$$

where $f_{ik} := f_k(x_{ik})$ and β_{ik} is the subgradient at point x_{ik}. We apparently need $O(n^2p)$ constraints to impose the supporting hyperplane constraints. In fact, only $O(np)$ constraints suffice, since univariate convex functions are characterized by the condition that the subgradient, which is a scalar, must increase monotonically. This observation leads to the optimization

$$\begin{aligned}
\min_{f,\beta,\mu} \ & \frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \mu - \sum_{k=1}^{p} f_{ik}\Big)^2 + \lambda \sum_{k=1}^{p} \|f_k\|_\infty \\
\text{subject to } & \text{for all } k = 1, \ldots, p\text{:} \\
& f_{(i+1)_k k} = f_{(i)_k k} + \beta_{(i)_k k}(x_{(i+1)_k k} - x_{(i)_k k}), \text{ for } i = 1, \ldots, n-1 \\
& \sum_{i=1}^{n} f_{ik} = 0, \\
& \beta_{(i+1)_k k} \geq \beta_{(i)_k k} \text{ for } i = 1, \ldots, n-1.
\end{aligned} \tag{5.2}$$

Here f_k denotes the vector $f_k = (f_{1k}, f_{2_k} \ldots, f_{nk})^T \in \mathbb{R}^n$ and $\{(1)_k, (2)_k, \ldots, (n)_k\}$ is a reordering of $\{1, 2, \ldots, n\}$ such that

$$x_{(1)_k k} \leq x_{(2)_k k} \leq \cdots \leq x_{(n)_k k}. \tag{5.3}$$

We can solve for μ explicitly as $\mu = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}$. This follows from the KKT conditions and the constraints $\sum_i f_{ki} = 0$. It is easy to verify that the constraints in (6.2) satisfy the supporting hyperplane constraints, since for all $j > i$

$$f_{(j)_k k} - f_{(i)_k k} = \sum_{t=i}^{j-1}(f_{(t+1)_k k} - f_{(t)_k k}) \tag{5.4}$$

$$= \sum_{t=i}^{j-1} \beta_{(t)_k k}(x_{(t+1)_k k} - x_{(t)_k k}) \tag{5.5}$$

$$\geq \beta_{(i)_k k} \sum_{t=i}^{j-1}(x_{(t+1)_k k} - x_{(t)_k k}) \tag{5.6}$$

$$= \beta_{(i)_k k}(x_{(j)_k k} - x_{(i)_k k}) \tag{5.7}$$

and for all $j < i$

$$f_{k(j)} - f_{k(i)} = \sum_{t=j}^{i-1}(f_{k(t)} - f_{k(t+1)}) \tag{5.8}$$

$$= \sum_{t=j}^{i-1} \beta_{k(t)}(x_{k(t)} - x_{k(t+1)}) \tag{5.9}$$

$$\geq \beta_{k(i)} \sum_{t=j}^{i-1}(x_{k(t)} - x_{k(t+1)}) \tag{5.10}$$

$$= \beta_{k(i)}(x_{k(j)} - x_{k(i)}). \tag{5.11}$$

The sparse convex additive model optimization in (6.2) is a quadratic program with $O(np)$ variables and $O(np)$ constraints. Directly applying a QP solver for f and β would be computationally

expensive for relatively large n and p. However, notice that variables in different feature dimensions are only coupled in the term $(Y_i - \mu - \sum_{k=1}^{p} f_{ik})^2$. Hence, we can apply the block coordinate descent method, where in each step we solve the following QP subproblem for $\{f_k, \beta_k\}$ with the other variables fixed. In matrix notation, the optimization is

$$\begin{aligned} \min_{f_k, \beta_k, \gamma_k} \quad & \frac{1}{2n}\|r_k - f_k\|_2^2 + \lambda\gamma_k \\ \text{such that } & Pf_k = \operatorname{diag}(P\boldsymbol{x}_k)\beta_k \\ & D\beta_k \le 0 \\ & -\gamma_k \mathbf{1}_n \le f_k \le \gamma_k \mathbf{1}_n \\ & \mathbf{1}_n^{\mathsf{T}} f_k = 0 \end{aligned} \tag{5.12}$$

where $\beta_k \in \mathbb{R}^{n-1}$ is the vector $\beta_k = (\beta_{1k}, \ldots, \beta_{(n-1)k})^T$, and $r_k \in \mathbb{R}^n$ is the residual vector $r_k = (y_i - \widehat{\mu} - \sum_{k' \neq k} f_{ik'})^T$. In addition, $P \in \mathbb{R}^{(n-1)\times n}$ is a permutation matrix where the i-th row is all zero except -1 at in position $(i)_k$ and 1 in position $(i+1)_k$, and $D \in \mathbb{R}^{(n-2)\times(n-1)}$ is another permutation matrix where the i-th row is all zero except for a 1 at position i and -1 at position $i+1$. We denote by $\operatorname{diag}(v)$ the diagonal matrix whose diagonal entries are v.

The extra variable γ_k is introduced to deal with the ℓ_∞ norm. This QP subproblem involves $O(n)$ variables, $O(n)$ constraints and a sparse structure, which can be solved efficiently using optimization packages (e.g., MOSEK: `http://www.mosek.com/`). We cycle through all feature dimensions (k) from 1 to p multiple times until convergence. Empirically, we observe that the algorithm converges in only a few cycles. We also implemented an ADMM solver for (6.2), but found that it is not as efficient as this QP solver.

After optimization, the function estimator for any input data $\boldsymbol{x}_j$ is, according to (6.1),

$$f(\boldsymbol{x}_j) = \sum_{k=1}^{p} f_k(x_{kj}) + \mu = \sum_{k=1}^{p} \max_i \{f_{ki} + \beta_{ki}(x_{kj} - x_{ki})\} + \mu.$$

The univariate concave function estimation is a straightforward modification of optimization 6.3. We need only modify one linear inequality constraint to enforce that the subgradients must be non-increasing: $\beta_{k(i+1)} \le \beta_{k(i)}$.

5.1 Alternative Formulation

Optimization (6.2) can be reformulated in terms of the 2nd derivatives, a form which we analyze in our theoretical analysis. The alternative formulation replaces the ordering constraints $\beta_{k(i+1)} \ge \beta_{k(i)}$ with positivity constraints, which simplifies theoretical analysis. Define $d_{k(i)}$ as the second derivative: $d_{k(1)} = \beta_{k(1)}$, and $d_{k(2)} = \beta_{k(2)} - \beta_{k(1)}$. The convexity constraint is equivalent to the constraint that $d_{k(i)} \ge 0$ for all $i > 1$.

It is easy to verify that $\beta_{k(i)} = \sum_{j \le i} d_{k(j)}$ and

$$\begin{aligned} f_k(x_{k(i)}) =& f_k(x_{k(i-1)}) + \beta_{k(i-1)}(x_{k(i)} - x_{k(i-1)}) \\ =& f_k(x_{k1}) + \sum_{j<i} \beta_{k(j)}(x_{k(j)} - x_{k(j-1)}) \\ =& f_k(x_{k1}) + \sum_{j<i} \sum_{j' \le j} d_{k(j')}(x_{k(j)} - x_{k(j-1)}) \\ =& f_k(x_{k1}) + \sum_{j'<i} d_{k(j')} \sum_{i>j\ge j'} (x_{k(j)} - x_{k(j-1)}) \\ =& f_k(x_{k1}) + \sum_{j'<i} d_{k(j')}(x_{k(i)} - x_{k(j')}) \end{aligned}$$

We can write this more compactly in matrix notations.

$$\begin{bmatrix} f_k(x_{k1}) \\ ... \\ f_k(x_{kn}) \end{bmatrix} = \begin{bmatrix} |x_{k1} - x_{k(1)}|_+ & ... & |x_{k1} - x_{k(n-1)}|_+ \\ ... & & \\ |x_{kn} - x_{k(1)}|_+ & ... & |x_{kn} - x_{k(n-1)}|_+ \end{bmatrix} \begin{bmatrix} d_{k(1)} \\ ... \\ d_{k(n-1)} \end{bmatrix} + \mu_k \equiv \Delta_k d_k + \mu_k$$

Where Δ_k is a $n \times n-1$ matrix such that $\Delta_k(i,j) = |x_{ki} - x_{k(j)}|_+$, $d_k = (d_{k(1)}, ..., d_{k(n-1)})$, and $\mu_k = f_k(x_{k1})$.

Because f_k has to be centered, $\mu_k = -\frac{1}{n}\mathbf{1}_n^\mathsf{T}\Delta_k d_k$, therefore:

$$\Delta_k d_k + \mu_k \mathbf{1}_n = \Delta_k d_k - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\Delta_k d_k = \overline{\Delta}_k d_k$$

where $\overline{\Delta}_k \equiv \Delta_k - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\Delta_k$ is Δ_k with the mean of each column subtracted.

We can now reformulate (6.2) as an equivalent optimization program with only centering and positivity constraints:

$$\begin{aligned} &\min_{d_k} \frac{1}{2n}\Big\|Y - \sum_{k=1}^p \overline{\Delta}_k d_k\Big\|_2^2 + \lambda_n \sum_{k=1}^p \|\overline{\Delta}_k d_k\|_\infty \\ &\text{s.t. } d_{k(2)}, \ldots, d_{k(n-1)} \geq 0 \qquad \text{(convexity)} \end{aligned} \tag{5.13}$$

The decoupled concave postprocessing stage optimization is again similar. Suppose $\widehat{d}_k$'s are the output of optimization 6.4, define $\widehat{r} = Y - \sum_{k=1}^p \overline{\Delta}_k \widehat{d}_k$.

$$\begin{aligned} &\text{for all } k \text{ such that } \widehat{d}_k = 0\text{:} \\ &\min_{c_k} \frac{1}{2n}\Big\|\widehat{r} - \Delta_k c_k\Big\|_2^2 + \lambda_n \|\Delta_k c_k\|_\infty \\ &\text{s.t.} c_{k(2)}, \ldots, c_{k(n-1)} \leq 0 \qquad \text{(concavity)} \end{aligned} \tag{5.14}$$

We can use either the off-centered Δ_k matrix or the centered $\overline{\Delta}_k$ matrix because the concave estimations are decoupled and hence suffer no identifiability problems.

Remark 5.1. in sparsistency analysis, we assume that L, an upper bound to the coordinate-wise Lipschitz smoothness of the true function f_0, is known and that we constrain our estimate $\widehat{f}$ to obey the same Lipschitz condition, that is, each $\widehat{f}_k$ must be L-Lipschitz. This Lipschitz constraint can be easily added to our optimization program. In optimization 6.4, we can enforce the Lipschitz condition by adding two constraints: $d_{k(1)} \geq -L$ and $\sum_{j=2}^{n-1} d_{k(j)} \leq L$ (similarly for optimization 6.5). We emphasize that we use the Lipschitz constraint only in our theoretical analysis; all of our experiments do not impose any Lipschitz condition.

6 Analysis of Variable Selection Consistency

We divide our analysis into two parts. We first establish a sufficient deterministic condition for sparsistency. We then consider the stochastic setting and argue that the deterministic conditions hold with high probability. In all of our results and analysis, we let c, C represent absolute constants; the actual values of c, C could change from line to line.

6.1 Deterministic Setting

We construct an additive convex solution $\{\widehat{d}_k\}_{k=1,\ldots,p}$ that is zero for $k \in S^c$ and show that it satisfies the KKT conditions for optimality of optimization 6.4. We define $\widehat{d}_k$ for $k \in S$ to be a solution to the restricted regression (defined below). We will then also show that $\widehat{c}_k = 0$ satisfies the optimality condition of optimization 6.5 for all $k \in S^c$.

Definition 6.1. We define as *restricted regression* where we restrict the indices k in optimization (6.4) to lie in the set S instead of ranging from $1, ..., p$:

$$\min_{d_k} \frac{1}{n} \Big\| Y - \sum_{k \in S} \overline{\Delta}_k d_k \Big\|_2^2 + \lambda_n \sum_{k \in S} \|\overline{\Delta}_k d_k\|_\infty \quad \text{such that } d_{k,1}, ..., d_{k,n-1} \geq 0$$

Theorem 6.1. *(Deterministic setting) Let $\{\widehat{d}_k\}_{k \in S}$ be a minimizer of the restricted regression as defined above. Let $\widehat{r} := Y - \sum_{k \in S} \overline{\Delta}_k \widehat{d}_k$ be the restricted regression residual.*

Suppose for all $k \in S^c$, for all $i = 1, ..., n$, $\lambda_n > |\frac{1}{2n} \widehat{r}^\mathsf{T} \mathbf{1}_{(i:n)}|$ where $\mathbf{1}_{(i:n)}$ is 1 on the coordinates of the i-th largest to the n-th largest entries of X_k and 0 elsewhere.

Then the following are true:

1. *Let $\widehat{d}_k = 0$ for $k \in S^c$, then $\{\widehat{d}_k\}_{k=1,...,p}$ is an optimal solution to optimization 6.4. Furthermore, any solution to the optimization program 6.4 must be zero on S^c.*
2. *For all $k \in S^c$, the solution $\widehat{c}_k$ to optimization 6.5 must be 0 and must be unique.*

This result holds regardless of whether we impose the Lipschitz conditions in optimization 6.4 and 6.5 or not. The full proof of Theorem 7.1 is in Section 10.1 of the Appendix.

Theorem 7.1 allows us to analyze false negative rates and false positive rates separately. To control false positives, we analyze when the condition on λ_n is fulfilled for all j and all $k \in S^c$. To control false negatives, we analyze the restricted regression.

The proof of theorem 7.1 looks at KKT conditions of optimization 6.4, similar to the now standard *primal-dual witness* technique Wainwright [2009]. We cannot derive analogous *mutual incoherence* conditions because the estimation is nonparametric – even the low dimensional restricted regression has $s(n-1)$ variables. The details of the proof are in section 10.1 of the Appendix.

6.2 Probabilistic Setting

We use the following statistical setting:

1. Let F be a distribution supported and positive on $\mathcal{X} = [-b, b]^p$. Let $X^{(1)}, ..., X^{(n)} \sim F$ be iid.
2. Let $Y = f_0(X) + w$ where w is zero-mean noise. Let $Y^{(1)}, ..., Y^{(n)}$ be iid.
3. Let $S_0 = \{1, ..., s_0\}$ denote the relevant variables where $s_0 \leq p$, i.e., $f_0(X) = f_0(X_{S_0})$.

Let $\mathcal{C}^1$ denote the set of univariate convex functions supported on $[-b, b]$. Define $\mathcal{C}^p$ as the set of convex additive function $\mathcal{C}^p \equiv \{f : f = \sum_{k=1}^p f_k,\ f_k \in \mathcal{C}^1\}$ Let $f^*(x) = \sum_{k=1}^p f_k^*(x_k)$ be the population risk minimizer: $f^* \equiv \arg\min_{f \in \mathcal{C}^p} \mathbb{E}\big(f_0(X) - f^*(X)\big)^2$.

Similarly, we define -$\mathcal{C}^1$ as the set of univariate concave functions supported on $[-b, b]$ and let $g_k^* = \arg\min_{g_k \in \text{-}\mathcal{C}^1} \mathbb{E}\big(f_0(X) - f^*(X) - g_k(X_k)\big)^2$.

We let $S = \{k = 1, ..., p : f_k^* \neq 0 \text{ or } g_k^* \neq 0\}$. By additive faithfulness (theorem 4.2), it must be that $S_0 \subset S$.

Each of our theorems will use a subset of the following assumptions:

A1: X_S, X_{S^c} are independent.

A2: f_0 is convex and twice-differentiable.

A3: $|\partial_{x_k} f_0| \leq L$ for all k

A4: w is mean-zero subgaussian, independent of X, with subgaussian scale σ, i.e. for all $t \in \mathbb{R}$, $\mathbb{E}e^{t\epsilon} \leq e^{\sigma^2 t^2/2}$.

By assumption A1, f_k^* is must be zero for $k \notin \{1, ..., s\}$.

We define α_f, α_g as a measure of the signal strength of the weakest variable:

$$\alpha_f = \inf_{f \in \mathcal{C}^p \,:\, \exists k,\, f_k^* \neq 0 \,\wedge\, f_k = 0} \Big\{ \mathbb{E}\big(f_0(X) - f(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X)\big)^2 \Big\}$$

$$\alpha_g = \min_{k \in S \,:\, g_k^* \neq 0} \Big\{ \mathbb{E}\big(f_0(X) - f^*(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X) - g_k^*(X_k)\big)^2 \Big\}$$

Intuitively, if α_f is smaller, then it is easier to make a false omission in the additive convex stage of the procedure. If α_g is smaller, then it is easier to make a false omission in the decoupled concave stage of the procedure.

Remark 6.1. We make strong assumptions on the covariates in A1 in order to make very weak assumptions on the true regression function f_0 in A2. In particular, we do not assume that f_0 is additive.

Theorem 6.2. *(Controlling false positives)*
Suppose assumptions A1-A4 hold.

Suppose $\lambda_n \geq csLb\sigma\sqrt{\frac{1}{n}\log^2 np}$*, then with probability at least* $1 - \frac{C}{n}$*, for all* $k \in S^c$*, and for all* $i' = 1, ..., n$*:*

$$\lambda_n > \Big| \frac{1}{2n} \widehat{r}^{\mathsf{T}} \mathbf{1}_{(i':n)} \Big|$$

And thereforefor all $k \in S^c$*, both the AC solution* $\widehat{f}_k$*, from optimization 6.4, and the DC solution* $\widehat{g}_k$*, from optimization 6.5 are zero.*

The proof of Theorem 7.2 exploits independence of $\widehat{r}$ and X_k from A1; when $\widehat{r}$ and X_k are independent, $\widehat{r}^{\mathsf{T}} \mathbf{1}_{(i':n)}$ is the sum of $n - i' + 1$ random coordinates of $\widehat{r}$. We can then uses the concentration of measure result for sampling without replacement to argue that $|\frac{1}{n}\widehat{r}^{\mathsf{T}} \mathbf{1}_{(i':n)}|$ is small with high probability. The full proof of Theorem 7.2 is in Section 10.2 of the Appendix.

Theorem 6.3. *(Controlling false negatives)*
Suppose assumptions A1-A4 hold. Let $\widehat{f}$ *be any AC solution to the restricted regression with L-Lipschitz constraint and let* $\widehat{g}_k$*'s be any DC solution to the restricted regression with L-Lipschitz constraint.*

Suppose $\lambda \leq csLb\sqrt{\frac{1}{n}\log^2 np}$ *and suppose* n *is large enough such that* $(Lb)^3\sigma\sqrt{\frac{s^5}{n^{4/5}}\log^2 np} \leq 1$*.*

Suppose $\frac{\alpha_f}{\sigma} \geq c(Lb)^3\sqrt{\frac{s^5}{n^{4/5}}\log^2 np}$ *and* $\frac{\alpha_g^2}{\sigma} \geq c(Lb)^3\sqrt{\frac{s^5}{n^{4/5}}\log^2 2np}$*.*

Then, with probability at least $1 - \frac{1}{n}$*, we have that for all* $k \in S$*,* $\widehat{f}_k \neq 0$ *or* $\widehat{g}_k \neq 0$*.*

This is a finite sample version of Theorem 4.1. We need stronger assumptions in Theorem 7.3 to use our additive faithfulness result, Theorem 4.1. We also include an extra Lipschitz constraint so that we can use existing covering number results Bronshtein [1976]. Recent work Guntuboyina and Sen [2013] shows that the Lipschitz constraint is not required with more advanced empirical process theory techniques; we leave the incorporation of this development as future work. We give the full proof of Theorem 7.3 in Section 10.3 of the Appendix.

Combining Theorem 7.2 and 7.3 we have the following result.

Corollary 6.1. *Suppose assumptions A1-A4 hold. Let* $\widehat{f}$ *be any AC solution and* $\widehat{g}_k$*'s be any DC solution, both with L-Lipschitz constraints, both with* $\lambda = \Theta\Big(sLb\sqrt{\frac{1}{n}\log^2 np}\Big)$*.*

Suppose $\frac{\alpha_f}{\sigma} \geq c(Lb)^3\sqrt{\frac{s^5}{n^{4/5}}\log^2 np}$ *and* $\frac{\alpha_g^2}{\sigma} \geq c(Lb)^3\sqrt{\frac{s^5}{n^{4/5}}\log^2 2np}$.

Then, for all large enough n*, we have that with probability at least* $1 - \frac{1}{n}$*:*

$$\widehat{f}_k \neq 0 \text{ or } \widehat{g}_k \neq 0 \text{ for all } k \in S$$
$$\widehat{f}_k = 0 \text{ and } \widehat{g}_k = 0 \text{ for all } k \notin S$$

The above corollary implies that sparsistency is achievable at the same exponential scaling of the ambient dimension $p = O(\exp(n^c)), c < 1$ rate as parametric models. The cost of nonparametric modeling is reflected in the scaling with respect to s, which can only scale at $o(n^{4/25})$.

Remark 6.2. Comminges and Dalalyan [2012] have shown that under tradtional smoothness constraints, variable selection is achievable only if $n > O(e^s)$. It is interesting to observe that because of additive faithfulness, the convexity assumption enables a much better scaling of $n = O(\text{poly}(s))$, demonstrating that geometric constraints can be quite different from the previously studied smoothness conditions.

7 Experiments

We first illustrate our methods using a simulation of the following regression problem

$$y_i = \boldsymbol{x}_{iS}^\top \boldsymbol{Q} \boldsymbol{x}_{iS} + \epsilon_i \quad (i = 1, 2, \ldots, n).$$

Here $\boldsymbol{x}_i$ denotes data sample i drawn from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, $\boldsymbol{x}_{iS}$ is a subset of $\boldsymbol{x}_i$ with dimension $|S| = 5$, where S represents the active feature set, and ϵ_i is the additive noise drawn from $\mathcal{N}(0, 1)$. $\boldsymbol{Q}$ is a symmetric positive definite matrix of dimension $|S| \times |S|$. Notice that if $\boldsymbol{Q}$ is diagonal, then the true function is convex additive; otherwise the true function is convex but not additive. For all the simulations in this section, we set $\lambda = 4\sqrt{\log(np)/n}$.

In the first simulation, we set $\boldsymbol{Q} = \boldsymbol{I}_{|S|}$ (the additive case), and choose $n = 100, 200, \ldots, 1000$ and $p = 64, 128, 256, 512$. For each (n, p) combination, we generate 200 independent data sets. For each data set we use SCAM to infer the model parameterized by $\boldsymbol{h}$ and $\boldsymbol{\beta}$; see equation (6.2). If $\|\boldsymbol{\beta}_{k\cdot}\|_\infty < 10^{-8}$ $(\forall k \notin S)$ and $\|\boldsymbol{\beta}_{k\cdot}\|_\infty > 10^{-8}$ $(\forall k \in S)$, then we declare correct support recovery. We then plot the probability of support recovery over the 200 data sets in Figure 2(a). We observe that SCAM performs consistent variable selection when the true function is convex additive. To give the reader a sense of the running speed, the code runs in about 2 minutes on one data set with $n = 1000$ and $p = 512$, on a MacBook with 2.3 GHz Intel Core i5 CPU and 4 GB memory.

In the second simulation, we study the case in which the true function is convex but not additive. We generate four $\boldsymbol{Q}$ matrices plotted in Figure 2(b), where the diagonal elements are all 1 and the off-diagonal elements are 0.5 with probability α ($\alpha = 0, 0.2, 0.5, 1$ for the four cases). We fix $p = 128$ and choose $n = 100, 200, \ldots, 1000$. We again run the SCAM optimization on 200 independently generated data sets and plot the probability of recovery in Figure 2(c). The results demonstrate that SCAM performs consistent variable selection even if the true function is not additive (but still convex).

In the third simulation, we study the case of correlated design, where $\boldsymbol{x}_i$ is drawn from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ instead of $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, with $\Sigma_{ij} = \nu^{|i-j|}$. We use the non-additive $\boldsymbol{Q}$ with $\alpha = 0.5$ and fix $p = 128$. The recovery curves for $\nu = 02, 0.4, 0.6, 0.8$ are depicted in Figure 2(d). As can be seen, for design of moderate correlation, SCAM can still select relevant variables well.

We next use the Boston housing data rather than simulated data. This data set contains 13 covariates, 506 samples and one response variable indicating housing values in suburbs of Boston. The data and detailed description can be found on the UCI Machine Learning Repository website [1].

We first use all $n = 506$ samples (with normalization) to train SCAM, using a set of candidate $\{\lambda^{(t)}\}$ with $\lambda^{(1)} = 0$ (no regularization). For each $\lambda^{(t)}$ we obtain a subgradient matrix $\boldsymbol{\beta}^{(t)}$ with

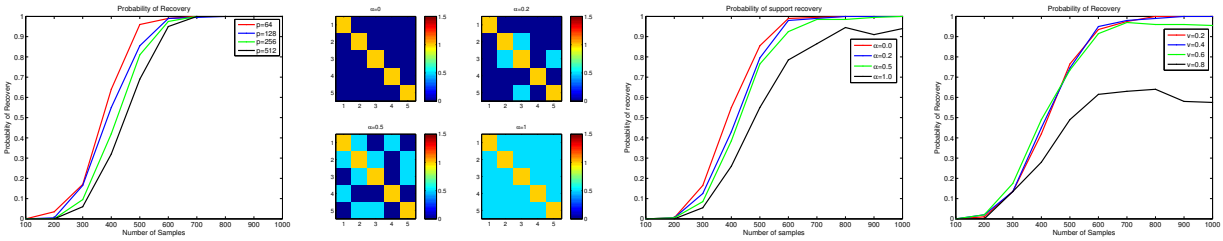
[1] `http://archive.ics.uci.edu/ml/datasets/Housing`

$p = 13$ rows. The non-zero rows in this matrix indicate the variables selected using $\lambda^{(t)}$. We plot $\|\boldsymbol{\beta}^{(t)}\|_\infty$ and the row-wise mean of $\boldsymbol{\beta}^{(t)}$ versus the normalized norm $\frac{\|\boldsymbol{\beta}^{(t)}\|_{\infty,1}}{\|\boldsymbol{\beta}^{(1)}\|_{\infty,1}}$ in Figures 3(a) and 3(b). As a comparison we plot the LASSO/LARS result in a similar way in Figure 3(c). From the figures we observe that the first three variables selected by SCAM and LASSO are the same: LSTAT, RM and PTRATIO, which is consistent with previous findings Ravikumar et al. [2007]. The fourth variable selected by SCAM is TAX (with $\lambda^{(t)} = 0.09$). We then refit SCAM with only these four variables without regularization, and plot the inferred additive functions in Figure 3(e). As can be seen, these functions contain clear nonlinear effects which cannot be captured by LASSO. The shapes of these functions are in agreement with those obtained by SpAM Ravikumar et al. [2007].

Next, in order to quantitatively study the predictive performance, we run 10 times 5-fold cross validation, following the same procedure described above (training, variable selection and refitting). A plot of the mean and standard deviation of the predictive Mean Squared Error (MSE) in in Figure 3(d). Since for SCAM the same $\lambda^{(t)}$ may lead to slightly different number of selected features in different folds and runs, the values on the x-axis (average number of selected features) for SCAM are not necessarily integers. Nevertheless, the figure clearly shows that SCAM has a much lower predictive MSE than LASSO. We also compared the performance of SCAM with that of Additive Forward Regression (AFR) presented in Liu and Chen [2009], and found that they are similar. The main advantages of SCAM compared with AFR and SpAM are 1) there are no other tuning parameters (such as bandwidth) besides λ; 2) SCAM is formulated as a convex program, which guarantees a global optimum.

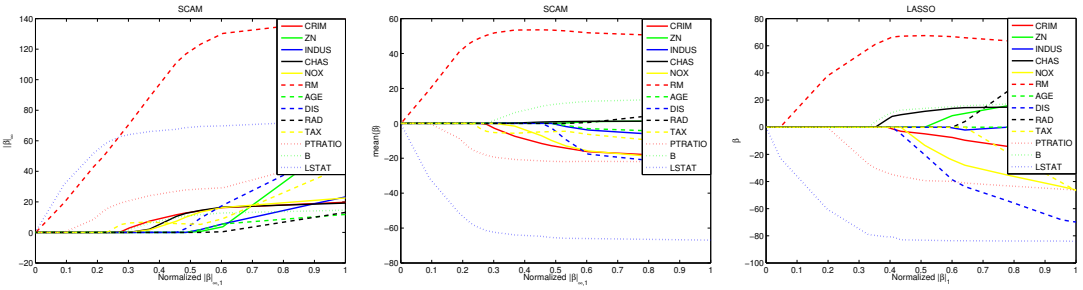
8 Discussion

We have introduced a framework for estimating high dimensional but sparse convex functions. Because of the special properties of convexity, variable selection for convex functions enjoys additive faithfulness—it suffices to carry out variable selection over an additive model, in spite of the approximation error this introduces. Sparse convex additive models can be optimized using block coordinate quadratic programming, which we have found to be effective and scalable. We established variable selection consistency results, allowing exponential scaling in the ambient dimension. We expect that the technical assumptions we have used in these analyses can be weakened; this is one direction for future work. Another interesting direction for building on this work is to allow for additive models that are a combination of convex and concave components. If the convexity/concavity of each component function is known, this again yields a convex program. The challenge is to develop a method to automatically detect the concavity or convexity pattern of the variables.

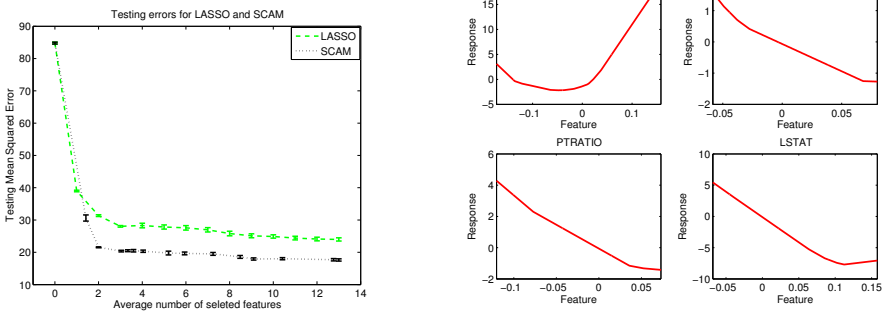


(a) additive model (b) four $\boldsymbol{Q}$ matrices (c) non-additive models (d) correlated design

Figure 3: Support recovery results where the additive assumption is correct (a), incorrect (b), (c), and with correlated design (d).



SCAM $\|\beta_k\|_\infty$ paths SCAM mean(β_k) paths LASSO β_k paths



predictive MSE inferred functions from SCAM

Figure 4: Results on Boston housing data, showing regularization paths, MSE and fitted functions.

References

Karine Bertin and Guillaume Lecué. Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics*, 2:1224–1241, 2008.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

E. M. Bronshtein. ε-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17: 393–398, 1976.

H. Chen and D. D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer-Verlag, 2001.

Laëtitia Comminges and Arnak S. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40(5):2667–2696, 2012.

Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Approximation of functions of few variables in high dimensions. *Constructive Approximation*, 33:125–143, 2011.

A. Goldenshluger and A. Zeevi. Recovering convex boundaries from blurred and noisy observations. *Ann. Statist*, 34:1375–1394, 2006.

A. Guntuboyina. Optimal rates of convergence for convex set estimation from support functions. *Annals of Statistics*, 40:385–411, 2012.

A. Guntuboyina and B. Sen. Covering numbers for convex functions. *IEEE Trans. Info. Theory*, 59:1957–1965, 2013.

L. A. Hannah and D. B. Dunson. Ensemble methods for convex regression with applications to geometric programming based circuit design. In *International Conference on Machine Learning (ICML)*, 2012.

R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press; Reprint edition, 1990.

Arlene KH Kim and Richard J Samworth. Global rates of convergence in log-concave density estimation. *arXiv preprint arXiv:1404.2298*, 2014.

Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 2010.

John Lafferty and Larry Wasserman. Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1):28–63, 2008.

A. S. Lele, S. R. Kulkarni, and A. S. Willsky. Convex-polygon estimation from support-line measurements and applications to target reconstruction from laser-radar data. *Journal of the Optical Society of America, Series A*, 9:1693–1714, 1992.

Eunji Lim and Peter W. Glynn. Consistency of multidimensional convex regression. *Operations Research*, 60(1):196–208, 2012.

H. Liu and X. Chen. Nonparametric greedy algorithm for the sparse learning problems. In *Advances in Neural Information Processing Systems*, 2009.

R. F. Meyer and J. W. Pratt. The consistent assessment and fairing of preference functions. *IEEE Trans. Systems Sci. Cybernetics*, 4(3):270–278, 1968.

E. Mossel, R. O'Donnell, and R. Servedio. Learning functions of k relevant variables. *Journal of Computer and System Sciences*, 69(3):421–434, 2004.

J. L. Prince and A. S. Willsky. Reconstructing convex sets from support line measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:377–389, 1990.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13:389–427, 2012.

P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Spam: Sparse additive models. In *Advances in Neural Information Processing Systems*, 2007.

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

Robert J Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48, 1974.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using¡ formula formulatype=. *Information Theory, IEEE Transactions on*, 55(5):2183–2202, 2009.

9 Appendix

9.1 Proof of the Deterministic Condition for Sparsistency

We restate Theorem 7.1 first for convenience.

Theorem 9.1. *The following holds regardless of whether we impose the Lipschitz condition in optimization 6.4 and optimization 6.5.*

Let $\{\widehat{d}_k\}_{k\in S}$ be a minimizer of the restricted regression, that is, the solution to optimization (6.4) where we restrict $k \in S$. Let $\widehat{r} := Y - \sum_{k\in S} \overline{\Delta}_k \widehat{d}_k$ be the restricted regression residual.

Suppose for all $k \in S^c$, for all $i = 1, ..., n$, $\lambda_n > |\frac{1}{2n}\widehat{r}^\mathsf{T}\mathbf{1}_{(i:n)}|$ where $\mathbf{1}_{(i:n)}$ is 1 on the coordinates of the i-th largest to the n-th largest entries of X_k and 0 elsewhere.

Then the following are true:

1. *Let $\widehat{d}_k = 0$ for $k \in S^c$, then $\{\widehat{d}_k\}_{k=1,...,p}$ is an optimal solution to optimization 6.4. Furthermore, any solution to the optimization program 6.4 must be zero on S^c.*
2. *For all $k \in S^c$, the solution to optimization 6.5 must be 0 and must be unique.*

Proof. We will omit the Lipschitz constraint in our proof here. It is easy to add that in and check that the result of the theorem still holds.

We first consider the first item in the conclusion of the theorem.

We will show that with $\{\widehat{d}_k\}_{k=1,..,p}$ as constructed, we can set the dual variables to satisfy complementary slackness and stationary conditions: $\nabla_{d_k}\mathcal{L}(\widehat{d}) = 0$ for all k.

The Lagrangian is

$$\mathcal{L}(\{d_k\}, \nu) = \frac{1}{2n}\Big\|Y - \sum_{k=1}^{p} \overline{\Delta}_k d_k\Big\|_2^2 + \lambda \sum_{k=1}^{p} \|\overline{\Delta}_k d_k\|_\infty - \sum_{k=1}^{p}\sum_{i=2}^{n-1} \nu_{ki} d_{ki} \tag{9.1}$$

with the constraint that $\nu_{ki} \geq 0$ for all k, i.

Because $\{\widehat{d}_k\}_{k\in S}$ is by definition the optimal solution of the restricted regression, it is a consequence that stationarity holds for $k \in S$, that is, $\partial_{\{d_k\}_{k\in S}}\mathcal{L}(d) = 0$, and that the dual variables ν_k for $k \in S$ satisfy complementary slackness.

We now verify that stationarity holds also for $k \in S^c$. We fix one dimension $k \in S^c$ and let $\widehat{r} = Y - \sum_{k'\in S} \overline{\Delta}_{k'}\widehat{d}_{k'}$.

The Lagrangian form of the optimization, in term of just d_k, is

$$\mathcal{L}(d_k, \nu_k) = \frac{1}{2n}\big\|Y - \sum_{k'\in S} \overline{\Delta}_{k'} d_{k'} - \overline{\Delta}_k d_k\big\|_2^2 + \lambda\|\overline{\Delta}_k d_k\|_\infty - \sum_{i=2}^{n-1} \nu_{ki} d_{ki}$$

with the constraint that $v_i \geq 0$ for all i.

The derivative of the Lagrangian is:

$$\partial_{d_k}\mathcal{L}(d_k) = -\frac{1}{n}\overline{\Delta}_k^\mathsf{T}(Y - \sum_{k'\in S} \overline{\Delta}_{k'} d_{k'} - \overline{\Delta}_k d_k) + \lambda\overline{\Delta}_k^\mathsf{T}\boldsymbol{u} - \nu_k$$

where $\boldsymbol{u}$ is the subgradient of $\|\overline{\Delta}_k d_k\|_\infty$, an n-vector such that $\|\boldsymbol{u}_T\|_1 = 1$ where $T = \{i : (\overline{\Delta}_k d_k)_i = \|\overline{\Delta}_k d_k\|_\infty\}$.

We now substitute in $d_{k'} = \widehat{d}_{k'}$ for $k' \in S$, $d_k = 0$ for $k \in S$, and $r = \widehat{r}$ and show that the duals can be set in a way to ensure that the derivatives are equal to 0.

$$\partial_{d_k}\mathcal{L}(\widehat{d}_k) = -\frac{1}{n}\overline{\Delta}_k^{\mathsf{T}}\widehat{r} + \lambda\overline{\Delta}_k^{\mathsf{T}}\boldsymbol{u} - \nu_k = 0$$

where $\|\boldsymbol{u}\| \leq 1$ and $\nu_k \geq 0$. It clear that to show stationarity, we only need to show that $-\frac{1}{n}\overline{\Delta}_k^{\mathsf{T}}\widehat{r} + \lambda\overline{\Delta}_k^{\mathsf{T}}\boldsymbol{u} \geq 0$ hwere the inequality is element-wise.

Let us reorder the samples so that the i-th sample is the i-smallest sample.

We will construct $\gamma = 0$, and $\boldsymbol{u} = (-a, 0, ..., a)$ for some $0 < a < 1/2$. (coordinates of $\boldsymbol{u}$ correspond to the new sample ordering) We then just need to show that

$$\begin{aligned}
-\frac{1}{n}\overline{\Delta}_k^{\mathsf{T}}\widehat{r} + \lambda\overline{\Delta}_k^{\mathsf{T}}\boldsymbol{u} \geq 0 \quad &\Leftrightarrow \\
-\frac{1}{n}\Delta_k^{\mathsf{T}}\widehat{r} + \lambda\Delta_k^{\mathsf{T}}\boldsymbol{u} \geq 0 \quad &\Leftrightarrow \\
-\frac{1}{n}\sum_{i>j}(X_{ki} - X_{kj})\widehat{r}_i + \lambda(X_{kn} - X_{kj})a \geq 0 \quad &\text{for each } j \\
-\frac{1}{n}\sum_{i>j}\sum_{j<i'\leq i}\mathsf{gap}_{i'}\widehat{r}_i + \lambda(X_{kn} - X_{kj})a \geq 0 \quad & \\
-\frac{1}{n}\sum_{i'>j}\mathsf{gap}_{i'}\sum_{i\geq i'}\widehat{r}_i + \lambda(X_{kn} - X_{kj})a \geq 0 \quad & \\
-\frac{1}{n}\sum_{i'>j}\mathsf{gap}_{i'}\mathbf{1}_{(i':n)}^{\mathsf{T}}\widehat{r} + \lambda(X_{kn} - X_{kj})a \geq 0 \quad &
\end{aligned}$$

where $\mathsf{gap}_i = X_{ki} - X_{k,i-1}$. If $\frac{1}{2n}|\mathbf{1}_{(i:n)}^{\mathsf{T}}\widehat{r}| \leq \lambda a$ for all $i = 1, ..., n$, then we have that:

$$\begin{aligned}
-\frac{1}{n}\sum_{i'>j}\mathsf{gap}_{i'}\mathbf{1}_{i':n}^{\mathsf{T}}\widehat{r} + \lambda(X_{kn} - X_{kj})a \geq 0 \quad &\Leftrightarrow \\
-\sum_{i'>j}\mathsf{gap}_{i'}\lambda a + \lambda(X_{kn} - X_{kj})a \geq 0 \quad & \\
-(X_{kn} - X_{kj})\lambda a + \lambda(X_{kn} - X_{kj})a \geq 0 \quad &
\end{aligned}$$

We have thus proven that there exist one solution $\{\widehat{d}_k\}_{k=1,...,p}$ such that $\widehat{d}_k = 0$ for all $k \in S^c$. Furthermore, we have shown that the subgradient variables $\boldsymbol{u}_k$ of the solution $\{\widehat{d}_k\}$ can be chosen such that $\|\boldsymbol{u}_k\|_1 < 1$ for all $k \in S^c$. We now prove that if $\{\widehat{d}'_k\}_{k=1,...,p}$ is another solution, then it must be that $\widehat{d}'_k = 0$ for all $k \in S^c$ as well.

We first claim that $\sum_{k=1}^p \overline{\Delta}_k\widehat{d}_k = \sum_{k=1}^p \overline{\Delta}_k\widehat{d}'_k$. If this were not true, then a convex combination of $\widehat{d}_k, \widehat{d}'_k$ would achieve a strictly lower objective on the quadratic term. More precisely, let $\zeta \in [0, 1]$. If $\sum_{k=1}^p \overline{\Delta}_k\widehat{d}'_k \neq \sum_{k=1}^p \overline{\Delta}_k\widehat{d}_k$, then $\|Y - \sum_{k=1}^p \overline{\Delta}_k\big(\widehat{d}_k + \zeta(\widehat{d}'_k - \widehat{d}_k)\big)\|_2^2$ is strongly convex as a function of ν. Thus, it cannot be that $\widehat{d}_k$ and $\widehat{d}'_k$ both achieve optimal objective and we have reached a contradiction.

Now, we look at the stationarity condition for both $\{\widehat{d}_k\}$ and $\{\widehat{d}'_k\}$. Let $\boldsymbol{u}_k \in \partial\|\overline{\Delta}_k\widehat{d}_k\|_\infty$ and let $\boldsymbol{u}'_k \in \partial\|\overline{\Delta}_k\widehat{d}'_k\|_\infty$ be the two sets of subgradients. Let $\{\nu_{ki}\}_{k=1,..,p,\, i=1,...n-1}$ and $\{\nu'_{ki}\}$ be the two sets of positivity dual variables. [2]

[2] since there is no positivity constraint on d_{k1}, we let $\nu_{k1} = 0$ always.

Let us define $\overline{\Delta}$, a $n \times p(n-1)$ matrix, to denote the column-wise concatenation of $\{\overline{\Delta}_k\}_k$ and $\widehat{d}$, a $p(n-1)$ dimensional vector, to denote the concatenation of $\{\widehat{d}_k\}_k$. With this notation, we can express $\sum_{k=1}^{p} \overline{\Delta}_k \widehat{d}_k = \overline{\Delta}\widehat{d}$.

Since both solutions $(\widehat{d}, \boldsymbol{u}, \nu)$ and $(\widehat{d'}, \boldsymbol{u'}, \nu')$ must satisfy the stationarity condition, we have that:

$$\overline{\Delta}^{\mathsf{T}}(Y - \overline{\Delta}\widehat{d}) + \lambda \sum_{k=1}^{p} \overline{\Delta}_k^{\mathsf{T}} \boldsymbol{u}_k - \nu = \overline{\Delta}^{\mathsf{T}}(Y - \overline{\Delta}\widehat{d'}) + \lambda \sum_{k=1}^{p} \overline{\Delta}_k^{\mathsf{T}} \boldsymbol{u}'_k - \nu' = 0$$

We multiply both sides of the above equation by $\widehat{d'}$:

$$\widehat{d'}^{\mathsf{T}}\overline{\Delta}^{\mathsf{T}}(Y - \overline{\Delta}\widehat{d}) + \lambda \sum_{k=1}^{p} \widehat{d'}_k^{\mathsf{T}}\overline{\Delta}_k^{\mathsf{T}} \boldsymbol{u}_k - \widehat{d'}^{\mathsf{T}}\nu = \widehat{d'}^{\mathsf{T}}\overline{\Delta}^{\mathsf{T}}(Y - \overline{\Delta}\widehat{d'}) + \lambda \sum_{k=1}^{p} \widehat{d'}_k^{\mathsf{T}}\overline{\Delta}_k^{\mathsf{T}} \boldsymbol{u}'_k - \widehat{d'}^{\mathsf{T}}\nu'$$

Since $\overline{\Delta}\widehat{d}_k = \overline{\Delta}\widehat{d}$, $\widehat{d'}^{\mathsf{T}}\nu' = 0$ (complementary slackness), and $\widehat{d'}_k^{\mathsf{T}}\overline{\Delta}_k^{\mathsf{T}}\boldsymbol{u}'_k = \|\widehat{f'}_k\|_\infty$ (where $\widehat{f'}_k = \overline{\Delta}_k\widehat{d'}_k$), we have that:

$$\lambda \sum_{k=1}^{p} \widehat{d'}_k^{\mathsf{T}}\overline{\Delta}_k^{\mathsf{T}} \boldsymbol{u}_k - \widehat{d'}^{\mathsf{T}}\nu = \lambda \sum_{k=1}^{p} \|\widehat{f'}_k\|_\infty$$

On one hand, $\widehat{d'}$ is a feasible solution so $\widehat{d'}^{\mathsf{T}}\nu \geq 0$ and so

$$\sum_{k=1}^{p} \widehat{d'}_k^{\mathsf{T}}\overline{\Delta}_k^{\mathsf{T}} \boldsymbol{u}_k \geq \sum_{k=1}^{p} \|\widehat{f'}_k\|_\infty.$$

On the other hand, by Holder's inequality:

$$\sum_{k=1}^{p} \widehat{d'}_k^{\mathsf{T}}\overline{\Delta}_k^{\mathsf{T}} \boldsymbol{u}_k \leq \sum_{k=1}^{p} \|\widehat{f'}_k\|_\infty \|\boldsymbol{u}_k\|_1$$

Since $\boldsymbol{u}_k$ can be chosen so that $\|\boldsymbol{u}_k\|_1 < 1$ for all $k \in S^c$, we would get a contradiction if $\|\widehat{f'}_k\|_\infty > 0$ for some $k \in S^c$. We thus conclude that $\widehat{d'}$ must follow the same sparsity pattern.

The second item in the theorem concerning optimization 6.5 is proven in exactly the same way.

The Lagrangian of optimization 6.5 is:

$$\mathcal{L}_{\text{cave}}(d_k, \nu_k) = \frac{1}{2n}\left\|\widehat{r} - \overline{\Delta}_k d_k\right\|_2^2 + \lambda\|\overline{\Delta}_k d_k\|_\infty + \sum_{k=1}^{p}\sum_{i=2}^{n-1} \nu_{ki} d_{ki}$$

The exact same reasoning applies to show that $\widehat{d}_k = 0$ satisfies KKT conditions sufficient for optimality.

□

9.2 Proof of False Positive Control

Note: the symbols c, C represent absolute constants. We will often abuse notation and "absorb" new absolute constants into c, C; the actual value of c, C could thus vary from line to line.

We first restate the theorem for convenience.

Theorem 9.2. *Suppose assumptions A1-A4 hold.*

Suppose $\lambda_n \geq csLb\sigma\sqrt{\frac{1}{n}\log^2 np}$, then with probability at least $1 - \frac{C}{n}$, for all $k \in S^c$, and for all $i' = 1, ..., n$:

$$\lambda_n > \left|\frac{1}{2n}\widehat{r}^{\mathsf{T}}\mathbf{1}_{(i':n)}\right|$$

And thereforefor all $k \in S^c$, both the AC solution $\widehat{f}_k$, from optimization 6.4, and the DC solution $\widehat{g}_k$, from optimization 6.5 are zero.

Proof. The key is to note that $\widehat{r}$ and $\Delta_{k,j}$ are independent for all $k \in S^c, j = 1, ..., n$ because $\widehat{r}$ is only dependent on X_S.

Fix j and i. $\widehat{r}^\mathsf{T}\mathbf{1}_{(i':n)}$ is then the sum of $n-i'+1$ random coordinates of $\widehat{r}$. We will then use Serfling's theorem on the concentration of measure of sampling without replacement. (corollary 10.2) We must first bound $\|\widehat{r}\|_\infty$ and $\frac{1}{n}\sum_{i=1}^n \widehat{r}_i$ before we can use Serfling's results however.

Step 1: Bounding $\|\widehat{r}\|_\infty$.

$\widehat{r}_i = f_0(x_i) + w_i - \widehat{f}(x_i)$ where $\widehat{f}(x_i) = \sum_{k\in S} \overline{\Delta}_k \widehat{d}_k$ is the convex additive function outputted by the restricted regression.

Both $f_0(x_i)$ and $\widehat{f}(x_i)$ are coordinate-wise L-Lipschitz and therefore are bounded by $2sLb$.

Because w_i is subgaussian, $|w_i| \le c\sigma\sqrt{\log\frac{2}{\delta}}$ with probability at most $1-\delta$. By union bound across $i = 1, ..., n$, we have that $\|w\|_\infty \le c\sigma\sqrt{\log\frac{2}{\delta}}$ with probability at most $1 - n\delta$.

We now put this together and take another union bound across all j and all i':

$$\begin{aligned}\|\widehat{r}\|_\infty &\le c(sLb + \sigma\sqrt{\log\frac{2}{\delta}}) \\ &\le csLb\sigma\sqrt{\log\frac{2}{\delta}}\end{aligned}$$

with probability at least $1 - n^2p\delta$. We supposed that both $sLb \ge 2$ and $\sigma\sqrt{\log\frac{2}{\delta}} \ge 2$.

Step 2: Bounding $|\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}|$.

$$\begin{aligned}\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1} &= \frac{1}{n}\sum_{i=1}^n f_0(x_i) + w_i - \widehat{f}(x_i) \\ &= \frac{1}{n}\sum_{i=1}^n f_0(x_i) + w_i \quad (\widehat{f} \text{ is centered})\end{aligned}$$

Because $|f_0(x_i)| \le sLb$, the first term $|\frac{1}{n}\sum_{i=1}^n f_0(x_i)|$ is at most $2sLb\sqrt{\frac{1}{n}\log\frac{2}{\delta}}$ with probability at most $1-\delta$ by Hoeffding Inequality.

Because w_i is subgaussian, the second term $|\frac{1}{n}\sum_{i=1}^n w_i|$ is at most $2\sigma\sqrt{\frac{1}{n}\log\frac{2}{\delta}}$ with probability at most $1-\delta$.

Taking an union bound, we have that

$$\begin{aligned}|\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}| &\le 2sLb\sqrt{\frac{1}{n}\log\frac{2}{\delta}} + 2\sigma\sqrt{\frac{1}{n}\log\frac{2}{\delta}} \\ &\le csLb\sigma\sqrt{\frac{1}{n}\log\frac{2}{\delta}}\end{aligned}$$

with probability at least $1 - 2\delta$.

Step 3: We now apply Serfling's theorem.

Serfling's theorem states that with probability at least $1-\delta$:

$$\left|\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}_{(i':n)}\right| \le 2\|\widehat{r}\|_\infty\sqrt{\frac{1}{n}\log\frac{2}{\delta}} + \left|\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}\right|$$

Taking an union bound across previous events, we have that with probability at least $1-3n^2p\delta$, for all $j \in S^c$, for all $i' = 1, ..., n$:

$$\left|\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}_{(i':n)}\right| \le csLb\sigma\sqrt{\frac{1}{n}\log\frac{2}{\delta}}$$

Setting $\delta = \frac{1}{n^3p}$ gives the desired result.

□

9.3 Proof of False Negative Control

Note: the symbols c, C represent absolute constants. We will often abuse notation and "absorb" new absolute constants into c, C; the actual value of c, C could thus vary from line to line.

We first introduce notations.

9.3.1 Notation

Let $f : \mathbb{R}^s \to \mathbb{R}$, we denote $\|f\|_P \equiv \mathbb{E}f(X)^2$.
Given samples $X_1, ..., X_n$, we denote $\|f\|_n \equiv \frac{1}{n}\sum_{i=1}^n f(X_i)^2$ and $\langle f, g\rangle_n \equiv \frac{1}{n}\sum_{i=1}^n f(X_i)g(X_i)$.

Let $\mathcal{C}^1$ denote the set of univariate convex functions supported on $[-b, b]$. Let $\mathcal{C}^1_L \equiv \{f \in \mathcal{C}^1 : \|\partial f\|_\infty \leq L\}$ denote the set of L-Lipschitz univariate convex functions.

Define $\mathcal{C}^s$ as the set of convex additive functions and $\mathcal{C}^s_L$ likewise as the set of L-Lipschitz convex additive functions.

$$\mathcal{C}^s \equiv \{f \,:\, f = \sum_{k=1}^s f_k,\ f_k \in \mathcal{C}^1\}$$

$$\mathcal{C}^s_L \equiv \{f \in \mathcal{C}^s \,:\, f = \sum_{k=1}^s f_k,\ \|\partial f_k\|_\infty \leq L\}$$

Let $f^*(x) = \sum_{k=1}^s f^*_k(x_k)$ be the population risk minimizer:

$$f^* = \arg\min_{f \in \mathcal{C}^s} \|f_0 - f^*\|_P^2$$

We let L be an upper bound on $\|\partial_{x_k} f_0\|_\infty$ and $\|\partial f^*_k\|_\infty$. Since f_0, f^* are supported on $[-b, b]^s$, it follows that $\|f_0\|_\infty, \|f^*\|_\infty \leq sLb$.

We define $\widehat{f}$ as the empirical risk minimizer:

$$\widehat{f} = \arg\min \left\{ \|y - f\|_n^2 + \lambda \sum_{k=1}^s \|f_k\|_\infty \,:\, f \in \mathcal{C}^s_L, \mathbf{1}_n^\mathsf{T} f_k = 0 \right\}$$

For $k \in \{1, ..., s\}$, define g^*_k to be decoupled concave population risk minimizer

$$g^*_k \equiv \arg\min_{g_k \in \text{-}\mathcal{C}^1} \|f_0 - f^* - g_k\|_P^2$$

In our proof, we will analyze g^*_k for k's such that $f^*_k = 0$. Likewise, we define the empirical version:

$$\widehat{g}_k \equiv \arg\min \left\{ \|f_0 - \widehat{f} - g_k\|_n^2 \,:\, g_k \in \text{-}\mathcal{C}^1_L\,, \mathbf{1}_n^\mathsf{T} g_k = 0 \right\}$$

By the definition of the ACDC procedure, $\widehat{g}_k$ exist only for k that have zero in their convex additive approximation.

9.3.2 Proof

By additive faithfulness of the ACDC procedure, it is necessary that $f^*_k \neq 0$ or $g^*_k \neq 0$ for all $k \in S$.

Intuitively, we would like to show the following:

$$\|f_0 - \widehat{f}\|_P \approx \|f_0 - f^*\|_P$$
$$\|f_0 - f^* - \widehat{g}_k\|_P \approx \|f_0 - f^* - g^*_k\|_P \quad \text{for all } k \in S \text{ where } f^*_k = 0$$

where the estimation error is a term that decreases with n.

Suppose $\widehat{f}_k = 0$ and $f_k^* \neq 0$, then, when n is large enough, there must exist a contradiction because the population risk of f^*, $\|f_0 - f^*\|_P$, is strictly larger than the population risk of the best approximation whose k-th component is constrained to be zero.

Suppose $f_k^* = 0$, then $g_k^* \neq 0$. When n is large enough, $\widehat{g}_k$ must not be zero or we would have another contradiction.

Theorem 9.3. *Let $\widehat{f}$ be the minimizer of the restricted regression with $\lambda \leq csB\sqrt{\frac{1}{n}\log^2 np}$.*
Then, with probability at least $1-\delta$,

$$\|f_0 - \widehat{f}\|_P^2 - \|f_0 - f^*\|_P^2 \leq cB^3\sigma\sqrt{\frac{s^5}{n^{4/5}}\log^2\frac{np}{\delta}} \tag{9.2}$$

Proof. **Step 1.** We start from the definition.

$$\|y - \widehat{f}\|_n^2 + \lambda\sum_{k=1}^s \|\widehat{f}_k\|_\infty \leq \|y - f^* + \overline{f}^*\|_n^2 + \lambda\sum_{k=1}^s \|f_k^* - \overline{f}^*\|_\infty$$

We plug in $y = f_0 + w$:

$$\|f_0 + w - \widehat{f}\|_n^2 + \lambda\sum_{k=1}^s \left(\|\widehat{f}_k\|_\infty - \|f_k^* - \overline{f}_k^*\|_\infty\right) \leq \|f_0 + w - f^* + \overline{f}^*\|_n^2$$

$$\|f_0 - \widehat{f}\|_n^2 + 2\langle w, f_0 - \widehat{f}\rangle_n + \lambda\sum_{k=1}^s \left(\|\widehat{f}_k\|_\infty - \|f_k^* - \overline{f}_k^*\|_\infty\right) \leq \|f_0 - f^* + \overline{f}^*\|_n^2 + 2\langle w, f_0 - f^* + \overline{f}^*\rangle$$

$$\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^* + \overline{f}^*\|_n^2 + \lambda\sum_{k=1}^s \left(\|\widehat{f}_k\|_\infty - \|f_k^* - \overline{f}_k^*\|_\infty\right) \leq 2\langle w, \widehat{f} - f^* + \overline{f}^*\rangle$$

The middle term can be bounded with the fact that $\|f_k^* - \overline{f}_k^*\|_\infty \leq 4B$ because f_k^* as is L-Lipschitz and supported on $[-b, b]$.

$$\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^* + \overline{f}^*\|_n^2 \leq 2\langle w, \widehat{f} - f^* + \overline{f}^*\rangle + \lambda 4sB$$

Using Lemma 10.2, we can remove $\overline{f}^*$ from the LHS. With probability at least $1-\delta$:

$$\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^*\|_n^2 \leq 2\langle w, \widehat{f} - f^* + \overline{f}^*\rangle + \lambda 4sB + c(sB)^2\frac{1}{n}\log\frac{2}{\delta} \tag{9.3}$$

Step 2. We upper bound $2\langle w, \widehat{f} - f^* + \overline{f}^*\rangle$ with bracketing entropy.

Define $\mathcal{G}$ as $\{f - f^* + \overline{f}^* : f \in \mathcal{C}^s\}$ as the set of convex additive functions centered around the function $f^* - \overline{f}^*$.

By Corollary 10.3, there is an ϵ-bracketing of $\mathcal{G}$ whose size is bounded by $\log N_{[]}(2\epsilon, \mathcal{G}, L_2(P)) \leq sK^{**}\left(\frac{2sbB}{\epsilon}\right)^{1/2}$, for all $\epsilon \in (0, sB\epsilon_3]$.

By Corollary 10.4, with probability at least $1-\delta$, each bracketing pair (h^U, h^L) is close in $L_1(P_n)$ norm, i.e., for all (h_U, h_L), $\frac{1}{n}\sum_{i=1}^n |h_U(X_i) - h_L(X_i)| \leq 2\epsilon + sB\sqrt{\frac{sK^{**}(2sbB)^{1/2} + \log\frac{1}{\delta}}{2\epsilon^{1/2}n}}$.

For each $h \in \mathcal{G}$, there exists a pair (h_U, h_L) such that $h_U(X_i) - h_L(X_i) \geq h(X_i) - h_L(X_i) \geq 0$. Therefore, with probability at least $1-\delta$, uniformly for all $h \in \mathcal{G}$:

$$\frac{1}{n}\sum_{i=1}^n |h(X_i) - h_L(X_i)| \leq \frac{1}{n}\sum_{i=1}^n |h_U(X_i) - h_L(X_i)| \leq 2\epsilon + (sB)\sqrt{\frac{sK^{**}(2sbBc_u)^{1/2} + \log\frac{1}{\delta}}{2\epsilon^{1/2}n}}$$

.

We denote $\epsilon_{n,\delta} \equiv (sB)\sqrt{\frac{sK^{**}(2sbBc_u)^{1/2}+\log\frac{1}{\delta}}{2\epsilon^{1/2}n}}$. Therefore,

$$|\langle w, h - h_L\rangle_n| \leq \|w\|_\infty \frac{1}{n}\sum_{i=1}^{n}|h(X_i) - h_L(X_i)| \leq \sigma\sqrt{\log\frac{n}{\delta}}\,(\epsilon + \epsilon_{n,\delta})$$

For the last inequality, we used the fact that w is a vector of independent subgaussian(σ) random variables and hence, by union bound, with probability at least $1-\delta$, $\|w\|_\infty \leq \sigma\sqrt{\log\frac{n}{\delta}}$.

By another property of subgaussian random variables, with probability at least $1-\delta$, $|\langle w, h_L\rangle_n| \leq \|h_L\|_n\sigma\sqrt{\frac{1}{cn}\log\frac{1}{\delta}}$. Applying an union bound, we have that $\sup_{h_L}|\langle w, h_L\rangle| \leq sB\sigma\sqrt{\frac{\log N_{[]}}{n}\log\frac{1}{\delta}}$.

Putting this together, we have that

$$\begin{aligned}
|\langle w, h\rangle_n| &\leq |\langle w, h_L\rangle_n| + |\langle w, h - h_L\rangle_n| \\
|\sup_{h\in\mathcal{G}}\langle w, h\rangle_n| &\leq |\sup_{h^L}\langle w, h^L\rangle_n| + \sigma\sqrt{\log\frac{n}{\delta}}(\epsilon + \epsilon_{n,\delta}) \\
&\leq sB\sigma\sqrt{\frac{\log N_{[]} + \log 1/\delta}{cn}} + \sigma\sqrt{\log\frac{n}{\delta}}(\epsilon + \epsilon_{n,\delta}) \\
&\leq sB\sigma\sqrt{\frac{sK^{**}(2sbB)^{1/2} + \log 1/\delta}{cn\epsilon^{1/2}}} + \sigma\sqrt{\log\frac{n}{\delta}}(\epsilon + \epsilon_{n,\delta}) \\
&\leq sB\sigma\sqrt{\frac{sK^{**}(2sbB)^{1/2} + \log 1/\delta}{cn\epsilon^{1/2}}} + \sigma\sqrt{\log\frac{n}{\delta}}\epsilon + sB\sigma\sqrt{\frac{sK^{**}(2sbB)^{1/2} + \log 1/\delta}{cn\epsilon^{1/2}}\log\frac{n}{\delta}} \\
&\leq \sigma\sqrt{\log\frac{n}{\delta}}\epsilon + sB\sigma\sqrt{\frac{sK^{**}(2sbB)^{1/2} + \log 1/\delta}{cn\epsilon^{1/2}}\log\frac{n}{\delta}}
\end{aligned}$$

The two terms are balanced when one sets $\epsilon = \sqrt{\frac{((sB)^2 sK^{**}(sBb)^{1/2})^{4/5}}{n^{4/5}}}$.

This is only valid if n is large enough so that $\epsilon \in (0, sB\epsilon_3]$.

We upper bound some terms to simplify the presentations again and end up with the following result:

$$|\sup_{h\in\mathcal{G}}\langle w, h\rangle| \leq sB\sigma\sqrt{\frac{sb^{1/2}\log\frac{1}{\delta}}{cn^{4/5}}}$$

Plugging this result into equation 10.3 and using an union bound, we get, with probability at least $1-2\delta$:

$$\begin{aligned}
\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^*\|_n^2 &\leq csB\sigma\sqrt{\frac{sb^{1/2}\log\frac{1}{\delta}}{cn^{4/5}}} + \lambda 4sB + c(sB)^2\frac{1}{n}\log\frac{2}{\delta} \\
\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^*\|_n^2 &\leq sB\sigma\sqrt{\frac{sb^{1/2}\log\frac{1}{\delta}}{cn^{4/5}}} + \lambda 4sB \\
&\leq cB\sigma\sqrt{\frac{s^3b^{1/2}}{n^{4/5}}\log\frac{C}{\delta}} + \lambda 4sB \qquad (9.4)
\end{aligned}$$

Step 3. We continue from equation 10.4, use lemma 10.1, use another union bound, with probability at least $1-3\delta$,

$$\|f_0 - \widehat{f}\|_P^2 - \|f_0 - f^*\|_P^2 \leq cB^2\sigma\sqrt{\frac{s^5(bc_u)^{2/5}}{n^{4/5}}\log\frac{C}{\delta}} + \lambda 4sB + cB^3\sqrt{\frac{s^5}{n^{4/5}}\log\frac{2}{\delta}}$$
$$\leq cB^3\sigma\sqrt{\frac{s^5(bc_u)^{2/5}}{n^{4/5}}\log\frac{2}{\delta}} + \lambda 4sB \tag{9.5}$$

Substituting in $\lambda \leq csB\sqrt{\frac{1}{n}\log^2 np}$ and we get the desired result.

□

Theorem 9.4. *Let $\widehat{g}_k$ denote the minimizer of the concave postprocessing with $\lambda \leq csLb\sqrt{\frac{1}{n}\log^2 np}$.*

Suppose n is large enough such that $(Lb)^3\sigma\sqrt{\frac{s^5}{n^{4/5}}\log^2\frac{4np}{\delta}} \leq 1$.

Then, with probability at least $1 - s\delta$, for all $k = 1, ..., s$:

$$\|f_0 - f^* - \widehat{g}_k\|_P^2 - \|f_0 - f^* - g_k^*\|_P^2 \leq c(Lb)^{2.5}\sigma^{0.5}\sqrt[4]{\frac{s^5}{n^{4/5}}\log^2\frac{4np}{\delta}}$$

Proof. The proof proceeds almost identically to that of theorem 10.3 because convex and concave functions have the same bracketing number.

Step 1. We start from the definition of $\widehat{g}_k$:

$$\|y - \widehat{f} - \widehat{g}_k\|_n^2 + \lambda\|\widehat{g}\|_\infty \leq \|y - \widehat{f} - g_k^*\|_n^2 + \lambda\|g^*\|_\infty$$
$$\|y - \widehat{f} - \widehat{g}_k\|_n^2 \leq \|y - \widehat{f} - g_k^*\|_n^2 + \lambda 2Lb$$

$$\|f_0 - \widehat{f} - \widehat{g}_k + w\|_n^2 \leq \|f_0 - \widehat{f} - g_k^* + w\|_n^2 + \lambda 2Lb$$
$$\|f_0 - \widehat{f} - \widehat{g}_k\|_n^2 - \|f_0 - \widehat{f} - g_k^*\|_n^2 \leq 2\langle w, \widehat{g}_k - g_k^*\rangle_n + \lambda 2Lb$$

Using identical analysis as Step 2 of the proof of Theorem 10.3 but setting $s = 1$, we have, with probability at least $1 - \delta$,

$$\|f_0 - \widehat{f} - \widehat{g}_k\|_n^2 - \|f_0 - \widehat{f} - g_k^*\|_n^2 \leq cB^2\sigma\sqrt{\frac{b^{1/2}}{n^{4/5}}\log\frac{C}{\delta}} + \lambda 2Lb$$

Using the uniform convergence result (lemma 10.1), we have, with probability at least $1 - 2\delta$:

$$\|f_0 - \widehat{f} - \widehat{g}_k\|_P^2 - \|f_0 - \widehat{f} - g_k^*\|_P^2 \leq c(Lb)^2\sigma\sqrt{\frac{1}{n}\log\frac{4}{\delta}} + \lambda 2Lb + c(Lb)^3\sqrt{\frac{s^5}{n^{4/5}}\log\frac{2}{\delta}}$$
$$\leq c(Lb)^3\sigma\sqrt{\frac{s^5}{n^{4/5}}\log\frac{4}{\delta}} + \lambda 2Lb$$

Plugging in $\lambda \leq \sqrt{\frac{1}{n}\log^2 np}$:

$$\|f_0 - \widehat{f} - \widehat{g}_k\|_P^2 - \|f_0 - \widehat{f} - g_k^*\|_P^2 \leq c(Lb)^3\sigma\sqrt{\frac{s^5}{n^{4/5}}\log\frac{4}{\delta}} + 2Lb\sqrt{\frac{1}{n}\log^2 np}$$
$$\|f_0 - \widehat{f} - \widehat{g}_k\|_P^2 - \|f_0 - \widehat{f} - g_k^*\|_P^2 \leq c(Lb)^3\sigma\sqrt{\frac{s^5}{n^{4/5}}\log^2\frac{4np}{\delta}}$$

Step 2. The goal is to bound the quality of approximation between $\|f_0 - \widehat{f} - g_k^*\|_P^2$ and $\|f_0 - f^* - g_k^*\|_P^2$ and likewise for $\widehat{g}_k$.

$$\begin{aligned}
\|f_0 - \widehat{f} - g_k^*\|_P^2 - \|f_0 - f^* - g_k^*\|_P^2 &\leq \|f_0 - \widehat{f}\|_P^2 - \|f_0 - f^*\|_P^2 - 2\langle f_0 - \widehat{f}, g_k^*\rangle + 2\langle f_0 - f^*, g_k^*\rangle \\
&\leq c(Lb)^3\sigma\sqrt{\frac{s^5}{n^{4/5}}\log^2\frac{4np}{\delta}} + 2|\langle \widehat{f} - f^*, g_k^*\rangle| \\
&\leq c(Lb)^3\sigma\sqrt{\frac{s^5}{n^{4/5}}\log^2\frac{4np}{\delta}} + 2\|\widehat{f} - f^*\|_P\|g_k^*\|_P \\
&\leq c(Lb)^3\sigma\sqrt{\frac{s^5}{n^{4/5}}\log^2\frac{4np}{\delta}} + c(Lb)\sqrt{(Lb)^3\sigma\sqrt{\frac{s^5}{n^{4/5}}\log^2\frac{4np}{\delta}}} \\
&\leq c(Lb)\sqrt{(Lb)^3\sigma\sqrt{\frac{s^5}{n^{4/5}}\log^2\frac{4np}{\delta}}}
\end{aligned}$$

The same bound likewise holds for $\widehat{g}_k$.

Step 3. Collecting the results from Step 1 and Step 2, we have, with probability at least $1 - 2\delta$:

$$\|f_0 - f^* - \widehat{g}_k\|_P^2 - \|f_0 - f^* - g_k^*\|_P^2 \leq c(Lb)^{2.5}\sigma^{0.5}\sqrt[4]{\frac{s^5}{n^{4/5}}\log^2\frac{4np}{\delta}}$$

Taking an union bound across $k = 1, ..., s$ dimensions completes the result.

□

9.3.3 Support Lemmas

Lemma 9.1. *With probability at least* $1 - \delta$*:*

$$\sup_{f \in \mathcal{C}_L^s}\left|\|f_0 - f\|_n^2 - \|f_0 - f\|_P^2\right| \leq cB^2\sqrt{\frac{s^5 b^{1/2}}{n^{4/5}}\log\frac{2}{\delta}}$$

Proof. Let $\mathcal{G}$ denote the off-centered set of convex functions, that is, $\mathcal{G} \equiv \mathcal{C}^s - f_0$. Note that if $h \in \mathcal{G}$, then $\|h\|_\infty = \|f_0 - f\|_\infty \leq 4sB$.

There exists an ϵ-bracketing of $\mathcal{G}$.

By Corollary 10.3, the bracketing has size at most $\log N_{[]}(2\epsilon, \mathcal{C}^s, L_1(P)) \leq sK^{**}\left(\frac{2sbB}{\epsilon}\right)^{1/2}$. By Corollary 10.4, we know that with probability at least $1-\delta$, we have that $\|h_U - h_L\|_{L_1(P_n)} \leq \epsilon + \epsilon_{n,\delta}$ for all pairs (h_U, h_L) in the bracketing, where $\epsilon_{n,\delta} = sB\sqrt{\frac{K^{**}(2sbB)^{1/2} + \log\frac{1}{\delta}}{2\epsilon^{1/2}n}}$.

For a particular function $h \in \mathcal{G}$, we can construct $\psi_L \equiv \min(|h_U|, |h_L|)$ and $\psi_U \equiv \max(|h_U|, |h_L|)$ so that

$$\psi_L^2 \leq h^2 \leq \psi_U^2$$

We can then bound the $L_1(P)$ norm of $\psi_U^2 - \psi_L^2$.

$$\begin{aligned}
\int(\psi_U^2(x) - \psi_L^2(x))p(x)dx &\leq \int|h_U^2(x) - h_L^2(x)|p(x)dx \\
&\leq \int|h_U(x) - h_L(x)|\,|h_U(x) + h_L(x)|p(x)dx \\
&\leq 2sB\epsilon
\end{aligned}$$

Now we can bound $\|h\|_n^2 - \|h\|_P^2$.

$$\frac{1}{n}\sum_{i=1}^{n}\psi_L(X_i)^2 - \mathbb{E}\psi_U(X)^2 \leq \|h\|_n^2 - \|h\|_P^2 \leq \frac{1}{n}\sum_{i=1}^{n}\psi_U(X_i)^2 - \mathbb{E}\psi_L(X)^2 \tag{9.6}$$

$\psi_L(X_i)^2$ and $\psi_U(X_i)^2$ are bounded random variables with upper bound sB. By uniform convergence, we have that, with probability at least $1-\delta$,, for all ψ_L (and likewise ψ_U):

$$\left|\frac{1}{n}\sum_{i=1}^{n}\psi_L(X_i)^2 - \mathbb{E}\psi_L(X)^2\right| \leq (sB)^2\sqrt{\frac{sK^{**}(sBb)^{1/2} + \log\frac{1}{\delta}}{\epsilon^{1/2}n}}$$

Plugging this into equation 10.6 above, we have that:

$$\begin{aligned}&\mathbb{E}\psi_L(X)^2 - \mathbb{E}\psi_U(X)^2 - (sB)^2\sqrt{\frac{sK^{**}(sBb)^{1/2} + \log\frac{1}{\delta}}{\epsilon^{1/2}n}}\\ &\leq \|h\|_n^2 - \|h\|_P^2 \leq \mathbb{E}\psi_U(X)^2 - \mathbb{E}\psi_L(X)^2 + (sB)^2\sqrt{\frac{sK^{**}(sBb)^{1/2} + \log\frac{1}{\delta}}{\epsilon^{1/2}n}}\end{aligned}$$

Using the $L_1(P)$ norm of $\psi_U^2 - \psi_L^2$ result, we have:

$$-sB\epsilon - (sB)^2\sqrt{\frac{sK^{**}(sBb)^{1/2} + \log\frac{1}{\delta}}{\epsilon^{1/2}n}} \leq \|h\|_n^2 - \|h\|_P^2 \leq sB\epsilon + (sB)^2\sqrt{\frac{sK^{**}(sBb)^{1/2} + \log\frac{1}{\delta}}{\epsilon^{1/2}n}}$$

We balance the terms by choosing $\epsilon = \left(\frac{(sB)^2 sK^{**}(sBb)^{1/2}}{n}\right)^{2/5}$.
We have then that, with probability at least $1-\delta$,

$$\left|\|h\|_n^2 - \|h\|_P^2\right| \leq B^2\sqrt{\frac{s^5 b^{1/2}\log\frac{1}{\delta}}{n^{4/5}}}$$

□

Lemma 9.2. *Let f_0, f^* be defined as in section 10.3.1. Define $\overline{f}^* = \frac{1}{n}\sum_{i=1}^{n} f^*(X_i)$.*
Then, with probability at least $1-2\delta$,

$$\left|\|f_0 - f^*\|_n^2 - \|f_0 - f^* + \overline{f}^*\|_n^2\right| \leq c(sLb)^2\frac{1}{n}\log\frac{4}{\delta}$$

Proof. (of lemma 10.2)

$$\begin{aligned}\|f_0 - f^* + \overline{f}^*\|_n^2 &= \|f_0 - f^*\|_n^2 + 2\langle f_0 - f^*, \overline{f}^*\rangle + \overline{f}^{*2}\\ &= \|f_0 - f^*\|_n^2 + 2\overline{f}^*\langle f_0 - f^*, \mathbf{1}\rangle_n + \overline{f}^{*2}\\ &= \|f_0 - f^*\|_n^2 + 2\overline{f}^*\overline{f}_0 - \overline{f}^{*2}\end{aligned}$$

$\overline{f}^* = \frac{1}{n}\sum_{i=1}^{n} f^*(X_i)$ is the average of n bounded mean-zero random variables and therefore, with probability at least $1-\delta$, $|\overline{f}^*| \leq 4sLb\sqrt{\frac{1}{n}\log\frac{2}{\delta}}$.
The same reasoning likewise applies to $\overline{f}_0 = \frac{1}{n}\sum_{i=1}^{n} f_0(X_i)$.
Taking a union bound and we have that, with probability at least $1-2\delta$,

$$\begin{aligned}|\overline{f}^*||\overline{f}_0| &\leq c(sLb)^2\frac{1}{n}\log\frac{2}{\delta}\\ \overline{f}^{*2} &\leq c(sLb)^2\frac{1}{n}\log\frac{2}{\delta}\end{aligned}$$

Therefore, with probability at least $1-2\delta$,

$$\|f_0 - f^*\|_n^2 - c(sLb)^2 \frac{1}{n}\log\frac{2}{\delta} \leq \|f_0 - f^* + \overline{f}^*\|_n^2 \leq \|f_0 - f^*\|_n^2 + c(sLb)^2 \frac{1}{n}\log\frac{2}{\delta}$$

□

9.4 Supporting Technical Material

9.4.1 Concentration of Measure

Sub-Exponential random variable is the square of a subgaussian random variableVershynin [2010].

Proposition 9.1. *(Subexponential Concentration Vershynin [2010]) Let $X_1, ..., X_n$ be zero-mean independent subexponential random variables with subexponential scale K.*

$$P(|\frac{1}{n}\sum_{i=1}^{n} X_i| \geq \epsilon) \leq 2\exp\left[-cn\min\left(\frac{\epsilon^2}{K^2}, \frac{\epsilon}{K}\right)\right]$$

where $c > 0$ is an absolute constant.

For uncentered subexponential random variables, we can use the following fact. If X_i subexponential with scale K, then $X_i - \mathbb{E}[X_i]$ is also subexponential with scale at most $2K$.

Restating. We can set

$$c\min\left(\frac{\epsilon^2}{K^2}, \frac{\epsilon}{K}\right) = \frac{1}{n}\log\frac{1}{\delta}.$$

Thus, with probability at least $1-\delta$, the deviation at most

$$K\max\left(\sqrt{\frac{1}{cn}\log\frac{C}{\delta}}, \frac{1}{cn}\log\frac{C}{\delta}\right)$$

Corollary 9.1. *Let $w_1, ..., w_n$ be n independent subgaussian random variables with subgaussian scale σ.*

Then, for all $n > n_0$, with probability at least $1 - \frac{1}{n}$,

$$\frac{1}{n}\sum_{i=1}^{n} w_i^2 \leq c\sigma^2$$

Proof. Using the subexponential concentration inequality, we know that, with probability at least $1 - \frac{1}{n}$,

$$|\frac{1}{n}\sum_{i=1}^{n} w_i^2 - \mathbb{E}w^2| \leq \sigma^2\max\left(\sqrt{\frac{1}{cn}\log\frac{C}{\delta}}, \frac{1}{cn}\log\frac{C}{\delta}\right)$$

First, let $\delta = \frac{1}{n}$. Suppose n is large enough such that $\frac{1}{cn}\log Cn < 1$. Then, we have, with probability at least $1 - \frac{1}{n}$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} w_i^2 &\leq c\sigma^2\Big(1 + \sqrt{\frac{1}{cn}\log Cn}\Big) \\ &\leq 2c\sigma^2\end{aligned}$$

□

9.4.2 Sampling Without Replacement

Lemma 9.3. *(Serfling Serfling [1974]) Let $x_1, ..., x_N$ be a finite list, $\overline{x} = \mu$. Let $X_1, ..., X_n$ be sampled from x without replacement.*

Let $b = \max_i x_i$ and $a = \min_i x_i$. Let $r_n = 1 - \frac{n-1}{N}$. Let $S_n = \sum_i X_i$. Then we have that

$$P(S_n - n\mu \geq n\epsilon) \leq \exp(-2n\epsilon^2 \frac{1}{r_n(b-a)^2})$$

Corollary 9.2. *Suppose $\mu = 0$.*

$$P(\frac{1}{N}S_n \geq \epsilon) \leq \exp(-2N\epsilon^2 \frac{1}{(b-a)^2})$$

And, by union bound, we have that

$$P(|\frac{1}{N}S_n| \geq \epsilon) \leq 2\exp(-2N\epsilon^2 \frac{1}{(b-a)^2})$$

A simple restatement. With probability at least $1 - \delta$, the deviation $|\frac{1}{N}S_n|$ is at most $(b - a)\sqrt{\frac{1}{2N}\log\frac{2}{\delta}}$.

Proof.

$$P(\frac{1}{N}S_n \geq \epsilon) = P(S_n \geq \frac{N}{n}n\epsilon) \leq \exp(-2n\frac{N^2}{n^2}\epsilon^2 \frac{1}{r_n(b-a)^2})$$

We note that $r_n \leq 1$ always, and $n \leq N$ always.

$$\exp(-2n\frac{N^2}{n^2}\epsilon^2 \frac{1}{r_n(b-a)^2}) \leq \exp(-2N\epsilon^2 \frac{1}{(b-a)^2})$$

This completes the proof.

□

9.4.3 Bracketing Number for Convex Functions

Definition 9.1. Let $\mathcal{C}$ be a set of functions. For a given ϵ and metric ρ (which we take to be L_2 or $L_2(P)$), we define a **bracketing** of $\mathcal{C}$ to be a set of pairs of functions $\{(f_L, f_U)\}$ satisfying (1) $\rho(f_L, f_U) \leq \epsilon$ and (2) for any $f \in \mathcal{C}$, there exist a pair (f_L, f_U) where $f^U \geq f \geq f^L$.

We let $N_{[]}(\epsilon, \boldsymbol{C}, \rho)$ denote the size of the smallest bracketing of $\mathcal{C}$

Proposition 9.2. *(Proposition 16 in Kim and Samworth [2014])*
Let $\mathcal{C}$ be the set of convex functions supported on $[-b, b]^d$ and uniformly bounded by B. Then there exist constants ϵ_3 and K^{**}, dependent on d, such that

$$\log N_{[]}(2\epsilon, \mathcal{C}, L_2) \leq K^{**}\left(\frac{2bB}{\epsilon}\right)^{d/2}$$

for all $\epsilon \in (0, B\epsilon_3]$.

It is trivial to extend Kim and Samworth's result to $L_2(P)$ norm for an absolutely continuous distribution P.

Proposition 9.3. *Let P be a distribution with a density p. Let $\mathcal{C}, b, B, \epsilon_3, K^{**}$ be defined as in Proposition 10.2. Then,*

$$\log N_{[]}(2\epsilon, \mathcal{C}, L_1(P)) \leq K^{**}\left(\frac{2bB}{\epsilon}\right)^{d/2}$$

For all $\epsilon \in (0, B\epsilon_3]$.

Proof. Let $\mathcal{C}_\epsilon$ be the bracketing the satisfies the size bound in Proposition 10.3.

Let $(f_L, f_U) \in \mathcal{C}_\epsilon$. Then we have that:

$$\begin{aligned}
\|f_L - f_U\|_{L_1(P)} &= \int |f_L(x) - f_U(x)| p(x) dx \\
&\leq \left(\int |f_L(x) - f_U(x)|^2 dx \right)^{1/2} \left(\int p(x)^2 dx \right)^{1/2} \\
&\leq \left(\int |f_L(x) - f_U(x)|^2 dx \right)^{1/2} \\
&\leq \|f_L - f_U\|_{L_2} \leq \epsilon
\end{aligned}$$

On the third line, we used the fact that $\int p(x)^2 dx \leq \left(\int p(x) dx \right)^2 \leq 1$. □

It is also simple to extend the bracketing number result to additive convex functions. As before, let $\mathcal{C}^s$ be the set of additive convex functions with s components.

Corollary 9.3. *Let P be a distribution with a density p upper bounded by c_u. Let b, B, ϵ_3, K^{**} be defined as in Proposition 10.2. Then,*

$$\log N_{[]}(2\epsilon, \mathcal{C}^s, L_1(P)) \leq s K^{**} \left(\frac{2sbB}{\epsilon} \right)^{1/2}$$

For all $\epsilon \in (0, sB\epsilon_3]$.

Proof. Let $f \in \mathcal{C}^s$. We can construct an ϵ-bracketing for f through ϵ/s-bracketings for each of the components $\{f_k\}_{k=1,...,s}$:

$$f_U = \sum_{k=1}^{s} f_{Uk} \qquad f_L = \sum_{k=1}^{s} f_{Lk}$$

It is clear that $f_U \geq f \geq f_L$. It is also clear that $\|f_U - f_L\|_{L_1(P)} \leq \sum_{k=1}^{s} \|f_{Uk} - f_{Lk}\|_{L_1(P)} \leq \epsilon$. □

The following result follows from Corollary 10.3 directly by an union bound.

Corollary 9.4. *Let $X_1, ..., X_n$ be random samples from a distribution P. Let $1 > \delta > 0$. Let $\mathcal{C}^s_\epsilon$ be an ϵ-bracketing of $\mathcal{C}^s$ with respect to the $L_1(P)$-norm whose size is at most $N_{[]}(2\epsilon, \mathcal{C}^s, L_1(P))$. Let $\epsilon \in (0, sB\epsilon_3]$.*

Then, with probability at least $1 - \delta$, for all pairs $(f_L, f_U) \in \mathcal{C}^s_\epsilon$, we have that

$$\frac{1}{n} \sum_{i=1}^{n} |f_L(X_i) - f_U(X_i)| \leq \epsilon + \epsilon_{n,\delta}$$

where $\epsilon_{n,\delta} \equiv sB \sqrt{\frac{\log N_{[]}(2\epsilon, \mathcal{C}^s, L_2(P)) + \log \frac{1}{\delta}}{2n}} = sB \sqrt{\frac{sK^{**}(2sbB)^{d/2} + \log \frac{1}{\delta}}{2\epsilon^{d/2} n}}$

Proof. $|f_L(X_i) - f_U(X_i)|$ is at most sB. There are $N_{[]}(2\epsilon, \mathcal{C}^s, L_1(P))$ pairs (f_L, f_U). The inequality follows from a direct application of union bound and Hoeffding Inequality.

□