Variable Selection in Convex Function Estimation

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

May 5, 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.

1 Introduction

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. Recent work Guntuboyina [2012], Guntuboyina and Sen [2013] shows that the minimax rate for convex function estimation in p dimensions is $n^{-4/(4+p)}$. Loosely speaking, this shows that the geometric convexity constraint is statistically equivalent to requiring two derivatives of the function, and thus is subject to the same curse of dimensionality. However, if the function is sparse, with $s \ll p$ relevant variables, then the faster rate $n^{-4/(4+s)}$ may be achievable if the s variables can be identified. To determine the relevant variables, we show that it suffices to estimate a sum of p one-dimensional convex functions, leading to significant computational and statistical advantages. In addition, we introduce algorithms and supporting statistical theory for a practical, effective approach to this variable selection problem.

The general sparse nonparametric regression problem is considered in Lafferty and Wasserman [2008], where it is shown that computationally efficient, near minimax-optimal estimation is possible, but in ambient dimensions that scale only as $p = O(\log n)$ instead of $p = O\big(e^{n^c}\big)$ as enjoyed by sparse linear models. Comminges and Dalalyan [2012] do achieve exponential scaling $p = O(e^n)$ under certain Fourier smoothness conditions but they show that variable selection is still hard in that the number of relevant variables s must be less than $\log n$.

Approximating the regression function by a sum of one-dimensional functions, known as sparse additive models, Ravikumar et al. [2009] is a practical alternative to fully nonparametric function estimation. But the additive assumption is limited. In particular, the natural idea of first selecting the single variable effects, then the pairwise effects, and so on, does not in general lead to consistent variable selection. In other words, the general nonparametric model is not additively faithful. Remarkably, the additional assumption of convexity does lead to consistent variable selection, as we show here. In addition, we show that the scaling $p = O(\log n)$ and $n = O\big(\text{poly}(s)\big)$ is achievable for sparse convex additive models. Thus, the geometric convexity constraint is quite different from the smoothness constraints imposed in traditional nonparametric regression.

A key to 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

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} (Y_i - m(x_i))^2 \\ \text{subject to} \quad & m : \mathbb{R}^p \to \mathbb{R} \text{ is convex} \end{aligned} \tag{1.1}$$

is precisely equivalent to the finite dimensional quadratic program

$$\begin{aligned} \text{minimize}_{h,\beta} \quad & \sum_{i=1}^{n} (Y_i - h_i)^2 \\ \text{subject to} \quad & h_j \geq h_i + \beta_i^T (x_j - x_i), \text{ for all } i, j. \end{aligned} \tag{1.2}$$

Here h_i is the estimated function value $m(x_i)$, and the vectors $\beta_i \in \mathbb{R}^d$ represent supporting hyperplanes to the epigraph of m. Importantly, this finite dimensional quadratic program does not have tuning parameters for smoothing the function. Such parameters are the bane of nonparametric estimation.

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, we offer that the convexity assumption is attractive as a tractable, nonparamametric relaxation of the linear model. In addition to the lack of tuning parameters, other than the regularization parameter λ to control the level of sparsity, the global convexity assumption leads to effective, scalable algorithms. We demonstrate use of our approach on experiments with standard regression data sets, in a comparison with sparse linear models (lasso).

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 Additive Faithfulness

For general regression, additive approximation may result in a relevant variable being incorrectly marked as irrelevant. Such mistakes are inherent to the approximation and may persist even with infinite samples. 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 3.1. *Let F be a distribution on $C = [0,1]^s$ with a positive density function p. Let $f : C \to \mathbb{R}$ be an integrable function*

Let $f_1^*, ..., f_s^*, \mu^* := \arg\min\{\mathbb{E}\big(f(X) - \sum_{k=1}^{s} f_k(X_k) - mu\big)^2 \,:\, \forall k, \mathbb{E}f_k(X_k) = 0\}$

Then

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

and $\mu^* = \mathbb{E}f(X)$ and this solution is unique.

Lemma 3.1 follows from the stationarity conditions of the optimal solution.

Proof. Let $f_1^*, ..., f_s^*, \mu^*$ be the minimizers as defined.

We first show that the optimal $\mu^* = \mathbb{E}f(X)$ 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 0 and strong convexity is guaranteed.

We now turn our attention toward the f_k^*'s.

It must be that f_k^* minimizes $\Big\{\mathbb{E}\big(f(X) - \mu^* - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - f_k(X_k)\big)^2 \,:\, \mathbb{E}f_k(X_k) = 0\Big\}$.

Fix x_k, we will show that the value $\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}(X_{k'}) \,|\, x_k] - \mu^*$, for all x_k, uniquely minimizes

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

It easily follows then that the function $x_k \mapsto \mathbb{E}[f(X) - \sum_{k'} f_{k'}(X_{k'}) \,|\, x_k] - \mu^*$ is the unique f_k^* that minimizes the expected square error.We focus our attention on f_K^*, and fix x_k.

The first-order optimality condition gives us:

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

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

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

The square error objective is strongly convex. 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.

Now, we note that as a function of x_k, $\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}(X_{k'}) | x_k] - \mathbb{E}f(X)$ has mean zero and we thus finish the proof. □

In the case that the distribution in Lemma 3.1 is a product distribution, we get particularly clean expressions for the additive components.

Corollary 3.1. *Let F be a product distribution on $\boldsymbol{C} = [0,1]^s$ with density function p which is positive on $\boldsymbol{C}$. Let $\mu^*, f^*_k(x_k)$ be defined as in Lemma 3.1. 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.*

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

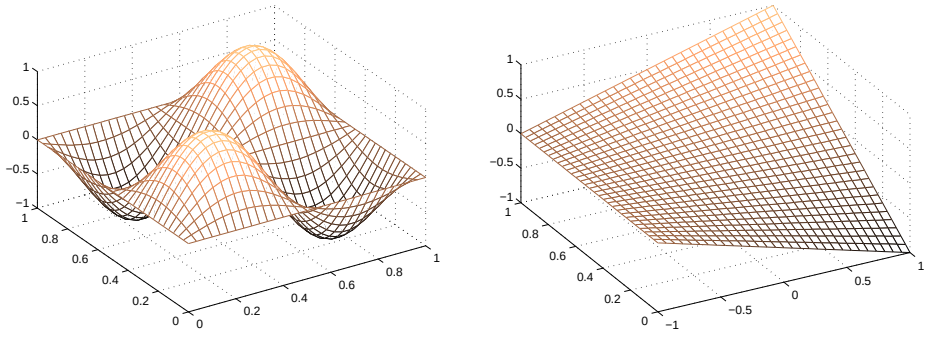
Example 3.1. Using Corollary 3.1, we give two examples of *additive unfaithfulness* under the uniform distribution, that is, examples where relevant variables are erroneously marked as irrelevant under an additive approximation. First, consider the following function:

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

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

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

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, 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 3.1. Let F be a distribution on $\boldsymbol{C} = [0,1]^s$, and $f : \boldsymbol{C} \to \mathbb{R}$.

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 1.

If f is differentiable, then f depends on k if $\partial_{x_k} f \neq 0$ with probability greater than 0.

Suppose we have the additive approximation:

$$f_k^*, \mu^* := \operatorname*{arg\,min}_{f_1,\ldots,f_s,\mu} \Big\{ \mathbb{E}(f(X) - \sum_{k=1}^{s} f_k(X_k) - \mu)^2 \,:\, \mathbb{E} f_k(X_k) = 0 \Big\}. \tag{3.1}$$

We say that f is **additively faithful** under F in case $f_k^* = 0 \Rightarrow f$ 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.

3.1 Additive Faithfulness of Convex Functions

Remarkably, under a general class of distributions which we characterize below, convex multivariate functions are additively faithful.

Definition 3.2. Let $p(\boldsymbol{x})$ be a density supported on $[0,1]^s$, p satisfies the *boundary-points condition* if, for all j, and for all $\boldsymbol{x}_{-j}$:

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

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 properties 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 the density of any product distributions.

The following theorem is the main result of this section.

Theorem 3.1. *Let p be a positive density supported on $C = [0,1]^s$ that satisfies the boundary-points property (definition 3.2). If f is convex and twice differentiable, then f is* additively faithful *under p.*

We pause to give some intuition before we present the full proof: suppose the underlying distribution is a product distribution for a second, then we know from lemma 3.1 that the additive approximation zeroes out k when, fixing x_k, every "slice" of f integrates to zero. We prove Theorem 3.1 by showing that "slices" of convex functions that integrate to zero cannot be "glued" together while still maintaining convexity.

Proof. (of Theorem 3.1)
Fix k. Using the result of Lemma 3.1, 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 then use the shorthand notation that $r(\boldsymbol{x}_{-k}) = \sum_{k' \neq k} f_{k'}(x_{k'})$ and assume without loss of generality that $\mu = 0$. We then assume that for all x_k,

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

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, which is valid because all functions are bounded.

$$\int_{\boldsymbol{x}_{-k}} p'(\boldsymbol{x}_{-k} \,|\, x_k)\big(f(\boldsymbol{x}) - r(\boldsymbol{x}_{-k})\big) + p(\boldsymbol{x}_{-k} \,|\, x_k) f'(x_k, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0 \tag{3.2}$$

$$\int_{\boldsymbol{x}_{-k}} p''(\boldsymbol{x}_{-k} \,|\, x_k)\big(f(\boldsymbol{x}) - r(\boldsymbol{x}_{-k})\big) + 2p'(\boldsymbol{x}_{-k} \,|\, x_k) f'(x_k, \boldsymbol{x}_{-k}) + p(\boldsymbol{x}_{-k} \,|\, x_k) f''(x_k, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0 \tag{3.3}$$

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

$$\int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k} \,|\, x_k^0) f'(x_k^0, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0 \tag{3.4}$$

$$\int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k} \,|\, x_k^0) f''(x_k^0, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0 \tag{3.5}$$

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 $p(\boldsymbol{x}_{-k} \,|\, x_k^0) > 0$ by 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 [1], which means that *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 3.4, we conclude then 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 $\boldsymbol{x}_{-k}$:

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

f thus does not depend on x_k.

□

Theorem 3.1 plays an important role in our sparsistency analysis, where we show that the additive approximation is variable selection consistent (or "sparsistent"), even when the true function is not additive.

Remark 3.1. We assume twice differentiability in Theorems 3.1 to simplify the proof. We believe this smoothness condition is not necessary because every non-smooth convex function can be approximated arbitrarily well by a smooth one.

Remark 3.2. It is difficult to prove the opposite direction of additive faithfulness, that is, if f does not depend on coordinate k, then f_k^* will be zero in the additive approximation. Consider as a conceptual example a 3D distribution over (X_1, X_2, X_3); suppose X_1, X_2 are independent, and f is only a function of X_1, X_2. We can then let $X_3 = f(X_1, X_2) - f_1^*(X_1) - f_2^*(X_2)$, that is, we let X_3 exactly capture the additive approximation error, then 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.

[1] See proposition 7.1.10 of Horn and Johnson [1990]

3.2 Convex Additive Model

Although convex functions are additively faithful, it is difficult to estimate the optimal additive functions f_k^*'s (as defined in equation 3.1) because f_k^* need not be a convex function.

Because the true regression function f is convex, it is natural to consider an convex additive model:

$$\{f_k^*\}_{k=1}^s = \arg\min \Big\{ \mathbb{E}\big(f(X) - \sum_{k=1}^s f_k(X_k)\big)^2 \,:\, f_k \in \mathcal{C}^1,\, \mathbb{E} f_k(X_k) = 0 \Big\} \tag{3.6}$$

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

The convex functions f_k^*'s are not additively faithful by themselves but faithfulness can be restored by coupling the f_k^*'s with a set of concave functions:

$$g_k^* = \arg\min \Big\{ \mathbb{E}\big(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k\big)^2 \,:\, g_k \in \text{-}\mathcal{C}^1, \mathbb{E} g_k(X_k) = 0 \Big\} \tag{3.7}$$

Theorem 3.2. *Suppose $p(\boldsymbol{x})$ is a positive density on $C = [0,1]^p$ and satisfies the boundary-points condition. Suppose that f is convex and twice-differentiable.*

Suppose 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.

Then, $f_k^* = 0$ and $g_k^* = 0$ (as defined in equation 3.6 and equation 3.7) implies that f does not depend on x_k, i.e., $\partial_{x_k} f(\boldsymbol{x}) = 0$ with probability 1.

Theorem 3.2 suggests a two-stage screening procedure for variable selection. We first fit a sparse *convex* additive model and then, for every variable marked as irrelevant in the first stage, we allow ourself to change our mind by fitting an univariate *concave* function. We refer to this procedure as AC/DC (additive convex / decoupled concave) and describe it in more details in section 4.

Proof. (of theorem 3.2)
Fix k. Let f_k^*, g_k^*'s be defined as equation 3.6 and equation 3.7.

By definition of f_k^* and g_k^*, we have that:

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

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

Therefore, $f_k^* = 0$ and $f_k^* = 0$ implies that $\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) \,|\, x_k] = 0$.

Now we use corollary 3.2. We let $\phi(\boldsymbol{x}_{-k}) = f(\boldsymbol{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'})$ and conclude that f does not depend on x_k. □

Corollary 3.2. *Suppose $p(x)$ is a positive density on $[0,1]^p$ and it satisfies the boundary-points condition.*

For any function $\phi(X_{-k})$ that does not depend on X_k:

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

We have that $f_k^* = 0 \Rightarrow \partial_{x_k} f(x) = 0$.

Proof. In the proof of theorem 3.1, 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 3.1 except we let $\phi(\boldsymbol{x}_{-k}) = r(\boldsymbol{x}_{-k})$. □

Proposition 3.1. *Let $C \subset \mathbb{R}^p$ be a compact set and let $h : C \to \mathbb{R}$. Let $p(x)$ be a positive density on C and suppose $\mathbb{E}h(X) = 0$. Suppose that $\partial_{x_k} h(x)$, $\partial_{x_k} p(x \,|\, x_k)$, and $\partial^2_{x_k} p(x \,|\, x_k)$ are all continuous as functions on C. Suppose that $\partial^2_{x_k} h(x) \geq 0$.*

Let

$$f_k^*(x_k) = \arg\min_{f_k} \left\{ \mathbb{E}\Big(h(X) - f_k(X_k)\Big)^2 \,:\, f_k \in \mathcal{C}^1,\, \mathbb{E}f_k = 0 \right\}$$

$$g_k^*(x_k) = \arg\min_{g_k} \left\{ \mathbb{E}\Big(h(X) - g_k(X_k)\Big)^2 \,:\, g_k \in \text{-}\mathcal{C}^1,\, \mathbb{E}g_k = 0 \right\}$$

be the best convex and concave univariate approximation respectively.

Then, $f_k^* = 0$ and $g_k^* = 0$ iff $h_k^*(x_k) \equiv \mathbb{E}[h(X) \,|\, x_k] = 0$.

Proof. First, we will establish that $h_k^*(x_k)$ is twice differentiable and that $\partial^2_{x_k} h_k^*(x_k)$ is lower bounded.

$$\begin{aligned} h_k^*(x_k) &= \mathbb{E}[h(X) \,|\, x_k] \\ &= \int_{x_{-k}} h(x) p(\boldsymbol{x}_{-k} \,|\, x_k) d\boldsymbol{x}_{-k} \\ \partial^2_{x_k} h_k^*(x_k) &= \int_{\boldsymbol{x}_{-k}} h''(\boldsymbol{x}) p(\boldsymbol{x}_{-k} \,|\, x_k) + 2h'(x) p'(\boldsymbol{x}_{-k} \,|\, x_k) + h(x) p''(\boldsymbol{x}_{-k} \,|\, x_k) d\boldsymbol{x}_{-k} \end{aligned}$$

The first term $h''(x)p(\boldsymbol{x}_{-k} \,|\, x_k)$ is strictly positive. By assumption, the remaining terms are continuous and hence bounded on a compact set. $\partial^2_{x_k} h_k^*(x_k)$ is therefore lower bounded. Before proceeding, we also note that because $\mathbb{E}h(X) = 0$, it must be that $\mathbb{E}h_k^*(X_k) = 0$.

Now suppose $f_k^* = 0$ and $g_k^* = 0$. Let σ_k^2 denote $\mathbb{E}X_k^2$. Then

$$\arg\min_{c \in \mathbb{R}} \mathbb{E}\Big(h(X) - c(X_k^2 - \sigma_k^2)\Big)^2 = 0$$

Since optimal $c^* = \frac{\mathbb{E}[h(X)(X_k^2 - \sigma_k^2)]}{\mathbb{E}[X_k^2]}$, we know $\mathbb{E}[h(X)X_k^2] = \mathbb{E}\Big[\mathbb{E}[h(X) \,|\, X_k]X_k^2\Big] = 0$.

Because $\partial^2_{x_k} h_k^*(x_k)$ is lower bounded, for large enough α, $h_k^*(x_k) + \alpha(x_k^2 - \sigma_k^2)$ has a non-negative second derivative and thus is convex. Then

$$\arg\min_{c \in \mathbb{R}} \mathbb{E}\Big(h(X) - c(h_k^*(X_k) + \alpha(X_k^2 - \sigma_k^2))\Big)^2 = 0$$

Again, $c^* = \frac{\mathbb{E}[g(X)\left(h_k^*(X_k) + \alpha(X_k^2 - \sigma_k^2)\right)]}{\mathbb{E}\left(h_k^*(X_k) + \alpha(X_k^2 - \sigma_k^2)\right)^2} = 0$, so

$$\begin{aligned} \mathbb{E}[h(X)\big(h_k^*(X_k) + \alpha X_k^2\big)] &= \mathbb{E}[h(X)h_k^*(X_k)] \\ &= \mathbb{E}\Big[\mathbb{E}[h(X)|X_k]h_k^*(X_k)\Big] \\ &= \mathbb{E}h_k^*(X_k)^2 = 0 \end{aligned}$$

Therefore, $h_k^*(x_k) = 0$.

□

4 Estimation Procedure

Theorem 3.2 suggest a two stage procedure for variable selection in convex regression–first learning a sparse convex additive model and then, on the residual, separately learning an univariate concave function for each of the dimensions.

More precisely, given samples $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$, we perform the following procedure, which we refer to as AC/DC (additively convex / decoupled concave):

1. Compute:
$$\widehat{f}_1, ..., \widehat{f}_p = \underset{f_1,...,f_p \in \mathcal{C}^1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \sum_{k=1}^{p} f_k(x_{ki}) \Big)^2 + \lambda \sum_{k=1}^{p} \|f_k\|_\infty \tag{4.1}$$

2. Compute, for each k such that $\|\widehat{f}_k\| = 0$:
$$\widehat{g}_k = \underset{g_k \in \mathcal{C}^1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \sum_{k'} \widehat{f}_{k'}(x_{k'i}) - g_k(x_{ki}) \Big)^2 + \lambda \|g_k\|_\infty \tag{4.2}$$

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

We added an ℓ_∞/ℓ_1 penalty in equation 4.1 and the ℓ_∞ penalty in equation 4.2 to encourage sparsisty. There are other forms of penalty that can also produce sparse estimates such as a penalty on the derivative of each of the component functions; we find the form we use to be convenient both for theoretical analysis and implementation.

5 Optimization

We describe in detail the optimization algorithm for the additive convex regression stage, the second decoupled concave regression stage follows almost identical procedure.

We let $\boldsymbol{x}_i \in \mathbb{R}^p$ be the covariate, Y_i be the response and 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.,

$$f_{kj} \geq f_{ki} + \beta_{ki}(x_{kj} - x_{ki}) \quad (\forall i, j) \tag{5.1}$$

where $f_{ki} := f_k(x_{ki})$ and β_{ki} is the subgradient at point x_{ki}. We apparently need $O(n^2 p)$ constraints to impose the supporting hyperplane constraints, which is computationally expensive for large scale problems. 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 our optimization program:

$$\begin{aligned} \min_{\boldsymbol{h},\boldsymbol{\beta},\mu} \quad & \frac{1}{2n} \sum_{i=1}^{n} \Big(Y_i - \sum_{k=1}^{p} f_{ki} - \mu \Big)^2 + \lambda \sum_{k=1}^{p} \|\boldsymbol{f}_{k\cdot}\|_\infty \\ \text{subject to} \quad & f_{k(i+1)} = f_{k(i)} + \beta_{k(i)}(x_{k(i+1)} - x_{k(i)}), \\ & \sum_{i=1}^{n} f_{ki} = 0, \\ & \beta_{k(i+1)} \geq \beta_{k(i)} \ (\forall k, i) \end{aligned} \tag{5.2}$$

Here $\{(1), (2), \ldots, (n)\}$ is a reordering of $\{1, 2, \ldots, n\}$ such that $x_{k(1)} \leq x_{k(2)} \leq \cdots \leq x_{k(n)}$.

We can solve for μ explicitly, as $\mu = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$ which follows from the KKT conditions and the constraints $\sum_i f_{ki} = 0$. It is easy to verify that the constraints in (5.2) satisfy the supporting

hyperplane constraints, as

$$\forall j \geq i,\ f_{k(j)} - f_{k(i)} = \sum_{t=i}^{j-1}(f_{k(t+1)} - f_{k(t)})$$
$$= \sum_{t=i}^{j-1} \beta_{k(t)}(x_{k(t+1)} - x_{k(t)}) \geq \beta_{k(i)} \sum_{t=i}^{j-1}(x_{k(t+1)} - x_{k(t)}) = \beta_{k(i)}(x_{k(j)} - x_{k(i)})$$
$$\forall j < i,\ f_{k(j)} - f_{k(i)} = \sum_{t=j}^{i-1}(f_{k(t)} - f_{k(t+1)})$$
$$= \sum_{t=j}^{i-1} \beta_{k(t)}(x_{k(t)} - x_{k(t+1)}) \geq \beta_{k(i)} \sum_{t=j}^{i-1}(x_{k(t)} - x_{k(t+1)}) = \beta_{k(i)}(x_{k(j)} - x_{k(i)}).$$

The sparse convex additive model optimization in (5.2) is a quadratic program (QP) with $O(np)$ variables and $O(np)$ constraints. Directly applying a QP solver for $\boldsymbol{h}, \boldsymbol{\beta}$ would be computationally expensive for relatively large n and p. However, notice that variables in different feature dimensions are only coupled in the term $(Y_i - \sum_{k=1}^{p} h_{ki})^2$. Hence, we can apply the block coordinate descent method, where in each step we solve the following QP subproblem for $\{\boldsymbol{h}_{k\cdot}, \boldsymbol{\beta}_{k\cdot}\}$ with the other variables fixed. We present the one-block optimization in matrix notation:

$$\min_{\boldsymbol{f}_k, \boldsymbol{\beta}_k, \gamma_k} \frac{1}{2n}\|\boldsymbol{r}_{\text{-}k} - \boldsymbol{f}_k\|_2^2 + \lambda\gamma_k \tag{5.3}$$
$$\text{s.t. } P\boldsymbol{f}_k = \text{diag}(P\boldsymbol{x}_k)\boldsymbol{\beta}_k$$
$$D\boldsymbol{\beta}_k \leq 0, \quad -\gamma_k \mathbf{1}_n \leq \boldsymbol{f}_k \leq \gamma_k \mathbf{1}_n, \quad \mathbf{1}_n^\top \boldsymbol{f}_k = 0$$

where $\boldsymbol{f}_k \in \mathbb{R}^n$ and equals $(f_{k1}, ..., f_{kn})$, $\boldsymbol{\beta}_k \in \mathbb{R}^{n-1}$ and equals $(\beta_{k1}, ..., \beta_{k(n-1)})$. $\boldsymbol{r}_{\text{-}k} \in \mathbb{R}^n$ is the residual vector: $\boldsymbol{r}_{\text{-}k}(i) = Y_i - \overline{Y} - \sum_{k' \neq k} f_{k'i}$. $\boldsymbol{x}_k \in \mathbb{R}^n$ and equals $(x_{k1}, ..., x_{kn})$.

$P \in \mathbb{R}^{(n-1)\times n}$ is a permutation matrix where the i'-th row $P_{i'\cdot}$ is all zero except -1 at the (i')-th smallest coordinate of $\boldsymbol{x}_k$ and 1 at the $(i'+1)$-th smallest coordinate of $\boldsymbol{x}_k$. $D \in \mathbb{R}^{(n-2)\times(n-1)}$ is another permutation matrix where the i'-th row is all zero except 1 at position i' and -1 at position $i'+1$. $\text{diag}(v)$ for a vector v is a 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 (5.2), 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 (5.1),

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

The univariate concave function estimation is a straightforward modification of optimization 5.3. We need only modify one linear inequality constraint to enforce that the subgradients must be non-increasing: $\beta_{k(i+1)} \leq \beta_{k(i)}$.

5.1 Alternative Formulation

Optimization (5.2) 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 $\beta_{k(i+1)} \geq \beta_{k(i)}$ 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 $d_{k(i)} \geq 0$ for all $i > 1$.

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

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

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 (5.2) as an equivalent optimization program with only centering and positivity constraints:

$$\begin{aligned}
&\min_{d_k} \frac{1}{2n}\left\|Y - \sum_{k=1}^p \overline{\Delta}_k d_k\right\|_2^2 + \lambda_n \sum_{k=1}^p \|\overline{\Delta}_k d_k\|_\infty \qquad (5.4) \\
&\text{s.t. } d_{k(2)}, \ldots, d_{k(n-1)} \geq 0 \qquad \text{(convexity)}
\end{aligned}$$

The decoupled concave postprocessing stage optimization is again similar. Suppose $\widehat{d}_k$'s are the output of optimization 5.4, define $\widehat{r} = Y - \sum_{k=1}^p \overline{\Delta}_k \widehat{d}_k$.

$$\begin{aligned}
&\text{for all } k \text{ such that } \widehat{d}_k = 0\text{:} \\
&\min_{c_k} \frac{1}{2n}\left\|\widehat{r} - \Delta_k c_k\right\|_2^2 + \lambda_n \|\Delta_k c_k\|_\infty \qquad (5.5) \\
&\text{s.t.} c_{k(2)}, \ldots, c_{k(n-1)} \leq 0 \qquad \text{(concavity)}
\end{aligned}$$

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 5.4, 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 5.5). 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 $\{\widehat{d}_k\}_{k=1,...,p}$ that is zero for $k \in S^c$ and show that it satisfies the KKT conditions for optimality of optimization 5.4. 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 5.5 for all $k \in S^c$.

Definition 6.1. We define as *restricted regression* where we restrict the indices k in optimization (5.4) to lie in the set S instead of ranging from $1, ..., p$:

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

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 5.4. Furthermore, any solution to the optimization program 5.4 must be zero on S^c.*
2. *For all $k \in S^c$, the solution $\widehat{c}_k$ to optimization 5.5 must be 0 and must be unique.*

This result holds regardless of whether we impose the Lipschitz conditions in optimization 5.4 and 5.5 or not. The full proof of Theorem 6.1 is in Section 9.1 of the Appendix.

Theorem 6.1 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 6.1 looks at KKT conditions of optimization 5.4, 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 9.1 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 3.2), 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 5.4, and the DC solution* $\widehat{g}_k$*, from optimization 5.5 are zero.*

The proof of Theorem 6.2 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 6.2 is in Section 9.2 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 3.1. We need stronger assumptions in Theorem 6.3 to use our additive faithfulness result, Theorem 3.1. 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 6.3 in Section 9.3 of the Appendix.

Combining Theorem 6.2 and 6.3 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 (5.2). 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 [2].

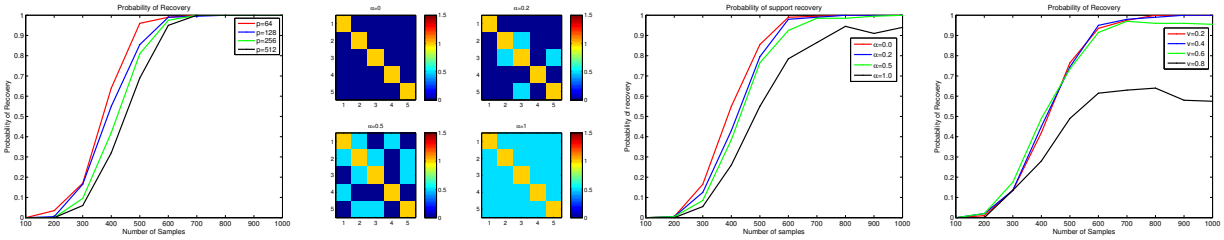
We first use all $n = 506$ samples (with normalization) to train SCAM, using a set of candidate $\{\lambda^{(t)}\}$ with $\lambda^{(1)} = 0$ (no regularization). For each $\lambda^{(t)}$ we obtain a subgradient matrix $\boldsymbol{\beta}^{(t)}$ with $p = 13$ rows. The non-zero rows in this matrix indicate 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

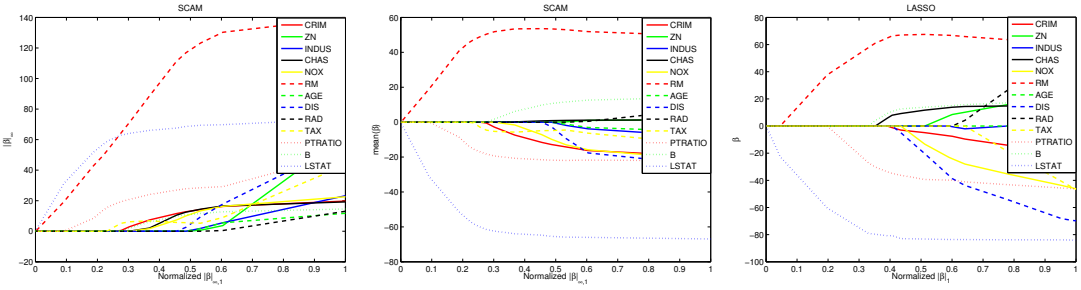
We have introduced a framework for estimating high dimensional but sparse convex functions. Because of the special properties of convexity, variable selection for convex functions enjoys additive faithfulness—it suffices to carry out variable selection over an additive model, in spite of the approximation 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

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

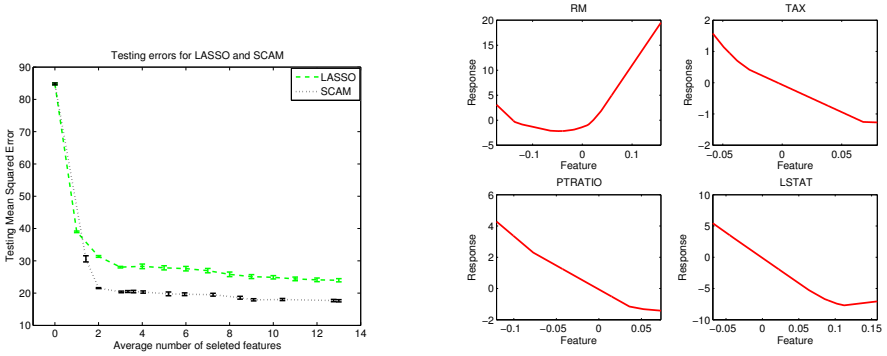


(a) additive model (b) four $\boldsymbol{Q}$ matrices (c) non-additive models (d) correlated design

Figure 2: 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 3: Results on Boston housing data, showing regularization paths, MSE and fitted functions.

challenge is to develop a method to automatically detect the concavity or convexity pattern of the variables.

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.

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 6.1 first for convenience.

Theorem 9.1. *The following holds regardless of whether we impose the Lipschitz condition in optimization 5.4 and optimization 5.5.*

Let $\{\widehat{d}_k\}_{k\in S}$ be a minimizer of the restricted regression, that is, the solution to optimization (5.4) 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 5.4. Furthermore, any solution to the optimization program 5.4 must be zero on S^c.*
2. *For all $k \in S^c$, the solution to optimization 5.5 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

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

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:

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

The second item in the theorem concerning optimization 5.5 is proven in exactly the same way.

The Lagrangian of optimization 5.5 is:

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

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 5.4, and the DC solution* $\widehat{g}_k$*, from optimization 5.5 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 9.2) 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, $|w_i| \leq c\sigma\sqrt{\log\frac{2}{\delta}}$ with probability at most $1 - \delta$. By union bound across $i = 1, ..., n$, we have that $\|w\|_\infty \leq c\sigma\sqrt{\log\frac{2}{\delta}}$ with probability at most $1 - n\delta$.

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

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

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

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 $|f_0(x_i)| \leq sLb$, 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

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

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

Step 3: We now apply Serfling's theorem.

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

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

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 $i' = 1, ..., n$:

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

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:

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

where the difference 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 $\lambda \leq csLb\sqrt{\frac{1}{n}\log^2 np}$.*
Then, with probability at least $1 - \delta$,

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

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 $\|f_k^* - \overline{f}_k^*\|_\infty \leq 4Lb$ because f_k^* as is L-Lipschitz and supported on $[-b, b]$.

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

Using Lemma 9.2, we can remove $\overline{f}^*$ from the LHS. With probability at least $1 - \delta$:

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

[TOOD: at this point, we still have to choose a value for λ]

Step 2. We upper bound $2\langle w, \widehat{f} - f^* + \overline{f}^*\rangle$ with theorem 9.5.

Define $\mathcal{G}$ as $\{f - f^* + \overline{f}^* : f \in \mathcal{C}_L^s\}$ as the set of convex additive functions centered around the function $f^* - \overline{f}^*$. Then, $\|h\|_n \le \|h\|_\infty \le 4sLb$ for all $h \in \mathcal{G}$.

According to theorem 9.5, for all $\epsilon > \frac{1}{\sqrt{n}}\sigma c \int_0^R \sqrt{\log N_2(t, \mathcal{G})}dt \vee R$,

$$P\Big(\sup_{h\in\mathcal{G}} \langle w, h\rangle_n \ge \epsilon\Big) \le 4\exp\Big(-\frac{n\epsilon^2}{cR^2\sigma^2}\Big)$$

where $R = 4sLb$ for our purpose.

Restated, we have that, with probablity at least $1 - \delta$,

$$\sup_{h\in\mathcal{G}} |\langle w, h\rangle_n| \le cR\sigma\sqrt{\frac{1}{n}\log\frac{4}{\delta}} + \Big(\int_0^R \sqrt{\log N_2(t, \mathcal{G})}dt \vee R\Big) c\sigma\sqrt{\frac{1}{n}}$$

Now we evaluate the integral. Since $N_{\|\cdot\|_n}(t, \mathcal{G}) \le N_\infty(t, \mathcal{G})$, we know that $\sqrt{\log N_{\|\cdot\|_n}(t, \mathcal{G})} \le \sqrt{Cs^{1.5}bL}t^{-1/4}$.

$$\begin{aligned}
\int_0^R \sqrt{\log N_{\|\cdot\|_n}(t, \mathcal{G})}dt &\le \sqrt{Cs^{1.5}bL}\int_0^R t^{-1/4}dt \\
&= \sqrt{Cs^{1.5}bL}\,\frac{4}{3}R^{3/4} \\
&= \sqrt{Cs^{1.5}bL}\,c(sLb)^{3/4} \\
&\le c(sbL)^2
\end{aligned}$$

Coming back, we have, with probability at least $1 - \delta$,

$$\begin{aligned}
\sup_{h\in\mathcal{G}} |\langle w, h\rangle| &\le csLb\sigma\sqrt{\frac{1}{n}\log\frac{4}{\delta}} + c(sLb)^2\sigma\sqrt{\frac{1}{n}} \\
&\le c(sLb)^2\sigma\sqrt{\frac{1}{n}\log\frac{4}{\delta}}
\end{aligned}$$

Plugging this result into equation 9.3 and using an union bound, we get, with probability at least $1 - 2\delta$:

$$\begin{aligned}
\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^*\|_n^2 &\le c(sLb)^2\sigma\sqrt{\frac{1}{n}\log\frac{4}{\delta}} + \lambda 4sLb + c(sLb)^2\frac{1}{n}\log\frac{2}{\delta} \\
\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^*\|_n^2 &\le c(sLb)^2\sigma\sqrt{\frac{1}{n}\log\frac{4}{\delta}} + \lambda 4sLb \qquad (9.4)
\end{aligned}$$

Step 3. We continue from equation 9.4, use lemma 9.1, use another union bound, with probability at least $1 - 3\delta$,

$$\begin{aligned}
\|f_0 - \widehat{f}\|_P^2 - \|f_0 - f^*\|_P^2 &\le c(sLb)^2\sigma\sqrt{\frac{1}{n}\log\frac{4}{\delta}} + \lambda 4sLb + c(Lb)^3\sqrt{\frac{s^5}{n^{4/5}}\log\frac{2}{\delta}} \\
&\le c(Lb)^3\sigma\sqrt{\frac{s^5}{n^{4/5}}\log\frac{2}{\delta}} + \lambda 4sLb \qquad (9.5)
\end{aligned}$$

Substituting in $\lambda \le csLb\sqrt{\frac{1}{n}\log^2 np}$ 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 9.3 because convex and concave functions have the same covering number.

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

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

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

Using the chaining result (theorem 9.5), setting $s = 1$, we have, with probability at least $1 - \delta$,

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

Using the uniform convergence result (lemma 9.1), we have, with probability at least $1 - 2\delta$:

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

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

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

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* $1 - \delta$*:*

$$\sup_{f \in \mathcal{C}_L^s}\left|\|f_0 - f\|_n^2 - \|f_0 - f\|_P^2\right| \leq c(Lb)^3\sqrt{\frac{s^5}{n^{4/5}}\log\frac{2}{\delta}}$$

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

For a $h \in \mathcal{G}$, let h_ϵ denote a function in the ϵ-cover of $\mathcal{G}$ closest to h. It obviously must be that $\|h - h_\epsilon\|_n \leq \|h - h_\epsilon\|_\infty \leq \epsilon$.

Because $\|h\|_n = \|h - h_\epsilon + h_\epsilon\|_n$, we have that

$$\begin{aligned}
\|h_\epsilon\|_n - \|h - h_\epsilon\|_n \leq \|h\|_n &\leq \|h_\epsilon\|_n + \|h - h_\epsilon\|_n \\
\|h_\epsilon\|_n - \epsilon \leq \|h\|_n &\leq \|h_\epsilon\|_n + \epsilon \\
\|h_\epsilon\|_n^2 - 8\epsilon(sLb) \leq \|h\|_n^2 &\leq \|h_\epsilon\|_n^2 + 8\epsilon(sLb)
\end{aligned}$$

where we used the fact that $\|h - h_\epsilon\|_n \leq \|h - h_\epsilon\|_\infty \leq \epsilon$ and $\|h_\epsilon\|_n \leq \|h_\epsilon\|_\infty \leq 4sLb$.

And likewise:

$$\|h_\epsilon\|_P^2 - 8\epsilon(sLb) \leq \|h\|_P^2 \leq \|h_\epsilon\|_P^2 + 8\epsilon(sLb)$$

Therefore,

$$\sup_{h \in \mathcal{G}}\left|\|h\|_n^2 - \|h\|_P^2\right| \leq \sup_{h_\epsilon}\left|\|h_\epsilon\|_n^2 - \|h_\epsilon\|_P^2\right| + \epsilon(16sLb)$$

Since $\|h_\epsilon\|_n^2 = \frac{1}{n}\sum_{i=1}^n h_\epsilon(X_i)^2$ is an average of bounded random variables, we have by Union Bound and Hoeffding Inequality that, with probability at most $1-\delta$,

$$\begin{aligned}\sup_{h_\epsilon}\left|\|h_\epsilon\|_n^2 - \|h_\epsilon\|_P^2\right| &\le (8sLb)^2\sqrt{\frac{1}{cn}\big(\log\frac{2}{\delta} + \log N_\infty(\epsilon, \mathcal{C}_L^s)\big)} \\ &\le (8sLb)^2\sqrt{\frac{1}{cn}\big(\log\frac{2}{\delta} + Cs^{1.5}Lb\epsilon^{-1/2}\big)}\end{aligned}$$

We will set $\epsilon = \frac{1}{n^{2/5}}(Cs^{0.5}Lb)^2$. Therefore:

$$\begin{aligned}\sup_{h_\epsilon}\left|\|h_\epsilon\|_n^2 - \|h_\epsilon\|_P^2\right| &\le (8sLb)^2\sqrt{\frac{1}{cn}\big(\log\frac{2}{\delta} + sn^{1/5}\big)} \\ &\le (8Lb)^2\sqrt{\frac{s^5}{cn^{4/5}}\log\frac{2}{\delta}}\end{aligned}$$

And

$$\begin{aligned}\sup_{h\in\mathcal{G}}\left|\|h\|_n^2 - \|h\|_P^2\right| &\le \sup_{h_\epsilon}\left|\|h_\epsilon\|_n^2 - \|h_\epsilon\|_P^2\right| + \frac{1}{n^{2/5}}C^2s^2(Lb)^3 \\ &\le (8Lb)^2\sqrt{\frac{s^5}{cn^{4/5}}\log\frac{2}{\delta}} + (CLb)^2\sqrt{\frac{s^4}{n^{4/5}}} \\ &\le c(Lb)^3\sqrt{\frac{s^5}{n^{4/5}}\log\frac{2}{\delta}}\end{aligned}$$

□

Lemma 9.2. *Let f_0, f^* be defined as in section 9.3.1. Define $\overline{f}^* = \frac{1}{n}\sum_{i=1}^n f^*(X_i)$. Then, with probability at least $1-2\delta$,*

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

Proof. (of lemma 9.2)

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

$\overline{f}^* = \frac{1}{n}\sum_{i=1}^n f^*(X_i)$ is the average of n bounded mean-zero random variables and therefore, with probability at least $1-\delta$, $|\overline{f}^*| \le 4sLb\sqrt{\frac{1}{n}\log\frac{2}{\delta}}$.

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 $1-2\delta$,

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

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

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

□

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 Covering Number for Lipschitz Convex Functions

Definition 9.1. $\{f_1, ..., f_N\} \subset \mathcal{C}[b, B, L]$ is an ϵ-covering of $\mathcal{C}[b, B, L]$ if for all $f \in \mathcal{C}[b, B, L]$, there exist f_i such that $\|f - f_i\|_\infty \leq \epsilon$.

We define $N_\infty(\epsilon, \mathcal{C}[b, B, L])$ as the size of the minimum covering.

Lemma 9.4. *(Bronshtein 1974)*

$$\log N_\infty(\epsilon, \mathcal{C}[b, B, L]) \leq C\left(\frac{bBL}{\epsilon}\right)^{1/2}$$

For some absolute constant C.

Lemma 9.5.

$$\log N_\infty(\epsilon, \mathcal{C}^s[b, B, L]) \leq Cs\left(\frac{bBLs}{\epsilon}\right)^{1/2}$$

For some absolute constant C.

Proof. Let $f = \sum_{k=1}^s f_k$ be a convex additive function. Let $\{f'_k\}_{k=1,..,s}$ be k functions from a $\frac{\epsilon}{s}$ L_∞ covering of $\mathcal{C}[b, B, L]$.

Let $f' := \sum_{k=1}^s f'_k$, then

$$\|f' - f\|_\infty \leq \sum_{k=1}^s \|f_k - f'_k\|_\infty \leq s\frac{\epsilon}{s} \leq \epsilon$$

Therefore, a product of s $\frac{\epsilon}{s}$-coverings of univariate functions induces an ϵ-covering of the additive functions. □

In our proofs, we will use the Dudley's chaining: [TODO:cite van de geer]

Theorem 9.5. *(Dudley's Chaining)*
Let $\mathcal{G} = \{g : \|g\|_n \leq R\}$. Let $M(\epsilon, R)$ be the size of the minimal ϵ-covering of $\mathcal{G}$ with respect to the $\|\cdot\|_n$ norm. Suppose $w = (w_1, ..., w_n)$ is a vector of i.i.d. subgaussian random variables with scale σ.

Suppose $\delta > 0$ is such that

$$\sqrt{n}\delta \geq \sigma\Big(14\sum_{s=0}^{\infty} 2^{-s}\sqrt{\log M(2^{-s}R, \mathcal{G})}\Big) \vee 70\log 2R$$

Then we have that

$$P\Big(\sup_{g\in\mathcal{G}}\langle w, g\rangle_n \geq \delta\Big) \leq 4\exp\Big(-\frac{n\delta^2}{(70R)^2\sigma^2}\Big)$$

For convenience, we can upper bound the metric-entropy sum with an integral:
$\sum_{s=0}^{\infty} 2^{-s}\sqrt{\log M(2^{-s}R, \mathcal{G})} \leq \int_0^R \sqrt{\log M(t, \mathcal{G})}dt$