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

November 15, 2013

Abstract

We consider the problem of estimating a sparse convex function of many variables. In constrast to classical nonparametric regression with smoothness constraints, we show that convexity is additively faithful—it suffices to estimate a convex additive model for variable selection. We develop algorithms for estimating sparse convex additive models, including an approach using iterative quadratic programming. Supporting experiments and statistical theory are presented, showing variable selection consistency in dimensions that can scale exponentially in the sample size. An attractive feature of this framework is the lack of tuning parameters for smoothness.

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.

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 $f : \boldsymbol{C} \to \mathbb{R}$.

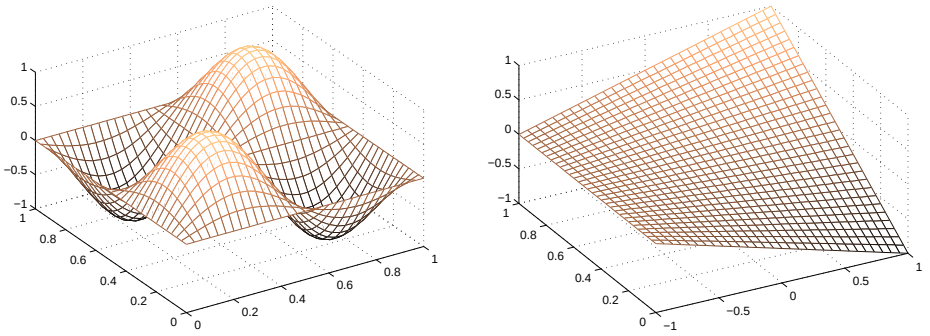
We say that f **depends on** coordinate k if, for all $x_k \in [0,1]$, the set $\left\{x_k' \in [0,1] \,:\, f(x_k, \boldsymbol{x}_{-k}) = f(x_k', \boldsymbol{x}_{-k}) \text{ for almost all } \boldsymbol{x}_{-k}\right\}$ has probability strictly less than 1.

Suppose we have the additive approximation:

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

We say that f has **additive recall** under F if $f_k^* = 0 \Rightarrow f$ does not depend on coordinate k.

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



(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.

Additive faithfulness is an attractive property because it implies that, in the population setting, the additive approximation yields consistent variable selection. Additive recall is a weaker condition but still useful: we can be sure that the additive approximation is conservative, that it does not exclude important variables.

3.1 Additive Faithfulness of Convex Functions

Remarkably, under a general class of distributions which we characterize below, convex multivariate functions has additive recall and under product distributions, convex functions are additively faithful.

Definition 3.2. Given a conditional density $p(\boldsymbol{x}_{-j} \mid x_j)$, we say that $\boldsymbol{X}_{-j}$ is **intermittently independent** of X_j if the set of x_j for which $\frac{\partial p(\boldsymbol{x}_{-j} \mid x_j)}{\partial x_j} = 0$ for all $\boldsymbol{x}_{-j}$ is either the whole line $[0, 1]$, or a collection of at least two disjoint closed intervals.

Let F be a distribution on $[0, 1]^s$. We say that F is an **intermittently independent** distribution if F has a non-zero density p with the property that, for all $j = 1, ..., s$, $\boldsymbol{X}_{-j}$ is intermittently independent of X_j.

Example 3.2. There are many examples of distributions which satisfy intermittent independence:

1. F is a product distribution.
2. F has a density p that is piece-wise constant on a gridding of $[0, 1]^s$.
3. F has a density p that is, for some $\frac{1}{2} > \epsilon > 0$, arbitrary on $(\epsilon, 1 - \epsilon)^s$ but constant elsewhere.

Theorem 3.1. *Let F be an **intermittently independent** distribution supported on $C = [0, 1]^s$ with positive density p. If f is convex and twice differentiable, then f has the* additive recall *property under F.*

Theorem 3.2. *Let F be a product distribution supported on $C = [0, 1]^s$ with positive density p. If f is convex and twice differentiable, then f is* additively faithful *under F.*

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 3.2 by showing that "slices" of convex functions that integrate to zero cannot be "glued" together while still maintaining convexity.

We give the proof of Theorem 3.1, which also proves one direction of Theorem 3.2. The other direction of Theorem 3.2 is trivial.

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.

Let us assume that for all x_k, $\mathbb{E}[f(X) - \sum_{k'} f_{k'}(X_{k'}) \,|\, x_k] - \mathbb{E}f(X) = 0$.

By the intermittent independence assumption, we know that there exists at least two disjoint closed intervals R_0, R_1 where $p(\boldsymbol{x}_{-k} \,|\, x_k)$ is independent of x_k. That is, there exist density functions $\phi_0(\boldsymbol{x}_{-k})$ and $\phi_1(\boldsymbol{x}_{-k})$ such that $p(\boldsymbol{x}_{-k} \,|\, x_k) = \phi_0(\boldsymbol{x}_{-k})$ for all $x_k \in R_0$ and $p(\boldsymbol{x}_{-k} \,|\, x_k) = \phi_1(\boldsymbol{x}_{-k})$ for all $x_k \in R_1$.

Let x_k^0 be an interior point of R_0.

For every $\boldsymbol{x}_{-k}$, we define g as the derivative of f with respect to x_k at the point $(\boldsymbol{x}_{-k}, x_k^0)$.

$$g(\boldsymbol{x}_{-k}) := \lim_{x_k \to x_k^0} \frac{f(x_k, \boldsymbol{x}_{-k}) - f(x_k^0, \boldsymbol{x}_{-k})}{x_k - x_k^0}$$

$g(\boldsymbol{x}_{-k})$ is well-defined by the assumption that f is everywhere differentiable.

Now, we use $r(\boldsymbol{x}_{-k})$ as a short-hand for $\sum_{k' \neq k} f_{k'}(x_{k'})$ and derive that for all $x_k \in R_0$, we have:

$$\begin{aligned}
&\int_{\boldsymbol{x}_{-k}} \phi_0(\boldsymbol{x}_{-k}) \Big(f(x_k, \boldsymbol{x}_{-k}) - f(x_k^0, \boldsymbol{x}_{-k}) \Big) d\boldsymbol{x}_{-k} \\
&= \int_{\boldsymbol{x}_{-k}} \phi_0(\boldsymbol{x}_{-k}) \Big(\big(f(x_k, \boldsymbol{x}_{-k}) - r(\boldsymbol{x}_{-k}) - \mu^* \big) - \big(f(x_k^0, \boldsymbol{x}_{-k}) - r(\boldsymbol{x}_{-k}) - \mu^* \big) \Big) d\boldsymbol{x}_{-k} \\
&= 0
\end{aligned}$$

The last equality follows because we assumed that $\mathbb{E}[f(X) - \sum_{k'} f_{k'}(X_{k'}) \,|\, x_k] - \mathbb{E}f(X) = 0$ for all x_k.

Since x_k^0 is in the interior of R_0, and since the domain of f is a compact set, we use the bounded convergence theorem and conclude that $\int_{\boldsymbol{x}_{-k}} \phi_0(\boldsymbol{x}_{-k}) g(\boldsymbol{x}_{-k}) = 0$ as well.

We now note that $g(\boldsymbol{x}_{-k})$ is a component of the subgradient $\partial_{\boldsymbol{x}} f(x_k^0, \boldsymbol{x}_{-k})$.

Therefore, the first-order characterization of convex functions gives us that

$$f(x_k, \boldsymbol{x}_{-k}) \geq f(x_k^0, \boldsymbol{x}_{-k}) + g(\boldsymbol{x}_{-k})(x_k - x_k^0) \quad \text{for all } x_k, \boldsymbol{x}_{-k}$$

Or, equivalently, for all $x_k, \boldsymbol{x}_{-k}$,

$$f(x_k, \boldsymbol{x}_{-k}) - f(x_k^0, \boldsymbol{x}_{-k}) - g(\boldsymbol{x}_{-k})(x_k - x_k^0) \geq 0$$

Now, the critical observation we make is that, for all $x_k \in R_0$,

$$\int_{\boldsymbol{x}_{-k}} \phi_0(\boldsymbol{x}_{-k}) \Big(f(x_k, \boldsymbol{x}_{-k}) - f(x_k^0, \boldsymbol{x}_{-k}) - g(\boldsymbol{x}_{-k})(x_k - x_k^0) \Big) d\boldsymbol{x}_{-k} = 0$$

Since ϕ_0 is positive by assumption and the second term is non-negative, we conclude that the second term must be exactly zero, that is, for all $x_k \in R_0$, for all $\boldsymbol{x}_{-k}$,

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

The Hessian of f at $(x_k, \boldsymbol{x}_{-k})$ for $x_k \in R$ 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 $g(\boldsymbol{x}_{-k})$ must be a constant function. Since g

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

integrates to 0 under $\phi_0(\boldsymbol{x}_{-k})$, it must be 0.

Thus, $f(x_k, \boldsymbol{x}_{-k}) = f(x_k^0, \boldsymbol{x}_{-k})$ for all $x_k \in R_0$ and for all $\boldsymbol{x}_{-k}$.

We can go through the same reasoning for R_1 and get that $f(x_k, \boldsymbol{x}_{-k})$ is a constant function of x_k for all $x_k \in R_1$, that is, $f(x_k, \boldsymbol{x}_{-k}) = f(x_k^1, \boldsymbol{x}_{-k})$ for all $x_k \in R_1$ for some interior point x_k^1 of R_1.

However, if $f(x_k, \boldsymbol{x}_{-k})$, as an one-dimensional convex function of x_k, is constant on two disjoint closed intervals, it must be constant for all $x_k \in [0, 1]$. Therefore, for all $\boldsymbol{x}_{-k}$, $f(x_k, \boldsymbol{x}_{-k})$ is a constant as a function of x_k and f does not depend on x_k.

□

Theorem 3.2 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 3.2 and 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. With only intermittent independence assumption, it is difficult to prove the other 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. This is, of course, unlikely to occur in practice and we leave as future work imposing reasonable assumptions that rules out such psychotic phenomenon.

4 Optimization for Sparse Convex Additive Models

We now consider the following nonparametric regression problem

$$Y_i = f(\boldsymbol{x}_i) + \epsilon_i = \sum_{k=1}^{p} f_k(x_{ki}) + \mu + \epsilon_i \quad i = 1, 2, \cdots, n$$

where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the covariate, Y_i is the response and ϵ_i is mean zero noise. The regression function $f(\cdot)$ 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 a univariate convex function, which can be represented by its supporting hyperplanes, i.e.,

$$h_{kj} \geq h_{ki} + \beta_{ki}(x_{kj} - x_{ki}) \quad (\forall i, j) \tag{4.1}$$

where $h_{ki} := f_k(x_{ki})$ 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:

$$\begin{aligned} \min_{\boldsymbol{h}, \boldsymbol{\beta}, \mu} \ & \frac{1}{2n} \sum_{i=1}^{n} \Big(Y_i - \sum_{k=1}^{p} h_{ki} - \mu \Big)^2 + \lambda \sum_{k=1}^{p} \|\boldsymbol{\beta}_{k\cdot}\|_\infty \\ \text{subject to } \ & h_{k(i+1)} = h_{k(i)} + \beta_{k(i)}(x_{k(i+1)} - x_{k(i)}), \\ & \sum_{i=1}^{n} h_{ki} = 0, \\ & \beta_{k(i+1)} \geq \beta_{k(i)} \ (\forall k, i) \end{aligned} \tag{4.2}$$

Here $\{(1),(2),\ldots,(n)\}$ is a reordering of $\{1,2,\ldots,n\}$ such that $x_{k(1)} \le x_{k(2)} \le \cdots \le x_{k(n)}$. 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 $\sum_i h_{ki} = 0$. It is easy to verify that the constraints in (4.2) satisfy the supporting hyperplane constraints, as

$$\begin{aligned}
\forall j \ge i,\ h_{k(j)} - h_{k(i)} &= \sum_{t=i}^{j-1}(h_{k(t+1)} - h_{k(t)}) \\
&= \sum_{t=i}^{j-1}\beta_{k(t)}(x_{k(t+1)} - x_{k(t)}) \ge \beta_{k(i)}\sum_{t=i}^{j-1}(x_{k(t+1)} - x_{k(t)}) \\
&= \beta_{k(i)}(x_{k(j)} - x_{k(i)}) \\
\forall j < i,\ h_{k(j)} - h_{k(i)} &= \sum_{t=j}^{i-1}(h_{k(t)} - h_{k(t+1)}) \\
&= \sum_{t=j}^{i-1}\beta_{k(t)}(x_{k(t)} - x_{k(t+1)}) \ge \beta_{k(i)}\sum_{t=j}^{i-1}(x_{k(t)} - x_{k(t+1)}) \\
&= \beta_{k(i)}(x_{k(j)} - x_{k(i)}).
\end{aligned}$$

The ℓ_∞/ℓ_1 penalty $\sum_{k=1}^p \|\boldsymbol{\beta}_{k\cdot}\|_\infty$ encourages group sparsity of the vectors $\boldsymbol{\beta}_{k\cdot}$, and thus performs variable selection. We refer to this framework as the sparse convex additive model (SCAM). While one can use supporting hyperplanes to the epigraph as in (1.2), SCAM uses the *inner piece-wise linear function* that approximates the graph with secant lines. Notice that if we replace $\beta_{k(i+1)} \ge \beta_{k(i)}$ with $\beta_{k(i+1)} = \beta_{k(i)}$, the optimization reduces to the lasso.

The SCAM optimization in (4.2) 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:

$$\begin{aligned}
\min_{\boldsymbol{h}_{k\cdot},\boldsymbol{\beta}_{k\cdot},\gamma_k}\ & \frac{1}{2n}\sum_{i=1}^n\Big((Y_i - \overline{Y} - \sum_{r\ne k} h_{ri}) - h_{ki}\Big)^2 + \lambda\gamma_k \\
\text{such that } & h_{k(i+1)} = h_{k(i)} + \beta_{k(i)}(x_{k(i+1)} - x_{k(i)}), \\
& \beta_{k(i+1)} \ge \beta_{k(i)},\ -\gamma_k \le \beta_{k(i)} \le \gamma_k \\
& \sum_{i=1}^n h_{ki} = 0,\ (\forall i).
\end{aligned}$$

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 (4.2), 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 (4.1),

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

4.1 Alternative Formulation

Optimization (4.2) 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 $\beta_{k(i)} = \sum_{j \leq i} d_{k(i)}$ and

$$\begin{aligned} f_k(x_{k(i)}) =& f_k(x_{k(1)}) + d_{k(1)}(x_{k(i)} \\ & - x_{k(1)}) + d_{k(2)}(x_{k(i)} - x_{k(2)}) + \cdots \\ & + d_{k(i-1)}(x_{k(i)} - x_{k(i-1)}) \end{aligned} \tag{4.3}$$

We can write this more compactly in matrix notations. First define $\Delta_{k(j)}(x_{ki}) = \max(x_{ki} - x_{k(j)}, 0)$.

$$(f_k(x_{k1}), ..., f_k(x_{kn}))^\mathsf{T} = \Delta_k d_k :=$$
$$\begin{bmatrix} \Delta_{k(1)}(x_{k1}) & ... & \Delta_{k(n-1)}(x_{k1}) \\ ... & & \\ \Delta_{k(1)}(x_{kn}) & ... & \Delta_{k(n-1)}(x_{kn}) \end{bmatrix} \begin{bmatrix} d_{k(1)} \\ ... \\ d_{k(n-1)} \end{bmatrix}$$

Where Δ_k is a $n \times n-1$ matrix such that $\Delta_k(i,j) = \Delta_{k(j)}(x_{ki})$ and $d_k = (d_{k(1)}, ..., d_{k(n-1)})$. We can now reformulate (4.2) as an equivalent optimization program with only centering and positivity constraints:

$$\begin{aligned} &\min_{d_k, c_k} \frac{1}{2n} \Big\| Y - \overline{Y}\mathbf{1}_n - \sum_{k=1}^{p} (\Delta_k d_k - c_k \mathbf{1}_n) \Big\|_2^2 + \lambda_n \sum_{k=1}^{p} \|d_k\|_1 \\ &\text{s.t. } d_{k(2)}, \ldots, d_{k(n-1)} \geq 0 \qquad \text{(convexity)} \\ &\quad c_k = \frac{1}{n} \mathbf{1}_n^\mathsf{T} \Delta_k d_k \qquad \text{(centering)} \end{aligned} \tag{4.4}$$

$\|d_k\|_1$ is not identical to $\|\boldsymbol{\beta}_{k\cdot}\|_\infty$, but it is easy to verify that $\|\boldsymbol{\beta}_{k\cdot}\|_\infty \leq \|d_k\|_1 \leq 4\|\boldsymbol{\beta}_{k\cdot}\|_\infty$.

Remark 4.1. For parts of our theoretical analysis, we will also impose onto (4.4) a boundedness constraint $-B\mathbf{1}_n \leq \Delta_k d_k + c_k \mathbf{1}_n \leq B\mathbf{1}_n$ which constrains that $\|f_k\|_\infty \leq B$, or a Lipschitz constraint $\|d_k\|_1 \leq L$ which constrains that f_k must be L-Lipschitz. We use these constraints only in the proof for technical reasons; we never need nor use these constraints in our experiments.

4.2 Convex-plus-concave Functions

Although convex regression enjoys good properties from both statistical and optimization perspective, convexity can be too strong of a shape constraint for some nonparametric regression problems. We can in fact significantly relax the shape constraint in the same optimization framework by learning a regression function f that is a sum of a convex and a concave functions. That is, there exist a convex function g and a concave function h such that $f = g + h$.

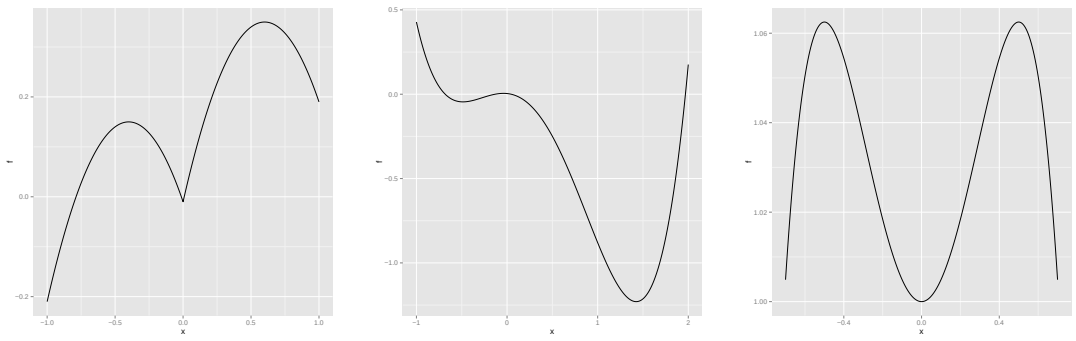
To learn a convex-plus-concave function, we need only modify the objective in optimization 4.2 to be

$$\min_{h,g,\boldsymbol{\beta},\boldsymbol{\theta},\mu} \frac{1}{2n} \sum_{i=1}^{n} \Big(Y_i - \sum_{k=1}^{p} (h_{ki} + g_{ki}) - \mu \Big)^2 + \lambda \sum_{k=1}^{p} \|\boldsymbol{\beta}_{k\cdot}\|_\infty + \lambda \sum_{k=1}^{p} \|\boldsymbol{\theta}_{k\cdot}\|_\infty$$

and add linear constraints to enforce the concavity of the g_k's with $\boldsymbol{\theta}$ representing the secants of g_k. The optimization is still convex and efficient; the runtime is on the same order as that of convex additive regression.

Convex-plus-concavity is a much less stringent condition than convexity; as shown in Figure 2, it encompasses a variety of interestingly shaped functions. And because convex-plus-concave functions can be efficiently fitted without the need to tune a smoothing parameter, it is an attract alternative to the traditional smoothing kernel approach in nonparametric regression.

Convex-plus-concave functions unfortunately do not inherit equally good additive faithfulness properties; the tilting slope example $f(x_1, x_2) = x_1 x_2 = \frac{1}{2}(x_1 + x_2)^2 - \frac{1}{2}(x_1^2 + x_2^2)$ is a concave-plus-convex function.



(a) $|x| - (x - 0.1)^2$ (b) $0.5(x - 0.3)^4 - x^2$ (c) $0.5x^2 + 1 - x^4$

Figure 2: Examples of one-dimensional concave-plus-convex functions.

5 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.

5.1 Deterministic Setting

We follow Wainwright [2009] and define the *restricted regression* purely for theoretical purposes.

Definition 5.1. In *restricted regression*, we restrict the indices k in optimization (4.4) to lie in the support S instead of ranging from $1, ..., p$.

Our analysis then differs from the now-standard "primal-dual witness technique" Wainwright [2009]. Primal-dual witness explicitly solves all the dual variables, but because our optimization is more complex, we do not solve the dual variables on S; we instead write the dual variables on S^c as a function of the restricted regression *residual*, which is implicitly a function of the dual variables on S.

Theorem 5.1. *(Deterministic setting) Let $\{\widehat{d}_k, \widehat{c}_k\}_{k \in S}$ be the minimizer of the restricted regression, that is, the solution to optimization (4.4) where we restrict $k \in S$. Let $\widehat{d}_k = 0$ and $\widehat{c}_k = 0$ for $k \in S^c$. Let $\widehat{r} := Y - \overline{Y}\mathbf{1}_n - \sum_{k \in S}(\Delta_k \widehat{d}_k - \widehat{c}_k \mathbf{1}_n)$ be the restricted regression residual. For $k \in \{1, ..., p\}$, Let $\Delta_{k,j} \in R^n$ be the j-th column of Δ_k, i.e. $\max(X_k - X_{k(j)}\mathbf{1}_n, 0)$.*

Suppose for all j and all $k \in S^c$, $\lambda_n > |\frac{1}{n}\widehat{r}^{\mathsf{T}}\Delta_{k,j}|$. Then $\widehat{\mu}$ and $\widehat{d}_k, \widehat{c}_k$ for $k = 1, ..., p$ is an optimal solution to the full regression 4.4. Furthermore, any solution to the optimization program 4.4 must be zero on S^c.

This result holds regardless of whether we impose the boundedness and Lipschitz conditions in optimization 4.4. The full proof of Theorem 5.1 is in Section 8.1 of the Appendix.

Remark 5.1. The incoherence condition of Wainwright [2009] is implicitly encoded in our condition on $\lambda_n, \widehat{r}, \Delta_{k,j}$. We can reconstruct the incoherence condition if we assume that the true function f_0 is linear and that our fitted functions $\widehat{f}_k$ are linear as well.

Theorem 5.1 allows us to analyze false negative rates and false positive rates separately. To control false positives, we study when the condition $\lambda_n > |\frac{1}{n}\widehat{r}^{\mathsf{T}}\Delta_{k,j}|$ is fulfilled for all j and all $k \in S^c$. To control false negatives, we study the restricted regression.

5.2 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 $Y = f_0(X) + \epsilon$ where ϵ is zero-mean noise. Let $Y^{(1)}, ..., Y^{(n)}$ be iid.
3. Let $S = \{1, ..., s\}$ denote the relevant variables where $s \leq p$, i.e., $f_0(X) = f_0(X_S)$.
4. Let $f_1^*, ..., f_s^* := \arg\min_{f_1,...,f_s}\{\mathbb{E}\Big(f_0(X) - \mathbb{E}f_0(X) - \sum_{k=1}^s f_k(X_k)\Big)^2 \mid \mathbb{E}[f_k(X_k)] = 0\}$.

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

A1: X_S, X_{S^c} are independent. A1': $\{X_k\}_{k\in S}$ are independent.

A2: $\|f_0\|_\infty \leq sB$ A2': f_0 is convex, twice-differentiable, and L-Lipschitz.

A3: Suppose ϵ is mean-zero sub-Gaussian, independent of X, with sub-Gaussian scale σ, i.e. for all $t \in \mathbb{R}$, $\mathbb{E}e^{t\epsilon} \leq e^{\sigma^2 t^2/2}$.

A4: For all $k = 1, ..., s$, $\mathbb{E}(f_s^*(X_k))^2 \geq \alpha$ for some positive constant α.

We will use assumptions A1, A2, A3 to control the probability of false positives and the stronger assumptions A1', A2', A3, A4 to control the probability of false negatives. Assumption A4 can be weakened so that the relevant functions satisfy $\mathbb{E}(f_s^*(X_k))^2 \geq \alpha_n$ for α_n decaying to zero at an appropriate rate.

Remark 5.2. 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 these assumptions is an interesting direction for future work.

Remark 5.3. Assumption A4 ensures that the relevant variables are "relevant enough". Under A4, the population risk of an additive function with $s-1$ components is at least α larger than the population risk of the optimal additive function with s components. See lemma 8.1 in section 8.3 of the appendix.

Theorem 5.2. *(Controlling false positives) Suppose assumptions A1, A2, A3 hold. Suppose also that we run optimization* (4.4) *with the B-boundness constraint. Let c, C be absolute constants. Suppose $\lambda_n \geq cb(sB + \sigma)\sqrt{\frac{s}{n}\log n \log(pn)}$. Then with probability at least $1 - \frac{C}{n}$, for all j, k, $\lambda_n > |\frac{1}{n}\widehat{r}^{\mathsf{T}}\Delta_{k,j}|$. Therefore, any solution to the full regression (4.4), with boundedness constraint, is zero on S^c.*

The proof of Theorem 5.2 exploits independence of $\widehat{r}$ and $\Delta_{k,j}$ from A1, and then uses concentration of measure results to argue that $|\frac{1}{n}\widehat{r}^{\mathsf{T}}\Delta_{k,j}|$ concentrates around zero at a desired rate. The fact that $\widehat{r}$ is a centered vector is crucial to our proof, and our theory thus further illustrates the importance of imposing the centering constraints in optimization (4.4). Our proof uses the concentration of the average of data sampled *without* replacement Serfling [1974]. The full proof of Theorem 5.2 is in Section 8.2 of the Appendix.

Theorem 5.3. *(Controlling false negatives) Suppose assumptions A1', A2', A3, A4 hold. Let* $\widehat{f} = \{\widehat{d}_k, \widehat{c}_k\}_{k \in S}$ *be any solution to the restricted regression with both the B-boundedness and L-Lipschitz constraint. Let* c, C *be absolute constants. Suppose* $sL\lambda_n \to 0$ *and* $Lb(sB + \sigma)sB\sqrt{\frac{s}{n^{4/5}} \log sn} \to 0$. *Then, for sufficiently large* n, $\widehat{f}_k = (\widehat{d}_k, \widehat{c}_k) \neq 0$ *for all* $k \in S$ *with probability at least* $1 - \frac{C}{n}$.

This is a finite sample version of Theorem 3.2. We need stronger assumptions in Theorem 5.3 to use our additive faithfulness result, Theorem 3.2. 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 5.3 in Section 8.3 of the Appendix.

Combining Theorem 5.2 and 5.3 and ignoring dependencies on b, B, L, σ, we have the following result.

Corollary 5.1. *Assume A1', A2', A3, A4. Let* $\lambda_n = \Theta\left(\sqrt{\frac{s^3}{n} \log n \log(pn)}\right)$. *Suppose* $s\lambda_n \to 0$ *and* $\sqrt{\frac{s^5}{n^{4/5}} \log sn} \to 0$. *Let* $\widehat{f_n}$ *be a solution to (4.4) with boundedness and Lipschitz constraints. Then* $\mathbb{P}(supp(\widehat{f_n}) = supp(f_0)) \to 1$.

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 5.4. 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.

6 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 (4.2). 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 3(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 3(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 3(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 $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ instead of $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$, 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 3(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].

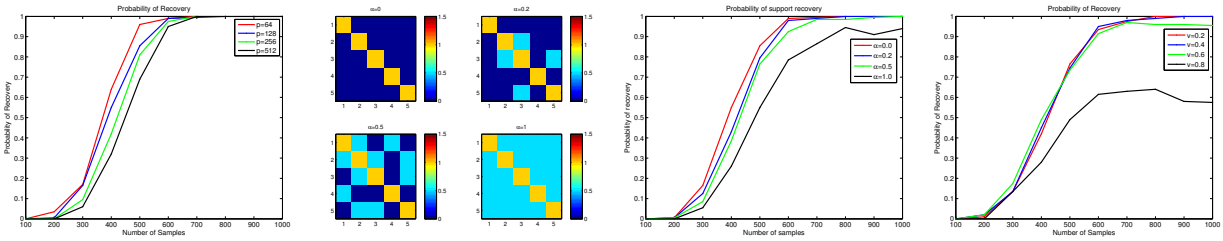
We first use all $n = 506$ samples (with normalization) to train SCAM, using a set of candidate $\{\lambda^{(t)}\}$ with $\lambda^{(1)} = 0$ (no regularization). For each $\lambda^{(t)}$ we obtain a subgradient matrix $\boldsymbol{\beta}^{(t)}$ with $p = 13$ rows. The non-zero rows in this matrix indicate 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 4(a) and 4(b). As a comparison we plot the LASSO/LARS result in a similar way in Figure 4(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 4(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 4(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.

7 Discussion

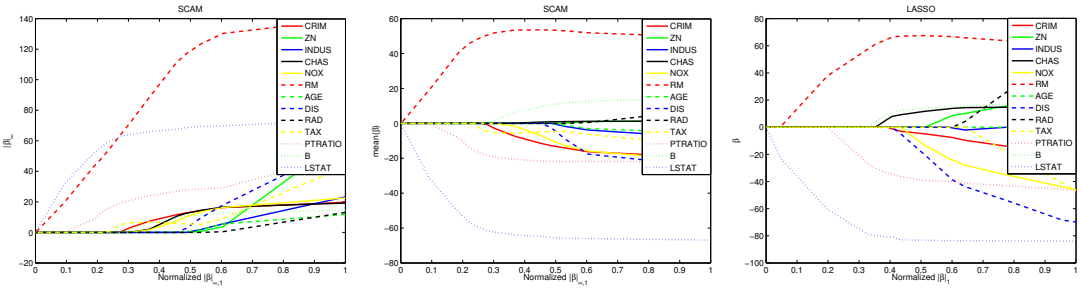
We have introduced a framework for estimating high dimensional but sparse convex functions. Because of the special properties of convexity, variable selection for convex functions enjoys additive faithfulness—it suffices to carry out variable selection over an additive model, in spite of the approximation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

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

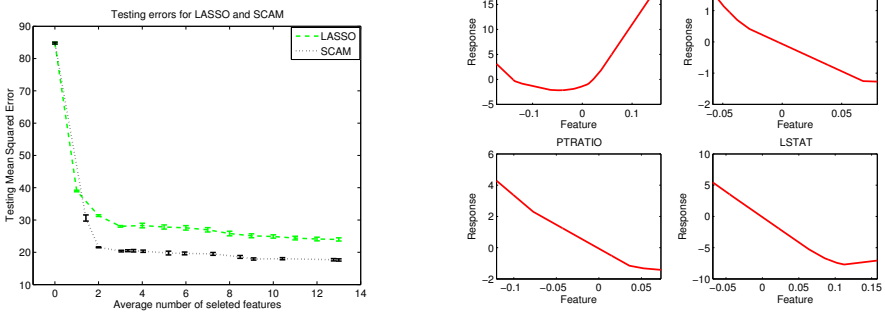


(a) additive model (b) four $\boldsymbol{Q}$ matrices (c) non-additive models (d) correlated design

Figure 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.

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 Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5): 2183–2202, May 2009.

8 Appendix

8.1 Proof of the Deterministic Condition for Sparsistency

We restate Theorem 5.1 first for convenience.

Theorem 8.1. *The following holds regardless of whether we impose the boundedness and smoothness condition in optimization 4.4 or not.*

For $k \in \{1, ..., p\}$, let $\Delta_{k,j}$ denote the n-dimensional vector $\max(X_k - X_{k(j)}\mathbf{1}, 0)$.

Let $\{\widehat{d}_k, \widehat{c}_k\}_{k \in S}$ be the minimizer of the restricted regression optimization program 4.4. Let $\widehat{d}_k = 0$ and $\widehat{c}_k = 0$ for $k \in S^c$.

Let $\widehat{r} := Y - \overline{Y}\mathbf{1}_n - \sum_{k \in S}(\Delta_k \widehat{d}_k - \widehat{c}_k \mathbf{1}_n)$ be the residual.

Suppose for all $j = 1, ..., n, k \in S^c$, $\lambda_n > |\frac{1}{n}\widehat{r}^\mathsf{T}\Delta_{k,j}|$, then $\widehat{d}_k, \widehat{c}_k$ for $k = 1, ..., p$ is an optimal solution to the full regression 4.4.

Furthermore, any solution to the optimization program 4.4 must be zero on S^c.

Proof. We will omit the boundedness and smoothness constraints in our proof here. It is easy to add those in and check that the result of the theorem still holds.

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

we can re-write the Lagrangian L, in term of just d_k, c_k, as the following.

$$\min_{d_k, c_k} \frac{1}{2n}\|r_k - \Delta_k d_k + c_k \mathbf{1}\|_2^2 + \lambda \sum_{i=2}^{n} d_{ki} + \lambda |d_{k1}| - \mu_k^\mathsf{T} d_k + \gamma_k (c_k - \mathbf{1}_n^\mathsf{T} \Delta_k d_k)$$

where $r_k := Y - \overline{Y}\mathbf{1}_n - \sum_{k' \in S, k' \neq k}(\Delta_{k'} d_{k'} - c_{k'}\mathbf{1}_n)$, and $\mu_k \in \mathbb{R}^{n-1}$ is a vector of dual variables where $\mu_{k,1} = 0$ and $\mu_{k,i} \geq 0$ for $i = 2, ..., n-1$.

First, note that by definition as solution of the restricted regression, for $k \in S$, $\widehat{d}_k, \widehat{c}_k$ satisfy stationarity with dual variables that satisfy complementary slackness.

Now, let us fix $k \in S^c$ and prove that $\widehat{d}_k = 0, \widehat{c}_k = 0$ is an optimal solution.

$$\begin{aligned} \partial d_k : \quad & -\frac{1}{n}\Delta_k^\mathsf{T}(r_k - \Delta_k d_k + c_k \mathbf{1}) + \lambda \boldsymbol{u}_k - \mu_k - \gamma_k \Delta_k^\mathsf{T}\mathbf{1} \\ \partial c_k : \quad & -\frac{1}{n}\mathbf{1}^\mathsf{T}(r_k - \Delta_k d_k + c_k \mathbf{1}) + \gamma_k \end{aligned}$$

In the derivatives, $\boldsymbol{u}_k$ is a $(n-1)$-vector whose first coordinate is $\partial|d_{k1}|$ and all other coordinates are 1.

We now substitute in $d_k = \widehat{d}_k = 0, c_k = \widehat{c}_k = 0, r_k = \widehat{r}_k = \widehat{r}$ and show that the duals can be set in a way to ensure that the derivatives are equal to 0.

$$\begin{aligned} -\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r} + \lambda \boldsymbol{u}_k - \mu_k - \gamma_k \Delta_k^\mathsf{T}\mathbf{1} = 0 \\ -\frac{1}{n}\mathbf{1}^\mathsf{T}\widehat{r} + \gamma_k = 0 \end{aligned}$$

where $\boldsymbol{u}_k$ is 1 in every coordinate except the first, where it can take any value in $[-1, 1]$.

First, we observe that $\gamma_k = 0$ because $\widehat{r}$ has empirical mean 0. All we need to prove then is that

$$\lambda \boldsymbol{u}_k - \mu_k = \frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}.$$

Suppose

$$\lambda \mathbf{1} > |\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}|,$$

then we easily see that the first coordinate of $\boldsymbol{u}_k$ can be set to some value in $(-1, 1)$ and we can set $\mu_{k,i} > 0$ for $i = 2, ..., n-1$.

Because we have strict inequality in the above equation, Lemma 1 from Wainwright [2009] show that all solutions must be zero on S^c. □

8.2 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 8.2. *Suppose assumptions A1, A2, A3 hold.*

Suppose $\lambda_n \geq cb(sB + \sigma)\sqrt{\frac{s}{n}\log n \log(pn)}$*, then with probability at least* $1 - \frac{C}{n}$*, for all* $j = 1, ..., n, k \in S^c$*,*

$$\lambda_n > |\frac{1}{n}\widehat{r}^\mathsf{T}\Delta_{k,j}|$$

And therefore, the solution to the optimization 4.4 is zero on S^c*.*

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.

We remind the reader that $\Delta_{k,j} = \max(X_k, -X_{k(j)}\mathbf{1}_n, 0)$. Because $\widehat{r}$ is empirically centered,

$$\begin{aligned}\frac{1}{n}\widehat{r}^\mathsf{T}\Delta_{k,j} &= \frac{1}{n}\widehat{r}^\mathsf{T}\max(X_k, X_{k(j)}\mathbf{1}_n) - \frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}_n X_{k(j)} \\ &= \frac{1}{n}\widehat{r}^\mathsf{T}\max(X_k, X_{k(j)}\mathbf{1}_n)\end{aligned}$$

Our goal in this proof is to bound $\frac{1}{n}\widehat{r}^\mathsf{T}\max(X_k, X_{k(j)})$ from above.

Step 1. We first get a high probability bound on $\|\widehat{r}\|_\infty$.

$$\begin{aligned}\widehat{r}_i &= Y_i - \overline{Y} - \sum_{k\in S}\widehat{f}_k(X_k^{(i)}) \\ &= f_0(X_S^{(i)}) + \epsilon_i - \overline{f_0} - \overline{\epsilon} - \sum_{k\in S}\widehat{f}_k(X_k^{(i)}) \\ &= f_0(X_S^{(i)}) - \overline{f_0} - \sum_{k\in S}\widehat{f}_k(X_k^{(i)}) + \epsilon_i - \overline{\epsilon}\end{aligned}$$

Where $\overline{f_0} = \frac{1}{n}\sum_{i=1}^n f_0(X_S^{(i)})$ and likewise for $\overline{\epsilon}$.

ϵ_i is subgaussian with subgaussian norm σ. For a single ϵ_i, we have that $P(|\epsilon_i| \geq t) \leq C\exp(-c\frac{1}{\sigma^2}t^2)$. Therefore, with probability at least $1-\delta$, $|\epsilon_i| \leq \sigma\sqrt{\frac{1}{c}\log\frac{C}{\delta}}$.

By union bound, with probability at least $1-\delta$, $\max_i|\epsilon_i| \leq \sigma\sqrt{\frac{1}{c}\log\frac{2nC}{\delta}}$.

Also, $|\overline{\epsilon}| \leq \sigma\sqrt{\frac{c}{n}\log\frac{C}{\delta}}$ with probability at least $1-\delta$.

We know that $|f_0(x)| \leq sB$ and $|\widehat{f}_k(x_k)| \leq B$ for all k.

Then $|\overline{f_0(x)}| \leq sB$ as well, and $|f^*(X_S^{(i)}) - \overline{f^*} - \sum_{k\in S}\widehat{f}_k(X_k^{(i)})| \leq 3sB$.

Therefore, taking an union bound, we have that with probability at least $1 - \frac{C}{n}$,

$$\|\widehat{r}\|_\infty \leq (3sB + c\sigma\sqrt{\log n})$$

Step 2. We now bound $\frac{1}{n}\widehat{r}^\mathsf{T}\max(X, X_{k(j)}\mathbf{1})$.

$$\frac{1}{n}\widehat{r}^\mathsf{T}\max(X_k, X_{k(j)}\mathbf{1}) = \frac{1}{n}\sum_{i=1}^{n}\widehat{r}_i\max(X_{ki}, X_{k(j)}) = \frac{1}{n}\sum_{i=1}^{n}\widehat{r}_i X_{ki}\delta(ord(i)\geq j) + \frac{1}{n}X_{k(j)}\mathbf{1}_A^\mathsf{T}\widehat{r}_A$$

Where $A = \{i : ord(i) \geq j\}$ and $ord(i)$ is the order of sample i where (1) is the smallest element. We will bound both terms.

Term 1.

$$\text{Want to bound} \qquad F(X_{k1}, ..., X_{kn}) := \frac{1}{n}\sum_{i=1}^{n}\widehat{r}_i X_{ki}\delta(ord(i)\geq j)$$

First, we note that X_{ki} is bounded in the range $[-b, b]$.

We claim then that F is coordinatewise-Lipschitz. Let $X_k = (X_{k1}, X_{k2}, ..., X_{kn})$ and $X'_k = (X'_{k1}, X_{k2}, ..., X_{kn})$ differ only on the first coordinate.

The order of coordinate i in X_k and X'_k can change by at most 1 for $i \neq 1$. Therefore, of the $j-1$ terms of the series, at most 2 terms differ from $F(X_k)$ to $F(X'_k)$ and

$$|F(X_{k1}, ..., X_{kn}) - F(X'_{k1}, ..., X'_{kn})| \leq \frac{4b\|\widehat{r}\|_\infty}{n}$$

By McDiarmid's inequality therefore,

$$P(|F(X_k) - \mathbb{E}F(X_k)| \geq t) \leq C\exp(-cn\frac{t^2}{(4b\|\widehat{r}\|_\infty)^2})$$

By symmetry and the fact that $\widehat{r}$ is centered, $\mathbb{E}F(X_k) = 0$.

We can fold the 4 into the constant c. With probability $1-\delta$, $|F(X_k)| \leq b\|\widehat{r}\|_\infty\sqrt{\frac{1}{cn}\log\frac{C}{\delta}}$.

Term 2:

$$\text{Want to bound}\frac{1}{n}X_{k(j)}\mathbf{1}_A^\mathsf{T}\widehat{r}_A$$

A is a random set and is probabilistically independent of $\widehat{r}$. $\mathbf{1}_A^\mathsf{T}\widehat{r}_A$ is the sum of a sample of $\widehat{r}$ without replacement. Therefore, according to Serfling's theorem (Corollary 8.2), with probability at least $1-\delta$, $|\frac{1}{n}\mathbf{1}_A^\mathsf{T}\widehat{r}_A|$ is at most $\|\widehat{r}\|_\infty\sqrt{\frac{1}{cn}\log\frac{C}{\delta}}$.

Since $|X_{k(j)}|$ is at most b, we obtain that with probability at least $1-\delta$, $|\frac{1}{n}X_{k(j)}\mathbf{1}_A^\mathsf{T}\widehat{r}_A| \leq b\|\widehat{r}\|_\infty\sqrt{\frac{1}{cn}\log\frac{C}{\delta}}$.

Now we put everything together.

Taking union bound across p and n, we have that with probability at least $1-\delta$,

$$|\frac{1}{n}\max(X_k, X_{k(j)}\mathbf{1})^\mathsf{T}\widehat{r}| \leq b\|\widehat{r}\|_\infty\sqrt{\frac{1}{c}\frac{1}{n}\log\frac{npC}{\delta}}$$

Taking union bound and substituting in the probabilistic bound on $\|\widehat{r}\|_\infty$, we get that with probability at least $1-\frac{C}{n}$,

$|\frac{1}{n}\max(X_k, X_{k(j)}\mathbf{1})^\mathsf{T}\widehat{r}|$ is at most

$$cb(sB+\sigma)\sqrt{\frac{s}{n}\log n\log(pn)}$$

□

8.3 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 will use covering number and uniform convergence and will thus need to first introduce some notations.

8.3.1 Notation

Given samples $X^{(1)}, ..., X^{(n)}$, let f, g be a function and w be a n-dimensional random vector, then we denote $\|f - g + w\|_n^2 := \frac{1}{n}\sum_{i=1}^n (f(X^{(i)}) - g(X^{(i)}) + w_i)^2$. We will also abuse notation and let $\|f + c\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f(X^{(i)}) + c)^2$ if c is a scalar.

We let $\langle f, g\rangle_n := \frac{1}{n}\sum_{i=1}^n f(X^{(i)})g(X^{(i)})$. It then follows that:

1. $\|f + g\|_n^2 = \|f\|_n^2 + 2\langle f, g\rangle_n + \|g\|_n^2$
2. $\langle f, g\rangle_n \leq \|f\|_n \|g\|_n$

For a function $g : \mathbb{R}^s \to \mathbb{R}$, define $\widehat{R}_s(g) := \|f_0 + w - \overline{f_0} - \overline{w} - g\|_n^2$ as the objective of the *restricted* regression and define $R_s(g) := \mathbb{E}|f_0(X) + w - \mu - g(X)|^2$ as the population risk, where $\overline{f_0} = \frac{1}{n}\sum_i f_0(X^{(i)})$ and $\overline{w} = \frac{1}{n}\sum_{i=1}^n w_i$ and $\mu = \mathbb{E}f_0(X)$. Note that we *subtract* out the mean in the two risk definitions.

For an additive function g, define $\rho_n(g) = \sum_{k=1}^s \|\partial g_k\|_\infty$. Because we always use the secant linear piece-wise function in our optimization program, we define $\|\partial g_k\|_\infty := \max_{i=1,...,n-1} |\frac{g_k(X^{(i)}) - g_k(X^{(i+1)})}{X^{(i)} - X^{(i+1)}}|$.

Let $\mathcal{C}[b, B, L]$ be the set of 1 dimensional convex functions on $[-b, b]$ that are bounded by B and L-Lipschitz.

Let $\mathcal{C}[s, b, B, L]$ be the set of additive functions with s components each of which is in $\mathcal{C}[b, B, L]$.

$$\mathcal{C}[s, b, B, L] := \{f : \mathbb{R}^s \to \mathbb{R} \,:\, f = \sum_{k=1}^s f_k(x_k),\ f_k \in \mathcal{C}[b, B, L]\}$$

Define $f^{*s} = \arg\min\{R_s(f) \,|\, f \in \mathcal{C}^s[b, B, L], \mathbb{E}f_k(X_k) = 0\}$.

Define $f^{*(s-1)} = \arg\min\{R_s(f) \,|\, f \in \mathcal{C}^{(s-1)}[b, B, L], \mathbb{E}f_k(X_k) = 0\}$, the optimal solution with only $s - 1$ components.

Note: By definition of the Lipschitz condition, $f_k \in \mathcal{C}[b, B, L]$ implies that $\|\partial f_k\|_\infty \leq L$. $f = \sum_k f_k \in \mathcal{C}[s, b, B, L]$ implies that $\rho_n(f) \leq sL$.

8.3.2 Proof

We first start with a lemma that converts assumption A4 into a more easily applicable condition.

Lemma 8.1. *Suppose assumptions A1' and A4 hold.*

Then $R(f^{*(s-1)}) - R(f^{*s}) \geq \alpha$, where α lower bounds the norm of the population optimal additive components as defined in assumpation A4.

Proof.

$$\begin{aligned}
R(f^{*(s-1)}) - R(f^{*s}) &= \mathbb{E}\Big(f^{*(s-1)}(X) - f_0(X) + \mu\Big)^2 - \mathbb{E}\Big(f^{*s}(X) - f_0(X) + \mu\Big)^2 \\
&= \mathbb{E}\Big(f^{*(s-1)}(X) - f^{*s}(X) + f^{*s}(X) - f_0(X) + \mu\Big)^2 - \mathbb{E}\Big(f^{*s}(X) - f_0(X) + \mu\Big)^2 \\
&= \mathbb{E}\Big(f^{*(s-1)}(X) - f^{*s}(X)\Big)^2 - 2\mathbb{E}\Big[(f^{*(s-1)}(X) - f^{*s}(X))(f^{*s}(X) - (f_0(X) - \mu))\Big]
\end{aligned}$$

We will argue that all the components of the additive function $f^{*(s-1)}$ are also in f^{*s}. Let us denote the components of $f^{*s} = \sum_{k=1}^s f_k^*$. We will now invoke Corollary 3.1, which is valid because

we assume $X_1, ..., X_s$ are independent by assumption A1'. By Corollary 3.1, if we set $f_k^* = 0$, the resulting additive function $\sum_{k' \neq k} f_{k'}^*$ minimizes the population risk subject to the constraint that $f_k = 0$. By definition, f^{*s} is $\arg\min_k \sum_{k' \neq k} f_{k'}^*$ and thus share components with f^{*s}.

Therefore, there exist some k such that $f^{*(s-1)} - f^{*s} = f_k^*$, and we can continue the bound

$$\begin{aligned} R(f^{*(s-1)}) - R(f^{*s}) &= \mathbb{E} f_k^*(X_k)^2 - 2\mathbb{E}[f_k^*(X_k)(f^{*s}(X) - (f_0(X) - \mu))] \\ &= \mathbb{E} f_k^*(X_k)^2 - 2\mathbb{E}[f_k^*(X_k) f^{*s}(X)] + 2\mathbb{E}[f_k^*(X_k)(f_0(X) - \mu)] \end{aligned}$$

$$\mathbb{E}[f_k^*(X_k) f^{*s}(X)] = \mathbb{E}\Big[f_k^*(X_k)\mathbb{E}[f^{*s}(X) \,|\, X_k]\Big] = \mathbb{E} f_k^*(X_k)^2$$
$$\mathbb{E}[f_k^*(X_k)(f_0(X) - \mu)] = \mathbb{E}\Big[f_k^*(X_k)\mathbb{E}[f_0(X) - \mu \,|\, X_k]\Big] = \mathbb{E} f_k^*(X_k)^2$$

Where we used the fact that $\mathbb{E} f_{k'}^*(X_{k'}) = 0$ for all k' and the fact that $\mathbb{E}[f_0(X) - \mu \,|\, X_k] = f_k^*(X_k)$ (Lemma 3.1).

Thus, $R(f^{*(s-1)}) - R(f^{*s}) \geq \mathbb{E} f_k^*(X_k)^2 \geq \alpha$ by Assumption A4.

□

We now restate the theorem in our newly defined notation.

Theorem 8.3. *Suppose assumptions A1', A2', A3, A4 hold.*
Let $\widehat{f} := \arg\min\{\widehat{R}_s(f) + \lambda_n \rho_n(f) \,:\, f \in \mathcal{C}[s, b, B, L], f_k \text{ centered}\}$.
Suppose that $csL\lambda_n \to 0$ *and* $cLb(sB + \sigma)sB\sqrt{\frac{s}{n^{4/5}} \log sn} \to 0$.
Then, for all large enough n, *with probability at least* $1 - \frac{C}{n}$, $\widehat{f}_k \neq 0$ *for all* $k = 1, ..., s$.

Proof. Let us first sketch out the rough idea of the proof. We know that in the population setting, the best approximate additive function f^{*s} has s non-zero components. We also know that the empirical risk approaches the population risk uniformly. Therefore, it cannot be that the empirical risk minimizer maintains a zero component for all n; if that were true, then we can construct a feasible solution to the empirical risk optimization, based on f^{*s}, that achieves lower empirical risk.

Step 1: f^{*s} is not directly a feasible solution to the empirical risk minimization program because it is not empirically centered. Given n samples, $f^{*s} - \overline{f}^{*s}$ is a feasible solution where $\overline{f}^{*s} = \sum_{k=1}^s \overline{f}_k^{*s}$ and $\overline{f}_k^{*s} = \frac{1}{n}\sum_{i=1}^n f_k^{*s}(X^{(i)})$.

$$\begin{aligned} |\widehat{R}_s(f^{*s} - \overline{f}^{*s}) - \widehat{R}_s(f^{*s})| &\leq \|f_0 + w - \overline{f}_0 - \overline{w} - f^{*s} + \overline{f}^{*s}\|_n^2 - \|f_0 + w - \overline{f}_0 - \overline{w} - f^{*s}\|_n^2 \\ &\leq 2\langle f_0 + w - \overline{f}_0 - \overline{w} - f^{*s}, \overline{f}^{*s}\rangle_n + \|\overline{f}^{*s}\|_n^2 \\ &\leq 2\|f_0 + w - \overline{f}_0 - \overline{w} - f^{*s}\|_n \|\overline{f}^{*s}\|_n + \|\overline{f}^{*s}\|_n^2 \\ &\leq 2|\overline{f}^{*s}|\|f_0 + w - \overline{f}_0 - \overline{w} - f^{*s}\|_n + |\overline{f}^{*s}|^2 \end{aligned}$$

Because each f^{*s} is bounded by sB and $\mathbb{E} f^{*s}(X) = 0$, by Hoeffding inequality, with probability at least $1 - \frac{C}{n}$, $|\overline{f}^{*s}| \leq sB\sqrt{\frac{1}{cn}\log n}$.

$$\|f_0 + w - \overline{f}_0 - \overline{w} - f^{*s}\|_n \leq \|f_0 - f^{*s}\|_n + \|w\|_n + |\overline{f}_0| + |\overline{w}|$$

$\|f_0 - f^{*s}\|_n \leq \|f_0 - f^{*s}\|_\infty$ is bounded by $2sB$ and w_i is zero-mean subgaussian with scale σ. Therefore, $\|w\|_n$ is at most $c\sigma$ with probability at least $1 - \frac{C}{n}$ for all $n > n_0$.

$|\overline{f}_0| \leq sB$ and $|\overline{w}| \leq c\sigma\sqrt{\frac{1}{n}}$ with probability at least $1 - \frac{C}{n}$ by Chernoff bound.

So we derive that, with probability at least $1 - \frac{C}{n}$, for all $n > n_0$,

$$|\widehat{R}_s(f^{*s} - \overline{f}^{*s}) - \widehat{R}_s(f^{*s})| \leq 2c(sB + \sigma)sB\sqrt{\frac{1}{cn}\log n}$$

Step 2: Now that we bounded the cost of approximating f^{*s} with the empirically centered $f^{*s} - \overline{f}^{*s}$, we move on to the proof of the main result.

Suppose $\widehat{f}$ has at most $s-1$ non-zero components. Then

$$\begin{aligned}
\widehat{R}_s(\widehat{f}) &\geq R_s(\widehat{f}) - \tau_n \\
&\geq R_s(f^{*(s-1)}) - \tau_n \\
&\geq R_s(f^{*s}) + \alpha - \tau_n \\
&\geq \widehat{R}_s(f^{*s}) + \alpha - 2\tau_n \\
&\geq \widehat{R}_s(f^{*s} - \overline{f}^{*s}) - \tau_n' + \alpha - 2\tau_n
\end{aligned}$$

The third line follows from Lemma 8.1. τ_n is the deviation between empirical risk and true risk and τ_n' is the approximation error incurred by empirically sampling f^{*s}.

Adding and subtracting $\lambda_n\rho_n(f^{*s} - \overline{f}^{*s})$ and $\lambda_n\rho_n(\widehat{f})$, we arrive at the conclusion that

$$\widehat{R}_s(\widehat{f}) + \lambda_n\rho_n(\widehat{f}) \geq \widehat{R}_s(f^{*s} - \overline{f}^{*s}) + \lambda_n\rho_n(f^{*s} - \overline{f}^{*s}) - (\lambda_n\rho_n(f^{*s} - \overline{f}^{*s}) + \lambda_n\rho_n(\widehat{f})) - \tau_n' + \alpha - 2\tau_n$$

Because we assume that we impose the Lipschitz constraint in our optimization, $\rho_n(\widehat{f}), \rho_n(f^{*s} - \overline{f}^{*s})$ are at most sL and so $|\lambda_n\rho_n(\widehat{f}) - \lambda_n\rho_n(f_s^*)| \leq 2sL\lambda_n$.

By Theorem 8.4, we know that under the condition of the theorem, $\tau_n \leq Lb(sB+\sigma)sB\sqrt{\frac{s}{cn^{4/5}}\log n}$.

τ_n', as shown above, is at most $2(sB+\sigma)sB\sqrt{\frac{1}{cn}\log sn}$ with probability at least $1-\frac{C}{n}$ for $n > n_0$.

For n large enough such that

$$csL\lambda_n < \frac{\alpha}{2} \text{ and } LbsB(sB+\sigma)\sqrt{\frac{s}{n^{4/5}}\log sn}) < \frac{\alpha}{4}$$

we get that $\widehat{R}_s(\widehat{f}) + \lambda_n\rho_n(\widehat{r}) > \widehat{R}_s(f_s^*) + \lambda_n\rho_n(f_s^*)$, which is a contradiction since we assumed that $\widehat{f}$ minimizes the regularized empirical risk.

□

Theorem 8.4. *(Uniform Risk Deviation) For all $n > n_0$, we have that, with probability at least $1 - \frac{C}{n}$,*

$$\sup_{f \in \mathcal{C}^s[b,B,L]} |\widehat{R}_s(f) - R_s(f)| \leq Lb(sB+\sigma)sB\sqrt{\frac{s}{cn^{4/5}}\log sn}$$

Proof. This proof uses a standard covering number argument.

Let $\mathcal{C}_\epsilon[s,b,B,L]$ be an ϵ-cover of $\mathcal{C}[s,b,B,L]$ such that for all $f \in \mathcal{C}[s,b,B,L]$, there exists $f' \in \mathcal{C}_\epsilon[s,b,B,L]$ such that $\|f - f'\|_\infty \leq \epsilon$.

For all $f \in \mathcal{C}[s,b,B,L]$,

$$\widehat{R}_s(f) - R_s(f) = \widehat{R}_s(f) - \widehat{R}_s(f') + \widehat{R}_s(f') - R_s(f') + R_s(f') - R_s(f)$$

where $f' \in \mathcal{C}_\epsilon[s,b,B,L]$ and $\|f - f'\|_\infty \leq \epsilon$.

Step 1. We first bound $\widehat{R}_s(f) - \widehat{R}_s(f')$.

$$\begin{aligned}
|\widehat{R}_s(f) - \widehat{R}_s(f')| &= |\|f_0 - \overline{f}_0 + w - \overline{w} - f\|_n^2 - \|f_0 - \overline{f}_0 + w - \overline{w} - f'\|_n^2| \\
&\leq 2\langle f_0 - \overline{f}_0 + w - \overline{w}, f' - f\rangle_n + \|f\|_n^2 - \|f'\|_n^2 \\
&\leq 2\|f_0 - \overline{f}_0 + w - \overline{w}\|_n\|f' - f\|_n + (\|f\|_n - \|f'\|_n)(\|f\|_n + \|f'\|_n)
\end{aligned}$$

We now want to bound $\|f_0 - \overline{f}_0 + w - \overline{w}\|_n \le \|f_0\|_n + \|w\|_n + |\overline{f}| + |\overline{w}|$.

$\|w\|_n^2 = \frac{1}{n}\sum_{i=1}^n w_i^2$ is the average of subexponential random variables. Therefore, for all n larger than some absolute constant n_0, with probability at least $1 - \frac{C}{n}$, $|\|w\|_n^2 - \mathbb{E}|w|^2| < \sigma^2\sqrt{\frac{1}{cn}\log n}$. The absolute constant n_0 is determined so that for all $n > n_0$, $\sqrt{\frac{1}{cn}\log n} < 1$. Since $\mathbb{E}w^2 \le \sigma^2$, for all $n > n_0$, with probability at least $1 - \frac{C}{n}$, for some constant c, $\|w\|_n^2 \le c\sigma^2$.

By Chernoff bound and the fact that $\mathbb{E}w = 0$, we know that $|\overline{w}| \le c\sigma\sqrt{\frac{1}{n}}$ with high probability.

$\|f_0\|_n^2 = \frac{1}{n}\sum_{i=1}^n f_0(X^{(i)})^2$. Since $f_0(X^{(i)})^2 \le s^2B^2$, $\|f_0\|_n^2 \le s^2B^2$.

$|\overline{f}_0| = |\frac{1}{n}\sum_{i=1}^n f_0(X^{(i)})| \le sB$.

Combining these together, We have that for all $n \ge n_0$, with probability at least $1 - \frac{C}{n}$, $\|f_0 - \overline{f}_0 + w - \overline{w}\|_n^2 \le c(s^2B^2 + \sigma^2)$, and so

$$\|f_0 - \overline{f}_0 + w - \overline{w}\|_n^2 \le c(sB + \sigma)$$

$\|f' - f\|_\infty \le \epsilon$ implies that $\|f' - f\|_n \le \epsilon$. And therefore, $\|f\|_n - \|f'\|_n \le \|f - f'\|_n \le \epsilon$.

f, f' are all bounded by sB, and so $\|f\|_n, \|f'\|_n \le sB$.

Thus, we have that, for all $n > n_0$,

$$|\widehat{R}_s(f) - \widehat{R}_s(f')| \le \epsilon c(sB + \sigma) \tag{8.1}$$

with probability at least $1 - \frac{C}{n}$.

Step 2: Now we bound $R_s(f') - R_s(f)$. The steps follow the bounds before, and we have that

$$|R_s(f') - R_s(f)| \le \epsilon c(sB + \sigma) \tag{8.2}$$

Step 3: Lastly, we bound $\sup_{f' \in \mathcal{C}_\epsilon[s,b,B,L]} \widehat{R}_s(f') - R_s(f')$.

For a fixed f', we have that, by definition

$$\begin{aligned}
\widehat{R}_s(f') &= \|f_0 + w - \overline{f}_0 - \overline{w} - f'\|_n^2 \\
&= \|f_0 - \overline{f}_0 - f'\|_n^2 + 2\langle w - \overline{w}, f_0 - \overline{f}_0 - f'\rangle_n + \|w - \overline{w}\|_n^2 \\
R_s(f') &= \mathbb{E}\big(f_0(X) + w - \mu - f'(X)\big)^2 \\
&= \mathbb{E}\big(f_0(X) - \mu - f'(X)\big)^2 + \mathbb{E}w^2
\end{aligned}$$

Therefore:

$$\begin{aligned}
\widehat{R}_s(f') - R_s(f') =& \|f_0 - \overline{f}_0 - f'\|_n^2 - \mathbb{E}(f_0(X) - \mu - f'(X))^2 \\
&+ \|w - \overline{w}\|_n^2 - \mathbb{E}w^2 \\
&+ 2\langle w - \overline{w}, f_0 - \overline{f}_0 - f'\rangle_n
\end{aligned}$$

Step 3.1: We first bound $2\langle w - \overline{w}, f_0 - \overline{f}_0 - f'\rangle_n$.

$$\langle w - \overline{w}, f_0 - \overline{f}_0 - f'\rangle_n = \langle w, f_0 - \overline{f}_0 - f'\rangle_n - \langle \overline{w}, f_0 - \overline{f}_0 - f'\rangle_n$$

The first term, fully expanded, is $\frac{1}{n}\sum_{i=1}^n w_i(f_0(X^{(i)}) - \overline{f}_0 - f'(X^{(i)}))$. Since w_i and $X^{(i)}$ are independent, we use the sub-Gaussian concentration inequality. Note that $|f_0(X^{(i)}) - \overline{f}_0 - f'(X^{(i)})| \le 3sB$, and so $|\langle w, f_0 - \overline{f}_0 - f'\rangle_n| > t$ with probability at most $C\exp\big(-cnt^2\frac{1}{\sigma^2(sB)^2}\big)$.

The second term, fully expanded, is $\overline{w}\overline{f'}$. $\overline{f'} \le sB$ and so $|\overline{w}\overline{f'}| > t$ with probability at most $C\exp\big(-cnt^2\frac{1}{\sigma^2(sB)^2}\big)$ as well.

Step 3.2 We now bound $\|w-\overline{w}\|_n^2 - \mathbb{E}w^2$.

$$\begin{aligned}\|w-\overline{w}\|_n^2 &= \|w\|_n^2 - 2\langle w, \overline{w}\rangle_n + \|\overline{w}\|_n \\ &= \|w\|_n^2 - \overline{w}^2\end{aligned}$$

Using sub-Exponential concentration, we know that $|\|w\|_n^2 - \mathbb{E}w^2| \geq t$ with probability at most $C\exp(-cn\frac{1}{\sigma^2})$.

$\overline{w} \leq \sigma\sqrt{\frac{1}{cn}}$ with probability at least $1-\frac{C}{n}$. Thus, $|\overline{w}|^2 \leq \sigma^2\frac{1}{cn}$ with high probability, has a second order effect, and can be safely ignored in the bound.

Step 3.3: We now bound $\|f_0 - \overline{f}_0 - f'\|_n^2 - \mathbb{E}(f_0(X)-\mu-f'(X))^2$.

$$\begin{aligned}\|f_0-\overline{f}_0-f'\|_n^2 &= \|f_0-\mu+\mu-\overline{f}_0-f'\|_n^2 \\ &= \|f_0-\mu-f'\|_n^2 + 2\langle \mu-\overline{f}_0, f_0-\mu-f'\rangle_n + \|\mu-\overline{f}_0\|_n^2\end{aligned}$$

Using similar reasoning as before, we know that $|\langle \mu-\overline{f}_0, f_0-\mu-f'\rangle_n| \geq t$ with probability at most $C\exp(-cnt^2\frac{1}{(sB)^4})$.

Likewise, $|\mu-\overline{f}_0| \leq sB\sqrt{\frac{1}{cn}}$ with probability at least $1-\frac{C}{n}$. Thus, $|\mu-\overline{f}_0|^2 \leq (sB)^2\frac{1}{cn}$, has a second order effect, and can be safely ignored in the bound.

Because $f_0(X^{(i)})-\mu-f'^(X^{(i)})$ is bounded by $3sB$, $\|f_0-\mu-f'\|_n^2$ is the empirical average of n random variables bounded by $9(sB)^2$.

Using Hoeffding Inequality then, we know that the probability $\left|\|f_0-\mu-f'\|_n^2 - \mathbb{E}(f_0(X)-\mu-f'(X))^2\right| \geq t$ is at most $C\exp(-cnt^2\frac{1}{(sB)^4})$.

Applying union bound, we have that $\sup_{f'\in\mathcal{C}_\epsilon[s,b,B,L]}|\widehat{R}_s(f') - R_s(f')| \geq t$ occurs with probability at most

$$C\exp\left(s\left(\frac{bBLs}{\epsilon}\right)^{1/2} - cnt^2\frac{1}{\sigma^2(sB)^2+(sB)^4}\right)$$

for all $n > n_0$.

Restating, we have that with probability at most $1-\frac{1}{n}$, the deviation is at most

$$(sB+\sigma)sB\sqrt{\frac{1}{cn}\left(\log Cn + s(\frac{bBLs}{\epsilon})^{1/2}\right)} \tag{8.3}$$

Substituting in $\epsilon = \frac{bBLs}{n^{2/5}}$, expression 8.3 can be upper bounded by $sB(\sigma+sB)\sqrt{\frac{s}{cn^{4/5}}\log Cn}$.

Expressions 8.1 and 8.2 from **Step 1** and **Step 2** become $\sqrt{\frac{(bBLs)^2}{cn^{4/5}}}(sB+\sigma)$.

We can arrive at the statement of the theorem by summing these up and absorbing any constants into the symbols c and C.

□

8.4 Supporting Technical Material

8.4.1 Concentration of Measure

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

Proposition 8.1. *(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 8.1. *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 $1 - \frac{1}{n}$,

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

□

8.4.2 Sampling Without Replacement

Lemma 8.2. *(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 8.2. *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.

□

8.4.3 Covering Number for Lipschitz Convex Functions

Definition 8.1. $\{f_1, ..., f_N\} \subset \mathcal{C}[b, B, L]$ 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 8.3. *(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 8.4.

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

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. □