Variable Selection in Convex Function Estimation

Min Xu, Minhua Chen, John Lafferty

September 29, 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 [5, 6] 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 sparse nonparametric regression problem is considered in [9], where it is shown that computationally efficient, near minimax-optimal estimation is possible, but in ambient dimensions that scale only as $p = O(\log n)$. This is in stark contrast to the exponential scaling $p = O\big(e^{n^c}\big)$ enjoyed by sparse linear models [19]. Sparse additive models [16] provide 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 additive faithfulness, as we show here. In addition, we show that exponential scaling 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

$$\text{minimize} \quad \sum_{i=1}^{n} (Y_i - m(x_i))^2 \tag{1.1}$$

$$\text{subject to} \quad m : \mathbb{R}^p \to \mathbb{R} \text{ is convex} \tag{1.2}$$

is precisely equivalent to the finite dimensional quadratic program

$$\text{minimize}_{h,\beta} \quad \sum_{i=1}^{n}(Y_i - h_i)^2 \tag{1.3}$$

$$\text{subject to} \quad h_j \geq h_i + \beta_i^T(x_j - x_i), \text{ for all } i, j. \tag{1.4}$$

See [1], Section 6.5.5. 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 [1], computed tomography [14], target reconstruction [10], image analysis [4] and circuit design [7]. Other applications include queuing theory [3] and economics, where it is of interest to estimate concave utility functions [13]. See [11] 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 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).

Notation. If $\boldsymbol{x}$ is a vector, we use $\boldsymbol{x}_{-k}$ to denote the vector with k-th coordinate removed. Let $\boldsymbol{v} \in \mathbb{R}^n$, 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 all ones. If $X \in \mathbb{R}^p$ and $S \subset \{1, ..., p\}$, then X_S is the subvector of X restricted to coordinates in S. Given n samples $X^{(1)}, ..., X^{(n)}$, we use $\overline{X}$ to denote empirical average.

2 Additive Faithfulness

The additive approximation may mark a relevant variable as irrelevant. Such mistakes are inherent to the approximation and may persist even with infinite samples. Lemma 2.1 below provides examples of errors made by the additive approximations that minimize the L_2 distance.

Lemma 2.1. *Let F be a product distribution on $\boldsymbol{C} = [0,1]^s$ with density function p which is positive on $\boldsymbol{C}$. Let $X = (X_1, ..., X_s) \sim F$. Let $f : \boldsymbol{C} \to \mathbb{R}$ be an integrable function and suppose, without loss of generality, that $\mathbb{E}f(X) = 0$. Let $f_1^*, ..., f_s^* := \arg\min\{\mathbb{E}\big(f(X) - \sum_{k=1}^{s} f_k(X_k)\big)^2 : \mathbb{E}f_k(X_k) = 0\}$. Then $f_k^*(x_k) = \mathbb{E}[f(X) \,|\, x_k]$ is unique.*

Lemma 2.1 follows from the stationarity condition of the optimal solution. If F is the uniform distribution, then $f_k^*(x_k) = \int_{\boldsymbol{x}_{-k}} f(x_k, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k}$.

Example 2.1. We give two examples of additive unfaithfulness under the uniform distribution. First, consider the following function:

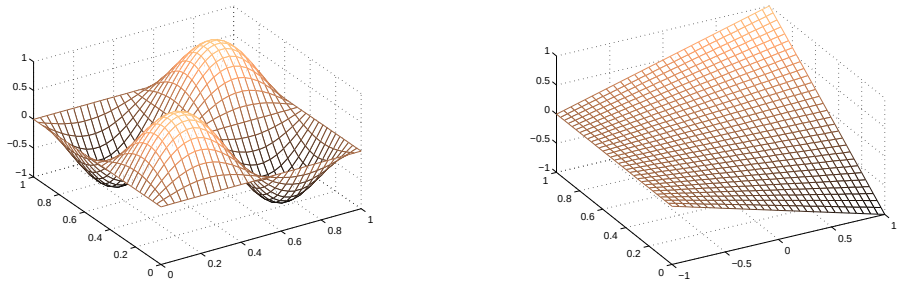
$$\text{(egg carton)} \quad f(x_1, x_2) = \sin(2\pi x_1)\sin(2\pi x_2) \qquad \text{for } (x_1, x_2) \in [0,1]^2$$

For all x_1, $\int_{x_2} f(x_1, x_2) dx_2 = 0$ and also, for all x_2, $\int_{x_1} f(x_1, x_2) dx_1 = 0$. An additive approximation would set $f_1 = 0$ and $f_2 = 0$. Next, consider the function

$$\text{(tilting slope)} \quad f(x_1, x_2) = x_1 x_2 \qquad \text{for } x_1 \in [-1, 1],\ x_2 \in [0, 1].$$

For all x_2, $\int_{x_1} f(x_1, x_2) dx_1 = 0$, 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 exactly x_2.

It is important to understand the kind of functions for which the additive approximation can accurately capture all of the relevant variables. We call this property *additive faithfulness*.



(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.

Definition 2.1. Let $\boldsymbol{C} = [0,1]^s$, and $f : \boldsymbol{C} \to \mathbb{R}$. We say that f *depends on* coordinate k if there exist $x_k' \neq x_k$ such that $f(x_k', \boldsymbol{x}_{-k})$ and $f(x_k, \boldsymbol{x}_{-k})$ are different functions of $\boldsymbol{x}_{-k}$.

Let F be a probability distribution on $\boldsymbol{C}$ and assume without loss of generality that $\mathbb{E}f(X) = 0$. Let

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

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

Remarkably, under product distributions, a convex multivariate function can always be faithfully approximated by an additive function.

Theorem 2.1. *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 give the full proof in Section 6.1 of the Appendix, but pause here to provide some intuition. From Lemma 2.1, we know intuitively that the additive approximation zeroes out k when, fixing x_k, every "slice" of f integrates to zero. We prove Theorem 2.1 by showing that "slices" of convex functions that integrate to zero cannot be "glued" together while still maintaining convexity.

Theorem 2.1 plays an important role in our sparsistency analysis, in which we prove that the additive approximation is variable selection consistent (or "sparsistent") *even when the true function is not additive.*

Remark 2.1. We assume twice differentiability in Theorem 2.1 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. Without restrictions on the distribution, a convex function may not be additively faithful. Intuitively, an arbitrarily shaped density p may "undo" the convexity of f so that the product $p(\boldsymbol{x})\, f(\boldsymbol{x})$ resembles an egg carton or a tilting slope. With appropriate conditions on the density p, however, it is possible to relax the independence assumption. We leave this to future work.

3 Optimization for Sparse Convex Additive Models

We consider the following nonparametric regression problem

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

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. 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.,

$$h_{kj} \geq h_{ki} + \beta_{ki}(x_{kj} - x_{ki}) \quad (\forall i, j) \tag{3.1}$$

where $h_{ki} := f_k(x_{ki})$ and β_{ki} is the subgradient at point x_{ki}. We apparently need $O(n^2p)$ 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:

$$\begin{aligned} &\min_{\boldsymbol{h},\boldsymbol{\beta},\mu} \ \frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \sum_{k=1}^{p} h_{ki} - \mu\Big)^2 + \lambda \sum_{k=1}^{p} \|\boldsymbol{\beta}_{k\cdot}\|_\infty \\ &\text{s.t.} \ \sum_{i=1}^{n} h_{ki} = 0, \ h_{k(i+1)} = h_{k(i)} + \beta_{k(i)}(x_{k(i+1)} - x_{k(i)}), \ \beta_{k(i+1)} \geq \beta_{k(i)} \ (\forall k, i) \end{aligned} \tag{3.2}$$

We introduce a mean parameter $\mu \in \mathbb{R}$ because f may not have zero-mean. We can in fact solve μ explicitly here: the optimal $\mu = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}$ because of KKT and the constraints that $\sum_i h_{ki} = 0$. $\{(1), (2), \ldots, (n)\}$ is a reordering of $\{1, 2, \ldots, n\}$ such that $x_{k(1)} \leq x_{k(2)} \leq \cdots \leq x_{k(n)}$. It is easy to verify that the constraints in (3.2) satisfy the supporting hyperplane constraints, as

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

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). Notice that if we replace $\beta_{k(i+1)} \geq \beta_{k(i)}$ with $\beta_{k(i+1)} = \beta_{k(i)}$, the optimization reduces to the lasso. Note while one can use supporting hyperplanes to the epigraph as in (1.4), SCAM uses the *inner piece-wise linear function* that approximates the graph with secant lines.

The SCAM optimization in (3.2) is a quadratic program (QP) with $O(np)$ variables and $O(np)$ constraints. The optimal μ has a closed form solution of $\mu = \frac{1}{n}\sum_{i=1}^{n} Y_i$, which is easy to derive from the KKT theorem and the constraints that $\sum_i h_{ki} = 0$ for all k. 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:

$$\begin{aligned} &\min_{\boldsymbol{h}_{k\cdot},\boldsymbol{\beta}_{k\cdot},\gamma_k} \ \frac{1}{2n}\sum_{i=1}^{n}\Big((y_i - \overline{y} - \sum_{r\neq k} h_{ri}) - h_{ki}\Big)^2 + \lambda\gamma_k \\ &\text{s.t.} \ \sum_{i=1}^{n} h_{ki} = 0, \ h_{k(i+1)} = h_{k(i)} + \beta_{k(i)}(x_{k(i+1)} - x_{k(i)}), \ \beta_{k(i+1)} \geq \beta_{k(i)}, \ -\gamma_k \leq \beta_{k(i)} \leq \gamma_k \ (\forall i). \end{aligned}$$

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 (3.2), 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 (3.1),

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

3.1 Alternative Formulation

Optimization (3.2) can be reformulated in terms of the 2nd derivatives. 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

$$f_k(x_{k(i)}) = f_k(x_{k(1)}) + d_{k(1)}(x_{k(i)} - x_{k(1)}) + d_{k(2)}(x_{k(i)} - x_{k(2)}) + \cdots + d_{k(i-1)}(x_{k(i)} - x_{k(i-1)})$$

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

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

Where Δ_k is a $n \times n-1$ matrix such that $\Delta_k(i,j) = \Delta_{k(j)}(x_{k(i)})$ and $d_k = (d_{k(1)}, ..., d_{k(n-1)})$. We can now reformulate (3.2) as an equivalent optimization program with only centering and positivity constraints:

$$\min_{d_k \in \mathbb{R}^{n-1}, c_k \in \mathbb{R}, \mu \in \mathbb{R}} \frac{1}{2n} \left\| Y - \overline{Y}\mathbf{1}_n - \sum_{k=1}^{p} (\Delta_k d_k - c_k \mathbf{1}_n) \right\|_2^2 + \lambda_n \sum_{k=1}^{p} \|d_k\|_1 \tag{3.3}$$

$$\text{s.t. } \forall k,\ d_{k(2)}, \dots, d_{k(n-1)} \geq 0 \qquad \text{(convexity)}$$

$$c_k = \frac{1}{n} \mathbf{1}_n^\top \Delta_k d_k \qquad \text{(centering)}$$

$\|d_k\|_1$ is not identical to $\|\boldsymbol{\beta}_{k\cdot}\|_\infty$, but it is easy to verify that $\|\boldsymbol{\beta}_{k\cdot}\|_\infty \leq \|d_k\|_1 \leq 2\|\boldsymbol{\beta}_{k\cdot}\|_\infty$.

Remark 3.1. For parts of our theoretical analysis, we will also impose onto (3.3) 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 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.

4.1 Deterministic Setting

We follow Wainwright [19] and define a purely theoretical construct called *restricted regression*.

Definition 4.1. In *restricted regression*, we restrict the indices k in optimization (3.3) to lie in the support S instead of ranging from $1, ..., p$.

Our analysis then diverges from the now-standard "primal-dual witness technique" [19]. 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 4.1. *(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 (3.3) where we restrict $k \in S$. Let $\widehat{d}_k = 0$ and $\widehat{c}_k = 0$ for $k \in S^c$. Let $\widehat{r} := Y - \overline{Y} - \sum_{k \in S}(\Delta_k \widehat{d}_k - \widehat{c}_k \mathbf{1})$ be the restricted regression residual. For $k \in \{1, ..., p\}$, Let $\Delta_{k,j}$ be the j-th column of Δ_k, i.e. the n-dimensional vector $\max(X_k - X_{k(j)}\mathbf{1}, 0)$.*

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 3.3. Furthermore, any solution to the optimization program 3.3 must be zero on S^c.

This result holds regardless of whether we impose the boundedness and Lipschitz conditions in optimization 3.3. The full proof of Theorem 4.1 is in Section 6.2 of the Appendix.

Remark 4.1. The incoherence condition of [19] 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 4.1 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.

4.2 Probabilistic Setting

We use the following statistical setting:

1. Let $\mathcal{X} = [-b, b]^p$. Let F be a distribution supported and positive on $\mathcal{X}$. Let $X^{(1)}, ..., X^{(n)}$ be iid samples from F.
2. Let $Y = f_0(X) + \epsilon$ where f_0 is the true function and ϵ is noise. Let $Y^{(1)}, ..., Y^{(n)}$ be iid samples.
3. Let $s \leq p$ be a positive integer. Let $S = \{1, ..., s\}$ denote the relevant variables, i.e., $f_0(X) = f_0(X_S)$.
4. Let $f_{01}, ..., f_{0s} := \arg\min_{f_1, ..., f_s}\{\mathbb{E}(f_0(X) - \sum_{k=1}^s f_k(X_k))^2 \,|\, \mathbb{E}[f_k(X_k)] = 0\}$ where $\mu = \mathbb{E}f(X)$.

Each of our statistical 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, L-Lipschitz, and $\text{supp}(f_0) = S$.

A3: Suppose ϵ is mean-zero sub-Gaussian, independent of X, with sub-Gaussian scale σ.

A4: For all $k = 1, ..., s$, $\mathbb{E}(f_{0k}(X_k))^2 \geq \alpha$ 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.

Remark 4.2. 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.

Theorem 4.2. *(Controlling false positives) Suppose assumptions A1, A2, A3 hold. Suppose also that we run optimization* (3.3) *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 (3.3), with boundedness constraint, is zero on S^c.*

The proof of Theorem 4.2 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 (3.3). Our proof uses the concentration of the average of data sampled *without* replacement [17], illustrating that the proof method is not a trivial application of existing techniques. The full proof of Theorem 4.2 is in Section 6.3 of the Appendix.

Theorem 4.3. *(Controlling false negatives) Suppose assumptions A1', A2', A3 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 $L\max\left(\lambda_n, b(B+\sigma)B\sigma\sqrt{\frac{s^5}{n^{4/5}}\log sn}\right) \to 0$. 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 2.1. We need stronger assumptions in Theorem 4.3 to use our additive faithfulness result, Theorem 2.1. We also include an extra Lipschitz constraint so that we can use existing covering number results [2]. Recent work [6] shows that the Lipschitz constraint is not required with more advanced empirical process theory techniques. We give the full proof of Theorem 4.3 in Section 6.4 of the Appendix.

Combining Theorem 4.2 and 4.3 and ignoring dependencies on b, B, L, σ, we have the following result.

Corollary 4.1. *Assume A1', A2', A3. Let $\lambda_n = \Theta\left(\sqrt{\frac{s^3}{n}\log n \log(pn)}\right)$. Suppose $\lambda_n \to 0$ and $\sqrt{\frac{s^5}{n^{4/5}}\log sn} \to 0$. Let $\widehat{f_n}$ be a solution to (3.3) 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 $o(n^{4/25})$.

5 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 (3.2). 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.

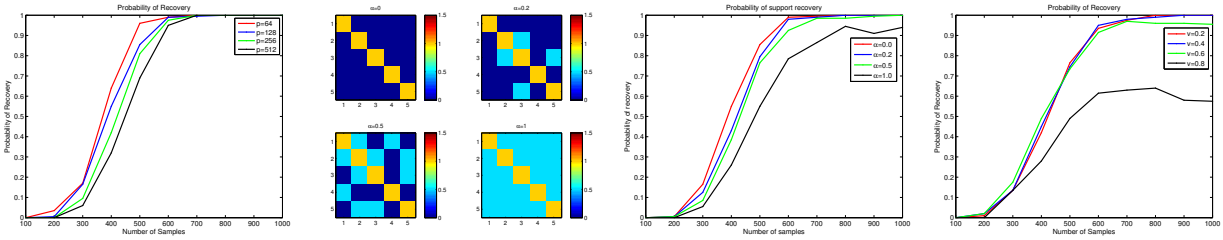
We next use the Boston housing data rather than simulated data. This data set contains 13 covariates, 506 samples and one response variable indicating housing values in suburbs of Boston. The data and detailed description can be found on the UCI Machine Learning Repository website `http://archive.ics.uci.edu/ml/datasets/Housing`.

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 [15]. 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 [15].

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 [12], 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.

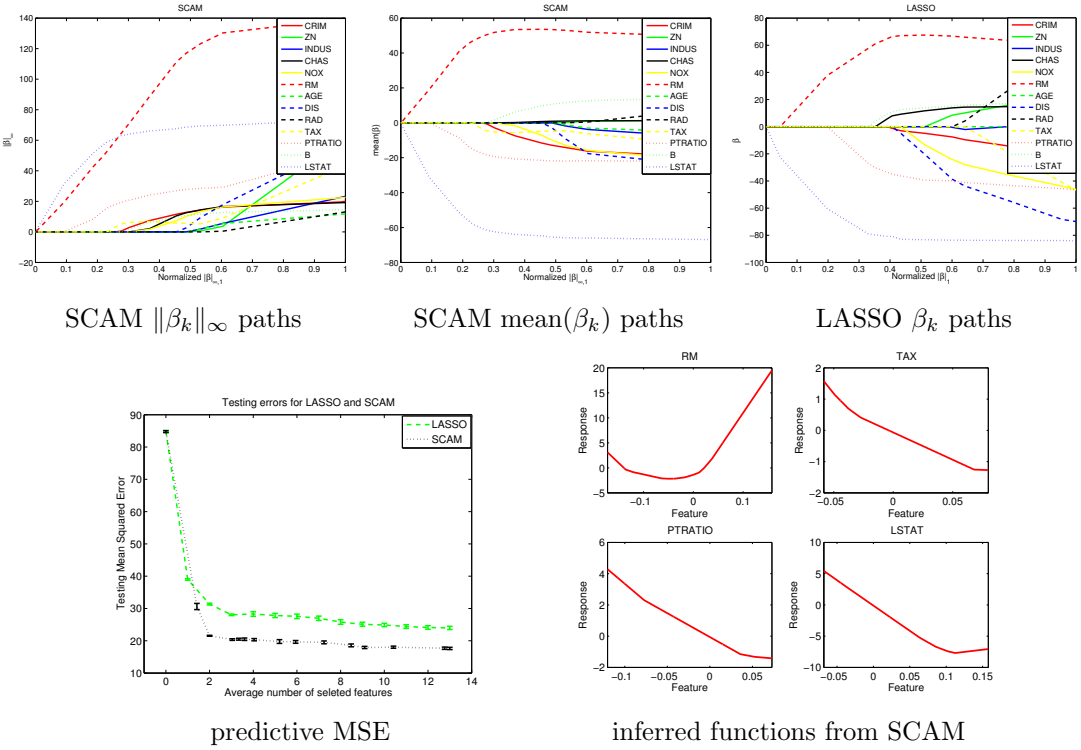
6 Discussion

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



(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

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

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

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

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

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

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

[7] 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.

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

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

[10] 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.

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

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

[13] 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.

[14] 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.

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

[16] 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.

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

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

[19] 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.

Appendix

6.1 Proof of Additive Faithfulness

We start with a lemma similar to Lemma 2.1 and whose proofs are identical.

Lemma 6.1. *Let F be a product distribution on $C = [0,1]^s$ with density function p. Let $f : C \to \mathbb{R}$ be a convex function. Suppose that $\mathbb{E}f(X) = 0$.*

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

Then $f_k^*(x_k) = \mathbb{E}[f(X) \,|\, x_k]$.

Proof. Let $f_1^*, ..., f_s^*$ be the minimizers as defined. It must be then that f_k^* minimizes $\{\mathbb{E}|f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - f_k(X_k)|^2 \,:\, f_k \text{ convex}, \mathbb{E}f_k(X_k) = 0\}$.

Fix x_k, we will show that the value $\mathbb{E}[f(X)|x_k]$, for all x_k, uniquely minimizes

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

If this is true, then the function $x_k \mapsto \mathbb{E}[f(X) \,|\, x_k]$ is the unique $f_k^*(x_k)$ that minimizes the expected square error.

Take the derivative with respect to the value $f_k(x_k)$ and set it equal to zero, we get that

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

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

$$f_k(x_k) = \int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k}) f(\boldsymbol{x}) d\boldsymbol{x}_{-k}$$

Where the second line follows because $p(\boldsymbol{x}) = \prod_k p(x_k)$ by product distribution assumption. The third line follows because $\int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k}) f_{k'}^*(x_{k'}) d\boldsymbol{x}_{-k} = 0$ for all $k' \neq k$ by the constraint that $\mathbb{E}[f_{k'}^*(X_{k'})] = 0$.

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 $f_k^*(x_k) = \mathbb{E}[f(X) \,|\, x_k]$ is unique.

Now, we verify that as a function of x_k, $\mathbb{E}[f(X)|x_k]$ has mean zero and is convex. The former is true because $\mathbb{E}f(X) = 0$; the latter is true because for every $\boldsymbol{x}_{-k}$, $f(x_k, \boldsymbol{x}_{-k})$ is a convex function with respect to x_k and therefore, $\int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k}) f(x_k, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k}$ is still convex.

□

The next proposition, combined with the previous Lemma, proves Theorem 2.1.

Proposition 6.1. *Let F be a product distribution supported on $C = [0,1]^s$ with positive density $p > 0$. Let $f : C \to \mathbb{R}$ be a convex function, twice differentiable.*

Let $f_1^*, ..., f_s^* := \arg\min\{\mathbb{E}|f(X) - \sum_k f_k(X_k)|^2 \,:\, \forall k, f_k \text{ convex}, \mathbb{E}f_k(X_k) = 0\}$.

The following are equivalent:

1. *f does not depends on coordinate k*
2. *For all x_k, $\mathbb{E}[f(X)|x_k] = 0$.*

Proof. The first condition trivially implies the second because $\mathbb{E}f(X) = 0$.

Fix k. Suppose that, for all x_k, $\mathbb{E}[f(X)|x_k] = 0$.

By the assumption that , we know that, for all x_k,

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

For every $\boldsymbol{x}_{-k}$, we define the derivative

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

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

We now describe two facts about g.

Fact 1. By exchanging limit with the integral, which is valid by bounded convergence theorem, we reason that

$$\int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k}) g(\boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0$$

Fact 2. Because f is convex, $g(\boldsymbol{x}_{-k})$ is a component of the subgradient $\partial_{\boldsymbol{x}} f(0, \boldsymbol{x}_{-k})$. (the subgradient coincides with the gradient by assumption that f is twice differentiable)

Therefore, using the first order characterization of a convex function, we have

$$f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + \partial_{\boldsymbol{x}} f(\boldsymbol{x})^{\mathsf{T}} (\boldsymbol{x}' - \boldsymbol{x}) \quad \text{for all } \boldsymbol{x}', \boldsymbol{x}$$
$$f(x_k, \boldsymbol{x}_{-k}) \geq f(0, \boldsymbol{x}_{-k}) + g(\boldsymbol{x}_{-k}) x_k \quad \text{for all } x_k, \boldsymbol{x}_{-k}$$

For all $x_k, \boldsymbol{x}_{-k}$,

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

and

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

Since $f(x_k, \boldsymbol{x}_{-k}) - f(0, \boldsymbol{x}_{-k}) - g(\boldsymbol{x}_{-k}) x_k$ is a continuous function of $\boldsymbol{x}_{-k}$ and the density $p(\boldsymbol{x}_{-k})$ is, by assumption, non-zero on the support, we conclude that for all $x_k, f(x_k, \boldsymbol{x}_{-k}) - f(0, \boldsymbol{x}_{-k}) - g(\boldsymbol{x}_{-k}) x_k = 0$ necessarily and thus $\boldsymbol{x}_{-k}$, $f(x_k, \boldsymbol{x}_{-k}) = f(0, \boldsymbol{x}_{-k}) + g(\boldsymbol{x}_{-k}) x_k$.

The Hessian of f then (guaranteed to exist by assumption) has a zero on the k-th main diagonal entry.

By proposition 7.1.10 from Horn and Johnson [8], such a matrix is positive semidefinite if and only if the k-th row and column are also zero. (to prove this, one can first work with a 2 by 2 matrix and then use the fact that every principal submatrix of a PSD matrix must be PSD)

Since k-th row and column correspond precisely to the gradient of $g(\boldsymbol{x}_{-k})$, we conclude that g must be a constant function. It follows therefore that $g = 0$ because it integrates to 0.

So we have that for all $x_k, \boldsymbol{x}_{-k}$, $f(x_k, \boldsymbol{x}_{-k}) = f(0, \boldsymbol{x}_{-k})$, which concludes our proof.

□

6.2 Proof of the Deterministic Condition for Sparsistency

We restate Theorem 4.1 first for convenience.

Theorem 6.1. *The following holds regardless of whether we impose the boundedness and smoothness condition in optimization 3.3 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 3.3. 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 residue.

Suppose for all j, k, $\lambda_n > |\frac{1}{n} \widehat{r}^{\mathsf{T}} \Delta_{k,j}|$, then $\widehat{d}_k, \widehat{c}_k$ for $k = 1, ..., p$ is an optimal solution to the full regression 3.3.

Furthermore, any solution to the optimization program 3.3 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 $\mu_k \in \mathbb{R}^n$ is a vector of dual variables where $\mu_{k,1} = 0$ and $\mu_{k,i} \geq 0$ for $i = 2, ..., n$.

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.

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

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

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

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

where $\boldsymbol{u}$ 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

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

Suppose

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

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

Because we have strict inequality in the above equation, Lemma 1 from [19] show that all solutions must be zero on S^c. □

6.3 Proof of False Positive Control

We first restate the theorem for convenience.

Theorem 6.2. *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, k*,*

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

And therefore, the solution to the optimization 3.3 is zero on S^c*.*

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

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

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

Where $\overline{f^*} = \frac{1}{n}\sum_{i=1}^n f^*(X_S^{(i)})$ 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 $|f^*(x)| \leq B$ and $|\widehat{f}_k(x_k)| \leq B$ for all k.

Then $|\overline{f^*}| \leq B$ 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 $\frac{1}{n}\widehat{r}^\mathsf{T}\max(X, X_{(j)}\mathbf{1})$.

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

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.

$$\text{Want to bound} \qquad F(X_1, ..., X_n) := \frac{1}{n}\sum_{i=1}^n \widehat{r}_i X_i \delta(ord(i)\geq j)$$

First, we note that X_i is bounded in the range $[-b, b]$.

We claim then that F is coordinatewise-Lipschitz. Let $X = (X_1, X_2, ..., X_n)$ and $X' = (X_1', X_2, ..., X_n)$ differ only on the first coordinate.

The order of coordinate i in X and X' 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 $F(X)$ to $F(X')$. Therefore,

$$|F(X_1, ..., X_n) - F(X_1', ..., X_n)| \leq \frac{4b\|\widehat{r}\|_\infty}{n}$$

By McDiarmid's inequality therefore,

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

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

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

Term 2:

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

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

Now we put everything together.

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

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

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

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

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

□

6.4 Proof of False Negative Control

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

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

For a function $g:\mathbb{R}^s\to\mathbb{R}$, define $\widehat{R}_s(g) := \|f_0+w-g\|_n^2$ as the objective of the *restricted* regression and define $R_s(g) := \mathbb{E}|f_0(X)+w-g(X)|^2$ as the population risk.

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 $\mathcal{C}[b,B,L]$.

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

We now restate the theorem in our newly defined notation.

Theorem 6.3. *Suppose assumptions A1', A2', A3 hold. Suppose f_0 depends on all s-variables.*
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 n is large enough such that $cL\max\left(\lambda_n, b(B+\sigma)B\sigma\sqrt{\frac{1}{n^{4/5}}s^5\log sn}\right)\to 0$.
Then, 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.

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

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

By Theorem 2.1, $R_s(f_s^*) - R_s(f_{(s-1)}^*) \geq \alpha > 0$.

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 $\overline{f}_k^{*s} = \frac{1}{n}\sum_{i=1}^{n} f_k^{*s}(X^{(i)})$.

$$\begin{aligned}|\widehat{R}_s(f^{*s} - \overline{f}^{*s}) - \widehat{R}_s(f^{*s})| &\leq \|y - f^{*s} + \overline{f}^{*s}\|_n^2 - \|y - f^{*s}\|_n^2 \\ &\leq 2\|y - f^{*s}\|_n\|\overline{f}^{*s}\|_n + \|\overline{f}^{*s}\|_n^2\end{aligned}$$

Because each f_k^{*s} is bounded by B, by Hoeffding inequality, with probability at least $1 - \frac{C}{n}$, $|\overline{f}_k^{*s}| \leq B\sqrt{\frac{1}{cn}\log n}$. By an union bound therefore, with probability at least $1 - \frac{C}{n}$, $\|\overline{f}^{*s}\|_n \leq B\sqrt{\frac{1}{cn}\log sn}$.

$$\begin{aligned}\|y - f^{*s}\|_n &= \|f_0 + w - f^{*s}\|_n \\ &\leq \|f_0 - f^{*s}\|_n + \|w\|_n\end{aligned}$$

$f_0 - f^{*s}$ 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 $n > n_0$.

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

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

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

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

Where 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

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

$\rho_n(\widehat{f}), \rho_n(f^{*s} - \overline{f}^{*s})$ are at most L. By Theorem 6.4, we know that under the condition of the theorem, $\tau_n \leq bLB\sigma(B + \sigma)\sqrt{\frac{1}{cn^{4/5}}s^5\log n}$.

$|\lambda_n\rho_n(\widehat{f}) - \lambda_n\rho_n(f_s^*)| \leq 2L\lambda_n$.

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

For n large enough such that

$$c\max(L\lambda_n, bLB\sigma(B + \sigma)\sqrt{\frac{1}{n^{4/5}}s^5\log sn}) < \alpha$$

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.

□

Theorem 6.4. *(Uniform Risk Deviation) For all $n > n_0$, we have that, with probability at least $1 - \frac{C}{n}$,*

$$\sup_{f \in \mathcal{C}^s[b,B,L]} |\widehat{R}_s(f) - R_s(f)| \leq B\sigma(B+\sigma)Lb\sqrt{\frac{1}{cn^{4/5}} s^5 \log sn}$$

Proof. Let $\mathcal{C}_\epsilon[s,b,B,L]$ be an ϵ-cover of $\mathcal{C}[s,b,B,L]$.

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 $\widehat{R}_s(f) - \widehat{R}_s(f')$.

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

$\|f_0 + w\|_n \leq \|f_0\|_n + \|w\|_n$. $\|w\|_n^2 = \frac{1}{n}\sum_{i=1}^n w_i^2$ 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 $n > n_0$, $\sqrt{\frac{1}{cn}\log n} < 1$.

$\|f_0\|_n^2$ is the average of random variables bounded by B^2 and therefore, with probability at least $1 - \frac{C}{n}$, $|\|f_0\|_n^2 - \mathbb{E}|f_0(X)|^2| \leq B^2\sqrt{\frac{1}{cn}\log n}$.

Since $\mathbb{E}|w|^2 \leq c\sigma^2$ and $\mathbb{E}|f_0(X)|^2 \leq B^2$, we have that for all $n \geq n_0$, with probability at least $1 - \frac{C}{n}$, $\|f_0 + w\|_n \leq c(B+\sigma)$.

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

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

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

$$|\widehat{R}_s(f) - \widehat{R}_s(f')| \leq \epsilon cs(B+\sigma) \tag{6.1}$$

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

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

$$|R_s(f') - R_s(f)| \leq \epsilon cs(B+\sigma) \tag{6.2}$$

Lastly, we bound $\sup_{f' \in \mathcal{C}_0^s[b,B,L]} \widehat{R}_s(f') - R_s(f')$.

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

$$\|f_0 + w - f'\|_n^2 = \|f_0 - f'\|_n^2 + 2\langle w, f_0 - f'\rangle_n + \|w\|_n^2$$

Because $f_0(X^{(i)}) - f'^(X^{(i)})$ is bounded by $2sB$, $\|f_0 - f'\|_n^2$ is the empirical average of n random variables bounded by $4(sB)^2$.

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

Consider now the term $2\langle w, f_0 - f'\rangle_n := \frac{2}{n}\sum_{i=1}^n w_i(f_0(X^{(i)}) - f'(X^{(i)}))$. We note that w_i and $X^{(i)}$ are independent, w_i is subgaussian.

The n-dimensional vector $\{\frac{1}{n}(f_0(X^{(i)}) - f'(X^{(i)}))\}_i$ has norm at most $\frac{sB}{\sqrt{n}}$. Therefore, $|2\langle w, f_0 - f'\rangle_n| \geq t$ with probability at most $C\exp(-cnt^2\frac{1}{\sigma^2(sB)^2})$.

The last term $\|w\|_n^2 = \frac{1}{n}\sum_{i=1}^n w_i^2$. Using subexponential concentration, we know that $|\|w\|_n^2 - \mathbb{E}|w|^2| \geq t$ occurs with probability at most $C\exp(-cn\frac{1}{\sigma^2})$ for n larger than some n_0.

Collecting all these results and applying union bound, we have that $\sup_{f' \in \mathcal{C}_0^s[b,B,L]} |\widehat{R}_s(f') - R_s(f')| \geq t$ occurs with probability at most

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

for all $n > n_0$.

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

$$\sqrt{\frac{1}{cn}\sigma^2(sB)^4\left(\log Cn + s(\frac{bBLs}{\epsilon})^{1/2}\right)} \tag{6.3}$$

Substituting in $\epsilon = \frac{bBLs}{n^{2/5}}$, expression 6.3 becomes $\sqrt{\frac{1}{cn^{4/5}}\sigma^2 s^5 B^4 \log Cn}$.

Expressions 6.1 and 6.2 become $\sqrt{\frac{(bBLs)^2}{cn^{4/5}}}(B+\sigma)$.

□

6.5 Supporting Technical Material

6.5.1 Concentration of Measure

Sub-Exponential random variable is the square of a subgaussian random variable[18].

Proposition 6.2. *(Subexponential Concentration [18]) 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 6.1. *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 $1 - \frac{1}{n}$,

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

□

6.5.2 Sampling Without Replacement

Lemma 6.2. *(Serfling [17]) 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 6.2. *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 $(b - a)\sqrt{\frac{1}{2N}\log\frac{2}{\delta}}$.s

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.

□

6.5.3 Covering Number for Lipschitz Convex Functions

Definition 6.1. $\{f_1, ..., f_N\} \subset \mathcal{C}[b, B, L]$ 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 6.3. *(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 6.4.

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

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. □