1 Setting and Notations

Setting:

1. Let $Y' = f_0(X) + \epsilon$ where f_0 is the true function and ϵ is noise. Given n samples $Y'^{(1)}, ..., Y'^{(n)}$, we input into our optimization $Y = Y' - \bar{Y}'$.
2. Let $\mathcal{X} = [-b, b]^p$. Let $\mathcal{P}$ be a distribution on $\mathcal{X}$. Let $X^{(1)}, ..., X^{(n)}$ be independent samples from $\mathcal{P}$.
3. Let $S \subset \{1, ..., p\}$ denote the set of relevant variables, that is, $f_0(X) = f_0(X_S)$ and let $s \coloneqq |S|$.

Assumptions:

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

A1' Suppose A1 is true and $\{X_k\}_{k \in S}$ are all independent.

A2 Suppose that $\|f_0\|_\infty \leq B$.

A2' Suppose that A2 is true and f_0 is convex and *L-Lipschitz*.

A3 Suppose ϵ is mean-zero subgaussian, independent of X, with subgaussian scale σ.

Notation: $\mathbf{1}_n$ is an all-ones vector of dimension n. If $X \in \mathbb{R}^p$ and $S \subset \{1, ..., p\}$, then X_S is a subvector of X restricted to coordinates in S. Let $v \in \mathbb{R}^n$, then $v_{(1)}$ denotes the largest coordinate of v in magnitude, $v_{(j)}$ denotes the j-th largest.

1.1 Reformulation

Let $x_{s(1)}, ..., x_{s(n)}$ be the n samples arranged from small to large. Define $\widehat{\beta}_{s1}$ as $\frac{\widehat{f}_s(x_{s(2)}) - \widehat{f}_s(x_{s(1)})}{x_{s(2)} - x_{s(1)}}$.

Then

$$\begin{aligned}
\widehat{f}_s(x_{s(1)}) &= \widehat{f}_s(x_{s(1)}) \qquad \text{constrained by centering} \\
\widehat{f}_s(x_{s(2)}) &= \widehat{f}_s(x_{s(1)}) + \widehat{\beta}_{s1}(x_{s(2)} - x_{s(1)}) \\
\widehat{f}_s(x_{s(t)}) &= \widehat{f}_s(x_{s(1)}) + \sum_{t'=1}^{t-1} \widehat{\beta}_{st'}(x_{s(t'+1)} - x_{s(t')})
\end{aligned}$$

We thus define the notation $\widehat{f}_s(x_{si}) = c_s + \widehat{\beta}_s^\intercal D_{si}$ where $D_{si} \in \mathbb{R}^{n-1}$, and is a vector

$$D_{si} = [x_{s(2)} - x_{s(1)}, x_{s(3)} - x_{s(2)}, ..., x_{s(t)} - x_{s(t-1)}, 0, ..., 0] \qquad \text{where } t = \mathit{order}(i)$$

And we have the constraint that $\sum_{i=1}^n \widehat{f}_s(x_{si}) = nc_s + \sum_{i=1}^n \widehat{\beta}_s^\intercal D_{si} = 0$, therefore, $c_s = -\left(\frac{1}{n}\sum_{i=1}^n D_{si}\right)^\intercal \widehat{\beta}_s$.

Some additional transformation. Let's define $\widehat{d}_{s(i)}$ as the gradient increment. $\widehat{d}_{s(1)} = \widehat{\beta}_{s(1)}$, and $\widehat{d}_{s(2)} = \widehat{\beta}_{s(2)} - \widehat{\beta}_{s(1)}$. The convexity constraint translates to the constraint that $\widehat{d}_{s(i)} \geq 0$ for all $i > 1$.

$\widehat{\beta}_{s(i)} = \sum_{j \leq i} \widehat{d}_{s(i)}$.

$$\widehat{f}_s(x_{s(2)}) = \widehat{f}_s(x_{s(1)}) + \widehat{d}_{s(1)}(x_{s(2)} - x_{s(1)})$$

$$\begin{aligned}
\widehat{f}_s(x_{s(3)}) &= \widehat{f}_s(x_{s(2)}) + \widehat{\beta}_{s(2)}(x_{s(3)} - x_{s(2)}) \\
&= \widehat{f}_s(x_{s(1)}) + \widehat{d}_{s(1)}(x_{s(2)} - x_{s(1)}) + (\widehat{d}_{s(2)} + \widehat{d}_{s(1)})(x_{s(3)} - x_{s(2)}) \\
&= \widehat{f}_s(x_{s(1)}) + \widehat{d}_{s(1)}(x_{s(3)} - x_{s(1)}) + \widehat{d}_{s(2)}(x_{s(3)} - x_{s(2)})
\end{aligned}$$

$$\widehat{f}_s(x_{s(i)}) = \widehat{f}_s(x_{s(1)}) + \widehat{d}_{s(1)}(x_{s(i)} - x_{s(1)}) + \widehat{d}_{s(2)}(x_{s(i)} - x_{s(2)}) + ... + \widehat{d}_{s(i-1)}(x_{s(i)} - x_{s(i-1)})$$

Define $\Delta(j, x_{si}) = 0$ if order$(i) \leq j$, $x_{si} - x_{s(j)}$ else. The j ranges from 1 to $n-1$. With this definition, we can re-write

$$\widehat{f}_s(x_{si}) = \widehat{d}_s^{\mathsf{T}} \Delta(x_{si}) \qquad \text{where } \Delta(x_{si}) \in \mathbb{R}^{n-1}.$$

With the simple constraint that all $\widehat{d}_{si} \geq 0$ for $i > 1$.

$$\begin{aligned}
&\min_{\{d_k, c_k\}} \frac{1}{2n} \|Y - \sum_{k=1}^{p} (\Delta_k d_k - c_k \mathbf{1}_n)\|_2^2 + \lambda_n \sum_{k=1}^{p} \|d_k\|_1 && (1.1) \\
&\text{s.t. } \forall k,\ d_{k2}, ..., d_{k(n-1)} \geq 0 && \text{(convexity)} \\
&\qquad c_k = \frac{1}{n} \mathbf{1}_n^{\mathsf{T}} \Delta_k d_k && \text{(centering)} \\
&\qquad -B\mathbf{1}_n \leq \Delta_k d_k + c_k \mathbf{1}_n \leq B\mathbf{1}_n && \text{(boundedness*)} \\
&\qquad \|d_k\|_1 \leq L && \text{(smoothness*)}
\end{aligned}$$

We will impose the boundness and smoothness constraints only in our theoretical analysis when we control the rate of false negatives.

$$\begin{aligned}
&\min_{\{d_k, c_k\}} \frac{1}{2n} \|Y - \sum_{k \in S} (\Delta_k d_k - c_k \mathbf{1}_n)\|_2^2 + \lambda_n \sum_{k=1}^{p} \|d_k\|_1 && (1.2) \\
&\text{s.t. } \forall k \in S,\ d_{k2}, ..., d_{k(n-1)} \geq 0 && \text{(convexity)} \\
&\qquad c_k = \frac{1}{n} \mathbf{1}_n^{\mathsf{T}} \Delta_k d_k && \text{(centering)} \\
&\qquad -B\mathbf{1}_n \leq \Delta_k d_k + c_k \mathbf{1}_n \leq B\mathbf{1}_n && \text{(boundedness*)} \\
&\qquad \|d_k\|_1 \leq L && \text{(smoothness*)}
\end{aligned}$$

In the proof, we will reason with the solution of the optimization 1.1 when we restricted k to be only in the subset S. This is of course a theoretical construct only and we refer to it as *restricted regression*.

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 $\widehat{R}_s(g) := \|f_0 + w - g\|_n^2$ and define $R_s(g) := \mathbb{E}|f_0(X) + w - g(X)|^2$.

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

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

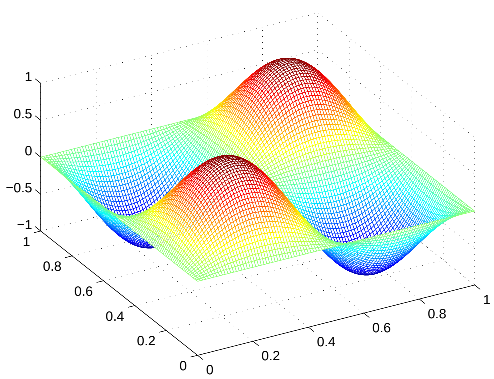
$$\mathcal{C}^s[b, B, L] := \{f : \mathbb{R}^s \to \mathbb{R} \,:\, f = \sum_{k=1}^s f_k(x_k),\ f_k \in \mathcal{C}[b, B, L]\}$$

2 Functions Which Are Not Additively Faithful

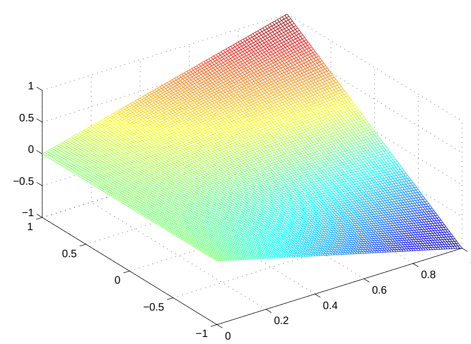
Example 1:

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

Note that 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 model would set $f_1 = 0$ and $f_2 = 0$.



(a) Example 1



(b) Example 2

Example 2:

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

Note that for all x_2, $\int_{x_1} f(x_1, x_2) dx_1 = 0$, therefore, we expect $f_2 = 0$ under the additive model.

This function, for every fixed x_2, is a zero-intercept linear function of x_1 with slope exactly x_2.

3 Convex Functions are Additively Faithful

Let μ be a probability measure on $C = [0, 1]^s$, $f(x)$ be a multivariate function on C. We say that f depends on coordinate i if there exist $x_i' \neq x_i$ such that $f(x_i', x_{-i})$ and $f(x_i, x_{-i})$ are different functions of x_{-i}. (on some measurable set)

Theorem 3.1. *Let p be a product probability distribution on $C = [0,1]^s$ so that $X_1, ..., X_s$ are all independent. Let $f : C \to \mathbb{R}$ be a convex function, twice differentiable.*

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

Then $f_1, ..., f_s$ are non-constant functions.

Lemma 3.1. *Let μ be a probability measure on $C = [0,1]^s$. Let $f : C \to \mathbb{R}$ be a convex function, twice differentiable. Suppose that $\mathbb{E} f(X) = 0$.*

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

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

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

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

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

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

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 $\mathbf{x}_{-k}$, $f(x_k, \mathbf{x}_{-k})$ is a convex function with respect to x_k and therefore, $\int_{x_{-k}} p(\mathbf{x}|x_k) f(x_k, \mathbf{x}_{-k}) d\mathbf{x}_{-k}$ is still convex.

□

Proposition 3.1. *Let p be a product probability distribution on $C = [0,1]^s$ so that $X_1, ..., X_s$ are all independent. Let $f : C \to \mathbb{R}$ be a convex function, twice differentiable.*

Let $f_1^*, ..., f_s^* := \arg\min\{\mathbb{E}|f(X) - \sum_k f_k(X_k)|^2 \,:\, \forall k, f_k$ 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 p is a product measure, we know that, for all x_k,

$$\begin{aligned} p(x_k)\mathbb{E}[f(X)|x_k] &= \int_{\mathbf{x}_{-k}} p(\mathbf{x}_{-\mathbf{k}} \,|\, x_k) f(x_k, \mathbf{x}_{-k}) d\mathbf{x}_{-k} \\ &= \int_{\mathbf{x}_{-k}} p(\mathbf{x}_{-\mathbf{k}}) f(x_k, \mathbf{x}_{-k}) d\mathbf{x}_{-k} = 0 \end{aligned}$$

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

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

$g(\mathbf{x}_{-k})$ 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, we reason that

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

Fact 2. Because f is convex, $g(\mathbf{x}_{-k})$ is a component of the subgradient $\partial_{\mathbf{x}} f(0, \mathbf{x}_{-k})$. (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

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

Because, for all $x_k, \mathbf{x}_{-k}$,

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

and

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

we conclude that for all $x_k, \mathbf{x}_{-k}$, $f(x_k, \mathbf{x}_{-k}) = f(0, \mathbf{x}_{-k}) + g(\mathbf{x}_{-k}) x_k$.

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

By proposition from Horn and Johnson [TODO ref], such a matrix is positive semidefinite if and only if the k-th row and column are also zero.

Since k-th row and column correspond precisely to the gradient of $g(\mathbf{x}_{-k})$, 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 $x_k, \mathbf{x}_{-k}$, $f(x_k, \mathbf{x}_{-k}) = f(0, \mathbf{x}_{-k})$, which concludes our proof.

$\square$

4 Deterministic Conditions

Theorem 4.1. *(Deterministic)*
The following holds regardless of whether we impose the boundness and smoothness condition in optimization 1.1 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 1.2. Let $\widehat{d}_k = 0$ and $\widehat{c}_k = 0$ for $k \in S^c$.

Let $\widehat{r} \coloneqq Y - \sum_{k \in S}(\Delta_k \widehat{d}_k - \widehat{c}_k \mathbf{1})$ 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 1.1.

Furthermore, any solution to the optimization program 1.1 must be zero on S^c.

Proof. We will omit the boundness 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.

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

where $r_k \coloneqq Y - \sum_{k \in S}(\Delta_k d_k - c_k \mathbf{1})$, 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.

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

In the derivatives, $\mathbf{u}$ 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.

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

where $\mathbf{u}$ 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

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

Suppose

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

then we easily see that the first coordinate of $\mathbf{u}$ can be set to some value in $(-1, 1)$ and $\mu_{k,i} > 0$ for $i = 2, ..., n$. Because we have strict inequality, Lemma [TODO:get wainwright lemma] shows that all solutions must be zero on S^c. □

4.1 Probabilistic Condition: Controlling False Positives

Theorem 4.2. *(Probabilistic: Controlling False Positives)*
Suppose assumptions A1, A2, A3 hold. Suppose also that we run optimization 1.1 with the B-boundness constraint.

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 1.1, with boundedness constraint, 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 not a function of X_{S^c} at all.

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

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

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

Suppose 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, $|\bar{\epsilon}| \leq \sigma\sqrt{\frac{c}{n}\log\frac{C}{\delta}}$ 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 $|\bar{f}^*| \leq B$ as well, and $|f^*(X_S^{(i)}) - \bar{f}^* - \sum_{k \in S}\widehat{f}_k(X_k^{(i)})| \leq 3sB$.
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 $\frac{1}{n}\widehat{r}^{\mathsf{T}}\max(X, X_{(j)}\mathbf{1})$.

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

Where $A = \{i : ord(i) \geq j\}$
We will bound both terms.

Term 1.

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

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, 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

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

□

4.2 Probabilistic Condition: Controlling False Negatives

Theorem 4.3. *(Probabilistic: Controlling False Negatives)*
Suppose assumptions A1', A2', A3 hold. Suppose we run optimization 1.1 with both the B-boundedness and L-Lipschitz constraint. Suppose f_0 depends on all s-variables.

Let $\widehat{f} \coloneqq \arg\min\{\widehat{R}_s(f) + \lambda_n \rho_n(f) \,:\, f \in \mathcal{C}^s[b,B,L], f_k \text{ centered}\}$.

Suppose n is large enough such that $cL \max\left(\lambda_n, b(B+\sigma)B\sigma\sqrt{\frac{1}{n^{4/5}} s^5 \log sn}\right) < C_{thresh}(f_0)$.

Then, with probability at least $1 - \frac{C}{n}$, $\widehat{f}_k \neq 0$ for all $k = 1, ..., s$ and therefore, the solution to optimization 1.1 is non-zero on 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. 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 $f^{*s} = \arg\min\{R_s(f) \,|\, f \in \mathcal{C}^s[b,B,L], \mathbb{E}f_k(X_k) = 0\}$.

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.

By [TODO:population no false negative theorem], $R_s(f^*_s) - R_s(f^*_{(s-1)}) \geq \alpha > 0$.

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

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

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

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

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

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

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

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

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

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

$\rho_n(\widehat{f}), \rho_n(f^{*s} - \bar{f}^{*s})$ are at most L. By Theorem 4.4, 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 4.4. *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 $\mathcal{C}^s_0[b,B,L]$ 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 $f' \in \mathcal{C}^s_0[b,B,L]$ and $\|f - f'\|_\infty \leq \epsilon$.

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

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

$\|f_0 + w\|_n \le \|f_0\|_n + \|w\|_n$. $\|w\|_n^2 = \frac{1}{n}\sum_{i=1}^n w_i^2$ 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 $1 - \frac{C}{n}$, $|\|f_0\|_n^2 - \mathbb{E}|f_0(X)|^2| \le B^2\sqrt{\frac{1}{cn}\log n}$.

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

$\|f' - f\|_\infty \le \epsilon$ 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 $n > n_0$,

$$|\widehat{R}_s(f) - \widehat{R}_s(f')| \le \epsilon cs(B + \sigma) \tag{4.1}$$

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

$$|R_s(f') - R_s(f)| \le \epsilon cs(B + \sigma) \tag{4.2}$$

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 $|\|f_0 - f'\|_n^2 - \mathbb{E}(f_0(X) - f'(X))^2| \ge t$ is at most $C\exp(-cnt^2\frac{1}{(sB)^4})$.

Consider now the term $2\langle w, f_0 - f'\rangle_n \coloneqq \frac{2}{n}\sum_{i=1}^n w_i(f_0(X^{(i)}) - f'(X^{(i)}))$. 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, $|2\langle w, f_0 - f'\rangle_n| \ge t$ 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 $|\|w\|_n^2 - \mathbb{E}|w|^2| \ge t$ 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 $\sup_{f' \in \mathcal{C}_0^s[b,B,L]} |\widehat{R}_s(f') - R_s(f')| \ge t$ 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

$$\sqrt{\frac{1}{cn}\sigma^2 (sB)^4 \left(\log Cn + s(\frac{bBLs}{\epsilon})^{1/2}\right)} \tag{4.3}$$

Substituting in $\epsilon = \frac{bBLs}{n^{2/5}}$, expression 4.3 becomes $\sqrt{\frac{1}{cn^{4/5}}\sigma^2 s^5 B^4 \log Cn}$.

Expressions 4.1 and 4.2 become $\sqrt{\frac{(bBLs)^2}{cn^{4/5}}}(B+\sigma)$.

□

5 Supporting Technical Results

5.1 Concentration of Measure

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

Proposition 5.1. *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 5.1. *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 $1 - \frac{1}{n}$,

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

□

5.1.1 Sampling Without Replacement

Lemma 5.1. *(Serfling) Let $x_1, ..., x_N$ be a finite list, $\bar{x} = \mu$. 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 5.2. *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.

□

5.2 Covering Number for Lipschitz Convex Functions

Definition 5.1. $\{f_1, ..., f_N\} \subset \mathcal{C}[b, B, L]$ 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 5.2. *(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 5.3.

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

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. □