1 No False Negatives

Let us intuitively describe the logic.

Let n be some large number and let $\widehat{f}_S$ be the optimal solution with a false negative.

Let f_S^* be the population optimal solution.

$$\begin{aligned}\widehat{R}(\widehat{f}_S) &\geq R(\widehat{f}_S) - gap_n \\ &\geq R(f_S^* miss) - gap_n \\ &\geq R(f_S^*) + gap - gap_n \\ &\geq \widehat{R}(f_S^*) + gap - 2gap_n \\ &\geq \widehat{R}(f_S^*) + pen(f_S^*) - pen(f_S^*) + pen(\widehat{f}_S) - pen(\widehat{f}_S) + gap - 2gap_n\end{aligned}$$

We have then that

$$\widehat{R}(\widehat{f}_S) + pen(\widehat{f}_S) \geq \widehat{R}(f_S^*) + pen(f_S^*) + (pen(\widehat{f}_S) - pen(f_S^*) + gap - 2gap_n)$$

Because pen and gap_n decreases with n and gap does not, for large enough n, we would reach a contradiction.

The only goal then is to figure out gap_n.

A more precisely statement is, with probability at least $1-\delta$,

$$\sup_{f_1,\ldots,f_s} |\widehat{R}(f_S) - R(f_S)| \leq gap_n$$

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 $f_2 = 0$.

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.

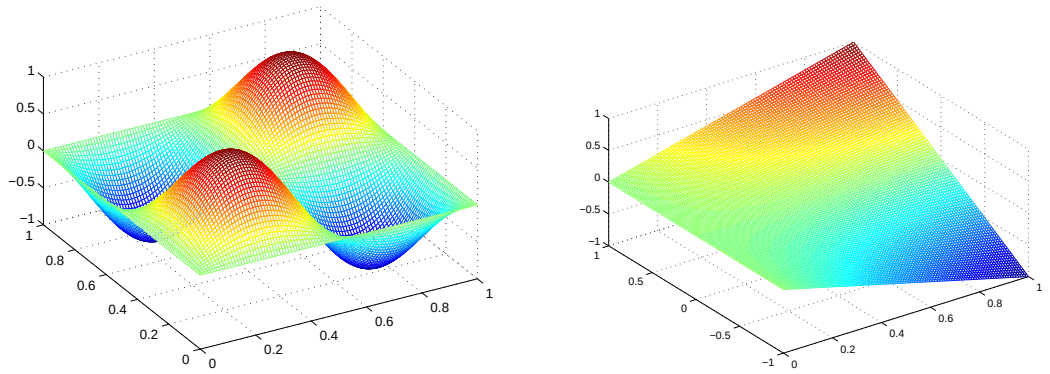
3 Convex Functions are Additively Faithful

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 μ be a probability measure on $C = [0,1]^s$. Let $f : C \to \mathbb{R}$ be a convex function, twice differentiable.*

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

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



(a) Example 1 (b) Example 2

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

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 $\mathbb{E}[f(X)|x_k]$ 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 probability distribution on $C = [0,1]^s$. Let $f : C \to \mathbb{R}$ be a convex function, twice differentiable. Suppose p is a product distribution.*

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

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 p is a product measure, we know that, for all x_k,

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

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

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

$g(\mathbf{x}_{-k})$ 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.

□

4 Technical Details

4.1 Norm of Subgaussian Random Vector

Let $V = [V_1, ..., V_s] \in \mathbb{R}^s$ be a subgaussian random vector, each with subgaussian norm at most σ.

Then, with probability at least $1-\delta$, $\|V\|_\infty$ is at most $\sigma\sqrt{\log \frac{s}{\delta}}$

Since $\|V\|_2 \le \sqrt{s}\|V\|_\infty$, with probability at least $1-\delta$, $\|V\|_2 \le \sigma\sqrt{\frac{s}{c}\log\frac{sC}{\delta}}$.

4.2 l_∞ norm bound on $\widehat{r}$

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

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

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

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

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

Lipschitz Function Bound Suppose f^* and $\widehat{f}_k$ are all Lipschitz with Lipschitz constant at most $L_{\max}$. The composite function $\sum_{k\in S}\widehat{f}_k$ is also $L_{\max}$-Lipschitz.

Let $u = \bar{f^*}$. $X_S \mapsto f^*(X_S) - \bar{f^*}$ is still $L_{\max}$-Lipschitz of course.

Define $h(X_S^{(i)}) \coloneqq f^*(X_S^{(i)}) - u - \sum_{k\in S}\widehat{f}_k(X_k^{(i)})$ as a short-hand. h is therefore Lipschitz with Lipschitz constant at least $2L_{\max}$.

Because $\sum_i h(X_S^{(i)}) = 0$, it must be that for some i, $h(X_S^{(i)}) \ge 0$ and for some i, $h(X_S^{(i)}) \le 0$.

$$\text{Therefore,} \qquad \max_i|h(X_S^{(i)})| \le \max_{i,j}|h(X_S^{(i)}) - h(X_S^{(j)})|$$

$$\begin{aligned}|h(X_S^{(i)}) - h(X_S^{(j)})| &\le 2L_{\max}\|X_S^{(i)} - X_S^{(j)}\| \\ &\le 2L_{\max}(\|X_S^{(i)}\| + \|X_S^{(j)}\|) \\ &\le 4L_{\max}\max_i\|X_S^{(i)}\|\end{aligned}$$

Therefore, $\max_i|h(X_S^{(i)})| \le 4L_{\max}\max_i\|X_S^{(i)}\|$.

Using the concentration inequality on norms of subgaussian random vectors and a union bound, we get that, with probability at least $1-\delta$, $\max_i\|X_S^{(i)}\| \le \sqrt{\frac{s}{c}\log\frac{nsC}{\delta}}$ (assuming that each coordinate of X_S has subgaussian norm 1.

Taking union bound across deviation for ϵ and the Lipschitz function term, we have that, with probability at least $1-\delta$,

$$\|\widehat{r}\|_\infty \le \sigma L_{\max}\sqrt{\frac{s}{c}(\log sn + \log\frac{C}{\delta})}$$

4.3 Sampling Without Replacement

Lemma 4.1. *(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 4.1. *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.

□

4.4 Bounding $\frac{1}{n}\widehat{r}^\mathsf{T} \max(X, X_{(j)}\mathbf{1})$

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

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 assume 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 $1-\delta$, $|\frac{1}{n}X_{(j)}\mathbf{1}_A^\mathsf{T}\widehat{r}_A| \leq b\|\widehat{r}\|_\infty\sqrt{\frac{1}{cn}\log\frac{C}{\delta}}$.

All 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 $1-\delta$,

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

$$\sigma L_{\max} b\sqrt{\frac{1}{c}\frac{s}{n}(\log(spn) + \log\frac{C}{\delta})^2}$$

DEPRECATED!

5 Additive Model

5.1 Proof Sketch

FACT: A bounded convex function is Lipschitz.

FACT: If $f_k(x_k)$ are L-Lipschitz, then $\sum_k f_k(x_k)$ is also L-Lipschitz.

$$|\sum_k f_k(x_k) - \sum_k f_k(x'_k)|^2 \leq \sum_k |f_k(x_k) - f_k(x'_k)|^2 \leq \sum_k L^2 |x_k - x'_k|^2 \leq L^2 \|x - x'\|_2^2$$

FACT: if X is sub-gaussian random vector, then $g(X)$ is also sub-gaussian if g is Lipschitz. This is true by McDiarmid's inequality.

ASSUMPTION: X_k for $k = 1, ..., p$ are independent. We can use weaker assumptions.

Let $\widehat{f}_k$ be the optimal solution to the optimization

$$\frac{1}{2n}\|Y - \sum_{s=1}^{p} \widehat{f}_s(X_s)\|_2^2 + \lambda \sum_{s=1}^{p} \|\partial \widehat{f}_s(X_s)\|_2 \qquad \widehat{f}_s \text{ convex and } \sum_i \widehat{f}_s(x_{is}) = 0.$$

ASSUMPTION: Y has mean zero, $\widehat{f}_k$ for all k has mean zero, and $\widehat{f}_k$ is Lipschitz with an universal upper bound on the Lipschitz constant.

CLAIM: If we set $\lambda = O\sqrt{\frac{\log(np)}{n}}$, then $\widehat{f}_k$ for an irrelevant k is zero.
We fix an irrelevant dimension k, and define the residue $\widehat{y} = y - \sum_{k' \neq k} \widehat{f}_{k'}(X_{k'})$.
Using the reformulation, we can re-write the optimization, in term of just f_k, as the following.

$$\min_{d,c} \frac{1}{2n}\|\widehat{y} - \Delta d - c\|_2^2 + \lambda \sum_{i=1}^{n} d_i + \lambda |d_0| \qquad \text{s.t.} \forall i > 0,\ d_i \geq 0$$

Where d is a n-dim vector, c scalar is the centering variable, and Δ is the $n \times n$ matrix of inter-sample distances.

Taking the subgradient with respect to d, we have that

$$\frac{1}{n}(-\widehat{y}^\mathsf{T}\Delta + d^\mathsf{T}\Delta^\mathsf{T}\Delta) + \lambda \mathbf{u} - \mu = 0$$

where $\mathbf{u}$ is a n-dim vector whose first coordinate is $\partial|d_0|$ and whose all other coordinates are 1. μ is a n-dim vector whose first coordinate is 0 and all other coordinates are the Lagrangian multipliers.

We now want to verify that if we substitute $d = 0$ into the subgradient, the subgradient equals 0. If $d = 0$, then we can substitute $\mathbf{u} = \mathbf{1}$.

All we need to verify then is that $\lambda \mathbf{1} \geq \frac{1}{n}\widehat{y}^{\mathsf{T}}\Delta$.

Note that $\widehat{y}$ is a sub-gaussian random vector with mean zero. Because $\widehat{y}$ and Δ are independent, $\widehat{y}^{\mathsf{T}}\Delta$ is also a sub-gaussian mean-zero random vector. Note also that Δ is entirely positive.

The empirical mean $\frac{1}{n}\widehat{y}^{\mathsf{T}}\Delta$ is then on the order of $\sqrt{\frac{1}{n}}$ with high probability.
We need to take union bound across the n-dimensions of $\widehat{y}^{\mathsf{T}}\Delta$ and across the p dimensions.
Therefore, setting $\lambda = O\sqrt{\frac{\log(np)}{n}}$ suffices.

5.2 Reformulation

The objective is, again

$$\frac{1}{2n}\|Y - \sum_{s=1}^{p} \widehat{f}_s(X_s)\|_2^2 + \sum_{s=1}^{p} \|\partial \widehat{f}_s(X_s)\|_2 \qquad \widehat{f}_s \text{ convex and } \sum_i \widehat{f}_s(x_{is}) = 0$$

Let us try to simplify this form. We can replace the identifiability constraint with $\widehat{f}_s(0) = 0$ or $\widehat{f}_s(-1) = 0$. Let's go with the latter formulation and define $x_{s(0)}$ as -1.
Let $x_{s(1)}, ..., x_{s(n)}$ be the n samples arranged from small to large.

Then $\widehat{f}_s(x_{s(1)}) = \widehat{f}_s(x_{s(0)}) + \widehat{\beta}_{s(0)}(x_{s(1)} - x_{s(0)})$
and $\widehat{f}_s(x_{s(2)}) = \widehat{\beta}_{s(0)}(x_{s(1)} - x_{s(0)}) + \widehat{\beta}_{s(1)}(x_{s(2)} - x_{s(1)})$

Some additional transformation. Let's define $\widehat{d}_{s(0)}$ as the gradient increments. $\widehat{d}_{s(0)} = \widehat{\beta}_{s(0)}$, and $\widehat{d}_{s(1)} = \widehat{\beta}_{s(1)} - \widehat{\beta}_{s(0)}$.

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

$$\widehat{f}_s(x_{s(1)}) = \widehat{f}_s(x_{s(0)}) + \widehat{d}_{s(0)}(x_{s(1)} - x_{s(0)})$$

$$\begin{aligned}
\widehat{f}_s(x_{s(2)}) &= \widehat{f}_s(x_{s(1)}) + \widehat{\beta}_{s(1)}(x_{s(2)} - x_{s(1)}) \\
&= \widehat{d}_{s(0)}(x_{s(1)} - x_{s(0)}) + (\widehat{d}_{s(0)} + \widehat{d}_{s(1)})(x_{s(2)} - x_{s(1)}) \\
&= \widehat{d}_{s(0)}(x_{s(2)} - x_{s(0)}) + \widehat{d}_{s(1)}(x_{s(2)} - x_{s(1)})
\end{aligned}$$

$$\widehat{f}_s(x_{s(i)}) = \widehat{d}_{s(0)}(x_{s(i)} - x_{s(0)}) + \widehat{d}_{s(1)}(x_{s(i)} - x_{s(1)}) + ... + \widehat{d}_{s(i-1)}(x_{s(i)} - x_{s(i-1)})$$

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

$$\widehat{f}_s(x_{si}) = \widehat{d}_s^{\mathsf{T}}\Delta(x_{si})$$

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

Notation: We will also define Δ_s as a $(n \times n)$ matrix whose i, j-th entry is $\Delta(j, x_{si})$.
With this short-hand, we can write $\widehat{f}_s(x_s) = \Delta_s \widehat{d}_s$.

5.3 Consistency with Short-hand

We can now attempt the consistency proof with our short-hand notation.

First, what is the objective? $\|\partial \widehat{f}_s(X_s)\|_\infty$ becomes $\|\widehat{d}_s\|_1$.

The objective is therefore

$$\frac{1}{2n}\|Y - \sum_{s=1}^{p} \Delta_s \widehat{d}_s\|_2^2 + \sum_{s=1}^{p} \|\widehat{d}_s\|_1.$$

Taking the same form as before, we have

$$\frac{1}{2n}\left\{\|\sum_s \Delta_s(\widehat{d}_s - d_s^*)\|_2^2 + \sum_s \langle \Delta_s(\widehat{d}_s - d_s^*), \epsilon\rangle\right\} \leq \lambda \sum_s \|d_s^*\|_1 - \lambda \sum_s \|\widehat{d}_s\|_1$$

The cross term can be bounded.

$$\begin{aligned}
\frac{1}{2n}|\sum_s \langle \Delta_s(\widehat{d}_s - d_s^*), \epsilon\rangle| &\leq \frac{1}{2}\sum_s \langle \widehat{d}_s - d_s^*, \frac{1}{n}\Delta_s^\intercal \epsilon\rangle \\
&\leq \frac{1}{2}\sum_s \|\widehat{d}_s - d_s^*\|_1 \|\frac{1}{n}\Delta_s^\intercal \epsilon\|_\infty \\
&\leq \frac{1}{2}\lambda \sum_s \|\widehat{d}_s - d_s^*\|_1
\end{aligned}$$

Continuing, we have

$$\frac{1}{2n}\|\sum_s \Delta_s(\widehat{d}_s - d_s^*)\|_2^2 + \frac{\lambda}{2}\sum_{s\in S^c} \|\widehat{d}_s\|_1 \leq \frac{3\lambda}{2}\sum_{s\in S} \|\widehat{d}_s - d_s^*\|_1$$

Inside the loss function, we have Δv where Δ is a $n \times n \times p$ Tensor and $v = \widehat{d} - d^*$ is a $p \times n$ matrix.

The tensor multiplication can be re-written

$$\sum_{s,i} \Delta_{si} v_{si} = \widetilde{\Delta}\widetilde{v} \qquad \text{where } \widetilde{\Delta} \text{ is } n \times np,\ \widetilde{v} \text{ is } np\text{-vector.}$$

5.4 Consistency

First we have to establish framework and notation. The optimization can be written in many different ways, with or without constraint.

$$\frac{1}{2n}\|Y - \sum_{s=1}^{p} \widehat{f}_s(X_s)\|_2^2 + \sum_{s=1}^{p} \|\partial \widehat{f}_s(X_s)\|_2 \qquad \widehat{f}_s \text{ convex}$$

$X \in \mathbb{R}^{n\times p}$ and $X_s \in \mathbb{R}^n$ where $s = 1, ..., p$. $\widehat{f}_s(X_s) = [\widehat{f}_s(X_{is})]_i$ is also a n-dim vector.

Step 1:

Transforming the objective. We claim that each $\widehat{f}_s$ is associated with pairs of real numbers $\{\widehat{\alpha}_{sk}, \widehat{\beta}_{sk}\}_k$ such that $\widehat{f}_s(X_{ks}) = \widehat{\alpha}_{sk} + \widehat{\beta}_{sk} X_{ks}$.

The alphas and betas must satisfy $\widehat{\alpha}_{sk'}, \widehat{\beta}_{sk'} = \text{argmax}_k\, \widehat{\alpha}_{sk} + \widehat{\beta}_{sk} X_{k's}$.

Our starting point will be the same as before.

$$Y = \sum_{s=1}^{p} f_s^*(X_s) + \epsilon$$

$$\frac{1}{2n}\|Y - \sum_{s=1}^{p} \widehat{f}_s(X_s)\|_2^2 + \sum_{s=1}^{p} \|\partial \widehat{f}_s(X_s)\|_{2,1} \geq \frac{1}{2n}\|Y - \sum_{s=1}^{p} f_s^*(X_s)\|_2^2 + \sum_{s=1}^{p} \|\partial f_s^*(X_s)\|_2$$

Note to self: f_s^*'s are not unique. Will this be a problem?

We can do algebra and get the following conclusion.

$$\frac{1}{2n}\left\{\|\sum_s f_s^*(X_s) - \sum_s \widehat{f}_s(X_s)\|_2^2 + \sum_s \langle f_s^*(X_s) - \widehat{f}_s(X_s), \epsilon\rangle\right\} \leq \lambda \sum_s \|\partial f_s^*(X_s)\|_2 - \lambda \sum_s \|\partial \widehat{f}_s(X_s)\|_2$$

Cross-term. We fix s.

$$\begin{aligned}
&\langle f_s^*(x_s) - \widehat{f}_s(x_s), \epsilon\rangle \\
=&\sum_{k=1}^{n} [\alpha_{sk}^* + \beta_{sk}^* x_{ks} - \widehat{\alpha}_{sk} - \widehat{\beta}_{sk} x_{ks}]\epsilon_k \\
=&\sum_{k=1}^{n} [\alpha_{sk}^* - \widehat{\alpha}_{sk}]\epsilon_k + [\beta_{sk}^* - \widehat{\beta}_{sk}]\epsilon_k x_{ks} \\
\leq&\|\alpha_s^* - \widehat{\alpha}_s\|_2 \|\epsilon\|_2 + \|\beta_s^* - \widehat{\beta}_s\|_2 \|x_s\epsilon\|_2 \quad x_s\epsilon \text{ is a } n\text{-dim vector}
\end{aligned}$$

It follows that $\|\epsilon\|_2$ and $\|x_s\epsilon\|_2$ are both on the order of $O(\sqrt{n})$.

Here lies the difficulty. In vanilla Lasso, there is only "one" parameter vector, β. Therefore, the inner product is easy to bound:

$$\frac{1}{n}\langle X\Delta, \epsilon\rangle = \langle \Delta, \frac{1}{n} X^\mathsf{T}\epsilon\rangle$$

And $\frac{1}{n} X^\mathsf{T}\epsilon$ is the empirical average of samples from a mean-zero distribution.

In new Lasso, the problem is that there are "as many" parameter vectors as there are samples. Let i index a sample and let s index a dimension.

$$\begin{aligned}
&\frac{1}{n}\sum_i X_i^\mathsf{T}\Delta_i\epsilon_i \\
=&\frac{1}{n}\sum_i\sum_s X_{is}\Delta_{is}\epsilon_i \\
=&\frac{1}{n}\sum_s\sum_i \Delta_{is}(X_{is}\epsilon_i)
\end{aligned}$$

The problem now is that we are forced to take the norm of $(\Delta_{is}X_{is})_i$, which is of too large of an order.

Continuing.

First, we have to formally define $\|\partial f_s^*(X_s)\|_2$, which we define as $\|\alpha_s^*\|_2 + \|\beta_s^*\|_2$.

Therefore, by setting λ appropriately, the derivation becomes:

$$\frac{1}{2n}\|\sum_s f_s^*(X_s) - \sum_s \widehat{f}_s(X_s)\|_2^2 - \frac{\lambda}{2}\|f^* - \widehat{f}\|_{2,1} \leq \lambda\|f^*\|_{2,1} - \lambda\|\widehat{f}\|_{2,1} \quad \text{note } \|f^*\|_{2,1} \equiv \sum_s \|f_s^*\|_2.$$

$$\frac{1}{2n}\|\sum_s f_s^*(X_s) - \sum_s \widehat{f}_s(X_s)\|_2^2 \leq \frac{3\lambda}{2}\|f^* - \widehat{f}_S\|_{2,1} - \frac{\lambda}{2}\|\widehat{f}_{S^c}\|_{2,1}$$

6 Simultaneous Update for h and S

6.1 Stopping Criteria

If we can condense the updates, then it is possible to have a sensible stopping criteria for ADMM.

The stopping criteria check the fulfillment of the stationarity condition and the primal feasibility condition.

6.2 The Derivations

The matrix of the form $\sum_{ij}(\mathbf{e}_j - \mathbf{e}_i)(\mathbf{e}_j - \mathbf{e}_i)^\mathsf{T}$ has an explicit form:

It is $2n-2$ on the diagonal and -2 everywhere else.

We do this for the vanilla implementation–no quadratic term, no additive asssumption etc.

We must solve the following linear equation:

$$(\frac{1}{\mu}\mathbf{I} + 2L)\mathbf{h} = \left(\frac{1}{\mu}\mathbf{y} + \sum_{ij}(\mathbf{e}_j - \mathbf{e}_i)((x_j - x_i)^\mathsf{T}B_i + S_{ij} - \frac{1}{\mu}W_{ji})\right)$$

$$S_{ji} = h_j - h_i - B_i^\mathsf{T}(x_j - x_i) + \frac{1}{\mu}W_{ji}$$

L is a matrix of the form $D - A$ where A is the all ones matrix and D is the degree matrix of A. Therefore, L is all negative ones everywhere except for the diagonal, where it is $n-1$.

Step 1. We separate the variables.

$$(\frac{1}{\mu}\mathbf{I} + 2L)\mathbf{h} - \sum_{ij}(\mathbf{e}_j - \mathbf{e}_i)S_{ij} = \left(\frac{1}{\mu}\mathbf{y} + \sum_{ij}(\mathbf{e}_j - \mathbf{e}_i)((x_j - x_i)^\mathsf{T}B_i - \frac{1}{\mu}W_{ji})\right)$$

$$S_{ji} - h_j + h_i = -B_i^\mathsf{T}(x_j - x_i) + \frac{1}{\mu}W_{ji}$$

Step 2. We describe the matrix form.

$$\boxed{\text{Big Matrix}}\begin{bmatrix}\mathbf{h}\\ \mathbf{S}\end{bmatrix} = \boxed{\text{Big Vector}}$$

The Big Matrix is $(n+n^2)\times(n+n^2)$ and is block structured:

$$\begin{bmatrix} \mu^{-1}\mathbf{I}+2L & \text{Operator h-S} \\ \text{Operator S-h} & \mathbf{I}_{|S|} \end{bmatrix}$$

Operator h-S is $n\times n^2$ and represented by $-\sum_{ij}(\mathbf{e}_j-\mathbf{e}_i)S_{ji}$, which is the output vector when we multiply by S

Operator S-h is $n^2\times n$ and represented by $-h_j+h_i$, which is the (j,i)-th entry of the output n^2-vector when we multiply by h.

We must also describe what the Big Vector is. Its form can be directly read from the equation array in step 1.

Step 3. We apply schur complement to invert.

Schur complement, abstractly written, is the following identity:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} I_p & 0 \\ -D^{-1}C & I_q \end{bmatrix} \begin{bmatrix} (A-BD^{-1}C)^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \begin{bmatrix} I_p & -BD^{-1} \\ 0 & I_q \end{bmatrix}$$

In our situation, D is identity.

Step 4. We perform step 1 of the Schur complement

We now apply a $n+n^2\times n+n^2$ matrix onto a $n+n^2$-vector and describe the resulting vector. We do not want to store a large matrix in memory.

Let $\begin{bmatrix} \mathbf{h}' \\ \mathbf{S}' \end{bmatrix}$ be the input vector.

The result is

$$\begin{bmatrix} \mathbf{h}'+\sum_{ij}(\mathbf{e}_j-\mathbf{e}_i)\mathbf{S}'_{ji} \\ \mathbf{S}' \end{bmatrix}$$

Step 5. We perform step 2 of the Schur complement

We must first compute the composition Operator h-S $\circ$ Operator S-h.

Let $\mathbf{h}'$ be a vector, then after Operator S-h, we have a matrix whose (j,i) entry is $-\mathbf{h}'_j+\mathbf{h}'_i$.

After an application of Operator h-S, we now have a vector of the form $-\sum_{ij}(\mathbf{e}_j-\mathbf{e}_i)(-\mathbf{h}'_j+\mathbf{h}'_i)$.

That can be re-written as: $\sum_{ij}(\mathbf{e}_j-\mathbf{e}_i)(\mathbf{e}_j-\mathbf{e}_i)^{\mathsf{T}}\mathbf{h}'$.

Hence, the combined operator is $2L$.

Because A is $\mu^{-1}\mathbf{I}+L$, the second step is just a multiplication of the first part by μ and an identity operation on the second part.

Thus, continuing our calculation from the previous step, the vector $\mathbf{h}'$ and $\mathbf{S}'$ have now become:

$$\begin{bmatrix} \mu\mathbf{h}'+\mu\sum_{ij}(\mathbf{e}_j-\mathbf{e}_i)\mathbf{S}'_{ji} \\ \mathbf{S}' \end{bmatrix}$$

Step 6. We perform step 3 of the Schur complement

We must now apply negative Operator S-h onto $\mu\mathbf{h}'+\mu\sum_{ij}(\mathbf{e}_j-\mathbf{e}_i)\mathbf{S}'_{ji}$

Applying Operator S-h onto $\mathbf{h}'$ results in $[-\mathbf{h}'_s+\mathbf{h}'_t]_{s,t}$

The second part is harder. A naive application gives $[\sum_{ij}(\mathbf{e}_j(s)-\mathbf{e}_i(s)-(\mathbf{e}_j(t)-\mathbf{e}_i(t)))\mathbf{S}'_{ji}]_{s,t}$

We may distribute the S'_{ji} into the differences and get:

$$\sum_i S'_{si}-\sum_j S'_{js}-\sum_i S'_{ti}+\sum_j S'_{jt}$$

To check for correctness, we remember that the constraints are, in order, $j = s$, $i = s$, $j = t$, $i = t$.

Implementation Note: We have: rowsum(s)-colsum(s)-rowsum(t)+colsum(t). For efficient implementation, we first compute all row and column sums of S'. We then construct a matrix whose (s,t)-th entry is $rowsum(s) - rowsum(t)$.

7 ADMM with Gap and Quadratic Term

Let B be a $p \times n$ matrix, and let D also be a $p \times n$ matrix. C is then a $p \times 2n$ matrix; let $C = [C_B, C_D]$.

$$\begin{aligned}
\min_{\mathbf{h},\mathbf{B},\mathbf{C},\mathbf{S},\mathbf{D}} & \frac{1}{2n}\sum_{i=1}^{n}(y_i - h_i)^2 + \lambda\|C\|_{\infty,1} + \gamma\frac{1}{n}\sum_{ij} S_{ji} \\
\text{s.t.} & C_B = B, C_D = D, S_{ji} \geq 0 \\
& S_{ji} = h_j - h_i - B_i^{\mathsf{T}}(X^j - X^i) - \sum_{k=1}^{p} D_{ik}(X_k^j - X_k^i)^2
\end{aligned}$$

The corresponding ADMM objective is similar to before:

$$\begin{aligned}
\min_{\mathbf{h},\mathbf{B},\mathbf{C},\mathbf{D},\mathbf{S},\mathbf{W},\mathbf{M}} & \frac{1}{2}\sum_{i=1}^{n}(y_i - h_i)^2 + \lambda\|C\|_{\infty,1} + \gamma\sum_{ij} S_{ji} \\
& + \sum_{ij}(W_{ji} \cdot [h_j - h_i - B_i^{\mathsf{T}}(X^j - X^i) - \sum_{k=1}^{p}\mathbf{D}_{ik}(X_k^j - X_k^i)^2 - S_{ji}]) \\
& + \frac{\mu}{2}\sum_{ij}(h_j - h_i - B_i^{\mathsf{T}}(X^j - X^i) - \sum_{k=1}^{p}\mathbf{D}_{ik}(X_k^j - X_k^i)^2 - S_{ji})^2 \\
& + \operatorname{tr}(M^{\mathsf{T}}(C - [B, D])) + \frac{\mu}{2}\|C - [B, D]\|^2 \text{ s.t. } S_{ji} \geq 0
\end{aligned}$$

I need to introduce additional columns to C, M, introduce D. The updates for $\mathbf{h}$ and $\mathbf{B}$ and $\mathbf{W}$ are easy. The updates for $\mathbf{C}, \mathbf{W}, \mathbf{M}$ are easy.

7.1 Update for C

The original update was:

$$\min_{\mathbf{C}} \lambda\|\mathbf{C}\|_{\infty,1} + \frac{\mu}{2}\|\mathbf{C} - (B - \frac{1}{\mu}M)\|_2^2$$

$\mathbf{C}$ is thus $\mathbf{C}$ subtracts a projection of $B - \frac{1}{\mu}M$ onto $\|\cdot\|_1$ ball of radius $\frac{\lambda}{\mu}$.
The new update is

$$\min_{\mathbf{C}} \lambda\|\mathbf{C}\|_{\infty,1} + \frac{\mu}{2}\|\mathbf{C} - ([B, D] - \frac{1}{\mu}M)\|_2^2$$

Same basic code. No change.

We take a brief digression to discuss how to solve problems of the form:

$$\min_{z\in\mathbb{R}^M} \|z-x\|_2^2 + \nu\|z\|_q$$

The subgradient optimality condition states that

$$x - z \in \frac{\nu}{2}\partial\|z\|_q$$

Define $\Psi^*(y) = \sup_z \langle z, y\rangle - \Psi(z)$ as the Fenchel conjugate of $\Psi(z) = \frac{\nu}{2}\|z\|_q$.
FACT: For any Fenchel conjugate pairs f, g, it holds that $y \in \partial f(x)$ iff $x \in \partial g(y)$.
Therefore, we have that

$$z \in \partial\Psi^*(x-z) \Rightarrow x \in x - z - \partial\Psi^*(x-z)$$

It is claimed that $\Psi^*(y) = \Xi_{B^{q'}(\nu/2)}(y)$ is the indicator function. Why is this? Let us look at $\sup_x \langle x, y\rangle - \Psi(x)$ for different values of y. If $\|y\|_{q'} \le \frac{\nu}{2}$, then $\frac{\nu}{2}\|x\|_q \ge \langle x, y\rangle$ for all x because $\|y\|_{q'} = \sup_x \langle \frac{x}{\|x\|_q}, y\rangle$. If $\|y\|_{q'} > \frac{\nu}{2}$, then there would exist an x such that $\langle x, y\rangle > \frac{\nu}{2}\|x\|_q$.

FACT: $y \in w + \partial\Xi_C(w)$ for an ∞-indicator function $\Xi_C(w)$ of a convex set C iff we have that $w = \arg\min_{w'\in C} \|w' - y\|_2^2$. PROOF:

Suppose the primus holds. Then $(y-w)^\mathsf{T}(w'-w) \le \Xi_C(w') - \Xi_C(w)$ for all $w' \in C$, this implies that $(y-w)^\mathsf{T}(w'-w) \le 0$ for all $w' \in C$.

The optimization can be restated with the indicator function: $\min_{w'} \|w'-y\|_2^2$ s.t. $\Xi_C(w') \le 0$. We can then form the Lagrangian.

Note: $v \in \partial\Xi_C(w)$ is equivalent to v in the tangent cone of C at w.

In any case, we establish the connection. $x - z$ is the projection of x onto the $\frac{\nu}{2}$ ball of the dual norm.

7.2 Update for B

Recall $\mathbf{B}$ is $p \times n$. The original update. B_i is column of $\mathbf{B}$

$$\min_{B_i} \sum_{j=1}^n \frac{\mu}{2}(h_j - (h_i + B_i^\mathsf{T}(X^j - X^i) + S_{ji}) + \frac{1}{\mu}W_{ji})^2 + \frac{\mu}{2}\|B_i - (\mathbf{C}_i + \frac{1}{\mu}\mathbf{M}_i)\|_2^2$$

$$B_i = \left(\mathbf{I} + \sum_{j=1}^n (X^j - X^i)(X^j - X^i)^\mathsf{T}\right)^{-1}\left(\mathbf{C}_i + \frac{1}{\mu}\mathbf{M}_i + \sum_{j=1}^n (X^j - X^i)(h_j - h_i - S_{ji} + \frac{1}{\mu}W_{ji})\right)$$

The new forms should be:

$$\min_{B_i} \sum_{j=1}^n \frac{\mu}{2}\left(h_j - (h_i + B_i^\mathsf{T}(X^j - X^i) + \sum_{k=1}^p \mathbf{D}_k^i (X_k^j - X_k^i)^2 + S_{ji}) + \frac{1}{\mu}W_{ji}\right)^2 + \frac{\mu}{2}\|B_i - (\mathbf{C}_{B,i} + \frac{1}{\mu}\mathbf{M}_{B,i})\|_2^2$$

$$B_i = \left(\mathbf{I} + \sum_{j=1}^n (X^j - X^i)(X^j - X^i)^\mathsf{T}\right)^{-1}\left(\mathbf{C}_{B,i} + \frac{1}{\mu}\mathbf{M}_{B,i} + \sum_{j=1}^n (X^j - X^i)(h_j - h_i - \sum_{k=1}^p \mathbf{D}_k^i (X_k^j - X_k^i)^2 + S_{ji} + \frac{1}{\mu}W_{ji})\right)$$

7.3 Updates for h

Before

$$\min_{\mathbf{h}} \frac{1}{2}\sum_{i=1}^{n}(y_i - h_i)^2 + \sum_{j=1}^{n}\frac{\mu}{2}\left(h_j - (h_i + B_i^{\mathsf{T}}(X^j - X^i) + S_{ji}) + \frac{1}{\mu}W_{ji}\right)^2$$

$$h = \left(\frac{1}{\mu}\mathbf{I} + \sum_{ij}(e_j - e_i)(e_j - e_i)^{\mathsf{T}}\right)^{-1} \left(\frac{1}{\mu}y + \sum_{ij}(e_j - e_i)((X^j - X^i)^{\mathsf{T}}B_i + S_{ji} - \frac{1}{\mu}W_{ji})\right)$$

After

$$\min_{\mathbf{h}} \frac{1}{2}\sum_{i=1}^{n}(y_i - h_i)^2 + \sum_{j=1}^{n}\frac{\mu}{2}\left(h_j - (h_i + B_i^{\mathsf{T}}(X^j - X^i) + \sum_{k=1}^{p}\mathbf{D}_k^i(X_k^j - X_k^i)^2 + S_{ji}) + \frac{1}{\mu}W_{ji}\right)^2$$

$$h = \left(\frac{1}{\mu}\mathbf{I} + \sum_{ij}(e_j - e_i)(e_j - e_i)^{\mathsf{T}}\right)^{-1} \left(\frac{1}{\mu}y + \sum_{ij}(e_j - e_i)((X^j - X^i)^{\mathsf{T}}B_i + \sum_{k=1}^{p}\mathbf{D}_k^i(X_k^j - X_k^i)^2 + S_{ji} - \frac{1}{\mu}W_{ji})\right)$$

Note: $\sum_j e_j M_{ij}$ is the i-th row of M.

7.4 Update for S

Before

$$\min_{S_{ji}} \frac{\mu}{2}\left(h_j - (h_i + B_i^{\mathsf{T}}(X^j - X^i) + S_{ji}) + \frac{1}{\mu}W_{ji}\right)^2 \text{ s.t. } S_{ji} \geq 0$$

$$S_{ji} = \max(h_j - (h_i + B_i^{\mathsf{T}}(X^j - X^i)) + \frac{1}{\mu}W_{ji}, 0)$$

After

$$\min_{S_{ji}} \frac{\mu}{2}\left(h_j - (h_i + B_i^{\mathsf{T}}(X^j - X^i) + \sum_{k=1}^{p}\mathbf{D}_k^i(X_k^j - X_k^i)^2 + S_{ji}) + \frac{1}{\mu}W_{ji}\right)^2 + \gamma S_{ji} \text{ s.t. } S_{ji} \geq 0$$

$$S_{ji} = \max(h_j - (h_i + B_i^{\mathsf{T}}(X^j - X^i) + \sum_{k=1}^{p}\mathbf{D}_k^i(X_k^j - X_k^i)^2) + \frac{1}{\mu}W_{ji} - \gamma, 0)$$

7.5 Updates for W, M

Before:

$$W_{ji} = W_{ji} + \mu(h_j - (h_i + B_i^\mathsf{T}(X^j - X^i) + S_{ji})), \;\; M = M + \mu(\mathbf{C} - [\mathbf{B}, \mathbf{D}])$$

After:

$$W_{ji} = W_{ji} + \mu(h_j - (h_i + B_i^\mathsf{T}(X^j - X^i) + \sum_{k=1}^{p} \mathbf{D}_k^i (X_k^j - X_k^i)^2 + S_{ji}))$$

$$M = M + \mu(\mathbf{C} - [\mathbf{B}, \mathbf{D}])$$

7.6 Updates for D

$$\begin{aligned}\min_D \sum_{ij} (W_{ji} \cdot [h_j - h_i - B_i^\mathsf{T}(X^j - X^i) - \sum_{k=1}^{p} \mathbf{D}_{ik}(X_k^j - X_k^i)^2 - S_{ji}]) \\ + \frac{\mu}{2} \sum_{ij} (h_j - h_i - B_i^\mathsf{T}(X^j - X^i) - \sum_{k=1}^{p} \mathbf{D}_{ik}(X_k^j - X_k^i)^2 - S_{ji})^2 \\ + \operatorname{tr}(M_D^\mathsf{T}(C_D - D)) + \frac{\mu}{2}\|C_D - D\|^2\end{aligned}$$

Simplifying things, we have

$$\min_{D_i} \sum_{j=1}^{n} \left(h_j - h_i - B_i^\mathsf{T}(X^j - X^i) - S_{ji} - D_i^\mathsf{T}\Delta_{ij} + \frac{1}{\mu} W_{ji} \right)^2 \frac{\mu}{2} + \frac{\mu}{2} \|D_i - (C_{D,i} + \frac{1}{\mu} M_{D,i})\|_2^2$$

where Δ_{ij} is a p-dim vector and $\Delta_{ij}(k) = (X_k^j - X_k^i)^2$.

Taking a cue from the update for B_i, I would guess that the update for D_i is the following:

$$\begin{aligned}D_i = &\left(\mathbf{I} + \sum_{j=1}^{n} \Delta_{ij} \Delta_{ij}^\mathsf{T} \right)^{-1} \\ &\left(\mathbf{C}_{D,i} + \frac{1}{\mu} \mathbf{M}_{D,i} + \sum_{j=1}^{n} \Delta_{ij} (h_j - h_i - S_{ji} - B_i^\mathsf{T}(X^j - X_i) + \frac{1}{\mu} W_{ji}) \right)\end{aligned}$$

8 High-Dimensional Consistency Analysis

This proof strategy first bounds denoising error and then use the fact that convex functions are Lipschitz on a compact set to bound the interpolation error.

We augment X, a $p \times n$ regressor matrix, with one more row of all ones. Then any convex function can be expressed as $\max B^\mathsf{T} X$ for some B matrix which is $(p+1) \times n$, the max-operation is taken column-wise, that is, over the n columns of B matrix.

Let $\widehat{B}$ be the empirical risk minimizer.

Note: Even without noise, there could be infinite number of convex functions that agrees with the ground truth on the n data points we sample. The equivalence class of such convex functions become smaller as n increases. We will compare our piece-wise linear convex function against an *arbitrary* piece-wise linear function, which we represent by B^*, that agrees with the true function g^* on $X_1, ..., X_n$.

$$\frac{1}{2n}\|Y - \max \widehat{B}^\mathsf{T} X\|_2^2 + \lambda \|\widehat{B}\|_{2,1} \leq \frac{1}{2n}\|Y - \max B^{*\mathsf{T}} X\|_2^2 + \lambda \|B^*\|_{2,1}$$

FACT 1: $\frac{1}{2n}\|Y - \max B^{*\mathsf{T}} X\|_2^2 = \frac{1}{2n}\|\epsilon\|_2^2$.
FACT 2:

$$\begin{aligned}\frac{1}{2n}\|Y - \max \widehat{B}^\mathsf{T} X\|_2^2 &= \frac{1}{2n}\|\max B^{*\mathsf{T}} X - \max \widehat{B}^\mathsf{T} X + \epsilon\|_2^2 \\ &= \frac{1}{2n}\left\{\|\max B^{*\mathsf{T}} X - \max \widehat{B}^\mathsf{T} X\|_2^2 + \|\epsilon\|_2^2 + \langle \max B^{*\mathsf{T}} X - \max \widehat{B}^\mathsf{T} X, \epsilon\rangle\right\}\end{aligned}$$

Putting these two facts back into the original inequality, we have:

$$\frac{1}{2n}\left\{\|\max B^{*\mathsf{T}} X - \max \widehat{B}^\mathsf{T} X\|_2^2 + \langle \max B^{*\mathsf{T}} X - \max \widehat{B}^\mathsf{T} X, \epsilon\rangle\right\} \leq \lambda \|B^*\|_{2,1} - \lambda \|\widehat{B}\|_{2,1}$$

Let us consider the cross term first. We write $[\max B^{*\mathsf{T}} X]_i = B_i^{*\mathsf{T}} X^i$, where B_i^* is a column of the B^* matrix.

Then we have, through two applications of Holder's inequality

$$\begin{aligned}\frac{1}{2n}|\langle \max B^{*\mathsf{T}} X - \max \widehat{B}^\mathsf{T} X, \epsilon\rangle| &\leq \frac{1}{2n}\sum_{i=1}^n \epsilon_i \sum_{s=1}^p (B^*_{si} - \widehat{B}_{si}) X^i(s) \\ &\leq \sum_{s=1}^p \frac{1}{2n}\sum_{i=1}^n (\epsilon_i X^i(s))(B^*_{si} - \widehat{B}_{si}) \\ &\leq \sum_{s=1}^p \|B^*_{s\cdot} - \widehat{B}_{s\cdot}\|_2 \frac{1}{2n}\sqrt{\sum_{i=1}^n \epsilon_i^2 X^{i2}(s)} \\ &\leq \|B^* - \widehat{B}\|_{2,1} \max_{s=1,..,p} \frac{1}{2n}\sqrt{\sum_{i=1}^n \epsilon_i^2 X^{i2}(s)}\end{aligned}$$

Let us temporarily impose the strong condition that the ϵ_i's are iid and that $X^i \sim N(0, I_p)$, then the second term $\frac{1}{2n}\sqrt{\sum_{i=1}^n \epsilon_i^2 X^{i2}(s)} \leq \sigma\sqrt{\frac{\log p}{n}}$ with high probability.

Returning to the main bound, we have:

$$\frac{1}{2n}\left\{\|\max B^{*\mathsf{T}} X - \max \widehat{B}^\mathsf{T} X\|_2^2 + \langle \max B^{*\mathsf{T}} X - \max \widehat{B}^\mathsf{T} X, \epsilon\rangle\right\} \leq \lambda \|B^*\|_{2,1} - \lambda \|\widehat{B}\|_{2,1}$$

Shuffling the terms, we have

$$\frac{1}{2n}\|\max B^{*\mathsf{T}} X - \max \widehat{B}^\mathsf{T} X\|_2^2 \leq \lambda \|B^*\|_{2,1} - \lambda \|\widehat{B}\|_{2,1} + \sigma\sqrt{\frac{\log p}{n}}\|B^* - \widehat{B}\|_{2,1}$$

Say $\frac{\lambda}{2} \equiv \sigma\sqrt{\frac{\log p}{n}}$. Now we break apart the norms. Let $S \subset \{1, ..., p\}$ be the set of relevant entries of B^*. Let B^*_S be a submatrix of B^* where *we restrict the rows* to only S–so B^*_S is a

FACT 1: $\|B^*\|_{2,1} - \|\widehat{B}\|_{2,1} \leq \|B^* - \widehat{B}_S\|_{2,1} + \|\widehat{B}_{S^c}\|_{2,1}$

FACT 2: $\|B^* - \widehat{B}\|_{2,1} = \|B^* - \widehat{B}_S\|_{2,1} + \|\widehat{B}_{S^c}\|_{2,1}$

These two facts combined give us an **important** intermediate result:

$$\begin{aligned}\frac{1}{2n}\|\max B^{*\mathsf{T}}X - \max \widehat{B}^{\mathsf{T}}X\|_2^2 &\leq \lambda\|B^*\|_{2,1} - \lambda\|\widehat{B}\|_{2,1} + \frac{\lambda}{2}\|B^* - \widehat{B}\|_{2,1} \\ &\leq \frac{3\lambda}{2}\|B^* - \widehat{B}_S\|_{2,1} - \frac{\lambda}{2}\|\widehat{B}_{S^c}\|_{2,1} \end{aligned} \tag{8.1}$$

In normal Lasso derivation, we would bound the RHS with the restricted eigenvalue (RE) condition. Let us make the following **assumption**, which is like the RE condition:

The KEY Assumption: Let $\{B_i^*\}$ be s-sparse vectors.

$$\frac{1}{2n}\sum_{i=1}^{n}((B_i - B_i^*)^{\mathsf{T}}X^i)^2 \geq \kappa^2\|B_S - B^*\|_F^2 \ ; \ \forall B \in \mathcal{B}$$

$$\mathcal{B} \equiv \{B \in \mathbb{R}^{(p+1)\times n} \ : \ (B_i - B_j)^{\mathsf{T}}X_i \geq 0 \forall i, j; \|B_{S^c}\|_{2,1} \leq 3\|B_S\|_{2,1}\}$$

κ is some constant, hopefully strictly positive, and hopefully does not decrease too quickly with n, p

Let us optimistically continue to tame the inequality:

$$\begin{aligned}\frac{1}{2n}\|\max B^{*\mathsf{T}}X - \max \widehat{B}^{\mathsf{T}}X\|_2^2 &\leq \frac{3\lambda}{2}\|B^* - \widehat{B}_S\|_{2,1} - \frac{\lambda}{2}\|\widehat{B}_{S^c}\|_{2,1} \\ &\leq \frac{3\lambda}{2}\sqrt{|S|}\|B^* - \widehat{B}_S\|_F - \frac{\lambda}{2}\|\widehat{B}_{S^c}\|_{2,1} \\ &\leq \frac{3\lambda}{2}\sqrt{\frac{1}{2n}}\frac{\sqrt{|S|}}{\kappa}\|\max B^{*\mathsf{T}}X - \max \widehat{B}^{\mathsf{T}}X\|_2 - \frac{\lambda}{2}\|\widehat{B}_{S^c}\|_{2,1}\end{aligned}$$

Reshuffling and discarding unnecessary terms, we get

$$\begin{aligned}\sqrt{\frac{1}{2n}}\|\max B^{*\mathsf{T}}X - \max \widehat{B}^{\mathsf{T}}X\|_2 &\leq \frac{3\lambda}{2}\sqrt{|S|}\frac{1}{\kappa} \\ &\leq \frac{3}{2}\sqrt{\frac{|S|\log p}{n}}\frac{1}{\kappa}\end{aligned}$$

Conclusion 1: We can denoise well. The estimated convex function $\widehat{g}$ and the true convex function g^* agree on the observed X^i's.

Going back to Equation 8.1 and the KEY assumption, we can derive other results:

$$\|\widehat{B}_S - B^*\|_F^2 \leq \frac{3}{2}\sqrt{\frac{|S|\log p}{n}}\frac{1}{\kappa^2}$$

Conclusion 2: Our estimated convex function $\widehat{g}$ agrees with g^* on the relevant dimensions.

$$\begin{aligned}\|\widehat{B}_{S^c}\|_{2,1} &\leq 3\|\widehat{B}_S - B^*\|_{2,1} \\ &\leq 3\sqrt{\frac{1}{2n}}\|\max B^{*\mathsf{T}}X - \max \widehat{B}^{\mathsf{T}}X\|_2 \\ &\leq \frac{9}{2}\sqrt{\frac{|S|\log p}{n}}\frac{1}{\kappa}\end{aligned}$$

Conclusion 3: Our estimated regression function is small on the irrelevant dimensions.

The next step of the analysis is to bound the interpolation error.

The Lipschitz assumption. (can be proven) We suppose that $\widehat{g}$ and g^* are both L-Lipschitz on a compact subset of $\mathbb{R}^{|S|}$, which is the space of the relevant dimensions; that implies that they must both be L-Lipschitz on a compact subset of $\mathbb{R}^{p+1}$ as well.

Note: We might only be able to prove a weaker condition described below; but that weaker condition may be sufficient.

$$|\widehat{g}(x) - \widehat{g}(x')| \leq L\|x_S - x'_S\|_2 + O\left(\sqrt{\frac{|S|\log p}{n}}\right)$$

Now we can begin to bound the interpolation error. Let x be an arbitrary point in the compact set on which $\widehat{g}, g^*$ are Lipschitz and let X^i be the sample in our training data that is closest to x.

$$\begin{aligned}\mathbb{E}|\widehat{g}(x) - g^*(x)| &\leq \mathbb{E}|\widehat{g}(x) - \widehat{g}(X^i) + \widehat{g}(X^i) - g^*(X^i) + g^*(X^i) - g^*(x)| \\ &\leq 2L\mathbb{E}\|x_S - X^i_S\|_2 + O\left(\sqrt{\frac{|S|\log p}{n}}\right)\end{aligned}$$

For many distributions, $\mathbb{E}\|x_S - X^i_S\|_2$ is on the order of $n^{-\frac{1}{s}}$.

8.1 Intuition on the KEY Assumption

Let us suppose $B^* = 0$ for now. Then the KEY assumption becomes:

$$\begin{aligned}&\frac{1}{2n}\sum_{i=1}^{n}(B_i^{\mathsf{T}}X^i)^2 \geq \kappa^2\|B_S\|_F^2 \ ; \ \forall B \in \mathcal{B} \\ &\mathcal{B} \equiv \{B \in \mathbb{R}^{(p+1)\times n} \ : \ (B_i - B_j)^{\mathsf{T}}X_i \geq 0 \forall i, j; \|B_{S^c}\|_{2,1} \leq 3\|B_S\|_{2,1};\}\end{aligned}$$

9 Looking at Subgradient

The linear inequality case:

$$\begin{aligned} &\min_{\beta} \|\beta\|_1 \\ &\text{s.t.} X\beta \leq y \end{aligned}$$

The Lagrangian optimization is

$$\max_{w} \min_{\beta} \|\beta\|_1 + w^{\mathsf{T}}(X\beta - y)$$

The KKT condition is that $-w^{\mathsf{T}}X \in \partial\|\beta\|_1$ and that $w^{\mathsf{T}}X\beta = w^{\mathsf{T}}y$.
w is a vector in $\mathbb{R}^n$ and the dual optimization is:

$$\begin{aligned} &\min_{w} w^{\mathsf{T}}y \\ &\text{s.t.} X^{\mathsf{T}}w \in [-1,1]^p \\ &\quad w \geq 0 \end{aligned}$$

FACT: Suppose I increase the penalty to λ–a large number–on the irrelevant coordinates, then the dual becomes

$$\begin{aligned} &\min_{w} w^{\mathsf{T}}y \\ &\text{s.t.} X_S^{\mathsf{T}}w \in [-1,1]^s \\ &\quad X_{S^c}^{\mathsf{T}}w \in [-\lambda,\lambda]^{p-s} \\ &\quad w \geq 0 \end{aligned}$$

FACT: if an oracle lights up the true support, then the dual becomes:

$$\begin{aligned} &\min_{w} w^{\mathsf{T}}y \\ &\text{s.t.} X_S^{\mathsf{T}}w \in [-1,1]^s \\ &\quad w \geq 0 \end{aligned}$$

This has strictly fewer constraints than full dimensional problem.

Lastly, we can also compare against

$$\begin{aligned} &\min_{w} w^{\mathsf{T}}X_S\beta^* \\ &\text{s.t.} X_S^{\mathsf{T}}w \in [-1,1]^s \\ &\quad w \geq 0 \end{aligned}$$

Let us consider an easy pair of problems:

$$\begin{aligned} &\min_{\beta} \|\beta\|_1 \\ &\text{s.t.} M\beta \leq M\beta^* \end{aligned}$$

$$\min_{\beta_S} \|\beta_S\|_1$$
$$\text{s.t.} M_S\beta_S \leq M_S\beta_S^*$$

where M_S is a subset of the columns of M.

Note that $M\beta = M_S\beta_S + M_{S^c}\beta_{S^c}$.

10 Using Second-order Information