1 (5/10/2013) Meeting Notes

1.1 Sparsistency

First, we briefly review our inner approximation short-hand notation:

$$f_k(X_{ki}) = \Delta_k d_k - c_k \mathbf{1}$$

where Δ_k is a $n \times n$ matrix: $\Delta_{k(ij)} = \begin{cases} 0 & \text{if } order(i) \leq j \\ X_{ki} - X_{k(j)} & \text{else} \end{cases}$

d_k is a n-dimensional vector. $d_{k1} = \beta_{k1}$, $d_{ki} = \beta_{ki} - \beta_{k(i-1)}$. Therefore, all coordinates of d_k except the first are positive.

c_k is a constant to ensure that $\sum_i f_k(X_{ki}) = 0$, that the fitted function is centered.

We will, for convenient, use the penalty $h(d_k) = |d_{k1}| + \sum_{i=2}^{n} d_{k2}$. The penalty that correspond to actual l_∞ norm is $h'(d_k) = \max(|d_{k1}|, d_{k1} + \sum_{i=2}^{n} d_{k2})$. We claim however that $h'(d_k) \leq h(d_k) \leq 2h'(d_k)$; this is because $h(d_k)$ is at most magnitude of the left-most derivative plus the right-most derivative.

1.1.1 Proof

FACT: A bounded convex function on a compact set is Lipschitz. The Lipschitz constant depends on $\max_{x \in C} f(x)$

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.

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

$$\frac{1}{2n} \|Y - \sum_{k=1}^{p} \widehat{f_k}(X_k)\|_2^2 + \lambda \sum_{k=1}^{p} \|\partial \widehat{f_k}(X_k)\|_\infty \qquad \widehat{f_k} \text{ convex and } \sum_i \widehat{f_k}(x_{ik}) = 0.$$

We note that using our reformulation, we can also write our objective as

$$F(d_1, ..., d_p) = \frac{1}{2n} \|Y - \sum_{k=1}^{p} \Delta_k d_k + c_k \mathbf{1}\|_2^2 + \lambda \sum_{k=1}^{p} (|d_{k1}| + d_{k2} ... d_{kn}).$$

We can also write the Lagrangian as

$$L(d_1, ..., d_p) = \frac{1}{2n} \|Y - \sum_{k=1}^{p} (\Delta_k d_k + c_k)\|_2^2 + \lambda \sum_{k=1}^{p} (|d_{k1}| + d_{k2} ... d_{kn}) - \sum_{k=1}^{p} \mu_k^\mathsf{T} d_k - \sum_{k=1}^{p} \gamma_k (c_k - \frac{1}{n} \mathbf{1}^\mathsf{T} \Delta_k d_k)$$

where for every k, μ_k has to be a positive vector.

ASSUMPTIONS:

1. $Y = f(X_S) + \epsilon$ for some Lipschitz function f and sub-gaussian ϵ. Y also has empirical mean 0. (note that the n coordinates of Y are no longer independent, but that's ok, it is still a sub-gaussian random vector)
2. The relevant and irrelevant dimensions X_S, X_{S^c} are independent. (We can relax this into a deterministic irrepresentability condition)

CLAIM: If we set $\lambda = O\sqrt{\frac{\log(np)}{n}}$, then (1) $\widehat{f}_{S^c} = 0$ with high probability and (2) $\widehat{f}_S$ is the solution to the *restricted* regression $\frac{1}{2n}\|Y - \sum_{s\in S}\widehat{f}_s(X_s)\|_2^2 + \lambda \sum_{s\in S}\|\partial \widehat{f}_s(X_s)\|_\infty$

NOTE:

1. We will assume for now, though I think it can be proven, that $\widehat{f}_k$ all have, with high probability, bounded Lipschitz constant independent of n, p.
2. We made very weak assumption on the true function f but very strong assumptions on the covariates. I think we can trade-off the two: we can still satisfy the deterministic irrespresentability condition with stronger assumptions on the true f and weaker assumptions on the covariates.
3. The proof is similar to Primal Dual Witness but we take gradient with respect to one dimension at a time. This is much more convenient for our case.

PROOF: We will show that there exists a sparse solution and that the irrelevant dual variables for this solution has magnitude strictly < 1. We will later argue that all solutions must then be sparse also.

Define $\widehat{f}$ as the solution described in the claim. We will prove stationarity, complementary slackness, on the Lagrangrian.

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

Define residue $\widehat{r} = Y - \sum_{s\in S}\widehat{f}_s(X_s)$.

We observe that $\widehat{r}$ is independent of X_{S^c} because $\widehat{f}_S$ is the solution to the restricted regression and is therefore independent of X_{S^c}. $\widehat{r}$ is also a sub-gaussian mean-zero random vector. (empirically mean 0 in fact)

Using the reformulation, we can re-write the Lagrangian L, in term of just f_k, as the following.

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

First, note that by definition as solution of the restricted regression, for $k \in S$, $\widehat{f}_k$ 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.

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

In the derivatives, $\mathbf{u}$ is a n-vector whose first coordinate is $\partial|d_{k0}|$ 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.

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

Note that $\mathbf{u}$ became $\mathbf{1}$.

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

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

We need strict inequality to later argue that ALL solutions are sparse.

Now, we recall that Δ_k and $\widehat{r}$ are independent, that Δ_k is a positive matrix with (whp) bounded entries and that $\widehat{r} = r_1 - \bar{r}_1$ where r_1 is a n-vector of independent zero-mean subgaussian random variables and $\bar{r}_1$ is the empirical mean of r_1.

Every coordinate of $\frac{1}{n}\Delta_k^\mathsf{T} r_1$ must then be of order $O(\sqrt{\frac{1}{n}})$ by standard concentration and $\frac{1}{n}\Delta_k^\mathsf{T}\mathbf{1}\bar{r}_1$ must also be of order $O(\sqrt{\frac{1}{n}})$.

We need to take union bounds across all n coordinates of $\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}$ as well as all $p - |S|$ irrelevant dimensions. Therefore, setting $\lambda = O\sqrt{\frac{\log(np)}{n}}$ suffices.

1.2 Further Developments

Suppose we impose an upper bound $|d_{k1}| + \sum_{i=2}^n d_{ki} \leq C_{max}$.

Then $\widehat{f}_s(x_{si})$ is upper bounded by $\log nC_{max}$.

We can look at $\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}$

2 (5/3/2013) Meeting Notes

2.1 Short-hand

The objective is

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

Let us try to simplify this form. 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 $\widehat{f}_s(x_{si}) = c_s + \widehat{\beta}_s^{\mathsf{T}} D_{si}$ where $D_{si} \in \mathbb{R}^{n-1}$, and is a vector

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

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

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 $i > 1$.

Notation: We will also define Δ_s as a $(n \times n-1)$ matrix whose i,j-th entry is $\Delta(j, x_{si})$. With this short-hand, we can write $\widehat{f}_s(x_s) = \Delta_s \widehat{d}_s$.

Row i of Δ_s is the vector $[\Delta(1, x_{si}), \Delta(2, x_{si})..., \Delta(n-1, x_{si})]$. It is $n-1$ dimensional.

Column j of Δ_s is the vector $[\Delta(j, x_{s1}), \Delta(j, x_{s2}), ..., \Delta(j, x_{sn})]$.

Examples: $\Delta(j, x_{si}) = \max(x_{si} - x_{s(j)}, 0)$

2.2 Consistency with Short-hand

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

First, what is the objective?

$$\|\partial \widehat{f}_s(X_s)\|_\infty = \|\widehat{\beta}_{si}\|_\infty$$

For simplicity, we can use a different but very similar constraint: $|\widehat{d}_{s0}| + \sum_{i=1}^n \widehat{d}_{si}$.

The objective is therefore

$$\frac{1}{2n}\|Y - \sum_{s=1}^p \Delta_s \widehat{d}_s\|_2^2 + \sum_{s=1}^p \left[|\widehat{d}_{s0}| + \sum_{i=1}^n \widehat{d}_{si} \right].$$

3 (4/19/2013) Meeting Notes

3.1 Main Result

Lemma 3.1. *Let $f(x_1, x_2)$ be a 2D convex function defined on $C = [0,1]^2$. Suppose $h(x_1), g(x_2)$ minimizes $\int_C |f(x_1, x_2) - h(x_1) - g(x_2)|^2 dx$ among all univariate convex functions.*

$h(x_1)$ is not a constant function iff $\int_{x_2} f(x_1, x_2) dx_2$ is not constant with respect to x_1.

This following proposition combined with Lemma 3.1 gives the desired result.

Proposition 3.1. *Let $f(x_1, x_2)$ be a bivariate convex function defined on $C = [0,1]^2$. Suppose f is* continuous *on C and is twice differentiable. (It must be continuous on the interior of C)*

The following are equivalent:

1. *For all x_2, $f(\cdot, x_2)$ as univariate functions of x_1 are identical.*
2. *For all x_2, the integrals $\int_{x_1} f(x_1, x_2) dx_1$ are identical.*

Proof. The first condition trivially implies the second condition.

Suppose that for all x_2, the integral $\int_{x_1} f(x_1, x_2) dx_1$ is, without loss of generality, equal to 0.

For a fixed x_1, define the *right directional derivative* as

$$g(x_1) \equiv \lim_{x_2 \to 0^+} \frac{f(x_1, x_2) - f(x_1, 0)}{x_2}$$

This limit has to exist by assumption of continuity and by the fact that $f(x_1, \cdot)$ is a univariate convex function with respect to x_2.

Fact 1: $\int_{x_1} g(x_1)dx_1 = 0$ because $\int_{x_1} f(x_1, x_2)dx_1 = 0$ for all x_2. We can invoke the bounded convergence theorem to make this deduction more formal.

Fact 2: Because f is convex, $g(x_1)$ is also a component of the 2-dimensional subgradient $\partial f(x_1, 0)$, that is, $g(x_1) = \partial f(x_1, 0)^\mathsf{T}\mathbf{e}_2$.

Therefore, for all x_1, $f(x_1, x_2) \geq f(x_1, 0) + x_2 g(x_1)$.

Since $\int_{x_1} f(x_1, 0) + x_2 g(x_1)dx_1 = 0$ and $\int_{x_1} f(x_1, x_2)dx_1 = 0$, it must be that $f(x_1, x_2) = f(x_1, 0) + x_2 g(x_1)$ for all x_1.

The Hessian of $f(x_1, x_2)$ at a point x_1, x_2 is of the form

$$\begin{bmatrix} a & b \\ b & 0 \end{bmatrix}$$

where $a \equiv \partial^2 f(x_1, 0)/\partial^2 x_1$ and $b \equiv \partial g(x_1)/\partial x_1$. Note that a, b depend on x_1 only. The characteristic polynomial is $-\lambda(a - \lambda) - b^2$.

Apply the quadratic formula to solve the characteristic polynomial, we get the solutions $\frac{a \pm \sqrt{a^2 + 4b^2}}{2}$. The Hessian is therefore NOT positive-semidefinite unless $b = 0$.

Thus, $g'(x_1) = 0$ for all x_1, and $g(x_1)$ is necessarily 0.

□

3.2 Lemmas

Lemma 3.2. *Let $f(x)$ be a univariate function, integrable. We approximate $f(x)$ with a constant: $c^* \equiv \min_c \int_C |c - f(x)|^2 dx$.*
Define $A = \int_C 1dx$ as the measure of C,

$$\textit{then we have that} \quad c^* = \frac{1}{A}\int_C f(x)dx$$

Proof. (of lemma)

Because the objective is convex with respect to c, we can use the KKT theorem. The stationarity condition implies that

$$2\int_C c^* - f(x)dx = 0$$
$$\int_C c^* dx = \int_C f(x)dx$$

□

Let us also assume that $C = [0,1]^2$ for simplicity; C has area 1 and when restricted to either x_1 or x_2, C is a line with length 1.

We claim that $h(x_1) = \int_{x_2} f(x_1, x_2)dx_2 - \alpha$ for some constant α. Why? If we solve $\min_h \int_C |f(x_1, x_2) - h(x_1) - g(x_2)|^2 dx$ *without* requiring that h is a convex function, the optimal solution is, by lemma, $x_1 \mapsto \int_{x_2} f(x_1, x_2) - g(x_2)dx_2$; but because the sum of convex functions is still convex, the solution actually satisfies the convex requirement; and because $\int_{x_2} g(x_2)dx_2$ is a constant, we can define it to be α.

Therefore, $h(x_1)$ is a constant function iff for all x_1, $\int_{x_2} f(x_1, x_2)dx_2$ is a constant. This proves Lemma 3.1.

A lemma that we might not use.

Lemma 3.3. *Let f be a convex function on $C \subset \mathbb{R}^n$. Let $x, y \in C$, then for all x' on the line from x to y:*

$$f(x') - f(x) \leq \frac{f(y) - f(x)}{(y - x)}(x' - x)$$

Simply put, if we draw a line, in R^{n+1}, from $(x, f(x))$ to $(y, f(y))$, then f applied to $[x, y]$ must lie "below" this line.

4 (4/12/2013) Meeting Notes

4.1 A Summary of Current Affairs

We have tried fitting piecewise linear convex (PLC) functions, reweighted PLC functions, additive PLC functions, PLC and quadratic functions.

Out of all these methods, only reweighted PLC function succeeds in exact variable selection. The reweighted PLC function is however difficult to analyze; we may need to analyze it as we would a "thresholded" PLC function.

All of these methods however give reasonable results: for reasonably sized n, the weights of the irrelevant variables are much smaller than the weights of the relevant variables.

This suggest then that we should use a weaker criterion for success, which Minhua has already begun to do.

Definition 4.1. (Consistency-like Criterion)
Let $f(n,p,s)$ be some scaling function, i.e. $f(n,p,s) = \sqrt{\frac{s \log p}{n}}$.
We declare success if, for some constant c independent of n, p:

$$\|\widehat{\beta}_j\|_\infty \geq c \quad \forall j \in S,\ n,p \to \infty\ f(n,p,s) \to 0$$
$$\|\widehat{\beta}_j\|_\infty \to 0 \quad \forall j \in S^c$$

I don't know for what scaling function $f(n,p,s)$ does the PLC fitting satisfy the above criterion. Perhaps we can search for an answer via experiments on the cluster.

I also suggested a criterion before–stronger than the one I write now–which I think does not hold for PLC methods.

$$\lim_{n\to\infty} \frac{\|B_S\|_{1,\infty}}{\|B\|_{1,\infty}} \to 1$$

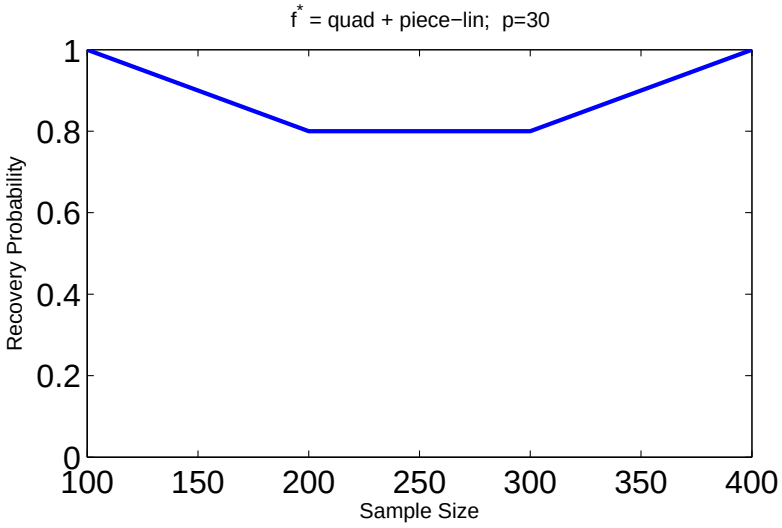
4.2 A Good Simpler Problem to Consider?

Multi-task lasso where each task has only one sample:
p dimensions, k tasks. *There is no n.*

B is $p \times k$. The rows of B are different but have same support.
$Y \in \mathbb{R}^k$. X is $k \times p$.
$Y = \text{diag}(XB) + \epsilon$

$$\min_B \frac{1}{2}\|Y - \text{diag}(XB)\|_2^2 + \lambda\|B\|_{1,\infty}$$

4.3 Reweighted PLC

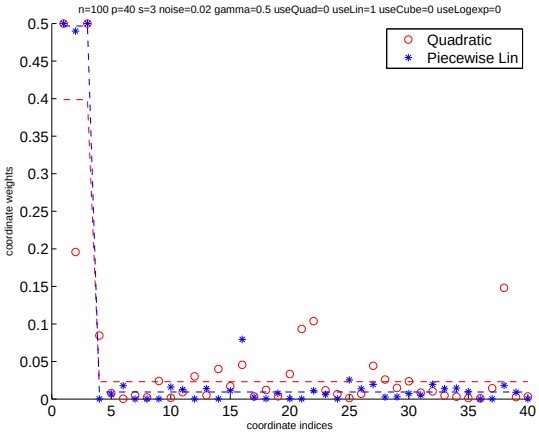


5 (4/5/2013) Meeting Notes

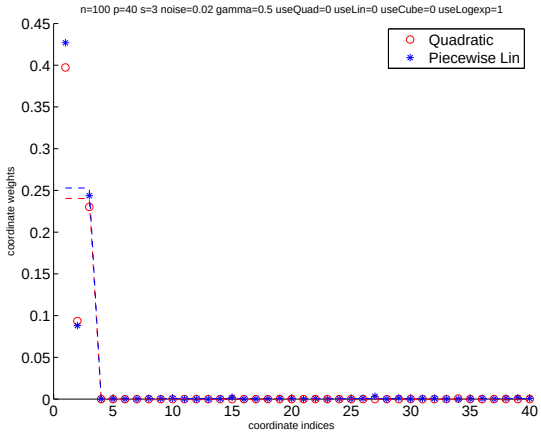
5.1 Quadratic Optimization Results

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

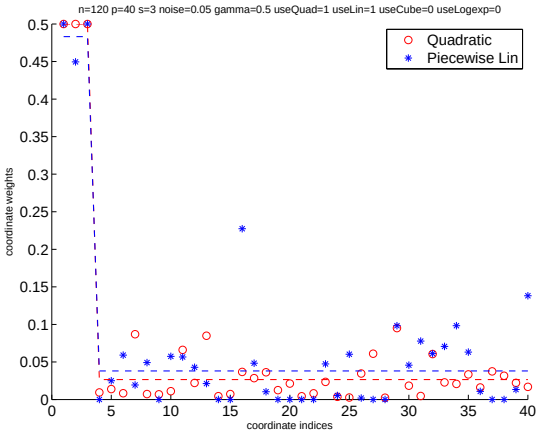
5.2 Reweighted Plot



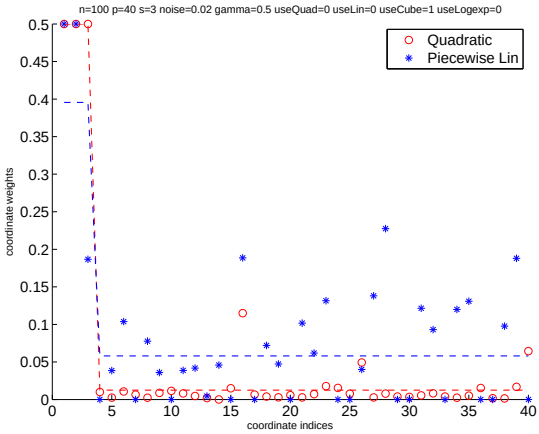
(a) f^* is piece-wise linear



(b) f^* is log-sum-exp

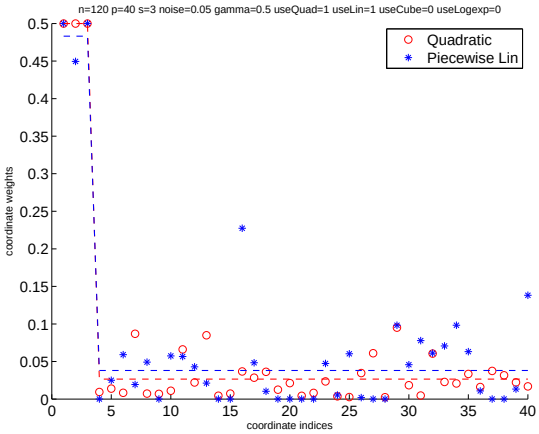


(c) f^* is a sum of quadratic and piece-wise -linear

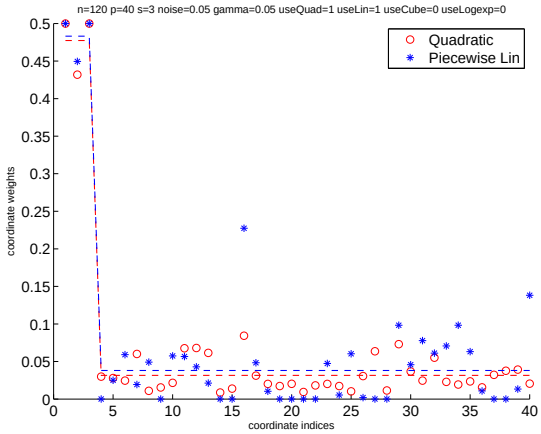


(d) f^* is a cubic

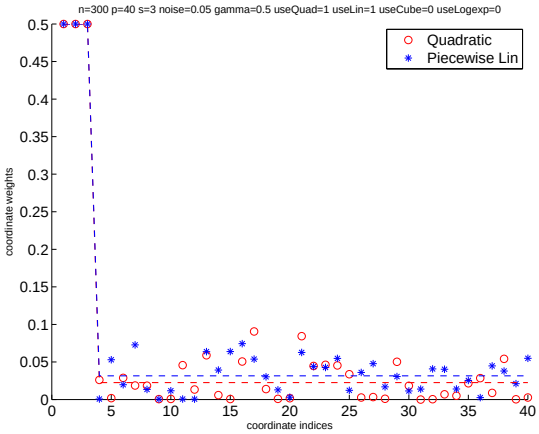
Figure 1: For all these figures, we have $n = 100, p = 40, s = 3$



(a) $n = 120, \gamma = 0.5$

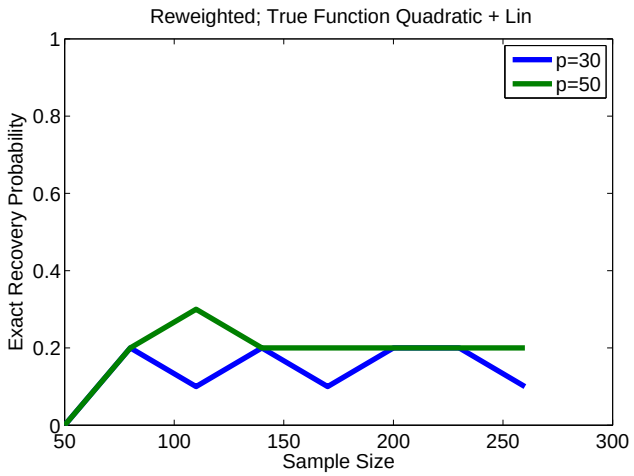


(b) $n = 120, \gamma = 0.05$



(c) $n = 300, \gamma = 0.5$

Figure 2: For all these figures, f^* is a sum of quadratic and piece-wise linear function.



6 (3/28/2013) Meeting Notes

We know that learning a sparse gradient at one point x is too formidable a task, that is why we impose the $l_\infty - l_1$ penalty to link and sync the learning of all the gradients.

The $l_\infty - l_1$ penalty seems to not be enough. One idea is to use reweighing. We brainstorm some other ideas in the following note.

6.1 K-means and Fused Lasso penalty for Piecewise-Linear Convex Functions

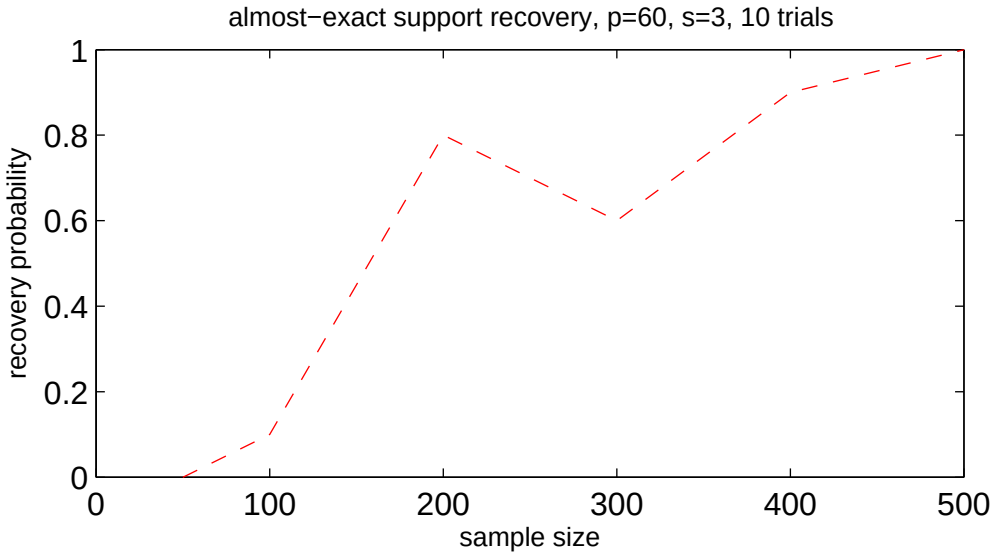
If the true function is piecewise-linear-convex, then the gradients for many of the x's are the same. Therefore, we can either use the K-means formulation or the convex fused lasso penalty $\sum_{i,j=1,\ldots,n} \|B_i - B_j\|_q$ for some norm.

Preliminary experiments with the K-means optimization suggest that this can be effective. We need the true number of "pieces" to be small and we need K-means to pick up the clusters correctly; the second requirement can be possibly waived if we use the fused lasso penalty.

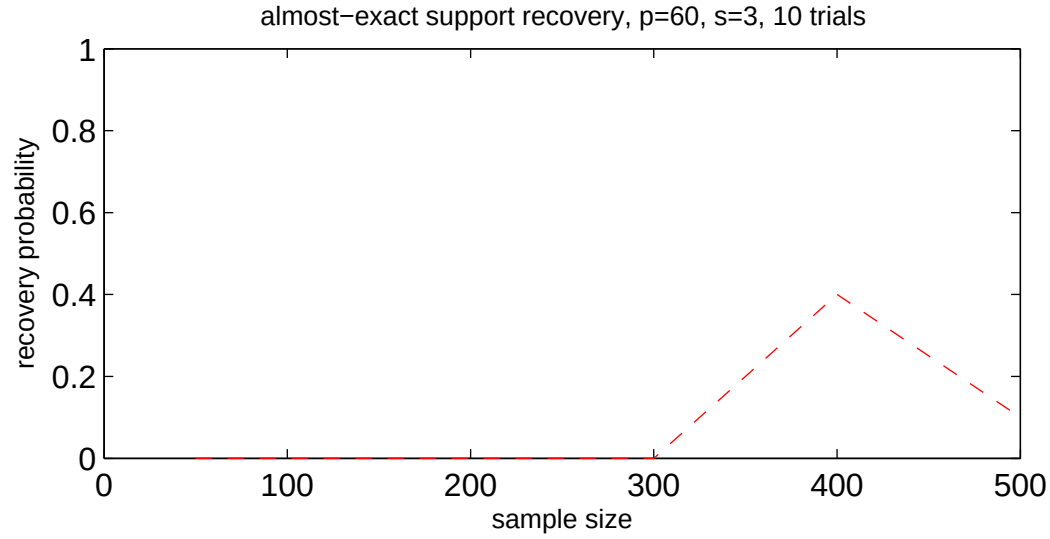
6.2 Quadratic Term

If a convex function f is strongly convex at a point x, then the linear form is a poor approximation, i.e. $f(y)$ can be much larger than $f(x) + \partial f(x)(y - x)$. The usefulness of the data point y in estimating the gradient at x diminishes as y distances itself from x and the approximation gap widens.

A natural method that arises out of this observation is to add an additional quadratic term into the optimization objective:



(a) k-means with $K* = 3$



(b) k-means with $K* = 7$

$$\min_{B,\mathbf{c}} \frac{1}{n}\sum_{i=1}^{n} |y_i - h_i|^2 + \|[B;\mathbf{c}]\|_{\infty,1} + S_{ij}$$

$$\text{s.t.} S_{ij} = h_i - h_j - B_j^{\mathsf{T}}(X^i - X^j) - \sum_{k=1}^{p} \mathbf{c}_k^j (X_k^i - X_k^j)^2$$

$$S_{ij} \geq 0$$

$$\mathbf{c}^j \geq 0$$

Each $\mathbf{c}^j$ is in $\mathbb{R}^p$ and share the same group penalty as the gradients B_j's.

The optimization is not difficult because the constraints are linear with respect to the new $\mathbf{c}^j$ variables.

7 Some Insight

Let us again consider this noiseless optimization:

$$\min_{B} \|B\|_{\infty,1}$$
$$\text{s.t. } y_j \geq y_i + B_i^{\mathsf{T}}(X^j - X^{i*}) \forall i, j$$

First, notice that, because (1) each constraint involves only one B^i and (2) the y_j's are not variables, the optimization can be decomposed into n sub-optimizations each done independently.

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_1$$
$$\text{s.t. } y_j \geq y_{i*} + \beta^{\mathsf{T}}(X^j - X^{i*}) \forall j$$

This optimization is analogous to finding a subgradient at just X^{i*}. This decomposition allows us to compare directly with the linear case because now, in both cases, we are solving for a single vector.

Let us simply the expression further. Define $\Delta \in \mathbb{R}^{p \times n}$ where the j-th row $\Delta_j = X^j - X^{i^*}$, and define $\mathbf{v} \in \mathbb{R}^n$ where $\mathbf{v}_j = y_j - y^{i*}$. Now the optimization takes on a familiar form.

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_1$$
$$\text{s.t. } \mathbf{v} \geq \beta^{\mathsf{T}} \Delta$$

The optimization is similar to the linear case, but we have an inequality instead of a strict equality here. The inequality is the culprit!

The solution to the optimization will lie on the intersection of the boundary of the constraint set and the boundary of a L_1 norm ball of some minimal radius.

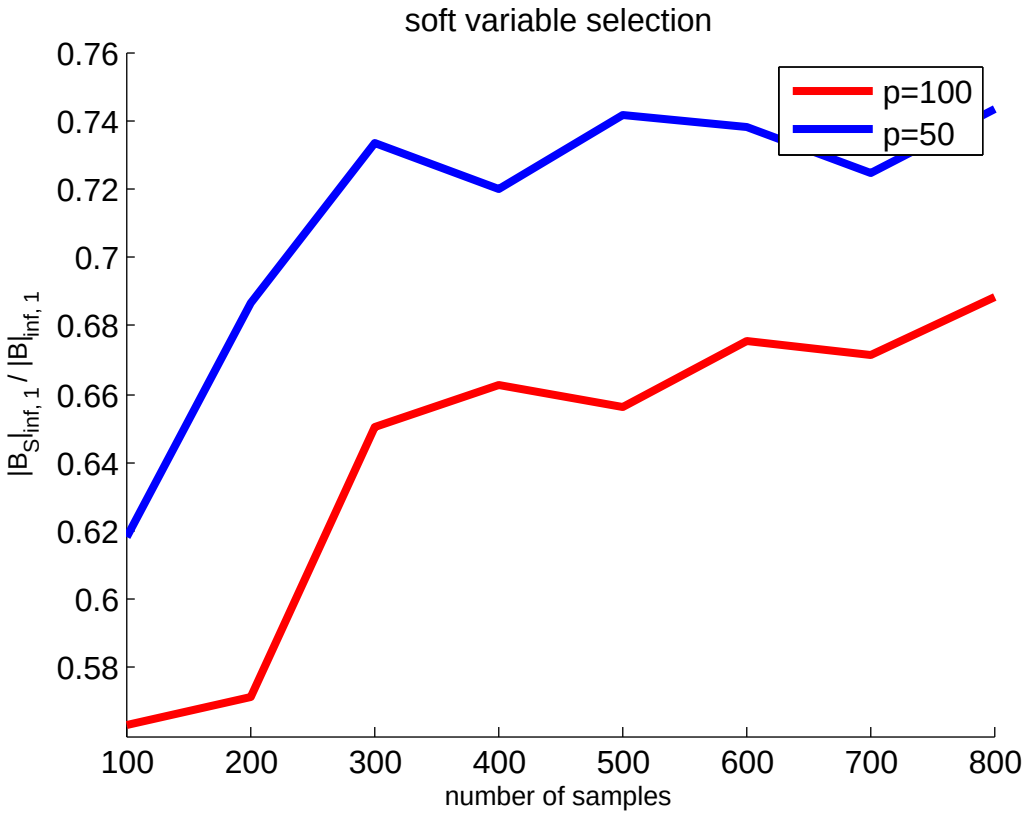
Let $\widehat{\beta}$ be the solution to the above optimization. Because $\widehat{\beta}$ lies on the boundary of the constraint set, some inequalities must be equalities. Let I be the set of inequalities which becomes equalities when we plug in $\widehat{\beta}$. Then $\widehat{\beta}$ is equivalently the solution of the following optimization:

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_1$$
$$\text{s.t. } \mathbf{v}_I = \beta^{\mathsf{T}} \Delta_I$$

where Δ_I is a $p \times |I|$ sub-matrix and $\mathbf{v}$ is a $|I|$-dimensional sub-vector. Now we see the problem: if $|I|$ is very small, then it is as if we are solving lasso with very few samples–$|I|$ number of samples instead of n!

How do we interpret I? Intuitively, I think of it as all j for which $y_j \approx y_{i*} + B_{i*}^{\mathsf{T}}(X^j - X^{i*})$. That is, it is the set of points for which the convex function is close to linear. For strongly convex functions like $\|x\|_2^2$, I would be very small. For linear functions like $\beta^{*\mathsf{T}} x$, $|I| = n$.

Therefore, this explains why our optimization gives sparse solution if the true function is linear and highly non-sparse solution if the true function is strongly convex.



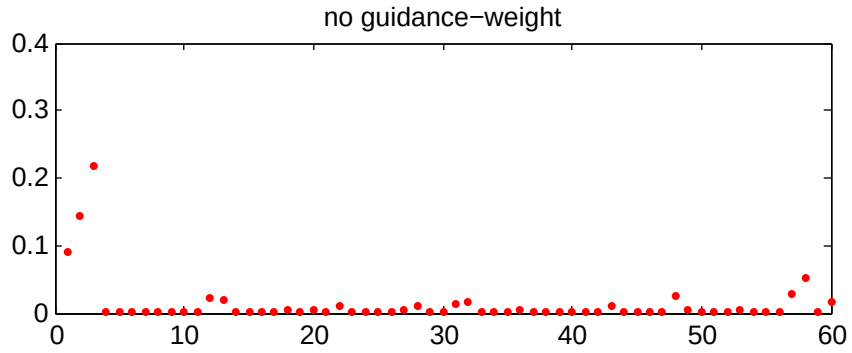
8 Experiments

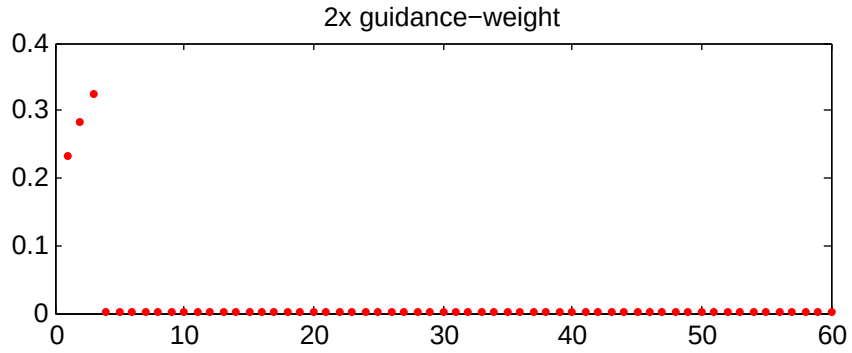
In the soft-recovery experiment, we measure $\frac{\|B_S\|_{\infty,1}}{\|B\|_{\infty,1}}$. The recovery curve plateaus. It should converge to 1 but appears to be doing so very slowly.

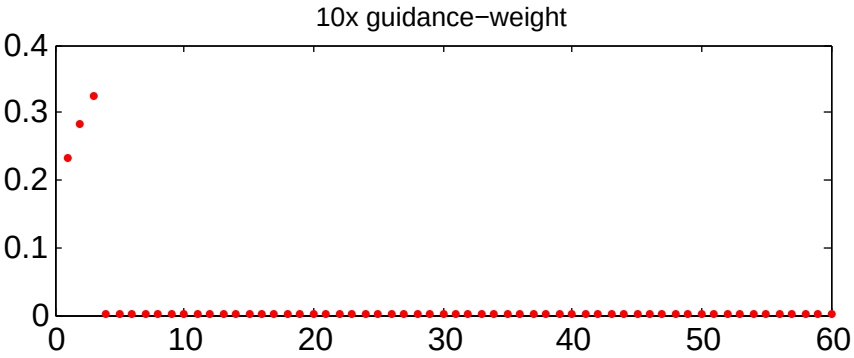
As a thought experiment, let's consider a noiseless support recovery optimization:

$$\min_B \|B\|_{\infty,1}$$
$$\text{s.t. } y_i \geq y_j + B_j^\mathsf{T}(X^i - X_j) \forall i, j$$

The optimum of this program should become exactly sparse as $n \to \infty$, but may become sparse at a very slow rate with respect to n. The penalized least-square solution then, like a heat-seeking missile thrown off target by a decoy, homes in to the non-sparse optimum rather than the actual sparse solution.







We have not implemented this optimization but can still approximate it with the penalized square error optimization by setting lambda very small.

What we see if that we must weigh the dimensions to exactly recover the support. We must use a weighed form of the objective.

$$\min_B \sum_{s=1}^{p} \lambda_s \|B_{s\cdot}\|_\infty$$
$$\text{s.t. } y_i \geq y_j + B_j^\intercal (X^i - X_j) \forall i, j$$

Where λ_s is small for $s \in S$ and large for $s \notin S$.