SUPPLEMENT TO “BEST SUBSET SELECTION VIA A MODERN OPTIMIZATION LENS”

By Dimitris Bertsimas, Angela King and Rahul Mazumder

Massachusetts Institute of Technology

This article contains additional technical material and supporting figures and tables supplementing the main paper: “Best Subset Selection via Modern Optimization Lens” by Bertsimas, King and Mazumder.

8. Additional Details for Section 2.

8.1. *Solving the convex quadratic optimization problems in main paper Section 2.3.2.* We show here that the convex quadratic optimization problems appearing in Section 2.3.2 are indeed quite simple and can be solved with small computational cost.

We first consider Problem (2.14), the computation of u_i^- which is a minimization problem. We assume without loss of generality that the feasible set of problem (2.14) is non-empty. Thus by standard results in quadratic optimization Boyd and Vandenberghe (2004), it follows that, there exists a τ such that:

$$\nabla\left(\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\tau\beta_i\right)=0,$$

where, ∇ denotes derivative wrt $\boldsymbol{\beta}$ and a $\boldsymbol{\beta}$ that satisfies the above gradient condition must also be feasible for Problem (2.14). Simplifying the above equation, we get:

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta}=\mathbf{X}'\mathbf{y}-\tau e_i,$$

where, e_i is a vector in $\Re^p$ such that its ith coordinate is one with the remaining equal to zero. Simplifying the above expression, we have

$$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2=\|(\mathbb{I}-P_X)\mathbf{y}+\tau q_i\|_2^2.$$

Above, $\mathbb{I}$ is the identity matrix of size $p\times p$ and P_X is the familiar projection matrix given by $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ [1] and $q_i=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}e_i$. Observing that $\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2=\text{UB}$, one can readily estimate τ that satisfies the above simple quadratic equation. This leads to the solution of τ, which subsequently leads to the optimal value that solves Problem (2.14). This readily leads to the optimum of Problem (2.14).

[1] Note that we assume here that $n>p$ which typically guarantees that $\mathbf{X}'\mathbf{X}$ is invertible with probability one, provided the entries of $\mathbf{X}$ are drawn from a continuous ditribution.

The above argument readily applies to Problem (2.14), for the computation of u_i^+ by writing it as an equivalent minimization problem and observing that:

$$-u_i^+ = \min_{\boldsymbol{\beta}} \quad -\beta_i \quad s.t. \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \leq \text{UB}.$$

The above derivation can also be adapted to the case of Problem (2.15). Towards this end, notice that for estimating v_i^- the above steps (for computing u_i^-) will be modified: e_i gets replaced by $\mathbf{x}_i \in \Re^p$ (the ith row of $\mathbf{X}$); and P_X denotes the projection matrix onto the column space of $\mathbf{X}$, even if the matrix $\mathbf{X}'\mathbf{X}$ is not invertible (since here, we consider arbitrary n, p).

In addition, the Problems (2.14) for the different features and (2.15) for the different samples; can be solved *completely* independently, in parallel.

8.2. *Details for Section 2.3.4 in main paper.* Note that in Problem (2.6) we consider a uniform bound on β_i's: $-\mathcal{M}_U \leq \beta_i \leq \mathcal{M}_U$, for all $i = 1, \ldots, p$. Note that some of the variables β_i may have larger amplitude than the others, thus it may be reasonable to have bounds depending upon the variable index i. Thusly motivated, for added flexibility, one can consider the following (adaptive) bounds on β_i's: $-\mathcal{M}_U^i \leq \beta_i \leq \mathcal{M}_U^i$ for $i = 1, \ldots, p$. The parameters $\mathcal{M}_U^i$ can be taken as $\max\{|u_i^+|, |u_i^-|\}$, as defined in (2.14).

More generally, one can also consider asymmetric bounds on β_i as: $u_i^- \leq \beta_i \leq u_i^+$ for all i.

Note that the above ideas for bounding β_i's can also be extended to obtain sample-specific bounds on $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle$ for $i = 1, \ldots, n$.

The bounds on $\|\widehat{\boldsymbol{\beta}}\|_1$ and $\|\mathbf{X}\widehat{\boldsymbol{\beta}}\|_1, \|\mathbf{X}\widehat{\boldsymbol{\beta}}\|_\infty$ can also be adapted to the above variable dependent bounds on β_i's.

While the above modifications may lead to marginally improved performances, we do not dwell much on these improvements mainly for the purpose of a more transparent exposition.

8.3. *Proof of Proposition 2 in main paper.*

Proof

(a) Given a set I, we define $\mathbf{G} := \mathbf{X}_I'\mathbf{X}_I - \mathbf{I}$, and let g_{ij} denote the (i, j)th entry of $\mathbf{G}$.

For any $\mathbf{u} \in \mathbb{R}^k$ we have

$$\begin{aligned}
\max_{\|\mathbf{u}\|_1=1} \|\mathbf{G}\mathbf{u}\|_1 &= \max_{\|\mathbf{u}\|_1=1} \left(\sum_{i=1}^k \left| \sum_{j=1}^k g_{ij} u_j \right| \right) \\
&\leq \max_{\|\mathbf{u}\|_1=1} \left(\sum_{i=1}^k \sum_{j=1}^k |u_j||g_{ij}| \right) \\
&= \max_{\|\mathbf{u}\|_1=1} \left(\sum_{j=1}^k |u_j| \sum_{i \neq j} |g_{ij}| \right) && (g_{jj} = 0) \\
&\leq \max_{\|\mathbf{u}\|_1=1} (\mu[k-1]\|\mathbf{u}\|_1) && \left(\sum_{i \neq j} |g_{ij}| \leq \mu[k-1] \right) \\
&= \mu[k-1].
\end{aligned}$$

(b) Using $\mathbf{X}_I'\mathbf{X}_I = \mathbf{I} + \mathbf{G}$ and standard power-series convergence (which is valid since $\|\mathbf{G}\|_{1,1} < 1$) we obtain

$$\|(\mathbf{X}_I'\mathbf{X}_I)^{-1}\|_{1,1} = \| (\mathbf{I}+\mathbf{G})^{-1} \|_{1,1} \leq \sum_{i=0}^{\infty} \|\mathbf{G}\|_{1,1}^i \leq \frac{1}{1 - \|\mathbf{G}\|_{1,1}} \leq \frac{1}{1-\mu[k-1]}. \qquad \square$$

8.4. *Proof of Theorem 2.1 in main paper.*

Proof

(a) Since $\widehat{\boldsymbol{\beta}}_I = (\mathbf{X}_I'\mathbf{X}_I)^{-1}\mathbf{X}_I'\mathbf{y}$ we have

$$\|\widehat{\boldsymbol{\beta}}\|_1 = \|\widehat{\boldsymbol{\beta}}_I\|_1 \leq \|(\mathbf{X}_I'\mathbf{X}_I)^{-1}\|_{1,1}\|\mathbf{X}_I'\mathbf{y}\|_1. \tag{8.1}$$

Note that

$$\|\mathbf{X}_I'\mathbf{y}\|_1 = \sum_{j\in I} |\langle \mathbf{X}_j, \mathbf{y}\rangle| \leq \max_{I,|I|=k} \sum_{j \in I} |\langle \mathbf{X}_j, \mathbf{y}\rangle| \leq \sum_{j=1}^k |\langle \mathbf{X}_{(j)}, \mathbf{y}\rangle|. \tag{8.2}$$

Applying Part (b) of Proposition 2 and (8.2) to (8.1), we obtain (2.10) .

(b) We write $\widehat{\boldsymbol{\beta}}_I = \mathbf{A}\mathbf{y}$ for $\mathbf{A} = (\mathbf{X}_I'\mathbf{X}_I)^{-1}\mathbf{X}_I'$. If $\mathbf{a}_i, i = 1, \ldots, k$ denote the rows of $\mathbf{A}$ we have:

$$\|\widehat{\boldsymbol{\beta}}_I\|_\infty = \max_{i=1,\ldots,k} |\langle \mathbf{a}_i, \mathbf{y}\rangle| \leq \left(\max_{i=1,\ldots,k} \|\mathbf{a}_i\|_2 \right) \|\mathbf{y}\|_2. \tag{8.3}$$

For every $i = 1, \ldots, k$ we have

$$\begin{aligned} \|\mathbf{a}_i\|_2 \leq & \max_{\|\mathbf{u}\|_2=1} \|\mathbf{A}\mathbf{u}\|_2 \\ = & \max_{\|\mathbf{u}\|_2=1} \|(\mathbf{X}_I'\mathbf{X}_I)^{-1}\mathbf{X}_I'\mathbf{u}\|_2 \\ = & \lambda_{\max}\left((\mathbf{X}_I'\mathbf{X}_I)^{-1}\mathbf{X}_I'\right) \\ = & \max\ \left\{\frac{1}{d_1}, \ldots, \frac{1}{d_k}\right\}, \end{aligned} \tag{8.4}$$

where $d_1, \ldots, d_k$ are the (nonzero) singular values of the matrix $\mathbf{X}_I$. To see how one arrives at (8.4) let us denote the singular value decomposition of $\mathbf{X}_I = \mathbf{UDV'}$ with $\mathbf{D} = \operatorname{diag}(d_1, d_2, \ldots, d_k)$. We then have

$$(\mathbf{X}_I'\mathbf{X}_I)^{-1}\mathbf{X}_I' = (\mathbf{V}\mathbf{D}^{-2}\mathbf{V}')(\mathbf{UDV}')' = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}'$$

and the singular values of $(\mathbf{X}_I'\mathbf{X}_I)^{-1}\mathbf{X}_I'$ are thus $1/d_i$, $i = 1, \ldots, k$.

The eigenvalues of $\mathbf{X}_I'\mathbf{X}_I$ are d_i^2 and from (2.8) we obtain that $d_i^2 \geq \gamma_k$. Using (8.4) we thus obtain

$$\max_{i=1,\ldots,k} \|\mathbf{a}_i\|_2 \leq \frac{1}{\sqrt{\gamma_k}}. \tag{8.5}$$

Substituting the bound (8.5) to (8.3) we obtain

$$\|\widehat{\boldsymbol{\beta}}_I\|_\infty \leq \frac{1}{\sqrt{\gamma_k}}\|\mathbf{y}\|_2. \tag{8.6}$$

Using the notation $\tilde{\mathbf{A}} = (\mathbf{X}_I'\mathbf{X}_I)^{-1}$, we have

$$\begin{aligned} \|\widehat{\boldsymbol{\beta}}_I\|_\infty = & \max_{i=1,\ldots,k} |\langle \tilde{\mathbf{a}}_i, \mathbf{X}_I'\mathbf{y}\rangle| \\ \leq & \left(\max_{i=1,\ldots,k|} \|\tilde{\mathbf{a}}_i\|_2\right) \|\mathbf{X}_I'\mathbf{y}\|_2 \\ \leq & \lambda_{\max}\left((\mathbf{X}_I'\mathbf{X}_I)^{-1}\right) \|\mathbf{X}_I'\mathbf{y}\|_2 \\ = & \left(\max_{i=1,\ldots,k} \frac{1}{d_i^2}\right) \cdot \sqrt{\sum_{j\in I} |\langle \mathbf{X}_j, \mathbf{y}\rangle|^2} \\ \leq & \frac{1}{\gamma_k}\sqrt{\sum_{j=1}^{k} |\langle \mathbf{X}_{(j)}, \mathbf{y}\rangle|^2}. \end{aligned} \tag{8.7}$$

Combining (8.6) and (8.7) we obtain (2.11).

(c) We have

$$\|\mathbf{X}_I\widehat{\boldsymbol{\beta}}_I\|_1 \leq \sum_{i=1}^{n} |\langle \mathbf{x}_i, \widehat{\boldsymbol{\beta}}_I\rangle| \leq \sum_{i=1}^{n} \|\mathbf{x}_i\|_\infty \|\widehat{\boldsymbol{\beta}}_I\|_1 = \sum_{i=1}^{n} \|\mathbf{x}_i\|_\infty \|\widehat{\boldsymbol{\beta}}_I\|_1. \tag{8.8}$$

Let $\mathbf{P}_I := \mathbf{X}_I(\mathbf{X}_I'\mathbf{X}_I)^{-1}\mathbf{X}_I'$ denote the projection onto the columns of $\mathbf{X}_I$. We have $\|\mathbf{P}_I\mathbf{y}\|_2 \leq \|\mathbf{y}\|_2$, leading to:

$$\|\mathbf{X}_I\widehat{\boldsymbol{\beta}}_I\|_1 = \|\mathbf{P}_I\mathbf{y}\|_1 \leq \sqrt{k}\|\mathbf{P}_I\mathbf{y}\|_2 \leq \sqrt{k}\|\mathbf{y}\|_2, \tag{8.9}$$

where we used that for any $\mathbf{a} \in \mathbb{R}^m$, we have $\sqrt{m}\|\mathbf{a}\|_2 \geq \|\mathbf{a}\|_1$. Combining (8.8) and (8.9) we obtain (2.12).

(d) For any vector $\boldsymbol{\beta}_I$ which has zero entries in the coordinates outside I, we have:

$$\|\mathbf{X}\boldsymbol{\beta}_I\|_\infty \leq \max_{i=1,\ldots,n} |\langle \mathbf{x}_i, \boldsymbol{\beta}_I\rangle| \leq \max_{i=1,\ldots,n} \|\mathbf{x}_i\|_{1:k}\|\boldsymbol{\beta}_I\|_\infty,$$

leading to (2.13). □

9. Proofs and Technical Details for Section 3 in main paper.

9.1. *Proof of Proposition 6 in main paper.*

Proof

(a) Let $\boldsymbol{\beta}$ be a vector satisfying $\|\boldsymbol{\beta}\|_0 \leq k$. Using the notation $\widehat{\boldsymbol{\eta}} \in \mathbf{H}_k\left(\boldsymbol{\beta} - \frac{1}{L}\nabla g(\boldsymbol{\beta})\right)$ we have the following chain of inequalities:

$$\begin{aligned}
g(\boldsymbol{\beta}) &= Q_L(\boldsymbol{\beta}, \boldsymbol{\beta}) \\
&\geq \inf_{\|\boldsymbol{\eta}\|_0 \leq k} \quad Q_L(\boldsymbol{\eta}, \boldsymbol{\beta}) \\
&= \inf_{\|\boldsymbol{\eta}\|_0 \leq k} \quad \left(\frac{L}{2}\|\boldsymbol{\eta} - \boldsymbol{\beta}\|_2^2 + \langle \nabla g(\boldsymbol{\beta}), \boldsymbol{\eta} - \boldsymbol{\beta}\rangle + g(\boldsymbol{\beta})\right) \\
&= \inf_{\|\boldsymbol{\eta}\|_0 \leq k} \quad \left(\frac{L}{2}\left\|\boldsymbol{\eta} - \left(\boldsymbol{\beta} - \frac{1}{L}\nabla g(\boldsymbol{\beta})\right)\right\|_2^2 - \frac{1}{2L}\|\nabla g(\boldsymbol{\beta})\|_2^2 + g(\boldsymbol{\beta})\right) \\
&= \quad \left(\frac{L}{2}\|\widehat{\boldsymbol{\eta}} - \left(\boldsymbol{\beta} - \frac{1}{L}\nabla g(\boldsymbol{\beta})\right)\|_2^2 - \frac{1}{2L}\|\nabla g(\boldsymbol{\beta})\|_2^2 + g(\boldsymbol{\beta})\right) \qquad \text{(From (3.7))} \\
&= \left(\frac{L}{2}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\beta}\|_2^2 + \langle \nabla g(\boldsymbol{\beta}), \widehat{\boldsymbol{\eta}} - \boldsymbol{\beta}\rangle + g(\boldsymbol{\beta})\right) \\
&= \left(\frac{L-\ell}{2}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\beta}\|_2^2 + \frac{\ell}{2}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\beta}\|_2^2 + \langle \nabla g(\boldsymbol{\beta}), \widehat{\boldsymbol{\eta}} - \boldsymbol{\beta}\rangle + g(\boldsymbol{\beta})\right) \\
&= \frac{L-\ell}{2}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\beta}\|_2^2 + \underbrace{\left(\frac{\ell}{2}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\beta}\|_2^2 + \langle \nabla g(\boldsymbol{\beta}), \widehat{\boldsymbol{\eta}} - \boldsymbol{\beta}\rangle + g(\boldsymbol{\beta})\right)}_{Q_\ell(\widehat{\boldsymbol{\eta}}, \boldsymbol{\beta})} \\
&\geq \frac{L-\ell}{2}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\beta}\|_2^2 + g(\widehat{\boldsymbol{\eta}}). \qquad \text{(From (3.6))}
\end{aligned}$$

This chain of inequalities leads to:

$$g(\boldsymbol{\beta}) - g(\widehat{\boldsymbol{\eta}}) \geq \frac{L-\ell}{2}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\beta}\|_2^2. \tag{9.1}$$

Applying (9.1) for $\boldsymbol{\beta} = \boldsymbol{\beta}_m$ and $\widehat{\boldsymbol{\eta}} = \boldsymbol{\beta}_{m+1}$, the vectors generated by Algorithm 1, we obtain (3.11). This implies that the objective values $g(\boldsymbol{\beta}_m)$ are decreasing and since the sequence is bounded below ($g(\boldsymbol{\beta}) \geq 0$), we obtain that $g(\boldsymbol{\beta}_m)$ converges as $m \to \infty$.

(b) If $L > \ell$ and from part (a), the result follows.

(c) The condition $\underline{\alpha}_k > 0$ means that for all m sufficiently large, the entry $|\beta_{(k),m}|$ will remain (uniformly) bounded away from zero. We will use this to prove that the support of $\boldsymbol{\beta}_m$ converges. For the purpose of establishing contradiction suppose that the support does not converge. Then, there are infinitely many values of m' such that $\mathbf{1}_{m'} \neq \mathbf{1}_{m'+1}$. Using the fact that $\|\boldsymbol{\beta}_m\|_0 = k$ for all large m we have

$$\|\boldsymbol{\beta}_{m'} - \boldsymbol{\beta}_{m'+1}\|_2 \geq \sqrt{\beta^2_{m',i} + \beta^2_{m'+1,j}} \geq \frac{|\beta_{m',i}| + |\beta_{m'+1,j}|}{\sqrt{2}}, \tag{9.2}$$

where i, j are such that $\beta_{m'+1,i} = \beta_{m',j} = 0$. As $m' \to \infty$, the quantity in the rhs of (9.2) remains bounded away from zero since $\underline{\alpha}_k > 0$. This contradicts the fact that $\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m \to \mathbf{0}$, as established in part (b). Thus, $\mathbf{1}_m$ converges, and since $\mathbf{1}_m$ is a discrete sequence, it converges after finitely many iterations, that is $\mathbf{1}_m = \mathbf{1}_{m+1}$ for all $m \geq M^*$. Algorithm 1 becomes a vanilla gradient descent algorithm, restricted to the space $\mathbf{1}_m$ for $m \geq M^*$. Since a gradient descent algorithm for minimizing a convex function over a closed convex set leads to a sequence of iterates that converge (Rockafellar, 1996; Nesterov, 2004), we conclude that Algorithm 1 converges. Therefore, the sequence $\boldsymbol{\beta}_m$ converges to $\boldsymbol{\beta}^*$, a first order stationarity point:

$$\boldsymbol{\beta}^* \in \mathbf{H}_k\left(\boldsymbol{\beta}^* - \frac{1}{L}\nabla g(\boldsymbol{\beta}^*)\right).$$

Boundedness follows from the convergence of $\boldsymbol{\beta}_m$.

(d) Let $\mathcal{I}_m \subset \{1, \ldots, p\}$ denote the set of k largest values of the vector $\left(\boldsymbol{\beta}_m - \frac{1}{L}\nabla g(\boldsymbol{\beta}_m)\right)$ in absolute value. By the definition of $\mathbf{H}_k\left(\boldsymbol{\beta}_m - \frac{1}{L}\nabla g(\boldsymbol{\beta}_m)\right)$, we have

$$\left|\left(\boldsymbol{\beta}_m - \frac{1}{L}\nabla g(\boldsymbol{\beta}_m)\right)_i\right| \geq \left|\left(\boldsymbol{\beta}_m - \frac{1}{L}\nabla g(\boldsymbol{\beta}_m)\right)_j\right|,$$

for all i, j with $i \in \mathcal{I}_m$ and $j \notin \mathcal{I}_m$. Thus,

$$\liminf_{m\to\infty} \min_{i\in\mathcal{I}_m} \left|\left(\boldsymbol{\beta}_m - \frac{1}{L}\nabla g(\boldsymbol{\beta}_m)\right)_i\right| \geq \liminf_{m\to\infty} \max_{j\notin\mathcal{I}_m} \left|\left(\boldsymbol{\beta}_m - \frac{1}{L}\nabla g(\boldsymbol{\beta}_m)\right)_j\right|. \tag{9.3}$$

Moreover,

$$\left(\boldsymbol{\beta}_m - \mathbf{H}_k\left(\boldsymbol{\beta}_m - \frac{1}{L}\nabla g(\boldsymbol{\beta}_m)\right)\right)_i = \begin{cases} \dfrac{1}{L}(\nabla g(\boldsymbol{\beta}_m))_i, & i \in \mathcal{I}_m, \\ \beta_{m,i}, & \text{otherwise.} \end{cases}$$

Using the fact that $\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m \to \mathbf{0}$ we have

$$(\nabla g(\boldsymbol{\beta}_m))_i \to 0, i \in \mathcal{I}_m \;\text{ and }\; \beta_{m,j} \to 0, j \notin \mathcal{I}_m$$

as $m \to \infty$. Combining with (9.3) we have that:

$$\liminf_{m\to\infty} \;\min_{i\in\mathcal{I}_m} \;|\boldsymbol{\beta}_{mi}| \geq \liminf_{m\to\infty} \;\max_{j\notin\mathcal{I}_m} \;\frac{1}{L}\left|(\nabla g(\boldsymbol{\beta}_m))_j\right| = \frac{1}{L}\liminf_{m\to\infty} \;\|\nabla g(\boldsymbol{\beta}_m)\|_\infty.$$

Since, $\liminf_{m\to\infty} \;\min_{i\in\mathcal{I}_m} \;|\boldsymbol{\beta}_{mi}| = \underline{\alpha}_k = 0$ (by hypothesis), the lhs of the above inequality equals zero, which leads to $\liminf_{m\to\infty} \;\|\nabla g(\boldsymbol{\beta}_m)\|_\infty = 0$.

(e) We build on the proof of Part (d).

It follows from equation (9.3) (by suitably modifying '$\liminf$' to '$\limsup$') that:

$$\underbrace{\limsup_{m\to\infty} \;\min_{i\in\mathcal{I}_m} \;|\boldsymbol{\beta}_{mi}|}_{\overline{\alpha}_k} \geq \limsup_{m\to\infty} \;\max_{j\notin\mathcal{I}_m} \;\frac{1}{L}\left|(\nabla g(\boldsymbol{\beta}_m))_j\right| = \frac{1}{L}\limsup_{m\to\infty} \;\|\nabla g(\boldsymbol{\beta}_m)\|_\infty.$$

Note that the lhs of the above inequality is $\overline{\alpha}_k$ which is zero (by hypothesis), thus $\|\nabla g(\boldsymbol{\beta}_m)\|_\infty \to 0$ as $m \to \infty$.

Suppose $\boldsymbol{\beta}_\infty$ is a limit point of the sequence $\boldsymbol{\beta}_m$. Thus there is a subsequence $m' \subset \{1, 2, \ldots, \}$ such that $\boldsymbol{\beta}_{m'} \to \boldsymbol{\beta}_\infty$ and $g(\boldsymbol{\beta}_{m'}) \to g(\boldsymbol{\beta}_\infty)$. Using the continuity of the gradient and hence the function $\cdot \mapsto \|\nabla g(\cdot)\|_\infty$ we have that $\|\nabla g(\boldsymbol{\beta}_{m'})\|_\infty \to \|\nabla g(\boldsymbol{\beta}_\infty)\|_\infty = 0$ as $m' \to \infty$. Thus $\boldsymbol{\beta}_\infty$ is a solution to the unconstrained (without cardinality constraints) optimization problem $\min \; g(\boldsymbol{\beta})$. Since $g(\boldsymbol{\beta}_m)$ is a decreasing sequence, $g(\boldsymbol{\beta}_m)$ converges to the minimum of $g(\boldsymbol{\beta})$. □

9.2. *Proof of Proposition 3 in main paper.*
Proof:
We provide a proof of Proposition 3, for the sake of completeness.

It suffices to consider $|c_i| > 0$ for all i. Let $\boldsymbol{\beta}$ be an optimal solution to Problem (3.3) and let $S := \{i : \beta_i \neq 0\}$. The objective function is given by $\sum_{i\notin S} |c_i|^2 + \sum_{i\in S}(\beta_i - c_i)^2$. Note that by selecting $\beta_i = c_i$ for $i \in S$, we can make the objective function $\sum_{i\notin S} |c_i|^2$. Thus, to minimize the objective function, S must correspond to the indices of the largest k values of $|c_i|, i \geq 1$. □

9.3. *Proof of Proposition 5 in main paper.*
Proof
If $\boldsymbol{\eta}$ is a first order stationary point with $\|\boldsymbol{\eta}\|_0 \leq k$, it follows from the argument following Definition 1, that there is a set $I \subset \{1, \ldots, p\}$ with $|I^c| = k$ such that $\nabla_i g(\boldsymbol{\eta}) = 0$ for all $i \notin I$ and $\eta_i = 0$ for all $i \notin I$. Let $\mu_i := \eta_i - \frac{1}{L}\nabla_i g(\boldsymbol{\eta})$ for $i = 1, \ldots, p$. Suppose I_k denotes the set of indices corresponding to the top k ordered values of $|\mu_i|$. Note that:

$$\mu_i = \eta_i, \;\; i \in I_k \quad \text{and} \quad |\mu_j| = |\frac{1}{L}\nabla_j g(\boldsymbol{\eta})|, \;\; j \notin I_k. \tag{9.4}$$

For $i \in I_k$ and $j \notin I_k$ we have $|\mu_i| \geq |\mu_j|$. This implies that $|\eta_i| \geq |\frac{1}{L}\nabla_j g(\boldsymbol{\eta})|$. Since $\boldsymbol{\eta} \in \mathbf{H}_k\left(\boldsymbol{\eta} - \frac{1}{L}\nabla g(\boldsymbol{\eta})\right)$ and $\|\boldsymbol{\eta}\|_0 < k$, it follows that $0 = \min_{i \in I_k} |\eta_i| = \min_{i \in I_k} |\mu_i|$. We thus have that $\nabla_j g(\boldsymbol{\eta}) = 0$ for all $j \notin I_k$. In addition, note that $\nabla_i g(\boldsymbol{\eta}) = 0$ for all $i \in I_k$. Thus it follows that $\nabla g(\boldsymbol{\eta}) = \mathbf{0}$ and hence $\boldsymbol{\eta} \in \arg\min_{\boldsymbol{\eta}} g(\boldsymbol{\eta})$. □

9.4. *Proof of Proposition 7 in main paper.*
Proof
Part (a): This follows from Proposition 6, Part (a), which implies that:

$$g(\boldsymbol{\eta}) - g(\widehat{\boldsymbol{\eta}}) \geq \frac{L-\ell}{2}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2^2,$$

for any $\widehat{\boldsymbol{\eta}} \in \mathbf{H}_k\left(\boldsymbol{\eta} - \frac{1}{L}\nabla g(\boldsymbol{\eta})\right)$. Now by the definition of $\mathbf{H}_k(\cdot)$, we have $g(\boldsymbol{\eta}) = g(\widehat{\boldsymbol{\eta}})$ which along with $L > \ell$ implies that the rhs of the above inequality is zero: thus $\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 = 0$, i.e., $\boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}$. Since the choice of $\widehat{\boldsymbol{\eta}}$ was arbitrary, it follows that $\boldsymbol{\eta}$ is the only element in the set $\mathbf{H}_k\left(\boldsymbol{\eta} - \frac{1}{L}\nabla g(\boldsymbol{\eta})\right)$. □

Part (b): The proof follows by noting that $\widehat{\boldsymbol{\beta}}$ is k-sparse along with Proposition 6, Part (a), which implies that:

$$g(\widehat{\boldsymbol{\beta}}) - g(\widehat{\boldsymbol{\eta}}) \geq \frac{L-\ell}{2}\left\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\eta}}\right\|_2^2,$$

for any $\widehat{\boldsymbol{\eta}} \in \mathbf{H}_k\left(\widehat{\boldsymbol{\beta}} - \frac{1}{L}\nabla g(\widehat{\boldsymbol{\beta}})\right)$. Now, by the definition of $\widehat{\boldsymbol{\beta}}$ we have $g(\widehat{\boldsymbol{\beta}}) = g(\widehat{\boldsymbol{\eta}})$ which along with $L > \ell$ implies that the rhs of the above inequality is zero: thus $\widehat{\boldsymbol{\beta}}$ is a first order stationary point. □

9.5. *Proof of Theorem 3.1 in main paper.*
Proof
Summing inequalities (3.11) for $1 \leq m \leq M$. we obtain

$$\sum_{m=1}^{M}\left(g(\boldsymbol{\beta}_m) - g(\boldsymbol{\beta}_{m+1})\right) \geq \frac{L-\ell}{2}\sum_{m=1}^{M}\|\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m\|_2^2, \tag{9.5}$$

leading to

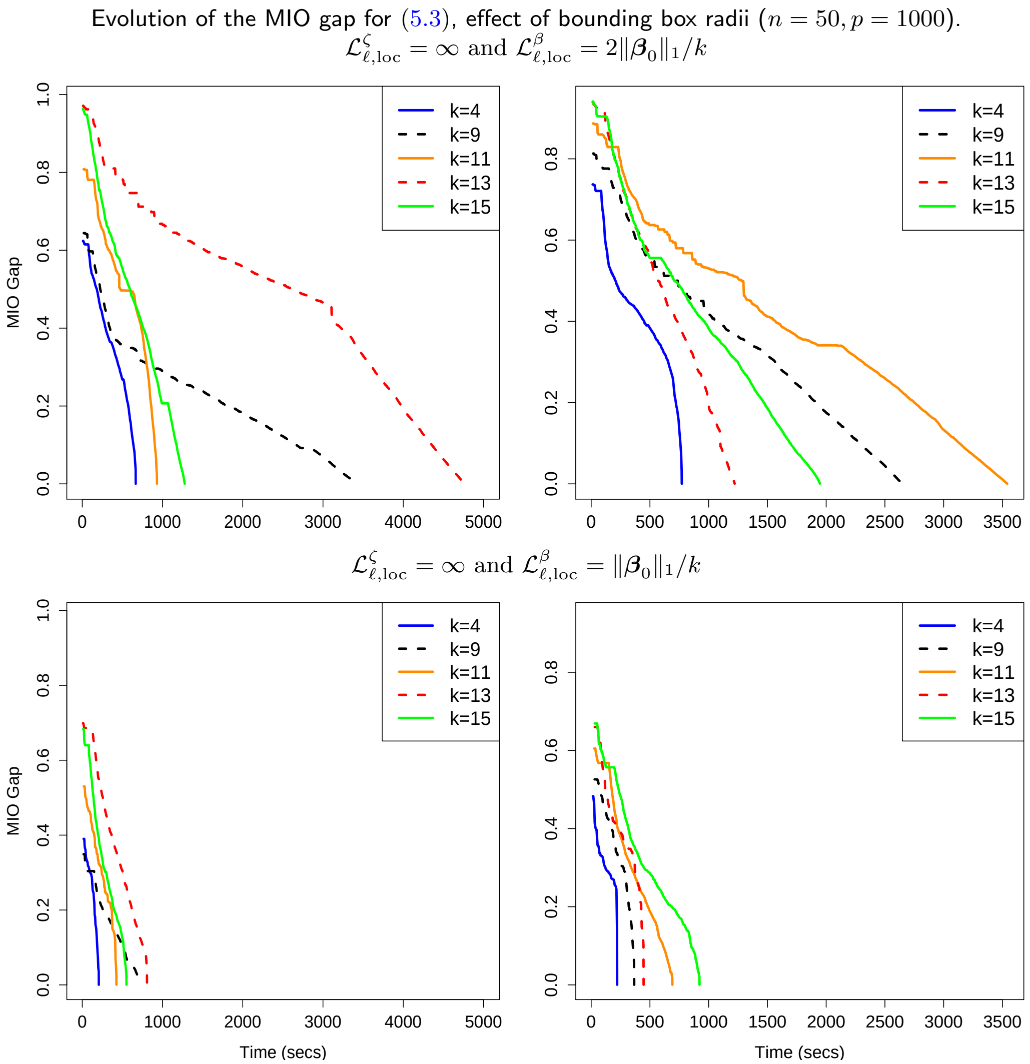
$$g(\boldsymbol{\beta}_1) - g(\boldsymbol{\beta}_{M+1}) \geq \frac{M(L-\ell)}{2}\min_{m=1,\ldots,M}\|\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m\|_2^2.$$

Since the decreasing sequence $g(\boldsymbol{\beta}_{m+1})$ converges to $g(\boldsymbol{\beta}^*)$ by Proposition 6 we have:

$$\frac{g(\boldsymbol{\beta}_1) - g(\boldsymbol{\beta}^*)}{M} \geq \frac{g(\boldsymbol{\beta}_1) - g(\boldsymbol{\beta}_{M+1})}{M} \geq \frac{(L-\ell)}{2}\min_{m=1,\ldots,M}\|\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m\|_2^2. \qquad \square$$

10. Additional Details on Experiments and Computations.

10.1. *Some additional figures related to the radii of bounding boxes.* Some figures illustrating the effect of the bounding box radii are presented in Figure 11.


Evolution of the MIO gap for (5.3), effect of bounding box radii ($n = 50, p = 1000$).

$\mathcal{L}^{\zeta}_{\ell,\mathrm{loc}} = \infty$ and $\mathcal{L}^{\beta}_{\ell,\mathrm{loc}} = 2\|\boldsymbol{\beta}_0\|_1/k$

$\mathcal{L}^{\zeta}_{\ell,\mathrm{loc}} = \infty$ and $\mathcal{L}^{\beta}_{\ell,\mathrm{loc}} = \|\boldsymbol{\beta}_0\|_1/k$

Fig 11: The evolution of the MIO gap with varying radii of bounding boxes for MIO formulation (5.3). The top panel has radii twice the size of the bottom panel. The dataset considered is generated as per Example 1 with $n = 50, p = 1000, \rho = 0.9$ and $k_0 = 5$ for different values of SNR: [Left Panel] SNR = 1, [Right Panel] SNR = 3. For each case, different values of k have been considered. The top panel has a bounding box radii which is twice the corresponding case in the lower panel. As expected, the times for the MIO gaps to close depends upon the radii of the boxes. The optimal solutions obtained were found to be insensitive to the choice of the bounding box radius.

10.2. *`Lasso`, unshrunk `Lasso` and MIO.* We present here comparisons of the *unshrunk* `Lasso` with MIO and `Lasso`.

Debiasing is often used to mitigate the shrinkage imparted by the `Lasso` regularization parameter. This is done by performing an unrestricted least squares on the support selected by the `Lasso`. Of course the results will depend upon the tuning parameter used for the problem. We use two methods towards this end. In the first method we find the best `Lasso` solution (by obtaining an optimal tuning parameter based on minimizing predictive error on a held out validation set); we then obtain the un-regularized least squares solution for that `Lasso` solution. This typically performed worse than `Lasso` in all the experiments we tried—see Tables 3 and 4. The unrestricted least squares solution on the optimal model selected by the `Lasso` (as shown in Figure 4) had worse predictive performance than the `Lasso`, with the same sparsity pattern, as shown in Table 3. This is probably due to overfitting since the model selected by the `Lasso` is quite dense compared to n, p. Table 4 presents the results for $50 = n \ll p = 1000$. We consider the same example presented in Figure 9, Example 1. First of all, Table 4 presents the prediction performance of `Lasso` after debiasing—we considered the same tuning parameter considered optimal for the `Lasso` problem. We see that as in the case of Table 3, the debiasing does not lead to improved performance in terms of prediction error.

We thus experimented with another variant of the *unshrunk* `Lasso`, where for every λ we computed the `Lasso` solution (1.2) and obtained $\widehat{\boldsymbol{\beta}}_{\mathrm{Deb},\lambda}$ by performing an unrestricted least squares fit on the support selected by the `Lasso` solution at λ. This method can be thought of delivering feasible solutions for Problem (1.1), for a value of $k := k(\lambda)$ determined by the `Lasso` solution at λ. The success of this method makes a case in support of using criterion (1.1). The tuning parameter was then selected by minimizing predictive performance on a held out test validation set. This method in general performed better than `Lasso` in delivering a sparser model with better predictive accuracy than the `Lasso`. The performance of the *unshrunk* `Lasso` was similar to `Sparsenet` and was in general inferior to MIO by orders of magnitude, especially for the problems where the pairwise correlations between the variables was large and SNR was low and $n \ll p$. The results are presented in Table 5,6 (for the case $n > p$) and 7 and 8 (for the case $n \ll p$).

Debiasing at optimal `Lasso` model, $n > p$

SNR	ρ	Ratio: `Lasso`/ *unshrunk* `Lasso`
6.33	0.5	0.33
3.17	0.5	0.54
1.58	0.5	0.53
6.97	0.8	0.67
3.48	0.8	0.64
1.74	0.8	0.63
8.73	0.9	1
4.37	0.9	0.58
2.18	0.9	0.61

TABLE 3

`Lasso` and unshrunk *`Lasso` corresponding to the numerical experiments of Figure 4, for Example 1 with $n = 500, p = 100, \rho \in \{0.5, 0.8, 0.9\}$ and $k_0 = 10$. Here, "Ratio" equals the ratio of the prediction error of the `Lasso` and the* unshrunk *`Lasso` at the optimal tuning parameter selected by the `Lasso`.*

The performance of this model was comparable with `Sparsenet`—it was better than `Lasso`

Debiasing at optimal Lasso model, $n \ll p$

SNR	ρ	Ratio: Lasso/ *unshrunk* Lasso
10	0.8	0.90
7	0.8	1.0
3	0.8	0.91

Table 4

Lasso and unshrunk *Lasso corresponding to the numerical experiments of Figure 9, for Example 1 with $n = 50, p = 1000, \rho = 0.8$ and $k_0 = 5$. Here, "Ratio" equals the ratio of the prediction error of the Lasso and the* unshrunk *Lasso at the optimal tuning parameter selected by the Lasso.*

in terms of obtaining a sparser model with better predictive accuracy. However, the performance of MIO was significantly better than the *unshrunk* version of the Lasso, especially for larger values of ρ and smaller SNR values.

Sparsity of Selected Models, $n > p$

SNR	ρ	Lasso	*unshrunk* Lasso	MIO
6.33	0.5	27.6 (2.122)	10.9 (0.65)	10.8 (0.51)
3.17	0.5	27.7 (2.045)	10.9 (0.65)	10.1 (0.1)
1.58	0.5	28.0 (2.276)	10.9 (0.65)	10.2 (0.2)
6.97	0.8	34.1 (3.60)	10.4 (0.15)	10 (0.0)
3.48	0.8	34.0 (3.54)	10.9 (0.55)	10.2 (0.2)
1.74	0.8	33.7 (3.49)	13.7 (1.50)	10 (0.0)
8.73	0.9	25.9 (0.94)	13.9 (0.68)	10.5 (0.17)
4.37	0.9	34.6 (3.23)	18.1 (1.30)	10.2 (0.25)
2.18	0.9	34.7 (3.28)	20.5 (1.85)	10.1 (0.10)

Table 5

Number of non-zeros in the selected model by Lasso, unshrunk *Lasso , and MIO corresponding to the numerical experiments of Figure 4, for Example 1 with $n = 500, p = 100, \rho \in \{0.5, 0.8, 0.9\}$ and $k_0 = 10$. The tuning parameters for all three models were selected separately based on the best predictive model on a held out validation set. Numbers within brackets denote standard-errors.* unshrunk *Lasso leads to less dense models than Lasso. When ρ is small and SNR is large, the model size of* unshrunk *Lasso is similar to MIO. However, for larger values of ρ and smaller values of SNR subset selection leads to orders of magnitude sparser solutions than* unshrunk *Lasso.*

We then follow the method described above (for the $n > p$ case), where we consider a sequence of models $\widehat{\boldsymbol{\beta}}_{\text{Deb},\lambda}$ and find the λ that delivers the best predictive model on a held out validation set.

Predictive Performance of Selected Models, $n > p$

SNR	ρ	`Lasso`	*unshrunk* `Lasso`	MIO	Ratio: *unshrunk* `Lasso` /MIO
6.33	0.5	0.0384 (0.001)	0.0255 (0.002)	0.0266 (0.001)	1.0
3.17	0.5	0.0768 (0.003)	0.0511 (0.004)	0.0478 (0.002)	1.0
1.58	0.5	0.1540 (0.007)	0.1021 (0.009)	0.0901 (0.009)	1.1
6.97	0.8	0.0389 (0.002)	0.0223 (0.001)	0.0231 (0.002)	1.0
3.48	0.8	0.0778 (0.004)	0.0464 (0.003)	0.0484 (0.004)	1.0
1.74	0.8	0.1557 (0.007)	0.1156 (0.008)	0.0795 (0.008)	1.5
8.73	0.9	0.0325 (0.001)	0.0220 (0.002)	0.0197 (0.002)	1.2
4.37	0.9	0.0632 (0.002)	0.0532 (0.003)	0.0427 (0.008)	1.3
2.18	0.9	0.1265 (0.005)	0.1254 (0.006)	0.0703 (0.011)	1.8

TABLE 6

Predictive Performance for tests of `Lasso`, unshrunk *`Lasso` , and MIO corresponding to the numerical experiments of Figure 4, for Example 1 with $n = 500, p = 100, \rho \in \{0.5, 0.8, 0.9\}$ and $k_0 = 10$. Numbers within brackets denote standard-errors. The tuning parameters for all three models were selected separately based on the best predictive model on a held out validation set. When ρ is small and SNR is large,* unshrunk *`Lasso` performance is similar to MIO. However, for larger values of ρ and smaller values of SNR subset selection performs better than* unshrunk *`Lasso` based solutions.*

Sparsity of Selected Models, $n \ll p$

SNR	ρ	`Lasso`	*unshrunk* `Lasso`	MIO
10	0.8	25.7 (1.73)	7.9 (0.43)	5 (0.12)
7	0.8	27.8 (2.69)	8.1 (0.43)	5 (0.16)
3	0.8	28.0 (2.72)	10.0 (0.88)	6 (1.18)

TABLE 7

Number of non-zeros in the selected model by `Lasso`, unshrunk *`Lasso` , and MIO corresponding to the numerical experiments of Figure 9, for Example 1with $n = 50, p = 1000, \rho = 0.8$ and $k_0 = 5$. Numbers within brackets denote standard-errors. The tuning parameters for all three models were selected separately based on the best predictive model on a held out validation set.* unshrunk *`Lasso` leads to less dense models than `Lasso` but more dense models than MIO. The performance gap between MIO and* unshrunk *`Lasso` becomes larger with lower values of SNR.*

Predictive Performance of Selected Models, $n \ll p$

SNR	ρ	Lasso	*unshrunk* Lasso	MIO	Ratio: *unshrunk* Lasso / MIO
10	0.8	0.084 (0.004)	0.046 (0.003)	0.014 (0.005)	3.3
7	0.8	0.122 (0.005)	0.070 (0.004)	0.020 (0.007)	3.5
3	0.8	0.257 (0.012)	0.185 (0.016)	0.151 (0.027)	1.2

Table 8

Predictive performances of Lasso, unshrunk *Lasso , and MIO corresponding to the numerical experiments of Figure 9, for Example 1with* $n = 50, p = 1000, \rho = 0.8$ *and* $k_0 = 5$*. Numbers within brackets denote standard-errors. The tuning parameters for all three models were selected separately based on the best predictive model on a held out validation set. MIO consistently leads to better predictive models than* unshrunk *Lasso and ordinary Lasso.* unshrunk *Lasso performs better than ordinary Lasso.*

Dimitris Bertsimas,
MIT Sloan School of Management and
Operations Research Center,
Massachusetts Institute of Technology,
Cambridge, MA.
E-mail: dbertsim@mit.edu

Angela King,
Operations Research Center,
Massachusetts Institute of Technology,
Cambridge, MA.
E-mail: aking10@mit.edu

Rahul Mazumder,
MIT Sloan School of Management and
Operations Research Center,
Massachusetts Institute of Technology,
Cambridge, MA.
E-mail: rahulmaz@mit.edu