SAFE SCREENING RULES FOR $\ell_0$-REGRESSION

ALPER ATAMTÜRK AND ANDRÉS GÓMEZ

ABSTRACT. We give safe screening rules to eliminate variables from regression with $\ell_0$ regularization or cardinality constraint. These rules are based on guarantees that a feature may or may not be selected in an optimal solution. The screening rules can be computed from a convex relaxation solution in linear time, without solving the $\ell_0$ optimization problem. Thus, they can be used in a preprocessing step to safely remove variables from consideration apriori. Numerical experiments on real and synthetic data indicate that, on average, 76% of the variables can be fixed to their optimal values, hence, reducing the computational burden for optimization substantially. Therefore, the proposed fast and effective screening rules extend the scope of algorithms for $\ell_0$-regression to larger data sets.

February 6, 2020

A. Atamtürk: Department of Industrial Engineering & Operations Research, University of California, Berkeley, CA 94720. atamturk@berkeley.edu
A. Gómez: Daniel J. Epstein Department of Industrial & Systems Engineering, University of Southern California, CA 90089. gomezand@usc.edu
1. Introduction

In machine learning and optimization communities, there is an increasing interest in regression models with $\ell_0$ and $\ell_2$ regularization:

\[
\text{(REG)} \quad \min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2 + \frac{1}{\gamma} \|x\|_2^2 + \mu \|x\|_0, \quad \text{and}
\]

\[
\text{(CARD)} \quad \min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2 + \frac{1}{\gamma} \|x\|_2^2 \quad \text{s.t.} \quad \|x\|_0 \leq k,
\]

where $A \in \mathbb{R}^{m \times n}$ is the model matrix, $y \in \mathbb{R}^m$ is the vector of response variables, and $x \in \mathbb{R}^n$ is the vector of decision variables, i.e., regression coefficients to be estimated. Problem (CARD) has an explicit cardinality constraint on the number of non-zeros of $x$, whereas (REG) is the regularized version of it. In these models, the $\ell_0$ terms impose sparsity (Miller 2002), which is a necessity for large-dimensional model inference (Hastie et al. 2001, 2015), and the $\ell_2$ (ridge) regularization (Hoerl & Kennard 1970) imposes bias/shrinkage in the regression coefficients. The $\ell_2$ regularization can be interpreted, from the robust optimization perspective, as a correction term to account for uncertainty in the model matrix $A$ (El Ghaoui & Lebret 1997, Xu et al. 2009), and has been shown to improve the performance of sparse regression models in high-noise regimes (Mazumder et al. 2017).

The popular $\ell_1$ (lasso, Tibshirani 1996) and $\ell_1-\ell_2$ (elastic net, Zou & Hastie 2005) regularizations perform shrinkage and model selection simultaneously and, as convex proxies for (REG), they are very fast. However, thanks to substantial progress in the field of mixed-integer optimization (MIO), there is an increasing interest in solving the non-convex problems (REG)–(CARD) directly. Indeed, several studies (Bertsimas et al. 2016, Cozad et al. 2014, Gómez & Prokopyev 2018, Miyashiro & Takano 2015, Park & Klabjan 2017) have shown that problems (REG)–(CARD) with hundreds of variables can be solved to optimality simply by employing general purpose MIO solvers, and the resulting estimators outperform their $\ell_1$ counterparts. Nonetheless, solving the $\ell_0$ problems in this manner is orders-of-magnitude slower than solving the $\ell_1$ approximations and does not scale to problems with $n \geq 1,000$. Therefore, fast heuristics such as $\ell_1$ approximations, thresholding, local (but combinatorial) search algorithms or greedy methods (Hastie et al. 2017, Hazimeh & Mazumder 2018, Xie & Deng 2020) may still be preferable in large-scale instances.

The gap in the performance between exact methods for (REG)–(CARD) and algorithms for a convex approximation is to be expected, as the $\ell_0$-regression is NP-hard. Moreover, there exist specialized software packages tailored to solving lasso and elastic net problems, such as glmnet (Friedman et al. 2010), which include a variety of techniques specific to $\ell_1$ inference problems. In contrast, general purpose MIO solvers are not tailored to tackle (REG)–(CARD). Researchers have recently experimented with implementing branch-and-bound methods tailored for (REG)–(CARD) (Bertsimas &
Van Parys [2017], Bertsimas et al. [2019], Dedieu et al. [2020], Kimura & Waki [2018], and the promising results indicate that there is substantial room for improvement for exact $\ell_0$-regression algorithms.

The purpose of this paper is to define screening rules for nonconvex $\ell_0$-regression problems (REG)–(CARD). El Ghaoui et al. [2010] propose safe rules for efficiently identifying regression variables that are guaranteed to be zero (null) in an optimal solution of the lasso problem, reducing the dimension of the problem to be solved a priori. Tibshirani et al. [2012] subsequently propose strong rules that may discard predictors that are part of an optimal lasso solution, but are quite effective in practice; these strong rules are incorporated into glmnet. Additional screening procedures have been proposed for other convex and lasso-type inference problems (Fercoq et al. 2015, Ogawa et al. 2013, Xiang & Ramadge 2012, Wang et al. 2013, Xiang et al. 2016). To the best of our knowledge, no such screening rule is given to-date for the nonconvex $\ell_0$-regression problems (REG)–(CARD).

In MIO community, screening rules are used as part of preprocessing in branch-and-bound solvers (Atamtürk et al. 2000, Savelsbergh 1994). In contrast to convex optimization, for MIO problems such as (REG)–(CARD), fixing a single binary variable to zero reduces the number of feasible solutions by half; thus, the expected speedup of enumerative methods such as the branch-and-bound method is exponential in the number of variables fixed. Therefore, effect of the screening rules on enumerative methods for non-convex optimization problems is significantly more than on polynomial-time algorithms for convex optimization problems. Unfortunately, the existing screening rules in MIO solvers are tailored for linear mixed-integer problems and, as such, they are ineffective for (REG)–(CARD).

Contributions and outline. In this paper we propose safe screening rules for nonconvex $\ell_0$-regression problems (REG)–(CARD). These rules can be applied to reduce the size of the problems, independent of the method used to solve them. Similar to the approach proposed by El Ghaoui et al. [2010] for lasso, the safe rules proposed are particularly effective in problems with large $\ell_0$–$\ell_2$ regularization terms, thus suitable for high noise regimes. The screening rules are obtained by exploiting convex perspective relaxations of the $\ell_0$ regression problems and using their Fenchel dual. The rules can be computed from a convex relaxation solution in linear time, without having to solve the $\ell_0$ optimization problem. In our computational experiments with benchmark instances, the screening rules have been able to fix, on average, 76% of the variables to their optimal values, and in some cases they have been sufficient to provably solve the problems outright. When used as preprocessing with a general purpose branch-and-bound solver, the screening procedure results in orders-of-magnitude speedups: instances previously requiring hours (or more) to prove optimality are solved in under 10 seconds with screening. Consequently, the speed and effectiveness of the safe screening rules extend the scope of algorithms for $\ell_0$-regression problems to larger data sets.
The rest of the paper is organized as follows. In Section 2 we describe mixed-integer formulations and convex perspective relaxations of problems \((\text{REG})\) and \((\text{CARD})\). In Section 3 we derive the safe screening rules for \((\text{REG})\) and \((\text{CARD})\) based on Fenchel duality of the perspective relaxations. In Section 4 we present our computational experiments with synthetic and real benchmark instances from the literature. We conclude in Section 5 with a few final remarks.

2. Mixed-integer & perspective formulations

Introducing indicator variables \(z \in \{0, 1\}^n\) such that \(z_i = 0 \implies x_i = 0\), problem \((\text{REG})\) can be naturally formulated as the quadratic mixed-integer optimization problem

\[
(1a) \quad \min_{x, z} \|y - Ax\|_2^2 + \frac{1}{\gamma} \sum_{i=1}^{n} x_i^2 + \mu \sum_{i=1}^{n} z_i
\]

\[
(1b) \quad \text{s.t. } x_i(1 - z_i) = 0, \quad i = 1, \ldots, n
\]

\[
(1c) \quad x \in \mathbb{R}^n, \ z \in \{0, 1\}^n.
\]

For each \(i\), the complementarity constraint \(x_i(1 - z_i) = 0\), ensures that \(x_i = 0\) whenever \(z_i = 0\). Such complementary constraints can be linearized via “big-M” constraints \(|x_i| \leq Mz_i\) (Bertsimas et al., 2016) for a suitably large value of \(M\). However, such formulations with large values of \(M\) are weak and may lead to poor performance as a consequence. A stronger formulation can be given by utilizing the perspective of the univariate quadratic function \(x_i^2\):

\[
(2a) \quad \zeta_R = \min_{x, z} \|y - Ax\|_2^2 + \frac{1}{\gamma} \sum_{i=1}^{n} \frac{x_i^2}{z_i} + \mu \sum_{i=1}^{n} z_i
\]

\[
(2b) \quad \text{(MIPR) } \quad \text{s.t. } x_i(1 - z_i) = 0, \quad i = 1, \ldots, n
\]

\[
(2c) \quad x \in \mathbb{R}^n, \ z \in \{0, 1\}^n,
\]

where we adopt the convention that \(x_i^2/z_i = 0\) if \(z_i = x_i = 0\), and \(x_i^2/z_i = +\infty\) if \(z_i = 0\) and \(x_i \neq 0\). The perspective function \(x_i^2/z_i\) significantly strengthens the convex relaxation and can be formulated with conic quadratic constraints (Aktürk et al., 2009), (Dong et al., 2015), (Frangioni & Gentile, 2006), (Günlük & Linderoth, 2010), (Xie & Deng, 2020). The perspective formulation is also at the core of recent specialized branch-and-bound methods for sparse regression (Bertsimas & Van Parys, 2017), (Bertsimas et al., 2019). A similar
SAFE SCREENING RULES FOR $\ell_0$-REGRESSION

A strong mixed-integer formulation of (CARD) is

\[
\begin{align*}
\zeta &= \min_{x,z} \|y - Ax\|_2^2 + \frac{1}{\gamma} \sum_{i=1}^{n} \frac{x_i^2}{z_i} \\
\text{s.t.} & \quad \sum_{i=1}^{n} z_i \leq k \\
& \quad x_i(1 - z_i) = 0, \quad i = 1, \ldots, n \\
& \quad x \in \mathbb{R}^n, \ z \in \{0,1\}^n.
\end{align*}
\]

Convex relaxation of the mixed-integer programs are obtained by dropping complementary constraints \((2b)\) and \((3c)\), and relaxing the integrality constraints in \((2c)\) and \((3d)\) to $z \in [0,1]^n$. Thus, we obtain the convex relaxation

\[
\begin{align*}
\zeta_{CR} &= \min_{x,z} \|y - Ax\|_2^2 + \frac{1}{\gamma} \sum_{i=1}^{n} \frac{x_i^2}{z_i} + \mu \sum_{i=1}^{n} z_i \\
\text{s.t.} & \quad \sum_{i=1}^{n} z_i \leq k, \ x \in \mathbb{R}^n, \ z \in [0,1]^n
\end{align*}
\]

of (MIPR), and the convex relaxation

\[
\begin{align*}
\zeta_{CC} &= \min_{x,z} \|y - Ax\|_2^2 + \frac{1}{\gamma} \sum_{i=1}^{n} \frac{x_i^2}{z_i} \\
\text{s.t.} & \quad \sum_{i=1}^{n} z_i \leq k, \ x \in \mathbb{R}^n, \ z \in [0,1]^n
\end{align*}
\]

of (MIPC).

The optimal solutions of (4) and (5) are good statistical estimators on their own right. Indeed, Pilanci et al. (2015) propose convex relaxations of (REG) – (CARD), which are later shown to be equivalent to perspective relaxations (Xie & Deng 2020), and study their strength and conditions for delivering optimal solutions.

3. SAFE SCREENING RULES FOR (REG) & (CARD)

In this section, we give safe screening rules for problems (MIPR) and (MIPC), to fix the binary indicator variables at their optimal values before solving them. The screening rules require an upper bound on the optimal objective value of the mixed-integer optimization problems (MIPR) or (MIPC) and an optimal solution of the perspective relaxation (CR) or (CC), respectively.

**Proposition 1** (Safe screening rules for REG). Let $x^*$ be an optimal solution to (CR) with objective value $\zeta_{CR}$, $\epsilon^* = y - Ax^*$, $\delta_i = (A_i^T \epsilon^*)^2$, $i = 1, \ldots, n$, and let $\zeta$ be an upper bound on $\zeta$. Then any optimal solution
to (MIPR) satisfies,
\[ z_i = \begin{cases} 
0, & \text{if } \zeta_{CR} + \mu - \gamma \delta_i > \bar{\zeta} \\
1, & \text{if } \zeta_{CR} - \mu + \gamma \delta_i > \bar{\zeta}. 
\end{cases} \]

**Proposition 2** (Safe screening rules for (CARD)). Let \( x^* \) be an optimal solution to (CC) with objective value \( \zeta_{CC} \), \( \varepsilon^* = y - Ax^* \), \( \delta_i = (A_i^t \varepsilon^*)^2 \), \( i = 1, \ldots, n \), \( \delta_{[k]} \) be the \( k \)-th largest value of vector \( \delta \), and let \( \bar{\zeta} \) be an upper bound on \( \zeta_C \). Then any optimal solution to (MIPC) satisfies,
\[ z_i = \begin{cases} 
0, & \text{if } \delta_i \leq \delta_{[k+1]} \text{ and } \zeta_{CC} - \gamma (\delta_i - \delta_{[k]}) > \bar{\zeta} \\
1, & \text{if } \delta_i \geq \delta_{[k]} \text{ and } \zeta_{CC} + \gamma (\delta_i - \delta_{[k+1]}) > \bar{\zeta}. 
\end{cases} \]

We prove Propositions 1 and 2 using Fenchel duality in §3.2. Before doing so, in §3.1 we discuss the computational cost of implementing the screening rules.

**3.1. Computational cost.** Computing optimal solutions to the convex perspective relaxations can be done in polynomial time, while finding upper bounds for the non-convex mixed-integer optimization can be accomplished via fast heuristics, thus the screening rules require substantially less time than solving (REG)-(CARD) to optimality. In this section we give pointers on how to do so effectively, and argue that in the context of branch-and-bound methods the overhead of the screening rules is linear in \( n \).

**Solving perspective relaxations.** Formulations (CR) and (CC) can be conveniently solved using off-the-shelf conic quadratic solvers (Aktürk et al. 2009, Günlük & Linderoth 2010) — this is the approach we use here. Pilanci et al. (2015) use a projected quasi-Newton method to solve (CC) which, they argue, is comparable in complexity to the lasso for low values of \( k \). Bertsimas & Van Parys (2017), Bertsimas et al. (2019) use a linear outer approximation method which they report performs faster than the lasso.

In fact, mixed-integer optimization methods based on formulations (MIPR) or (MIPC) will solve problems (CR) or (CC) at the root node of the branch-and-bound tree anyway. Thus, in this context, an optimal solution of the perspective relaxation can be obtained without an additional cost.

**Obtaining upper bounds.** There exist extensive work on heuristics for sparse regression, including stepwise selection methods (Efroymson 1966) and other methods mentioned in §1. Branch-and-bound methods, both based on off-the-shelf solvers or recent specialized implementations, use heuristics to warm-start the solvers and may even require them to initialize big-\( M \) values (Bertsimas et al. 2016, Dedieu et al. 2020). Thus, upper bounds in this context are available without incurring in additional costs.

In addition, feasible solutions for sparse regression problems can be obtained directly from convex relaxations. For example, Pilanci et al. (2015) use randomized rounding to obtain high quality feasible solutions of perspective relaxations. In our computations with cardinality constrained problems, we use a simpler rounding mechanism informed by Proposition 2 given
an optimal solution for (CC), we set \( z_i = 1 \) for the \( k \) largest values of \( \delta \) (breaking ties arbitrarily), and set \( x \) equal to the least squares estimator corresponding to the chosen variables.

**Additional operations.** It is easy to see that for problem \( \text{(REG)} \), given a convex relaxation solution and upper bound, the screening rule of Proposition 1 can be computed in \( O(n) \) time with a single pass along the variables. For \( \text{(CARD)} \), given a convex relaxation solution and upper bound, \( \delta_{[k]} \) and \( \delta_{[k+1]} \) can be selected in \( O(n) \) (without the need for sorting) and then the screening rule of Proposition 2 can be computed in \( O(n) \) time as well.

### 3.2. Derivation of the screening rules

We now derive the screening rules using Fenchel duality. Note that, whereas Pilanci et al. (2015) and Bertsimas & Van Parys (2017) derive their methods based on the Fenchel dual of the error term \( \|y - Ax\|_2^2 \), we instead use the dual of the perspective terms.

#### 3.2.1. Derivation of Proposition 1

Let \( h^*(p,q) \) be the bivariate convex conjugate of the perspective function \( x^2/z \), i.e.,

\[
(6) \quad h^*(p,q) = \max_{x,z} px + qz - \frac{x^2}{z}.
\]

From Fenchel’s inequality, we have

\[
(7) \quad px + qz - h^*(p,q) \leq h(x,z)
\]

for any \( p,q,x,z \in \mathbb{R} \). Employing (7) for each term to get a lower bound on (CR) and maximizing the lower bound, we obtain the Fenchel dual for (2):

\[
\begin{align}
\max_{p,q \in \mathbb{R}^n} & \min_{x,z} \|y - Ax\|_2^2 + \mu \sum_{i=1}^{n} z_i \\
&p - \frac{2x}{z} = 0 \\
&q + \left(\frac{x}{z}\right)^2 = 0,
\end{align}
\]

Indeed, the conjugate function \( h^* \) can be computed in closed form. Since (6) is concave in both \( x \) and \( z \), by taking derivatives with respect to \( x \) and \( z \) and setting to zero, we find the optimality conditions:

\[
\begin{align}
(9) & \quad p - \frac{2x}{z} = 0 \\
(10) & \quad q + \left(\frac{x}{z}\right)^2 = 0,
\end{align}
\]

since, otherwise, (6) is unbounded. The optimality conditions imply that

\[
\frac{p^2}{4} = -q \quad \text{and} \quad px + qz - \frac{x^2}{z} = 0,
\]
where the second inequality is obtained by multiplying (9) by $x$ and (10) by $y$, and summing them up. Thus,

$$h^*(p, q) = \begin{cases} 
0, & \text{if } q = -p^2/4 \\
+\infty, & \text{otherwise.} 
\end{cases}$$

Therefore, we find that (8) reduces to

$$\zeta_{FR} = \max_{p \in \mathbb{R}^n} \min_{x, z} \|y - Ax\|^2_2 + \mu \sum_{i=1}^n z_i$$

(11a) \hspace{2cm} \text{(FDR)}

$$+ \frac{1}{\gamma} \sum_{i=1}^n \left( p_i x_i - \frac{p_i^2}{4} z_i \right)$$

(11b)

s.t. $x \in \mathbb{R}^n$, $z \in [0, 1]^n$.

(11c)

In fact, if max and min are interchanged in (11), then $p_i^* = 2\frac{x_i}{z_i}$ (if $x_i$ and $z_i$ are both non-zero) and we recover precisely (CR); thus, there is no duality gap between (CR) and (FDR) and we have $\zeta_{CR} = \zeta_{FR}$.

In optimal solutions of the inner minimization problem we have

$$z_i = \begin{cases} 
0, & \text{if } \mu - \frac{p_i^2}{4\gamma} > 0 \\
1, & \text{if } \mu - \frac{p_i^2}{4\gamma} < 0 \\
\in [0, 1] & \text{otherwise.} 
\end{cases}$$

and $A'Ax = A'y - \frac{1}{2\gamma}p$. Note that if $\mu - \frac{p_i^2}{4\gamma} \neq 0$ for all $i = 1, \ldots, n$, then the optimal solution of the inner minimization problem in (FDR) is unique; in this case, by strong duality, that solution is also optimal for (CR) and, since it is integral, it is in fact optimal for (MIPR) as well. However, if $\mu - \frac{p_i^2}{4\gamma} = 0$ for some $i$, then the inner minimization problem in (FDR) has an infinite number of optimal solutions and the solution of (CR) may not be integral.

Now, let $x^*$ be an optimal solution of (CR) and $\varepsilon^* = y - Ax^*$ be the vector of residuals. Given $x^*$, a corresponding optimal dual solution $p^*$ can be recovered as $A'Ax^* = A'y - \frac{1}{2\gamma}p^*$, or $p^* = 2\gamma A'\varepsilon^*$. Moreover, we find that

$$\mu - \frac{(p_i^*)^2}{4\gamma} = \mu - \gamma(A_i'\varepsilon^*)^2 = \mu - \gamma \delta_i,$$

where $A_i$ is the $i$-th column of $A$. Consequently, optimal $(p^*, z^*)$ for (FDR) can be recovered from $\varepsilon^*$. We can now give the proof of Proposition [1]

**Proof of Proposition [1].** Suppose $\mu - \gamma \delta_i > 0$ and thus $z_i = 0$ in an optimal solution to (FDR). Note that in this case the inequality $\zeta_{CR} - \mu + \gamma \delta_i > \bar{\zeta}$ is never satisfied. Let $\zeta_{FR}(z_i = 1)$ be the optimal objective value of the Fenchel dual with the additional constraint $z_i = 1$. Note that

$$\zeta_{FR} + \mu - \gamma \delta_i = \zeta_{FR} + \mu - \frac{(p_i^*)^2}{4\gamma} \leq \zeta_{FR}(z_i = 1),$$

where the second inequality is obtained by multiplying (9) by $x$ and (10) by $y$, and summing them up. Thus,

$$h^*(p, q) = \begin{cases} 
0, & \text{if } q = -p^2/4 \\
+\infty, & \text{otherwise.} 
\end{cases}$$

Therefore, we find that (8) reduces to

$$\zeta_{FR} = \max_{p \in \mathbb{R}^n} \min_{x, z} \|y - Ax\|^2_2 + \mu \sum_{i=1}^n z_i$$

(11a) \hspace{2cm} \text{(FDR)}

$$+ \frac{1}{\gamma} \sum_{i=1}^n \left( p_i x_i - \frac{p_i^2}{4} z_i \right)$$

(11b)

s.t. $x \in \mathbb{R}^n$, $z \in [0, 1]^n$.

(11c)

In fact, if max and min are interchanged in (11), then $p_i^* = 2\frac{x_i}{z_i}$ (if $x_i$ and $z_i$ are both non-zero) and we recover precisely (CR); thus, there is no duality gap between (CR) and (FDR) and we have $\zeta_{CR} = \zeta_{FR}$.

In optimal solutions of the inner minimization problem we have

$$z_i = \begin{cases} 
0, & \text{if } \mu - \frac{p_i^2}{4\gamma} > 0 \\
1, & \text{if } \mu - \frac{p_i^2}{4\gamma} < 0 \\
\in [0, 1] & \text{otherwise.} 
\end{cases}$$

and $A'Ax = A'y - \frac{1}{2\gamma}p$. Note that if $\mu - \frac{p_i^2}{4\gamma} \neq 0$ for all $i = 1, \ldots, n$, then the optimal solution of the inner minimization problem in (FDR) is unique; in this case, by strong duality, that solution is also optimal for (CR) and, since it is integral, it is in fact optimal for (MIPR) as well. However, if $\mu - \frac{p_i^2}{4\gamma} = 0$ for some $i$, then the inner minimization problem in (FDR) has an infinite number of optimal solutions and the solution of (CR) may not be integral.

Now, let $x^*$ be an optimal solution of (CR) and $\varepsilon^* = y - Ax^*$ be the vector of residuals. Given $x^*$, a corresponding optimal dual solution $p^*$ can be recovered as $A'Ax^* = A'y - \frac{1}{2\gamma}p^*$, or $p^* = 2\gamma A'\varepsilon^*$. Moreover, we find that

$$\mu - \frac{(p_i^*)^2}{4\gamma} = \mu - \gamma(A_i'\varepsilon^*)^2 = \mu - \gamma \delta_i,$$

where $A_i$ is the $i$-th column of $A$. Consequently, optimal $(p^*, z^*)$ for (FDR) can be recovered from $\varepsilon^*$. We can now give the proof of Proposition [1]

**Proof of Proposition [1].** Suppose $\mu - \gamma \delta_i > 0$ and thus $z_i = 0$ in an optimal solution to (FDR). Note that in this case the inequality $\zeta_{CR} - \mu + \gamma \delta_i > \bar{\zeta}$ is never satisfied. Let $\zeta_{FR}(z_i = 1)$ be the optimal objective value of the Fenchel dual with the additional constraint $z_i = 1$. Note that

$$\zeta_{FR} + \mu - \gamma \delta_i = \zeta_{FR} + \mu - \frac{(p_i^*)^2}{4\gamma} \leq \zeta_{FR}(z_i = 1),$$
and the inequality is tight if the dual variables $p^*$ are still optimal after introducing the constraint $z_i = 1$. Thus, if $\zeta_{FR} + \mu - \gamma \delta_i > \bar{\zeta}$, we conclude that any feasible solution for (CR) with $z_i = 1$ has an objective worse than the upper bound and, in particular, there exists no optimal solution of (MIPR) with $z_i = 1$.

Similarly, suppose $\mu - \gamma \delta_i < 0$ and $z_i = 1$ in an optimal solution to (FDR). Since $\zeta_{FR} - \mu - \gamma \delta_i \leq \zeta_{FR}(z_i = 0)$, if the lower bound $\zeta_{FR} + \mu - (p_i^*)^2 / 4 \gamma > \bar{\zeta}$, we conclude that there exists no optimal MIP solution with $z_i = 0$. □

Remark 1. If $A'A$ is invertible, then an explicit formulation of the dual problem (11) can be obtained as

$$\max_{p \in \mathbb{R}^n} \|y\|^2 - \left( A'y - \frac{1}{2\gamma} p \right)' (A'A)^{-1} \left( A'y - \frac{1}{2\gamma} p \right) + \sum_{i=1}^n \min \left\{ 0, \mu - \frac{p_i^2}{4\gamma} \right\}.$$  

3.2.2. Derivation of Proposition 2. Using identical arguments as in §3.2.1, we find the Fenchel dual of (CC) as

$$\zeta_{FC} = \max_{p \in \mathbb{R}^n, x, z} \|y - Ax\|^2 + \frac{1}{\gamma} \sum_{i=1}^n (p_i x_i - \frac{p_i^2}{4\gamma} z_i)$$

(12a)

(FDC) s.t. $\sum_{i=1}^n z_i \leq k$, $x \in \mathbb{R}^n$, $z \in [0, 1]^n$.

As for (FDR) if max and min are interchanged, then $p_i^* = 2 \frac{x_i}{z_i}$ (if $x_i$ and $z_i$ are both non-zero) and we recover precisely (CC) ; thus, there is no duality gap between (CC) and (FDC) and we have $\zeta_{CC} = \zeta_{FC}$.

Observe that for the inner minimization problem, an optimal solution satisfies $z_i = 1$ for indices with the largest $k$ values of $p_i^2 / 4\gamma$ and $z_i = 0$ otherwise. Moreover, if there is no tie between the $k$-th and $(k+1)$-st largest value in an optimal solution of (FDC) , then this solution is unique and is also optimal for (CC) and (MIPC) . Otherwise, if there is a tie, then (CC) may not have optimal solutions integral in $z$.

Now, let $x^*$ be an optimal solution of the convex relaxation of (MIPC) , and let $\varepsilon^* = y - Ax^*$ be the vector of residuals. Then, the corresponding optimal dual solution $p^*$ can be recovered as $A'Ax^* = A'y - \frac{1}{2\gamma} p^*$, or $p^* = 2\gamma A'\varepsilon^*$. Moreover, we find that

$$-\frac{(p_i^*)^2}{4\gamma} = -\gamma (A_i'\varepsilon^*)^2 = -\gamma \delta_i.$$

Proof of Proposition 2. Suppose $\delta_i \leq \delta_{[k+1]}$. Then, $z_i = 0$ in an optimal solution of the inner minimization in (FDC) ; let $z_{[k]}$ be the indicator variables corresponding to the term $\delta_{[k]}$. Let $\zeta_{FC}(z_i = 1)$ be the optimal objective

1A similar result is given in (Pilanci et al. 2015, Prop. 1).
value of Fenchel dual with the additional constraint \( z_i = 1 \). The cardinality constraint implies that \( z_k = 0 \) for an optimal solution of this problem. Since \( \zeta_{CC} - \gamma \delta_i + \delta_k \leq \zeta_{FC}(z_i = 1) \), if the lower bound \( \zeta_{CC} - \gamma \delta_i + \delta_k > \bar{\zeta} \), we conclude that there exists no optimal solution to \( (MIPC) \) with \( z_i = 1 \).

Similarly, suppose \( \delta_i \geq \delta_k \); then, we have \( z_i = 1 \) in an optimal solution of the inner minimization of \( (FDR) \). Let \( \zeta_{FC}(z_i = 0) \) be the objective value of the Fenchel dual with the additional constraint \( z_i = 0 \). Since \( \zeta_{CC} + \gamma \delta_i - \delta_{k+1} \leq \zeta_{FC}(z_i = 0) \), if the lower bound \( \zeta_{CC} + \gamma \delta_i - \delta_{k+1} > \bar{\zeta} \), we conclude that there exists no optimal solution to \( (MIPC) \) with \( z_i = 0 \).

\[ \square \]

4. Computational experiments

In this section we report on our computational experiments to test the effectiveness of the screening rules for the cardinality constrained sparse regression problem \( (CARD) \). As the statistical merits of solving \( (CARD) \) are, by now, extensively documented in the literature \cite{Atamturk2019, Bertsimas2016, Bertsimas2019, Hastie2017, Hazimeh2018, Mazumder2017}, we focus on the impact of the safe screening rules on solving \( (MIPC) \) efficiently.

In our computations we use CPLEX 12.8 mixed-integer optimizer. All experiments are performed on a laptop with eight Intel(R) Core(TM) i7-8550 CPUs and 16GB RAM. In §4.1 we test the screening rules on “standard” synthetic data sets \cite{Atamturk2019, Bertsimas2016, Bertsimas2019, Hastie2017, Xie2020}, and in §4.2 we use the real data sets reported in Table 1. The “Diabetes” data set is first used by Efron et al. \cite{Efron2004}, whereas the other data sets are obtained from the UCI Machine Learning Repository \cite{Dua2017}.

**Table 1. Real data sets used.**

<table>
<thead>
<tr>
<th>Name</th>
<th>( n )</th>
<th>( m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>64</td>
<td>442</td>
</tr>
<tr>
<td>Crime</td>
<td>100</td>
<td>1993</td>
</tr>
<tr>
<td>Parkinsons</td>
<td>753</td>
<td>756</td>
</tr>
<tr>
<td>CNAE</td>
<td>856</td>
<td>1,081</td>
</tr>
<tr>
<td>Micromass</td>
<td>1,300</td>
<td>360</td>
</tr>
</tbody>
</table>

4.1. Synthetic data. We follow the data generation methodology of Bertsimas et al. \cite{Bertsimas2019}, where instances are generated according to a number of features of \( n \), number of rows \( m \), true sparsity \( k \), regularization parameter \( \gamma \), autocorrelation parameter \( \rho \), and signal noise ratio (SNR). In our experiments, we let \( n = 1,000, m = 500, k \in \{10,30,50\}, \gamma = 10^{i} \gamma_0 \) with \( i \in \{-1,0,2,4\} \) and \( \gamma_0 = \frac{n}{mk \max_i \|a_i\|^2} \) (where \( a_i \) denotes the \( i \)-th row of \( A \)), \( \rho \in \{0.2,0.5,0.7\} \), and \( \text{SNR} \in \{0.05,1.00,6.00\} \). The parameters \( m, \gamma, \rho \) and \( \text{SNR} \) coincide with the values used in Bertsimas et al. \cite{Bertsimas2019}. Our instances
are smaller with \( n = 1,000 \) and \( k \in \{10,30,50\} \) as we use a general purpose mixed-integer solver rather than a tailored solution method for (MIPC) as in Bertsimas et al. (2019). Several other papers in the literature generate data similarly. Finally, we set the time limit to ten minutes.

Figures 1 and 2 show aggregated results over all 540 synthetic instances tested. Figure 1 depicts the performance profiles of CPLEX with and without the safe screening rules proposed in the paper. We see that default CPLEX struggles with instances of this size, and is able to solve only 14% of the instances within the time limit; similar performance for general purpose MIP
Table 2. Number of variables fixed in synthetic instances with \( n = 1,000 \).

<table>
<thead>
<tr>
<th>( \gamma )</th>
<th>SNR</th>
<th>( \rho )</th>
<th>( k )</th>
<th>(.2)</th>
<th>(.5)</th>
<th>(.7)</th>
<th>( k )</th>
<th>(.2)</th>
<th>(.5)</th>
<th>(.7)</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 2^{-1} \gamma_0 )</td>
<td>1.00</td>
<td>1.000</td>
<td>0.997</td>
<td>0.897</td>
<td>0.791</td>
<td>0.693</td>
<td>1.000</td>
<td>0.997</td>
<td>0.897</td>
<td>0.791</td>
<td>930 ± 219</td>
</tr>
<tr>
<td>( 2^{-1} \gamma_0 )</td>
<td>6.00</td>
<td>1.000</td>
<td>0.997</td>
<td>0.897</td>
<td>0.791</td>
<td>0.693</td>
<td>1.000</td>
<td>0.997</td>
<td>0.897</td>
<td>0.791</td>
<td>930 ± 219</td>
</tr>
<tr>
<td>( 2^{0} \gamma_0 )</td>
<td>1.00</td>
<td>0.998</td>
<td>0.997</td>
<td>0.995</td>
<td>0.993</td>
<td>0.991</td>
<td>0.998</td>
<td>0.997</td>
<td>0.995</td>
<td>0.993</td>
<td>867 ± 232</td>
</tr>
<tr>
<td>( 2^{0} \gamma_0 )</td>
<td>6.00</td>
<td>1.000</td>
<td>0.999</td>
<td>0.996</td>
<td>0.995</td>
<td>0.993</td>
<td>1.000</td>
<td>0.999</td>
<td>0.997</td>
<td>0.995</td>
<td>867 ± 232</td>
</tr>
<tr>
<td>( 2^{2} \gamma_0 )</td>
<td>1.00</td>
<td>0.988</td>
<td>0.987</td>
<td>0.985</td>
<td>0.983</td>
<td>0.981</td>
<td>0.988</td>
<td>0.987</td>
<td>0.985</td>
<td>0.983</td>
<td>816 ± 226</td>
</tr>
<tr>
<td>( 2^{2} \gamma_0 )</td>
<td>6.00</td>
<td>1.000</td>
<td>0.998</td>
<td>0.997</td>
<td>0.996</td>
<td>0.995</td>
<td>1.000</td>
<td>0.998</td>
<td>0.997</td>
<td>0.996</td>
<td>816 ± 226</td>
</tr>
<tr>
<td>( 2^{4} \gamma_0 )</td>
<td>1.00</td>
<td>0.977</td>
<td>0.976</td>
<td>0.974</td>
<td>0.972</td>
<td>0.971</td>
<td>0.977</td>
<td>0.976</td>
<td>0.975</td>
<td>0.974</td>
<td>765 ± 391</td>
</tr>
<tr>
<td>( 2^{4} \gamma_0 )</td>
<td>6.00</td>
<td>1.000</td>
<td>0.979</td>
<td>0.978</td>
<td>0.976</td>
<td>0.975</td>
<td>1.000</td>
<td>0.979</td>
<td>0.978</td>
<td>0.976</td>
<td>765 ± 391</td>
</tr>
</tbody>
</table>

solvers has been observed in the literature for instances with \( n = 1,000 \) (Hastie et al. 2017, Xie & Deng 2020). In contrast, when the screening rules are incorporated, the performance improves substantially: it only takes 11 seconds to solve the same 14% of the instances, and 75% of the instances are provably solved to optimality within the ten-minuted time limit. Thus, for the synthetic instances that are solved to optimality by both methods, the screening procedure results in a \( 60 \times \) speedup. In fact, as Figure 2 shows, the screening procedures alone are sufficient to prove optimality for 23% of the instances, and are able to fix 75% or more of the variables in an additional 52% of the instances. There is, however, a small portion of the instances where few or no variables were fixed by the screening procedure.

Table 2 presents detailed information on the number of variables fixed as a function of the parameters \( k, \gamma, \rho, \) and SNR. Each entry in the table corresponds to an average over five identically generated instances. As the parameter \( k \) decreases (imposing higher \( \ell_0 \) regularization) and the parameter \( \gamma \) increases (imposing higher \( \ell_2 \) regularization), the screening procedures become more effective at fixing variables. We also observe that the screening rules are more effective when the signal-noise ratio is large, while the parameter \( \rho \) plays a relatively minor role.

4.2. Real data. We test the safe screening procedure in the data sets given in Table 1. For each data set, we solve problem (MIPC) with \( k \in \{10, 20, 30\} \). Bertsimas et al. (2019) indicate in the documentation of their code\(^2\) that setting \( \gamma = 1/\sqrt{m} \) is an appropriate scaling for regression problems. For this value of \( \gamma \), on average, 98.2% of the variables are fixed by the screening

\(^2\)https://github.com/jeanpauphilet/SubsetSelectionCIO.jl
procedure, and all instances are solved in four seconds. To better understand the effectiveness of the screening procedures for a broader set of parameters, we let $\gamma = 2^i \gamma_0$ with $i \in \{-1, 0, 1, 2, 3, 4, 5, 6, 7, 8\}$ and $\gamma_0$ as described in §4.1.

**Figure 3.** Number of instances solved as a function of the time (sec.)

**Figure 4.** Distribution of the number of variables fixed on instances with real data.

Figures 3 and 4 display the aggregated results over 150 instances tested with a time limit of one hour. The performance profile in Figure 3 shows that default CPLEX is able to solve 55% of the instances in one hour. When the screening rules are incorporated, the same 55% of the instances are solved in under 10 seconds, and 87% of the instances are solved within the time
limit of one hour. Therefore, for the instances that are solved to optimality by both methods, the screening procedure results in a $360 \times$ speedup. The distribution of the percentage of variables fixed (Figure 4) is similar to the one reported in §4.1 and 75% or more of the variables are fixed in 73% of the instances.

Figure 5 depicts the number of variables fixed for each data set and each value of $\gamma$; the points in the graph represent the average of three instances with different cardinalities. We observe that the screening procedure is able to fix most of the variables for $\gamma \leq 2^6 \gamma_0$. As $\gamma$ increases further, the strength of the perspective relaxation decreases and the screening procedure is unable to fix as many variables.

**Figure 5.** Proportion of variables fixed in instances with real data, where $\gamma = 2^i \gamma_0$. Each point is an average of three instances with different cardinalities $k$.

Finally, Table 3 shows four instances with the Diabetes data set where the screening procedure is able to fix only a small percentage of the variables, yet it results in substantial reduction in solution times. The table shows the time in seconds and the number of branch-and-bound nodes required to solve the problems to optimality, and the % of variables fixed by the screening procedure. Observe that even by fixing fewer than 20% of the variables, the screening rule leads to a substantial reduction in running times. In some cases, instances that are not solved to optimality within the one-hour time limit are solved in under 15 seconds with screening.

---

3 Instances with $k = 10$ on this dataset are solved in five seconds or less independently of the use of the screening procedure, and are omitted. Similarly, instances with $\gamma \geq 2^7 \gamma_0$ are solved in under 6 seconds, independent of the use of the screening procedure. Instances on datasets with $n \geq 100$ are rarely solved to optimality unless at least 50% of the variables are fixed.
Table 3. Sample instances with the Diabetes dataset illustrating impact of fixing a small number of variables.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\gamma$</th>
<th>CPLEX time</th>
<th>CPLEX+screening % fixed time nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>$\gamma_0/2$</td>
<td>968</td>
<td>10.9% 303 10,552</td>
</tr>
<tr>
<td>20</td>
<td>$\gamma_0$</td>
<td>2,080</td>
<td>14.1% &lt;1 0</td>
</tr>
<tr>
<td>30</td>
<td>$\gamma_0/2$</td>
<td>2,791</td>
<td>20.3% 444 30,903</td>
</tr>
<tr>
<td>30</td>
<td>$\gamma_0$</td>
<td>1hr limit</td>
<td>9.4% 12 272</td>
</tr>
</tbody>
</table>

5. Conclusion

We give a simple, yet very effective safe screening procedure for non-convex $\ell_0$ regression problems. Computational on synthetic and real data sets show that when used as preprocessing before solving the problems, the screening rules eliminate, on average, 76% of the binary variables, and consequently lead to substantial reduction in solution times. Additional research on strong convex relaxations should lead to even more effective safe screening rules.

Acknowledgments

Andrés Gómez is supported, in part, by the National Science Foundation under Grant 1818700. Alper Atamtürk is supported, in part, by grants from ARPA-E, NSF, and ONR.

References


