Safe Screening Rules for ℓ_0-Regression from Perspective Relaxations

Alper Atamtürk [1] **Andrés Gómez** [2]

Abstract

We give safe screening rules to eliminate variables from regression with ℓ_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 ℓ_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 a significant number of the variables can be removed quickly, hence reducing the computational burden for optimization substantially. Therefore, the proposed fast and effective screening rules extend the scope of algorithms for ℓ_0-regression to larger data sets.

1. Introduction

In machine learning and optimization communities, there is an increasing interest in regression models with ℓ_0 and ℓ_2 regularization:

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

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

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 ℓ_0 terms impose sparsity (Miller, 2002), which is a necessity for large-dimensional model inference (Hastie et al., 2001; 2015), and the ℓ_2 (ridge) regularization (Hoerl & Kennard, 1970) imposes bias/shrinkage in the regression coefficients. The ℓ_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 ℓ_1 (lasso, Tibshirani, 1996) and ℓ_1–ℓ_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 ℓ_1 counterparts. Nonetheless, solving the ℓ_0 problems in this manner is orders-of-magnitude slower than solving the ℓ_1 approximations and does not scale to problems with $n \geq 1{,}000$. Therefore, fast heuristics such as ℓ_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 ℓ_0-regression is NP-hard (Chen et al., 2019). 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 ℓ_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 ℓ_0-regression algorithms.

The purpose of this paper is to define *screening rules* for non-

*Equal contribution [1]IEOR, University of California, Berkeley, USA 94720 [2]ISE, University of Southern California, Los Angeles, USA 90089. Correspondence to: Alper Atamtürk <atamturk@berkeley.edu>, Andrés Gómez <gomezand@usc.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

convex ℓ_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; Ndiaye et al., 2017; 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 ℓ_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 nonconvex 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 ℓ_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 ℓ_0–ℓ_2 regularization terms, thus suitable for high noise regimes. The screening rules are obtained by exploiting convex perspective relaxations of the ℓ_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 ℓ_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 ℓ_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 (REG)-(CARD). In Section 3, we derive the safe screening rules for (REG) and (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 (REG) can be naturally formulated as the quadratic mixed-integer optimization problem

$$\min_{x,z} \|y - Ax\|_2^2 + \frac{1}{\gamma}\sum_{i=1}^n x_i^2 + \mu \sum_{i=1}^n z_i \tag{1a}$$

$$\text{s.t. } x_i(1-z_i) = 0, \qquad i = 1,\dots,n \tag{1b}$$

$$x \in \mathbb{R}^n,\ z \in \{0,1\}^n. \tag{1c}$$

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 $|x_i| \le M z_i$ (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:

$$\zeta_R = \min_{x,z} \|y - Ax\|_2^2 + \frac{1}{\gamma}\sum_{i=1}^n \frac{x_i^2}{z_i} + \mu \sum_{i=1}^n z_i \tag{2a}$$

$$\text{(MIPR)} \quad \text{s.t. } x_i(1-z_i) = 0, \qquad i = 1,\dots,n \tag{2b}$$

$$x \in \mathbb{R}^n,\ z \in \{0,1\}^n, \tag{2c}$$

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 strong mixed-integer formulation of (CARD) is

$$\zeta_C = \min_{x,z} \|y - Ax\|_2^2 + \frac{1}{\gamma}\sum_{i=1}^n \frac{x_i^2}{z_i} \tag{3a}$$

$$\text{(MIPC)} \quad \text{s.t.} \sum_{i=1}^n z_i \le k \tag{3b}$$

$$x_i(1-z_i) = 0, \qquad i = 1,\dots,n \tag{3c}$$

$$x \in \mathbb{R}^n,\ z \in \{0,1\}^n. \tag{3d}$$

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

$$\zeta_{CR} = \min_{x,z} \|y - Ax\|_2^2 + \frac{1}{\gamma}\sum_{i=1}^n \frac{x_i^2}{z_i} + \mu\sum_{i=1}^n z_i \tag{4a}$$

$$\text{(CR)} \qquad x \in \mathbb{R}^n,\ z \in [0,1]^n, \tag{4b}$$

of (MIPR) , and the convex relaxation

$$\zeta_{CC} = \min_{x,z} \|y - Ax\|_2^2 + \frac{1}{\gamma}\sum_{i=1}^n \frac{x_i^2}{z_i} \tag{5a}$$

$$\text{(CC)} \qquad \sum_{i=1}^n z_i \le k,\ x \in \mathbb{R}^n,\ z \in [0,1]^n \tag{5b}$$

of (MIPC) . Note that the z-variables can be easily projected out in formulation (4), resulting in a formulation involving the *reverse Huber penalty* (Pilanci et al., 2015).

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. We denote by A_i the i-th column of A.

Proposition 1 (Safe screening rules for REG)**.** *Let x^* be an optimal solution to (CR) with objective value ζ_{CR}, $\varepsilon^* = y - Ax^*$, $\delta_i = (A_i'\varepsilon^*)^2$, $i = 1,\dots,n$, and let $\bar\zeta$ be an upper bound on ζ_R. 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 ζ_{CC}, $\varepsilon^* = y - Ax^*$, $\delta_i = (A_i'\varepsilon^*)^2$, $i = 1,\dots,n$, $\delta_{[k]}$ be the k-th largest value of vector δ, and let $\bar\zeta$ be an upper bound on ζ_C. Then any optimal solution to (MIPC) satisfies,*

$$z_i = \begin{cases} 0, \text{ if } \delta_i \le \delta_{[k+1]} \text{ and } \zeta_{CC} - \gamma(\delta_i - \delta_{[k]}) > \bar\zeta \\ 1, \text{ if } \delta_i \ge \delta_{[k]} \text{ and } \zeta_{CC} + \gamma(\delta_i - \delta_{[k+1]}) > \bar\zeta. \end{cases}$$

Remark 1. Suppose that the optimal residual ε^* are unknown but approximated by $\bar\varepsilon$ such that $\|\varepsilon^* - \bar\varepsilon\|_2^2 \le \tau$. Letting $\bar\delta_i = (A_i'\bar\varepsilon)^2$, we find that

$$\delta_i - \bar\delta_i = \Big(A_i'(\varepsilon^* - \bar\varepsilon)\Big)^2 \le \|A_i\|_2^2\tau.$$

Therefore, by adding $\gamma\|A_i\|_2^2\tau$ to the right hand side of the screening rules in Proposition 1, it is still possible to ensure that z variables are fixed to their optimal values. Similar arguments can be made concerning screening rules in Proposition 2, as well as uncertainties concerning the exact value of ζ_{CR} or ζ_{CC}.

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. Finally, Hazimeh & Mazumder (2019) and Hazimeh et al. (2020) have reported promising results with solving the regularized problem (CR) using first-order coordinate descent methods.

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 δ (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 (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 (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.

The screening rules presented in this paper can also be used as *dynamic safe rules* (Bonnefoy et al., 2014; 2015). Note that branch-and-bound methods also solve restricted versions of (CR) or (CC) while exploring the search space. Thus, thanks to their fast computational time, the screening outlined in Propositions 1-2 can be used throughout the branch-and-bound algorithm to fix variables corresponding to a partial tree, resulting in more aggressive pruning of the search space.

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.,

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

From Fenchel's inequality, we have

$$px + qz - h^*(p, q) \leq h(x, z) \tag{7}$$

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):

$$\max_{p,q\in\mathbb{R}^n} \min_{x,z} \|y - Ax\|_2^2 + \mu \sum_{i=1}^n z_i \tag{8a}$$

$$+ \frac{1}{\gamma} \sum_{i=1}^n \Big(p_i x_i + q_i z_i - h^*(p_i, q_i) \Big) \tag{8b}$$

$$\text{s.t. } x \in \mathbb{R}^n,\ z \in [0, 1]^n. \tag{8c}$$

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:

$$p - \frac{2x}{z} = 0 \tag{9}$$

$$q + \left(\frac{x}{z}\right)^2 = 0, \tag{10}$$

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

$$\frac{p^2}{4} = -q \text{ and } px + qz - \frac{x^2}{z} = 0,$$

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

$$\zeta_{FR} = \max_{p\in\mathbb{R}^n} \min_{x,z} \|y - Ax\|_2^2 + \mu \sum_{i=1}^n z_i \tag{11a}$$

$$\text{(FDR)} \qquad + \frac{1}{\gamma} \sum_{i=1}^n \Big(p_i x_i - \frac{p_i^2}{4} z_i \Big) \tag{11b}$$

$$\text{s.t. } x \in \mathbb{R}^n,\ z \in [0, 1]^n. \tag{11c}$$

In fact, max and min can be interchanged in (11), since setting $p_i^* = 2\frac{x_i}{z_i}$ (if x_i and z_i are both non-zero) 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 ε^*. 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

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

and the inequality is tight if the dual variables p^* are still optimal 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 2. If $A'A$ is invertible, then an explicit formulation of the dual problem (11) can be obtained as

$$\max_{p\in\mathbb{R}^n} \|y\|_2^2 - \left(A'y - \frac{1}{2\gamma}p\right)' (A'A)^{-1}\left(A'y - \frac{1}{2\gamma}p\right) + \sum_{i=1}^n \min\left\{0, \mu - \frac{p_i^2}{4\gamma}\right\}.$$

3.2.2. Derivation of Proposition 2

Using identical arguments as in §3.2.1, we find the Fenchel dual of (CC) as

$$\zeta_{FC} = \max_{p\in\mathbb{R}^n} \min_{x,z} \|y - Ax\|_2^2 + \frac{1}{\gamma}\sum_{i=1}^n \left(p_i x_i - \frac{p_i^2}{4} z_i\right) \quad (12a)$$

$$\text{(FDC)} \qquad \text{s.t.} \sum_{i=1}^n z_i \leq k,\ x \in \mathbb{R}^n, z \in [0,1]^n. \quad (12b)$$

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 $\frac{p_i^2}{4\gamma}$ 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[1] 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 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 $z_i = 0$. □

4. Computational experiments

In this section we report on our computational experiments to test the effectiveness of the screening rules for the cardi-

[1] A similar result is given in (Pilanci et al., 2015, Prop. 1).

nality constrained sparse regression problem (CARD) . As the statistical merits of solving (CARD) are, by now, extensively documented in the literature (Atamtürk & Gómez, 2019; Bertsimas et al., 2016; Bertsimas & Van Parys, 2017; Bertsimas et al., 2019; Hastie et al., 2017; Hazimeh & Mazumder, 2018; Mazumder et al., 2017), 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 (Atamtürk & Gómez, 2019; Bertsimas et al., 2016; 2019; Hastie et al., 2017; Xie & Deng, 2020), 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. (2004), whereas the other data sets are obtained from the UCI Machine Learning Repository (Dua & Graff, 2017).

Table 1: Real data sets used.

Name	n	m
Diabetes	64	442
Autos	74	193
Crime	100	1993
UJIndoorLoc	520	19,937
Micromass	1,300	360

4.1. Synthetic data

We follow the data generation methodology of Bertsimas et al. (2019), where instances are generated according to a number of features of n, number of rows m, true sparsity k, regularization parameter γ, autocorrelation parameter ρ, and signal noise ratio (SNR). In our experiments, we let $n = 1{,}000$, $m = 500$, $k \in \{10, 30, 50\}$, $\gamma = 2^i\gamma_0$ with $i \in \{-1, 0, 2, 4\}$ and $\gamma_0 = \frac{n}{mk \max_i \|a_i\|_2^2}$ (where a_i denotes the i-th row of A), $\rho \in \{0.2, 0.5, 0.7\}$, and SNR $\in \{0.05, 1.00, 6.00\}$. The parameters m, γ, ρ and SNR coincide with the values used in Bertsimas et al. (2019). 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 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.

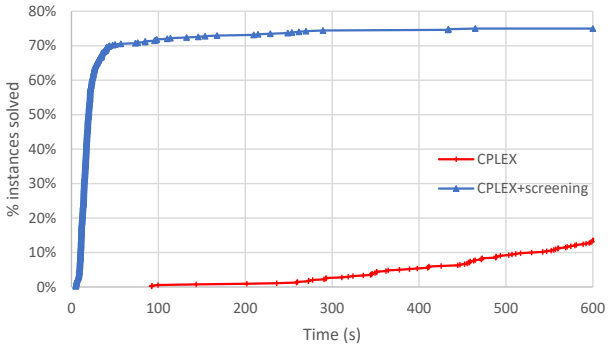

Figure 1: Number of synthetic instances solved as a function of the time in seconds.

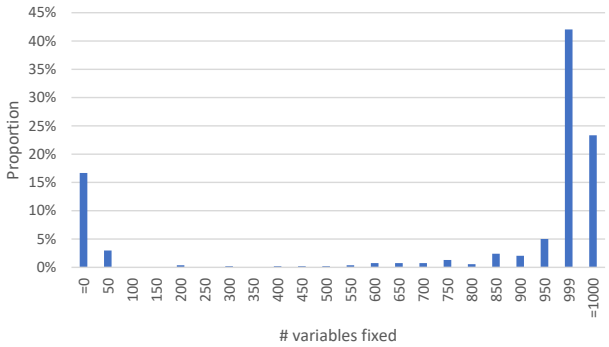

Figure 2: Distribution of the number of variables fixed across all synthetic instances.

Table 2 presents detailed information on the number of variables fixed as a function of the parameters k, γ, ρ, and SNR. Each entry in the table corresponds to an average over five identically generated instances. As the parameter k decreases (imposing higher ℓ_0 regularization) and the parameter γ increases (imposing higher ℓ_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 ρ plays a relatively minor role.

Table 2: Number of variables fixed in synthetic instances with $n = 1,000$.

γ	SNR	k / ρ	10 .2	10 .5	10 .7	30 .2	30 .5	30 .7	50 .2	50 .5	50 .7	Average
$2^{-1}\gamma_0$	0.05		995	996	993	997	791	983	996	792	934	930 ± 219
	1.00		1,000	998	999	997	912	473	906	982	950	
	6.00		1,000	1,000	997	993	988	987	1000	752	692	
$2^{0}\gamma_0$	0.05		983	986	991	980	972	988	988	989	800	967 ± 147
	1.00		998	997	996	958	977	988	973	789	994	
	6.00		1,000	999	995	997	993	997	785	997	992	
$2^{2}\gamma_0$	0.05		553	245	621	893	640	751	804	952	902	886 ± 232
	1.00		991	988	977	971	969	940	968	976	962	
	6.00		1,000	1,000	983	978	980	974	962	975	968	
$2^{4}\gamma_0$	0.05		0	0	0	0	0	0	1	111	0	276 ± 410
	1.00		577	194	174	302	455	457	40	166	144	
	6.00		1,000	999	597	939	379	109	40	297	471	
Average			801 ± 385			770 ± 377			723 ± 409			765 ± 391

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 γ, on average, 98.2% of the variables are fixed by the screening 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 γ_0 as described in §4.1.

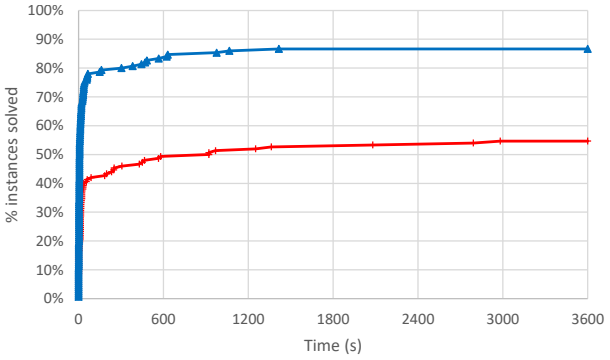

Figure 3: Number of real data instances solved as a function of the time in seconds.

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

[2] https://github.com/jeanpauphilet/SubsetSelectionCIO.jl.

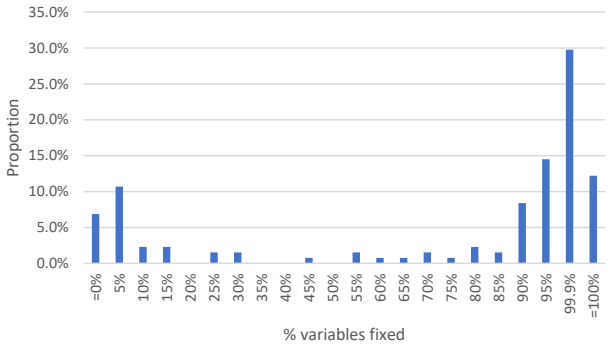

Figure 4: Distribution of the number of variables fixed across all instances with real data.

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× *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 70% of the instances.

Figure 5 depicts the number of variables fixed for each data set and each value of γ; 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 $\gamma \leq 2^5\gamma_0$. As γ increases further, the strength of the perspective relaxation decreases and the screening procedure is unable to fix as many variables.

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[3]. The table shows the time in sec-

[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

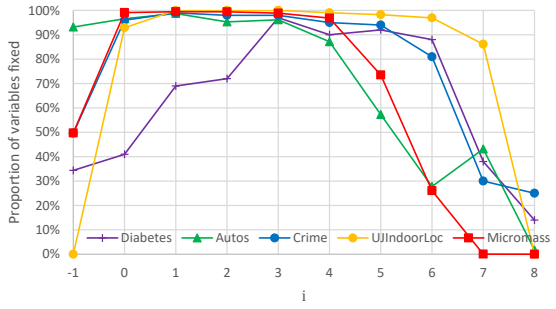

Figure 5: Proportion of variables fixed in instances with real data, where $\gamma = 2^i\gamma_0$. Each point is an average of three instances with different cardinalities k.

onds 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.

Table 3: Sample instances with the Diabetes dataset illustrating impact of fixing a small number of variables.

k	γ	CPLEX		CPLEX+screening		
		time	nodes	% fixed	time	nodes
20	$\gamma_0/2$	968	48,050	10.9%	303	10,552
20	γ_0	2,080	80,095	14.1%	<1	0
30	$\gamma_0/2$	2,791	119,638	20.3%	444	30,903
30	γ_0	1hr limit	168,311	9.4%	12	272

5. Conclusion

We give a simple, yet very effective safe screening procedure for non-convex ℓ_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. Strong convex relaxations of ℓ_0 formulations are key to the success of the safe screening rules. Screening rules based on stronger relaxations (Atamtürk et al., 2018; Atamtürk & Gómez, 2018; Atamtürk & Gómez, 2019; Han et al., 2020) than the perspective formulations considered in this paper should lead to even more effective safe screening rules.

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.

Acknowledgements

Alper Atamtürk is supported, in part, by NSF award 1807260, DOE ARPA-E grant 260801540061, and DOD ONR grant 12951270. Andrés Gómez is supported, in part, by grants 1818700 and 1930582 from the National Science Foundation.

References

Aktürk, M. S., Atamtürk, A., and Gürel, S. A strong conic quadratic reformulation for machine-job assignment with controllable processing times. *Operations Research Letters*, 37(3):187–191, 2009.

Atamtürk, A. and Gómez, A. Strong formulations for quadratic optimization with m-matrices and indicator variables. *Mathematical Programming*, 170(1):141–176, 2018.

Atamtürk, A. and Gómez, A. Rank-one convexification for sparse regression. *arXiv preprint arXiv:1901.10334*, 2019.

Atamtürk, A., Nemhauser, G. L., and Savelsbergh, M. W. Conflict graphs in solving integer programming problems. *European Journal of Operational Research*, 121(1):40–55, 2000.

Atamtürk, A., Gómez, A., and Han, S. Sparse and smooth signal estimation: Convexification of L0 formulations. *arXiv preprint arXiv:1811.02655*, 2018.

Bertsimas, D. and Van Parys, B. Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *arXiv preprint arXiv:1709.10029*, 2017.

Bertsimas, D., King, A., Mazumder, R., et al. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

Bertsimas, D., Pauphilet, J., and Van Parys, B. Sparse regression: Scalable algorithms and empirical performance. *arXiv preprint arXiv:1902.06547*, 2019.

Bonnefoy, A., Emiya, V., Ralaivola, L., and Gribonval, R. A dynamic screening principle for the lasso. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pp. 6–10. IEEE, 2014.

Bonnefoy, A., Emiya, V., Ralaivola, L., and Gribonval, R. Dynamic screening: Accelerating first-order algorithms for the lasso and group-lasso. *IEEE Transactions on Signal Processing*, 63(19):5121–5132, 2015.

Chen, Y., Ye, Y., and Wang, M. Approximation hardness for a class of sparse optimization problems. *Journal of Machine Learning Research*, 2019.

Cozad, A., Sahinidis, N. V., and Miller, D. C. Learning surrogate models for simulation-based optimization. *AIChE Journal*, 60(6):2211–2227, 2014.

Dedieu, A., Hazimeh, H., and Mazumder, R. Learning sparse classifiers: Continuous and mixed integer optimization perspectives. *arXiv preprint arXiv:2001.06471*, 2020.

Dong, H., Chen, K., and Linderoth, J. Regularization vs. relaxation: A conic optimization perspective of statistical variable selection. *arXiv preprint arXiv:1510.06083*, 2015.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. Least angle regression. *The Annals of Statistics*, 32(2): 407–499, 2004.

Efroymson, M. Stepwise regression–a backward and forward look. *Florham Park, New Jersey*, 1966.

El Ghaoui, L. and Lebret, H. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications*, 18(4):1035–1064, 1997.

El Ghaoui, L. E., Viallon, V., and Rabbani, T. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*, 2010.

Fercoq, O., Gramfort, A., and Salmon, J. Mind the duality gap: Safer rules for the lasso. *arXiv preprint arXiv:1505.03410*, 2015.

Frangioni, A. and Gentile, C. Perspective cuts for a class of convex 0–1 mixed integer programs. *Mathematical Programming*, 106(2):225–236, 2006.

Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/.

Gómez, A. and Prokopyev, O. A mixed-integer fractional optimization approach to best subset selection. *Optimization-online*, 2018.

Günlük, O. and Linderoth, J. Perspective reformulations of mixed integer nonlinear programs with indicator variables. *Mathematical Programming*, 124(1-2):183–205, 2010.

Han, S., Gómez, A., and Atamtürk, A. 2x2 convexifications for convex quadratic optimization with indicator variables. *arXiv preprint arXiv:2004.07448*, 2020.

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: Data mining, inference, and prediction*, volume 1. Springer series in statistics New York, NY, USA, 2001.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical learning with sparsity: The lasso and generalizations*. CRC press, 2015.

Hastie, T., Tibshirani, R., and Tibshirani, R. J. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*, 2017.

Hazimeh, H. and Mazumder, R. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *arXiv preprint arXiv:1803.01454*, 2018.

Hazimeh, H. and Mazumder, R. Learning hierarchical interactions at scale: A convex optimization approach. *arXiv preprint arXiv:1902.01542*, 2019.

Hazimeh, H., Mazumder, R., and Saab, A. Sparse regression at scale: Branch-and-bound rooted in first-order optimization. *arXiv preprint arXiv:2004.06152*, 2020.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Kimura, K. and Waki, H. Minimization of akaike's information criterion in linear regression analysis via mixed integer nonlinear program. *Optimization Methods and Software*, 33(3):633–649, 2018.

Mazumder, R., Radchenko, P., and Dedieu, A. Subset selection with shrinkage: Sparse linear modeling when the snr is low. *arXiv preprint arXiv:1708.03288*, 2017.

Miller, A. *Subset selection in regression*. CRC Press, 2002.

Miyashiro, R. and Takano, Y. Subset selection by Mallows' Cp: A mixed integer programming approach. *Expert Systems with Applications*, 42(1):325–331, 2015.

Ndiaye, E., Fercoq, O., Gramfort, A., and Salmon, J. Gap safe screening rules for sparsity enforcing penalties. *The Journal of Machine Learning Research*, 18(1):4671–4703, 2017.

Ogawa, K., Suzuki, Y., and Takeuchi, I. Safe screening of non-support vectors in pathwise SVM computation. In *International Conference on Machine Learning*, pp. 1382–1390, 2013.

Park, Y. W. and Klabjan, D. Subset selection for multiple linear regression via optimization. *arXiv preprint arXiv:1701.07920*, 2017.

Pilanci, M., Wainwright, M. J., and El Ghaoui, L. Sparse learning via boolean relaxations. *Mathematical Programming*, 151(1):63–87, 2015.

Savelsbergh, M. W. P. Preprocessing and probing techniques for mixed integer programming problems. *ORSA J. on Computing*, 6(4):445–454, 1994.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74 (2):245–266, 2012.

Wang, J., Zhou, J., Wonka, P., and Ye, J. Lasso screening rules via dual polytope projection. In *Advances in Neural Information Processing Systems*, pp. 1070–1078, 2013.

Xiang, Z. J. and Ramadge, P. J. Fast lasso screening tests based on correlations. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2137–2140. IEEE, 2012.

Xiang, Z. J., Wang, Y., and Ramadge, P. J. Screening tests for lasso problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):1008–1027, 2016.

Xie, W. and Deng, X. Scalable algorithms for the sparse ridge regression. 2020.

Xu, H., Caramanis, C., and Mannor, S. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(Jul):1485–1510, 2009.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodology)*, 67(2):301–320, 2005.