Safe Screening for Logistic Regression with ℓ_0–ℓ_2 Regularization

Anna Deza[a] and Alper Atamtürk[b]
Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA, U.S.A.

Keywords: Screening Rules, Sparse Logistic Regression.

Abstract: In logistic regression, it is often desirable to utilize regularization to promote sparse solutions, particularly for problems with a large number of features compared to available labels. In this paper, we present screening rules that safely remove features from logistic regression with $\ell_0-\ell_2$ regularization before solving the problem. The proposed safe screening rules are based on lower bounds from the Fenchel dual of strong conic relaxations of the logistic regression problem. Numerical experiments with real and synthetic data suggest that a high percentage of the features can be effectively and safely removed apriori, leading to substantial speed-up in the computations.

1 INTRODUCTION

Logistic regression is a classification model used to predict the probability of a binary outcome from a set of features. Its use is prevalent in a large variety of domains, from diagnostics in healthcare (Gramfort et al., 2013; Shevade and Keerthi, 2003; Cawley and Talbot, 2006) to sentiment analysis in natural language processing (Wang and Park, 2017; Yen et al., 2011) and consumer choice modeling in economics (Kuswanto et al., 2015).

Given a data matrix $A \in \mathbb{R}^{m\times n}$ of m observations, each with n features and binary labels $y \in \{-1,1\}^m$, the logistic regression model seeks regression coefficients $x \in \mathbb{R}^n$ that minimize the convex loss function

$$\mathcal{L}(x) = \frac{1}{m}\sum_{i=1}^{m} \log\Big(1+\exp(-y_i A_i x)\Big).$$

We use A_i to denote the i-th row of matrix A and A^j to denote the j-th column of A. When the number of available features is large compared to the number of the observations (labels), i.e., $m << n$, logistic regression models are prone to overfitting. Such cases require pruning the features to mitigate the risk of overfitting. Regularization is a natural approach for this purpose. Convex ℓ_2-regularization (ridge) (Hoerl and Kennard, 1970) imposes bias by shrinking the regression coefficients x_i, $i \in [n]$, toward zero. The ℓ_1-regularization (lasso) (Tibshirani, 1996) and ℓ_1–ℓ_2-regularization (elastic net) (Zou and Hastie, 2005) perform shrinkage of the coefficients and selection of the features simultaneously. Recently, there has been a growing interested in utilizing the exact ℓ_0-regularization (Miller, 2002; Bertsimas et al., 2016) for selecting features in linear regression. Although ℓ_0-regularization introduces non-convexity to regression models, significant progress has been done to develop strong models and specialized algorithms to solve medium to large scale instances recently (e.g. Bertsimas and Van Parys, 2017; Atamtürk and Gómez, 2019; Hazimeh and Mazumder, 2020; Han et al., 2020).

We consider logistic regression with ℓ_0–ℓ_2 regularization:

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

$$\min_{x\in\mathbb{R}^n} \mathcal{L}(x) + \frac{1}{\gamma}\|x\|_2^2 \text{ s.t. } \|x\|_0 \le k. \qquad \text{(CARD)}$$

Whereas the ℓ_2-regularization penalty term above encourages shrinking the coefficients, which helps counter effects of noise present in the data matrix A, the ℓ_0-regularization penalty term in (REG) encourages sparsity, selecting a small number of key features to be used for prediction, which is modeled as an explicit cardinality constraint in (CARD). Due to the ℓ_0-regularization terms, (REG) and (CARD) are non-convex optimization problems.

Screening rules refer to preprocessing procedures that discard certain features, leading to a reduction in the dimension of the problem, which, in turn, improves the solution times of the employed algorithms. For ℓ_1-regularized linear regression, El Ghaoui et al. (2010) introduce safe screening rules that guarantee to remove only features that are not selected in the solution. Strong rules (Tibshirani, 2011), on the other

[a] https://orcid.org/0000-0002-4849-683X
[b] https://orcid.org/0000-0003-1220-808X

Deza, A. and Atamtürk, A.
Safe Screening for Logistic Regression with – Regularization.
DOI: 10.5220/0011578100003335
In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - Volume 1: KDIR*, pages 119-126
ISBN: 978-989-758-614-9; ISSN: 2184-3228

hand, are heuristics with no guarantee but able to prune a large number of features fast. A large body of work exists on screening rules for ℓ_1-regularized regression (Wang et al., 2013; Liu et al., 2014; Fercoq et al., 2015; Ndiaye et al., 2017; Dantas et al., 2021), including some for logistic regression (Wang et al., 2014). However, little attention has been given to the ℓ_0-regularized regression problem, where dimension reduction by screening rules can have substantially larger impact due to the higher computational burden for solving the non-convex regression problems. Bounds from strong conic relaxations of ℓ_0-regularized problems (Atamtürk et al., 2021; Atamtürk and Gómez, 2019) substantially reduce the computational burden with effective pruning strategies. Recently, Atamtürk and Gómez (2020) propose safe screening rules for the ℓ_0-regularized linear regression problem from perspective relaxations. To the best of our knowledge, no screening rule exists in the literature for the logistic regression problems (REG) and (CARD) with ℓ_0–ℓ_2 regularization, studied in this paper.

Outline. In Section 2, we give strong conic mixed 0-1 formulations for logistic regression problems (REG) and (CARD) with ℓ_0–ℓ_2 regularization. In Section 3, we derive the safe screening rules for them based on bounds from Fenchel duals of their conic relaxations and in Section 4, we summarize the computational experiments performed for testing the effectiveness of the proposed screening rules for ℓ_0–ℓ_2 logistic regression problems with synthetic as well as real data. Finally, we conclude with a few final remarks in Section 5.

2 CONIC REFORMULATIONS

In this section, we present strong conic formulations for (REG) and (CARD). First, we state convex logistic regression loss minimization as a conic optimization problem. Writing the epigraph of the softplus function $\log(1+\exp(x)) \leq s$ as an upper bound on the sum of two exponential functions $\exp(x-s)+\exp(-s) \leq 1$, it follows that the logistics regression loss $\mathcal{L}(x)$ minimization problem can be formulated as an exponential cone optimization problem

$$\begin{aligned}\min_{x\in\mathbb{R}^n} \mathcal{L}(x) = \min_{x,s,u,v} \ & \frac{1}{m}\sum_{i=1}^{m} s_i \\ \text{s.t. } & u_i \geq \exp(-y_i A_i x - s_i), && i\in[m] \\ & v_i \geq \exp(s_i), && i\in[m] \\ & u_i + v_i \leq 1, && i\in[m] \\ & x\in\mathbb{R}^n,\ s,u,v\in\mathbb{R}^m,\end{aligned}$$

which is readily solvable by modern conic optimization solvers.

Introducing binary indicator variables $z\in\{0,1\}^n$ to model the ℓ_0-regularization terms, (REG) can be formulated as a mixed-integer conic optimization problem:

$$\eta_R = \min \mathcal{L}(x) + \frac{1}{\gamma}\sum_{j=1}^{n}\frac{x_j^2}{z_j} + \mu\sum_{j=1}^{n} z_j \tag{2a}$$

$$\text{s.t. } x_j(1-z_j)=0, \quad j\in[n] \tag{2b}$$

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

Here, we adopt the convention $x_j^2/z_j = 0$ if $z_j = 0$, $x_j = 0$ and $x_j^2/z_j = \infty$ if $z_j = 0$, $x_j \neq 0$. Constraint (2b) ensures that $x_j = 0$ whenever $z_j = 0$. This constraint can be linearized using the "big-M" technique by replacing it with $-Mz_j \leq x_j \leq Mz_j$, where M is a large enough positive scalar. However, such big-M constraints lead to very weak convex relaxations as we show in the computational experiments in Section 4.

Instead, we use the conic formulation of the perspective function of x_j^2 to model them more effectively. Replacing x_j^2 in the objective with its perspective function x_j^2/z_j significantly strengthens the convex relaxation when $0 < z_j < 1$, and introducing $t_j \geq 0$, the perspective can be stated as a rotated second-order cone constraint $x_j^2 \leq z_j t_j$ Aktürk et al. (2009). Dropping the complementary constraints (2b) as well as the integrality constraints on z, we arrive at the respective conic (convex) relaxation for (REG):

$$\eta_{CR} = \min \mathcal{L}(x) + \frac{1}{\gamma}\sum_{j=1}^{n} t_j + \mu\sum_{j=1}^{n} z_j \tag{3a}$$

$$x_j^2 \leq z_j t_j,\ j\in[n] \tag{3b}$$

$$x\in\mathbb{R}^n, t\in{\mathbb{R}_+}^n, z\in[0,1]^n. \tag{3c}$$

Note that constraint (3b) is valid for $z\in\{0,1\}^n$: as $z_j = 0$ implies $x_j = 0$ and $z_j = 1$ implies simply $x_j^2 \leq t_j$, $j\in[n]$.

Similarly, one can write the cardinality-constrained version (CARD) as a mixed integer non-linear model with the perspective reformulation:

$$\eta_C = \min \mathcal{L}(x) + \frac{1}{\gamma}\sum_{j=1}^{n}\frac{x_j^2}{z_j} \tag{4a}$$

$$\text{s.t. } \sum_{j=1}^{n} z_j \leq k \tag{4b}$$

$$x_j(1-z_j)=0, \quad j\in[n] \tag{4c}$$

$$x\in\mathbb{R}^n, z\in\{0,1\}^n. \tag{4d}$$

Dropping (4c) and integrality constraints, and stating the perspectives as rotated cone constraints, we arrive at the conic relaxation for (CARD):

$$\eta_{CC} = \min \mathcal{L}(x) + \frac{1}{\gamma}\sum_{j=1}^{n} t_j \tag{5a}$$

$$\text{s.t.} \sum_{j=1}^{n} z_j \leq k \tag{5b}$$

$$x_j^2 \leq z_j t_j,\ j \in [n] \tag{5c}$$

$$x \in \mathbb{R}^n, t \in \mathbb{R}_+{}^n, z \in [0,1]^n. \tag{5d}$$

3 SAFE SCREENING RULES

In this section, we first present the safe screening rules for logistic regression with ℓ_0–ℓ_2 regularization and then discuss their derivation.

Proposition 1 (Safe Screening Rule for Regularized Logistic Regression (REG))**.** *Let x^* be an optimal solution to* (3)*, with objective value η_{CR}, $\alpha_i = y_i/\big(1+\exp(y_i A_i x^*)\big)$, $i \in [m]$, $\delta_j = \frac{1}{4}(\alpha' A^j)^2$, $j \in [n]$, and $\bar{\eta}_R$ be an upper bound on η_R. Then any optimal solution to* (2) *satisfies*

$$z_j = \begin{cases} 0, & \text{if } \eta_{CR} + \mu - \gamma\delta_j > \bar{\eta}_R \\ 1, & \text{if } \eta_{CR} - \mu + \gamma\delta_j > \bar{\eta}_R. \end{cases}$$

Proposition 2 (Safe Screening Rule for Cardinality–constrained Logistic Regression (CARD))**.** *Let x^* be an optimal solution to* (5)*, with objective value η_{CC}, $\alpha_i = y_i/(1+exp(y_i A_i x^*))$, $i \in [m]$, $\delta_j = \frac{1}{4}(\alpha' A^j)^2$, $j \in [n]$, $\delta_{[k]}$ denote the k-th largest value of δ, and $\bar{\eta}_C$ be an upper bound on η_C. Then any optimal solution to* (4) *satisfies*

$$z_j = \begin{cases} 0, & \text{if } \delta_j \leq \delta_{[k+1]} \text{ and } \eta_{CC} - \gamma(\delta_j - \delta_{[k]}) > \bar{\eta}_C \\ 1, & \text{if } \delta_j \geq \delta_{[k]} \text{ and } \eta_{CC} + \gamma(\delta_j - \delta_{[k+1]}) > \bar{\eta}_C. \end{cases}$$

3.1 Derivation of Proposition 1

In this section, we present the derivation for the screening rule for (REG) via Fenchel duality. Similar to Atamtürk and Gómez (2020), we utilize the dual of the perspective terms. In particular, for $p, q \in \mathbb{R}$, consider the convex conjugate, $h^*(p,q)$ of the perspective function $h(x,z) = x^2/z$:

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

By Fenchel's inequality, we have $px + qz - h^*(p,q) \leq \frac{x^2}{z}$. Therefore, for any $p, q \in \mathbb{R}^n$, we can replace the perspective terms in the objective of (3) to derive a lower bound on η_{CR}. Then, the Fenchel dual of (3) is obtained by maximizing the lower bound:

$$\max_{p,q} \min_{x,z\in[0,1]^n} \mathcal{L}(x) + \mu\sum_{j=1}^{n} z_j + \frac{1}{\gamma}\Big(p'x + q'z - \sum_{j=1}^{n} h^*(p_j, q_j)\Big). \tag{7}$$

Observing that $px + qz - \frac{x^2}{z}$ is concave in x and z, allows one to get a closed form solution for (6). Indeed, by simply setting the partial derivatives to zero, we obtain

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

Then, replacing q_j with $-p_j^2/4$ and using the closed form solution for h^*, we obtain from (7) the simplified form of the Fenchel dual:

$$\eta_{FR} = \max_{p} \min_{x,z\in[0,1]^n} \mathcal{L}(x) + \sum_{j=1}^{n}\Big(\mu z_j + \frac{p_j}{\gamma}x_j - \frac{p_j^2}{4\gamma}z_j\Big). \tag{8}$$

Note that (8) is concave in p. Taking the derivative of (8) with respect to p_j, we obtain the optimal $p_j^* = 2x_j/z_j$, $j \in [n]$. Plugging p^* into (8), we see that it is equivalent to (3), implying that the dual is tight, i.e., $\eta_{CR} = \eta_{FR}$

For the inner minimization problem, taking the derivative with respect to z_j, we find the optimality conditions

$$z_j = \begin{cases} 0, & \mu - \frac{p_j^2}{4} > 0 \\ 1, & \mu - \frac{p_j^2}{4} < 0. \end{cases}$$

If $\mu - \frac{p_j^2}{4} = 0$, then $z_j \in [0,1]$. On the other hand, taking the derivative with respect to x_j we derive the following optimality condition:

$$\frac{p_j}{\gamma} = \sum_{i=1}^{m} \frac{y_i A_{ij}}{1+\exp(y_i A_i x)}.$$

Let x^* be the optimal solution, and, for $i \in [m]$, define

$$\alpha_i := y_i/(1+\exp(y_i A_i x^*)),\ \text{ for } i \in [m]$$

and

$$\delta_j := \frac{1}{4}(\alpha' A^j)^2,\ \text{for } j \in [n].$$

Then, $p^* = \gamma A^T \alpha$. Furthermore,

$$\mu - \frac{(p_j^*)^2}{4\gamma} = \mu - \frac{\gamma(\alpha' A^j)^2}{4} = \mu - \gamma\delta_j,$$

Using this closed form solution, we can obtain p^* for (8) from the optimal solution of (3) via α, which in turn can be used to recover $z_j^*,\ j \in [n]$.

Proof of Proposition 1. Suppose $\mu - \gamma\delta_j > 0$. Then $z_j^* = 0$ in (8), and further $\eta_{CR} - (\mu - \gamma\delta_j) < \bar{\eta}_R$. Suppose we add a constraint $z_j = 1$ to (8). Let the optimal objective value for this problem be $\eta_{FR}(z_j = 1)$. Since $\eta_{FR} + \mu - \gamma\delta_j \leq \eta_{FR}(z_j = 1)$, then if $\eta_{FR} + \mu - \gamma\delta_j > \bar{\eta}_R$, there exists no feasible solution for (3) with $z_j = 1$ that has a lower objective than $\bar{\eta}_R$. But, this implies that no optimal solution for (2) has $z_j = 1$, and thus it must be that $z_j = 0$.

The same argument is used for the case that $\mu - \gamma\delta_j < 0$ and $z_j^* = 1$ in an optimal solution to (8). Since $\eta_{FR} - (\mu - \gamma\delta_j) = \eta_{FR} - \mu + \gamma\delta_j \leq \eta_{FR}(z_j = 0)$, if $\eta_{FR} - \mu + \gamma\delta_j > \bar{\eta}_R$, then the optimal solution for (2) must have $z_j = 1$.

3.2 Derivation of Proposition 2

Using steps similar to in Section 3.1 we derive the Fenchel dual for (5):

$$\eta_{FC} = \max_p \min_{x,z \in [0,1]^n} \mathcal{L}(x) + \frac{1}{\gamma} \sum_{j=1}^{n} \left(p_j x_j - \frac{p_j^2}{4} z_j \right) \quad \text{s.t.} \sum_{i=1}^{n} z_j \leq k. \tag{9}$$

Similarly it can be shown that $p_j^* = 2x_j/z_j,\ j \in [n]$, and thus there is no duality gap and $\eta_{CC} = \eta_{FC}$. Again, taking the derivative we see that for the minimization problem, the optimal solution for (9) has $z_j = 1$ for the k most negative values of $\mu - \frac{p_j^2}{4}$ which simply translates to the z_j with the k largest values of $\frac{p_j^2}{4}$, with the rest of the indicator variables being equal to zero. In the case that there is no tie between the k-th and $(k+1)$-th most largest values, then there is a unique optimal solution for (9) which is integer in z, which is therefore the unique optimal solution for (4). Again, we can recover $p^* = \gamma A^T \alpha$, and find that $-\frac{(p_j^*)^2}{4\gamma} = -\gamma\delta_j$.

Proof for Proposition 2. Suppose $\delta_j \leq \delta_{[k+1]}$. Then $x_j = 0$ in an optimal solution for (9). Adding the constraint $z_j = 1$, one obtains a solution where the $(k-1)$ indicators with the largest values of δ are set to 1, as well as z_j, implying $z_{[k]} = 0$ by the cardinality constraint. But since $\eta_{FC} - \gamma\delta_j + \gamma\delta_{[k]} \leq \eta_{FC}(z_j = 1)$, there exists no optimal solution for (4) with $z_j = 1$ if $\eta_{FC} - \gamma(\delta_j + \delta_{[k]}) > \bar{\eta}_C$.

Using the same argument, if $\delta_j \geq \delta_{[k]}$, then $z_j = 1$ in an optimal solution for (9). Adding the constraint $z_j = 0$, we obtain a solution with $\delta_{[k+1]} = 1$ as the solution sets the indicator with the next largest δ to one. Therefore, $\eta_{FC} + \gamma\delta_j - \gamma\delta_{[k+1]} \leq \eta_{FC}(z_j = 0)$, and thus there exists no solution for (4) with $z_j = 1$ if $\eta_{FC} + \gamma(\delta_j - \delta_{[k]}) > \bar{\eta}_C$.

4 COMPUTATIONAL RESULTS

In this section, we present the computational experiments performed to test the effectiveness of the safe screening rules described in Section 3 for the $\ell_0 - \ell_2$ regularized and cardinality-constrained logistic regression problem. We test the proposed screening methods on synthetic datasets as well as on real datasets.

4.1 Experimental Setup

The real data instances of varying sizes are obtained from the UCI Machine Learning Repository (Dua et al., 2017) as well as genomics data from the Gene Expression Omnibus Database (Edgar et al., 2002).

Synthetic datasets are generated using the methodology described in Dedieu et al. (2021). Given a number of features n and a number of observations m, we generate a data matrix $A \sim \mathcal{N}_n(0, \Sigma)$, and a sparse binary vector $\tilde{x}$, representing the "true" features, which has k equi-spaced entries equal to one and the remaining entries equal to zero. For each observation $i \in [m]$, we generate a binary label y_i, where $Pr(y_i = 1|A_i) = (1 + \exp(-sA_i\tilde{x}))^{-1}$. The covariance matrix Σ controls the correlations between features, and s can be viewed as the signal-to-noise ratio. For each experimental setting, we generate ten random instances and report the average of the results for these ten instances for experiments with synthetic data.

We compare the performance of solving (REG) and (CARD) using MOSEK ApS (2021) mixed-integer conic branch-and-bound algorithm with and without screening. For consistency of the runs, we fix the solver options as follows: the branching strategy is set to pseudocost method, node selection is set to best bound method, and presolve and heuristics that add random factors to the experiments are turned off. Upper bounds used for the screening rules are obtained by simply rounding the conic relaxation solution to a nearest feasible integer solution.

4.2 Results on Synthetic Data

We first present the experimental results with screening procedure applied to the synthetic datasets. We test the regularized logistic regression (REG) with

$n = 500$, $s = 1000$, $k = 50$ as a function of the number of observations, $m \in \{200, 500, 1000\}$, the strength of the ℓ_2 regularization, $\gamma \in \{1, 1.5, 1.8\}$, and the ℓ_0 regularization, $\mu \in \{5e^{-4}, 1e^{-3}\}$. For the cardinality-constrained model (CARD), we use the same setting and vary γ in the same way while changing the ratio $k/n \in \{0.25, 0.05, 0.017\}$ by fixing $k = 50$ and varying n. In both experiments, $\Sigma = I$, which corresponds to generating features that are independent of one another.

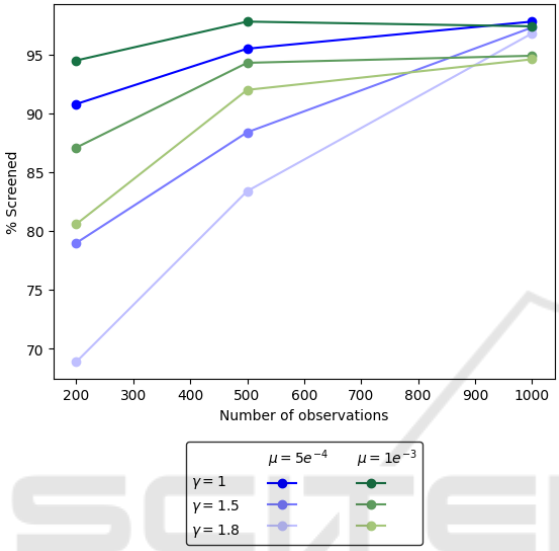

Figure 1: Percentage of features screened as a function of the number of observations in the dataset and regularization strength for (REG).

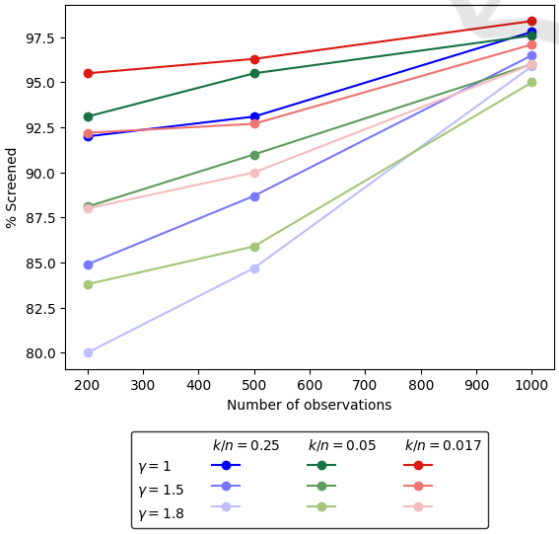

Figure 2: Percentage of features screened as a function of the number of observations in the dataset and regularization strength for (CARD).

Figures 1 and 2 show the percentage of features eliminated from the regression by the screening procedure for different regularization strengths for (REG) and (CARD), respectively. As the number of observations increases, the number of screened features increases as well. We observe the same trend as the strength of the regularization increases, i.e., higher values of μ and lower values of γ lead to better screening. The reason for improved screening with larger number of observations and stronger regularization can be explained by the smaller integrity gap of the conic relaxations, as shown in Tables 1 and 2. Integrality gap of a relaxation is the relative gap between the optimal objective value of the mixed-integer problem and the relaxation. Smaller integrality gaps lead to the satisfaction of a higher number of screening rules in Propositions 1 and 2.

Table 1: Integrality gap of big-M and conic formulations for (REG).

		Big-M relaxation			Conic relaxation		
μ	m \ γ	1	1.5	1.8	1	1.5	1.8
$5e^{-4}$	200	12.91	15.06	16.05	0.01	0.02	0.04
	500	8.43	10.18	10.97	$7e^{-3}$	0.02	0.03
	1,000	6.00	7.15	7.69	$7e^{-3}$	$9e^{-3}$	0.01
$1e^{-3}$	200	15.81	18.94	20.34	0.02	0.04	0.06
	500	9.88	12.38	13.51	0.01	0.03	0.04
	1,000	7.39	9.23	10.01	0.01	0.03	0.03
Average		10.07	12.16	13.10	0.01	0.03	0.04

Table 2: Integrality gap of big-M and conic formulations for (CARD).

		Big-M relaxation			Conic relaxation		
k/n	m \ γ	1	1.5	1.8	1	1.5	1.8
0.250	200	10.23	13.32	14.86	0.02	0.05	0.07
	500	5.27	7.12	8.08	0.01	0.03	0.04
	1,000	2.87	3.91	4.46	$4e^{-3}$	0.01	0.02
0.050	200	19.49	24.33	26.61	0.03	0.07	0.10
	500	10.48	13.82	15.51	0.01	0.03	0.05
	1,000	6.14	8.25	9.33	$6e^{-3}$	0.02	0.02
0.017	200	41.74	47.70	50.19	0.05	0.10	0.11
	500	26.58	32.73	-	0.02	0.06	-
	1,000	16.67	21.47	23.78	0.01	0.03	0.04
Average		15.50	19.18	19.10	0.02	0.04	0.06

In Tables 1 and 2, we also compare the strength of the conic formulation with the big-M formulation. Observe that the integrality gaps produced by the conic relaxation are very small, on average 0.03% for the regularized model and 0.04% for the cardinality-constrained model. On the other hand, the big-M formulation has a much weaker gap, 12% and 18% for the regularized and constrained models, respectively. The tighter gaps with the conic formulation significantly help speed up the solution time of the branch-and-bound algorithm, as well as lead to the elimination of more variables with the screening rules, further

Table 3: Solution times for the regularized logistic regression (REG) with and without screening rules.

		Time (sec.)						Speed-up		
		BnB			BnB + Screening					
μ	m \ γ	1	1.5	1.8	1	1.5	1.8	1	1.5	1.8
$5e^{-4}$	200	16	136	264	5	58	127	2.9	2.4	2.1
	500	25	69	174	6	19	57	4.3	3.7	3.2
	1,000	30	35	49	5	6	9	5.3	5.7	5.5
$1e^{-3}$	200	10	31	69	3	10	25	3.4	3.0	2.9
	500	9	29	38	2	6	8	4.2	4.7	4.8
	1,000	39	66	71	6	9	10	6.3	7.1	6.7
Average		21	61	111	5	18	39	4.4	4.4	4.2

Table 4: Solution times for the cardinality-constrained logistic regression (CARD) with and without screening rules.

		Time (sec.)						Speed-up		
		BnB			BnB + Screening					
k/n	m \ γ	1	1.5	1.8	1	1.5	1.8	1	1.5	1.8
0.250	200	16	40	69	4	11	20	4.2	3.6	3.5
	500	41	110	256	7	23	68	5.8	4.9	4.2
	1,000	30	47	52	4	7	7	6.7	7.0	7.1
0.050	200	73	200	410	12	38	92	6.2	5.5	4.6
	500	102	407	1,056	13	61	234	8.1	6.8	5.3
	1,000	159	242	287	14	23	28	10.8	10.3	10.0
0.017	200	912	2,267	1,457	92	1,313	1,703	10.2	8.0	6.4
	500	1,267	3,548	-	167	1,144	1,971	12.6	9.3	-
	1,000	1,166	1,806	2,327	57	153	368	19.9	15.3	14.4
Average		418	963	740	41	308	499	9.4	7.9	6.9

speeding up the optimization.

In order to see the impact of screening procedure on the overall solution times, we solve the logistic regression problem using the branch-and-bound algorithm with and without screening, and compare the solution times and speed-up due to screening variables. The branch-and-bound algorithm for solving the big-M formulation exceeds our time limit of 12 hours for the larger instances; therefore, we report results for the perspective formulation only. These results are shown in Tables 3 and 4. The computation time for the screening procedure is included when reporting the solution times for branch-and-bound with screening. The reported times are rounded to the nearest second. On average, we observe a 4.3× and 8.1× speed-up in computations due to the proposed screening procedure for (REG) and (CARD), respectively. The improvement in solution times increases with the number of observations. Again, a trend of increased speed-up as the strength of regularization penalty increases is seen, since more features are eliminated a priori.

4.3 Results on Real Data

In order to test the effectiveness of the proposed screening procedures on real data, we solve problems from the UCI Machine Learning Repository (Dua et al., 2017) (`arcene` and `newsgroups`) and genomic data from the Gene Expression Omnibus Database (Edgar et al., 2002) (`genomic`). In particular we focus on these larger instances of the repository with a high ratio of features to observations for which regularization is more important to avoid overfitting. We solve these instances using the regularized logistic regression model (REG), varying the strength of the regularization. As before, a time limit of 12 hours is set for each run.

The results are summarized in Table 5. For each instance, at least 92% of the features are screened, and particularly for the genomic dataset, 99.9% of the features are screened for each parameter setting. Over all instances, on average, 98% of the features are eliminated by the screening procedure before the branch-and-bound algorithm. Seven out of the 18 runs did not complete in 12 hours without screening. On the other hand, with screening, all but one run is completed within the time limit and always much faster. For the instances where branch-and-bound with and without screening both terminate within the time limit, screening leads to on average 13.8× speed-up, with larger speed-up (up to 25.6×) for the more difficult instances. For instances where only the branch-and0bound does not terminate, there is an average of

Table 5: Results for screening on real datasets using regularized logistic regression (REG).

	μ	γ	% Screened	Time (sec.) BnB	Time (sec.) BnB + Screening	Speed-up
genomic $n = 22,883$ $m = 107$	$5e^{-4}$	0.5	99.9	104	19	5.5
		1	99.9	182	17	11.0
		1.5	99.9	184	33	5.5
	$1e^{-3}$	0.5	99.9	152	14	11.0
		1	99.9	445	32	13.8
		1.5	99.9	384	54	7.1
arcene $n = 10,000$ $m = 100$	$5e^{-4}$	0.5	97	25,963	1,013	25.6
		1	97	6,999	336	20.8
		1.5	92	>12hr	10,925	>4.0
	$1e^{-3}$	0.5	99	477	32	14.8
		1	96	10,044	467	21.5
		1.5	95	22,466	1,425	15.7
newsgroups $n = 28,467$ $m = 1,977$	$5e^{-4}$	0.5	99.9	>12hr	1,135	>38.1
		0.7	99.9	>12hr	8,701	>5.0
		1	99	>12hr	>12hr	n.a
	$1e^{-3}$	0.5	99.9	>12hr	401	>107.7
		0.7	99.9	>12hr	522	>82.8
		1	99.7	>12hr	7,439	>5.8

at least 40.5× speed-up. These experimental results clearly indicate that the proposed screening rules are very effective in pruning a large number of features and result in substantial savings in computational effort for the real datasets as well.

5 CONCLUSION

In this work, we present safe screening rules for $\ell_0 - \ell_2$ regularized and cardinality-constrained logistic regression. Our numerical experiments show that a large percentage of features can be eliminated efficiently and safely via this preprocessing step before employing branch-and-bound algorithms, particularly when regularization is strong, leading to significant computational speed-up. The strength of the conic relaxations contribute significantly to the effectiveness of the screening rules in pruning a large number of features. We show the conic formulation provides much smaller integrality gaps compared to the big-*M* formulation, making it more suitable for solving ℓ_0–ℓ_2-regularized logistic regression with a branch-and-bound algorithm and also for the derived screening rules.

REFERENCES

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

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

Atamtürk, A. and Gómez, A. (2020). Safe screening rules for ℓ_0-regression from perspective relaxations. In *International Conference on Machine Learning*, pages 421–430. PMLR.

Atamtürk, A., Gómez, A., and Han, S. (2021). Sparse and smooth signal estimation: Convexification of ℓ_0-formulations. *Journal of Machine Learning Research*, 22(52):1–43.

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

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

Cawley, G. C. and Talbot, N. L. (2006). Gene selection in cancer classification using sparse logistic regression with bayesian regularization. *Bioinformatics*, 22(19):2348–2355.

Dantas, C., Soubies, E., and Févotte, C. (2021). Expanding boundaries of gap safe screening. *arXiv preprint arXiv:2102.10846*.

Dedieu, A., Hazimeh, H., and Mazumder, R. (2021). Learning sparse classifiers: Continuous and mixed integer optimization perspectives. *Journal of Machine Learning Research*, 22(135):1–47.

Dua, D., Graff, C., et al. (2017). UCI machine learning repository.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210.

El Ghaoui, L., Viallon, V., and Rabbani, T. (2010). Safe

feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*.

Fercoq, O., Gramfort, A., and Salmon, J. (2015). Mind the duality gap: safer rules for the lasso. In *International Conference on Machine Learning*, pages 333–342. PMLR.

Gramfort, A., Strohmeier, D., Haueisen, J., Hämäläinen, M. S., and Kowalski, M. (2013). Time-frequency mixed-norm estimates: Sparse m/eeg imaging with non-stationary source activations. *NeuroImage*, 70:410–422.

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

Hazimeh, H. and Mazumder, R. (2020). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5):1517–1537.

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

Kuswanto, H., Asfihani, A., Sarumaha, Y., and Ohwada, H. (2015). Logistic regression ensemble for predicting customer defection with very large sample size. *Procedia Computer Science*, 72:86–93.

Liu, J., Zhao, Z., Wang, J., and Ye, J. (2014). Safe screening with variational inequalities and its application to lasso. In *International Conference on Machine Learning*, pages 289–297. PMLR.

Miller, A. (2002). *Subset Selection in Regression*. CRC Press.

MOSEK ApS, . (2021). *MOSEK Optimizer API for Python. Release 9.3.13*.

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

Shevade, S. K. and Keerthi, S. S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253.

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

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:273–282.

Wang, J. and Park, E. (2017). Active learning for penalized logistic regression via sequential experimental design. *Neurocomputing*, 222:183–190.

Wang, J., Zhou, J., Liu, J., Wonka, P., and Ye, J. (2014). A safe screening rule for sparse logistic regression. *Advances in Neural Information Processing Systems*, 27:1053–1061.

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

Yen, S.-J., Lee, Y.-S., Ying, J.-C., and Wu, Y.-C. (2011). A logistic regression-based smoothing method for chinese text categorization. *Expert Systems with Applications*, 38(9):11581–11590.

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