Safe Screening with Variational Inequalities and Its Application to Lasso

Jun Liu, Zheng Zhao {JUN.LIU,ZHENG.ZHAO}@SAS.COM
SAS Institute Inc., Cary, NC 27513

Jie Wang, Jieping Ye {JIE.WANG.USTC,JIEPING.YE}@ASU.EDU
Arizona State University, Tempe, AZ 85287

Abstract

Sparse learning techniques have been routinely used for feature selection as the resulting model usually has a small number of non-zero entries. Safe screening, which eliminates the features that are guaranteed to have zero coefficients for a certain value of the regularization parameter, is a technique for improving the computational efficiency. Safe screening is gaining increasing attention since 1) solving sparse learning formulations usually has a high computational cost especially when the number of features is large and 2) one needs to try several regularization parameters to select a suitable model. In this paper, we propose an approach called "Sasvi" (Safe screening with variational inequalities). Sasvi makes use of the variational inequality that provides the sufficient and necessary optimality condition for the dual problem. Several existing approaches for Lasso screening can be casted as relaxed versions of the proposed Sasvi, thus Sasvi provides a stronger safe screening rule. We further study the monotone properties of Sasvi for Lasso, based on which a sure removal regularization parameter can be identified for each feature. Experimental results on both synthetic and real data sets are reported to demonstrate the effectiveness of the proposed Sasvi for Lasso screening.

1. Introduction

Sparse learning (Candes & Wakin, 2008; Tibshirani, 1996) is an effective technique for analyzing high dimensional data. It has been applied successfully in various areas, such as machine learning, signal processing, image processing, medical imaging, and so on. In general, the ℓ_1-regularized sparse learning can be formulated as:

$$\min_{\boldsymbol{\beta}} \quad \text{loss}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ contains the model coefficients, $\text{loss}(\boldsymbol{\beta})$ is a loss function defined on the design matrix $X \in \mathbb{R}^{n \times p}$ and the response $\mathbf{y} \in \mathbb{R}^n$, and λ is a positive regularization parameter that balances the tradeoff between the loss function and the ℓ_1 regularization. Let $\mathbf{x}^i \in \mathbb{R}^p$ denote the i-th sample that corresponds to the transpose of the i-th row of X, and let $\mathbf{x}_j \in \mathbb{R}^n$ denote the j-th feature that corresponds to the j-th column of X. We use $\text{loss}(\boldsymbol{\beta}) = \frac{1}{2}\|X\boldsymbol{\beta} - \mathbf{y}\|_2^2 = \frac{1}{2}\sum_{i=1}^n (y_i - \boldsymbol{\beta}^T \mathbf{x}^i)^2$ in Lasso (Tibshirani, 1996) and $\text{loss}(\boldsymbol{\beta}) = \sum_{i=1}^n \log(1 + \exp(-y_i \boldsymbol{\beta}^T \mathbf{x}^i))$ in sparse logistic regression (Koh et al., 2007).

Since the optimal λ is usually unknown in practical applications, we need to solve formulation (1) corresponding to a series of regularization parameter $\lambda_1 > \lambda_2 > \ldots > \lambda_k$, obtain the solutions $\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \ldots, \boldsymbol{\beta}_k^*$, and then select the solution that is optimal in terms of a pre-specified criterion, e.g., Schwarz Bayesian information criterion (Schwarz, 1978) and cross-validation. The well-known LARS approach (Efron et al., 2004) can be modified to obtain the full piecewise linear Lasso solution path. Other approaches such as interior point (Koh et al., 2007), coordinate descent (Friedman et al., 2010) and accelerated gradient descent (Nesterov, 2004) usually solve formulation (1) corresponding to a series of pre-defined parameters.

The solutions $\boldsymbol{\beta}_k^*, k = 1, 2, \ldots,$ are sparse in that many of their coefficients are zero. Taking advantage of the nature of sparsity, the screening techniques have been proposed for accelerating the computation. Specifically, given a solution $\boldsymbol{\beta}_1^*$ at the regularization parameter λ_1, if we can identify the features that are guaranteed to have zero coefficients in $\boldsymbol{\beta}_2^*$ at the regularization parameter λ_2, then the cost for computing $\boldsymbol{\beta}_2^*$ can be saved by excluding those inactive features. There are two categories of screening techniques: 1) the safe screening techniques (Ghaoui et al., 2012; Wang et al., 2013; Ogawa et al., 2013; Zhen et al., 2011) with which our obtained solution is exactly the same as the one obtained by directly solving (1), and 2) the heuristic rule such as the strong rules (Tibshirani et al.,

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

2012) which can eliminate more features but might mistakenly discard active features.

In this paper, we propose an approach called "Sasvi" (Safe screening with variational inequalities) and take Lasso as an example in the analysis. Sasvi makes use of the variational inequality which provides the sufficient and necessary optimality condition for the dual problem. Several existing approaches such as SAFE (Ghaoui et al., 2012) and DPP (Wang et al., 2013) can be casted as relaxed versions of the proposed Sasvi, thus Sasvi provides a stronger screening rule. The monotone properties of Sasvi for Lasso are studied based on which a sure removal regularization parameter can be identified for each feature. Empirical results on both synthetic and real data sets demonstrate the effectiveness of the proposed Sasvi for Lasso screening. Extension of the proposed Sasvi to the generalized sparse linear models such as logistic regression is briefly discussed.

Notations Throughout this paper, scalars are denoted by italic letters, and vectors by bold face letters. Let $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$ denote the ℓ_1 norm, the Euclidean norm, and the infinity norm, respectively. Let $\langle \mathbf{x}, \mathbf{y} \rangle$ denote the inner product between $\mathbf{x}$ and $\mathbf{y}$.

2. The Proposed Sasvi

Our proposed approach builds upon an analysis on the following simple problem:

$$\min_{\beta} \{-\beta b + |\beta|\}. \tag{2}$$

We have the following results:
1) If $|b| \leq 1$, then the minimum of (2) is 0;
2) If $|b| > 1$, then the minimum of (2) is $-\infty$; and
3) If $|b| < 1$, then the optimal solution $\beta^* = 0$.

The dual problem usually can provide a good insight about the problem to be solved. Let $\boldsymbol{\theta}$ denote the dual variable of Eq. (1). In light of Eq. (2), we can show that β_j^*, the j-th component of the optimal solution to Eq. (1), optimizes

$$\min_{\beta_j} \{-\beta_j \langle \mathbf{x}_j, \boldsymbol{\theta}^* \rangle + |\beta_j|\}, \tag{3}$$

where $\mathbf{x}_j$ denotes the j-th feature and $\boldsymbol{\theta}^*$ denotes the optimal dual variable of Eq. (1). From the results to Eq. (2), we need $|\langle \mathbf{x}_j, \boldsymbol{\theta}^* \rangle| \leq 1$ to ensure that Eq. (3) does not equal to $-\infty$[1], and we have

$$|\langle \mathbf{x}_j, \boldsymbol{\theta}^* \rangle| < 1 \Rightarrow \beta_j^* = 0. \tag{4}$$

Eq. (4) says that, the j-th feature can be safely eliminated in the computation of $\boldsymbol{\beta}^*$ if $|\langle \mathbf{x}_j, \boldsymbol{\theta}^* \rangle| < 1$.

Let λ_1 and λ_2 be two distinct regularization parameters that satisfy

$$\lambda_{\max} \geq \lambda_1 > \lambda_2 > 0, \tag{5}$$

[1]This is used in deriving the last equality of Eq. (6).

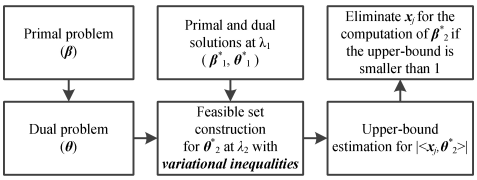

Figure 1. The work flow of the proposed Sasvi. The purpose is to discard the features that can be safely eliminated in computing $\boldsymbol{\beta}_2^*$ with the information obtained at λ_1.

where $\lambda_{\max}$ denotes the value of λ above which the solution to Eq. (1) is zero. Let $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ be the optimal primal variables corresponding to λ_1 and λ_2, respectively. Let $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2^*$ be the optimal dual variables corresponding to λ_1 and λ_2, respectively. Figure 1 illustrates the work flow of the proposed Sasvi. We firstly derive the dual problem of Eq. (1). Suppose that we have obtained the primal and dual solutions $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\theta}_1^*$ for a given regularization parameter λ_1, and we are interested in solving Eq. (1) with $\lambda = \lambda_2$ by using Eq. (4) to screen the features to save computational cost. However, the difficulty lies in that, we do not have the dual optimal $\boldsymbol{\theta}_2^*$. To deal with this, we construct a feasible set for $\boldsymbol{\theta}_2^*$, estimate an upper-bound of $|\langle \mathbf{x}_j, \boldsymbol{\theta}_2^* \rangle|$, and safely remove $\mathbf{x}_j$ if this upper-bound is smaller than 1.

The construction of a tight feasible set for $\boldsymbol{\theta}_2^*$ is key to the success of the screening technique. If the constructed feasible set is too loose, the estimated upper-bound of $|\langle \mathbf{x}_j, \boldsymbol{\theta}_2^* \rangle|$ is over 1, and thus only a few features can be discarded. In this paper, we propose to construct the feasible set by using the variational inequalities that provide the sufficient and necessary optimality conditions for the dual problems with $\lambda = \lambda_1$ and λ_2. Then, we estimate the upper-bound of $|\langle \mathbf{x}_j, \boldsymbol{\theta}_2^* \rangle|$ in the constructed feasbile set, and discard the j-th feature if the upper-bound is smaller than 1. For discussion convenience, we focus on Lasso in this paper, but the underlying methodology can be extended to the general problem in Eq. (1). Next, we elaborate the three building blocks that are illustrated in the bottom row of Figure 1.

2.1. The Dual Problem of Lasso

We follow the discussion in Section 6 of (Nesterov, 2013) in deriving the dual problem of Lasso as follows:

$$\begin{aligned}
&\min_{\boldsymbol{\beta}} \left[\frac{1}{2} \|X\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right] \\
&= \min_{\boldsymbol{\beta}} \max_{\boldsymbol{\theta}} \left[\langle \mathbf{y} - X\boldsymbol{\beta}, \lambda\boldsymbol{\theta} \rangle - \frac{1}{2}\|\lambda\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right] \\
&= \max_{\boldsymbol{\theta}} \min_{\boldsymbol{\beta}} \lambda \left[\langle \mathbf{y}, \boldsymbol{\theta} \rangle - \frac{\lambda\|\boldsymbol{\theta}\|_2^2}{2} - \langle X^T\boldsymbol{\theta}, \boldsymbol{\beta} \rangle + \|\boldsymbol{\beta}\|_1 \right] \\
&= \max_{\boldsymbol{\theta}: \|X^T\boldsymbol{\theta}\|_\infty \leq 1} \lambda^2 \left[-\frac{1}{2} \left\| \boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda} \right\|_2^2 + \frac{1}{2} \left\| \frac{\mathbf{y}}{\lambda} \right\|_2^2 \right].
\end{aligned} \tag{6}$$

A dual variable $\boldsymbol{\theta}$ is introduced in the first equality, and the equivalence can be verified by setting the derivative with regard to $\boldsymbol{\theta}$ to zero, which leads to the following relationship between the optimal primal variable ($\boldsymbol{\beta}^*$) and the optimal dual variable ($\boldsymbol{\theta}^*$):

$$\lambda\boldsymbol{\theta}^* = \mathbf{y} - X\boldsymbol{\beta}^*. \tag{7}$$

In obtaining the last equality of Eq. (6), we make use of the results to Eq. (2).

The dual problem of Eq. (1) can be formulated as:

$$\min_{\boldsymbol{\theta}:\|X^T\boldsymbol{\theta}\|_\infty \leq 1} \frac{1}{2}\left\|\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda}\right\|_2^2. \tag{8}$$

For Lasso, the $\lambda_{\max}$ in Eq. (5) can be analytically computed as $\lambda_{\max} = \|X^T\mathbf{y}\|_\infty$. In applying Sasvi, we might start with $\lambda_1 = \lambda_{\max}$, since the primal and dual optimals can be computed analytically as: $\boldsymbol{\beta}_1^* = \mathbf{0}$ and $\boldsymbol{\theta}_1^* = \frac{\mathbf{y}}{\lambda_{\max}}$.

2.2. Feasible Set Construction

Given λ_1, $\boldsymbol{\theta}_1^*$ and λ_2, we aim at estimating the upper-bound of $|\langle \mathbf{x}_j, \boldsymbol{\theta}_2^*\rangle|$ without the actual computation of $\boldsymbol{\theta}_2^*$. To this end, we construct a feasible set for $\boldsymbol{\theta}_2^*$, and then estimate the upper-bound in the constructed feasible set. To construct the feasible set, we make use of the variational inequality that provides the sufficient and necessary condition of a constrained convex optimization problem.

Lemma 1 *(Nesterov, 2004) For the constrained convex optimization problem:*

$$\min_{\mathbf{x}\in G} f(\mathbf{x}), \tag{9}$$

with G being convex and closed and $f(\cdot)$ being convex and differentiable, $\mathbf{x}^* \in G$ is an optimal solution of Eq. (9) if and only if

$$\langle f'(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle \geq 0, \forall \mathbf{x} \in G. \tag{10}$$

Eq. (10) is the so-called variation inequality for the problem in Eq. (9). Applying Lemma 1 to the Lasso dual problem in Eq. (8), we can represent the optimality conditions for $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2^*$ using the following two variational inequalities:

$$\left\langle \boldsymbol{\theta}_1^* - \frac{\mathbf{y}}{\lambda_1}, \boldsymbol{\theta} - \boldsymbol{\theta}_1^* \right\rangle \geq 0, \forall \boldsymbol{\theta} : \|X^T\boldsymbol{\theta}\|_\infty \leq 1, \tag{11}$$

$$\left\langle \boldsymbol{\theta}_2^* - \frac{\mathbf{y}}{\lambda_2}, \boldsymbol{\theta} - \boldsymbol{\theta}_2^* \right\rangle \geq 0, \forall \boldsymbol{\theta} : \|X^T\boldsymbol{\theta}\|_\infty \leq 1. \tag{12}$$

Plugging $\boldsymbol{\theta} = \boldsymbol{\theta}_2^*$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1^*$ into Eq. (11) and Eq. (12) respectively, we have

$$\left\langle \boldsymbol{\theta}_1^* - \frac{\mathbf{y}}{\lambda_1}, \boldsymbol{\theta}_2^* - \boldsymbol{\theta}_1^* \right\rangle \geq 0, \tag{13}$$

$$\left\langle \boldsymbol{\theta}_2^* - \frac{\mathbf{y}}{\lambda_2}, \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_2^* \right\rangle \geq 0. \tag{14}$$

With Eq. (13) and Eq. (14), we can construct the following feasible set for $\boldsymbol{\theta}_2^*$ as:

$$\Omega(\boldsymbol{\theta}_2^*) = \{\boldsymbol{\theta} : \langle \boldsymbol{\theta}_1^* - \frac{\mathbf{y}}{\lambda_1}, \boldsymbol{\theta} - \boldsymbol{\theta}_1^*\rangle \geq 0, \langle \boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda_2}, \boldsymbol{\theta}_1^* - \boldsymbol{\theta}\rangle \geq 0\}. \tag{15}$$

For an illustration of the feasible set, please refer to Figure 2. Generally speaking, the closer λ_2 is to λ_1, the tighter the feasible set for $\boldsymbol{\theta}_2^*$ is. In fact, when λ_2 approaches to λ_1, $\Omega(\boldsymbol{\theta}_2^*)$ concentrates to a singleton set that only contains $\boldsymbol{\theta}_2^*$. Note that one may use additional $\boldsymbol{\theta}$'s in Eq. (12) for improving the estimation of the feasible set of $\boldsymbol{\theta}_2^*$. Next, we discuss how to make use of the feasible set defined in Eq. (15) for estimating an upper-bound for $|\langle \mathbf{x}_j, \boldsymbol{\theta}_2^*\rangle|$.

2.3. Upper-bound Estimation

Since $\boldsymbol{\theta}_2^* \in \Omega(\boldsymbol{\theta}_2^*)$, we can estimate an upper-bound of $|\langle \mathbf{x}_j, \boldsymbol{\theta}_2^*\rangle|$ by solving

$$\max_{\boldsymbol{\theta}\in\Omega(\boldsymbol{\theta}_2^*)} |\langle \mathbf{x}_j, \boldsymbol{\theta}\rangle|. \tag{16}$$

Next, we show how to solve Eq. (16). For discussion convenience, we introduce the following three variables:

$$\begin{aligned} \mathbf{a} &= \frac{\mathbf{y}}{\lambda_1} - \boldsymbol{\theta}_1^* = \frac{X\boldsymbol{\beta}_1^*}{\lambda_1}, \\ \mathbf{b} &= \frac{\mathbf{y}}{\lambda_2} - \boldsymbol{\theta}_1^* = \mathbf{a} + (\frac{\mathbf{y}}{\lambda_2} - \frac{\mathbf{y}}{\lambda_1}), \\ \mathbf{r} &= 2\boldsymbol{\theta} - (\boldsymbol{\theta}_1^* + \frac{\mathbf{y}}{\lambda_2}), \end{aligned} \tag{17}$$

where $\mathbf{a}$ denotes the prediction based on $\boldsymbol{\beta}_1^*$ scaled by $\frac{1}{\lambda_1}$, and $\mathbf{b}$ is the summation of $\mathbf{a}$ and the change of the inputs to the dual problem in Eq. (8) from λ_1 to λ_2.

Figure 2 illustrates $\mathbf{a}$ and $\mathbf{b}$ by lines EB and EC, respectively. For the triangle EBC, the following theorem shows that the angle between $\mathbf{a}$ and $\mathbf{b}$ is acute.

Theorem 1 *Let $\mathbf{y} \neq \mathbf{0}$, and $\|X^T\mathbf{y}\|_\infty \geq \lambda_1 > \lambda_2 > 0$. We have*

$$\mathbf{b} \neq \mathbf{0}, \langle \mathbf{b}, \mathbf{a}\rangle \geq 0, \tag{18}$$

and $\langle \mathbf{b}, \mathbf{a}\rangle = 0$ if and only if $\lambda_1 = \|X^T\mathbf{y}\|_\infty$. In addition, if $\lambda_1 < \|X^T\mathbf{y}\|_\infty$, then $\mathbf{a} \neq \mathbf{0}$.

The proof of Theorem 1 is given in Supplement A. With the notations in Eq. (17), Eq. (16) can be rewritten as

$$\begin{aligned} \max_{\mathbf{r}} \quad & \frac{1}{2}\left|\left\langle \mathbf{x}_j, \boldsymbol{\theta}_1^* + \frac{\mathbf{y}}{\lambda_2}\right\rangle + \langle \mathbf{x}_j, \mathbf{r}\rangle\right| \\ \text{subject to} \quad & \langle \mathbf{a}, \mathbf{r} + \mathbf{b}\rangle \leq 0, \|\mathbf{r}\|_2^2 \leq \|\mathbf{b}\|_2^2. \end{aligned} \tag{19}$$

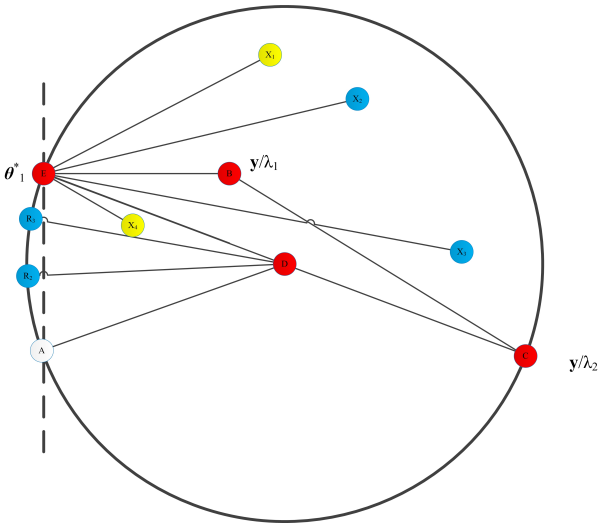

Figure 2. Illustration of the feasible set used in Sasvi and Theorem 3. The points in the figure are explained as follows. E: $\boldsymbol{\theta}^*_1$, B: $\frac{\mathbf{y}}{\lambda_1}$, C: $\frac{\mathbf{y}}{\lambda_2}$, D: $\frac{\boldsymbol{\theta}^*_1+\frac{\mathbf{y}}{\lambda_2}}{2}$. The left hand side of the dash line represents the half space $\{\boldsymbol{\theta} : \langle \boldsymbol{\theta}^*_1 - \frac{\mathbf{y}}{\lambda_1}, \boldsymbol{\theta} - \boldsymbol{\theta}^*_1 \rangle \geq 0\}$, and the ball centered at D with radius ED represents $\{\boldsymbol{\theta} : \langle \boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda_2}, \boldsymbol{\theta}^*_1 - \boldsymbol{\theta} \rangle \geq 0\}$. For Theorem 3, EX_1, EX_2, EX_3 and EX_4 denote $\pm\mathbf{x}_j$ in two subcases: 1) the angle between EB and EX_1 (EX_4) is larger than the angle between EB and EC, and 2) the angle between EB and EX_2 (EX_3) is smaller than the angle between EB and EC. R_2 (R_3) is the maximizer to Eq. (16) with EX_2 (EX_3) denoting $\pm\mathbf{x}_j$. With EX_1 (EX_4) denoting $\pm\mathbf{x}_j$, the maximizer to Eq. (16) is on the intersection between the dashed line and the ball centered at D with radius ED.

The objective function of Eq. (19) can be represented by half of the following form:

$$\max\left(\langle \mathbf{x}_j, \boldsymbol{\theta}^*_1 + \frac{\mathbf{y}}{\lambda_2}\rangle + \langle \mathbf{x}_j, \mathbf{r}\rangle, -\langle \mathbf{x}_j, \boldsymbol{\theta}^*_1 + \frac{\mathbf{y}}{\lambda_2}\rangle - \langle \mathbf{x}_j, \mathbf{r}\rangle\right)$$

which indicates that Eq. (19) can be computed by maximizing $\langle \mathbf{x}_j, \mathbf{r}\rangle$ and $-\langle \mathbf{x}_j, \mathbf{r}\rangle$ over the feasible set in the same equation. Maximizing $\langle \mathbf{x}_j, \mathbf{r}\rangle$ and $-\langle \mathbf{x}_j, \mathbf{r}\rangle$ can be computed by minimizing $\langle -\mathbf{x}_j, \mathbf{r}\rangle$ and $\langle \mathbf{x}_j, \mathbf{r}\rangle$, which can be solved by the following minimization problem:

$$\begin{aligned} \min_{\mathbf{r}} \quad & \langle \mathbf{x}, \mathbf{r}\rangle \\ \text{subject to} \quad & \langle \mathbf{a}, \mathbf{r}+\mathbf{b}\rangle \leq 0, \|\mathbf{r}\|_2^2 \leq \|\mathbf{b}\|_2^2. \end{aligned} \tag{20}$$

We assume that $\mathbf{x}$ is a non-zero vector. Let

$$\mathbf{x}^\perp = \mathbf{x} - \mathbf{a}\langle \mathbf{x}, \mathbf{a}\rangle / \|\mathbf{a}\|_2^2, \tag{21}$$

$$\mathbf{x}_j^\perp = \mathbf{x}_j - \mathbf{a}\langle \mathbf{x}_j, \mathbf{a}\rangle / \|\mathbf{a}\|_2^2, \tag{22}$$

$$\mathbf{y}^\perp = \mathbf{y} - \mathbf{a}\langle \mathbf{y}, \mathbf{a}\rangle / \|\mathbf{a}\|_2^2, \tag{23}$$

which are the orthogonal projections of $\mathbf{x}$, $\mathbf{x}_j$, and $\mathbf{y}$ onto the null space of $\mathbf{a}$, respectively. Our next theorem says that Eq. (20) admits a closed form solution.

Theorem 2 *Let* $0 < \lambda_1 \leq \|X^T\mathbf{y}\|_\infty$, $0 < \lambda_2 < \lambda_1$, $\mathbf{x} \neq \mathbf{0}$ *and* $\mathbf{y} \neq \mathbf{0}$. *Eq. (20) equals to* $-\|\mathbf{x}\|_2\|\mathbf{b}\|_2$, *if* $\frac{\langle \mathbf{b},\mathbf{a}\rangle}{\|\mathbf{b}\|_2} \leq \frac{\langle \mathbf{x},\mathbf{a}\rangle}{\|\mathbf{x}\|_2}$, *and* $-\|\mathbf{x}^\perp\|_2\sqrt{\|\mathbf{b}\|_2^2 - \frac{\langle \mathbf{b},\mathbf{a}\rangle^2}{\|\mathbf{a}\|_2^2}} - \frac{\langle \mathbf{a},\mathbf{b}\rangle\langle \mathbf{x},\mathbf{a}\rangle}{\|\mathbf{a}\|_2^2}$ *otherwise.*

The proof of Theorem 2 is given in Supplement B. With Theorem 2, we can obtain the upper-bound of $|\langle \mathbf{x}_j, \boldsymbol{\theta}^*_2\rangle|$ in the following theorem.

Theorem 3 *Let* $\mathbf{y} \neq 0$, *and* $\|X^T\mathbf{y}\|_\infty \geq \lambda_1 > \lambda_2 > 0$. *Denote*

$$u_j^+(\lambda_2) = \max_{\boldsymbol{\theta} \in \Omega(\boldsymbol{\theta}^*_2)} \langle \mathbf{x}_j, \boldsymbol{\theta}\rangle, \tag{24}$$

$$u_j^-(\lambda_2) = \max_{\boldsymbol{\theta} \in \Omega(\boldsymbol{\theta}^*_2)} \langle -\mathbf{x}_j, \boldsymbol{\theta}\rangle. \tag{25}$$

We have:

1) If $\mathbf{a} \neq \mathbf{0}$ *and* $\frac{\langle \mathbf{b},\mathbf{a}\rangle}{\|\mathbf{b}\|_2\|\mathbf{a}\|_2} > \frac{|\langle \mathbf{x}_j,\mathbf{a}\rangle|}{\|\mathbf{x}_j\|_2\|\mathbf{a}\|_2}$ *then*

$$u_j^+(\lambda_2) = \langle \mathbf{x}_j, \boldsymbol{\theta}^*_1\rangle + \frac{\frac{1}{\lambda_2} - \frac{1}{\lambda_1}}{2}\left[\|\mathbf{x}_j^\perp\|_2\|\mathbf{y}^\perp\|_2 + \langle \mathbf{x}_j^\perp, \mathbf{y}^\perp\rangle\right], \tag{26}$$

$$u_j^-(\lambda_2) = -\langle \mathbf{x}_j, \boldsymbol{\theta}^*_1\rangle + \frac{\frac{1}{\lambda_2} - \frac{1}{\lambda_1}}{2}\left[\|\mathbf{x}_j^\perp\|_2\|\mathbf{y}^\perp\|_2 - \langle \mathbf{x}_j^\perp, \mathbf{y}^\perp\rangle\right]. \tag{27}$$

2) If $\langle \mathbf{x}_j, \mathbf{a}\rangle > 0$ *and* $\frac{\langle \mathbf{b},\mathbf{a}\rangle}{\|\mathbf{b}\|_2\|\mathbf{a}\|_2} \leq \frac{\langle \mathbf{x}_j,\mathbf{a}\rangle}{\|\mathbf{x}_j\|_2\|\mathbf{a}\|_2}$, *then* $u_j^+(\lambda_2)$ *satisfies Eq. (26), and*

$$u_j^-(\lambda_2) = -\langle \mathbf{x}_j, \boldsymbol{\theta}^*_1\rangle + \frac{1}{2}\left[\|\mathbf{x}_j\|_2\|\mathbf{b}\|_2 - \langle \mathbf{x}_j, \mathbf{b}\rangle\right]. \tag{28}$$

3) If $\langle \mathbf{x}_j, \mathbf{a}\rangle < 0$ *and* $\frac{\langle \mathbf{b},\mathbf{a}\rangle}{\|\mathbf{b}\|_2\|\mathbf{a}\|_2} \leq \frac{-\langle \mathbf{x}_j,\mathbf{a}\rangle}{\|\mathbf{x}_j\|_2\|\mathbf{a}\|_2}$, *then*

$$u_j^+(\lambda_2) = \langle \mathbf{x}_j, \boldsymbol{\theta}^*_1\rangle + \frac{1}{2}\left[\|\mathbf{x}_j\|_2\|\mathbf{b}\|_2 + \langle \mathbf{x}_j, \mathbf{b}\rangle\right]. \tag{29}$$

and $u_j^-(\lambda_2)$ *satisfies Eq. (27).*

4) If $\mathbf{a} = \mathbf{0}$, *then Eq. (28) and Eq. (29) hold.*

The proof of Theorem 3 is given in Supplement C. An illustration of Theorem 3 for different cases can be found in Figure 2. It follows from Eq. (4) that, if $u_j^+(\lambda_2) < 1$ and $u_j^-(\lambda_2) < 1$, then the j-th feature can be safely eliminated for the computation of $\boldsymbol{\beta}^*_2$. We provide the following analysis to the established upper-bound. Firstly, we have

$$\lim_{\lambda_2 \to \lambda_1} u_j^+(\lambda_2) = \langle \mathbf{x}_j, \boldsymbol{\theta}^*_1\rangle, \lim_{\lambda_2 \to \lambda_1} u_j^-(\lambda_2) = -\langle \mathbf{x}_j, \boldsymbol{\theta}^*_1\rangle,$$

which attributes to the fact that $\lim_{\lambda_2 \to \lambda_1} \Omega(\boldsymbol{\theta}^*_2) = \{\boldsymbol{\theta}^*_1\}$. Secondly, in the extreme case that $\mathbf{x}_j$ is orthogonal to the scaled prediction $\mathbf{a} = \frac{X\boldsymbol{\beta}^*_1}{\lambda_1}$ which is nonzero, Theorem 3 leads to $\mathbf{x}_j^\perp = \mathbf{0}$, $u_j^+(\lambda_2) = \langle \mathbf{x}_j, \boldsymbol{\theta}^*_1\rangle$ and $u_j^-(\lambda_2) = -\langle \mathbf{x}_j, \boldsymbol{\theta}^*_1\rangle$. Thus, the j-th feature can be safely removed for any positive λ_2 that is smaller than λ_1 so long as

$|\langle \mathbf{x}_j, \boldsymbol{\theta}_1^* \rangle| < 1$. Thirdly, in the case that $\mathbf{x}_j$ has low correlation with the prediction $\mathbf{a} = \frac{X\boldsymbol{\beta}_1^*}{\lambda_1}$, Theorem 3 indicates that the j-th feature is very likely to be safely removed for a wide range of λ_2 if $|\langle \mathbf{x}_j, \boldsymbol{\theta}_1^* \rangle| < 1$. The monotone properties of the upper-bound established in Theorem 3 is given Section 4.

3. Comparison with Existing Approaches

Our proposed Sasvi differs from the existing screening techniques (Ghaoui et al., 2012; Tibshirani et al., 2012; Wang et al., 2013; Zhen et al., 2011) in the construction of the feasible set for $\boldsymbol{\theta}_2^*$.

3.1. Comparison with the Strong Rule

The strong rule (Tibshirani et al., 2012) works on $0 < \lambda_2 < \lambda_1$ and makes use of the assumption

$$|\lambda_2 \langle \mathbf{x}_j, \boldsymbol{\theta}_2^* \rangle - \lambda_1 \langle \mathbf{x}_j, \boldsymbol{\theta}_1^* \rangle| \leq |\lambda_2 - \lambda_1|, \tag{30}$$

from which we can obtain an estimated upper-bound for $|\langle \mathbf{x}_j, \boldsymbol{\theta}_2^* \rangle|$ as:

$$\begin{aligned} |\langle \mathbf{x}_j, \boldsymbol{\theta}_2^* \rangle| &\leq \frac{|\lambda_1 \langle \mathbf{x}_j, \boldsymbol{\theta}_1^* \rangle| + |\lambda_2 \langle \mathbf{x}_j, \boldsymbol{\theta}_2^* \rangle - \lambda_1 \langle \mathbf{x}_j, \boldsymbol{\theta}_1^* \rangle|}{\lambda_2} \\ &\leq \frac{|\lambda_1 \langle \mathbf{x}_j, \boldsymbol{\theta}_1^* \rangle| + (\lambda_1 - \lambda_2)}{\lambda_2} \\ &= \frac{\lambda_1}{\lambda_2} |\langle \mathbf{x}_j, \boldsymbol{\theta}_1^* \rangle| + \left[\frac{\lambda_1}{\lambda_2} - 1 \right] \end{aligned} \tag{31}$$

A comparison between Eq. (31) and the upper-bound established in Theorem 3 shows that, 1) both are dependent on $\langle \mathbf{x}_j, \boldsymbol{\theta}_1^* \rangle$, the inner product between the j-th feature and the dual variable $\boldsymbol{\theta}_1^*$ obtained at λ_1, but note that $\frac{\lambda_1}{\lambda_2} > 1$, 2) in comparison with the data independent term $\frac{\lambda_1}{\lambda_2} - 1$ used in the strong rule, Sasvi utilizes a data dependent term as shown in Eqs. (26)-(29). We note that, 1) when a feature $\mathbf{x}_j$ has low correlation with the prediction $\mathbf{a} = \frac{X\boldsymbol{\beta}_1^*}{\lambda_1}$, the upper-bound for $|\langle \mathbf{x}_j, \boldsymbol{\theta}_2^* \rangle|$ estimated by Sasvi might be lower than the one by the strong rule [2], and 2) as pointed out in (Tibshirani et al., 2012), Eq. (30) might not always hold, and the same applies to Eq. (31).

Next, we compare Sasvi with the SAFE approach (Ghaoui et al., 2012) and the DPP approach (Wang et al., 2013), and the differences in terms of the feasible sets are shown in Figure 3.

[2] According to the analysis given at the end of Section 2.3, this argument is true for the extreme case that $\mathbf{x}_j$ is orthogonal to the nonzero prediction $\mathbf{a} = \frac{X\boldsymbol{\beta}_1^*}{\lambda_1}$.

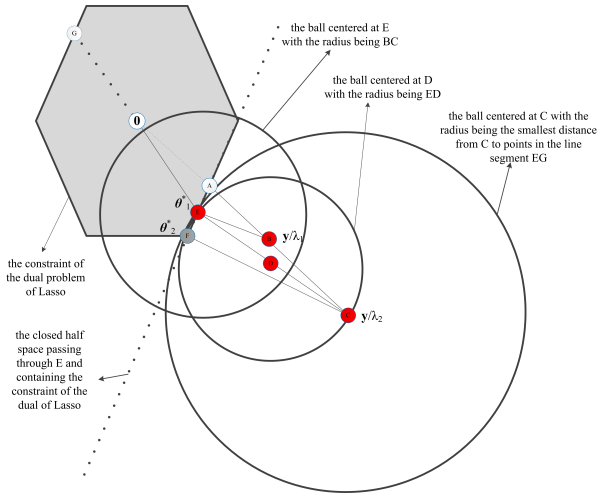

Figure 3. Comparison of Sasvi with existing safe screening approaches. The points in the figure are as follows. A: $\frac{\mathbf{y}}{\lambda_{\max}}$, B: $\frac{\mathbf{y}}{\lambda_1}$, C: $\frac{\mathbf{y}}{\lambda_2}$, D: the middle point of C and E, E: $\boldsymbol{\theta}_1^*$, F: $\boldsymbol{\theta}_2^*$, and G: $-\boldsymbol{\theta}_1^*$. The feasible set for $\boldsymbol{\theta}_2^*$ used by the proposed Sasvi approach is the intersection between the ball centered at D with radius being half EC and the closed half space passing through E and containing the constraint of the dual of Lasso. The feasible set for $\boldsymbol{\theta}_2^*$ used by the SAFE (Ghaoui et al., 2012) approach is the ball centered at C with radius being the smallest distance from C to the points in the line segment EG. The feasible set for $\boldsymbol{\theta}_2^*$ used by the DPP (Wang et al., 2013) approach is the ball centered at E with radius BC.

3.2. Comparison with the SAFE approach

Denote $G(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\|\lambda_2\boldsymbol{\theta} - \mathbf{y}\|_2^2$. The SAFE approach makes use of the so-called "dual" scaling, and compute the upper-bound of the $G(\boldsymbol{\theta})$ for λ_2 as

$$\gamma(\lambda_2) = \max_{s:|s|\leq 1} G(s\boldsymbol{\theta}) = \max_{s:|s|\leq 1} \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{1}{2}\|s\lambda_2\boldsymbol{\theta}_1^* - \mathbf{y}\|_2^2, \tag{32}$$

Note that, compared to the SAFE paper, the dual variable $\boldsymbol{\theta}$ has been scaled in the formulation in Eq. (32), but this scaling does not influence of the following result for the SAFE approach. Denote s^* as the optimal solution. Solving Eq. (32), we have $s^* = \max\left(\min\left(\frac{\langle \boldsymbol{\theta}_1^*, \mathbf{y} \rangle}{\lambda_2 \|\boldsymbol{\theta}_1^*\|_2}, 1\right), -1\right)$ when $\boldsymbol{\theta}_1 \neq \mathbf{0}$. The SAFE approach computes the upper-bound for $|\langle \mathbf{x}_j, \boldsymbol{\theta}_2^* \rangle|$ as follows:

$$\begin{aligned} |\langle \mathbf{x}_j, \boldsymbol{\theta}_2^* \rangle| &\leq \max_{\boldsymbol{\theta}: G(\boldsymbol{\theta}) \geq \gamma(\lambda_2)} |\langle \mathbf{x}_j, \boldsymbol{\theta} \rangle| \\ &= \max_{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda_2}\|_2 \leq \|s^*\boldsymbol{\theta}_1^* - \frac{\mathbf{y}}{\lambda_2}\|_2} |\langle \mathbf{x}_j, \boldsymbol{\theta} \rangle| \\ &= \frac{|\langle \mathbf{x}_j, \mathbf{y} \rangle|}{\lambda_2} + \|\mathbf{x}_j\|_2 \left\| s^*\boldsymbol{\theta}_1^* - \frac{\mathbf{y}}{\lambda_2} \right\|_2. \end{aligned} \tag{33}$$

Next, we show that the feasible set for $\boldsymbol{\theta}_2^*$ used in Eq. (33) can be derived from the variational inequality in Eq. (12) followed by relaxations.

Utilizing $\|X^T\boldsymbol{\theta}_1^*\|_\infty \leq 1$ and $|s^*| \leq 1$, we can set $\boldsymbol{\theta} = s^*\boldsymbol{\theta}_1^*$ in Eq. (12) and obtain

$$\left\langle \boldsymbol{\theta}_2^* - \frac{\mathbf{y}}{\lambda_2}, s^*\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_2^* \right\rangle \geq 0, \tag{34}$$

which leads to

$$\begin{aligned} &\left\langle \boldsymbol{\theta}_2^* - \frac{\mathbf{y}}{\lambda_2}, \boldsymbol{\theta}_2^* - \frac{\mathbf{y}}{\lambda_2} \right\rangle - \left\langle \boldsymbol{\theta}_2^* - \frac{\mathbf{y}}{\lambda_2}, s^*\boldsymbol{\theta}_1^* - \frac{\mathbf{y}}{\lambda_2} \right\rangle \\ &= \left\langle \boldsymbol{\theta}_2^* - \frac{\mathbf{y}}{\lambda_2}, \boldsymbol{\theta}_2^* - \frac{\mathbf{y}}{\lambda_2} + \frac{\mathbf{y}}{\lambda_2} - s^*\boldsymbol{\theta}_1^* \right\rangle \leq 0. \end{aligned} \tag{35}$$

Since

$$\left\langle \boldsymbol{\theta}_2^* - \frac{\mathbf{y}}{\lambda_2}, s^*\boldsymbol{\theta}_1^* - \frac{\mathbf{y}}{\lambda_2} \right\rangle \leq \left\| \boldsymbol{\theta}_2^* - \frac{\mathbf{y}}{\lambda_2} \right\|_2 \left\| s^*\boldsymbol{\theta}_1^* - \frac{\mathbf{y}}{\lambda_2} \right\|_2, \tag{36}$$

we have

$$\left\| \boldsymbol{\theta}_2^* - \frac{\mathbf{y}}{\lambda_2} \right\|_2 \leq \left\| s^*\boldsymbol{\theta}_1^* - \frac{\mathbf{y}}{\lambda_2} \right\|_2, \tag{37}$$

which is the feasible set used in Eq. (33). Note that, the ball defined by Eq. (37) has higher volume than the one defined by Eq. (34) due to the relaxation used in Eq. (36), and it can be shown that the ball defined by Eq. (34) lies within the ball defined by Eq. (37).

3.3. Comparison with the DPP approach

The feasible set for $\boldsymbol{\theta}_2^*$ used in the DPP approach is

$$\left\| \frac{\mathbf{y}}{\lambda_2} - \frac{\mathbf{y}}{\lambda_1} \right\|_2 \geq \|\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_1^*\|_2, \tag{38}$$

which can be obtained by

$$\left\langle \frac{\mathbf{y}}{\lambda_2} - \frac{\mathbf{y}}{\lambda_1}, \boldsymbol{\theta}_2^* - \boldsymbol{\theta}_1^* \right\rangle \geq \langle \boldsymbol{\theta}_2^* - \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^* - \boldsymbol{\theta}_1^* \rangle. \tag{39}$$

and

$$\left\langle \frac{\mathbf{y}}{\lambda_2} - \frac{\mathbf{y}}{\lambda_1}, \boldsymbol{\theta}_2^* - \boldsymbol{\theta}_1^* \right\rangle \leq \left\| \frac{\mathbf{y}}{\lambda_2} - \frac{\mathbf{y}}{\lambda_1} \right\|_2 \|\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_1^*\|_2, \tag{40}$$

where Eq. (39) is a result of adding Eq. (13) and Eq. (14). Therefore, although the authors in (Wang et al., 2013) motivates the DPP approach from the viewpoint of Euclidean projection, the DPP approach can indeed be treated as generating the feasible set for $\boldsymbol{\theta}_2^*$ using the variational inequality in Eq. (11) and Eq. (12) followed by relaxation in Eq. (40). Note that, the ball specified by Eq. (38) has higher volume than the one specified by Eq. (39) due to the relaxation used in Eq. (40), and it can be shown that the ball defined by Eq. (39) lies within the ball defined by Eq. (38).

4. Feature Sure Removal Parameter

In this subsection, we study the monotone properties of the upper-bound established in Theorem 3 with regard to the regularization parameter λ_2. With such study, we can identify the feature sure removal parameter—the smallest value of λ above which a feature is guaranteed to have zero coefficient and thus can be safely removed.

Without loss of generality, we assume $\langle \mathbf{x}_j, \mathbf{a} \rangle \geq 0$ and the results can be easily extended to the case $\langle \mathbf{x}_j, \mathbf{a} \rangle < 0$. In addition, we assume that if $\lambda_1 \neq \|X^T\mathbf{y}\|_\infty$ then $\boldsymbol{\theta}_1^* \neq \frac{\mathbf{y}}{\|X^T\mathbf{y}\|_\infty}$. This is a valid assumption for real data.

Let $\mathbf{y} \neq 0$, and $\lambda_{\max} = \|X^T\mathbf{y}\|_\infty \geq \lambda_1 \geq \lambda > 0$[3]. We introduce the following two auxiliary functions:

$$f(\lambda) = \frac{\langle \frac{\mathbf{y}}{\lambda} - \boldsymbol{\theta}_1^*, \mathbf{a} \rangle}{\|\frac{\mathbf{y}}{\lambda} - \boldsymbol{\theta}_1^*\|_2} \tag{41}$$

$$g(\lambda) = \frac{\langle \frac{\mathbf{y}}{\lambda} - \boldsymbol{\theta}_1^*, \mathbf{y} \rangle}{\|\frac{\mathbf{y}}{\lambda} - \boldsymbol{\theta}_1^*\|_2} \tag{42}$$

We show in Supplement D that $f(\lambda)$ is strictly increasing with regard to λ in $(0, \lambda_1]$ and $g(\lambda)$ is strictly decreasing with regard to λ in $(0, \lambda_1]$. Such monotone properties, which are illustrated geometrically in the first plot of Figure 4, guarantee that $f(\lambda) = \frac{\langle \mathbf{x}_j, \mathbf{a} \rangle}{\|\mathbf{x}_j\|_2}$ and $g(\lambda) = \frac{\langle \mathbf{x}_j, \mathbf{y} \rangle}{\|\mathbf{x}_j\|_2}$ have unique roots with regard to λ when some conditions are satisfied.

Our main results are summarized in the following theorem:

Theorem 4 *Let $\mathbf{y} \neq 0$ and $\|X^T\mathbf{y}\|_\infty \geq \lambda_1 > \lambda_2 > 0$. Let $\langle \mathbf{x}_j, \mathbf{a} \rangle \geq 0$. Assume that if $\lambda_1 \neq \|X^T\mathbf{y}\|_\infty$ then $\boldsymbol{\theta}_1^* \neq \frac{\mathbf{y}}{\|X^T\mathbf{y}\|_\infty}$.*

Define $\lambda_{2,a}$ as follows: If $\frac{\langle \mathbf{y}, \mathbf{a} \rangle}{\|\mathbf{y}\|_2} \geq \frac{\langle \mathbf{x}_j, \mathbf{a} \rangle}{\|\mathbf{x}_j\|_2}$, then let $\lambda_{2,a} = 0$; otherwise, let $\lambda_{2,a}$ be the unique value in $(0, \lambda_1]$ that satisfies $f(\lambda_{2,a}) = \frac{\langle \mathbf{x}_j, \mathbf{a} \rangle}{\|\mathbf{x}_j\|_2}$.

Define $\lambda_{2,y}$ as follows: If $\mathbf{a} = \mathbf{0}$ or if $\mathbf{a} \neq \mathbf{0}$ and $\frac{\langle \mathbf{a}, \mathbf{y} \rangle}{\|\mathbf{a}\|_2} \geq \frac{\langle \mathbf{x}_j, \mathbf{y} \rangle}{\|\mathbf{x}_j\|_2}$, then let $\lambda_{2,y} = \lambda_1$; otherwise, let $\lambda_{2,y}$ be the unique value in $(0, \lambda_1]$ that satisfies $g(\lambda_{2,y}) = \frac{\langle \mathbf{x}_j, \mathbf{y} \rangle}{\|\mathbf{x}_j\|_2}$.

We have the following monotone properties:

1. *$u_j^+(\lambda_2)$ is monotonically decreasing with regard to λ_2 in $(0, \lambda_1]$.*
2. *If $\lambda_{2,a} \leq \lambda_{2,y}$, then $u_j^-(\lambda_2)$ is monotonically decreasing with regard to λ_2 in $(0, \lambda_1]$.*

[3] If $\lambda_1 \geq \lambda_{\max}$, we have $\boldsymbol{\beta}_1^* = \mathbf{0}$ and thus we focus on $\lambda_1 \leq \lambda_{\max}$. In addition, for given λ_1, we are interested in the screening for a smaller regularization parameter, i.e., $\lambda < \lambda_1$.

3. *If $\lambda_{2,a} > \lambda_{2,y}$, then $u_j^-(\lambda_2)$ is monotonically decreasing with regard to λ_2 in $(0, \lambda_{2,y})$ and $(\lambda_{2,a}, \lambda_1)$, but monotonically increasing with regard to λ_2 in $[\lambda_{2,y}, \lambda_{2,a}]$.*

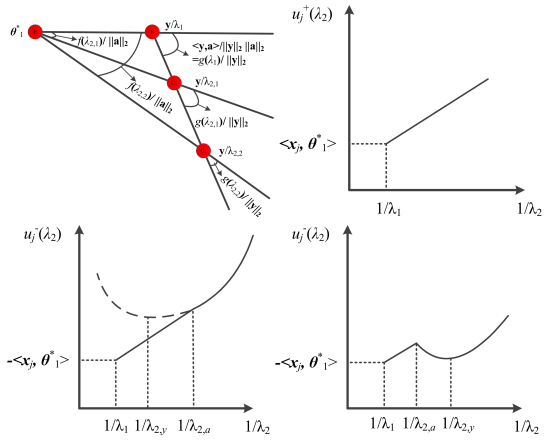

Figure 4. Illustration of the monotone properties of Sasvi for Lasso with the assumption $\langle \mathbf{x}_j, \mathbf{a} \rangle \geq 0$. The first plot geometrically shows the monotone properties of $f(\lambda)$ and $g(\lambda)$, respectively. The last three plots correspond to the three cases in Theorem 4. For illustration convenience, the x-axis denotes $\frac{1}{\lambda_2}$ rather than λ_2.

The proof of Theorem 4 is given in Supplement D. Note that, $\lambda_{2,a}$ and $\lambda_{2,y}$ are dependent on the index j, which is omitted for discussion convenience. Figure 4 illustrates results presented in Theorem 4. The first two cases of Theorem 4 indicate that, if the j-th feature $\mathbf{x}_j$ can be safely removed for a regularization parameter $\lambda = \lambda_2$, then it can also be safely discarded for any regularization parameter λ larger than λ_2. However, the third case in Theorem 4 says that this is not always true. This somehow coincides with the characteristic of Lasso that, a feature that is inactive for a regularization parameter $\lambda = \lambda_2$ might become active for a larger regularization parameter $\lambda > \lambda_2$. In other words, when following the Lasso solution path with a decreasing regularization parameter, a feature that enters into the model might get removed.

By using Theorem 4, we can easily identify for each feature a sure removable parameter λ_s that satisfies $u_j^+(\lambda) < 1$ and $u_j^-(\lambda) < 1$, $\forall \lambda > \lambda_s$. Note that Theorem 4 assumes $\langle \mathbf{x}_j, \mathbf{a} \rangle \geq 0$, but it can be easily extended to the case $\langle \mathbf{x}_j, \mathbf{a} \rangle < 0$ by replacing $\mathbf{x}_j$ with $-\mathbf{x}_j$.

5. Experiments

In this section, we conduct experiments to evaluate the performance of the proposed Sasvi in comparison with the sequential SAFE rule (Ghaoui et al., 2012), the sequential strong rule (Tibshirani et al., 2012), and the sequential DPP

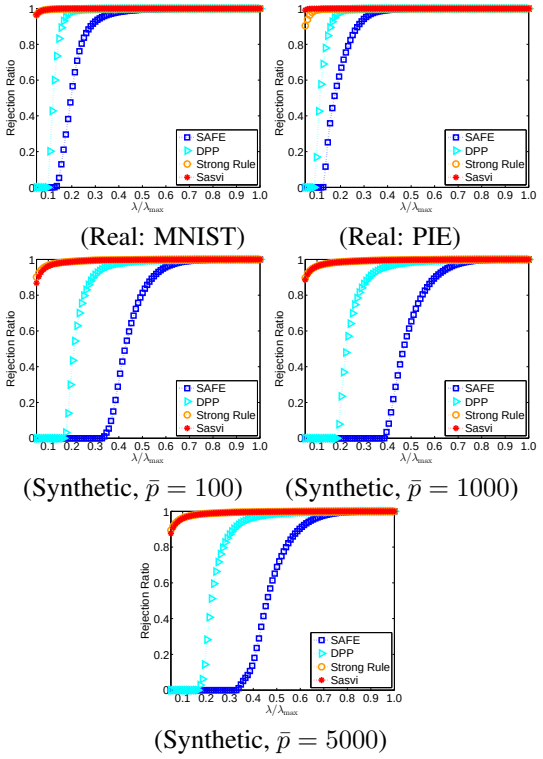

Figure 5. The rejectioin ratios—the ratios of the number features screened out by SAFE, DPP, the strong rule and Sasvi on synthetic and real data sets.

(Wang et al., 2013). Note that, SAFE, Sasvi and DPP methods are "safe" in the sense that the discarded features are guaranteed to have 0 coefficients in the true solution, and the strong rule—which is a heuristic rule—might make error and such error was corrected by a KKT condition check as suggested in (Tibshirani et al., 2012).

Synthetic Data Set We follow (Bondell & Reich, 2008; Zou & Hastie, 2005; Tibshirani, 1996) in simulating the data as follows:

$$\mathbf{y} = X\boldsymbol{\beta}^* + \sigma\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, 1), \tag{43}$$

where X has 250×10000 entries. Similar to (Bondell & Reich, 2008; Zou & Hastie, 2005; Tibshirani, 1996), we set the pairwise correlation between the i-th feature and the j-th feature to $0.5^{|i-j|}$ and draw X from a Gaussian distribution. In constructing the ground truth $\boldsymbol{\beta}^*$, we set the number of non-zero components to $\bar{p}$ and randomly assign the values from a uniform $[-1, 1]$ distribution. We set $\sigma = 0.1$ and generate the response vector $\mathbf{y} \in \mathbb{R}^{250}$ using Eq. (43). For the value of $\bar{p}$, we try 100, 1000, and 5000.

PIE Face Image Data Set The PIE face image data set used in this experiment [4] contains 11554 gray face images

[4] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

Method	Synthetic with $\bar{p}$			Real	
	100	1000	5000	MINST	PIE
solver	88.55	101.00	101.55	2683.57	617.85
SAFE	73.37	88.42	90.21	651.23	128.54
DPP	44.00	49.57	50.15	328.47	79.84
Strong	2.53	3.00	2.92	5.57	2.97
Sasvi	2.49	2.77	2.76	5.02	1.90

Table 1. Running time (in seconds) for solving the Lasso problems along a sequence of 100 tuning parameter values equally spaced on the scale of λ/λ_{max} from 0.05 to 1 by the solver (Liu et al., 2009) without screening, and the solver combined with different screening methods.

of 68 people, taken under different poses, illumination conditions and expressions. Each of the images has 32×32 pixels. To use the regression model in Eq. (43), we first randomly pick up an image as the response $\mathbf{y} \in \mathbb{R}^{1024}$, and then set the remaining images as the data matrix $X \in \mathbb{R}^{1024 \times 11553}$.

MNIST Handwritten Digit Data Set This data set contains grey images of scanned handwritten digits, including $60,000$ for training and $10,000$ for testing. The dimension of each image is 28×28. To use the regression model in Eq. (43), we first randomly select 5000 images for each digit from the training set (and in total we have 50000 images) and get a data matrix $X \in \mathbb{R}^{784 \times 50000}$, and then we randomly select an image from the testing set and treat it as the response vector $\mathbf{y} \in \mathbb{R}^{784}$.

Experimental Settings For the Lasso solver, we make use of the SLEP package (Liu et al., 2009). For a given generated data set (X and $\mathbf{y}$), we run the solver with or without screening rules to solve the Lasso problems along a sequence of 100 parameter values equally spaced on the λ/λ_{max} scale from 0.05 to 1.0. The reported results are averaged over 100 trials of randomly drawn X and $\mathbf{y}$.

Results Table 1 reports the running time by different screening rules, and Figure 5 presents the corresponding rejection ratios—the ratios of the number features screened out by the screening approaches. It can be observed that the propose Sasvi significantly outperforms the safe screening rules such as SAFE and DPP. The reason is that, Sasvi is able to discard more inactive features as discussed in Section 3. In addition, the rejection ratios of the strong rule and Sasvi are comparable, and both of them are more effective in discarding inactive features than SAFE and DPP. In terms of the speedup, Sasvi provides better performance than the strong rule. The reason is that the strong rule is a heuristic screening method, i.e., it may mistakenly discard active features which have nonzero components in the solution, and thus the strong rule needs to check the KKT conditions to make correction if necessary to ensure the correctness of the result. In contrast, Sasvi does not need to check the KKT conditions or make correction since the discarded features are guaranteed to be absent from the resulting sparse representation.

6. Conclusion and Discussion

The safe screening is a technique for improving the computational efficiency by eliminating the inactive features in sparse learning algorithms. In this paper, we propose a novel approach called Sasvi (Safe screening with variational inequalities). The proposed Sasvi has three modules: dual problem derivation, feasible set construction, and upper-bound estimation. The key contribution of the proposed Sasvi is the usage of the variational inequality which provides the sufficient and necessary optimality conditions for the dual problem. Several existing approaches can be casted as relaxed versions of the proposed Sasvi, and thus Sasvi provides a stronger screening rule. The monotone properties of the established upper-bound are studied based on a sure removal regularization parameter which can be identified for each feature.

The proposed Sasvi can be extended to solve the generalized sparse linear models, by filling in Figure 1 with the three key modules. For example, the sparse logistic regression can be written as

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{\beta}^T \mathbf{x}_i)) + \lambda \|\boldsymbol{\beta}\|_1. \qquad (44)$$

We can derive its dual problem as

$$\min_{\boldsymbol{\theta}: \|X^T\boldsymbol{\theta}\|_\infty \leq 1} -\sum_{i=1}^{n} \left(\log\left(\frac{\frac{y_i}{\lambda}}{\frac{y_i}{\lambda} - \theta_i} \right) + \frac{\theta_i}{\frac{y_i}{\lambda}} \log\left(\frac{\frac{y_i}{\lambda} - \theta_i}{\theta_i} \right) \right).$$

According to Lemma 1, for the dual optimal θ_i^*, the optimality condition via the variational inequality is

$$\sum_{i=1}^{n} \frac{1}{\frac{y_i}{\lambda}} \log\left(\frac{\frac{y_i}{\lambda} - \theta_i^*}{\theta_i^*} \right) (\theta_i - \theta_i^*) \leq 0, \forall \boldsymbol{\theta} : \|X^T\boldsymbol{\theta}\|_\infty \leq 1.$$

Then, we can construct the feasible set for $\boldsymbol{\theta}_2^*$ at the regularization parameter λ_2 in a similar way to the $\Omega(\boldsymbol{\theta}_2^*)$ in Eq. (15). Finally, we can estimate the upper-bound of $|\langle \mathbf{x}_j, \boldsymbol{\theta}_2^* \rangle|$ by Eq. (16), and discard the j-th feature if such upper-bound is smaller than 1. Note that, compared to the Lasso case, Eq. (16) is much more challenging for the logistic loss case. We plan to replace the feasible set $\Omega(\boldsymbol{\theta}_2^*)$ by its quadratic approximation so that Eq. (16) has an easy solution. We also plan to apply the proposed Sasvi to solving the Lasso solution path using LARS.

Acknowledgments

This work was supported in part by NSFC (61035003), NIH (LM010730) and NSF (IIS-0953662, CCF-1025177).

References

Bondell, H. and Reich, B. Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics*, 64:115–123, 2008.

Candes, E. and Wakin, M. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25:21–30, 2008.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

Friedman, J. H., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Ghaoui, L., Viallon, V., and Rabbani, T. Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization*, 8:667–698, 2012.

Koh, K., Kim, S., and Boyd, S. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research*, 8:1519–1555, 2007.

Liu, J., Ji, S., and Ye, J. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009. URL http://www.public.asu.edu/~jye02/Software/SLEP.

Nesterov, Y. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., 2004.

Nesterov, Y. Gradient methods for minimizing composite objective function. *Mathematical Programming*, 140: 125–161, 2013.

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

Schwarz, G. estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

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

Tibshirani, R., Bien, J., Friedman, J. H., 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*, 74:245–266, 2012.

Wang, J., Lin, B., Gong, P., Wonka, P., and Ye, J. Lasso screening rules via dual polytope projection. In *Advances in Neural Information Processing Systems*, 2013.

Zhen, J. X., Hao, X., and Peter, J. R. Learning sparse representations of high dimensional data on large scale dictionaries. In *Advances in Neural Information Processing Systems*, 2011.

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