Dynamic Sasvi: Strong Safe Screening for Norm-Regularized Least Squares

Anonymous Author(s)
Affiliation
Address
email

Abstract

A recently introduced technique, called "safe screening," for a sparse optimization problem allows us to identify irrelevant variables in the early stages of optimization. In this paper, we first propose a flexible framework for safe screening based on the Fenchel–Rockafellar duality and then derive a strong safe screening rule for norm-regularized least squares using the proposed framework. We refer to the proposed screening rule for norm-regularized least squares as "dynamic Sasvi" because it can be interpreted as a generalization of Sasvi. Unlike the original Sasvi, it does not require the exact solution of a more strongly regularized problem; hence, it works safely in practice. We show that our screening rule always eliminates more features compared with the existing state-of-the-art methods.

1 Introduction

Sparse models such as Lasso [Tibshirani, 1996] and group Lasso [Yuan and Lin, 2006] have been widely studied in the areas of statistics and machine learning and used for various applications such as compressed sensing [Donoho, 2006] and biomarker discovery [Climente-González et al., 2019]. Although sparse models can be formulated as a simple convex optimization problem, the computational cost can be large if the numbers of samples and dimensions are extremely large.

To address this problem, a technique called safe screening has been introduced [Ghaoui et al., 2010] for Lasso problems. Specifically, it eliminates variables that are guaranteed to be zero in the Lasso solution before solving the original Lasso optimization problem. Many safe screening methods have been proposed for various problems [Ghaoui et al., 2010, Ogawa et al., 2013, Wang et al., 2015, Liu et al., 2014, Xiang et al., 2017]. These are called sequential screening rules because they require the solution to a more strongly regularized problem. A recent technique, called dynamic screening, has been proposed to eliminate variables through an estimated solution in an iterative solver [Bonnefoy et al., 2015]. In particular, Gap Safe [Fercoq et al., 2015, Ndiaye et al., 2015], a dynamic screening framework, is widely used owing to its generality and efficiency [Ndiaye et al., 2017, Shibagaki et al., 2016, Bao et al., 2020, Raj et al., 2016, Ndiaye et al., 2020].

In this paper, we propose a dynamic safe screening algorithm that is stronger than Gap Safe for the *Lasso-Like* problem, which includes norm-regularized least squares. To this end, we first propose a general screening framework based on the Fenchel–Rockafellar duality and then derive *Dynamic Sasvi*, a strong safe screening rule for *Lasso-like* problems. Our framework can be regarded as a generalization of the Gap Safe framework, and thus, we can derive Gap Safe simply using our results. Moreover, owing to this generalization, we can use the strong problem adaptive inequality. The derived screening rule for *Lasso-like* problems can be seen as a dynamic variant of safe screening with variational inequalities (Sasvi) [Liu et al., 2014], which is a sequential screening rule for Lasso. Therefore, we refer to this as Dynamic Sasvi. Unlike the original Sasvi, dynamic Sasvi does not require an exact solution to the problem with another hyper-parameter and, hence, operates safely in

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

practice. Moreover, we propose the use of dynamic enhanced dual polytope projection (EDPP) [Wang et al., 2015], which is a relaxation of dynamic Sasvi due to the introduction of a minimum radius sphere. We show, both theoretically and experimentally, that the screening power and computational costs of Dynamic Sasvi and Dynamic EDPP compare favorably with those of other state-of-the-art Gap Safe methods.

Contribution: The contributions of our study are summarized as follows.

- We propose a flexible screening framework based on the Fenchel–Rockafellar duality, which is a generalization of the Gap Safe framework [Ndiaye et al., 2017].
- We propose two novel dynamic screening rules for norm-regularized least squares: a dynamic variant of Sasvi [Liu et al., 2014] and a dynamic variant of EDPP.
- We show that Dynamic Sasvi always eliminates more features and increases the speed of the solver compared with Gap Safe [Fercoq et al., 2015, Ndiaye et al., 2017] .

2 Preliminary

In this section, we first formulate the problem and introduce the key techniques used in this study.

2.1 Notation

Given $h : \mathbb{R}^m \to [-\infty, \infty]$, the domain of h is defined by $\mathrm{dom}(h) := \{\boldsymbol{z} \in \mathbb{R}^m \mid |h(\boldsymbol{z})| < \infty\}$ and $h^\star : \mathbb{R}^m \to [-\infty, \infty]$, which is the Fenchel conjugate of h, is defined as $h^\star(\boldsymbol{v}) := \sup_{\boldsymbol{z}\in\mathbb{R}^m} \boldsymbol{v}^\top \boldsymbol{z} - h(\boldsymbol{z})$. If h is proper, the Fenchel–Young inequality,

$$h(\boldsymbol{z}) + h^\star(\boldsymbol{v}) \geq \boldsymbol{v}^\top \boldsymbol{z}, \tag{1}$$

can be proven directly based on definition of the Fenchel conjugate. The subdifferential of a proper function, $h : \mathbb{R}^m \to (-\infty, \infty]$, at $\boldsymbol{z}$ is given as

$$\partial h(\boldsymbol{z}) := \{\boldsymbol{v} \in \mathbb{R}^m \mid \forall \boldsymbol{w} \in \mathbb{R}^m \ \boldsymbol{v}^\top(\boldsymbol{w} - \boldsymbol{z}) + h(\boldsymbol{z}) \leq h(\boldsymbol{w})\}.$$

The next proposition is important for deriving safe-screening algorithms.

Proposition 1 *Assume that $h : \mathbb{R}^m \to (-\infty, \infty]$ is a proper lower semi-continuous convex function and $\boldsymbol{z}, \boldsymbol{v} \in \mathbb{R}^m$. We then have*

$$\boldsymbol{v} \in \partial h(\boldsymbol{z}) \iff h(\boldsymbol{z}) + h^\star(\boldsymbol{v}) = \boldsymbol{v}^\top \boldsymbol{z} \iff \boldsymbol{z} \in \partial h^\star(\boldsymbol{v}).$$

See Section 16 of [Bauschke et al., 2011] for the proof. For convex set $C \subset \mathbb{R}^m$, the relative interior of C is defined as

$$\mathrm{relint}(C) := \{v \in C \mid \forall w \in C \ \exists \epsilon > 0 \text{ s.t. } v + \epsilon(v - w) \in C\}.$$

2.2 Problem Formulation

In this study, we consider an optimization problem, formulated as

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \ \ f(\boldsymbol{X}\boldsymbol{\beta}) + g(\boldsymbol{\beta}), \tag{2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^d$ is the optimization variable, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ is a constant matrix, and $f : \mathbb{R}^n \to (-\infty, \infty]$ and $g : \mathbb{R}^d \to (-\infty, \infty]$ are proper lower semi-continuous convex functions. We assume

$$\exists \boldsymbol{\beta} \in \mathrm{relint}(\mathrm{dom}(g)) \text{ s.t. } \boldsymbol{X}\boldsymbol{\beta} \in \mathrm{relint}(\mathrm{dom}(f))$$

and the existence of the optimal point, i.e.,

$$\exists \hat{\boldsymbol{\beta}} \in \mathrm{dom}(P) \text{ s.t. } P(\hat{\boldsymbol{\beta}}) = \inf_{\boldsymbol{\beta}\in\mathbb{R}^d} P(\boldsymbol{\beta}),$$

where $P : \mathbb{R}^d \to \mathbb{R}$ is defined as $P(\boldsymbol{\beta}) = f(\boldsymbol{X}\boldsymbol{\beta}) + g(\boldsymbol{\beta})$. Moreover, we focus on the cases in which g induces sparsity. Although all theorems in this paper hold, we cannot eliminate any variables without sparsity.

This is a popular class of optimization problem, with the most popular example being Lasso [Tibshirani, 1996]:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1.$$

Many extensions of Lasso, including Group-Lasso [Yuan and Lin, 2006], elastic net [Zou and Hastie, 2005], and sparse logistic regression [Meier et al., 2008], are in this class. The dual problems of a support vector machine [Cortes and Vapnik, 1995] and a support vector regression [Smola and Schölkopf, 2004] are also in this class.

2.3 Dual Problem

In the derivation of a safe screening rule for the optimization problem, Eq. (2), the Fenchel–Rockafellar dual formulation, plays an important role.

Theorem 2 *(Fenchel–Rockafellar Duality) If all assumptions for the optimization problem* (2) *are satisfied, we have the following:*

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} f(\boldsymbol{X}\boldsymbol{\beta})+g(\boldsymbol{\beta})=\max_{\boldsymbol{\theta}\in\mathbb{R}^n} -f^\star(-\boldsymbol{\theta})-g^\star(\boldsymbol{X}^\top\boldsymbol{\theta}). \tag{3}$$

The proof of Theorem 2 is provided in the appendix. We denote $-f^\star(-\boldsymbol{\theta})-g^\star(\boldsymbol{X}^\top\boldsymbol{\theta})$ by $D(\boldsymbol{\theta})$. For primal/dual solutions, there are many conditions that are equivalent to optimality. Herein, we provide a list of such conditions for convenience.

Proposition 3 *(Optimal Condition) If all assumptions for the optimization problem* (2) *are satisfied, the following are equivalent.*

(a) $\hat{\boldsymbol{\beta}}\in\operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} P(\boldsymbol{\beta})\wedge\hat{\boldsymbol{\theta}}\in\operatorname{argmax}_{\boldsymbol{\theta}\in\mathbb{R}^n} D(\boldsymbol{\theta})$

(b) $P(\hat{\boldsymbol{\beta}})=D(\hat{\boldsymbol{\theta}})$

(c) $f(\boldsymbol{X}\hat{\boldsymbol{\beta}})+f^\star(-\hat{\boldsymbol{\theta}})=-\hat{\boldsymbol{\theta}}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}=-g(\hat{\boldsymbol{\beta}})-g^\star(\boldsymbol{X}^\top\hat{\boldsymbol{\theta}})$

(d) $-\hat{\boldsymbol{\theta}}\in\partial f(\boldsymbol{X}\hat{\boldsymbol{\beta}})\wedge\boldsymbol{X}^\top\hat{\boldsymbol{\theta}}\in\partial g(\hat{\boldsymbol{\beta}})$

(e) $\boldsymbol{X}\hat{\boldsymbol{\beta}}\in\partial f^\star(-\hat{\boldsymbol{\theta}})\wedge\hat{\boldsymbol{\beta}}\in\partial g^\star(\boldsymbol{X}^\top\hat{\boldsymbol{\theta}})$

(Proof) (a) $\iff$ (b) is directly derived from the strong duality. (b) $\iff$ (c) is derived from the Fenchel–Young inequality (1). (c) $\iff$ (d) $\iff$ (e) are derived from Proposition 1. □

2.4 Relationship of Dual Safe Region and Screening

For the optimization problem (2), we can eliminate some features by constructing a simple region that contains $\hat{\boldsymbol{\theta}}$. Assume that the dual optimal point, $\hat{\boldsymbol{\theta}}$, is within region $\mathcal{R}$. According to Proposition 3, we have

$$\hat{\boldsymbol{\beta}}\in\partial g^\star(\boldsymbol{X}^\top\hat{\boldsymbol{\theta}})\subset\bigcup_{\boldsymbol{\theta}\in\mathcal{R}}\partial g^\star(\boldsymbol{X}^\top\boldsymbol{\theta}).$$

Hence, if $\bigcup_{\boldsymbol{\theta}\in\mathcal{R}}\partial g^\star(\boldsymbol{X}^\top\boldsymbol{\theta})\subset\{\boldsymbol{\beta}\mid\boldsymbol{\beta}_i=0\}$ holds, we obtain $\hat{\boldsymbol{\beta}}_i=0$. A simple example is the following corollary.

Example 4 *Consider an optimization problem, i.e., Eq.* (2) *with* $g(\boldsymbol{\beta})=\|\boldsymbol{\beta}\|_1$. *Assume that* $\hat{\boldsymbol{\theta}}\in\mathcal{R}$. *Then, we have*

$$\max_{\boldsymbol{\theta}\in\mathcal{R}}|\boldsymbol{x}_i^\top\boldsymbol{\theta}|<1 \implies \hat{\boldsymbol{\beta}}_i=0.$$

(Proof) Based on the definition of g, we have $\partial g^\star(\boldsymbol{X}^\top\boldsymbol{\theta})\subset\{\boldsymbol{\beta}\mid\boldsymbol{\beta}_i=0\}\iff|\boldsymbol{x}_i^\top\boldsymbol{\theta}|<1$. When $\max_{\boldsymbol{\theta}\in\mathcal{R}}|\boldsymbol{x}_i^\top\boldsymbol{\theta}|<1$, we have $\hat{\boldsymbol{\beta}}\in\bigcup_{\boldsymbol{\theta}\in\mathcal{R}}\partial g^\star(\boldsymbol{X}^\top\boldsymbol{\theta})\subset\{\boldsymbol{\beta}\mid\boldsymbol{\beta}_i=0\}$. □

To eliminate features safely, we should first construct a simple region $\mathcal{R} \subset \mathbb{R}^n$ which contains $\hat{\boldsymbol{\theta}}$ and then determine whether $\bigcup_{\boldsymbol{\theta} \in \mathcal{R}} \partial g^\star(\boldsymbol{X}^\top \boldsymbol{\theta}) \subset \{\boldsymbol{\beta} \in \mathbb{R}^d \mid \boldsymbol{\beta}_i = 0\}$ is guaranteed. In this study, we provide a novel general framework for constructing a simple safe region, $\mathcal{R} \subset \mathbb{R}^n$. By combining it with existing methods to calculate an upper bound of $\bigcup_{\boldsymbol{\theta} \in \mathcal{R}} \partial g^\star(\boldsymbol{X}^\top \boldsymbol{\theta})$ (cf. [Ndiaye et al., 2017] for the un-overlapping group L1 norm and sparse group L1 norm, [Bao et al., 2020] for the ordered weighted L1 norm), we can formulate strong safe screening rules.

3 General Framework for Constructing Safe Region

Herein, we propose a general framework for constructing a dual region that contains the solution to the optimization problem expressed by Eq. (3). Our framework consists of a general lower bound and a problem-adaptive upper bound of the optimal value. Hence, we can derive a narrower region than the framework with a general upper bound under certain situations.

The general lower-bound is derived from the optimal condition and the L-strong convexity. Assume that $f^\star$ is L-strongly convex ($L \geq 0$). Then, as D is L-strongly concave and $\hat{\boldsymbol{\theta}}$ is the optimal point, we have $D(\tilde{\boldsymbol{\theta}}) \leq D(\hat{\boldsymbol{\theta}}) - \frac{L}{2}\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\|_2^2$ for $\forall \tilde{\boldsymbol{\theta}} \in \mathbb{R}^n$. Thus, we have

$$\hat{\boldsymbol{\theta}} \in \{\boldsymbol{\theta} \mid \frac{L}{2}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 + D(\tilde{\boldsymbol{\theta}}) \leq D(\boldsymbol{\theta})\}.$$

Because this region is too complicated for screening, we use a simple upper bound of $D(\boldsymbol{\theta})$ to construct a simple safe region.

Theorem 5 *Consider the optimization problem expressed by Eq.* (3) *and assume that $f^\star$ is L-strongly convex ($L \geq 0$). Let $\hat{\boldsymbol{\theta}}$ be the solution to Eq.* (3)*. Assume that $D(\boldsymbol{\theta})$ is upper-bounded by $u(\boldsymbol{\theta})$, i.e., $\forall \boldsymbol{\theta} \in \mathbb{R}^n$ $D(\boldsymbol{\theta}) \leq u(\boldsymbol{\theta})$. Then, for $\forall \tilde{\boldsymbol{\theta}} \in \mathbb{R}^n$, we have*

$$\hat{\boldsymbol{\theta}} \in \mathcal{R}(\tilde{\boldsymbol{\theta}}, u) = \{\boldsymbol{\theta} \mid \frac{L}{2}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 + D(\tilde{\boldsymbol{\theta}}) \leq u(\boldsymbol{\theta})\}.$$

The complexity of $\mathcal{R}(\tilde{\boldsymbol{\theta}}, u)$ depends on the complexity of u. For example, if u is linear, then $\mathcal{R}(\tilde{\boldsymbol{\theta}}, u)$ is a sphere. We can construct a narrow, simple, and safe region with a tight simple upper-bound u. In fact, the Gap Safe Sphere region [Fercoq et al., 2015, Ndiaye et al., 2017] can be derived easily from this theorem and weak duality.

Corollary 6 *(Gap Safe Sphere) Consider the optimization problem described by Eq.* (3) *and assume that $f^\star$ is L-strongly convex ($L \geq 0$). Let $\hat{\boldsymbol{\theta}}$ be the solution to Eq.* (3)*. For $\forall \tilde{\boldsymbol{\beta}} \in \mathbb{R}^d$ and $\forall \tilde{\boldsymbol{\theta}} \in \mathbb{R}^n$, we have*

$$\hat{\boldsymbol{\theta}} \in \{\boldsymbol{\theta} \mid \frac{L}{2}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 + D(\tilde{\boldsymbol{\theta}}) \leq P(\tilde{\boldsymbol{\beta}})\}.$$

(Proof) Based on a weak duality, we have $\forall \boldsymbol{\theta}$ $D(\boldsymbol{\theta}) \leq P(\tilde{\boldsymbol{\beta}})$. Using this constant function as an upper bound in Theorem 5, we can derive the corollary directly. □

Hence, our framework can be seen as a generalization of Gap Safe. Owing to this generalization, we can use a stronger problem-adaptive upper bound rather than a weak duality. In the next section, we derive specific regions for the *Lasso-Like* problem. Some regions for other problems are presented in the appendix.

4 Safe Region for Lasso-like Problem

In this section, we introduce a strong upper bound for the dual problems of Lasso and similar problems. The dome region derived from it can be seen as a generalization of Sasvi [Liu et al., 2014] and is narrower than Gap Safe region.

4.1 Norm-regularized Least Squares Problem and its Generalization

Norm-regularized least squares is an optimization problem and is formulated as $\underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + g(\boldsymbol{\beta})$ where g is a norm. This is a subset of problem 2. Although this formulation includes Lasso [Tibshirani, 1996], (overlapping) group-Lasso [Yuan and Lin, 2006, Jacob et al., 2009], and ordered weighted L1 regression [Figueiredo and Nowak, 2016], the non-negative Lasso is not included. To unify them, we define the *Lasso-like* problem as follows:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + g(\boldsymbol{\beta}), \tag{4}$$

where the problem satisfies all assumptions for Eq. (2) and g satisfies

$$\forall k \geq 0, \boldsymbol{\beta} \in \mathbb{R}^d \ g(k\boldsymbol{\beta}) = kg(\boldsymbol{\beta}). \tag{5}$$

For the Lasso-like problem, the Fenchel conjugate functions of f and g are given as

$$f^\star(-\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 - \boldsymbol{y}^\top\boldsymbol{\theta} \qquad g^\star(\boldsymbol{X}^\top\boldsymbol{\theta}) = \begin{cases} 0 & (\forall\boldsymbol{\beta} \ \ \boldsymbol{\theta}^\top\boldsymbol{X}\boldsymbol{\beta} - g(\boldsymbol{\beta}) \leq 0) \\ \infty & (\exists\boldsymbol{\beta} \ \ \boldsymbol{\theta}^\top\boldsymbol{X}\boldsymbol{\beta} - g(\boldsymbol{\beta}) > 0). \end{cases} \tag{6}$$

Note that $\{\boldsymbol{\theta} \mid g^\star(\boldsymbol{X}^\top\boldsymbol{\theta}) = 0\}$ is a closed convex set. Hence, the Lasso-like problem is a class of problems whose Fenchel–Rockafellar dual can be seen as a convex projection.

4.2 Proposed Dome Region for Lasso-like Problem

Using Theorem 5, we can construct a safe region by proposing an upper bound, $u(\boldsymbol{\theta})$. In this section, we propose a tight upper bound for Lasso-like problems. The direct expression of $f^\star$ in Eq. (6) is sufficiently simple. We only need an upper bound of $-g^\star$ to construct a simple region. The upper bound is given as follows.

Lemma 7 *For Lasso-like problems* (4)*, for* $\forall\tilde{\boldsymbol{\beta}} \in \mathbb{R}^d$ *and* $\forall\boldsymbol{\theta} \in \mathbb{R}^n$*, we have*

$$D(\boldsymbol{\theta}) \leq -f^\star(-\boldsymbol{\theta}) + \inf_{k\geq 0} g(k\tilde{\boldsymbol{\beta}}) - \boldsymbol{\theta}^\top\boldsymbol{X}(k\tilde{\boldsymbol{\beta}}) = \begin{cases} -f^\star(-\boldsymbol{\theta}) & (g(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\theta}^\top\boldsymbol{X}\tilde{\boldsymbol{\beta}} \geq 0) \\ -\infty & (g(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\theta}^\top\boldsymbol{X}\tilde{\boldsymbol{\beta}} < 0). \end{cases} \tag{7}$$

The proof of Lemma 7 is given in the appendix. Using the right-hand side of Eq. (7) as an upper bound of D, we can construct a simple and safe region.

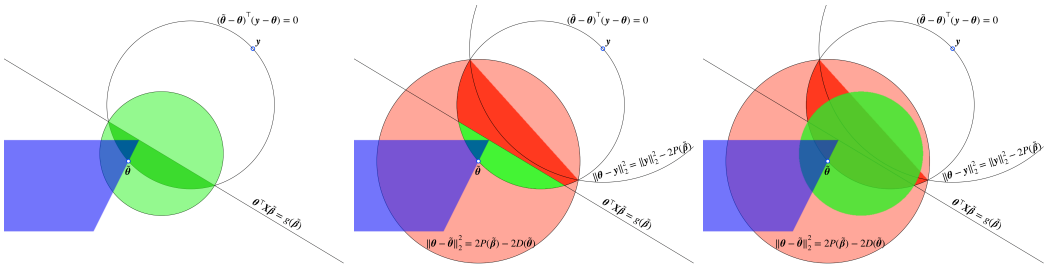
Theorem 8 *Consider the Lasso-like problem described by Eq.* (4) *and its Fenchel–Rockafellar dual problem presented in Eq.* (3)*. Let* $\hat{\boldsymbol{\theta}}$ *be the dual optimal point. We assume that* $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^d$ *and* $\tilde{\boldsymbol{\theta}} \in \mathrm{dom}(D)$*. Then,* $\hat{\boldsymbol{\theta}}$ *is within the Dynamic Sasvi region, which is given as the intersection of a sphere and a half space:*

$$\begin{aligned}\mathcal{R}^{\mathrm{DS}}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}}) :=& \{\boldsymbol{\theta} \mid \frac{1}{2}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 + D(\tilde{\boldsymbol{\theta}}) \leq -f^\star(-\boldsymbol{\theta}) \wedge 0 \leq g(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\theta}^\top\boldsymbol{X}\tilde{\boldsymbol{\beta}}\} \\ =& \{\boldsymbol{\theta} \mid (\boldsymbol{y}-\boldsymbol{\theta})^\top(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq 0 \wedge 0 \leq g(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\theta}^\top\boldsymbol{X}\tilde{\boldsymbol{\beta}}\}.\end{aligned}$$

The proof of Theorem 8 is given in the appendix. As proven in section 4.5, $\mathcal{R}^{\mathrm{DS}}(\boldsymbol{\beta}^k, \boldsymbol{\theta}^k)$ converges to $\{\hat{\boldsymbol{\theta}}\}$ when $\lim_{k\to\infty} P(\boldsymbol{\beta}^k) - D(\boldsymbol{\theta}^k) = 0$. When we have good primal/dual feasible points, we can derive a very narrow safe region and eliminate almost all irrelevant features.

4.3 Relation to Sasvi

In this section, we show that safe screening with variational inequality (Sasvi) [Liu et al., 2014] is a special case of our screening rule. First, we review Sasvi. The target task of Sasvi is to minimize $\frac{1}{2}\left\|\lambda^{-1}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|_2^2 + \|\boldsymbol{\beta}\|_1$ with many λs. Sasvi is called "sequential" screening because it is designed to solve a sequence of problems with a sequence of penalty parameters, $\lambda_1 > \lambda_2 > \ldots > \lambda_M$. Let $\hat{\boldsymbol{\beta}}^{(\lambda)}$ and $\hat{\boldsymbol{\theta}}^{(\lambda)}$ be the optimal points of the primal problem and its Fenchel–Rockafellar dual problem,



(a) Regions of dynamic Sasvi (dark green) and dynamic EDPP (light green).

(b) Regions of dynamic Sasvi (green), Gap Safe Sphere (light red) and Gap Safe Dome (dark red).

(c) Regions of dynamic EDPP (green), Gap Safe Sphere (light red), Gap Safe Dome (dark red).

Figure 1: Comparisons of various safe regions for Lasso ($\boldsymbol{X} = \begin{pmatrix} 2 & 0 \\ -1 & 3 \end{pmatrix}$, $\boldsymbol{y} = \begin{pmatrix} 1.5 \\ 1 \end{pmatrix}$). The blue region represents the feasible region. $\tilde{\boldsymbol{\beta}}$ is obtained by using a cycle of coordinate descent. $\tilde{\boldsymbol{\theta}} = \phi(\tilde{\boldsymbol{\beta}})$.

respectively. Because the dual problem can be interpreted as a projection from $\lambda^{-1}\boldsymbol{y}$ to a closed convex set, $\{\boldsymbol{\theta} \in \mathbb{R}^n \mid \|\boldsymbol{X}^\top\boldsymbol{\theta}\|_\infty \leq 1\}$, the following two variational inequalities hold:

$$0 \geq ({\lambda_2}^{-1}\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{(\lambda_2)})^\top(\hat{\boldsymbol{\theta}}^{(\lambda_1)} - \hat{\boldsymbol{\theta}}^{(\lambda_2)}) \qquad 0 \geq ({\lambda_1}^{-1}\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{(\lambda_1)})^\top(\hat{\boldsymbol{\theta}}^{(\lambda_2)} - \hat{\boldsymbol{\theta}}^{(\lambda_1)}). \tag{8}$$

Given $\hat{\boldsymbol{\theta}}^{(\lambda_1)}$, these inequalities provide a safe region for $\hat{\boldsymbol{\theta}}^{(\lambda_2)}$. Sasvi finds some zero elements of $\hat{\boldsymbol{\beta}}^{(\lambda_2)}$ based on this region. Note that $-\hat{\boldsymbol{\theta}}^{(\lambda_1)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda_1)} - {\lambda_1}^{-1}\boldsymbol{y}$ and $-\hat{\boldsymbol{\theta}}^{(\lambda_1)\top}\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda_1)} = -g(\hat{\boldsymbol{\beta}}^{(\lambda_1)})$ are derived from Proposition 3. Then, we have

$$({\lambda_1}^{-1}\boldsymbol{y} - \hat{\boldsymbol{\theta}}^{(\lambda_1)})^\top(\hat{\boldsymbol{\theta}}^{(\lambda_2)} - \hat{\boldsymbol{\theta}}^{(\lambda_1)}) = \hat{\boldsymbol{\theta}}^{(\lambda_2)\top}\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda_1)} - g(\hat{\boldsymbol{\beta}}^{(\lambda_1)}).$$

Because ${\lambda_2}^{-1}\boldsymbol{y}$ in Eq. (8) corresponds to $\boldsymbol{y}$ in Theorem 8, $\mathcal{R}^{\mathrm{DS}}(\hat{\boldsymbol{\beta}}^{(\lambda_1)}, \hat{\boldsymbol{\theta}}^{(\lambda_1)})$ is equal to the Sasvi region. Hence, we can state that our method is a generalization of Sasvi. Our method requires primal/dual feasible points, unlike the original Sasvi, which requires the dual optimal point of a problem with another parameter. Such screening rules are called "dynamic" because they dynamically provide safe regions as iterative optimization proceeds. For this reason, we have labeled it as "Dynamic Sasvi." Owing to its "dynamic" property, Dynamic Sasvi eliminates almost all irrelevant features in the late step of the iterative optimization. As reported in [Fercoq et al., 2015], some sequential safe screening rules, including Sasvi, are not safe in practice because we cannot provide an exact solution for λ_1. Dynamic Sasvi overcomes this problem because it does not require the solution of a problem with another parameter.

4.4 Comparison to Gap Safe

Here, we show that the proposed method is stronger than Gap Safe Dome [Fercoq et al., 2015] and Gap Safe Sphere [Fercoq et al., 2015], [Ndiaye et al., 2017] for Lasso-like problems. As shown in [Fercoq et al., 2015], for Lasso, the regions of the Gap Safe Dome and Gap Safe Sphere are the relaxation of the intersection of a sphere and the contra of another sphere:

$$\{\boldsymbol{\theta} \mid (\boldsymbol{y} - \boldsymbol{\theta})^\top(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq 0 \wedge -f^\star(-\boldsymbol{\theta}) \leq P(\tilde{\boldsymbol{\beta}})\},$$

where $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\theta}}$ are primal/dual feasible vectors, respectively. $(\boldsymbol{y} - \hat{\boldsymbol{\theta}})^\top(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \leq 0$ is the variational inequality, and $-f^\star(-\hat{\boldsymbol{\theta}}) \leq P(\tilde{\boldsymbol{\beta}})$ is derived from weak duality. Using a simple transformation, we have

$$\begin{aligned} -f^\star(-\boldsymbol{\theta}) \leq P(\tilde{\boldsymbol{\beta}}) &\iff -\frac{1}{2}\|\boldsymbol{\theta}\|_2^2 + \boldsymbol{y}^\top\boldsymbol{\theta} \leq \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|_2^2 + g(\tilde{\boldsymbol{\beta}}) \\ &\iff -\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{\theta}\|_2^2 \leq g(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\theta}^\top\boldsymbol{X}\tilde{\boldsymbol{\beta}}. \end{aligned}$$

Hence, we have $0 \leq g(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\theta}^\top\boldsymbol{X}\tilde{\boldsymbol{\beta}} \implies -f^\star(-\boldsymbol{\theta}) \leq P(\tilde{\boldsymbol{\beta}})$. This means that the region of dynamic Sasvi is a subset of the region of Gap Safe Dome. Our screening is always stronger than Gap Safe Dome and Gap Safe Sphere. Figure 1b shows the regions of Dynamic Sasvi, Gap Safe Dome, and Gap Safe Sphere.

4.5 Sphere Relaxation (Dynamic EDPP)

In some situations, even a dome region is too complicated to calculate $\bigcup_{\boldsymbol{\theta}\in\mathcal{R}} \partial g^\star(\boldsymbol{X}^\top\boldsymbol{\theta})$. For such cases, we propose using a minimum radius sphere that includes the dynamic Sasvi region. This method can be seen as a dynamic variant of the EDPP [Wang et al., 2015] because the EDPP is the minimum radius sphere relaxation of Sasvi.

Theorem 9 *Consider the Lasso-like problem presented in Eq.* (4) *and its Fenchel–Rockafellar dual problem in Eq.* (3)*. We assume that* $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^d$ *and* $\tilde{\boldsymbol{\theta}} \in \mathrm{dom}(D)$*. If* $n \geq 2$*, the minimum radius sphere including* $\mathcal{R}^{\mathrm{DS}}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$ *is*

$$\mathcal{R}^{\mathrm{DE}}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}}) := \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \frac{1}{2}(\tilde{\boldsymbol{\theta}} + \boldsymbol{y}) + \alpha \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|_2^2 \leq \frac{1}{4}\|\tilde{\boldsymbol{\theta}} - \boldsymbol{y}\|_2^2 - \alpha^2 \|\boldsymbol{X}\tilde{\boldsymbol{\beta}}\|_2^2\}, \tag{9}$$

where

$$\alpha = \max\left(0, \|\boldsymbol{X}\tilde{\boldsymbol{\beta}}\|_2^{-2}\left(\frac{1}{2}(\tilde{\boldsymbol{\theta}} + \boldsymbol{y})^\top \boldsymbol{X}\tilde{\boldsymbol{\beta}} - g(\tilde{\boldsymbol{\beta}})\right)\right).$$

The proof of Theorem 9 is provided in the appendix. Figures 1a and 1c show the dynamic EDPP region and other regions. To compare the Dynamic EDPP region and the Gap Safe Sphere region in general, we present the next theorem.

Theorem 10 *Let* r *be the radius of* $\mathcal{R}^{\mathrm{DE}}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$*. We then have*

$$r < \sqrt{P(\tilde{\boldsymbol{\beta}}) - D(\tilde{\boldsymbol{\theta}})}.$$

The proof is given in the appendix. Because the radius of the Gap Safe Sphere region is $\sqrt{2(P(\tilde{\boldsymbol{\beta}}) - D(\tilde{\boldsymbol{\theta}}))}$, the squared radius of the Dynamic EDPP region is always less than half that of Gap Safe Sphere. Furthermore, Theorem 10 shows that $\mathcal{R}^{\mathrm{DE}}(\boldsymbol{\beta}^k, \boldsymbol{\theta}^k)$ and $\mathcal{R}^{\mathrm{DS}}(\boldsymbol{\beta}^k, \boldsymbol{\theta}^k)$ converge to $\{\hat{\boldsymbol{\theta}}\}$ when $\lim_{k\to\infty} P(\boldsymbol{\beta}^k) - D(\boldsymbol{\theta}^k) = 0$.

5 Implementation for Lasso

In this section, we provide a specific solver based on Theorem 8. Because the algorithm used to calculate $\bigcup_{\boldsymbol{\theta}\in\mathcal{R}} \partial g^\star(\boldsymbol{X}^\top\boldsymbol{\theta})$ depends on g, we introduce a Lasso solver as an example. We must choose an iterative solver to combine with the screening methods because they cannot estimate the solution alone. Although our methods can work with any iterative method, we use coordinate descent, which is recommended in [Friedman et al., 2007].

5.1 Choice of a Dual Feasible Point

As shown in the previous section, $\lim_{t\to\infty} \mathcal{R}^{\mathrm{DS}}(\boldsymbol{\beta}^t, \boldsymbol{\theta}^t)$ converges to $\{\hat{\boldsymbol{\theta}}\}$ when $\lim_{t\to\infty} P(\boldsymbol{\beta}^t) - D(\boldsymbol{\theta}^t) = 0$ holds. Because the iterative solver provides a sequence of primal points whose $P(\boldsymbol{\beta}^t)$ converges to $D(\hat{\boldsymbol{\theta}})$, we only need a converging sequence of dual points to obtain a converging safe region. The next theorem provides such a sequence.

Theorem 11 *(Converging* $\boldsymbol{\theta}^t$*) Consider the optimization problem in Eq.* (4) *with* $g(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$*. Let* $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ *and* $\hat{\boldsymbol{\theta}} \in \mathbb{R}^n$ *be the primal/dual solution. Assume that* $\lim_{t\to\infty} \boldsymbol{\beta}^t = \hat{\boldsymbol{\beta}}$*. Let us define* $\phi : \mathbb{R}^d \to \mathbb{R}^n$ *as*

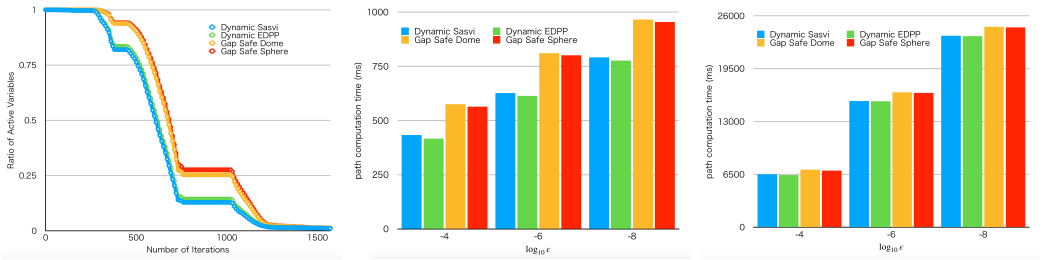
$$\phi(\boldsymbol{\beta}) := \frac{1}{\max(1, \|\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\|_\infty)}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Then, $\forall \boldsymbol{\beta}\ \phi(\boldsymbol{\beta}) \in \mathrm{dom}(D)$ *and* $\lim_{t\to\infty} \phi(\boldsymbol{\beta}^t) = \hat{\boldsymbol{\theta}}$ *hold.*

(Proof) $\phi(\boldsymbol{\beta}) \in \mathrm{dom}(D)$ is directly derived from $\|\boldsymbol{X}^\top\phi(\boldsymbol{\beta})\|_\infty = \min(\|\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\|_\infty, 1) \leq 1$. Because $\{\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t\}$ converges to $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\theta}}$, $\lim_{t\to\infty} \phi(\boldsymbol{\beta}^t) = \hat{\boldsymbol{\theta}}$ also holds. □

Algorithm 1 Coordinate descent with Dynamic Sasvi for Lasso

Input: $\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}^0, T, c, \epsilon$
Initialize $\tilde{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}^0$, $\mathcal{A} \leftarrow [\![d]\!]$
for $t \in [\![T]\!]$ **do**
  if $t \mod c = 1$ **then**
    Compute $\tilde{\boldsymbol{\theta}} = \phi_{\mathcal{A}}(\tilde{\boldsymbol{\beta}})$
    if $P(\tilde{\boldsymbol{\beta}}) - D(\tilde{\boldsymbol{\theta}}) \leq \frac{1}{2}\|\boldsymbol{y}\|_2^2 \epsilon$ **then**
      break
    end if
    $\mathcal{R} \leftarrow \mathcal{R}^{\mathrm{DS}}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$
    $\mathcal{A} \leftarrow \{j \in \mathcal{A} : \max_{\boldsymbol{\theta} \in \mathcal{R}} |\boldsymbol{x}_j^\top \boldsymbol{\theta}| \geq 1\}$
    for $j \in [\![d]\!] - \mathcal{A}$ **do**
      $\tilde{\boldsymbol{\beta}}_j \leftarrow 0$
    end for
  end if
  for $j \in \mathcal{A}$ **do**
    $u \leftarrow \tilde{\boldsymbol{\beta}}_j \|\boldsymbol{x}_j\|_2^2 - \boldsymbol{x}_j^\top (\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{y})$
    $\tilde{\boldsymbol{\beta}}_j \leftarrow \frac{1}{\|\boldsymbol{x}_j\|_2^2} \mathrm{sign}(u) \max(0, |u| - 1)$
  end for
end for
Output: $\tilde{\boldsymbol{\beta}}$



(a) Feature remaining rate (Leukemia). (b) Computational time (Leukemia). (c) Computational time (RCV1).

Figure 2: (a): Feature remaining rate of each iteration for Lasso on leukemia (density of $n = 72, d = 7128$). (b) Average computational time of the Lasso path on subsampled leukemia (density of $n = 50, d = 7128$). (c): Average computational time of the Lasso path on subsampled RCV1 (sparsity of $n = 20000, d = 47236$)

If $\mathcal{A}$ is the set of features that is not yet eliminated, we can use

$$\phi_{\mathcal{A}}(\boldsymbol{\beta}) := \frac{1}{\max(1, \max_{j \in \mathcal{A}} \boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}))} (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})$$

instead of $\phi(\boldsymbol{\beta})$. Although $\phi_{\mathcal{A}}(\boldsymbol{\beta}) \in \mathrm{dom}(D)$ is not guaranteed, $\phi_{\mathcal{A}}(\boldsymbol{\beta})$ is guaranteed to satisfy all the constraints that are active in the dual solution. In other words, $\phi_{\mathcal{A}}(\boldsymbol{\beta})$ is in the domain of the dual problem of the small primal problem without eliminated features. Now, we can optimize the problem using the proposed screening method. The pseudocode is described in Algorithm 1. The direct expression of $\max_{\boldsymbol{\theta} \in \mathcal{R}^{\mathrm{DS}}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})} |\boldsymbol{x}_j^\top \boldsymbol{\theta}|$ is given in the appendix.

5.2 Computational Cost

In the screening part of Algorithm 1, only multiplying $\boldsymbol{X}$ or $\boldsymbol{X}^\top$ costs $O(nd)$. Because $\tilde{\boldsymbol{\theta}} = \phi_{\mathcal{A}}(\tilde{\boldsymbol{\beta}})$ is a linear combination of $\boldsymbol{X}\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{y}$, we can obtain $\boldsymbol{X}^\top \boldsymbol{\theta}$ from $\boldsymbol{X}^\top \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{X}^\top \boldsymbol{y}$. Because $\boldsymbol{X}^\top \boldsymbol{y}$ is constant, only the calculations of $\boldsymbol{X}\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{X}^\top \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ cost $O(nd)$. Hence, the screening cost is almost the same for all methods, which require $\boldsymbol{X}^\top \boldsymbol{X}\tilde{\boldsymbol{\beta}}$, including Gap Safe.

Table 1: Logarithm of acceleration ratio. The larger values indicate a greater speed-up.

Dataset	-log ϵ	Dynamic Sasvi	Dynamic EDPP	Gap Safe Dome	Gap Safe Sphere
Leukemia	4	0.468 ± 0.066	0.487 ± 0.066	0.349 ± 0.066	0.358 ± 0.060
	6	0.828 ± 0.072	0.838 ± 0.067	0.719 ± 0.074	0.725 ± 0.072
	8	0.987 ± 0.057	0.997 ± 0.056	0.902 ± 0.066	0.907 ± 0.066
RCV1	4	0.257 ± 0.0056	0.263 ± 0.0078	0.221 ± 0.0081	0.228 ± 0.0089
	6	0.373 ± 0.0065	0.374 ± 0.0085	0.345 ± 0.0099	0.346 ± 0.0114
	8	0.417 ± 0.0065	0.418 ± 0.0079	0.397 ± 0.0086	0.398 ± 0.0092

6 Experiments

In this section, we show the efficacy of the proposed methods using real-world data. We compared the proposed methods with Gap Safe Sphere and Gap Safe Dome [Fercoq et al., 2015, Ndiaye et al., 2017], which are state-of-the-art dynamic safe screening methods. All methods were run on a Macbook Air with a 1.1 GHz quad-core Intel Core i5 CPU with 16 GB of RAM. All methods were implemented in C++ using the Accelerate framework.

6.1 Number of Screened Variables

First, we compared the number of screened variables among the four dynamic safe screening methods. We solved the Lasso problem using the leukemia dataset [1] [Golub et al., 1999] (dense data with 72 samples and 7128 features) and $\lambda = \frac{1}{100}\|\boldsymbol{X}^\top \boldsymbol{y}\|_\infty$. We used cyclic coordinate descent as the iterative algorithm and screened the variables for 10 iterations each. Figure 2a shows the ratio of the uneliminated features at each iteration. As guaranteed theoretically, we can see that Dynamic Sasvi eliminates more variables in earlier steps compared with both Gap Safe Dome and Gap Safe Sphere. The figure also shows that the Dynamic EDPP, which is a relaxed version of Dynamic Sasvi, eliminates almost the same number of features as Dynamic Sasvi.

6.2 Gains in the Computation of Lasso Paths

Next, we compared the computation time of the path of the Lasso solutions for various values of λ. We used $\lambda_j = 100^{-\frac{j}{99}}\|\boldsymbol{X}^\top \boldsymbol{y}\|_\infty$ $(j = 0, \dots, 99)$. We used the estimated primal solution for λ_j with scaling as the initial vector in the solver for the problem with λ_{j+1}. The iterative solver stops when the duality gap is smaller than $\epsilon(P(\mathbf{0}) - D(\mathbf{0}))$. Note that $P(\mathbf{0}) - D(\mathbf{0})$ makes the stopping criterion independent of the data scale. We used the leukemia and RCV1 [2] [Lewis et al., 2004] datasets (sparse data with 23149 samples and 47236 features). We subsampled the data 50 times and ran all the methods for the same 50 subsamples. The subsampled data size was 50 for leukemia and 20000 for RCV1. Figures 2b and 2c show the average computation times of the Lasso path for the leukemia and RCV1 datasets, respectively. For all settings, dynamic Sasvi and dynamic EDPP outperform Gap Safe Dome and Gap Safe Sphere. Table 1 lists the average values and standard deviations of the negative of the logarithm of the computational time ratio with respect to that for the same subsample without screening. The proposed methods are significantly faster than the Gap Safe methods.

7 Conclusion

In this study, we proposed a framework for safe screening based on the Fenchel–Rockafellar duality and derived Dynamic Sasvi and Dynamic EDPP, which are specific safe screening methods for Lasso-like problems. Dynamic Sasvi and Dynamic EDPP can be regarded as dynamic feature elimination variants of Sasvi and EDPP, respectively. We proved that Dynamic Sasvi always eliminates more features than both Gap Safe Sphere and Gap Safe Dome. Dynamic EDPP is based on the sphere relaxation of the Dynamic Sasvi region and eliminates almost the same number of features as Dynamic Sasvi. We also showed experimentally that the computational costs of the proposed methods are lower than those of Gap Safe Sphere and Gap Safe Dome.

[1] https://leo.ugr.es/elvira/DBCRepository/Leukemia/ALLAML.html

[2] https://scikit-learn.org/0.18/datasets/rcv1.html

References

Runxue Bao, Bin Gu, and Heng Huang. Fast oscar and owl regression via safe screening rules. In *ICML*, 2020.

Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

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

Héctor Climente-González, Chloé-Agathe Azencott, Samuel Kaski, and Makoto Yamada. Block hsic lasso: model-free biomarker detection for ultra-high dimensional data. *Bioinformatics*, 35(14): i427–i435, 2019.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Mind the duality gap: safer rules for the lasso. In *ICML*, 2015.

Mario Figueiredo and Robert Nowak. Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects. In *AISTATS*, 2016.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 12 2007.

Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*, 2010.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439): 531–537, 1999.

Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *ICML*, 2009.

David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

Jun Liu, Zheng Zhao, Jie Wang, and Jieping Ye. Safe screening with variational inequalities and its application to lasso. In *ICML*, 2014.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

Eugène Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparse multi-task and multi-class models. In *NIPS*, 2015.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, 18(1):4671–4703, 2017.

Eugene Ndiaye, Olivier Fercoq, and Joseph Salmon. Screening rules and its complexity for active set identification, 2020.

Kohei Ogawa, Yoshiki Suzuki, and Ichiro Takeuchi. Safe screening of non-support vectors in pathwise svm computation. In *ICML*, 2013.

Anant Raj, Jakob Olbrich, Bernd Gärtner, Bernhard Schölkopf, and Martin Jaggi. Screening rules for convex problems. *arXiv preprint arXiv:1609.07478*, 2016.

Atsushi Shibagaki, Masayuki Karasuyama, Kohei Hatano, and Ichiro Takeuchi. Simultaneous safe screening of features and samples in doubly sparse modeling. In *ICML*, 2016.

Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

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

Jie Wang, Peter Wonka, and Jieping Ye. Lasso screening rules via dual polytope projection. *Journal of Machine Learning Research*, 16(1):1063–1101, 2015.

Zhen James Xiang, Yun Wang, and Peter J. Ramadge. Screening tests for lasso problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):1008–1027, May 2017.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49–67, 02 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We described the target problems in Section 2 and Section 4.1
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work makes sparse convex optimizers faster. It is quite unlikely to have a negative effect.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We described the target problems in Section 2 and Section 4.1
   (b) Did you include complete proofs of all theoretical results? [No] Complete proofs of some theoretical results are in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 6
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 6
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 6 and Table 1
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 6
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 6
   (b) Did you mention the license of the assets? [Yes] See Section 6
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]