Journal of Convex Analysis
Volume 28 (2021), No. 4, 1053–1072

Screening Rules and its Complexity for Active Set Identification

Eugene Ndiaye
Riken AIP, Japan; eugene.ndiaye@riken.jp

Olivier Fercoq
LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France

Joseph Salmon
IMAG,Univ Montpellier, CNRS, Montpellier, France

Received: September 5, 2020
Accepted: October 10, 2020

Screening rules were recently introduced as a technique for explicitly identifying active structures such as sparsity, in optimization problem arising in machine learning. This has led to new methods of acceleration based on a substantial dimension reduction. We show that screening rules stem from a combination of natural properties of subdifferential sets and optimality conditions, and can hence be understood in a unified way. Under mild assumptions, we analyze the number of iterations needed to identify the optimal active set for any converging algorithm. We show that it only depends on its convergence rate.

Keywords: Convex optimization, screening rules, active set identification, nonsmooth optimization.
2010 Mathematics Subject Classification: 49M29, 65K05, 90C06, 90C25.

1. Introduction

In learning problems involving a large number of variables, sparse models such as Lasso and Support Vector Machines (SVM) allow to select the most important variables. For instance, the Lasso estimator depends only on a subset of features that have a maximal absolute correlation with the residual; whereas the SVM classifier depends only on a subset of sample (the support vectors) that characterize the margin. The remaining features/variables have no contribution to the optimal solution. Thus, early detection of those non influential variables may lead to significant simplifications of the problem, memory and computational resources saving. Some noticeable examples are the *facial reduction* preprocessing steps used for accelerating the linear programming solvers [4, 22] and conic programming [3], we refer to [25, 6] for recent reviews. Other applications can also be found in [26] for projecting onto the simplex and ℓ_1 ball in [5] or data preprocessing before application of statistical methods [10] in high dimensional settings. In an optimization problem, screening rules eliminate the variables that have no influence on the set of optimal coefficients. Recently, [9] have introduced the *safe screening rules*, which ignore non-active variables in Lasso or non-support vectors for SVM, without false exclusions. It is a versatile optimization technique useful for many machine learning tasks, see [11, 14, 32, 34, 40, 45, 46] to name a few. However, in the existing safe screening studies, a case by case safe screening rule is developed for each problem, and there is

ISSN 0944-6532 / $ 2.50

no unified understanding for which class of optimization problems safe screening are possible and how efficient the screening can be. We summarize our contributions as follows:

- We provide a simple setting for explicitly identifying optimal active sets in convex composite optimization problems with separable regularization. Our approach combines a natural property of the subdifferential set with optimality conditions. It allows us to subsume the previously introduced screening rules in a unified framework.
- Relying on a recent work on the convergence of the duality gap [7], we can easily provide (for safe screening of features using smooth losses or for safe screening of observations with a strongly-convex penalty) an algorithm-independent complexity analysis of the (finite) active set identification. Then, the latter holds for any converging optimization scheme as soon as it is endowed with screening rules.
- We discuss some acceleration strategies based on a combination of safe and relaxed safe rules. Several popular strategies such as strong rules [44] and some recent working sets methods [14, 15, 24] can then be generalized to a larger set of optimization problems.

Notation. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a function not identically equal to infinity and such that there is an affine function minorizing it.
Denoting $\operatorname{dom} f = \{x \in \mathbb{R}^n : f(x) < +\infty\} \neq \emptyset$, the *Fenchel-Legendre conjugate* of f is the function $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined by $f^*(x^*) = \sup_{x \in \operatorname{dom} f} \langle x^*, x\rangle - f(x)$. The *subdifferential* of f at x is the set

$$\partial f(x) = \{v \in \mathbb{R}^n : f(z) \geq f(x) + \langle v, z - x\rangle, \forall z \in \mathbb{R}^n\} \ . \tag{1}$$

A convex function f is *μ-strongly convex* (with $\mu > 0$) if for all $z, x \in \operatorname{dom} f$,

$$f(z) \geq f(x) + \langle v, z - x\rangle + \frac{\mu}{2} \|z - x\|^2 \text{ for all } v \in \partial f(x) \ .$$

We recall that a function f is strongly concave if $-f$ is strongly convex. A differentiable convex function is *L-smooth* (with $L > 0$) if its gradient mapping is L-Lipschitz. The *support function* of a nonempty set C is defined as $\mathcal{S}_C(x) = \sup_{c \in C} \langle c, x\rangle$. If C is closed, convex and contains 0, we define its *polar function* as $\mathcal{S}_C^\circ(x^*) = \sup_{\mathcal{S}_C(x) \leq 1} \langle x^*, x\rangle$. The interior (resp. boundary) of a set C is denoted $\operatorname{int} C$ (resp. $\operatorname{bd} C$). Its *indicator function* is defined as $\iota_C(x) = 0$ if $x \in C$ and $+\infty$ otherwise. We denote by $[T]$ the set $\{1, \dots, T\}$ for any non zero integer T. The vector of observations is $y \in \mathbb{R}^n$ and the design matrix $X = [x_1, \dots, x_n]^\top = [X_1, \dots, X_p] \in \mathbb{R}^{n \times p}$ has n observations row-wise, and p features (column-wise). A group of features is a subset $g \subset [p]$ and $|g|$ is its cardinality. The set of groups is denoted by $\mathcal{G}$. We denote by β_g the vector in $\mathbb{R}^{|g|}$ which is the restriction of β to the indices in g. We also use the notation $X_g \in \mathbb{R}^{n \times |g|}$ to refer to the sub-matrix of X assembled from the columns with indices $j \in g$.

2. General framework

We consider composite optimization problems involving a sum of a data fitting function plus a separable group regularization function $\Omega(\beta) = \sum_{g \in \mathcal{G}} \Omega_g(\beta_g)$ that

enforces specific structure such as sparsity in the optimal solutions:

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} f(X\beta) + \Omega(\beta) = P(\beta) \ . \tag{2}$$

We will assume that the functions f and Ω are proper (i.e., strictly larger than $-\infty$ with a non empty domain), lower-semicontinuous and convex. We also assume that $\mathrm{Im}(X) \cap \mathrm{dom}(f)$ is non empty, the primal problem admits a solution and there exists a vector β in $\mathrm{dom}(\Omega)$ such that f is continuous at $X\beta$. Such formulations often arise in statistical learning in a context of regularized empirical risk minimization [39].

The main purpose of screening rules is to identify and eliminate the irrelevant features/samples during (or before) an optimization process for solving Problem (2), to reduce the memory and computational footprint. For instance, many features X_j for some j in $[p]$ are expected to be irrelevant in the Lasso estimator [43], defined as $\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|y - X\beta\|^2 + \lambda \|\beta\|_1$ (for some observation vector $y \in \mathbb{R}^n$ and regularization parameter $\lambda > 0$). They correspond to coordinate where $\hat{\beta}_j = 0$. Exploiting the known sparsity of the solution, safe screening rules [9] discard features prior to starting a sparse solver. Computational gains are obtained from the reduction of the dimension. A similar example is the SVM classifier relying on the hinge loss, defined as $\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \sum_{i \in [n]} \max(0, 1 - y_i x_i^\top \beta) + \frac{\lambda}{2} \|\beta\|^2$ in which some sample x_i are expected to be irrelevant: $\hat{\theta}_i = 0$ for some i in $[n]$ where $\hat{\theta}$ is the associated dual solution. Identifying irrelevant variables in optimization is not restricted to sparse problems. We show how it is intimately related to the non smoothness of the objective function and present a simple and self contained framework for developing screening rules.

2.1. Identification rules

We introduce the main lemma that captures a natural property of the subdifferential (1), that allows us to obtain a simple and unified presentation of many screening rules recently introduced in the literature.

Lemma 2.1. (Separation of subdifferentials) *Let P be a proper function and z such that* $\mathrm{int}\partial P(z) \neq \emptyset$. *Then we have* $\mathrm{int}(\partial P(z)) \cap \partial P(z') = \emptyset$ *for all* $z' \neq z$.

Proof. Let z' be such that there exists g in $\mathrm{int}\partial P(z) \cap \partial P(z')$. Now g in the open set $\mathrm{int}\partial P(z)$ implies that there exists $\alpha > 0$ such that $g_\alpha := g + \alpha(z' - z) \in \partial P(z)$.

$$\begin{aligned} \text{Then} \qquad P(z) &\geq P(z') + \langle g, z - z' \rangle \geq P(z) + \langle g_\alpha, z' - z \rangle + \langle g, z - z' \rangle \\ &= P(z) + \alpha \|z' - z\|_2^2 \ , \end{aligned}$$

where the first (resp. the second) inequality comes from $g \in \partial P(z')$ (resp. $g_\alpha \in \partial P(z)$). Thus $z' = z$. □

Remark 2.2. Without precautions, the interior cannot be replaced by the relative interior in Lemma 2.1. For example, the function $P(z) = |z_1 - z_2|$ defined on $\mathbb{R}^2$ does not satisfy the assumption of the Lemma. Indeed, any $z = (z_1, z_2) \in \mathbb{R}^2$ for $z_1 = z_2$, the subdifferential $\partial P(z) = \{(a, -a) : a \in [-1, 1]\}$ is a one dimensional line segment in $\mathbb{R}^2$ with empty interior. Also for non-convex function, note that the lemma is applicable only when the subdifferential set defined in (1) is non-empty.

The contrapositive of Lemma 2.1 leads to a simple identification rule:

$$\text{int}\partial P(z) \cap \partial P(z') \neq \emptyset \implies z = z' \ . \tag{3}$$

Fermat's rule provides the optimality condition:

$$0 \in \partial P(\hat{\beta}) = X^\top \partial f(X\hat{\beta}) + \partial\Omega(\hat{\beta}) \ ,$$

$$\text{which is equivalent to} \qquad X^\top\hat{\theta} \in \partial\Omega(\hat{\beta}) \text{ for some } \hat{\theta} \in -\partial f(X\hat{\beta}) \ . \tag{4}$$

Suppose now that there exists $\beta_g^\star$ such that $\text{int}\partial\Omega_g(\beta_g^\star)$ is non empty. In the case $\Omega_g(\beta_g) = \|\beta_g\|$ for instance, we can take $\beta_g^\star = 0$. For any group of features g in $\mathcal{G}$ such that $X_g^\top\hat{\theta} \in \partial\Omega_g(\hat{\beta}_g)$, and any vector $\beta_g^\star$ with $\text{int}\partial\Omega_g(\beta_g^\star)$ is non empty, we have

$$X_g^\top\hat{\theta} \in \text{int}\partial\Omega_g(\beta_g^\star) \Longrightarrow \text{int}\partial\Omega_g(\beta_g^\star) \cap \partial\Omega_g(\hat{\beta}_g) \neq \emptyset \overset{(3)}{\Longrightarrow} \beta_g^\star = \hat{\beta}_g \ .$$

This relation means that the group of feature X_g is irrelevant and can be discarded in the Problem (2) i.e., $\hat{\beta}_g$ is identified to be equal to $\beta_g^\star$, whenever $X_g^\top\hat{\theta}$ belongs to $\text{int}\partial\Omega_g(\beta_g^\star)$. However, since $\hat{\theta}$ depends on the *unknown* solution $\hat{\beta}$, this rule is of limited use. Fortunately, it is often possible to construct a set $\mathcal{R} \subset \mathbb{R}^n$, called a *safe region*, that contains such a $\hat{\theta}$. This observation leads to the following result.

Proposition 2.3. (Feature-wise screening rule) *Let $\beta_g^\star$ such that $\text{int}\partial\Omega_g(\beta_g^\star) \neq \emptyset$ and let $\mathcal{R}$ be a set containing $\hat{\theta}$. Then*

$$X_g^\top\mathcal{R} \subset \text{int}\partial\Omega_g(\beta_g^\star) \Longrightarrow \beta_g^\star = \hat{\beta}_g \ .$$

Proposition 2.3 provides a general recipe for applying (safe) screening rule in optimization Problem (2) without assumption on the algebraic expression of functions f and Ω.

Choice of the safe region $\mathcal{R}$. When f is L-Lipschitz, for any $v \in -\partial f(X\hat{\beta})$ as in (4), we have $\|v\| \leq L\,\|X\|$ (see [38, Lemma 2.6]). The set $\mathcal{R}$ can be taken as a ball of radius $L\,\|X\|$. When f is differentiable with Lipschitz gradient, we will see that $\hat{\theta}$ belongs to a ball centered around any dual estimate θ and whose radius quantifies the approximation quality. Thus, the better the estimate θ, the smaller the safe region $\mathcal{R}$, i.e., the more efficient the screening rule is.

Dual formulation. Similar results naturally hold in the dual space. Under the classical Fenchel-Rockafellar duality [36, Chapter 31], the dual formulation of Problem (2) reads:

$$\hat{\theta} \in \underset{\theta\in\mathbb{R}^n}{\arg\max} -f^*(-\theta) - \Omega^*(X^\top\theta) = D(\theta) \ . \tag{5}$$

The solutions $\hat{\beta}$ in the primal and $\hat{\theta}$ in the dual are linked by the optimality condition in (4). In (5), the role of f and Ω are flipped and we have the next result.

Proposition 2.4. (Sample-wise screening rule) *Suppose that $f(x) = \sum_{i=1}^n f_i(x_i)$. Let $\theta^\star \in \mathbb{R}^n$ be a vector such that for some $i \in [n]$, $\text{int}\partial f_i^*(\theta_i^\star) \neq \emptyset$ and let $\mathcal{R} \subset \mathbb{R}^p$ be a set containing a solution $\hat{\beta}$ of Problem (2). Then*

$$x_i^\top\mathcal{R} \subset \text{int}\partial f_i^*(\theta_i^\star) \Longrightarrow -\hat{\theta}_i = \theta_i^\star \ .$$

For simplicity, we will mostly restrict the discussion on the primal formulation since the same properties can be recovered by duality.

2.2. Examples

To apply the screening rule within the introduced setting, we only have to identify the subdifferential of Ω and points where it has a non empty interior. We present a few examples popular in machine learning and statistics, as well as other ones commonly met in convex optimization [13].

Sparsity inducing regularization. The Lasso and Elastic net are examples where each group reduces to a single feature $g = \{j\}$ for $j \in [p]$. In both cases the feature-wise penalties can be written respectively $\Omega_g(\beta_g) = |\beta_j|$, and $\Omega_g(\beta_g) = |\beta_j| + \alpha\beta_j^2/2$ for $\alpha \geq 0$. For $j \in [p]$ and $\beta_j^\star = 0$, we have $\partial\Omega_j(\beta_j^\star) = [-1, 1]$ which has a non empty interior equal to $(-1, 1)$. We also have cases where, $\Omega_g = \|\cdot\|$ is a norm. It includes examples such as Group Lasso [47] and Sparse-Group Lasso [41] for instance. In that case, $\beta_g^\star = 0 \in \mathbb{R}^{|g|}$ and $\partial\Omega_g(\beta_g^\star)$ is equal to the unit ball with respect to the dual norm of Ω_g, which also have a non empty interior. These examples also easily extend to gauges and support functions (see [13] for definitions). For instance, Let $Q \in \mathbb{R}^{|g|\times|g|}$ be a symmetric positive semi-definite matrix and define $\Omega_g(\beta_g) = \sqrt{\langle Q\beta_g, \beta_g\rangle}$. Then $\beta_g^\star$ can be any element of $\mathrm{Ker}(Q)$ and $\partial\Omega_g(\beta_g^\star) = Q^{1/2}\mathcal{B}(0, 1)$, where $\mathcal{B}(0, 1)$ is the unit ball for $\|\cdot\|$ in $\mathbb{R}^{|g|}$.

Non negativity constraint: $\Omega_j(\beta_j) = \iota_{[0,+\infty)}(\beta_j)$ with $g = \{j\}$. In this case, $\beta_j^\star = 0$ and $\partial\Omega(\beta_j^\star) = (-\infty, 0]$. This formulation appears in non-negative Least-square, non-negative Lasso and simplex constrained problems.

Box constraint: $\Omega_j(\beta_j) = \iota_{[a,b]}(\beta_j)$ with $g = \{j\}$. In this case $\beta_j^\star$ belongs to $\{a, b\}$ and $\partial\Omega_j(\beta_j^\star) = (-\infty, 0]$ (resp. $[0, +\infty)$) if $\beta_j^\star$ is equal to a (resp. b).

Hinge: $\Omega_j(\beta_j) = \max(0, 1-\beta_j)$ with $g = \{j\}$, we have $\beta_j^\star = 1$ and $\partial\Omega(\beta_j^\star) = [-1, 0]$ which have a non empty interior $(-1, 0)$. This is used in ϵ-insensitive loss which flattened the ℓ_1 norm around zero i.e., $\Omega_j(\beta_j) = \max(0, |\beta_j| - \epsilon)$ for $\epsilon \geq 0$.

Distance function: We consider a closed convex subset C of $\mathbb{R}^{|g|}$ and the set $\Omega_g(\beta_g) = \min\{\|z - \beta_g\| : z \in C\}$. In that case, $\partial\Omega_g(\beta_g^\star) = N_C(\beta_g^\star) \cap \mathcal{B}(0, 1)$ where $N_C(w)$ is the normal cone of C at $w \in \mathbb{R}^{|g|}$. Hence, $\beta_g^\star$ can be any vector in $\mathrm{bd}\, C$ such that $N_C(\beta_g^\star)$ has a non empty interior.

Piecewise affine functions: Let us consider a set of real values $\{r_1, \ldots, r_m\}$ and vectors $\{s_1, \ldots, s_m\}$ in $\mathbb{R}^{|g|}$ for an integer $m \geq 2$. One can define the set $\Omega_g(\beta_g) = \max\{r_j + \langle s_j, \beta_g\rangle : j \in [m]\}$, then $\partial\Omega_g(\beta_g^\star)$ is the convex hull of the set $\{s_j : j \in J(\beta_g^\star)\}$ where $J(w) = \{j \in [m] : \Omega_g(w) = r_j + \langle s_j, w\rangle\}$ and $\beta_g^\star$ can be chosen as any vector such that the matrix $(s_j)_{j\in J(\beta_g^\star)}$ has full rank. This generalizes the previous examples of ℓ_1 regularization and hinge loss. We refer to [13, Chapter D] for a further generalization to supremum over a collection of convex function, more sophisticated examples and detailed subdifferential calculus rules.

3. Explicit active set identification

The main interest of using screening rules in optimization algorithms is to focus the computational efforts on the most important variables (or samples in the dual). To do so, one needs to explicitly and safely identify parts of the solution vector i.e., detect some group g such that we can guarantee that $\hat{\beta}_g = \beta^\star_g$ where the latter is such that $\mathrm{int}\partial\Omega(\beta^\star_g)$ is not empty (which is a necessary condition for identifiability in our framework). This is particularly well suited for proximal (block) coordinate descent method as this type of method can easily ignore non influential (block) coordinates.

Any time a safe region $\mathcal{R}$ is considered for a safe screening test (following Proposition 2.3), one can associate to it a *safe active set* consisting of the features that cannot yet be removed.

Definition 3.1. (Feature-wise (safe) active sets) Let $\beta^\star_g$ be a vector such that $\mathrm{int}\partial\Omega_g(\beta^\star_g)$ is non empty. The set of *(group) active features* at $\beta^\star_g$ is defined as:

$$\mathcal{A} := \mathcal{A}(\hat{\theta}) = \left\{ g \in \mathcal{G} : \ X_g^\top \hat{\theta} \notin \mathrm{int}\partial\Omega_g(\beta^\star_g) \right\} \ ,$$

where $\hat{\theta}$ is a dual optimal solution in (5). Moreover, if $\mathcal{R}$ is a safe region, its corresponding set of *(group) safe active features* at $\beta^\star_g$ is defined as

$$\mathcal{A}_\mathcal{R} := \left\{ g \in \mathcal{G} : \ X_g^\top \mathcal{R} \not\subset \mathrm{int}\partial\Omega_g(\beta^\star_g) \right\} \ .$$

The complements (i.e., the set of non active groups) of $\mathcal{A}$ and $\mathcal{A}_\mathcal{R}$ are denoted by $\mathcal{Z}$ and $\mathcal{Z}_\mathcal{R}$.

3.1. Computation of screening tests

The set inclusion tests presented in Propositions 2.3 and 2.4 have a simple numerical implementation. Since the subdifferential $\partial\Omega_g(\beta^\star_g)$ is a closed convex set, for any region $\mathcal{R}$, the screening test $X_g^\top \mathcal{R} \subset \mathrm{int}\partial\Omega_g(\beta^\star_g)$ can be evaluated computationally thanks to the following classical lemma that allows to check whether a point belongs to the interior of a closed convex set by means of support function. We recall that for a nonempty set C, it is defined as $\mathcal{S}_C(x) = \sup_{c\in C}\langle c, x\rangle$.

Lemma 3.2. [13, Theorem C-2.2.3] *Let C_1 and C_2 be nonempty closed convex subsets of $\mathbb{R}^n$. Then, we have for any direction $d \in \mathrm{bd}\,\mathcal{B}(0,1)$,*

$$\mathcal{S}_{C_1}(d) < \mathcal{S}_{C_2}(d) \Longleftrightarrow C_1 \subset \mathrm{int}C_2 \ .$$

By applying Lemma 3.2 to the closed convex sets $C_1 = X_g^\top \mathcal{R}$ and $C_2 = \partial\Omega_g(\beta^\star_g)$, we obtain a computational tool for evaluating the screening tests.

Proposition 3.3. *Let $\mathcal{R}$ be a closed convex set that contains the dual optimal solution $\hat{\theta}$ in (5). For all group g in $\mathcal{G}$ and for any vector $\beta^\star_g$ such that $\mathrm{int}\partial\Omega_g(\beta^\star_g)$ is non empty, $\hat{\beta}_g = \beta^\star_g$ whenever for any direction $d \in \mathrm{bd}\,\mathcal{B}(0,1)$*

$$\textbf{\textit{Screening:}} \qquad \mathcal{S}_{\{X_g^\top\hat{\theta}\}}(d) < \mathcal{S}_{\partial\Omega_g(\beta^\star_g)}(d) \ .$$

$$\textbf{\textit{Safe screening:}} \qquad \mathcal{S}_{X_g^\top\mathcal{R}}(d) < \mathcal{S}_{\partial\Omega_g(\beta^\star_g)}(d) \ .$$

The *safe screening* rule consists in removing the g-th group from the optimization process whenever the previous test is satisfied, since then $\hat{\beta}_g$ is guaranteed to

be equal to $\beta^\star_g$. Should $\mathcal{R}$ be small enough to screen many groups, one can observe considerable speed-ups in practice as long as the testing can be performed efficiently. Thus a natural goal is to find safe regions as narrow as possible and only cheap computations are needed to check if $X_g^\top \mathcal{R} \subset \text{int}\partial\Omega_g(\beta^\star_g)$.

Relying on optimality conditions, one can easily find a set containing the dual optimal solution $\hat{\theta}$. For a pair of vector $(\beta, \theta) \in \text{dom}\, P \times \text{dom}\, D$, the duality gap is defined as the difference between the primal and dual objectives:

$$\text{Gap}(\beta, \theta) = P(\beta) - D(\theta) \ .$$

For such (β, θ) pair, weak duality holds: $P(\beta) \geq D(\theta)$, namely $P(\beta) - P(\hat{\beta}) \leq \text{Gap}(\beta, \theta)$. At optimal values and under mild conditions, we have the strong duality $\text{Gap}(\hat{\beta}, \hat{\theta}) = 0$. This allows us to exploit the duality gap as an optimality certificate or as an algorithmic stopping criterion for solving problems (2) and (5). For any dual feasible vector θ, we have $D(\hat{\theta}) \geq D(\theta)$. Moreover by weak duality, for any $\beta \in \text{dom}\, P$ (primal feasible), we have $P(\beta) \geq D(\hat{\theta})$. Whence,

$$\hat{\theta} \in \{\zeta \in \text{dom}\, D : \ P(\beta) \geq D(\zeta) \geq D(\theta)\} \ .$$

Nevertheless, the screening test corresponding to such set may be hard to compute explicitly. Various shapes have been considered in practice as a safe region $\mathcal{R}$. Here we consider *sphere regions* following the terminology introduced by [9] i.e., choosing a ball $\mathcal{R} = \mathcal{B}(c, r)$ as a safe region. In that case, by positive homogeneity of the support function, we have for any direction $d \in \mathbb{R}^{|g|}$,

$$\mathcal{S}_{X_g^\top \mathcal{B}(c,r)}(d) = \mathcal{S}_{X_g^\top c}(d) + \mathcal{S}_{X_g^\top \mathcal{B}(0,r)}(d) \leq \mathcal{S}_{X_g^\top c}(d) + r \sup_{\|u\|=1} \mathcal{S}_{X_g^\top u}(d) \ .$$

For any group g in $\mathcal{G}$, $\hat{\beta}_g = \beta^\star_g$ when for any d, the following *sphere test* holds

$$\mathcal{S}_{X_g^\top c}(d) + r \sup_{\|u\|=1} \mathcal{S}_{X_g^\top u}(d) < \mathcal{S}_{\partial\Omega_g(\beta^\star_g)}(d) \ . \tag{6}$$

Note that when Ω_g is a norm the sphere test reduces to $\Omega_g^\circ(X_g^\top c) + r\Omega_g^\circ(X_g) < 1$, where Ω_g° is the norm dual to Ω_g. For the non negativity constraint, the test reduces to $-X_j^\top c + r|X_j| < 0$.

3.2. Gap safe rules

Under the smoothness assumption on the loss function, it was shown by [11, 34, 29] that one can rely on the duality gap to construct a safe region, whenever a dual feasible point can be constructed. Here, we recall this construction and show how to generalize it to construct a dual feasible point for a wider class of problems (2).

Proposition 3.4. *If the dual objective D is μ_D-strongly concave with respect to a norm $\|\cdot\|$, for any primal/dual feasible vectors $(\beta, \theta) \in \text{dom}\, P \times \text{dom}\, D$, we have:*

$$\|\hat{\theta} - \theta\|^2 \leq \frac{2}{\mu_D} \text{Gap}(\beta, \theta) \ , \tag{7}$$

where $(\hat{\beta}, \hat{\theta})$ is any primal/dual optimal solution.

Thus, the ball $\mathcal{B}\left(\theta, \sqrt{\frac{2}{\mu_D} \text{Gap}(\beta, \theta)}\right)$ is a safe region (called gap safe sphere).

We call *gap safe rule* the sphere test in (6) applied with the gap safe sphere in (3.4).

Construction of a dual feasible vector. To build a center for the safe sphere, we map a primal vector onto the dual space thanks to the gradient[1] mapping $\nabla f(\cdot)$. However, the obtained dual vector is not necessarily feasible for the dual problem. When the projection on the feasible set is hard, a generic procedure consists in performing a rescaling step so that it belongs to the dual set. Precisely, we want to build $\theta \in \mathbb{R}^n$ such that

$$-\theta \in \operatorname{dom} f^* \text{ and } X^\top \theta \in \operatorname{dom} \Omega^* \ . \tag{8}$$

Given a vector β in $\mathbb{R}^p$, we build a dual point by rescaled gradient mapping i.e.,

$$\theta = \frac{-\nabla f(X\beta)}{\alpha} \tag{9}$$

$$\text{with} \quad \alpha = \max\left(1, \mathcal{S}^\circ_{\operatorname{dom}\Omega^*}\left(-X^\top \nabla f(X\beta)\right)\right) \ . \tag{10}$$

This choice is motivated by the primal-dual optimality link (4): $\hat{\theta} = -\nabla f(X\hat{\beta})$. Building a dual feasible vector by scaling is often used in the literature (see for instance [9]) and reduces to residual rescaling, see [21, 23] when f is quadratic.

Proposition 3.5. *If f and Ω are bounded from below, then the dual vector θ in* (9) *satisfies the feasibility condition* (8)*.*

Proof. Since Ω is bounded from below then $\Omega^*(0) = -\inf_z \Omega(z) < +\infty$ which is equivalent to $\operatorname{dom}\Omega^*$ contains 0. Since it is also closed and convex, we have $\mathcal{S}^\circ_{\operatorname{dom}\Omega^*}$ is positively homogeneous. Hence the vector θ in (9) satisfies $\mathcal{S}^\circ_{\operatorname{dom}\Omega^*}(X^\top\theta) \leq 1$ which is equivalent to $X^\top\theta$ in $\operatorname{dom}\Omega^*$. Moreover, by denoting $s = 1/\alpha \in [0,1]$, we have $-\theta = s\nabla f(X\beta) = s\nabla f(X\beta) + (1-s)0$. Since $\operatorname{dom} f^*$ is convex, it remains to show that it contains the vectors $\nabla f(X\beta)$ and 0, thus it will necessarily contains $-\theta$ by convex combination.

Since f is bounded from below, we have $f^*(0) = -\inf_z f(z) < +\infty$ which is equivalent to $0 \in \operatorname{dom} f^*$. Moreover, the equality case in the Fenchel-Young inequality shows that $f(X\beta) + f^*(\nabla f(X\beta)) = \langle \nabla f(X\beta), X\beta\rangle < +\infty$. Hence $f^*(\nabla f(X\beta))$ is also finite. □

Remark 3.6. The scaling constant α in (10) is equal to 1 if $\operatorname{dom}\Omega^*$ is unbounded.

4. How long does it take to identify the optimal active set with a screening rule?

We recall the notion of converging safe regions introduced in [11] that helps to reach exact active set identification in a finite number of steps.

Definition 4.1. (Converging Safe Region) Let $(\mathcal{R}_k)_{k\in\mathbb{N}}$ be a sequence of compact convex sets containing the dual optimal solution $\hat{\theta}$. It is a *converging sequence of safe regions* if the diameters of the sets converge to zero. The associated safe screening rules are referred to as converging.

[1] any subgradient can be used when f is not differentiable

The following proposition asserts that after a finite number of steps, the active set is exactly identified. Such a property is sometimes referred to as finite identification of the support [20].

Proposition 4.2. *Let $(\mathcal{R}_k)_{k\in\mathbb{N}}$ be a sequence of compact convex set containing $\hat{\theta}$ for each k in $\mathbb{N}$. If $(\mathcal{R}_k)_k$ is converging in the sense of Definition 4.1, then it exists an integer k_0 such that $\mathcal{A}_{\mathcal{R}_k} = \mathcal{A}$ for any $k \geq k_0$.*

Proof. We proceed in two steps.

Firstly, we show that for any direction $d \in \mathbb{R}^{|g|}$, $\max_{\theta\in\mathcal{R}_k} \mathcal{S}_{X_g^\top\theta}(d) \to_k \mathcal{S}_{X_g^\top\hat{\theta}}(d)$. Indeed, for any $k \in \mathbb{N}$ and $\theta \in \mathcal{R}_k$ we have from the sublinearity and positive homogeneity of the support function, and since $\hat{\theta}$ in $\mathcal{R}_k$:

$$\mathcal{S}_{X_g^\top\hat{\theta}}(d) \leq \max_{\theta\in\mathcal{R}_k} \mathcal{S}_{X_g^\top\theta}(d) \leq \mathcal{S}_{X_g^\top\hat{\theta}}(d) + \operatorname{diam}(\mathcal{R}_k) \sup_{\|u\|=1} \mathcal{S}_{X_g^\top u}(d) \ ,$$

The conclusion follows from the fact that $\mathcal{R}_k$ is a converging sequence, since we have $\lim_{k\to\infty} \operatorname{diam}(\mathcal{R}_k) = 0$.

Secondly, we proceed by double inclusion. First, remark that $\mathcal{A} = \mathcal{A}_{\mathcal{R}_\infty}$ where $\mathcal{R}_\infty := \{\hat{\theta}\}$. So for all $k \in \mathbb{N}$, we have $\mathcal{A} \subseteq \mathcal{A}_{\mathcal{R}_k}$ since $(\mathcal{A}_{\mathcal{R}_k})_k$ are nested sequence of sets. Reciprocally, suppose that there exists a non active group $g \in \mathcal{G}$ i.e., $\mathcal{S}_{X_g^\top\hat{\theta}}(d) < 1$ that remains in the active set $\mathcal{A}_{\mathcal{R}_k}$ for all iterations i.e., $\forall k \in \mathbb{N}$, $\max_{\theta\in\mathcal{R}_k} \mathcal{S}_{X_g^\top\theta}(d) \geq 1$. Since $\lim_{k\to\infty} \max_{\theta\in\mathcal{R}_k} \mathcal{S}_{X_g^\top\theta}(d) = \mathcal{S}_{X_g^\top\hat{\theta}}(d)$, we obtain $\mathcal{S}_{X_g^\top\hat{\theta}}(d) \geq 1$ by passing to the limit. Hence, by contradiction, there exits an integer $k_0 \in \mathbb{N}$ such that $[p]\backslash\mathcal{A} \subseteq \mathcal{A}^c_{\mathcal{R}_k}$ for all $k \geq k_0$. □

One can note that the rate of identification of the active set is strongly related to the rate at which the sequence of diameters $\operatorname{diam}(\mathcal{R}_k)$ goes to zero during the optimization process. We quantify this in the next section.

4.1. Complexity of safe active set identification

Dynamic safe screening rules have practical benefits since they increase the number of screened out variables as the algorithm proceeds. The next proposition states that if one relies on a primal converging algorithm, then the dual sequence we propose is also converging. It only requires uniqueness of $X\hat{\beta}$ (not that of $\hat{\beta}$). In the following, β_k is the current estimate of a primal solution $\hat{\beta}$ and $\theta_k = -\nabla f(X\beta_k)/\alpha_k$, with $\alpha_k = \max(1, \mathcal{S}^\circ_{\operatorname{dom}\Omega^*}(X^\top\nabla f(X\beta_k)))$, be the current estimate of the dual solution $\hat{\theta}$.

Lemma 4.3. *It holds* $\lim_{k\to\infty} X\beta_k = X\hat{\beta}$ *implies* $\lim_{k\to\infty} \theta_k = \hat{\theta}$.

Proof. Let $\alpha_k = \max(1, \mathcal{S}^\circ_{\operatorname{dom}\Omega^*}(X^\top\nabla f(X\beta_k)))$, we have:

$$\begin{aligned}\|\theta_k - \hat{\theta}\|_2 &= \left\|\nabla f(X\hat{\beta}) - \frac{\nabla f(X\beta_k)}{\alpha_k}\right\|_2 \\ &\leq \left|1 - \frac{1}{\alpha_k}\right| \|\nabla f(X\beta_k)\|_2 + \left\|\nabla f(X\hat{\beta}) - \nabla f(X\beta_k)\right\|_2 .\end{aligned}$$

If $X\beta_k \to X\hat{\beta}$ holds, then

$$\alpha_k \to \max(1, \mathcal{S}^\circ_{\operatorname{dom}\Omega^*}(X^\top\nabla f(X\hat{\beta}))) = \max(1, \mathcal{S}^\circ_{\operatorname{dom}\Omega^*}(X^\top\hat{\theta})) = 1,$$

since $\nabla f(X\hat{\beta}) = -\hat{\theta}$ thanks to the optimality condition and the feasibility of $\hat{\theta}$: $\mathcal{S}^{\circ}_{\operatorname{dom}\Omega^*}(X^\top\hat{\theta}) \leq 1$. Hence the right hand side of the previous inequality converges to zero, and the conclusion holds. □

From Lemma 4.3 and by strong duality we conclude that the sequence of radius $r_k = (2\operatorname{Gap}(\beta_k, \theta_k)/\mu_D)^{1/2}$ converges to 0 as k goes to ∞. Hence, the sequence of safe balls $\mathcal{B}(\theta_k, r_k)$ converges to $\{\hat{\theta}\}$. Whence, we deduce the following property.

Proposition 4.4. *The Gap Safe rules produce converging safe regions.*

The results in Propositions 4.2 and 4.4 ensure that screening rules, applied iteratively with the duality gap based safe region, will identify the active set after a finite number of iterations.

For any safe ball $\mathcal{B}(c_k, r_k)$, we have the inequalities

$$\mathcal{S}_{X_g^\top\hat{\theta}}(d) \leq \max_{\theta\in\mathcal{B}(c_k,r_k)} \mathcal{S}_{X_g^\top\theta}(d) \leq \mathcal{S}_{X_g^\top\hat{\theta}}(d) + 2r_k \sup_{\|u\|=1} \mathcal{S}_{X_g^\top u}(d) \ ,$$

and the identification of the active set occurs when for all group g in $\mathcal{Z}$ and any direction d, we have $\mathcal{S}_{X_g^\top\hat{\theta}}(d) < \mathcal{S}_{\partial\Omega_g(\beta^\star_g)}(d)$. The latter holds as soon as[2]

$$r_k < \frac{1}{2} \min_{\substack{g\in\mathcal{Z} \\ d\in\operatorname{bd}\mathcal{B}(0,1)}} \frac{\mathcal{S}_{\partial\Omega_g(\beta^\star_g)}(d) - \mathcal{S}_{X_g^\top\hat{\theta}}(d)}{\sup_{\|u\|=1}\mathcal{S}_{X_g^\top u}(d)} =: \delta_{\mathcal{Z}} \ .$$

Whence, the identification of the active set using a safe ball of radius r_k occurs after k_0 iterations where

$$k_0 := \inf\{k \in \mathbb{N} : \ r_k < \delta_{\mathcal{Z}}\} \ . \tag{11}$$

Remark 4.5. (Non-degeneracy condition) By definition, the set $\mathcal{Z}$ is empty if $\delta_{\mathcal{Z}}$ is equal to zero. Thus all the complexity bounds are equal to infinity and then $\delta_{\mathcal{Z}} > 0$ is a necessary non-degeneracy condition to ensure finite identifications of the active set.

4.2. Duality gap certificates

Recently, a complexity analysis of the convergence of the duality gap, used as an optimality certificate as been proposed [7]. This analysis is important for deriving the complexity of active set identification that depends only on the rate of convergence of the algorithm. The next lemma adapts the proposed analysis to that take dual rescaling into account.

Lemma 4.6. *Let f be ν_f-smooth and Ω be μ_Ω-strongly convex ($\mu_\Omega = 0$ is allowed when Ω^* is subdifferentiable on its domain[3]). Since the dual vector θ in (9) is feasible, we can choose $u \in \partial\Omega^*(X^\top\theta) \neq \emptyset$. For all s in $[0, 1]$, it holds*

$$P(\beta) - P(\hat{\beta}) \geq s(\operatorname{Gap}(\beta, \theta) + \Delta(\alpha)) + s^2\left[\frac{(1-s)\mu_\Omega}{s}\|\beta - u\|^2 - \frac{\nu_f}{2}\|X(u-\beta)\|^2\right] \tag{12}$$

with $\Delta(\alpha) = f^*(\nabla f(X\beta)) - f^*(-\theta) + (\alpha - 1)\langle\theta, Xu\rangle$. The scaling α is defined in (10).

[2] We implicitly avoid the trivial case when there exists some group g with $\sup_{\|u\|=1}\mathcal{S}_{X_g^\top u}(d) = 0$.

[3] Note that in particular, the rescaled gradient mapping allows $\mu_\Omega = 0$ without restricting Ω to have a bounded support.

Proof. By optimality of $\hat{\beta}$, for any β and u in $\operatorname{dom} P$, we have:

$$\begin{aligned} P(\beta) - P(\hat{\beta}) &\geq P(\beta) - P(\beta + s(u-\beta)) \\ &= [\Omega(\beta) - \Omega(\beta + s(u-\beta))] + [f(X\beta) - f(X(\beta + s(u-\beta)))] \ . \end{aligned} \tag{13}$$

By the strong convexity of Ω, we have:

$$\Omega(\beta) - \Omega(\beta + s(u-\beta)) \geq s(\Omega(\beta) - \Omega(u)) + \frac{s(1-s)\mu_\Omega}{2} \|u-\beta\|^2 \ . \tag{14}$$

From the smoothness of f, we have:

$$f(X\beta) - f(X\beta + sX(u-\beta)) \geq s\langle \nabla f(X\beta), X(u-\beta)\rangle - \frac{s^2\nu_f}{2} \|X(u-\beta)\|^2 \ . \tag{15}$$

Then, plugging (14) and (15) to (13), yields:

$$P(\beta) - P(\hat{\beta}) \geq s\Gamma + \frac{s^2}{2}\left[\frac{(1-s)\mu_\Omega}{s} \|u-\beta\|^2 - \nu_f \|X(u-\beta)\|^2\right] \ ,$$

where $\Gamma = \Omega(\beta) - \Omega(u) - \langle \nabla f(X\beta), X(u-\beta)\rangle$.

The choice of the scaling α in (10), we have $X^\top\theta \in \operatorname{dom}\Omega^*$ which implies that $\partial\Omega^*(X^\top\theta)$ is non empty.

Thus we can choose $u \in \partial\Omega^*(X^\top\theta)$ which ensure that $u \in \operatorname{dom}\Omega$. Also, f is smooth if and only if f^* is strongly convex which implies that $\operatorname{dom} f$ is the whole space. Thus $Xu \in \operatorname{dom} f$. Whence $u \in \operatorname{dom}\Omega$ and $Xu \in \operatorname{dom} f$ implies $u \in \operatorname{dom} P$. Let $\beta \in \operatorname{dom} P$, the for any $s \in [0,1]$ on can check that $\beta + s(u-\beta) \in \operatorname{dom} P$.

For $u \in \partial\Omega^*(X^\top\theta)$, the equality case in the Fenchel-Young inequality reads:

$$\Omega(u) = \langle u, X^\top\theta\rangle - \Omega^*(X^\top\theta) \ .$$

Whence,

$$\begin{aligned} \Gamma &= \Omega(\beta) + \Omega^*(X^\top\theta) - \langle u, X^\top\theta\rangle - \langle \nabla f(X\beta), X(u-\beta)\rangle \\ &= \mathrm{Gap}(\beta,\theta) - f(X\beta) - f^*(-\theta) - \langle u, X^\top\theta\rangle - \langle \nabla f(X\beta), X(u-\beta)\rangle \ . \end{aligned}$$

From the equality case in the Fenchel-Young inequality, we obtain

$$f(X\beta) = \langle \nabla f(X\beta), X\beta\rangle - f^*(\nabla f(X\beta)).$$

Thanks to the last display and to the definition of θ, we get

$$-f(X\beta) - f^*(-\theta) - \langle u, X^\top\theta\rangle - \langle \nabla f(X\beta), X(u-\beta)\rangle = \Delta(\alpha),$$

hence the result. □

Let us denote the sub-optimality gap by

$$\mathcal{E}_k = P(\beta_k) - P(\hat{\beta}), \text{ for } k \in \mathbb{N} \ . \tag{16}$$

Cases where $\mu_\Omega > 0$. In such a case, $\operatorname{dom}\Omega^*$ is the whole dual space and we can choose $\alpha = 1$ (see Remark 3.6) whence $\Delta(\alpha) = 0$. Now choosing $s = \mu_\Omega/(\sigma_X\nu_f + \mu_\Omega)$ where σ_X is the spectral norm of the design matrix X (see also [7]), then the last term in (12) vanishes.

Thus, $$\frac{\mu_\Omega}{\sigma_X \nu_f + \mu_\Omega} \mathrm{Gap}(\beta_k, \theta_k) \leq \mathcal{E}_k \leq \mathrm{Gap}(\beta_k, \theta_k) \ .$$

This guarantees that the duality gap converges at the same rates as the sub-optimality gap. Along with (11), we obtain the following proposition.

Proposition 4.7. *For $\mu_\Omega > 0$ and any linearly converging primal algorithm i.e., with* $\mathrm{Rate}(k) = \exp(-\kappa k)$*, the active set will be identified after at most k_0 iterations where*

$$k_0 \leq \frac{1}{\kappa} \log \left(\frac{C_{f,\Omega,X}}{\delta_{\mathcal{Z}}^2} \frac{2}{\mu_D} \mathcal{E}_0 \right) \ ,$$

for some κ in $(0, 1]$ and the constant $C_{f,\Omega,X} := \frac{\sigma_X \nu_f + \mu_\Omega}{\mu_\Omega}$ depends only on the conditioning of the design matrix X and on the regularity of f and Ω.

Case where $\mu_\Omega = 0$. One possibility, here, is to modify Ω by adding a small strongly convex term (e.g. smoothing). Then, the previous result still holds for the modified problem. However, this will slightly modify the iterates of the algorithm. Otherwise, one can assume that Ω has a bounded support i.e., $\mathrm{dom}\,\Omega$ is included in a ball of radius L. In such a case, Ω^* is finite everywhere and we can still choose $\alpha = 1$ whence $\Delta(\alpha) = 0$ while having

$$\|X(u_k - \beta_k)\| \leq 2\sigma_X L \ . \tag{17}$$

Plugging it into Lemma 4.6, we obtain $\mathrm{Gap}(\beta_k, \theta_k) \leq \frac{1}{s}\mathcal{E}_k + 2\nu_f \sigma_X^2 L^2 s$.

Minimizing the upper bound in s onto $(0, 1]$, we have $\mathrm{Gap}(\beta_k, \theta_k) \leq \sqrt{8\nu_f \sigma_X^2 L^2 \mathcal{E}_k}$.

When the optimization algorithm converges linearly, the complexity in Prop. 4.7 is preserved up to some constants because the logarithmic term is not affected by the square-root. But, it leads to a suboptimal bound in the sub-linear regime,

Proposition 4.8. *For $\mu_\Omega = 0$ and any sub-linearly primal convergent algorithm i.e., with* $\mathrm{Rate}(k) = C/k^\gamma$ *where $\gamma > 0$, the active set will be identified after at most k_0 iterations where*

$$k_0 \leq \left(\frac{8\nu_f \sigma_X^2 L^2 C}{(\mu_D \delta_{\mathcal{Z}}^2)^2} \right)^{\frac{1}{\gamma}} \ .$$

To exactly match the rate of the algorithm (i.e., to remove the squared term (4.8)) we propose to additionally assume Lipschitz continuity of the sub-differential $\partial\Omega^*$ and a quadratic error bound on the objective function P. More precisely, we suppose that there exists some constants L_* and $\gamma_P > 0$ such that for a selection of u in $\partial\Omega^*(\zeta)$ and $\hat{u}$ in $\partial\Omega^*(\hat{\zeta})$, we have[4]:

$$\|u - \hat{u}\| \leq L_* \|\zeta - \hat{\zeta}\| \tag{18}$$

$$\frac{\gamma_P}{2} \|\beta - \hat{\beta}\|^2 \leq P(\beta) - P(\hat{\beta}) \ . \tag{19}$$

[4] This condition is required only for $\hat{\zeta} = -X^\top \nabla f(X\hat{\beta})$ and when $\zeta = \hat{\zeta}$, the choice restricts to $u = \hat{u} = \hat{\beta}$, which can be ensured by selecting u as the projection of β onto $\partial\Omega(X^\top\theta)$.

The reason is that u_k is expected to converge to $\hat{\beta}$ and so the bound in (17) may be too crude. In a sub-linear regime, the following lemma shows that assumptions (18) and (19) are sufficient conditions to improve the previous analysis in [7].

Remark 4.9. The quadratic error bound condition in (19) was proven to be satisfied for a large class of optimization problems. One can refer to [1] where it was used to analyze the complexity of first order optimization methods. Similar Lipschitz continuity assumptions on the subdifferential in (18) were made in [16], see also [37, Chapter 9.E]. However, it is not straightforward to explicitly compute these constants for practical applications.

Lemma 4.10. *Under assumptions* (18) *and* (19)*, for any integer k, it holds*

$$\mathrm{Gap}(\beta_k, \theta_k) \leq \sqrt{2\nu_f C'_{f,\Omega,X}}\, \mathcal{E}_k \ ,$$

where $C'_{f,\Omega,X} = \dfrac{4\sigma_X^2(L_*^2\sigma_X^2\nu_f^2+1)}{\gamma_P}$ *is non negative and finite.*

Proof. First note that $\|X(u_k-\beta_k)\|^2 \leq 2\sigma_X^2(\|u_k-\hat{\beta}\|^2 + \|\beta_k-\hat{\beta}\|^2)$.

When $\alpha = 1$, we have $u_k \in \partial\Omega^*(-X^\top\nabla f(\beta_k))$. Moreover, $\hat{\beta} \in \partial\Omega^*(-X^\top\nabla f(\hat{\beta}))$ and we have:

$$\begin{aligned}\|u_k-\hat{\beta}\| &\leq L_*\|X^\top\nabla f(X\beta_k) - X^\top\nabla f(X\hat{\beta})\| \leq L_*\sigma_X\nu_f\|\beta_k-\hat{\beta}\| \\ &\leq L_*\sigma_X\nu_f\sqrt{\frac{2}{\gamma_P}\mathcal{E}_k} \ ,\end{aligned}$$

where the first inequality results from the assumption (18), the second from the smoothness of f (f is ν_f smooth so the gradient is ν_f Lipschitz), and the third from the quadratic error bound (19).

Thus, for $C'_{f,\Omega,X} = \dfrac{4\sigma_X^2(L_*^2\sigma_X^2\nu_f^2+1)}{\gamma_P}$, we have $\|X(u_k-\beta_k)\|^2 \leq C'_{f,\Omega,X}\mathcal{E}_k$.

Plugging it into Lemma 4.6, we obtain

$$\mathcal{E}_k \geq P(\beta_k) - P(\hat{\beta}) \geq s\,\mathrm{Gap}(\beta,\theta) - s^2\frac{\nu_f C'_{f,\Omega,X}}{2}\mathcal{E}_k \ .$$

Whence $$\mathrm{Gap}(\beta_k,\theta_k) \leq \left(\frac{1}{s} + s\frac{\nu_f C'_{f,\Omega,X}}{2}\right)\mathcal{E}_k \ .$$

Minimizing the upper bound in s onto $(0,1]$, we obtain the result. □

Proposition 4.11. *For $\mu_\Omega = 0$, under assumptions* (18) (19) *and any sub-linearly primal convergent algorithm i.e., with* $\mathrm{Rate}(k) = C/R^\gamma$ *where $\gamma > 0$, the active set will be identified after at most k_0 iterations where*

$$k_0 \leq \left(\frac{\sqrt{8\nu_f C'_{f,\Omega,X}}C}{\mu_D\delta_{\mathcal{Z}}^2}\right)^{\frac{1}{\gamma}} \ .$$

Finally, when the domain of Ω is not bounded, the algorithm can be equipped with a modified duality gap which enforces the bounded domain assumption (this is known as the *Lipschitzing Trick* [7] in the litterature). Then, the previous result still holds without modifying the iterates of the algorithm.

Related works. To our knowledge, this paper is the first one to discuss the complexity of active set identification with screening rules. Our results match the existing results on active set identification in [20, 31, 42] for proximal algorithms. Interestingly, our result uniformly holds for any converging algorithm not only proximal methods and illustrates the benefits obtained as screening rules explicitly and definitely eliminate non-active variables along the algorithmic progress.

5. Acceleration Strategies

We discuss some practical methods for efficiently using screening rules to speed up optimization processes for solving (2) and show how some popular previous acceleration heuristics such as *strong rules* [44] or recent *working sets* [14, 23] can be extended in our framework.

Static (Pre-processing). A natural strategy is to set, once for all, a gap safe radius using some initial fixed vectors $\theta = \theta_0$ and $\beta = \beta_0$. The resulting static safe region $\mathcal{B}\left(\theta_0, \sqrt{\frac{2}{\mu_D}\operatorname{Gap}(\beta_0, \theta_0)}\right)$ is used in (6). Such a strategy is only efficient when (β_0, θ_0) are good enough estimate of the optimal solutions, and have limited scope in practice.

Dynamic. One could rather use the information gained during an optimization process to obtain a smaller safe region therefore a greater elimination of inactive variables. Whence, we consider $\mathcal{B}\left(\theta_k, \sqrt{\frac{2}{\mu_D}\operatorname{Gap}(\beta_k, \theta_k}\right)$. Dynamic safe region was initially suggested in [2] and further used in the duality gap based region in [11, 18, 29, 40].

Sequential (Homotopy Continuation). Sequential screening is motivated by the intuition that, often, the duality gap grows continuously with respect to the regularization parameter [12, 30].

It basically states that when λ close to λ_t, the duality gap $\operatorname{Gap}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ tends to $\operatorname{Gap}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$. As a by product, given a sufficiently fine grid of parameter $(\lambda_t)_{t\in[T]}$, the sequential screenings based on balls $\mathcal{B}\left(\theta^{(\lambda_t)}, \sqrt{\frac{2}{\mu_D}\operatorname{Gap}_{\lambda_{t-1}}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})}\right)$, will be small enough to efficiently remove non active variables.

Active Warm Start (aka strong rules). This method was introduced in [44] as a heuristic relaxation of the safe rules to discard features more aggressively in ℓ_1 regularized optimization problem. We generalize it into our framework. Let $F(\beta) = f(X\beta)$. We have for any $d \in \mathbb{R}^{|g|}$

$$\begin{aligned}\mathcal{S}_{X_g^\top \hat{\theta}^{(\lambda)}}(d) &= \mathcal{S}_{X_g^\top \hat{\theta}^{(\lambda')}}(d) + \mathcal{S}_{\{X_g^\top \hat{\theta}^{(\lambda)} - X_g^\top \hat{\theta}^{(\lambda')}\}}(d) \\ &= \mathcal{S}_{X_g^\top \hat{\theta}^{(\lambda')}}(d) + \mathcal{S}_{\{\nabla_g F(\hat{\beta}^{(\lambda')}) - \nabla_g F(\hat{\beta}^{(\lambda)})\}}(d) \ .\end{aligned}$$

If ∇F is group-wise non-expansive along the regularization path i.e.,

$$\|\nabla_g F(\hat{\beta}^{(\lambda')}) - \nabla_g F(\hat{\beta}^{(\lambda)})\| \leq |\lambda' - \lambda|,$$

the screening holds whenever the *(generalized) strong rule* holds:

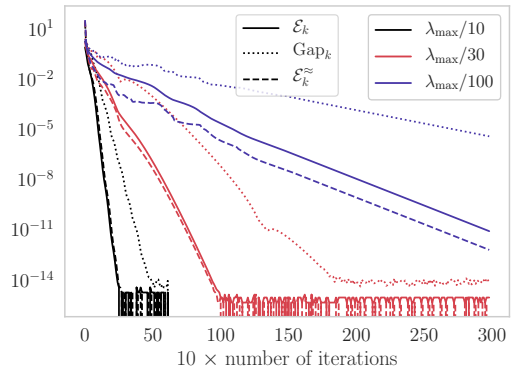
$$\mathcal{S}_{X_g^\top \hat{\theta}^{(\lambda')}}(d) + |\lambda' - \lambda| < \mathcal{S}_{\partial\Omega_g(\beta_g^\star)}(d) \ .$$

The strong rules are un-safe because the non-expansiveness condition on ∇F is usually not satisfied without stronger assumptions on the design matrix X (e.g. X has full column rank and $(X^\top X)^{-1}$ is diagonally dominant). Moreover, the *exact* solution $\hat{\theta}^{(\lambda')}$ is usually not available.

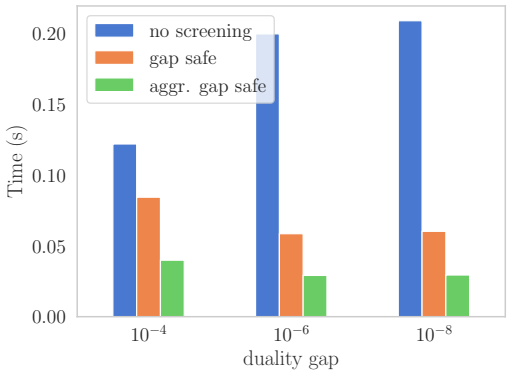
As a simpler rule, specially when the previous regularity condition cannot be verified, we rather suggest to use the previous active set

$$\mathcal{S}_{X_g^\top \hat{\theta}^{(\lambda')}}(d) < \mathcal{S}_{\partial\Omega_g(\beta_g^\star)}(d) \ .$$

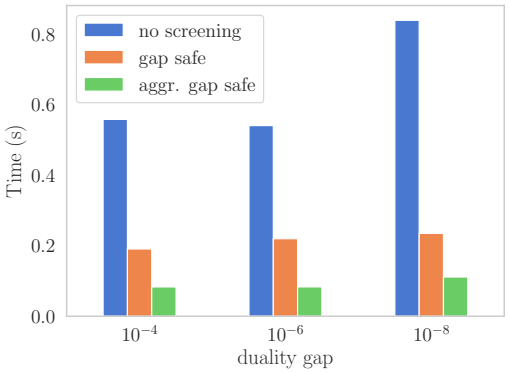
The rational behind these heuristics is that, often, the active set is stable along the regularization path, a crucial argument used to build the `Lars` algorithm [8] and variants [33].



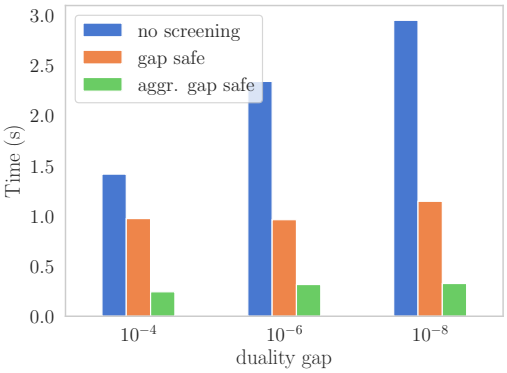
(a) Convergence rate.



(b) Times for $\lambda = \lambda_{\max}/10$.



(c) Times for $\lambda = \lambda_{\max}/30$.



(d) Times for $\lambda = \lambda_{\max}/100$.

Figure 5.1: Illustrations on Lasso using (cyclic) coordinate descent on Leukemia dataset ($n = 72$ observations and $p = 7129$ features. Here $\lambda_{\max} = \|X^\top y\|_\infty$ is the smallest λ such that $\hat{\beta} = 0$ is a primal optimal solution.

Aggressive Active Warm Start. The gap safe screening rule relies on an upper estimate the suboptimal gap by the duality gap $\mathrm{Gap}(\beta_k, \theta_k)$. This can be conservative for the screening rules since no false elimination is allowed. Here we suggest a new heuristic in order to remove more variables at an early stage of an optimization

process. At any iteration k, use $\mathcal{E}_k^{\approx} = |P(\beta_{k-s}) - P(\beta_k)|$ as an unsafe estimate of the suboptimal gap. This will eliminate more variables depending on the choice of the delaying parameter s. In practice, we delay β_k and β_{k-s} with 10 epochs for instance for the Lasso case, when using coordinate descent as a solver. To avoid a severe underestimation, one can instead use $(1-\eta)\mathcal{E}_k^{\approx} + \eta\,\mathrm{Gap}(\beta_k, \theta_k)$. We set a default value $\eta = 10^{-3}$. See the numerical illustrations in Figure 5.1 and the Appendix.

Remark 5.1. Since these rules are unsafe i.e., they can wrongly remove some variables, they must be accompanied with a post-precessing step. For instance by adding back the variables that violates the KKT conditions. We rather suggest to use the solution obtained in these steps as a warm start for the dynamic safe rules with a converging algorithm. In this way, a low computational complexity can be maintained when passing over the entire problem with a better initialization of gap safe screening rules.

Working Sets. Following the suggestions made in [14, 24], one can consider, for any group g in $\mathcal{G}$

$$d_g(\theta) = \frac{\mathcal{S}_{\partial\Omega_g(\beta_g^\star)}(d) - \mathcal{S}_{X_g^\top\theta}(d)}{\sup_{\|u\|=1} \mathcal{S}_{X_g^\top u}(d)}\ , \tag{20}$$

as a measure of the importance of feature X_g. Thus, one can design a working set i.e., a set of group g in which to restrict the optimization problem, by selecting the groups that have a higher value $d_g(\theta)$. These methods fit naturally in our framework.

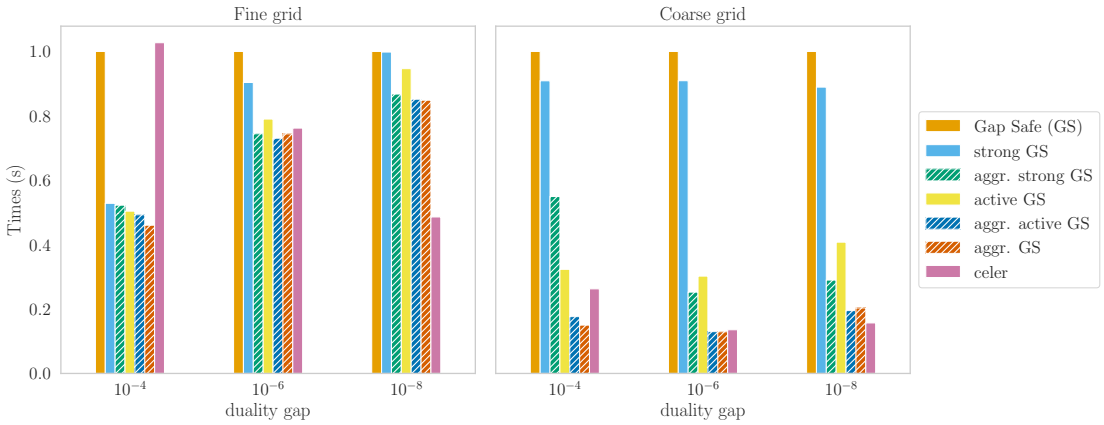


Figure 5.2: Lasso on the `Leukemia` (dense data with $n = 72$ observations and $p = 7129$ features). Computation times needed to solve the Lasso regression path to desired accuracy for a grid of λ from $\lambda_{\max} = \left\|X^\top y\right\|_\infty$ to $\lambda_{\max}/100$. The size of the dense grid (resp. sparse grid) is 100 (resp. 10).

6. Numerical Experiments

We consider simple examples to illustrate the performance of different acceleration strategies with screening rules on the Lasso problem with real datasets. We use a cyclic coordinate descent solver[5] as a shared standard algorithm for all methods. All methods are stopped when the duality gap reaches a prescribed tolerance $\epsilon \|y\|^2$ where ϵ is set to 10^{-4}, 10^{-6} or 10^{-8}. For readability, the execution times of the

[5] The implementation is available at `https://github.com/EugeneNdiaye/Gap_Safe_Rules`.

algorithms are normalized with respect to the running time of coordinate descent with the gap safe screening rule baseline as done in [29]. Evaluations of the performance of safes rules for other problems such as logistic regression, Sparse-Group Lasso, SVM etc are available in the literature, e.g. [27, 28, 40].

Although safe rules can save a significant amount of computational time, they should be conservative so as not to wrong eliminate relevant variables. In our numerical experiments, we observe that this constraint can limit their efficiency. By reducing this safety constraint, one can greatly improve their efficiency by combining them with a simple heuristic like the one introduced in Section 5.

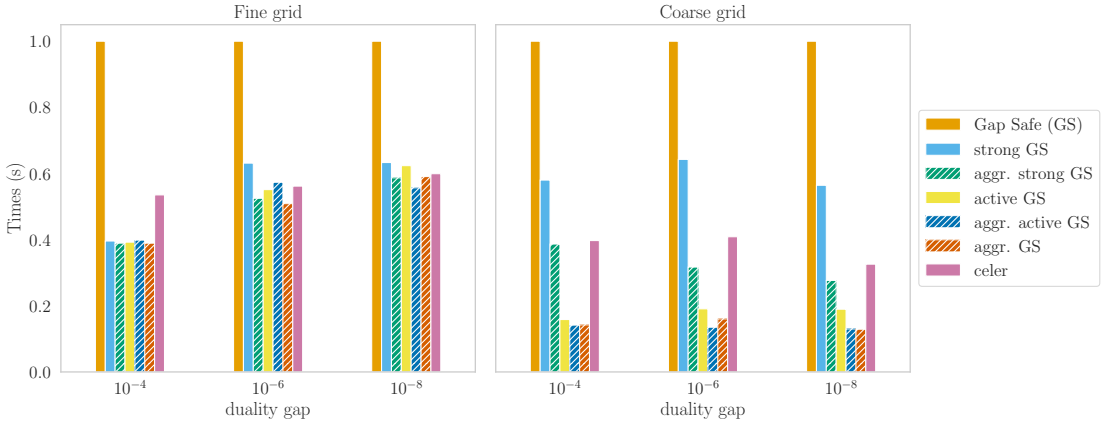


Figure 6.1: Lasso on the `climate NCEP/NCAR Reanalysis 1` (dense data with $n = 814$ observations and $p = 73570$ features) see [17]. Computation times needed to solve the Lasso regression path to desired accuracy for a grid of λ from $\lambda_{\max} = \left\|X^\top y\right\|_\infty$ to $\lambda_{\max}/100$. The size of the dense grid (resp. sparse grid) is 100 (resp. 10)

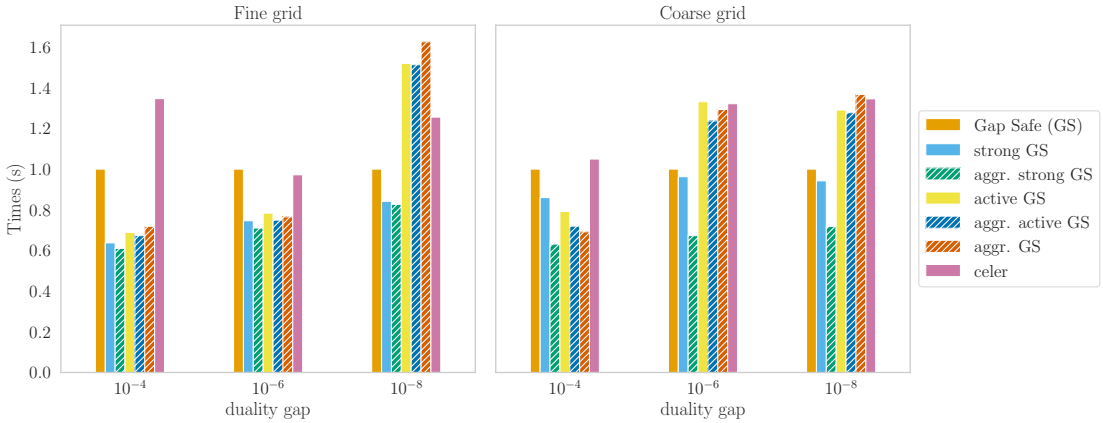


Figure 6.2: Lasso on the `rcv1_train` (sparse data with $n = 20242$ observations and $p = 19960$ features) available in libsvm. Computation times needed to solve the Lasso regression path to desired accuracy for a grid of λ from $\lambda_{\max} = \left\|X^\top y\right\|_\infty$ to $\lambda_{\max}/100$. The size of the dense grid (resp. sparse grid) is 100 (resp. 10).

7. Conclusion

We have presented a simple way to unify various contributions that explicitly identify active variables, especially in sparse regression problems. For this, we have relied on optimality conditions and the fact that the subdifferentials of a function evaluated at

two distinct points can not be overlapped. It should be noted that this remarkable property is not limited to convex functions (e.g. it holds for non-convex setting as soon as the set (1) is non empty).

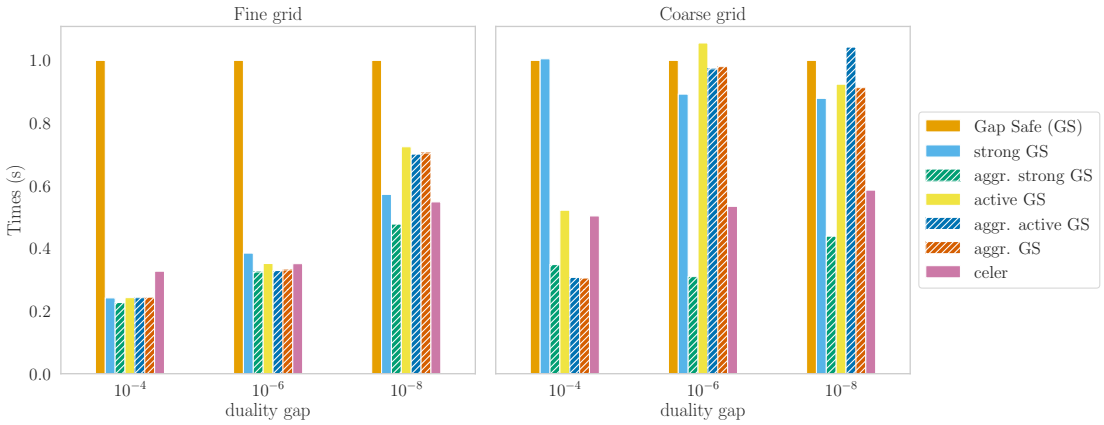


Figure 7.1: Lasso on the `news20` (sparse data with $n = 19996$ observations and $p = 632983$ features) available in libsvm. Computation times needed to solve the Lasso regression path to desired accuracy for a grid of λ from $\lambda_{\max} = \left\|X^\top y\right\|_\infty$ to $\lambda_{\max}/100$. The size of the dense grid (resp. sparse grid) is 100 (resp. 10).

Extending the identification rules to subdifferential in the sense of Fréchet or Clarke would be a natural venue for future works. Promising results have been shown in [19, 35]. However, it is still open to get a unified framework for non convex optimization problems and non separable regularization function.

When an optimization algorithm can benefit from screening rules, we have also shown that the number of iterations to identify the active set can be accurately estimated, and depends only on the rate of convergence of the (converging) algorithm used. Numerical experiments of some heuristic acceleration rules have been provided, showing their interest for (block) coordinate descent algorithms.

References

[1] J. Bolte, T. P. Nguyen, J. Peypouquet, B. W. Suter: *From error bounds to the complexity of first-order descent methods for convex functions*, Math. Programming 165 (2017) 471-–507.

[2] A. Bonnefoy, V. Emiya, L. Ralaivola, R. Gribonval: *A dynamic screening principle for the lasso*, EUSIPCO (2014) 6–10.

[3] J. M. Borwein, H. Wolkowicz: *Facial reduction for a cone-convex programming problem*, J. Australian Math. Soc. 30/3 (1981) 369–380.

[4] A. L. Brearley, G. Mitra, H. P. Williams: *Analysis of mathematical programming problems prior to applying the simplex algorithm*, Math. Programming 8/1 (1975) 54–83.

[5] L. Condat: *Fast projection onto the simplex and the l_1 ball*, Math. Programming A 158/1-2 (2016) 575–585.

[6] D. Drusvyatskiy, H. Wolkowicz: *The many faces of degeneracy in conic optimization*, Found. Trends Optimization 3/2 (2017) 77–170.

[7] C. Dünner, S. Forte, M. Takáč, M. Jaggi: *Primal-dual rates and certificates*, Proc. Machine Learning Research 48 (2016) 783–792.

[8] B. Efron, T. J. Hastie, I. M. Johnstone, R. Tibshirani: *Least angle regression*, Ann. Statistics 32/2 (2004) 407–499.

[9] L. El Ghaoui, V. Viallon, T. Rabbani: *Safe feature elimination in sparse supervised learning*, J. Pacific Optimimization 8/4 (2012) 667–698.

[10] J. Fan, J. Lv: *Sure independence screening for ultrahigh dimensional feature space*, J. R. Stat. Soc. Ser. B Stat. Method. 70/5 (2008) 849–911.

[11] O. Fercoq, A. Gramfort, J. Salmon: *Mind the duality gap: safer rules for the lasso*, Proc. Machine Learning Research 37 (2015) 333–342.

[12] J. Giesen, J. K. Müller, S. Laue, S. Swiercy: *Approximating concavely parameterized optimization problems*, Neural Inform. Processing Systems 25 (2012) 2105–2113.

[13] J-B. Hiriart-Urruty, C. Lemaréchal: *Fundamentals of Convex Analysis*, Springer, Berlin (2001).

[14] T. B. Johnson, C. Guestrin: *Blitz: A principled meta-algorithm for scaling sparse optimization*, Proc. Machine Learning Research 37 (2015) 1171–1179.

[15] T. B. Johnson, C. Guestrin: *Unified methods for exploiting piecewise linear structure in convex optimization*, Neural Inform. Processing Systems 29 (2016) 4754–4762.

[16] A. Jourani, L. Thibault, D. Zagrodny: *$C^{1,\omega(\cdot)}$-regularity and Lipschitz-like properties of subdifferential*, Proc. Lond. Math. Soc. (3) 105/1 (2012) 189–223.

[17] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, D. Joseph: *The NCEP/NCAR 40-year reanalysis project*, Bull. Amer. Meteor. Soc. 77/3 (1996) 437–472.

[18] M. Le Morvan, J.-P. Vert: *WHinter: A working set algorithm for high-dimensional sparse second order interaction models*, Proc. Machine Learning Research 80 (2018) 3632–3641.

[19] S. Lee, P. Breheny: *Strong rules for nonconvex penalties and their implications for efficient algorithms in high-dimensional regression*, J. Comp. Graph. Statistics 24/4 (2015) 1074–1091.

[20] J. Liang, J. Fadili, G. Peyré: *Activity identification and local linear convergence of forward-backward-type methods*, SIAM J. Optimization 27/1 (2017) 408–437.

[21] J. Mairal: *Sparse Coding for Machine Learning, Image Processing and Computer Vision*, PhD Thesis, École Normale Supérieure de Cachan (2010).

[22] H. Markowitz: *The optimization of a quadratic function subject to linear constraints*, Naval Res. Logist. Quart. 3 (1956) 111–133.

[23] M. Massias, A. Gramfort, J. Salmon: *Celer: a Fast Solver for the Lasso with Dual Extrapolation*, Proc. Machine Learning Research 80 (2018) 3315–3324.

[24] M. Massias, S. Vaiter, A. Gramfort, J. Salmon: *Dual extrapolation for sparse GLMs*, J. Machine Learning Research 21/234 (2020) 1–33.

[25] C. Mészáros, U. H. Suhl: *Advanced preprocessing techniques for linear and quadratic programming*, OR Spectrum 25/4 (2003) 575–595.

[26] C. Michelot: *A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$*, J. Optimization Theory Appl. 50/1 (1986) 195–200.

[27] E. Ndiaye, O. Fercoq, A. Gramfort, J. Salmon: *GAP safe screening rules for sparse multi-task and multi-class models*, in: NIPS 2015, Proc. 28th Int. Conf. Neural Inform. Processing Systems 1 (2015) 811–819.

[28] E. Ndiaye, O. Fercoq, A. Gramfort, J. Salmon: *GAP safe screening rules for Sparse-Group Lasso*, in: NIPS 2016, Advances in Neural Inform. Processing Systems 29 (2016) 1–9.

[29] E. Ndiaye, O. Fercoq, A. Gramfort, J. Salmon: *Gap safe screening rules for sparsity enforcing penalties*, J. Machine Learning Research 18/128 (2017) 1–33.

[30] E. Ndiaye, T. Le, O. Fercoq, J. Salmon, I. Takeuchi: *Safe grid search with optimal complexity*, Proc. Machine Learning Research 97 (2019) 4771–4780.

[31] J. Nutini, I. Laradji, M. Schmidt: *Let's make block coordinate descent go fast: Faster greedy rules, message-passing, active-set complexity, and superlinear convergence*, arXiv: 1712.08859 (2017).

[32] K. Ogawa, Y. Suzuki, I. Takeuchi: *Safe screening of non-support vectors in pathwise SVM computation*, Proc. Machine Learning Research 28 (2013) 1382–1390.

[33] M. R. Osborne, B. Presnell, B. A. Turlach: *A new approach to variable selection in least squares problems*, IMA J. Numer. Analysis 20/3 (2000) 389–403.

[34] A. Raj, J. Olbrich, B. Gärtner, B. Schölkopf, M. Jaggi: *Screening rules for convex problems*, arXiv: 1609.07478 (2016).

[35] A. Rakotomamonjy, G. Gasso, J. Salmon: *Screening rules for lasso with non-convex sparse regularizers*, Proc. Machine Learning Research 97 (2019) 5341–5350.

[36] R. T. Rockafellar: *Convex Analysis*, Princeton University Press, Princeton (1997).

[37] R. T. Rockafellar, R. Wets: *Variational Analysis*, Grundlehren der mathematischen Wissenschaften 317, Springer, Berlin (2009).

[38] S. Shalev-Shwartz. *Online learning and online convex optimization*, Found. Trends Machine Learning 4/2 (2012) 107–194.

[39] S. Shalev-Shwartz, S. Ben-David: *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, Cambridge (2014).

[40] A. Shibagaki, M. Karasuyama, K. Hatano, I. Takeuchi: *Simultaneous safe screening of features and samples in doubly sparse modeling*, Proc. Machine Learning Research 48 (2016) 1577–1586.

[41] N. Simon, J. Friedman, T. Hastie, R. Tibshirani: *A sparse-group lasso*, J. Comput. Graph. Statistics 22/2 (2013) 231–245.

[42] Y. Sun, H. Jeong, J. Nutini, M. Schmidt: *Are we there yet? Manifold identification of gradient-related proximal methods*, Proc. Machine Learning Research 89 (2019) 1110–1119.

[43] R. Tibshirani: *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B Stat. Method. 58/1 (1996) 267–288.

[44] R. Tibshirani, J. Bien, J. Friedman, T. J. Hastie, N. Simon, J. Taylor, R. J. Tibshirani: *Strong rules for discarding predictors in lasso-type problems*, J. R. Stat. Soc. Ser. B Stat. Method. 74/2 (2012) 245–266.

[45] J. Wang, J. Zhou, J. Liu, P. Wonka, J. Ye: *A safe screening rule for sparse logistic regression*, in: NIPS 2014, Proc. 27th Int. Conf. Neural Inform. Processing Systems 1 (2014) 1053–1061.

[46] Z. J. Xiang, H. Xu, P. J. Ramadge: *Learning sparse representations of high dimensional data on large scale dictionaries*, in: NIPS 2011 (2011) 900–908.

[47] M. Yuan, Y. Lin: *Model selection and estimation in regression with grouped variables*, J. R. Stat. Soc. Ser. B Stat. Method. 68/1 (2006).