SAFE RULES FOR THE IDENTIFICATION OF ZEROS IN THE SOLUTIONS OF THE SLOPE PROBLEM*

CLÉMENT ELVIRA† AND CÉDRIC HERZET‡

Abstract. In this paper we propose a methodology to accelerate the resolution of the so-called "Sorted L-One Penalized Estimation" (SLOPE) problem. Our method leverages the concept of "safe screening", well-studied in the literature for *group-separable* sparsity-inducing norms, and aims at identifying the zeros in the solution of SLOPE. More specifically, we derive a set of $\frac{n(n+1)}{2}$ inequalities for each element of the n-dimensional primal vector and prove that the latter can be safely screened if some subsets of these inequalities are verified. We propose moreover an efficient algorithm to jointly apply the proposed procedure to all the primal variables. Our procedure has a complexity $\mathcal{O}(n \log n + LT)$ where $T \leq n$ is a problem-dependent constant and L is the number of zeros identified by the test. Numerical experiments confirm that, for a prescribed computational budget, the proposed methodology leads to significant improvements of the solving precision.

Key words. SLOPE, safe screening, acceleration techniques, convex optimization

AMS subject classifications. 68Q25, 68U05

1. Introduction.

During the last decades, sparse linear regression has attracted much attention in the field of statistics, machine learning and inverse problems. It consists in finding an approximation of some input vector $\mathbf{y} \in \mathbb{R}^m$ as the linear combination of a few columns of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ (often called dictionary). Unfortunately, the general form of this problem is NP-hard and convex relaxations have been proposed in the literature to circumvent this issue. The most popular instance of convex relaxation for sparse linear regression is undoubtedly the so-called "LASSO" problem where the coefficients of the regression are penalized by an ℓ_1 norm, see [11]. Generalized versions of LASSO have also been introduced to account for some possible structure in the pattern of the nonzero coefficients of the regression, see [2].

In this paper, we focus on the following generalization of LASSO:

$$\min_{\mathbf{x} \in \mathbb{R}^n} P(\mathbf{x}) \triangleq \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\, r_{\text{SLOPE}}(\mathbf{x}), \quad \lambda > 0 \tag{1.1}$$

where

$$r_{\text{SLOPE}}(\mathbf{x}) \triangleq \sum_{k=1}^{n} \gamma_k |\mathbf{x}|_{[k]} \tag{1.2}$$

with

$$\gamma_1 > 0, \qquad \gamma_1 \geq \cdots \geq \gamma_n \geq 0, \tag{1.3}$$

and $|\mathbf{x}|_{[k]}$ is the kth largest element of $\mathbf{x}$ in absolute value, that is

$$\forall \mathbf{x} \in \mathbb{R}^n : \ |\mathbf{x}|_{[1]} \geq |\mathbf{x}|_{[2]} \geq \ldots \geq |\mathbf{x}|_{[n]}. \tag{1.4}$$

---

* Submitted to the editors DATE.
The research presented in this paper is reproducible. Code and data are available at https://gitlab-research.centralesupelec.fr/2020elvirac/slope-screening

† IETR UMR CNRS 6164, CentraleSupelec Rennes Campus, 35576 Cesson Sévigné, France (clement.elvira@centralesupelec.fr, https://c-elvira.github.io/).

‡ Inria centre Rennes - Bretagne Atlantique, Rennes, France (cedric.herzet@inria.fr, http://people.rennes.inria.fr/Cedric.Herzet/).

Problem (1.1) is commonly referred to as *"Sorted L-One Penalized Estimation" (SLOPE)* or *"Ordered Weighted L-One Linear Regression"* in the literature and has been introduced in two parallel works [5, 49].[1] The first instance of a problem of the form (1.1) (for some nontrivial choice of the parameters γ_k's) is due to Bondell and Reich in [7]. The authors considered a problem similar to (1.1), named "Octagonal Shrinkage and Clustering Algorithm for Regression" (OSCAR), where the regularization function is a linear combination of an ℓ_1 norm and a sum of pairwise ℓ_∞ norms of the elements of $\mathbf{x}$, that is

$$r_{\text{OSCAR}}(\mathbf{x}) = \beta_1\|\mathbf{x}\|_1 + \beta_2 \sum_{j'>j} \max(|\mathbf{x}_{(j')}|, |\mathbf{x}_{(j)}|), \tag{1.5}$$

for some $\beta_1 \in \mathbb{R}_+^*$, $\beta_2 \in \mathbb{R}_+$. It is not difficult to see that r_{OSCAR} can be expressed as a particular case of r_{SLOPE} with the following choice $\gamma_k = \beta_1 + \beta_2(n-k)$. We note that some authors have recently considered "group" versions of the SLOPE problem where the ordered ℓ_2 norm of subsets of $\mathbf{x}$ is penalized by a decreasing sequence of parameters γ_k, see *e.g.*, [9, 25, 26].

SLOPE enjoys several desirable properties which have attracted many researchers during the last decade. First, it was shown in several works that, for some proper choices of parameters γ_k's, SLOPE promotes *sparse* solutions with some form of *"clustering"*[2] of the nonzero coefficients, see *e.g.*, [7, 21, 30, 39]. This feature has been exploited in many application domains: portfolio optimization [31, 47], genetics [26], magnetic-resonance imaging [16], subspace clustering [38], deep neural networks [50], etc. Moreover, it has been pointed out in a series of works that SLOPE has very good statistical properties: it leads to an improvement of the false detection rate (as compared to LASSO) for moderately-correlated dictionaries [6, 25] and is minimax optimal in some asymptotic regimes, see [33, 40].

Another desirable feature of SLOPE is its convexity. In particular, it was shown in [6, Proposition 1.1] and [48, Lemma 2] that r_{SLOPE} is a norm as soon as (1.3) holds. As a consequence, several numerical procedures have been proposed in the literature to find the global minimizer(s) of problem (1.1). In [6] and [51], the authors considered an accelerated gradient proximal implementation for SLOPE and OSCAR, respectively. In [31], the authors tackled problem (1.1) via an alternating-direction method of multipliers [8]. An approach based on an augmented Lagrangian method was considered in [35]. In [48], the authors expressed r_{SLOPE} as an atomic norm and particularized a Frank-Wolfe minimization procedure [23] to problem (1.1). An efficient algorithm to compute the Euclidean projection onto the unit ball of the SLOPE norm was provided in [14]. Finally, in [10] a heuristic "message-passing" method was proposed.

In this paper, we introduce a new "safe screening" procedure to accelerate the resolution of SLOPE. The concept of "safe screening" is well known in the LASSO literature: it consists in performing simple tests to identify the zero elements of the minimizers; this knowledge can then be exploited to reduce the problem dimensionality by discarding the columns of the dictionary weighted by the zero coefficients. Safe screening for LASSO has been first introduced by El Ghaoui *et al.* in the seminal paper [24] and extended to *group-separable* sparsity-inducing norm in [36]. Safe screening has rapidly been recognized as a simple yet effective procedure to accelerate the resolution of LASSO, see *e.g.*, [12, 20, 27–29, 34, 42, 43, 45]. The term "safe" refers to

[1] We will stick to the former denomination in the following.
[2] More specifically, groups of nonzero coefficients tend to take on the same value.

the fact that all the elements identified by a safe screening procedure are theoretically guaranteed to correspond to zeros of the minimizers. In contrast, *unsafe* versions of screening for LASSO (often called "strong screening rules") also exist, see [41]. More recently, screening methodologies have been extended to detect saturated components in different convex optimization problems, see [17, 18].

In this paper, we derive *safe* screening rules for SLOPE and emphasize that their implementation enables significant improvements of the solving precision when addressing SLOPE with a prescribed computational budget. We note that the SLOPE norm is not group-separable and the methodology proposed in [36] does therefore not trivially apply here. Prior to this work, we identified two contributions addressing screening for SLOPE. In [32], the authors proposed an extension of the *strong* screening rules derived in [41] to the SLOPE problem. In [3], the authors suggested a simple test to identify some zeros of the SLOPE solutions. Although the derivations made by these authors have been shown to contain several technical flaws [19], their test can be cast as a particular case of our result in Theorem 4.3 (and is therefore quite unexpectedly safe).

The paper is organized as follows. We introduce the notational conventions used throughout the paper in Section 2 and recall the main concepts of safe screening for LASSO in Section 3. Section 4 contains our proposed safe screening rules for SLOPE. Section 5 illustrates the effectiveness of the proposed approach through numerical simulations. All technical details and mathematical derivations are postponed to Appendices A and B.

2. Notations. Unless otherwise specified, we will use the following conventions throughout the paper. Vectors are denoted by lowercase bold letters (*e.g.*, $\mathbf{x}$) and matrices by uppercase bold letters (*e.g.*, $\mathbf{A}$). The "all-zero" vector of dimension n is written $\mathbf{0}_n$. We use symbol $^{\mathrm{T}}$ to denote the transpose of a vector or a matrix. $\mathbf{x}_{(j)}$ refers to the jth component of $\mathbf{x}$. When referring to the sorted entries of a vector, we use bracket subscripts; more precisely, the notation $\mathbf{x}_{[k]}$ refers to the kth largest value of $\mathbf{x}$. For matrices, we use $\mathbf{a}_j$ to denote the jth column of $\mathbf{A}$. We use the notation $|\mathbf{x}|$ to denote the vector made up of the absolute value of the components of $\mathbf{x}$. The sign function is defined for all scalars x as $\mathrm{sign}\,(x) = x/|x|$ with the convention $\mathrm{sign}\,(x) = 0$. [CE: sign (0) non?] Calligraphic letters are used to denote sets (*e.g.*, $\mathcal{J}$) and $\mathrm{card}\,(\cdot)$ refers to their cardinality. If $a < b$ are two integers, $[\![a, b]\!]$ is used as a shorthand notation for the set $\{a, a+1, \ldots, b\}$. Given a vector $\mathbf{x} \in \mathbb{R}^n$ and a set of indices $\mathcal{J} \subseteq [\![1, n]\!]$, we let $\mathbf{x}_{\mathcal{J}}$ be the vector of components of $\mathbf{x}$ with indices in $\mathcal{J}$. Similarly, $\mathbf{A}_{\mathcal{J}}$ denotes the submatrix of $\mathbf{A}$ whose columns have indices in $\mathcal{J}$. $\mathbf{A}_{\backslash \ell}$ corresponds to matrix $\mathbf{A}$ deprived of its ℓth column.

3. Screening: main concepts. "Safe screening" has been introduced by El Ghaoui *et al.* in [24] for ℓ_1-penalized problems:

$$\min_{\mathbf{x} \in \mathbb{R}^n} P(\mathbf{x}) \triangleq f(\mathbf{A}\mathbf{x}) + \lambda \|\mathbf{x}\|_1, \quad \lambda > 0 \tag{3.1}$$

where $f \colon \mathbb{R}^m \to \mathbb{R}$ is a closed convex function. It is grounded on the following ideas.

First, it is well-known that ℓ_1-regularization favors sparsity of the minimizers of (3.1). For instance, if $f = \frac{1}{2}\|\cdot\|_2^2$ and the solution of (3.1) is unique, it can be shown that the minimizer contains at most m nonzero coefficients, see *e.g.*, [22, Theorem 3.1]. Second, if some zeros of the minimizers are identified, (3.1) can be shown to be equivalent to a problem of *reduced* dimension. More precisely, let $\mathcal{L} \subseteq [\![1, n]\!]$ be a set of indices such that we have for any minimizer $\mathbf{x}^\star$ of (3.1):

$$\forall \ell \in \mathcal{L}: \; \mathbf{x}^\star_{(\ell)} = 0 \tag{3.2}$$

and let $\bar{\mathcal{L}} = [\![1, n]\!] \backslash \mathcal{L}$. Then the following problem

$$\min_{\mathbf{z} \in \mathbb{R}^{\mathrm{card}(\bar{\mathcal{L}})}} f(\mathbf{A}_{\bar{\mathcal{L}}}\mathbf{z}) + \lambda \|\mathbf{z}\|_1, \quad \lambda > 0 \tag{3.3}$$

admits the same optimal value as (3.1) and there exists a simple bijection between the minimizers of (3.1) and (3.3). We note that $\mathbf{x}$ belongs to an n-dimensional space whereas $\mathbf{z}$ is a $\mathrm{card}(\bar{\mathcal{L}})$-dimensional vector. Hence, solving (3.3) rather than (3.1) may lead to dramatic memory and computational savings if $\mathrm{card}(\bar{\mathcal{L}}) \ll n$.

The crux of screening consists therefore in identifying (some) zeros of the minimizers of (3.1) with marginal cost. El Ghaoui *et al.* emphasized that this is possible by relaxing some primal-dual optimality condition of problem (3.1). More precisely, let

$$\mathbf{u}^\star \in \operatorname*{arg\,max}_{\mathbf{u} \in \mathbb{R}^m} \; D(\mathbf{u}) \triangleq -f^*(-\mathbf{u}) \quad \text{s.t. } \|\mathbf{A}^{\mathrm{T}}\mathbf{u}\|_\infty \leq \lambda \tag{3.4}$$

be the dual problem of (3.1), where f^* denotes the Fenchel conjugate. Then, by complementary slackness, we must have for any minimizer $\mathbf{x}^\star$ of (3.1):

$$\forall \ell \in [\![1, n]\!]: \; (|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star| - \lambda)\, \mathbf{x}^\star_{(\ell)} = 0. \tag{3.5}$$

Since dual feasibility imposes that $|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star| \leq \lambda$, we obtain the following implication:

$$|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star| < \lambda \implies \mathbf{x}^\star_{(\ell)} = 0. \tag{3.6}$$

Hence, if $\mathbf{u}^\star$ is available, the left-hand side of (3.6) can be used to detect if the ℓth component of $\mathbf{x}^\star$ is equal to zero.

Unfortunately, finding a maximizer of dual problem (3.4) is generally as difficult as solving primal problem (3.1). This issue can nevertheless be circumvented by identifying some region $\mathcal{R}$ of the dual space (commonly referred to as *"safe region"*) such that $\mathbf{u}^\star \in \mathcal{R}$. Indeed, since

$$\max_{\mathbf{u} \in \mathcal{R}} |\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}| < \lambda \implies |\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star| < \lambda, \tag{3.7}$$

the left-hand side of (3.7) constitutes an alternative (weaker) test to detect the zeros of $\mathbf{x}^\star$. For proper choices of $\mathcal{R}$, the maximization over $\mathbf{u}$ admits a simple analytical solution. For example, if $\mathcal{R}$ is a ball, that is

$$\mathcal{R} = \mathcal{S}(\mathbf{c}, R) \triangleq \{\mathbf{u} \in \mathbb{R}^m : \|\mathbf{u} - \mathbf{c}\|_2 \leq R\}, \tag{3.8}$$

then $\max_{\mathbf{u} \in \mathcal{R}} |\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}| = |\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}| + R\|\mathbf{a}_\ell\|_2$ and the relaxation of (3.7) leads to

$$|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}| < \lambda - R\|\mathbf{a}_\ell\|_2 \implies \mathbf{x}^\star_{(\ell)} = 0. \tag{3.9}$$

In this case, the screening test is straightforward to implement since it only requires the evaluation of one inner product between $\mathbf{a}_\ell$ and $\mathbf{c}$.[3]

[3] We note that the ℓ_2-norm appearing in the expression of the test is usually considered as "known" since it can be evaluated offline.

Many procedures have been proposed in the literature to construct safe spheres [20, 36, 46] or safe regions with refined geometries [12, 42, 44, 45]. If f^* is a ζ-strongly convex function, a popular approach to construct a safe region is the so-called "GAP sphere" [36] whose center and radius are defined as follows:

$$\begin{aligned} \mathbf{c} &= \mathbf{u} \\ R &= \sqrt{\tfrac{2}{\zeta}(P(\mathbf{x}) - D(\mathbf{u}))} \end{aligned} \tag{3.10}$$

where $(\mathbf{x}, \mathbf{u})$ is any primal-dual feasible couple. This approach has gained in popularity because of its good behavior when $(\mathbf{x}, \mathbf{u})$ is close to optimality. In particular, if f is proper lower semi-continuous, $\mathbf{x} = \mathbf{x}^\star$ and $\mathbf{u} = \mathbf{u}^\star$, then $P(\mathbf{x}) - D(\mathbf{u}) = 0$ by strong duality [4, Proposition 15.22]. In this case, screening test (3.9) reduces to (3.6) and, except in some degenerated cases, all the zero components of $\mathbf{x}^\star$ can be identified by the screening test. Interestingly, this behavior also provably occurs for sufficiently small values of the dual gap [37, Propositions 8 and 9] and has been observed in many numerical experiments, see *e.g.*, [17, 20, 28, 36].

As a final remark, let us mention that the framework presented in this section extends to optimization problems where the (sparsity-promoting) penalty function describes a group-separable norm, see *e.g.*, [13, 36]. In particular, the complementary slackness condition (3.5) still holds (up to a minor modification), thus allowing to design safe screening tests based on the same rationale. We note that, since the SLOPE penalization does not feature such a separability property, the methodology presented in this section does unfortunately not apply.

4. Safe screening rules for SLOPE.

In this section, we propose a new procedure to extend the concept of safe screening to SLOPE. Our exposition is organized as follows. In Subsection 4.1 we describe our working assumptions and in Subsection 4.2 we present a family of screening tests for SLOPE (see Theorem 4.3). Each test is defined by a set of parameters $\{p_q\}_{q\in[\![1,n]\!]}$ and takes the form of a series of inequalities. We show that a simple test of the form (3.9) can be recovered for some particular values of the parameters $\{p_q\}_{q\in[\![1,n]\!]}$, although this choice does not correspond to the most effective test in the general case. In Subsection 4.3, we finally propose an efficient numerical procedure to verify simultaneously *all* the proposed screening tests.

4.1. Working hypotheses.

In this section, we present two working assumptions which are assumed to hold in the rest of the paper even when not explicitly mentioned.

We first suppose that the regularization parameter λ satisfies

$$0 < \lambda < \lambda_{\max} \triangleq \max_{q\in[\![1,n]\!]} \left(\sum_{k=1}^{q} \left|\mathbf{A}^{\mathrm{T}}\mathbf{y}\right|_{[k]} \Big/ \sum_{k=1}^{q} \gamma_k \right). \tag{4.1}$$

In particular, the hypothesis $\lambda_{\max} > 0$ is tantamount to assuming that $\mathbf{y} \notin \ker(\mathbf{A}^{\mathrm{T}})$. On the other hand, $\lambda < \lambda_{\max}$ prevents the vector $\mathbf{0}_n$ from being a minimizer of the SLOPE problem (1.1). More precisely, it can be shown that under condition (1.3),

$$\lambda \text{ and } \{\gamma_k\}_{k=1}^{n} \text{ verify (4.1)} \iff \mathbf{0}_n \text{ is not a minimizer of (1.1)}. \tag{4.2}$$

A proof of this result is provided in Appendix A.2.

Second, we assume that the columns of the dictionary $\mathbf{A}$ are unit-norm, *i.e.*,

$$\forall j \in [\![1, n]\!]: \quad \|\mathbf{a}_j\|_2 = 1. \tag{4.3}$$

Assumption (4.3) simplifies the statement of our results in the next subsection. However, all our subsequent derivations can be easily extended to the general case where (4.3) does not hold.

4.2. Safe screening rules. In this section, we derive a family of safe screening rules for SLOPE.

Let us first note that (1.1) admits at least one minimizer and our screening problem is therefore well-posed. Indeed, the primal cost function in (1.1) is continuous and coercive since r_{SLOPE} is a norm (see *e.g.*, [6, Proposition 1.1] or [48, Lemma 2]); the existence of a minimizer then follows from Weierstrass theorem [4, Theorem 1.29]. In the following, we will assume that the minimizer is unique to simplify our statements. Nevertheless, all our results extend to the general case where there exist more than one minimizer by replacing "$\mathbf{x}^\star_{(\ell)} = 0$" by "$\mathbf{x}^\star_{(\ell)} = 0$ for any minimizer of (1.1)" in all our subsequent statements.

Our starting point to derive our safe screening rules is the following primal-dual optimality condition:

THEOREM 4.1. *Let*

$$\mathbf{u}^\star = \operatorname*{arg\,max}_{\mathbf{u} \in \mathcal{U}} \; D(\mathbf{u}) \triangleq \tfrac{1}{2}\|\mathbf{y}\|_2^2 - \tfrac{1}{2}\|\mathbf{y} - \mathbf{u}\|_2^2, \tag{4.4}$$

where

$$\mathcal{U} = \left\{ \mathbf{u} \colon \sum_{k=1}^{q} \left|\mathbf{A}^{\mathrm{T}}\mathbf{u}\right|_{[k]} \leq \lambda \sum_{k=1}^{q} \gamma_k,\, q \in [\![1, n]\!] \right\}. \tag{4.5}$$

Then, for all integers $\ell \in [\![1, n]\!]$:

$$\forall q \in [\![1, n]\!]: \; \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star\right| + \sum_{k=1}^{q-1} \left|\mathbf{A}_{\backslash \ell}^{\mathrm{T}}\mathbf{u}^\star\right|_{[k]} < \lambda \sum_{k=1}^{q} \gamma_k \implies \mathbf{x}^\star_{(\ell)} = 0. \tag{4.6}$$

A proof of this result is provided in Appendix B.1. We mention that, although it differs quite significantly in its formulation, Theorem 4.1 is closely related to [32, Proposition 1].[4] We also note that (4.4) corresponds to the dual problem of (1.1), see *e.g.*, [6, Section 2.5]. Moreover, $\mathbf{u}^\star$ exists and is unique because D is a continuous strongly-concave function and $\mathcal{U}$ a closed convex set. The equality in (4.4) is therefore well-defined.

Theorem 4.1 provides a condition similar to (3.6) relating the dual optimal solution $\mathbf{u}^\star$ to the zero components of the primal minimizer $\mathbf{x}^\star$. Unfortunately, evaluating the dual solution $\mathbf{u}^\star$ requires a computational load comparable to the one needed to solve the SLOPE problem (1.1). Similarly to ℓ_1-penalized problems, tractable screening rules can nevertheless be devised if "easily-computable" upper bounds on the

[4] We refer the reader to Section SM1 of the electronic supplementary material of this paper for a detailed description and a proof of the connection between these two results.

left-hand side of (4.6) can be found. In particular, for any set $\{B_{q,\ell} \in \mathbb{R}\}_{q\in[\![1,n]\!]}$ verifying

$$\forall q \in [\![1,n]\!]: \; \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star\right| + \sum_{k=1}^{q-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}^\star\right|_{[k]} \leq B_{q,\ell}, \tag{4.7}$$

we readily have that

$$\forall q \in [\![1,n]\!]: B_{q,\ell} < \lambda \sum_{k=1}^{q} \gamma_k \implies \mathbf{x}^\star_{(\ell)} = 0. \tag{4.8}$$

The next lemma provides several instances of such upper bounds:

Lemma 4.2. *Let $\mathbf{u}^\star \in \mathcal{S}(\mathbf{c}, R)$. Then $\forall \ell \in [\![1,n]\!]$ and $\forall q \in [\![1,n]\!]$, we have that*

$$B_{q,\ell} \triangleq \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}\right| + \sum_{k=p}^{q-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{c}\right|_{[k]} + (q-p+1)R + \lambda \sum_{k=1}^{p-1} \gamma_k$$

verifies (4.7) for any $p \in [\![1,q]\!]$.

A proof of this result is available in Appendix B.2. We note that Lemma 4.2 defines *one* particular family of upper bounds on the left-hand side of (4.7). The derivation of these upper bounds is based on the knowledge of a safe spherical region and partially exploits the definition of the dual feasible set, see Appendix B.2. We nevertheless emphasize that other choices of safe regions or majorization techniques can be envisioned and possibly lead to more favorable upper bounds.

Defining

$$\kappa_{q,p} \triangleq \lambda \left(\sum_{k=p}^{q} \gamma_k \right) - (q-p+1)R, \tag{4.9}$$

a straightforward particularization of (4.8) then leads to the following safe screening rules for SLOPE:

Theorem 4.3. *Let $\{p_q\}_{q\in[\![1,n]\!]}$ be a sequence such that $p_q \in [\![1,q]\!]$ for all $q \in [\![1,n]\!]$. Then, the following statement holds:*

$$\forall q \in [\![1,n]\!]: \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}\right| + \sum_{k=p_q}^{q-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{c}\right|_{[k]} < \kappa_{q,p_q} \implies \mathbf{x}^\star_{(\ell)} = 0. \tag{4.10}$$

We mention that the notation "p_q" is here introduced to stress the fact that a different value of p can be used for each q in (4.10). Since $q \in [\![1,n]\!]$ and each parameter p_q can take on q different values in Theorem 4.3, (4.10) thus defines $n!$ different screening tests for SLOPE where $\frac{n(n+1)}{2}$ distinct inequalities are involved. We discuss two particular choices of parameters $\{p_q\}_{q\in[\![1,n]\!]}$ below and propose an efficient procedure to jointly evaluate all the tests defined by feasible sequences $\{p_q\}_{q\in[\![1,n]\!]}$ in the next section.

Let us first consider the case where

$$\forall q \in [\![1,n]\!]: \; p_q = 1. \tag{4.11}$$

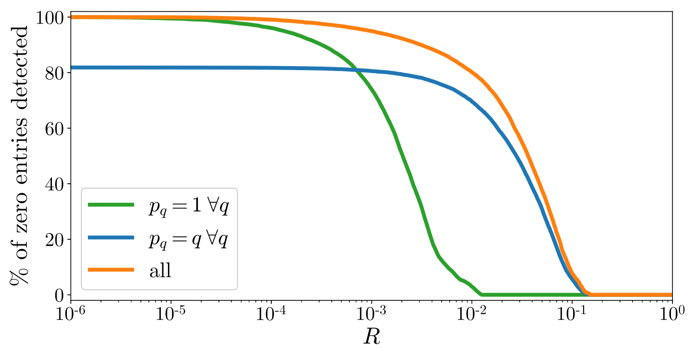

FIG. 1. *Percentage of zero entries in $\mathbf{x}^\star$ detected by the safe screening tests as a function of R, the radius of the safe sphere. Each curve corresponds to a different implementation of the safe screening test* (4.10)*: $p_q = 1\ \forall q$, see* (4.12) *(green curve), $p_q = q\ \forall q$, see* (4.14) *(blue curve), and all possible choices for $\{p_q\}_{q\in[\![1,n]\!]}$ (orange curve). The results are generated by using the* **OSCAR-1** *sequence for $\{\gamma_k\}_{k=1}^n$, the Toeplitz dictionary and the ratio $\lambda/\lambda_{\max} = 0.5$, see* Subsection 5.1.

Screening test (4.10) then particularizes as

$$\forall q \in [\![1,n]\!]:\ \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}\right| + \sum_{k=1}^{q-1}\left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{c}\right|_{[k]} < \lambda\left(\sum_{k=1}^{q}\gamma_k\right) - qR \implies \mathbf{x}^\star_{(\ell)} = 0. \tag{4.12}$$

Interestingly, (4.12) shares the same mathematical structure as optimality condition (4.6). In particular, (4.12) reduces to (4.6) when $\mathbf{c} = \mathbf{u}^\star$ and $R = 0$. In this case, it is easy to see that (4.12) is the best[5] screening test within the family of tests defined in Theorem 4.3 since an equality occurs in (4.7).

In practice, we may expect this conclusion to remain valid when R is "sufficiently" close to zero. This behavior is illustrated in Figure 1. The figure represents the proportion of zeros entries of $\mathbf{x}^\star$ detected by screening test (4.10) for different "qualities" of the safe region and different choices of parameters $\{p_q\}_{q\in[\![1,n]\!]}$. We refer the reader to Subsection 5.1 for a detailed description of the simulation setup. The center of the safe sphere used to apply (4.10) is assumed to be equal (up to machine precision) to $\mathbf{u}^\star$ and the x-axis of the figure represents the radius R of the sphere region. The green curve corresponds to test (4.12); the orange curve represents the screening performance achieved when test (4.10) is implemented for all possible choices for $\{p_q\}_{q\in[\![1,n]\!]}$. We note that, as expected, the green curve attains the best screening performance as soon as R becomes close to zero.

At the other extreme of the spectrum, another case of interest reads as:

$$\forall q \in [\![1,n]\!]:\ p_q = q. \tag{4.13}$$

Using our initial hypothesis (1.3), the screening test (4.10) rewrites[6]

$$|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}| < \lambda\gamma_n - R \implies \mathbf{x}^\star_{(\ell)} = 0. \tag{4.14}$$

Interestingly, this test has the same mathematical structure as (3.9) with the exception that λ is multiplied by the value of the smallest weighting coefficient γ_n. In particular,

[5]In the following sense: if test (4.10) passes for some choice of the parameters $\{p_q\}_{q\in[\![1,n]\!]}$, then test (4.12) also necessarily succeeds.

[6] More precisely, (4.10) reduces to "$\forall q \in [\![1,n]\!]:\ |\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}| < \lambda\gamma_q - R \implies \mathbf{x}^\star_{(\ell)} = 0$" which, in view of (1.3), is equivalent to (4.14).

if $\gamma_k = 1\ \forall k \in [\![1, n]\!]$ SLOPE reduces to LASSO and test (4.14) is equivalent to (3.9); Theorem 4.3 thus encompasses standard screening rule (3.9) for LASSO as a particular case. The following result emphasizes that (4.14) is in fact the best screening rule within the family of tests defined by Theorem 4.3 when $\gamma_k = 1\ \forall k \in [\![1, n]\!]$:

LEMMA 4.4. *If $\gamma_k = 1\ \forall k \in [\![1, n]\!]$ and test (4.10) passes for some choice of parameters $\{p_q\}_{q\in[\![1,n]\!]}$, then test (4.14) also succeeds.*

A proof of this result is available in Appendix B.3.

As a final remark, let us mention that, although we just emphasized that some choices of parameters $\{p_q\}_{q\in[\![1,n]\!]}$ can be optimal (in terms of screening performance) in some situations, no conclusion can be drawn in the general case. In particular, we found in our numerical experiments that the best choice for $\{p_q\}_{q\in[\![1,n]\!]}$ depends on many factors: the weights $\{\gamma_k\}_{k=1}^n$, the radius of the safe sphere R, the nature of the dictionary, the atom to screen, etc. This is illustrated in Fig. 1: we see that the blue and green curves deviate from the orange curve for certain values of R, that is the best screening performance is not necessarily achieved for $p_q = 1$ or $p_q = q\ \forall q \in [\![1, n]\!]$.

4.3. Efficient implementation. Since the best values for $\{p_q\}_{q\in[\![1,n]\!]}$ cannot be foreseen, it is desirable to evaluate the screening rule (4.10) for *any* choice of these parameters. Formally, this ideal test reads:

$$\forall q \in [\![1, n]\!], \exists p_q \in [\![1, q]\!] : \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}\right| + \sum_{k=p_q}^{q-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{c}\right|_{[k]} < \kappa_{q,p_q} \implies \mathbf{x}^\star_{(\ell)} = 0. \tag{4.15}$$

Since verifying this test for a *given* index ℓ involves the evaluation of $\mathcal{O}(n^2)$ inequalities, a brute-force evaluation of (4.15) for all atoms of the dictionary requires $\mathcal{O}(n^3)$ operations. In this section, we present a procedure to perform this task with a complexity scaling as $\mathcal{O}(n \log n + TL)$ where $T \le n$ is some problem-dependent constant (to be defined later on) and L is the number of atoms of the dictionary passing test (4.15). Our procedure is summarized in Algorithms 4.1 and 4.2, and is grounded on the following nesting properties.

Nesting of the tests for different atoms. We first emphasize that there exists an implication between the failures of test (4.15) for some group of indices. In particular, the following result holds:

LEMMA 4.5. *Let $B_{q,\ell}$ be defined as in Lemma 4.2 and assume that*

$$\left|\mathbf{a}_1^{\mathrm{T}}\mathbf{c}\right| \ge \ldots \ge \left|\mathbf{a}_n^{\mathrm{T}}\mathbf{c}\right|. \tag{4.16}$$

Then $\forall q \in [\![1, n]\!]$:

$$\ell < \ell' \implies B_{q,\ell} \ge B_{q,\ell'}. \tag{4.17}$$

A proof of this result is provided in Appendix B.4. Lemma 4.5 has the following consequence: if (4.16) holds, the failure of test (4.15) for some $\ell' \in [\![2, n]\!]$ implies the failure of the test for any index $\ell \in [\![1, \ell' - 1]\!]$. This immediately suggests a backward strategy for the evaluation of (4.15), starting from $\ell = n$ and going backward to smaller indices. This is the sense of the main recursion in Algorithm 4.1.

We note that hypothesis (4.16) can always be verified by a proper reordering of the elements of $|\mathbf{A}^{\mathrm{T}}\mathbf{c}|$. This can be achieved by state-of-the-art sorting procedures

Algorithm 4.1 Fast implementation of SLOPE screening test (4.15)

```
Require: radius R ≥ 0, sorted elements {|A^T c|_[k]}_{k=1}^n
ℒ = ∅ {Set of screened atoms: init}
ℓ = n {Index of atom under testing: init}
Evaluate {g(p)}_{p=1}^n, {p*(q)}_{q=1}^n, {q*(k)}_{k=1}^n
run = 1
while run == 1 and ℓ > 0 do
  test = Algorithm 4.2(R,ℓ,{g(p)}_{p=1}^n,{p*(q)}_{q=1}^n,{q*(k)}_{k=1}^n)
  if test == 1 then
    ℒ = ℒ ∪ {ℓ}
    ℓ = ℓ − 1
  else
    run = 0 {Stop testing as soon as one atom does not pass the test}
  end if
end while
return ℒ (Set of indices passing test (4.15))
```

with a complexity of $\mathcal{O}(n \log n)$. Therefore, in the sequel we will assume that (4.16) holds even if not explicitly mentioned.

Nesting of some inequalities. We next show that the number of inequalities to be verified may possibly be substantially smaller than $\mathcal{O}(n^2)$. We first focus on the case "$\ell = n$" and then extend our result to the general case "$\ell < n$".

Let us first note that under hypothesis (4.16):

$$\forall k \in [\![1, n-1]\!] : \ |\mathbf{A}_{\backslash n}^{\mathrm{T}}\mathbf{c}|_{[k]} = |\mathbf{A}_{\backslash n}^{\mathrm{T}}\mathbf{c}|_{(k)}, \tag{4.18}$$

that is the kth largest element of $|\mathbf{A}_{\backslash n}^{\mathrm{T}}\mathbf{c}|$ is simply equal to its kth component. The particularization of (4.15) to $\ell = n$ can then be rewritten as:

$$\forall q \in [\![1, n]\!], \exists p_q \in [\![1, q]\!] : \ \left|\mathbf{a}_n^{\mathrm{T}}\mathbf{c}\right| < \tau_{q,p_q} \tag{4.19}$$

where $\tau_{q,p}$ is defined $\forall q \in [\![1, n]\!]$ and $p \in [\![1, q]\!]$ as

$$\tau_{q,p} \triangleq \kappa_{q,p} - \sum_{k=p}^{q-1} \left|\mathbf{A}^{\mathrm{T}}\mathbf{c}\right|_{(k)} = \sum_{k=p}^{q-1} (\lambda\gamma_k - \left|\mathbf{A}^{\mathrm{T}}\mathbf{c}\right|_{(k)} - R) + (\lambda\gamma_q - R). \tag{4.20}$$

We show hereafter that (4.19) can be verified by only considering a "well-chosen" subset of thresholds $\mathcal{T} \subseteq \{\tau_{q,p} \colon q \in [\![1, n]\!], p \in [\![1, q]\!]\}$, see Lemma 4.6 below.

If

$$p^\star(q) \triangleq \operatorname*{arg\,max}_{p \in [\![1,q]\!]} \tau_{q,p}, \tag{4.21}$$

we obviously have

$$\left|\mathbf{a}_n^{\mathrm{T}}\mathbf{c}\right| < \tau_{q,p^\star(q)} \iff \exists p_q \in [\![1, q]\!] : \ \left|\mathbf{a}_n^{\mathrm{T}}\mathbf{c}\right| < \tau_{q,p_q}. \tag{4.22}$$

In other words, for each $q \in [\![1, n]\!]$, satisfying the inequality "$\left|\mathbf{a}_n^{\mathrm{T}}\mathbf{c}\right| < \tau_{q,p}$" for $p = p^\star(q)$ is necessary and sufficient to ensure that it is verified for some $p_q \in [\![1, q]\!]$.

Motivated by this observation, we show the following items below: *i)* $p^\star(q)$ can be evaluated $\forall q \in [\![1,n]\!]$ with a complexity $\mathcal{O}(n)$; *ii)* similarly to p, only a subset of values of $q \in [\![1,n]\!]$ are of interest to implement (4.19).

Let us define the function:

$$\begin{aligned} g\colon [\![1,n]\!] &\to \mathbb{R} \\ p &\mapsto \textstyle\sum_{k=p}^{n} (\lambda\gamma_k - \left|\mathbf{A}^{\mathrm{T}}\mathbf{c}\right|_{(k)} - R)\,. \end{aligned} \tag{4.23}$$

We then have $\forall q \in [\![1,n]\!]$ and $p \in [\![1,q]\!]$:

$$\tau_{q,p} = g(p) - (g(q) - \lambda\gamma_q) - R. \tag{4.24}$$

In view of (4.24), the optimal value $p^\star(q)$ can be computed as

$$p^\star(q) = \operatorname*{arg\,max}_{p \in [\![1,q]\!]} g(p). \tag{4.25}$$

Considering (4.23), we see that the evaluation of $g(p)$ $\forall p \in [\![1,n]\!]$ (and therefore $p^\star(q)$ $\forall q \in [\![1,n]\!]$) can be done with a complexity scaling as $\mathcal{O}(n)$. This proves item *i)*.

Let us now show that only some specific indices $q \in [\![1,n]\!]$ are of interest to implement (4.19). Let

$$q^\star(k) \triangleq \operatorname*{arg\,max}_{q \in [\![1,k]\!]} g(q) - \lambda\gamma_q, \tag{4.26}$$

and define the sequence $\{q^{(t)}\}_t$ as

$$\begin{cases} q^{(1)} &= q^\star(n) \\ q^{(t)} &= q^\star(p^\star(q^{(t-1)}) - 1) \end{cases} \tag{4.27}$$

where the recursion is applied as long as $p^\star(q^{(t-1)}) > 1$.[7] We then have the following result whose proof is available in Appendix B.5:

Lemma 4.6. *Let* $\mathcal{T} \triangleq \left\{\tau_{q,p^\star(q)} \colon q \in \{q^{(t)}\}_t\right\}$ *where* $\{q^{(t)}\}_t$ *is defined in* (4.27). *Test* (4.19) *is passed if and only if*

$$\forall \tau \in \mathcal{T}:\ |\mathbf{a}_n^{\mathrm{T}}\mathbf{c}| < \tau. \tag{4.28}$$

Lemma 4.6 suggests the procedure described in Algorithm 4.2 (with $\ell = n$) to verify if (4.19) is passed. In a nutshell, the lemma states that only $\operatorname{card}(\mathcal{T})$ inequalities need to be taken into account to implement (4.19). We note that $\operatorname{card}(\mathcal{T}) \leq n$ since only one value of p (that is $p^\star(q)$) has to be considered for any $q \in [\![1,n]\!]$. This is in contrast with a brute-force evaluation of (4.19) which requires the verification of $\mathcal{O}(n^2)$ inequalities.

We finally emphasize that the procedure described in Algorithm 4.2 also applies to $\ell < n$ as long as the screening test is passed for all $\ell' > \ell$. More specifically, if test (4.15) is passed for all $\ell' \in [\![\ell+1,n]\!]$, then its particularization to atom $\mathbf{a}_\ell$ reads

$$\forall \tau \in \mathcal{T}':\ \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}\right| < \tau \tag{4.29}$$

for some $\mathcal{T}' \subseteq \mathcal{T}$.

[7] We note that the sequence $\{q^{(t)}\}_t$ is strictly decreasing and thus contains at most n elements.

Algorithm 4.2 Check if test (4.15) is passed for ℓ if it is passed for $\ell' > \ell$

Require: radius $R \geq 0$, index $\ell \in [\![1, n]\!]$, $\{g(p)\}_{p=1}^{n}$, $\{p^\star(q)\}_{q=1}^{n}$,$\{q^\star(k)\}_{k=1}^{n}$

$q = q^\star(\ell)$
test $= 1$
run $= 1$
while run $== 1$ **do**
  $\tau = g(p^\star(q)) - g(q) + (\lambda\gamma_q - R)$ {Evaluation of current threshold, see (4.24)}
  if $|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}| \geq \tau$ **then**
    test $= 0$ {Test failed}
    run $= 0$ {Stops the recursion}
  end if
  if $p^\star(q) > 1$ **then**
    $q = q^\star(p^\star(q) - 1)$ {Next value of q to test, see (4.27)}
  else
    run $= 0$ {Stops the recursion}
  end if
end while
return test ($= 1$ if test passed and 0 otherwise)

Indeed, if screening test (4.15) is passed for all $\ell' \in [\![\ell + 1, n]\!]$, the corresponding elements can be discarded from the dictionary and we obtain a reduced problem only involving atoms $\{\mathbf{a}_{\ell'}\}_{\ell' \in [\![1,\ell]\!]}$. Since (4.16) is assumed to hold, $\mathbf{a}_\ell$ attains the smallest absolute inner product with $\mathbf{c}$ and we end up with the same setup as in the case "$\ell = n$". In particular, if screening test (4.15) is passed for all $\ell' \in [\![\ell + 1, n]\!]$, Lemma 4.6 still holds for $\mathbf{a}_\ell$ by letting $q^{(1)} = q^\star(\ell)$ in the definition of the sequence $\{q^{(t)}\}_t$ in (4.27).

To conclude this section, let us summarize the complexity needed to implement Algorithms 4.1 and 4.2. First, Algorithm 4.1 requires the entries $|\mathbf{A}^{\mathrm{T}}\mathbf{c}|$ to be sorted to satisfy hypothesis (4.5). This involves a complexity $\mathcal{O}(n \log n)$. Moreover, the sequences $\{g(p)\}_{p=1}^{n}$, $\{p^\star(q)\}_{q=1}^{n}$, $\{q^\star(k)\}_{k=1}^{n}$ can be evaluated with a complexity $\mathcal{O}(n)$. Finally, the main recursion in Algorithm 4.1 implies to run Algorithm 4.2 L times, where L is the number of atoms passing test (4.15). Since Algorithm 4.2 requires to verify at most $T = \operatorname{card}(\mathcal{T})$ inequalities, the overall complexity of the main recursion scales as $\mathcal{O}(LT)$. Overall, the complexity of Algorithm 4.1 is therefore $\mathcal{O}(n \log n + LT)$.

5. Numerical simulations.

We present hereafter several simulation results demonstrating the effectiveness of the proposed screening procedure to accelerate the resolution of SLOPE. This section is organized as follows. In Subsection 5.1, we present the experimental setups considered in our simulations. In Subsection 5.2 we compare the effectiveness of different screening strategies. In Subsection 5.3, we show that our methodology enables to reach better convergence properties for a given computational budget.

5.1. Experimental setup.

We detail below the experimental setups used in all our numerical experiments.

Dictionaries and observation vectors: New realizations of $\mathbf{A}$ and $\mathbf{y}$ are drawn for each trial as follows. The observation vector is generated according to a uniform distribution on the m-dimensional sphere. The elements of $\mathbf{A}$ obey one of the following models:

1. the entries are i.i.d. realizations of a centered Gaussian,
2. the entries are i.i.d. realizations of a uniform distribution on $[0, 1]$,
3. the columns are shifted versions of a Gaussian curve.

For all distributions, the columns of $\mathbf{A}$ are normalized to have unit ℓ_2-norm. In the following, these three options will be respectively referred to as "Gaussian", "Uniform" and "Toeplitz".

Regularization parameters: We consider three different choices for the sequence $\{\gamma_k\}_{k=1}^n$, each of them corresponding to a different instance of the well-known OSCAR problem [7, Eq. (3)]. More specifically, we let

$$\forall k \in [\![1, n]\!] : \ \gamma_k \triangleq \beta_1 + \beta_2(n-k) \tag{5.1}$$

where β_1, β_2 are nonnegative parameters chosen so that $\gamma_1 = 1$ and $\gamma_n \in \{.9, .1, 10^{-3}\}$. In the sequel, these parametrizations will respectively be referred to as "`OSCAR-1`", "`OSCAR-2`" and "`OSCAR-3`".

5.2. Performance of screening strategies. We first compare the effectiveness of different screening strategies described in Section 4. More specifically, we evaluate the proportion of zero entries in $\mathbf{x}^\star$ – the solution of SLOPE problem (1.1) – that can be identified by tests (4.12), (4.14) and (4.15) as a function of the "quality" of the safe sphere. These tests will respectively be referred to as "`test-p=1`", "`test-p=q`" and "`test-all`" in the following. Figures 1 (see Subsection 4.2) and 2 represent this criterion of performance as a function of some parameter R_0 (described below) and different values of the ratio $\lambda/\lambda_{\max}$. The results are averaged over 50 realizations. For each simulation trial, we draw a new realization of $\mathbf{y} \in \mathbb{R}^{100}$ and $\mathbf{A} \in \mathbb{R}^{100\times 300}$ according to the distributions described in Subsection 5.1. We consider Toeplitz dictionaries in Figure 1 and Gaussian dictionaries in Figure 2.

The safe sphere used in the screening tests is constructed as follows. A primal-dual solution $(\mathbf{x}_a, \mathbf{u}_a)$ of problems (1.1) and (4.4) is evaluated with "high-accuracy", *i.e.*, with a duality GAP of 10^{-14} as stopping criterion. More precisely, $\mathbf{x}_a$ is first evaluated by solving the SLOPE problem (1.1) with the algorithm proposed in [5]. To evaluate $\mathbf{u}_a$, we extend the so-called "dual scaling" operator [24, Section 3.3] to the SLOPE problem: we let $\mathbf{u}_a = (\mathbf{y} - \mathbf{A}\mathbf{x}_a)/\beta(\mathbf{y} - \mathbf{A}\mathbf{x}_a)$ where

$$\forall \mathbf{z} \in \mathbb{R}^m : \ \beta(\mathbf{z}) \triangleq \max\left(1, \max_{q\in[\![1,n]\!]} \frac{\sum_{k=1}^q |\mathbf{A}^{\mathrm{T}}\mathbf{z}|_{[k]}}{\lambda \sum_{k=1}^q \gamma_k}\right). \tag{5.2}$$

The couple $(\mathbf{x}_a, \mathbf{u}_a)$ is then used to construct a sphere $\mathcal{S}(\mathbf{c}_a, R_a)$ in $\mathbb{R}^m$ whose parameters are given by

$$\mathbf{c} = \mathbf{u}_a \tag{5.3a}$$

$$R = R_0 + \sqrt{2(P(\mathbf{x}_a) - D(\mathbf{u}_a))} \tag{5.3b}$$

where R_0 is a nonnegative scalar. We note that for $R_0 = 0$, the latter sphere corresponds to the GAP safe sphere described in (3.10).[8] Hence, (5.3a) and (5.3b) define

[8] We note that the GAP safe sphere derived in [36] for problem (3.1) extends to SLOPE since 1) the dual problem has the same mathematical form and 2) its derivation does not leverage the definition of the dual feasible set.

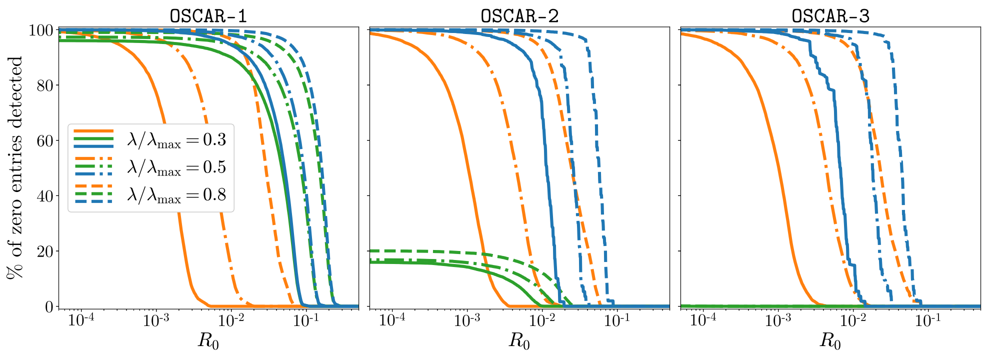

FIG. 2. *Percentage of zero entries in the solution of the SLOPE problem identified by* `test-p=1` *(orange lines),* `test-p=q` *(green lines) and* `test-all` *(blue lines) as a function of R_0 for the Gaussian dictionary, three values of $\lambda/\lambda_{\max}$ and three parameter sequences $\{\gamma_k\}_{k=1}^n$.*

a safe sphere for any choice of the nonnegative scalar $R_0 \geq 0$.

Figure 1 concentrates on the sequence `OSCAR-1` whereas each subfigure corresponds to a different choice for $\{\gamma_k\}_{k=1}^n$ in Figure 2. For the three considered screening strategies, we observe that the detection performance decreases as R_0 increases. Interestingly, different behaviors can be noticed. For all simulation setups, `test-p=1` reaches a detection rate of 100% whenever R_0 is sufficiently small. The performance of `test-p=q` varies from one sequence to another: it outperforms `test-p=1` for `OSCAR-1`, is able to detect at most 20% of the zeros for `OSCAR-2` and fail for all values of R_0 for `OSCAR-3`. Finally, `test-all` outperforms quite logically the two other strategies. The gap in performance depends on both the considered setup and the radius R_0 but can be quite significant in some cases. For example, when $\lambda/\lambda_{\max} = 0.5$ and $R_0 = 10^{-2}$, there is 80% more entries passing `test-all` than `test-p=1` for all parameter sequences.

These results may be explained as follows. First, we already mentioned in Section 4 that when the radius of the safe sphere is sufficiently small (that is, when R_0 is close to zero), `test-p=1` is expected to be the best[9] screening test within the family of tests defined in Theorem 4.3. Similarly, if the SLOPE weights satisfy $\gamma_1 = \gamma_n$, we showed in Lemma 4.4 that no test in Theorem 4.3 can outperform `test-p=q`. Hence, one may reasonably expect that this conclusion remains valid whenever $\gamma_1 \simeq \gamma_n$, as observed for the sequence `OSCAR-1` in our simulations. On the other hand, passing `test-p=q` becomes more difficult as parameter γ_n is small. As a matter of fact, the test will *never* pass when $\gamma_n = 0$. In our experiments, the sequences $\{\gamma_k\}_{k=1}^n$ are such that γ_n is close to zero for `OSCAR-2` and `OSCAR-3`. Finally, since `test-all` encompasses the two other tests, it is expected to always perform at least as well as the latter.

5.3. Benchmarks. As far as our simulation setup is concerned, the results presented in the previous section show a significant advantage in implementing `test-all` in terms of detection performance. However, this conclusion does not include any consideration about the numerical complexity of the tests. We note that, although the proposed screening rules can lead to a significant reduction of the problem dimen-

[9] in the sense defined in Footnote 5 page 8.

sions, our tests also induce some additional computational burden. In particular, we emphasized in Subsection 4.3 that `test-all` can be verified for all atoms of the dictionary with a complexity $\mathcal{O}(n \log n + TL)$ where $T \leq n$ is a problem-dependent parameter and L is the number of atoms passing the test. Moreover, we also note that, as far as a GAP safe sphere is considered in the implementation of the tests, its construction requires the identification of a dual feasible point $\mathbf{u}$ and this operation typically induces a computational overhead of $\mathcal{O}(n \log n)$ (see below for more details).

In this section, we therefore investigate the benefits (from a "complexity-accuracy trade-off" point of view) of interleaving the proposed safe screening methodology with the iterations of an accelerated proximal gradient algorithm [5]. In all our tests, we consider the GAP safe sphere defined in (3.10). The primal point used in the construction of the GAP sphere corresponds to the current iterate of the solving procedure, say $\mathbf{x}^{(t)}$. A dual feasible point $\mathbf{u}^{(t)}$ is constructed as

$$\mathbf{u}^{(t)} = \frac{\mathbf{y} - \mathbf{A}\mathbf{x}^{(t)}}{\beta(\mathbf{y} - \mathbf{A}\mathbf{x}^{(t)})} \tag{5.4}$$

where $\beta \colon \mathbb{R}^m \to \mathbb{R}^m$ is either defined as in (5.2) or as follows:

$$\forall \mathbf{z} \in \mathbb{R}^m : \ \beta(\mathbf{z}) \triangleq \max\left(1, \max_{k \in [\![1,n]\!]} \frac{\left|\mathbf{A}^{\mathrm{T}}\mathbf{z}\right|_{[k]}}{\lambda \gamma_k}\right). \tag{5.5}$$

(5.2) matches the standard definition of the "dual scaling" operator proposed in [24, Section 3.3] whereas (5.5) corresponds to the option considered in [3].[10] We notice that the two options require to sort the elements of $\left|\mathbf{A}^{\mathrm{T}}\mathbf{z}\right|$ and thus lead to a complexity overhead scaling as $\mathcal{O}(n \log n)$.

In our simulations, we consider the four following solving strategies:

1. Run the proximal gradient procedure [5] with *no* screening.
2. Interleave some iterations of the proximal gradient algorithm with `test-p=q` and construct the dual feasible point with (5.2).
3. Interleave some iterations of the proximal gradient algorithm with `test-p=q` and construct the dual feasible point with (5.5).
4. Interleave some iterations of the proximal gradient algorithm with `test-all` and construct the dual feasible point with (5.2).

These strategies will respectively be denoted "`PG-no`", "`PG-p=q`", "`PG-Bao`" and "`PG-all`" in the sequel. We note that `PG-Bao` closely matches the solving procedure considered in [3].

We compare the performance of these solving strategies by resorting to Dolan-Moré profiles [15]. More precisely, we run each procedure for a given budget of time (that is the algorithm is stopped after a predefined amount of time) on $I = 50$ different instances of the SLOPE problems. In `PG-p=q`, `PG-Bao` and `PG-all`, the screening procedure is applied once every 20 iterations. Each problem instance is generated by drawing a new dictionary $\mathbf{A} \in \mathbb{R}^{100 \times 300}$ and observation vector $\mathbf{y} \in \mathbb{R}^{100}$ according to the distributions described in Subsection 5.1. We then compute the following performance profile for each solver $\texttt{solv} \in \{$`PG-no`, `PG-p=q`, `PG-Bao`, `PG-all`$\}$:

$$\rho_{\texttt{solv}}(\delta) \triangleq 100 \, \frac{\mathrm{card}\left(\{i \in [\![1, I]\!] \colon d_{i,\texttt{solv}} \leq \delta\}\right)}{I} \quad \forall \delta \in \mathbb{R}_+ \tag{5.6}$$

[10] See companion code of [3] available at https://github.com/brx18/Fast-OSCAR-and-OWL-Regression-via-Safe-Screening-Rules/tree/1e08d14c56bf4b6293899ae2092a5e0238d27bf6.

where $d_{i,\texttt{solv}}$ denotes the dual gap attained by solver `solv` for problem instance i. $\rho_{\texttt{solv}}(\delta)$ thus represents the (empirical) probability that solver `solv` reaches a dual gap no greater than δ for the considered budget of time.

Figure 3 presents the performance profiles obtained for three types of dictionaries (Gaussian, Uniform and Toeplitz) and three different weighting sequences $\{\gamma_k\}_{k=1}^n$ (`OSCAR-1`, `OSCAR-2` and `OSCAR-3`). The results are displayed for $\lambda/\lambda_{\max} = 0.5$ but similar performance profiles have been obtained for other values of the ratio $\lambda/\lambda_{\max}$. All algorithms are implemented in Python with Cython bindings and experiments are run on a Dell laptop, 1.80 GHz, Intel Core i7. For each setup, we adjusted the time budget so that $\rho_{\texttt{PG-all}}(10^{-8}) \simeq 50\%$ for the sake of comparison.

As far as our simulation setup is concerned, these results show that the proposed screening methodologies improve the solving accuracy as compared to a standard proximal gradient. `PG-all` improves the average accuracy over `PG-no` in all the considered settings. The gap in performance depends on the setup but is generally quite significant. `PG-p=q` also enhances the average accuracy in most cases and performs at least comparably to `PG-Bao` in all setups. As expected, the behavior of `PG-p=q` and `PG-Bao` is more sensitive to the choice of the weighting sequence $\{\gamma_k\}_{k=1}^n$. In particular, the screening performance of these strategies decreases when $\gamma_n \simeq 0$ as emphasized in Subsection 5.2. This results in no accuracy gain over `PG-no` for the sequence `OSCAR-3` as illustrated in Figure 3. Nevertheless, we note that, even in absence of gain, `PG-p=q` and `PG-Bao` do not seem to significantly degrade the performance as compared to `PG-no`.

6. Conclusions. In this paper we proposed a new methodology to safely identify the zeros of the solutions of the SLOPE problem. In particular, we introduced a family of screening rules indexed by some parameters $\{p_q\}_{q=1}^n$ where n is the dimension of the primal variable. Each test of this family takes the form of a series of n inequalities which, when verified, imply the nullity of some coefficient of the minimizers. Interestingly, the proposed tests encompass standard “sphere” screening rule for LASSO as a particular case for some $\{p_q\}_{q=1}^n$, although this choice does not correspond to the most effective test in the general case. We then introduced an efficient numerical procedure to jointly evaluate all the tests in the proposed family. Our algorithm has a complexity $\mathcal{O}(n \log n + TL)$ where $T \leq n$ is some problem-dependent constant and L is the number of elements passing at least one test of the family. We finally assessed the performance of our screening strategy through numerical simulations and showed that the proposed methodology leads to significant improvements of the solving accuracy for a prescribed computational budget.

Acknowledgments. The authors would like to thank the anonymous reviewers for their thoughtful comments and for pointing out one technical flaw in the first version of the manuscript.

Appendix A. Miscellaneous results. Appendix A.1 reminds some useful results from convex analysis applied to the SLOPE problem (1.1). Appendix A.2 provides a proof of (4.2). In all the statements below, $\partial r_{\text{SLOPE}}(\mathbf{x})$ denotes the subdifferential of r_{SLOPE} evaluated at $\mathbf{x}$.

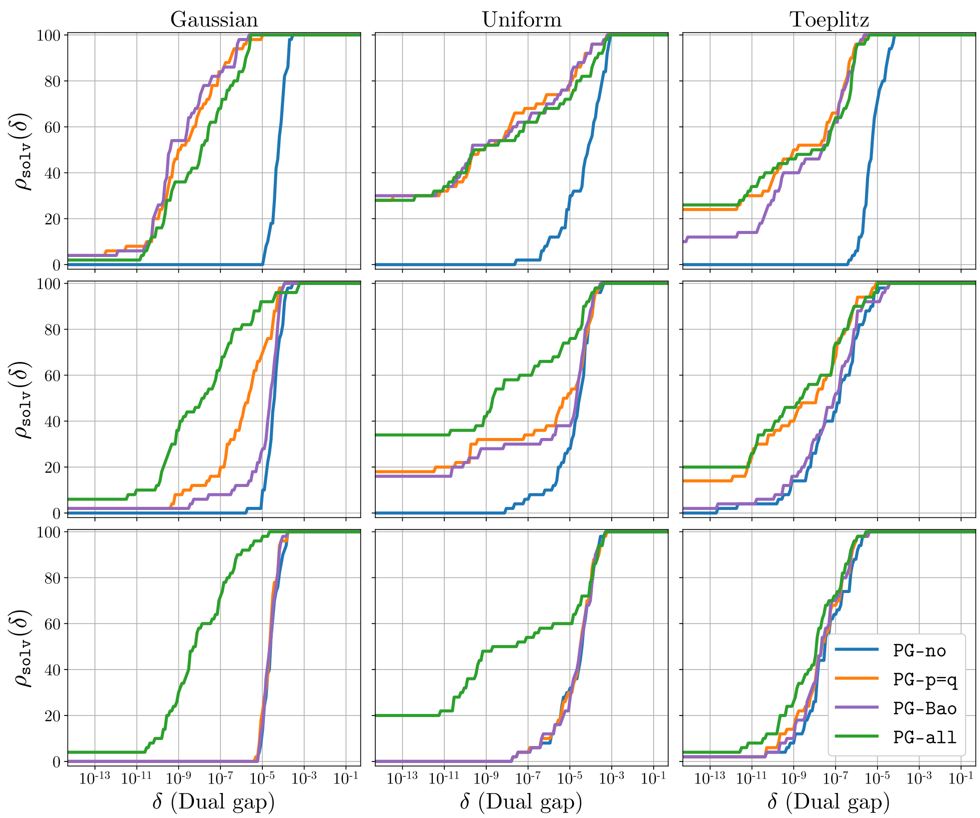

Fig. 3. *Performance profiles of* `PG-no`*,* `PG-p=q`*,* `PG-Bao` *and* `PG-all` *obtained for the "Gaussian" (column 1), "Uniform" (column 2) and "Toeplitz" (column 3) dictionaries and* $\lambda/\lambda_{\max} = 0.5$ *with a budget of time. First row:* `OSCAR-1`*, second row:* `OSCAR-2` *and third row:* `OSCAR-3`*.*

A.1. Some results of convex analysis. We remind below several results of convex analysis that will be used in our subsequent derivations. The first lemma provides a necessary and sufficient condition for $\mathbf{x}^\star \in \mathbb{R}^n$ to be a minimizer of the SLOPE problem (1.1):

Lemma A.1. $\mathbf{x}^\star$ *is a minimizer of* (1.1) $\iff \lambda^{-1}\mathbf{A}^{\mathrm{T}}(\mathbf{y} - \mathbf{A}\mathbf{x}^\star) \in \partial r_{\text{SLOPE}}(\mathbf{x}^\star)$.

Lemma A.1 follows from a direct application of Fermat's rule [4, Proposition 16.4] to problem (1.1). We note that under condition (1.3), r_{SLOPE} defines a norm on $\mathbb{R}^n$, see *e.g.*, [6, Proposition 1.1] or [48, Lemma 2]. The subdifferential $\partial r_{\text{SLOPE}}(\mathbf{x})$ is therefore well defined for all $\mathbf{x} \in \mathbb{R}^n$ and writes as

$$\partial r_{\text{SLOPE}}(\mathbf{x}) = \left\{\mathbf{g} \in \mathbb{R}^n \colon \mathbf{g}^{\mathrm{T}}\mathbf{x} = r_{\text{SLOPE}}(\mathbf{x}) \text{ and } r_{\text{SLOPE},*}(\mathbf{g}) \le 1\right\}, \tag{A.1}$$

where

$$r_{\text{SLOPE},*}(\mathbf{g}) \triangleq \sup_{\mathbf{x}\in\mathbb{R}^n} \mathbf{g}^{\mathrm{T}}\mathbf{x} \;\text{ s.t. } r_{\text{SLOPE}}(\mathbf{x}) \le 1 \tag{A.2}$$

is the dual norm of r_{SLOPE}, see *e.g.*, [1, Eq. (1.4)].

The next lemma states a technical result which will be useful in the proof of Theorem 4.1 in Appendix B:

Lemma A.2. *If* $\mathbf{g} \in \partial r_{\text{SLOPE}}(\mathbf{x})$*, then* $\mathbf{x}^{\mathrm{T}}(\mathbf{g}-\mathbf{g}') \geq 0 \,\forall \mathbf{g}' \in \mathbb{R}^n$ *s.t.* $r_{\text{SLOPE},*}(\mathbf{g}') \leq 1$.

Proof. Let $\mathbf{g} \in \partial r_{\text{SLOPE}}(\mathbf{x})$. One has

$$\begin{aligned} \mathbf{g} \in \partial r_{\text{SLOPE}}(\mathbf{x}) &\iff \mathbf{x} \in \partial r^*_{\text{SLOPE}}(\mathbf{g}) \\ &\iff \forall \mathbf{g}' \in \mathbb{R}^n,\ r^*_{\text{SLOPE}}(\mathbf{g}') \geq r^*_{\text{SLOPE}}(\mathbf{g}) + \langle \mathbf{x}, \mathbf{g}' - \mathbf{g} \rangle \end{aligned} \tag{A.3}$$

where r^*_{SLOPE} refers to the Fenchel conjugate of r_{SLOPE}. The first equivalence is a consequence of [4, Theorem 16.29] and the second of the definition of the subdifferential set. Lemma A.2 follows by noticing that $r^*_{\text{SLOPE}}(\mathbf{g}') = 0 \;\forall \mathbf{g}' \in \mathbb{R}^n$ such that $r_{\text{SLOPE},*}(\mathbf{g}') \leq 1$ by property of r^*_{SLOPE} [4, Item (v) of Example 13.3]. ∎

In the last lemma of this section, we provide a closed-form expression of the subdifferential and the dual norm of r_{SLOPE}:[11]

Lemma A.3. *The dual norm and the subdifferential of* $r_{\text{SLOPE}}(\mathbf{x})$ *respectively write:*

$$\begin{aligned} r_{\text{SLOPE},*}(\mathbf{g}) &= \max_{q \in [\![1,n]\!]} \ \frac{1}{\sum_{k=1}^{q} \gamma_k} \sum_{k=1}^{q} |\mathbf{g}|_{[k]}, \\ \partial r_{\text{SLOPE}}(\mathbf{x}) &= \left\{ \mathbf{g} \in \mathbb{R}^n \colon \mathbf{g}^{\mathrm{T}}\mathbf{x} = r_{\text{SLOPE}}(\mathbf{x}) \ \textit{and} \ \forall q \in [\![1,n]\!] : \ \sum_{k=1}^{q} |\mathbf{g}|_{[k]} \leq \sum_{k=1}^{q} \gamma_k \right\}. \end{aligned}$$

Proof. The expression of the dual norm is a direct consequence of [48, Lemma 4]. More precisely, the authors showed that

$$r_{\text{SLOPE},*}(\mathbf{g}) = \max_{\mathbf{v} \in \bigcup_{q=1}^{n} \mathcal{V}_q} \ \mathbf{g}^{\mathrm{T}}\mathbf{v} \tag{A.4}$$

where $\mathcal{V}_q \triangleq \left\{ \frac{1}{\sum_{k=1}^{q}\gamma_k}\mathbf{s} \colon \mathbf{s} \in \{0,-1,+1\}^n, \operatorname{card}\left(\{j : \mathbf{s}_{(j)} \neq 0\}\right) = q \right\}$ for all $q \in [\![1,n]\!]$. The expression of $r_{\text{SLOPE},*}$ given in Lemma A.3 is a compact rewriting of (A.4) that can be obtained as follows. See first that for all $q \in [\![1,n]\!]$,

$$\max_{\mathbf{v} \in \mathcal{V}_q} \ \mathbf{g}^{\mathrm{T}}\mathbf{v} \leq \frac{1}{\sum_{k=1}^{q}\gamma_k} \sum_{k=1}^{q} |\mathbf{g}|_{[k]}. \tag{A.5}$$

Second, for $q \in [\![1,n]\!]$, let $\mathcal{J}_q \subset [\![1,n]\!]$ be a set q distinct indices such that $|\mathbf{g}_{(j)}| \geq |\mathbf{g}|_{[q]}$ for all $j \in \mathcal{J}_q$. Then, the upper bound in (A.5) is attained by evaluating the left-hand side at $\mathbf{v} \in \mathcal{V}_q$ defined as

$$\forall j \in [\![1,n]\!] : \quad \mathbf{v}_{(j)} = \begin{cases} \frac{1}{\sum_{k=1}^{q}\gamma_k} \operatorname{sign}\left(\mathbf{g}_{(j)}\right) & \text{if } j \in \mathcal{J}_q \\ 0 & \text{otherwise.} \end{cases} \tag{A.6}$$

The expression of the subdifferential follows from (A.1) by plugging the expression of the dual norm in the inequality "$r_{\text{SLOPE},*}(\mathbf{g}) \leq 1$". ∎

[11] We note that an expression of the subdifferential of r_{SLOPE} has already been derived in [10, Fact A.2 in supplementary material]. However, the expression of the subdifferential proposed in Lemma A.3 has a more compact form and is better suited to our subsequent derivations.

A.2. Proof of (4.2). We first observe that

$$\mathbf{0}_n \text{ is not a minimizer of (1.1)} \iff \lambda^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y} \notin \partial r_{\text{SLOPE}}(\mathbf{0}_n), \tag{A.7}$$

as a direct consequence of Lemma A.1. Particularizing the expression of $\partial r_{\text{SLOPE}}(\mathbf{x})$ in Lemma A.3 to $\mathbf{x} = \mathbf{0}_n$, the right-hand side of (A.7) can equivalently be rewritten as

$$\exists q \in [\![1, n]\!] : \ \lambda^{-1} \sum_{k=1}^{q} \left|\mathbf{A}^{\mathrm{T}}\mathbf{y}\right|_{[k]} > \sum_{k=1}^{q} \gamma_k. \tag{A.8}$$

Since $\gamma_1 > 0$ and the sequence $\{\gamma_k\}_{k=1}^{n}$ is nonnegative by hypothesis (1.3), (A.8) can also be rewritten as

$$\exists q \in [\![1, n]\!] : \ \lambda < \frac{\sum_{k=1}^{q} \left|\mathbf{A}^{\mathrm{T}}\mathbf{y}\right|_{[k]}}{\sum_{k=1}^{q} \gamma_k}. \tag{A.9}$$

The statement in (4.2) then follows by noticing that the right-hand side of (4.1) is a compact reformulation of (A.9).

Appendix B. Proofs related to screening tests.

B.1. Proof of Theorem 4.1. In this section, we provide the technical details leading to (4.6). Our derivation leverages the Fermat's rule and the expression of the subdifferential derived in Lemma A.3.

We prove (4.6) by contraposition. More precisely, we show that if $\mathbf{x}^\star_{(\ell)} \neq 0$ for some $\ell \in [\![1, n]\!]$, then

$$\exists q_0 \in [\![1, n]\!], \ \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star\right| + \sum_{k=1}^{q_0-1} \left|\mathbf{A}_{\setminus \ell}^{\mathrm{T}}\mathbf{u}^\star\right|_{[k]} = \lambda \sum_{k=1}^{q_0} \gamma_k. \tag{B.1}$$

Using Lemma A.1 and the following connection between primal-dual solutions (see [6, Section 2.5])

$$\mathbf{u}^\star = \mathbf{y} - \mathbf{A}\mathbf{x}^\star, \tag{B.2}$$

we have that $\mathbf{x}^\star$ is a minimizer of (1.1) if and only if

$$\mathbf{g}^\star \triangleq \lambda^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{u}^\star \in \partial r_{\text{SLOPE}}(\mathbf{x}^\star). \tag{B.3}$$

In the rest of the proof, we will use Lemma A.2 with $\mathbf{x} = \mathbf{x}^\star$, $\mathbf{g} = \mathbf{g}^\star$ and different instances of vector $\mathbf{g}'$ to prove our statement. First, let us define $\mathbf{g}' \in \mathbb{R}^n$ as

$$\begin{aligned} \mathbf{g}'_{(j)} &= \mathbf{g}^\star_{(j)} \quad \forall j \in [\![1, n]\!] \setminus \{\ell\}, \\ \mathbf{g}'_{(\ell)} &= 0. \end{aligned}$$

It is easy to verify that $r_{\text{SLOPE},*}(\mathbf{g}') \leq 1$. Applying Lemma A.2 then leads to

$$\mathbf{g}^\star_{(\ell)}\mathbf{x}^\star_{(\ell)} \geq 0. \tag{B.4}$$

Since $\mathbf{x}^\star_{(\ell)}$ is assumed to be nonzero, we then have

$$\operatorname{sign}\big(\mathbf{g}^\star_{(\ell)}\big) \operatorname{sign}\big(\mathbf{x}^\star_{(\ell)}\big) \geq 0, \tag{B.5}$$

where the equality holds if and only if $\mathbf{g}^\star_{(\ell)} = 0$.

Second, let us consider the following choice for $\mathbf{g}' \in \mathbb{R}^n$:

$$\text{(B.6)} \qquad \begin{aligned} \mathbf{g}'_{(j)} &= \mathbf{g}^\star_{(j)} \quad \forall j \in [\![1, n]\!] \setminus \{\ell\}, \\ \mathbf{g}'_{(\ell)} &= \mathbf{g}^\star_{(\ell)} + s\delta, \end{aligned}$$

where

$$\text{(B.7)} \qquad s \triangleq \begin{cases} \operatorname{sign}\big(\mathbf{g}^\star_{(\ell)}\big) & \text{if } \mathbf{g}^\star_{(\ell)} \neq 0 \\ \operatorname{sign}\big(\mathbf{x}^\star_{(\ell)}\big) & \text{otherwise,} \end{cases}$$

and δ is any nonnegative scalar such that

$$\text{(B.8)} \qquad r_{\text{SLOPE},*}(\mathbf{g}') \leq 1.$$

On the one hand, we note that (B.8) is verified for $\delta = 0$. On the other hand, it can be seen that (B.8) is violated as soon as $\delta > 0$ by using the following arguments. First, applying Lemma A.2 with $\mathbf{g}'$ defined as in (B.6) leads to

$$\text{(B.9)} \qquad -s\mathbf{x}^\star_{(\ell)}\delta \geq 0.$$

Second, using (B.5) and the definition of s in (B.7), we must have $s\mathbf{x}^\star_{(\ell)} > 0$. Hence, satisfying inequality (B.8) necessarily implies that $\delta = 0$. The contraposition of this result implies:

$$\text{(B.10)} \qquad \forall \delta > 0, \exists q_0 \in [\![1, n]\!]: \ \sum_{k=1}^{q_0} |\mathbf{g}^\star|_{[k]} + \delta > \sum_{k=1}^{q_0} \gamma_k$$

or equivalently

$$\text{(B.11)} \qquad \exists q_0 \in [\![1, n]\!]: \ \sum_{k=1}^{q_0} |\mathbf{g}^\star|_{[k]} = \sum_{k=1}^{q_0} \gamma_k.$$

Let us next emphasize that the range of values for q_0 can be restricted by choosing some suitable value for δ. In particular, define $q_0' \in [\![1, n]\!]$ as

$$\text{(B.12)} \qquad q_0' \triangleq \min\Big\{ q \in [\![1, n]\!] \colon |\mathbf{g}^\star_{(\ell)}| = |\mathbf{g}^\star|_{[q]} \Big\}$$

and let

$$\text{(B.13)} \qquad 0 < \delta < |\mathbf{g}^\star|_{[q_0'-1]} - |\mathbf{g}^\star|_{[q_0']}$$

with the convention $\mathbf{g}^\star_{[0]} = +\infty$. Considering $\mathbf{g}'$ as defined in (B.6) with δ satisfying (B.13), we have that the first $q_0' - 1$ largest absolute elements of $\mathbf{g}'$ and $\mathbf{g}^\star$ are the same. Since $r_{\text{SLOPE},*}(\mathbf{g}^\star) \leq 1$, the inequality in the right-hand side of (B.10) can therefore not be verified for $q_0 \in [\![1, q_0' - 1]\!]$. Hence, considering δ as in (B.13), we have

$$\text{(B.14)} \qquad \exists q_0 \in [\![q_0', n]\!]: \ \sum_{k=1}^{q_0} |\mathbf{g}^\star|_{[k]} = \sum_{k=1}^{q_0} \gamma_k.$$

We finally obtain our original assertion (B.1) by using the definition of $\mathbf{g}^\star$ in (B.3) and the fact that

$$\sum_{k=1}^{q_0} \left|\mathbf{A}^{\mathrm{T}}\mathbf{u}^\star\right|_{[k]} = \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star\right| + \sum_{k=1}^{q_0-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}^\star\right|_{[k]} \tag{B.15}$$

since $|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star| = |\mathbf{A}^{\mathrm{T}}\mathbf{u}^\star|_{[q_0']}$ by definition of q_0' in (B.12) and $|\mathbf{A}^{\mathrm{T}}\mathbf{u}^\star|_{[q_0']} \geq |\mathbf{A}^{\mathrm{T}}\mathbf{u}^\star|_{[q_0]}$ by definition of $q_0 \geq q_0'$.

B.2. Proof of Lemma 4.2. We first state and prove the following technical lemma:

LEMMA B.1. *Let* $\mathbf{g} \in \mathbb{R}^n$ *and* $\mathbf{h} \in \mathbb{R}^n$ *be such that* $\mathbf{g}_{(j)} \leq \mathbf{h}_{(j)}\ \forall j \in [\![1,n]\!]$. *Then*

$$\mathbf{g}_{[k]} \leq \mathbf{h}_{[k]} \quad \forall k \in [\![1,n]\!]. \tag{B.16}$$

Proof. Let $k \in [\![1,n]\!]$. We have by definition

$$\begin{aligned}
\mathbf{h}_{[k]} &= \max_{\substack{\mathcal{J}\subseteq[\![1,n]\!]\\ \mathrm{card}(\mathcal{J})=k}} \min_{j\in\mathcal{J}} \mathbf{h}_{(j)},\\
&\geq \max_{\substack{\mathcal{J}\subseteq[\![1,n]\!]\\ \mathrm{card}(\mathcal{J})=k}} \min_{j\in\mathcal{J}} \mathbf{g}_{(j)},\\
&= \mathbf{g}_{[k]},
\end{aligned}$$

where the inequality follows from our assumption $\mathbf{g}_{(j)} \leq \mathbf{h}_{(j)}\ \forall j \in [\![1,n]\!]$. ∎

We are now ready to prove Lemma 4.2. For any $p \in [\![1,q]\!]$, we can write:

$$\left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star\right| + \sum_{k=1}^{q-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}^\star\right|_{[k]} = \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star\right| + \sum_{k=1}^{p-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}^\star\right|_{[k]} + \sum_{k=p}^{q-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}^\star\right|_{[k]}. \tag{B.17}$$

First, since $\mathbf{u}^\star$ is dual feasible, we have:

$$\sum_{k=1}^{p-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}^\star\right|_{[k]} \leq \lambda \sum_{k=1}^{p-1} \gamma_k. \tag{B.18}$$

We next show that if $\mathbf{u}^\star \in \mathcal{S}(\mathbf{c}, R)$, then

$$\left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star\right| + \sum_{k=p}^{q-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}^\star\right|_{[k]} \leq \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}\right| + \sum_{k=p}^{q-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{c}\right|_{[k]} + (q-p+1)R. \tag{B.19}$$

We then obtain the result stated in the lemma by combining (B.18)-(B.19).

Inequality (B.19) can be shown as follows. First,

$$\forall j \in [\![1,n]\!] : \max_{\mathbf{u}\in\mathcal{S}(\mathbf{c},R)} |\mathbf{a}_j^{\mathrm{T}}\mathbf{u}| = |\mathbf{a}_j^{\mathrm{T}}\mathbf{c}| + R. \tag{B.20}$$

Hence,

$$\left(\max_{\mathbf{u}\in\mathcal{S}(\mathbf{c},R)} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}\right|\right)_{[k]} = \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{c}\right|_{[k]} + R \tag{B.21}$$

where the maximum is taken component-wise in the left-hand side of the equation. Applying Lemma B.1 with $\mathbf{g} = |\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}|$ and $\mathbf{h} = \max_{\tilde{\mathbf{u}}\in\mathcal{S}(\mathbf{c},R)} |\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\tilde{\mathbf{u}}|$, we have

$$\forall \mathbf{u}\in\mathcal{S}(\mathbf{c},R):\ \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}\right|_{[k]} \leq \left(\max_{\tilde{\mathbf{u}}\in\mathcal{S}(\mathbf{c},R)} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\tilde{\mathbf{u}}\right|\right)_{[k]} \tag{B.22}$$

and therefore

$$\max_{\mathbf{u}\in\mathcal{S}(\mathbf{c},R)}\left(\left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}\right|_{[k]}\right) \leq \left(\max_{\mathbf{u}\in\mathcal{S}(\mathbf{c},R)} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}\right|\right)_{[k]}. \tag{B.23}$$

Combining these results leads to

$$\begin{aligned}
\left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}^\star\right| + \sum_{k=p}^{q-1}\left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}^\star\right|_{[k]} &\leq \max_{\mathbf{u}\in\mathcal{S}(\mathbf{c},R)}\left(\left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}\right| + \sum_{k=p}^{q-1}\left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}\right|_{[k]}\right)\\
&\leq \max_{\mathbf{u}\in\mathcal{S}(\mathbf{c},R)}\left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}\right| + \sum_{k=p}^{q-1}\max_{\mathbf{u}\in\mathcal{S}(\mathbf{c},R)}\left(\left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}\right|_{[k]}\right)\\
&\leq \max_{\mathbf{u}\in\mathcal{S}(\mathbf{c},R)}\left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{u}\right| + \sum_{k=p}^{q-1}\left(\max_{\mathbf{u}\in\mathcal{S}(\mathbf{c},R)}\left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{u}\right|\right)_{[k]}\\
&\leq \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}\right| + \sum_{k=p}^{q-1}\left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{c}\right|_{[k]} + (q-p+1)R.
\end{aligned}$$

B.3. Proof of Lemma 4.4. We want to show that if test (4.10) is passed for some $\{p_q\}_{q\in[\![1,n]\!]}$, then test (4.14) is also passed when $\gamma_k = 1\ \forall k\in[\![1,n]\!]$.

Assume (4.10) holds for some $\{p_q\}_{q\in[\![1,n]\!]}$, that is $\forall q\in[\![1,n]\!]$, $\exists p_q\in[\![1,q]\!]$ such that

$$\left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}\right| + \sum_{k=p_q}^{q-1}\left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{c}\right|_{[k]} < \kappa_{q,p_q}, \tag{B.24}$$

where $\kappa_{q,p} \triangleq \lambda\left(\sum_{k=p}^{q}\gamma_k\right) - (q-p+1)R$. Considering the case "$q=1$", we have $p_1 = 1$, $\kappa_{1,1} = \lambda\gamma_1 - R$ and (B.24) thus particularizes to

$$\left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}\right| < \lambda\gamma_1 - R. \tag{B.25}$$

Since $\gamma_k = 1\ \forall k\in[\![1,n]\!]$ by hypothesis, the latter inequality is equal to (4.14) and the result is proved.

B.4. Proof of Lemma 4.5. We prove the result by showing that $\forall q\in[\![1,n]\!]$ the sequence $\{B_{q,\ell}\}_{\ell\in[\![1,n]\!]}$ is non-increasing. To this end, we first rewrite $B_{q,\ell}$ in a slightly different manner, easier to analyze. Let

$$\begin{aligned}
C_{q,p} &\triangleq\ (q-p+1)R + \lambda\left(\textstyle\sum_{k=1}^{p-1}\gamma_k\right) && \forall q\in[\![1,n]\!], \forall p\in[\![1,q]\!]\\
\sigma_q &\triangleq\ \textstyle\sum_{k=1}^{q}\left|\mathbf{a}_k^{\mathrm{T}}\mathbf{c}\right| && \forall q\in[\![0,n]\!]
\end{aligned} \tag{B.26}$$

with the convention $\sigma_0 \triangleq 0$. Using these notations and hypothesis (4.16), $B_{q,\ell}$ can be rewritten as

$$B_{q,\ell} - C_{q,p} = \left|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}\right| + \sum_{k=1}^{q-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{c}\right|_{(k)} - \sum_{k=1}^{p-1} \left|\mathbf{A}_{\backslash\ell}^{\mathrm{T}}\mathbf{c}\right|_{(k)} \tag{B.27}$$

$$= \begin{cases} |\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}| + \sigma_{q-1} - \sigma_{p-1} & \text{if } q < \ell \\ \sigma_q - \sigma_{p-1} & \text{if } p-1 < \ell \leq q \\ |\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}| + \sigma_q - \sigma_p & \text{if } \ell \leq p-1. \end{cases} \tag{B.28}$$

We next show that $\forall q \in [\![1, n]\!]$ the sequence $\{B_{q,\ell}\}_{\ell \in [\![1,n]\!]}$ is non-increasing. We first notice that $C_{q,p}$ does not depend on ℓ and we can therefore focus on (B.28) to prove our claim. Using the fact that $|\mathbf{a}_\ell^{\mathrm{T}}\mathbf{c}| \geq |\mathbf{a}_{\ell+1}^{\mathrm{T}}\mathbf{c}|$ by hypothesis, we immediately obtain that $B_{q,\ell} \geq B_{q,\ell+1}$ whenever $\ell \notin \{p-1, q\}$. We conclude the proof by treating the cases "$\ell = p-1$" and "$\ell = q$" separately.

If $\ell = q$ we have from (B.28):

$$B_{q,\ell+1} - B_{q,\ell} = |\mathbf{a}_{q+1}^{\mathrm{T}}\mathbf{c}| + \sigma_{q-1} - \sigma_q = |\mathbf{a}_{q+1}^{\mathrm{T}}\mathbf{c}| - |\mathbf{a}_q^{\mathrm{T}}\mathbf{c}| \leq 0, \tag{B.29}$$

where the last inequality holds true by virtue of (4.16).

If $\ell = p-1$ (and provided that $p \geq 2$) the same rationale leads to

$$B_{q,\ell+1} - B_{q,\ell} = |\mathbf{a}_p^{\mathrm{T}}\mathbf{c}| - |\mathbf{a}_{p-1}^{\mathrm{T}}\mathbf{c}| \leq 0. \tag{B.30}$$

B.5. Proof of Lemma 4.6. The necessity of (4.28) can be shown as follows. Assume $|\mathbf{a}_n^{\mathrm{T}}\mathbf{c}| \geq \tau$ for some $\tau \in \mathcal{T}$ and let $q \in [\![1, n]\!]$ be such that $\tau = \tau_{q,p^\star(q)}$. From (4.22) we then have

$$\forall p \in [\![1, q]\!]: \ |\mathbf{a}_n^{\mathrm{T}}\mathbf{c}| \geq \tau_{q,p} \tag{B.31}$$

and test (4.19) therefore fails.

To prove the sufficiency of (4.28), let us first notice that the definition of $\tau_{q,p}$ given in (4.24) can be naturally extended to any arbitrary couple of indices $q, p \in [\![1, n]\!]$, *i.e.*,

$$\forall q, p \in [\![1, n]\!]: \quad \tau_{q,p} = g(p) - (g(q) - \lambda\gamma_q) - R. \tag{B.32}$$

On the other hand, the index $q^{(1)}$ has been defined as

$$q^{(1)} \triangleq q^\star(n) = \operatorname*{arg\,max}_{q \in [\![1,n]\!]} g(q) - \lambda\gamma_q, \tag{B.33}$$

see (4.26) and (4.27). Combining (B.32) and (B.33), one obtains $\forall p \in [\![1, n]\!]$:

$$\tau_{q^{(1)},p} = \operatorname*{arg\,min}_{q \in [\![1,n]\!]} \tau_{q,p}. \tag{B.34}$$

In particular, letting $p = p^{(1)}$, we have

$$\forall q \in [\![p^{(1)}, n]\!]: \ \tau_{q^{(1)},p^{(1)}} \leq \tau_{q,p^{(1)}}. \tag{B.35}$$

Hence,

$$|\mathbf{a}_n^{\mathrm{T}}\mathbf{c}| < \tau_{q^{(1)},p^{(1)}} \implies \forall q \in [\![p^{(1)}, n]\!]: \ |\mathbf{a}_n^{\mathrm{T}}\mathbf{c}| < \tau_{q,p^{(1)}}. \tag{B.36}$$

In other words, satisfying the left-hand side of (B.36) implies that test (4.19) is verified for each $q \in [\![p^{(1)}, n]\!]$.

We can apply the same reasoning iteratively to show that $\forall t \in [\![1, \mathrm{card}\,(\mathcal{T})]\!]$:

$$|\mathbf{a}_n^{\mathrm{T}}\mathbf{c}| < \tau_{q^{(t)},p^{(t)}} \implies \forall q \in [\![p^{(t)}, p^{(t-1)} - 1]\!] : \; |\mathbf{a}_n^{\mathrm{T}}\mathbf{c}| < \tau_{q,p^{(t)}}. \tag{B.37}$$

Since $p^{(\mathrm{card}(\mathcal{T}))} = 1$, we obtain that (4.28) implies that (4.19) is verified $\forall q \in [\![1, n]\!]$.

REFERENCES

[1] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, *Convex optimization with sparsity-inducing norms*, in Optimization for Machine Learning, Neural information processing series, MIT Press, 2011, pp. 19–49, https://doi.org/10.7551/mitpress/8996.003.0004.

[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, *Optimization with sparsity-inducing penalties*, Foundations and Trends®in Machine Learning, 4 (2012), pp. 1–106, https://doi.org/10.1561/2200000015.

[3] R. Bao, B. Gu, and H. Huang, *Fast OSCAR and OWL with safe screening rules*, in Proceedings of the 37th International Conference on Machine Learning, 2020.

[4] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer International Publishing, 2017, https://doi.org/10.1007/978-3-319-48311-5.

[5] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès, *SLOPE—adaptive variable selection via convex optimization*, The Annals of Applied Statistics, 9 (2015), pp. 1103–1140, https://doi.org/10.1214/15-aoas842.

[6] M. Bogdan, E. van den Berg, W. Su, and E. Candes, *Statistical estimation and testing via the sorted l1 norm*, 2013, https://arxiv.org/abs/1310.1969.

[7] H. D. Bondell and B. J. Reich, *Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR*, Biometrics, 64 (2007), pp. 115–123, https://doi.org/10.1111/j.1541-0420.2007.00843.x.

[8] N. Boyd, G. Schiebinger, and B. Recht, *The alternating descent conditional gradient method for sparse inverse problems*, SIAM Journal on Optimization, 27 (2017), pp. 616–639.

[9] D. Brzyski, A. Gossmann, W. Su, and M. Bogdan, *Group SLOPE –adaptive selection of groups of predictors*, Journal of the American Statistical Association, 114 (2019), pp. 419–433, https://doi.org/10.1080/01621459.2017.1411269.

[10] Z. Bu, J. Klusowski, C. Rush, and W. Su, *Algorithmic analysis and statistical estimation of SLOPE via approximate message passing*, in Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 9366–9376.

[11] S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by Basis Pursuit*, SIAM J. Sci. Comp., 20 (1999), pp. 33–61.

[12] L. Dai and K. Pelckmans, *An ellipsoid based, two-stage screening test for BPDN*, in European Signal Processing Conference (EUSIPCO), IEEE, August 2012, pp. 654–658.

[13] C. F. Dantas, E. Soubies, and C. Févotte, *Expanding Boundaries of Gap Safe Screening*, Journal of Machine Learning Research, 22 (2021), pp. 1–57, http://jmlr.org/papers/v22/21-0179.html.

[14] D. Davis, *An $o(n\log(n))$ algorithm for projecting onto the ordered weighted ℓ_1 norm ball*. arXiv-1505.00870, 2015, https://arxiv.org/abs/1505.00870.

[15] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Mathematical Programming, 91 (2002), pp. 201–213, https://doi.org/10.1007/s101070100263.

[16] L. El Gueddari, E. Chouzenoux, A. Vignaud, and P. Ciuciu, *Calibration-less parallel imaging compressed sensing reconstruction based on OSCAR regularization*. Preprint hal-02292372, Sept. 2019, https://hal.inria.fr/hal-02292372.

[17] C. Elvira and C. Herzet, *Safe squeezing for antisparse coding*, IEEE Transactions on Signal Processing, 68 (2020), pp. 3252–3265, https://doi.org/10.1109/TSP.2020.2995192.

[18] C. Elvira and C. Herzet, *Short and squeezed: Accelerating the computation of antisparse representations with safe squeezing*, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 5615–5619, https://doi.org/10.1109/ICASSP40776.2020.9053156.

[19] C. Elvira and C. Herzet, *A response to "fast OSCAR and OWL regression via safe screening rules" by bao et al.*, tech. report, October 2021, https://c-elvira.github.io/pdf/tecreports/Elvira2021_responseSLOPE.pdf. Technical report.

[20] O. Fercoq, A. Gramfort, and J. Salmon, *Mind the duality gap: safer rules for the Lasso*, in Proceedings of the 32nd International Conference on Machine Learning, vol. 37 of Proceedings of Machine Learning Research, Lille, France, 07–09 Jul 2015, PMLR, pp. 333–342.

[21] M. Figueiredo and R. Nowak, *Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, vol. 51, Cadiz, Spain, May 2016, PMLR, pp. 930–938, http://proceedings.mlr.press/v51/figueiredo16.html.

[22] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing.*, Ap-

plied and Numerical Harmonic Analysis, Birkhaüser, 2013, http://www.springer.com/birkhauser/mathematics/book/978-0-8176-4947-0.

[23] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Research Logistics (NRL), 3 (1956), pp. 95–110.

[24] L. Ghaoui, V. Viallon, and T. Rabbani, *Safe feature elimination in sparse supervised learning*, Pacific Journal of Optimization, 8 (2010).

[25] A. Gossmann, S. Cao, D. Brzyski, L. J. Zhao, H. W. Deng, and Y. P. Wang, *A sparse regression method for group-wise feature selection with false discovery rate control*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15 (2018), pp. 1066–1078, https://doi.org/10.1109/TCBB.2017.2780106.

[26] A. Gossmann, S. Cao, and Y.-P. Wang, *Identification of significant genetic variants via SLOPE, and its extension to group SLOPE*, in Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics, BCB '15, New York, USA, 2015, pp. 232–240, https://doi.org/10.1145/2808719.2808743.

[27] T. Guyard, C. Herzet, and C. Elvira, *Screen & relax: Accelerating the resolution of Elastic-Net by safe identification of the solution support*, in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 5443–5447, https://doi.org/10.1109/ICASSP43922.2022.9747412.

[28] C. Herzet, C. Dorffer, and A. Drémeau, *Gather and conquer: Region-based strategies to accelerate safe screening tests*, IEEE Transactions on Signal Processing, 67 (2019), pp. 3300–3315, https://doi.org/10.1109/TSP.2019.2914885.

[29] C. Herzet and A. Malti, *Safe screening tests for LASSO based on firmly non-expansiveness*, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2016, pp. 4732–4736, https://doi.org/10.1109/ICASSP.2016.7472575.

[30] P. Kremer, D. Brzyski, M. Bogdan, and S. Paterlini, *Sparse index clones via the sorted l1-norm*. en. SSRN Scholarly Paper ID 3412061. Rochester, NY: Social Science Research Network, June 2019.

[31] P. J. Kremer, S. Lee, M. Bogdan, and S. Paterlini, *Sparse portfolio selection via the sorted ℓ_1-norm*, Journal of Banking & Finance, 110 (2020).

[32] J. Larsson, M. Bogdan, and J. Wallin, *The strong screening rule for SLOPE*, 2020, https://arxiv.org/abs/2005.03730.

[33] G. Lecué and S. Mendelson, *Regularization and the small-ball method I: sparse recovery*. arXiv-1601.05584, 2017.

[34] J. Liu, Z. Zhao, J. Wang, and J. Ye, *Safe screening with variational inequalities and its application to Lasso*, in Proceedings of the 31st International Conference on Machine Learning, JMLR Workshop and Conference Proceedings, 2014, pp. 289–297, http://jmlr.org/proceedings/papers/v32/liuc14.pdf.

[35] Z. Luo, D. Sun, K. Chuan Toh, and N. Xiu, *Solving the OSCAR and SLOPE models using a semismooth newton-based augmented lagrangian method*, Journal of Machine Learning Research, 20 (2019), pp. 1–25.

[36] E. Ndiaye, O. Fercoq, Alex, re Gramfort, and J. Salmon, *Gap safe screening rules for sparsity enforcing penalties*, Journal of Machine Learning Research, 18 (2017), pp. 1–33, http://jmlr.org/papers/v18/16-577.html.

[37] Ndiaye, Eugene and Fercoq, Olivier and Salmon, Joseph, *Screening rules and its complexity for active set identification*, 28, https://www.heldermann.de/JCA/JCA28/jca28.htm#jca284.

[38] U. Oswal and R. Nowak, *Scalable sparse subspace clustering via ordered weighted l1 regression*, in 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2018, pp. 305–312, https://doi.org/10.1109/ALLERTON.2018.8635965.

[39] U. Schneider and P. Tardivel, *The geometry of uniqueness, sparsity and clustering in penalized estimation*, 2020, https://arxiv.org/abs/2004.09106.

[40] W. Su and E. Candes, *SLOPE is adaptive to unknown sparsity and asymptotically minimax*, Ann. Statist., 44 (2016), pp. 1038–1068, https://doi.org/10.1214/15-AOS1397.

[41] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, *Strong rules for discarding predictors in lasso-type problems*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74 (2012), pp. 245–266, https://doi.org/10.1111/j.1467-9868.2011.01004.x.

[42] T.-L. Tran, C. Elvira, H.-P. Dang, and C. Herzet, *Beyond GAP screening for Lasso by exploiting new dual cutting half-spaces with supplementary material*. arXiv 2203.00987, Feb. 2022.

[43] J. Wang, P. Wonka, and J. Ye, *Lasso screening rules via dual polytope projection*, Journal of Machine Learning Research, (2015).

[44] Z. J. Xiang and P. J. Ramadge, *Fast lasso screening tests based on correlations*, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, march 2012, pp. 2137–2140, https://doi.org/10.1109/icassp.2012.6288334.

[45] Z. J. Xiang, Y. Wang, and P. J. Ramadge, *Screening tests for Lasso problems*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (2017), pp. 1008–1027, https://doi.org/10.1109/TPAMI.2016.2568185.

[46] Z. J. Xiang, H. Xu, and P. J. Ramadge, *Learning sparse representations of high dimensional data on large scale dictionaries*, in Advances in Neural Information Processing Systems, 2011, pp. 900–908, http://books.nips.cc/papers/files/nips24/NIPS2011_0578.pdf.

[47] X. Xing, J. Hu, and Y. Yang, *Robust minimum variance portfolio with l-infinity constraints*, Journal of Banking & Finance, 46 (2014), pp. 107–117, https://doi.org/10.1016/j.jbankfin.2014.05.004.

[48] X. Zeng and M. A. T. Figueiredo, *The atomic norm formulation of OSCAR regularization with application to the frank-wolfe algorithm*, in European Signal Processing Conference (EUSIPCO), 2014, pp. 780–784.

[49] X. Zeng and M. A. T. Figueiredo, *Decreasing weighted sorted ℓ_1 regularization*, IEEE Signal Processing Letters, 21 (2014), pp. 1240–1244, https://doi.org/10.1109/lsp.2014.2331977, https://doi.org/10.1109/lsp.2014.2331977.

[50] D. Zhang, H. Wang, M. Figueiredo, and L. Balzano, *Learning to share: simultaneous parameter tying and sparsification in deep learning*, in International Conference on Learning Representations, 2018, https://openreview.net/forum?id=rypT3fb0b.

[51] W. Zhong and J. Kwok, *Efficient sparse modeling with automatic feature grouping*, in Proceedings of the 28th International Conference on Machine Learning, ICML '11, New York, USA, June 2011, ACM, pp. 9–16.