Gather and Conquer: Region-based Strategies to Accelerate Safe Screening Tests

Cédric Herzet, Clément Dorffer, Angélique Drémeau

Abstract—In this paper, we propose new methodologies to decrease the computational cost of safe screening tests for LASSO. We first introduce a new screening strategy, dubbed "joint screening test", which allows the rejection of a *set* of atoms by performing *one* single test. Our approach enables to find good compromises between complexity of implementation and effectiveness of screening. Second, we propose two new methods to decrease the computational cost inherent to the construction of the (so-called) "safe region". Our numerical experiments show that the proposed procedures lead to significant computational gains as compared to standard methodologies.

I. INTRODUCTION

In the last decade, sparse representations have proven to be powerful tools to solve many problems in signal processing, machine learning, etc. A standard methodology to find a sparse representation of some $\mathbf{y} \in \mathbb{R}^m$ in a dictionary $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_n] \in \mathbb{R}^{m \times n}$ is the so-called (nonnegative) LASSO problem:[1]

$$\mathbf{x}_\lambda^\star \in \arg\min_{\mathbf{x} \geq \mathbf{0}} P_\lambda(\mathbf{y}, \mathbf{x}), \tag{1}$$

where $P_\lambda(\mathbf{y}, \mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$ and $\mathbf{x} \in \mathbb{R}^n$. Without loss of generality, we will assume hereafter that $\|\mathbf{a}_i\|_2 = 1$ for $i = 1 \ldots n$.

Solving (1) may require a heavy computational load when the dimension of $\mathbf{x}$ becomes large. Therefore, the conception of computationally-efficient techniques to solve (1) has become an active field of research. Among the most popular approaches addressing (1), one can mention interior-point methods [2], proximal gradient algorithms ([3], [4]), homotopy procedures ([5], [6]), splitting methods ([1], [7]) or Frank-Wolfe methodologies [8].

One important contribution in this field is the so-called "safe"[2] screening technique proposed by El Ghaoui *et al.* in [9]. The idea of safe screening consists in identifying *some* of the zeros of $\mathbf{x}_\lambda^\star$ with a low computational burden. These elements can then be withdrawn from the optimization variables, resulting in a problem of smaller dimension.

The identification of zeros is achieved through the implementation of simple tests. The overall computational gain reached by screening techniques is thus a compromise between their induced computational burden and the number of zeros

they can identify. To find a good compromise between these two objectives, the main trend in the current literature has been to improve the definition of the so-called "safe region", a set of the dual space containing the dual optimal solution of (1). In particular, different geometries and definitions for the safe region have been proposed in a series of contributions [10]–[17].

In this paper, we consider another avenue of research. We first note that, in order to implement their screening tests, all contributions mentioned above require to compute (at least) one inner product in $\mathbb{R}^m$ *per* atom in the dictionary. Hence, the cost associated to the implementation of the screening test may become prohibitive when the size of the dictionary becomes very large. As a limit case, one can think of the problem of sparse representations in "continuous" dictionaries, where the number of atoms is infinite uncountable, see *e.g.*, [18]–[21].[3] In this context, standard screening methodologies simply become inapplicable.

In this paper, we propose a new screening methodology to deal with this issue. We first introduce a new procedure, dubbed "joint screening", allowing to safely screen a *set* of atoms while performing one single test. In particular, we show that a careful design of the proposed method allows to achieve a better compromise between the effectiveness and the efficiency of the screening tests. Following the same line of thought, we then introduce new methodologies to alleviate the complexity associated to the construction of the safe region.

This work is an extended version of [22] in which we laid the main foundations of joint screening. The present paper contains all the proofs of our results, extended discussions and developments, as well as extensive simulation results. We also note that the material presented hereafter share some connections with a parallel work [23]. In particular, the authors of [23] came to the same expressions as those presented in Theorems 1 and 2 below in the case of "sphere" regions. Nevertheless, apart from this connection, the two works are distinct since [23] consider the problem of screening with approximate dictionaries (but having the same size as the original dictionary) whereas this paper focus on the problem of screening *sets* of atoms.

The paper is organized as follows. In Section II, we recall the main principles underlying safe screening methodologies. In Section III we introduce the proposed joint screening tests and discuss the tuning of their parameters. In Section IV, we

---

[1]The standard LASSO can be written as a particular case of (1), see [1].

[2]The term "safe" refers to the fact that the elements identified by the screening method always correspond to zeros in $\mathbf{x}_\lambda^\star$.

[3]Let us mention that, although the methodologies proposed in this paper carry over to "continuous" dictionaries, hereafter we restrict our attention to the case of discrete dictionaries. Rigorous derivations in the more-complex formalism of continuous dictionaries is left for future work.

tackle the problem of identifying a safe region. In particular, we propose several approaches to alleviate the complexity of the so-called "dual scaling" method. Finally, in Section V we assess numerically the effectiveness of the proposed methodologies.

II. STANDARD PROCEDURES FOR SAFE SCREENING

In this section, we give a brief reminder of the rationale underlying safe screening methods (Sections II-A, II-B and II-C) and emphasize some of their computational bottlenecks (Section II-D). In a last subsection (Section II-E) we motivate the working hypotheses that will be used throughout the paper.

A. Some convex considerations

Problem (1) is convex and always admits (at least) one solution. The dual problem associated to (1) can be written as (see for example [24])

$$\mathbf{z}_\lambda^\star = \arg\max_{\mathbf{z} \in \mathcal{Z}} D_\lambda(\mathbf{y}, \mathbf{z}), \tag{2}$$

where

$$D_\lambda(\mathbf{y}, \mathbf{z}) = \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2}\left\|\frac{\mathbf{y}}{\lambda} - \mathbf{z}\right\|_2^2, \tag{3}$$

$$\mathcal{Z} = \{\mathbf{z} \in \mathbb{R}^m : \langle \mathbf{a}_i, \mathbf{z} \rangle \leq 1, \ i = 1\ldots n\}, \tag{4}$$

and $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^m$. Since $D_\lambda(\mathbf{y}, \mathbf{z})$ is a strictly concave coercive function and $\mathcal{Z}$ is a closed set, problem (2) admits a unique solution $\mathbf{z}_\lambda^\star$, see [25, Proposition A.8].

The primal and dual solutions $(\mathbf{x}_\lambda^\star, \mathbf{z}_\lambda^\star)$ are related through the well-known Karush-Kuhn-Tucker (KKT) conditions [25, Proposition 5.1.5]:

$$\mathbf{x}_\lambda^\star \geq \mathbf{0}, \ \langle \mathbf{a}_i, \mathbf{z}_\lambda^\star \rangle \leq 1 \text{ for all } i, \tag{5}$$

$$(\langle \mathbf{a}_i, \mathbf{z}_\lambda^\star \rangle - 1)\, \mathbf{x}_\lambda^\star(i) = 0 \ \text{ for all } i, \tag{6}$$

$$\mathbf{y} = \lambda \mathbf{z}_\lambda^\star + \mathbf{A}\mathbf{x}_\lambda^\star, \tag{7}$$

where $\mathbf{x}_\lambda^\star(i)$ denotes the ith component of $\mathbf{x}_\lambda^\star$. The KKT conditions are necessary and sufficient optimality conditions for $(\mathbf{x}_\lambda^\star, \mathbf{z}_\lambda^\star)$. We will see in the next section that safe screening methods exploit the necessity of these conditions to identify some zeros of $\mathbf{x}_\lambda^\star$.

B. Safe screening: standard methodologies

Safe screening procedures, first proposed in [9], leverage the following observation: if $\mathcal{S} \subset \mathbb{R}^m$ is a region such that $\mathbf{z}_\lambda^\star \in \mathcal{S}$,[4] then the following inequality trivially holds

$$\langle \mathbf{a}_i, \mathbf{z}_\lambda^\star \rangle \leq \max_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{a}_i, \mathbf{z} \rangle,$$

and from (6), we thus have

$$\max_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{a}_i, \mathbf{z} \rangle < 1 \Rightarrow \mathbf{x}_\lambda^\star(i) = 0. \tag{8}$$

In other words, if the inequality in the left-hand side of (8) is satisfied, one is ensured that the ith component of the solution vector $\mathbf{x}_\lambda^\star$ is equal to zero.

[4]Such a region is commonly referred to as "safe region" in the screening literature.

Since the seminal work [9] by El Ghaoui and co-authors, different screening tests, based on different choices of $\mathcal{S}$, have been proposed in the literature, see [10]–[15]. The most popular ones are probably the tests based on "sphere" regions,[5] that is

$$\mathcal{S} = \mathcal{B}(\mathbf{c}, 1 - \tau) \triangleq \{\mathbf{z} : \|\mathbf{z} - \mathbf{c}\|_2 \leq 1 - \tau\}, \tag{9}$$

for some parameters $\mathbf{c} \in \mathbb{R}^m$ and $\tau \leq 1$. Interestingly, for this particular choice, the general screening test (8) takes the following simple form:

$$\langle \mathbf{a}_i, \mathbf{c} \rangle < \tau \Rightarrow \mathbf{x}_\lambda^\star(i) = 0. \tag{10}$$

Other geometries for $\mathcal{S}$ (e.g., domes), leading to slightly more involved screening tests, have also been considered in [16], [17]. In this paper, we exclusively focus on spherical safe regions, although the material presented hereafter can also be extended to more complex definitions of $\mathcal{S}$.

C. Finding a safe region

A crucial step in the design of a safe screening test is the identification of a safe region $\mathcal{S}$, that is a region such that $\mathbf{z}_\lambda^\star \in \mathcal{S}$. Many methods addressing this task have been proposed in the literature, see e.g., [9]–[17]. As far as spherical regions are concerned, we can for example mention the "ST1" sphere introduced in [9], whose center $\mathbf{c}$ and radius $1 - \tau$ are defined as:

$$\begin{aligned} \mathbf{c} &= \frac{\mathbf{y}}{\lambda}, \\ 1 - \tau &= \left\|\frac{\mathbf{y}}{\lambda} - \mathbf{z}\right\|_2, \end{aligned} \tag{11}$$

where $\mathbf{z} \in \mathcal{Z}$ is some dual feasible point. Another (more recent) example is the so-called "GAP" sphere [14] for which

$$\begin{aligned} \mathbf{c} &= \mathbf{z}, \\ 1 - \tau &= \sqrt{2\frac{P_\lambda(\mathbf{y}, \mathbf{x}) - D_\lambda(\mathbf{y}, \mathbf{z})}{\lambda^2}}, \end{aligned} \tag{12}$$

where $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{z} \in \mathcal{Z}$ can be any primal-dual feasible points.

A common ingredient to all the methodologies proposed so far to define a safe region $\mathcal{S}$, is the identification of a dual feasible point $\mathbf{z} \in \mathcal{Z}$,[6] that is

$$\langle \mathbf{a}_i, \mathbf{z} \rangle \leq 1 \quad \text{for all } i = 1\ldots n. \tag{13}$$

Unfortunately, in the vast majority of cases, such a dual feasible point is not directly available and its identification must therefore be part of the construction of the safe region. In [9], El Ghaoui et al. suggested a simple procedure, dubbed *"dual scaling"*, to build a dual feasible point from any vector $\tilde{\mathbf{z}} \in \mathbb{R}^m$. Their method is based on the following simple observation: given any $\tilde{\mathbf{z}} \in \mathbb{R}^m$, we have that

$$\mathbf{z} = \left(\max_{1 \leq i \leq n} \langle \mathbf{a}_i, \tilde{\mathbf{z}} \rangle\right)^{-1} \tilde{\mathbf{z}} \tag{14}$$

[5]According to definition (9), the term "ball" would certainly be more appropriate than "sphere". However, since the terminology "sphere" is firmly rooted in the screening community, we will also use it in our article.

[6]For example, (11) and (12) provide two instances of safe regions requiring the knowledge of some dual feasible point $\mathbf{z}$.

is dual feasible. In other words, as long as the inner products between $\tilde{\mathbf{z}} \in \mathbb{R}^m$ and the atoms of the dictionary can be evaluated, one can always compute a dual feasible point via (14). To the best of our knowledge, dual scaling is the standard procedure used in all the contributions of the literature to identify a dual feasible point.

D. Computational cost of standard methodologies

In this section, we comment on the complexity required to implement the safe screening techniques discussed above. Let us first notice that screening can be applied in either a "static" or a "dynamic" way. Static screening means that the screening procedure is applied once before the application of the optimization algorithm addressing (1). Dynamic screening [13] corresponds to the repeated application of a screening technique throughout the iterations of the optimization algorithm. In the latter case, the size of the dictionary can thus be reduced all along the optimization process. In order to address jointly these two cases, we assume in the rest of this paper (with a slight abuse of notation) that n represents the number of atoms in the dictionary before the application of the screening method. In the static case, n thus corresponds to the initial number of atoms in the dictionary; in the dynamic case, n is the number of atoms which have not yet been screened at the current iteration of the optimization process.

With this convention in mind, we have that the computational cost associated to the implementation of safe screening techniques typically evolves linearly with the size n of the dictionary. This order of complexity stems from two particular steps in the screening process.

First, it can be seen from (8) that the implementation of the standard screening test necessitates the evaluation of the quantity "$\max_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{a}_i, \mathbf{z} \rangle$" for *each* atom of the dictionary. Hence, the computational load required to perform this operation obviously evolves linearly with the total number of atoms. For example, in the case where $\mathcal{S}$ is a sphere region, (10) involves the computation of the inner products between the center of the sphere $\mathbf{c}$ and each atom of the dictionary, leading to a complexity scaling as $\mathcal{O}(mn)$.

A second computational bottleneck is related to the definition of the safe region $\mathcal{S}$, more specifically the identification of a dual feasible point. As mentioned in Section II-C, the standard technique to construct such a point is "dual scaling", see (14). Now, the naive application of this method entails the computation of n inner products $\langle \mathbf{a}_i, \mathbf{z} \rangle$, leading to a complexity scaling as $\mathcal{O}(mn)$.

This linear evolution of the complexity may become prohibitive in some applications, in particular those involving very large dictionaries. In this paper, we propose new procedures to reduce this computational cost. We address the two issues mentioned above separately. In Section III, we introduce new screening tests allowing to jointly identify a *set* of zeros in $\mathbf{x}_\lambda^\star$. In Section IV, we address the problem of computing efficiently a dual feasible point. These methodologies are assessed numerically in Section V.

E. Working hypotheses

In this section, we introduce two working assumptions that will be considered in the rest of the paper. These two hypotheses will always be assumed verified even when not explicitly mentioned in the statement of our results.

As a first working hypothesis, we suppose that the penalization parameter λ appearing in the LASSO problem (1) satisfies

$$\lambda < \lambda_{\max}, \qquad (15)$$

where $\lambda_{\max} \triangleq \max_i \langle \mathbf{a}_i, \mathbf{y} \rangle$. This assumption makes sense from a practical point of view since $\mathbf{x}_\lambda^\star \neq \mathbf{0}$ if and only if (15) is satisfied. In other words, considering (1) with $\lambda \geq \lambda_{\max}$ always leads to the all-zero solution.

Our second working hypothesis reads

$$\tau > -\|\mathbf{c}\|_2, \qquad (16)$$

where $\mathbf{c}$ and τ are respectively the center and the radius parameter of the safe sphere considered in the screening test. In order to motivate our assumption, let us first notice that any atom belonging to the safe region $\mathcal{S}$ will never satisfy test (8).[7] Hypothesis (16) then simply ensures that

$$\mathbb{S}_m \not\subseteq \mathcal{B}(\mathbf{c}, \mathbf{1} - \tau), \qquad (17)$$

where $\mathbb{S}_m$ is the m-dimensional unit sphere. In other words, if (16) holds then there exists *at least one* unit-norm vector not included in the safe sphere $\mathcal{B}(\mathbf{c}, \mathbf{1} - \tau)$. This assumption makes sense from a practical point of view since if (17) does not hold, then no atom can be screened anyway.

A consequence of hypotheses (15) and (16) is that we must necessarily have $\mathbf{c} \neq \mathbf{0}$. Indeed, if $\mathbf{c} = \mathbf{0}$, satisfying (16) requires $\tau > 0$. On the other hand, the first hypothesis (15) combined to $\mathbf{c} = \mathbf{0}$ imposes that $\tau \leq 0$ as shown in Lemma 4, Appendix A. Since $\tau > 0$ and $\tau \leq 0$ cannot obviously hold simultaneously, we have that the combination of (15) and (16) discards the case $\mathbf{c} = \mathbf{0}$.

III. JOINT SCREENING TESTS

In this section, we introduce a new screening procedure having a complexity not depending on the number of atoms in the dictionary (or only in a logarithmic factor for the procedure described in Section III-C). We dub our methodology "joint screening test" because it allows to screen a *set* of atoms by carrying out one *single* test. In a first subsection, we derive tests allowing to screen any atom belonging to some specific region $\mathcal{R} \subset \mathbb{R}^m$. In a second subsection, we elaborate on the relative effectiveness of the proposed test for different choices of $\mathcal{R}$. Finally, in the last two subsections, we discuss the choice of the parameters defining the screening region $\mathcal{R}$.

A. Joint screening of a region $\mathcal{R}$

Let $\mathcal{A} = \{\mathbf{a}_i\}_{i=1}^n$ denote the set of atoms of the dictionary and let $\mathcal{S} \subset \mathbb{R}^m$ be a safe region (that is $\mathbf{z}_\lambda^\star \in \mathcal{S}$). The "joint" screening procedure proposed in this paper is a direct consequence of the following observation:[8]

---

[7]This comes straightforwardly from the observation that for any $\mathbf{a}_i \in \mathcal{S}$, $\max_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{a}_i, \mathbf{z} \rangle \geq \langle \mathbf{a}_i, \mathbf{a}_i \rangle = 1 \geq \tau$.

[8]The validity of (18) directly follows from (8).

$$\max_{\mathbf{a} \in \mathcal{R}} \max_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{a}, \mathbf{z} \rangle < 1 \Rightarrow \mathbf{x}_\lambda^\star(i) = 0 \quad \forall i : \mathbf{a}_i \in \mathcal{A} \cap \mathcal{R}. \quad (18)$$

In order words, if the inequality in the left-hand side of (18) is satisfied, all the atoms $\mathbf{a}_i \in \mathcal{A} \cap \mathcal{R}$ can be *safely* and *jointly* screened from problem (1).

In what follows, we will see that the verification of the inequality in the left-hand side of (18) can be done very efficiently for some specific choices of regions $\mathcal{S}$ and $\mathcal{R}$. First, we will assume that $\mathcal{S}$ is a sphere region (9). The joint screening test (18) then takes the simple form:

$$\max_{\mathbf{a} \in \mathcal{R}} \langle \mathbf{a}, \mathbf{c} \rangle < \tau \Rightarrow \mathbf{x}_\lambda^\star(i) = 0 \quad \forall i : \mathbf{a}_i \in \mathcal{A} \cap \mathcal{R}. \quad (19)$$

Moreover, we will consider the two following options for $\mathcal{R}$:

$$\mathcal{R} = \mathcal{B}(\mathbf{t}, \epsilon) \triangleq \{\mathbf{a} : \|\mathbf{a} - \mathbf{t}\|_2 \le \epsilon\}, \quad (20)$$
$$\mathcal{R} = \mathcal{D}(\mathbf{t}, \delta) \triangleq \{\mathbf{a} : \langle \mathbf{a}, \mathbf{t} \rangle \ge \delta, \|\mathbf{a}\|_2 \le 1\}, \quad (21)$$

where $\mathbf{t} \in \mathbb{R}^m$ and ϵ, δ are some parameters. The regions $\mathcal{B}(\mathbf{t}, \epsilon)$ and $\mathcal{D}(\mathbf{t}, \delta)$ have some easy geometric interpretations: $\mathcal{B}(\mathbf{t}, \epsilon)$ corresponds to the set of vectors located in a ball of radius ϵ centered on $\mathbf{t}$; $\mathcal{D}(\mathbf{t}, \delta)$ is a dome including all the vectors of norm smaller than one and having an inner product with $\mathbf{t}$ greater than or equal to δ.

For these two choices of regions, the joint screening test defined in (19) admits very simple analytical solutions, as shown by the following result:

Theorem 1. $\max_{\mathbf{a} \in \mathcal{B}(\mathbf{t}, \epsilon)} \langle \mathbf{a}, \mathbf{c} \rangle < \tau$ *if and only if*

$$\langle \mathbf{t}, \mathbf{c} \rangle < \tau - \epsilon \|\mathbf{c}\|_2. \quad (22)$$

Moreover, $\max_{\mathbf{a} \in \mathcal{D}(\mathbf{t}, \delta)} \langle \mathbf{a}, \mathbf{c} \rangle < \tau$ *if and only if* [9]

$$\langle \mathbf{t}, \mathbf{c} \rangle < \tau, \quad (23)$$

and

$$\delta > \frac{\langle \mathbf{t}, \mathbf{c} \rangle \tau + \sqrt{\|\mathbf{c}\|_2^2 - \langle \mathbf{t}, \mathbf{c} \rangle^2} \sqrt{\|\mathbf{c}\|_2^2 - \tau^2}}{\|\mathbf{c}\|_2^2}. \quad (24)$$

A proof of this result can be found in Appendix B. We note from (22)-(24) that the joint screening tests (19) based on a sphere region $\mathcal{B}(\mathbf{t}, \epsilon)$ or a dome region $\mathcal{D}(\mathbf{t}, \delta)$ only requires the evaluation of *one* inner product, namely $\langle \mathbf{t}, \mathbf{c} \rangle$; nonetheless, if these tests are passed, they allow to screen *all* the atoms in the regions $\mathcal{B}(\mathbf{t}, \epsilon)$ and $\mathcal{D}(\mathbf{t}, \delta)$. In this paper, we will often refer to $\mathbf{t}$ as "test vector" since it is the only vector appearing in the joint test associated to region $\mathcal{R}$.

Motivated by Theorem 1, we propose the screening procedure described in Algorithm 1. The complexity[10] of this algorithm is dominated by the operations carried out in steps 4, 6 and 7. If the regions $\{\mathcal{R}_l\}_{l=1}^L$ are prespecified once for all

---

Algorithm 1 Joint Screening Test

1: **inputs:** L : number of regions, $\mathcal{A}$: set of atoms
2: **procedure** JOINT SCREENING($L, \mathcal{A}$)
3:     **init:** $\mathcal{A}_{\mathrm{scr}} = \emptyset$
4:     Select a set of regions $\{\mathcal{R}_l\}_{l=1}^L$
5:     **for all** $1 \le l \le L$ **do**
6:         **if** $\max_{\mathbf{a} \in \mathcal{R}_l} \langle \mathbf{a}, \mathbf{c} \rangle < \tau$ **then**
7:             $\mathcal{A}_{\mathrm{scr}} = \mathcal{A}_{\mathrm{scr}} \cup (\mathcal{A} \cap \mathcal{R}_l)$
8:         **end if**
9:     **end for**
10: **end procedure**
11: **outputs:** $\mathcal{A}_{\mathrm{scr}}$: set of screened atoms

---

in advance, the complexity of this procedure is concentrated in the verification of the tests "$\max_{\mathbf{a} \in \mathcal{R}_l} \langle \mathbf{a}, \mathbf{c} \rangle < \tau$". Now, if one considers sphere or dome regions, we have from Theorem 1 that this operation only requires the evaluation of L inner products. In this case, the overall complexity of the proposed methodology thus scales as $\mathcal{O}(Lm)$.

If the choice of the regions $\{\mathcal{R}_l\}_{l=1}^L$ is changed "dynamically" (for example, one may want to adapt the choice of $\{\mathcal{R}_l\}_{l=1}^L$ to the safe region $\mathcal{S}$), the overall complexity of the algorithm must bear additional costs, namely the construction of the region $\mathcal{R}_l$ and the identification of the set of atoms belonging to it, that is $\mathcal{A} \cap \mathcal{R}_l$. In Section III-C, we propose a procedure to adapt the size of the regions $\{\mathcal{R}_l\}_{l=1}^L$ only leading to an increase of the computational cost of the order of $\mathcal{O}(L \log_2 n)$. In this case, the overall complexity of the proposed procedure thus scales as $\mathcal{O}(Lm + L \log_2 n)$.

These orders of complexity have to be compared to the computational cost of standard screening procedures, that is $\mathcal{O}(mn)$. We thus see that the screening procedure advocated in this section may lead to huge computational savings when the size n of the dictionary becomes large. However, the price to pay to achieve such a complexity reduction is a decrease of the effectiveness of the screening tests. In particular, as discussed in Section III-B below, the joint screening test for a region $\mathcal{R}$ can only be verified if the standard screening test (8) is passed for all the atoms $\mathbf{a}_i \in \mathcal{A} \cap \mathcal{R}$. The choice of the regions $\{\mathcal{R}_l\}_{l=1}^L$ will therefore play a major role in the ability of the proposed screening method to identify the zeros of $\mathbf{x}_\lambda^\star$. We discuss the latter issue in Sections III-C and III-D.

B. Relative effectiveness of the screening tests

It is easy to see from (18) that the choice of region $\mathcal{R}$ is a compromise between the number of atoms that can be jointly screened and the ease of passing the test. Indeed, although large regions allow to screen more atoms, they are also less likely to pass the joint screening test since, for any $\mathcal{R}_1 \subseteq \mathcal{R}_2$, we have

$$\max_{\mathbf{a} \in \mathcal{R}_1} \max_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{a}, \mathbf{z} \rangle \le \max_{\mathbf{a} \in \mathcal{R}_2} \max_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{a}, \mathbf{z} \rangle. \quad (25)$$

In particular, letting $\mathcal{R}_1 = \{\mathbf{a}_i\}$ and $\mathcal{R}_2 = \mathcal{R}$ in the above inequality, we see that passing the joint screening test (18) requires that the standard screening test (8) is verified for

---

[9] We assume that $\|\mathbf{t}\|_2 = 1$.

[10] In our discussion of the complexity, we only consider the operations which have to be carried out for each new proposal of a safe region $\mathcal{S}$. For example, if the regions $\{\mathcal{R}_l\}_{l=1}^L$ are prespecified, the set of atoms belonging to each region $\mathcal{R}_l$, that is $\mathcal{A} \cap \mathcal{R}_l$, may be precomputed. This task has therefore not to be repeated for each new screening test and is not taken into account in the "on-line" complexity required to implement the screening procedure.

any atom $\mathbf{a}_i \in \mathcal{A} \cap \mathcal{R}$. Hence, quite logically, joint screening test (18) can only lead to inferior screening performance as compared to standard screening test (8).

Another question of interest is the relative effectiveness of the joint sphere and dome tests proposed in (22) and (23)-(24), respectively. The next lemma provides some insights into this question.

Lemma 1. *The smallest[11] dome containing a set of unit-norm vectors $\mathcal{U}$ is always contained in the smallest sphere containing $\mathcal{U}$.*

A proof of this lemma can be found in Appendix C. In view of (25), a direct consequence of Lemma 1 is as follows: if one wishes to jointly screen a set of unit-norm atoms, there always exists a joint dome test leading to screening performance at least as good as the "best"[12] joint sphere test.

C. Choosing the "size" of $\mathcal{R}$

In this section, we discuss a particular strategy to adapt the size of the regions $\{\mathcal{R}_l\}_{l=1}^{L}$ in step 4 of Algorithm 1. More specifically, we assume that $\mathcal{R}_l$ corresponds to either a sphere or a dome region and we discuss the choice of the parameters ϵ and δ.

Let us first consider the case when $\mathcal{R}_l$ is a sphere, that is $\mathcal{R}_l = \mathcal{B}(\mathbf{t}_l, \epsilon_l)$. We assume that the test vector $\mathbf{t}_l$ is given and want to tune the value of ϵ_l so that $\mathcal{R}_l$ is the largest sphere passing (if possible) the joint screening test (22). Noticing that the joint sphere test (22) is satisfied as soon as the radius ϵ_l verifies

$$\epsilon_l < \epsilon_{\mathbf{t}_l,\mathbf{c}} \triangleq \frac{\tau - \langle \mathbf{t}_l, \mathbf{c} \rangle}{\|\mathbf{c}\|_2}, \tag{26}$$

an ideal choice for ϵ_l therefore consists in setting the latter as close as possible to $\epsilon_{\mathbf{t}_l,\mathbf{c}}$.

We note that computing the value of $\epsilon_{\mathbf{t}_l,\mathbf{c}}$ only requires the evaluation of one inner product $\langle \mathbf{t}_l, \mathbf{c} \rangle$ and does therefore not imply any computational increase with respect to the case where the regions are predefined once for all.[13] However, as mentioned previously, another issue arises when varying the size of $\mathcal{R}_l$: the efficient identification of the set of atoms $\mathcal{A} \cap \mathcal{R}_l$. We show hereafter that this operation can actually be carried out with a small computational overhead.

Letting the radius ϵ_l tend to its largest value $\epsilon_{\mathbf{t},\mathbf{c}}$, we have that any atom $\mathbf{a}_i \in \mathcal{A}$ having a distance to $\mathbf{t}_l$ strictly smaller than $\epsilon_{\mathbf{t}_l,\mathbf{c}}$ will be screened by test (22), that is

$$\|\mathbf{a}_i - \mathbf{t}_l\|_2 < \epsilon_{\mathbf{t}_l,\mathbf{c}} \Rightarrow \mathbf{x}_\lambda^\star(i) = 0. \tag{27}$$

Now, the quantities $\{\|\mathbf{a}_i - \mathbf{t}_l\|_2\}_{i=1}^{n}$ can be precomputed and sorted once for all in advance, so that the identification of the atoms verifying (27) can be done very efficiently (one can for example achieve a complexity scaling as $\mathcal{O}(\log_2 n)$ by using state-of-the-art sorting algorithms [26]). As a consequence, the

"on-line" complexity associated to the implementation of (27) is of the order of $\mathcal{O}(m + \log_2 n)$.[14]

We can apply the same kind of reasoning when $\mathcal{R}_l$ is a dome region, that is $\mathcal{R}_l = \mathcal{D}(\mathbf{t}_l, \delta_l)$. If we assume that $\mathbf{t}_l$ is given, a tight lower bound on the value of δ_l verifying the joint dome test is trivially given by the right-hand side of (24), that is

$$\delta_l > \delta_{\mathbf{t}_l,\mathbf{c}}, \tag{28}$$

where

$$\delta_{\mathbf{t}_l,\mathbf{c}} \triangleq \frac{\langle \mathbf{t}_l, \mathbf{c} \rangle \tau + \sqrt{\|\mathbf{c}\|_2^2 - \langle \mathbf{t}_l, \mathbf{c} \rangle^2} \sqrt{\|\mathbf{c}\|_2^2 - \tau^2}}{\|\mathbf{c}\|_2^2}. \tag{29}$$

Hence, provided that $\langle \mathbf{t}_l, \mathbf{c} \rangle < \tau$, letting parameter δ_l tend to its smallest value $\delta_{\mathbf{t}_l,\mathbf{c}}$ will lead to the screening of any atom $\mathbf{a}_i \in \mathcal{A}$ having an inner product with $\mathbf{t}_l$ strictly greater than $\delta_{\mathbf{t}_l,\mathbf{c}}$, that is:

$$\begin{cases} \langle \mathbf{t}_l, \mathbf{c} \rangle < \tau \\ \langle \mathbf{t}_l, \mathbf{a}_i \rangle > \delta_{\mathbf{t}_l,\mathbf{c}} \end{cases} \Rightarrow \mathbf{x}_\lambda^\star(i) = 0. \tag{30}$$

Again, the quantities $\{\langle \mathbf{t}_l, \mathbf{a}_i \rangle\}_{i=1}^{n}$ can be precomputed and sorted once for all in advance, so that a complexity scaling as $\mathcal{O}(m + \log_2 n)$ can also be achieved here.

Going back to the screening procedure advocated in Algorithm 1, we see that adapting the parameter ϵ_l (resp. δ_l) for each region $\mathcal{R}_l$ as discussed above is equivalent to applying test (27) (resp. (30)) for each of the L different test vectors $\mathbf{t}_l$ specifying the regions $\mathcal{R}_l$. The overall complexity of this procedure thus scales as $\mathcal{O}(Lm + L\log_2 n)$.

D. Choosing the test vectors

In the previous section, we assumed that the test vector $\mathbf{t}_l$ was given and we discussed the tuning of the parameters ϵ_l and δ_l. In this section, we elaborate on the choice of the test vectors $\{\mathbf{t}_l\}_{l=1}^{L}$. The devise of algorithms optimizing the set of test vectors being a broad topic of research, we leave this task for future works. Here, we keep our discussion to a more conceptual level and highlight some guidelines according to which the choice of the test vectors should/could be made.

In our discussion in Section III-B, we emphasized that the joint test of region $\mathcal{R}_l$ in (18) can only be passed if the standard test (8) is successful for any atom in $\mathcal{A} \cap \mathcal{R}_l$. Standard tests thus characterize the best screening performance achievable by the joint procedures advocated in this paper. In what follows, our discussion will be motivated by the latter observation. We introduce the next lemma to support our discussion.

Lemma 2. *Let $\mathcal{S} = \mathcal{B}(\mathbf{c}, 1 - \tau)$ be a safe sphere for problem (1)-(2). The standard sphere test (10) based on $\mathcal{S}$ fails for atom $\mathbf{a}_i \in \mathbb{S}^m$ if and only if*

$$\mathbf{a}_i \in \mathcal{D}\left(\frac{\mathbf{c}}{\|\mathbf{c}\|_2}, \frac{\tau}{\|\mathbf{c}\|_2} \right). \tag{31}$$

---

[11]"Smallest" should be understood as the region with the smallest volume.

[12]"Best" joint sphere test should be understood as the one involving the sphere of smallest volume.

[13]The inner product $\langle \mathbf{t}_l, \mathbf{c} \rangle$ has anyway to be computed for the evaluation of the tests (22)-(24).

[14]$\mathcal{O}(m)$ to evaluate the inner product $\langle \mathbf{t}_l, \mathbf{c} \rangle$ and $\mathcal{O}(\log_2 n)$ to identify the elements of $\{\|\mathbf{a}_i - \mathbf{t}_l\|_2\}_{i=1}^{n}$ passing the test (27).

Proof: By definition, the set of unit-norm atoms which do not pass test (10) is given by

$$\{\mathbf{a} : \langle \mathbf{a}, \mathbf{c} \rangle \geq \tau, \|\mathbf{a}\|_2 = 1\} = \left\{\mathbf{a} : \left\langle \mathbf{a}, \frac{\mathbf{c}}{\|\mathbf{c}\|_2} \right\rangle \geq \frac{\tau}{\|\mathbf{c}\|_2}\right\} \cap \mathbb{S}^m$$

$$= \mathcal{D}\left(\frac{\mathbf{c}}{\|\mathbf{c}\|_2}, \frac{\tau}{\|\mathbf{c}\|_2}\right) \cap \mathbb{S}^m.$$

Since $\mathbf{a}_i \in \mathbb{S}^m$, we obtain the result. Note that we were allowed to divide by $\|\mathbf{c}\|_2$ in the above expressions because our working hypotheses (see Section II-E) ensure that $\mathbf{c} \neq \mathbf{0}$. $\square$

Lemma 2 states that the set of unit-norm vectors which cannot be screened by a (standard) sphere test corresponds to a well-defined region, that is $\mathcal{D}\left(\frac{\mathbf{c}}{\|\mathbf{c}\|_2}, \frac{\tau}{\|\mathbf{c}\|_2}\right) \cap \mathbb{S}^m$. Now, since the joint test (18) can only be passed if the standard test (8) is verified for all the vectors belonging to $\mathcal{R}_l$, any non-empty intersection between $\mathcal{R}_l$ and $\mathcal{D}\left(\frac{\mathbf{c}}{\|\mathbf{c}\|_2}, \frac{\tau}{\|\mathbf{c}\|_2}\right) \cap \mathbb{S}^m$ will lead to a failure of the joint test. We build hereafter on this observation to provide two rules of thumb for the choice of the tests vectors $\{\mathbf{t}_l\}_{l=1}^L$.

Let us first consider the case where the sizes of the regions $\{\mathcal{R}_l\}_{l=1}^L$ are chosen once for all in advance. In this context, a first strategy to select the vectors $\{\mathbf{t}_l\}_{l=1}^L$ (and the size of the regions) could be based on the following (reasonable) requirements: *i)* each atom $\mathbf{a}_i \in \mathcal{A}$ must belong to at least one region $\mathcal{R}_l$; *ii)* each region must be as small as possible. The first requirement ensures that each atom may potentially be (jointly) screened by some test. The second one intends to "optimize" the probability of passing the joint test, since one may argue that small regions are less likely to have an intersection with the dome $\mathcal{D}\left(\frac{\mathbf{c}}{\|\mathbf{c}\|_2}, \frac{\tau}{\|\mathbf{c}\|_2}\right) \cap \mathbb{S}^m$. This suggest the set of test vectors $\{\mathbf{t}_l\}_{l=1}^L$ should be distributed evenly in $\mathrm{cone}(\mathcal{A}) \cap \mathbb{S}^m$.

A second option to select the set of test vectors is based on the following connection between the joint screening strategy proposed in (30) and the standard screening procedure (10): considering (30) with $\mathbf{t}_l = -\frac{\mathbf{c}}{\|\mathbf{c}\|_2}$ is exactly equivalent to applying the standard test (10) on *all* the atoms of the dictionary $\mathcal{A}$. This can be seen as follows. First, rewriting Lemma 2 in a slightly different way, we have that the set of unit-norm atoms passing test (10) can be written as

$$\left\{\mathbf{a} : \left\langle \mathbf{a}, \frac{\mathbf{c}}{\|\mathbf{c}\|_2} \right\rangle < \frac{\tau}{\|\mathbf{c}\|_2}\right\} \cap \mathbb{S}^m. \tag{32}$$

Moreover, particularizing (30) to the specific choice $\mathbf{t}_l = -\frac{\mathbf{c}}{\|\mathbf{c}\|_2}$, we obtain

$$\begin{cases} -\|\mathbf{c}\|_2 < \tau \\ \left\langle \mathbf{a}_i, -\frac{\mathbf{c}}{\|\mathbf{c}\|_2} \right\rangle > -\frac{\tau}{\|\mathbf{c}\|_2} \end{cases} \Rightarrow \mathbf{x}_\lambda^\star(i) = 0. \tag{33}$$

The first inequality in (33) is always satisfied by virtue of our working assumptions (see Section II-E); the second one is equivalent to

$$\mathbf{a}_i \in \left\{\mathbf{a} : \left\langle \mathbf{a}, \frac{\mathbf{c}}{\|\mathbf{c}\|_2} \right\rangle < \frac{\tau}{\|\mathbf{c}\|_2}\right\} \cap \mathbb{S}^m. \tag{34}$$

Hence, in view of (32), (33) is equivalent to applying the standard test (10) on the atoms of the dictionary $\mathcal{A}$.

This observation leads to our second rule of thumb to select $\{\mathbf{t}_l\}_{l=1}^L$: when test (30) is considered, the test vectors should be chosen "as close as possible" to $-\frac{\mathbf{c}}{\|\mathbf{c}\|_2}$. In particular, as emphasized above, if $\mathbf{t}_l = -\frac{\mathbf{c}}{\|\mathbf{c}\|_2}$ then one single joint test (30) leads to the screening of all the atoms "screenable" by the standard test (8). Of course, in practice, the actual value of the safe-sphere center $\mathbf{c}$ is not known in advance and depend on many parameters (*e.g.*, the LASSO problem at stake, the method used to construct the safe sphere, etc).

However, our above observation gives us good clues on how the test vectors should be chosen. When we have at our disposal a "training set" of sphere center $\{\mathbf{c}\}$, we could for example imagine to "learn" the test vectors from this dataset by imposing that elements of $\{\mathbf{t}_l\}_{l=1}^L$ are not "too far" from those in the training set.

IV. FINDING A SAFE REGION

In this section we elaborate on the problem of evaluating efficiently a dual feasible point. As pointed out in Section II-D, the standard "dual scaling" methodology entails a complexity scaling as $\mathcal{O}(mn)$. We propose hereafter two strategies based on "regions" to alleviate this order of complexity. Although presented separately, these two procedures can be combined in practice.

A. Relaxed dual scaling

A first strategy is based on the following observation: the standard dual scaling procedure can be relaxed by noticing that if some $\beta \in \mathbb{R}$ is such that

$$\beta \geq \max_{1 \leq i \leq n} \langle \mathbf{a}_i, \tilde{\mathbf{z}} \rangle, \tag{35}$$

then the point

$$\mathbf{z} = \beta^{-1} \tilde{\mathbf{z}} \tag{36}$$

is dual feasible.

We propose hereafter a strategy based on the use of regions to compute efficiently such an upper bound β. More specifically, let $\{\mathcal{R}_l\}_{l=1}^L$ be a set of (closed) regions of $\mathbb{R}^m$ such that[15]

$$\mathcal{A} \subseteq \cup_{l=1}^L \mathcal{R}_l. \tag{37}$$

Then, obviously, a valid choice for β is as follows:

$$\beta = \max_{1 \leq l \leq L} \max_{\mathbf{a} \in \mathcal{R}_l} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle. \tag{38}$$

Now, for the sphere and dome regions specified in (20)-(21), the inner maximum "$\max_{\mathbf{a} \in \mathcal{R}_l} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle$" admits a simple closed-form expression as shown in the following result:

---

[15]Note the regions $\{\mathcal{R}_l\}_{l=1}^L$ considered here may be different from those considered in the joint screening tests described in Section III.

Theorem 2. $\forall \tilde{\mathbf{z}} \in \mathbb{R}^m$ *we have*

$$\max_{\mathbf{a}\in\mathcal{B}(\mathbf{t},\epsilon)} \langle \mathbf{a},\tilde{\mathbf{z}}\rangle = \langle \mathbf{t},\tilde{\mathbf{z}}\rangle + \epsilon \|\tilde{\mathbf{z}}\|_2, \qquad (39)$$

$$\max_{\mathbf{a}\in\mathcal{D}(\mathbf{t},\delta)} \langle \mathbf{a},\tilde{\mathbf{z}}\rangle = \begin{cases} \|\tilde{\mathbf{z}}\|_2 & \text{if } \delta \le \frac{\langle\mathbf{t},\tilde{\mathbf{z}}\rangle}{\|\tilde{\mathbf{z}}\|_2} \\ f(\delta) & \text{otherwise} \end{cases}, \qquad (40)$$

where

$$f(\delta) \triangleq \delta \langle \mathbf{t},\tilde{\mathbf{z}}\rangle + \sqrt{1-\delta^2}\sqrt{\|\tilde{\mathbf{z}}\|_2^2 - \langle \mathbf{t},\tilde{\mathbf{z}}\rangle^2}. \qquad (41)$$

We refer the reader to Appendix B for a proof of this result. We see that the expressions (39)-(40) only depend on *one* inner product, namely $\langle \mathbf{t},\tilde{\mathbf{z}}\rangle$. The identification of a dual feasible point via (36)-(38) can therefore be done through the evaluation of L inner products. This first strategy to evaluate a dual feasible point has therefore a complexity scaling as $\mathcal{O}(Lm)$ when sphere and dome regions are considered.

The same remark as in Section III applies here: smaller regions usually lead to tightest bounds β but also typically involve larger complexities since more regions may be needed to cover the dictionary $\mathcal{A}$, see (37). As an extreme example, for the particular choice $\mathcal{R}_l = \{\mathbf{a}_l\}$, $l = 1 \dots n$, one recovers the standard (brute-force) dual scaling approach.

B. "Dual" screening

Our second strategy is inspired from the joint screening methodologies introduced in Section III. More specifically, our approach can be seen as a "screening method" for dual scaling: for a given $\tilde{\mathbf{z}} \in \mathbb{R}^m$, we propose the use of a joint screening procedure to identify (some of) the atoms $\mathbf{a}_i \in \mathcal{A}$ which *do not* attain the maximum value of $\langle \mathbf{a}_i,\tilde{\mathbf{z}}\rangle$.

Our procedure is based on the following simple observations. First, for any subset $\tilde{\mathcal{A}} \subseteq \mathcal{A}$ we obviously have

$$\max_{\mathbf{a}\in\tilde{\mathcal{A}}} \langle \mathbf{a},\tilde{\mathbf{z}}\rangle \le \max_{\mathbf{a}\in\mathcal{A}} \langle \mathbf{a},\tilde{\mathbf{z}}\rangle. \qquad (42)$$

Moreover, any region $\mathcal{R} \subset \mathbb{R}^m$ such that $\mathcal{R} \cap \mathcal{A} \ne \emptyset$ leads to

$$\max_{\mathbf{a}\in\mathcal{R}\cap\mathcal{A}} \langle \mathbf{a},\tilde{\mathbf{z}}\rangle \le \max_{\mathbf{a}\in\mathcal{R}} \langle \mathbf{a},\tilde{\mathbf{z}}\rangle. \qquad (43)$$

Therefore, combining (42) and (43), we have

$$\max_{\mathbf{a}\in\mathcal{R}} \langle \mathbf{a},\tilde{\mathbf{z}}\rangle < \max_{\mathbf{a}\in\tilde{\mathcal{A}}} \langle \mathbf{a},\tilde{\mathbf{z}}\rangle \Rightarrow \arg\max_{\mathbf{a}\in\mathcal{A}} \langle \mathbf{a},\tilde{\mathbf{z}}\rangle \notin \mathcal{R}\cap\mathcal{A}. \quad (44)$$

Interestingly, we note that the inequality in the left-hand side of (44) is equivalent to the joint screening test (19) introduced in Section III by making the following substitutions:

$$\begin{aligned} \mathbf{c} &\leftrightarrow \tilde{\mathbf{z}} \\ \tau &\leftrightarrow \max_{\mathbf{a}\in\tilde{\mathcal{A}}} \langle \mathbf{a},\tilde{\mathbf{z}}\rangle. \end{aligned} \qquad (45)$$

Nevertheless, passing test (44) allows us to draw a quite different conclusion from that obtained for conventional screening: if the test is verified, one is ensured that $\max_{\mathbf{a}\in\mathcal{A}} \langle \mathbf{a},\tilde{\mathbf{z}}\rangle$ is not attained in $\mathcal{R}\cap\mathcal{A}$. We will therefore refer to the methodologies based on (44) as "dual screening".

Because of the similarity between (19) and (44), all the results presented in Section III may be reused to implement efficiently dual-scaling screening with the substitution (45). In particular, if one chooses $\mathcal{R}$ to be a sphere or a dome, one may exploit the expressions stated in Theorem 1 to

---

Algorithm 2 Dual Screening Test

1: **inputs:** L : number of regions, $\mathcal{A}$: set of atoms, $\tilde{\mathcal{A}}$: subset of atoms, $\tilde{\mathbf{z}}$: dual variable
2: **procedure** DUAL SCREENING(L, $\mathcal{A}$, $\tilde{\mathcal{A}}$)
3:     **init:** $\mathcal{A}_{\text{scr}} = \emptyset$
4:     Evaluate $\max_{\mathbf{a}\in\tilde{\mathcal{A}}} \langle \mathbf{a},\tilde{\mathbf{z}}\rangle$
5:     Select a set of regions $\{\mathcal{R}_l\}_{l=1}^L$
6:     **for all** $1 \le l \le L$ **do**
7:         **if** $\max_{\mathbf{a}\in\mathcal{R}_l}\langle\mathbf{a},\tilde{\mathbf{z}}\rangle < \max_{\mathbf{a}\in\tilde{\mathcal{A}}}\langle\mathbf{a},\tilde{\mathbf{z}}\rangle$ **then**
8:             $\mathcal{A}_{\text{scr}} = \mathcal{A}_{\text{scr}} \cup (\mathcal{A} \cap \mathcal{R}_l)$
9:         **end if**
10:     **end for**
11: **end procedure**
12: **outputs:** $\mathcal{A}_{\text{scr}}$: set of screened atoms

---

evaluate "$\max_{\mathbf{a}\in\mathcal{R}}\langle\mathbf{a},\tilde{\mathbf{z}}\rangle$" by computing one single inner product. Evaluating "$\max_{\mathbf{a}\in\mathcal{R}}\langle\mathbf{a},\tilde{\mathbf{z}}\rangle$" for L different (sphere or dome) regions thus leads to a complexity scaling as $\mathcal{O}(Lm)$. On the other hand, computing the other side of the inequality in (44), that is "$\max_{\mathbf{a}\in\tilde{\mathcal{A}}}\langle\mathbf{a},\tilde{\mathbf{z}}\rangle$", requires the evaluation of $\text{card}(\tilde{\mathcal{A}})$ inner products, resulting in a complexity scaling as $\mathcal{O}(\text{card}(\tilde{\mathcal{A}})m)$. Finally, if the size of the regions are dynamically adapted as suggested in Section III-C, an additional cost of $\mathcal{O}(L\log n)$ has to be added to the overall complexity. "Dual screening" can thus be implemented with a complexity scaling at most as $\mathcal{O}(Lm + \text{card}(\tilde{\mathcal{A}})m + L\log n)$. If the test vectors $\{\mathbf{t}_l\}_{l=1}^L$ defining the L regions are included in $\tilde{\mathcal{A}}$, the complexity can even be reduced to $\mathcal{O}(\text{card}(\tilde{\mathcal{A}})m + L\log n)$ since the same inner products do not have to be evaluated twice. A summary of the operations performed by dual screening is provided in Algorithm 2.

V. NUMERICAL EXPERIMENTS

In this section we evaluate the effectiveness of the proposed methodologies via numerical simulations. We consider a noiseless Gaussian deconvolution problem where the atoms of the dictionaries are defined as:

$$\mathbf{a}_i(j) \propto \exp\left(-\frac{1}{2\sigma^2}\left(\frac{j-1}{m-1} - \frac{i-1}{n-1} \right)^2 \right), \quad j = 1 \dots m.$$

In our simulations, we set $\sigma = 10^{-1}$, $m = 50$ and $n = 1024$.

The evaluation of the proposed procedures is organized as follows. In section V-A, we study the tightness of the upper bounds proposed in Section IV-A for relaxed dual scaling, and illustrate its impact on the radius of the ST and GAP safe spheres. In section V-B, we assess the performance of the dual screening methodology advocated in Section IV-B. In the next subsection, we illustrate the screening ability of the joint procedures described in Section III. Finally, in section V-D, we study the computational savings which can be obtained in the LASSO implementation by combining the different methodologies proposed in this paper.
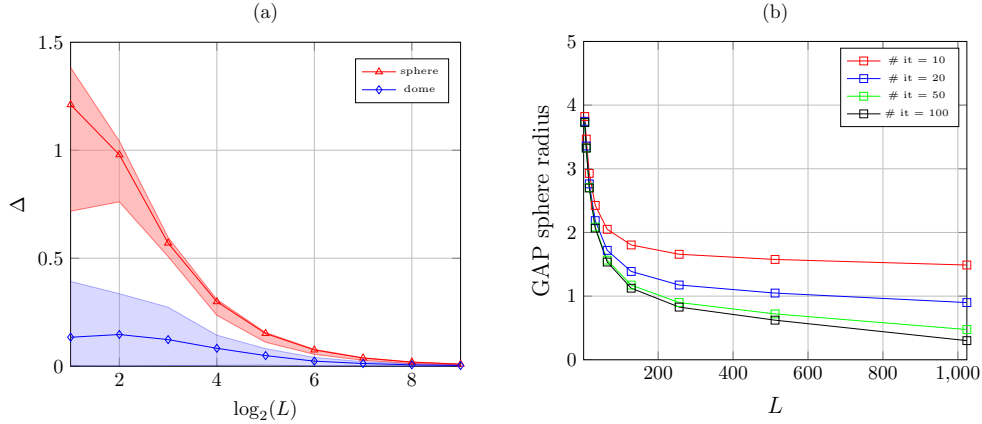
Figure 1. (a) Evaluation of the tightness of the upper bound β defined (38): sphere (red) and dome (blue) regions are considered. The plain curves represent the averaged value of $\Delta = \beta - \max_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle$ over 200 realizations of $\tilde{\mathbf{z}}$; the shaded areas represent the set of realizations of Δ. (b) Evolution of the radius of the GAP safe sphere as a function of the number L of regions. The different curves represent the radius of the sphere constructed from the output of FISTA at a given iteration.

A. Relaxed dual scaling

We first discuss the effectiveness of the relaxed dual scaling presented in Section IV-A. We remind the reader that "relaxed dual scaling" is based on the evaluation of an upper bound β verifying (35); this upper bound is then used to find a dual feasible point as in (36) which, in turn, may serve in the construction of a safe region (see for example (11) and (12) in Section II-C).

In Fig. 1a, we illustrate the tightness of the upper bound presented in Theorem 2. We consider the following metric:

$$\Delta = \beta - \max_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle, \qquad (46)$$

where β is defined in (38). We generate $\tilde{\mathbf{z}}$ as the random linear combination of 5 atoms of $\mathcal{A}$. The results are averaged over 200 realizations. The plain curves correspond to the averaged value of Δ for sphere (red) and dome (blue) regions; the shaded areas represent the set of realizations of Δ. The horizontal axis corresponds to the number of regions L considered in the evaluation of β. The center of the regions $\{\mathcal{R}_l\}_{l=1}^{L}$ are chosen uniformaly over $[0, 1]$:

$$\mathbf{t}_l = \mathbf{a}_{l \frac{n}{L}} \quad \text{for } l = 1 \ldots L. \qquad (47)$$

The size of the regions are set so that $\mathcal{A} \subseteq \cup_{l=1}^{L} \mathcal{R}_l$. We see in Fig. 1a that the precision of the upper bound improves as the number of regions increase. We also note that the dome region leads to more accurate results than spherical regions.

In Fig. 1b, we illustrate the impact of the proposed relaxation on the radius of the GAP sphere (12). We generated the results as follows. We build vectors $\mathbf{y}$ as random combinations of 10 atoms of $\mathcal{A}$. We run a FISTA algorithm [3] to solve the LASSO problem (1) with $\mathbf{y}$ as input and $\lambda = 10^{-1} \max_{\mathbf{a} \in \mathcal{A}} |\langle \mathbf{a}, \mathbf{y} \rangle|$. We finally construct a GAP safe sphere at each iteration of FISTA as follows: i) we set $\tilde{\mathbf{z}} = \lambda^{-1} (\mathbf{y} - \mathbf{A} \hat{\mathbf{x}})$ where $\hat{\mathbf{x}}$ is the current iterate of FISTA; ii) we identify a dual feasible point via (36) and use (12) to compute the radius of the GAP sphere. Fig. 1b displays the radius obtained as a function of the number of regions used in the computation of the upper bound β. The results are averaged

over 200 realizations of $\mathbf{y}$. We see that, at a given iteration of FISTA, the radius of the sphere quite logically decreases with the number of regions used in the computation of β.

B. Dual screening

We now illustrate the performance of the dual screening methodology presented in Section IV-B. We apply Algorithm 2 to random realizations of the dual variable: $\tilde{\mathbf{z}}$ is generated as a random linear combination of 5 atoms of $\mathcal{A}$. Both the elements of $\tilde{\mathcal{A}}$ and the test vectors $\{\mathbf{t}_l\}_{l=1}^{L}$ defining the regions used in Algorithm 2 are chosen as a regular subsampling of L atoms of $\mathcal{A}$, i.e.,

$$\tilde{\mathcal{A}} = \{\mathbf{t}_l\}_{l=1}^{L} = \{\mathbf{a}_{l \frac{n}{L}} : l = 1 \ldots L\}. \qquad (48)$$

The following figures of merit are represented in Fig. 2:
- Fig. 2a: the percentage of atoms not attaining $\max_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle$ identified by dual screening,
- Fig. 2b: the computational gain induced by dual screening.

The results are averaged over 50 realizations of $\tilde{\mathbf{z}}$. The performance of dual screening is shown for both dome and sphere regions.

Regarding Fig. 2b, we consider the following procedure to generate the figure of merit: i) we apply Algorithm 2 to identify a set of atoms, say $\mathcal{A}_{\text{scr}}$, which do not attain the maximum value of $\langle \mathbf{a}, \tilde{\mathbf{z}} \rangle$; ii) we evaluate $\max_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle$ as

$$\max_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle = \max_{\mathbf{a} \in \mathcal{A} \backslash \mathcal{A}_{\text{scr}}} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle. \qquad (49)$$

We note that the right-hand side of (49) only requires the evaluation of $\text{card}\{\mathcal{A} \backslash \mathcal{A}_{\text{scr}}\}$ inner products. If L regions are considered in Algorithm 2, the above procedure then only requires the computation of $L + \text{card}\{\mathcal{A} \backslash \mathcal{A}_{\text{scr}}\}$ inner products to evaluate $\max_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle$. This complexity has to be compared to the n inner products needed for a brute-force evaluation of $\max_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle$. The figure of merit represented in Fig. 2b illustrates this ratio and is defined as

$$\text{computational gain} = \frac{L + \text{card}\{\mathcal{A} \backslash \mathcal{A}_{\text{scr}}\}}{n}. \qquad (50)$$
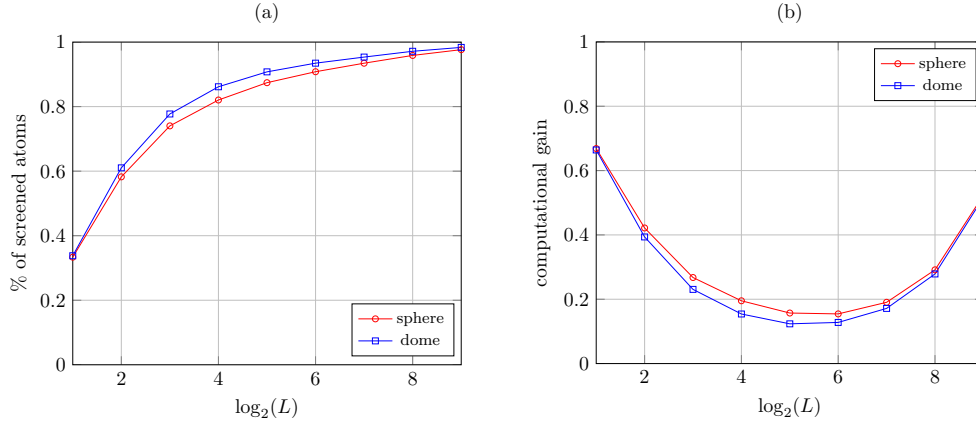
Figure 2. Performance of the dual screening procedure : (a) percentage of atoms screened by the proposed procedure; (b) computational gain allowed by dual scaling. The results are shown for sphere (circle mark) and dome (square mark) regions as a function of (the logarithm of) the number of regions L.

In Fig. 2a, we see that the percentage of atoms identified by dual screening logically increases as the number of regions increases. More than 90% of the atoms not attaining the maximum value of $\langle \mathbf{a}, \tilde{\mathbf{z}} \rangle$ can be identified as soon as $\log_2(L) \geq 4$. We also notice that the dome regions lead to sightly more favorable results than the sphere regions.

In Fig. 2b, we note that the ability of dual screening to reduce the number of atoms allows for important savings in terms of computational complexity. In particular, considering the case where $\log_2(L) \geq 4$, we see that the two-step procedure discussed above is able to evaluate $\max_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle$ with a complexity roughly ten times smaller than the one required by a brute-force approach.

C. Joint screening for LASSO

In this section, we evaluate the performance of the joint screening procedures for LASSO presented in Section III.

We first notice that the ability of a screening method to identify zeros of the primal vector $\mathbf{x}_\lambda^\star$ depends on the "quality" of the safe region $\mathcal{S}$ considered in the test. In our simulations hereafter, we will characterize the latter quality via the following parameter: $\tilde{\tau} \triangleq \frac{\tau}{\|\mathbf{c}\|_2}$. From our working assumptions in Section II-E, we have that $\tilde{\tau} \in (-1, 1]$. The joint screening test (19) can be expressed as a function of $\tilde{\tau}$ as

$$\max_{\mathbf{a} \in \mathcal{R}} \left\langle \mathbf{a}, \frac{\mathbf{c}}{\|\mathbf{c}\|_2} \right\rangle < \tilde{\tau}. \qquad (51)$$

The qualitative interpretation of $\tilde{\tau}$ is as follows: if $\tilde{\tau} = 1$, *all* the atoms of the dictionary (except $\mathbf{a}_i = \mathbf{c}$, if it exists) will pass the standard screening test (10) (or equivalently (51) with $\mathcal{R} = \{\mathbf{a}_i\}$); on the other hand, if $\tilde{\tau} = -1$ *no* atoms can pass the standard test.[16] Values of $\tilde{\tau}$ between 1 and -1 correspond to intermediate scenarios between these two extreme cases: qualitatively speaking, the quality of the safe region degrades for $\tilde{\tau}$ going from 1 to -1.

In Fig. 3, we illustrate the screening performance achieved by the dome joint screening test presented in Section III

[16]The two latter assertions can easily be seen from (51) and the Cauchy-Schwarz inequality.

as a function of $\tilde{\tau}$ (horizontal axis) and the (logarithm of the) number of regions L. The results for the sphere region are sensibly similar and therefore not reported for the sake of conciseness. We also focus on the range $\tilde{\tau} \in [-0.5, 1]$ which is the one of most interest for the representation of the method's performance. The automatic choice of the region size as discussed in Section III-C is implemented in the screening procedure. The test vectors defining the L regions are chosen uniformly on $[1, n]$, *i.e.*,

$$\mathbf{t}_l = \mathbf{a}_{l\frac{n}{L}} \quad \text{for } l = 1 \ldots L. \qquad (52)$$

The center of the safe sphere $\mathbf{c}$ is chosen as a random linear combination of $k = 5$ elements of $\mathcal{A}$. All the results are averaged over 50 experiments. The following figures of merit are represented:

- Fig. 3a: the percentage of zeros of $\mathbf{x}_\lambda^\star$ identified by the screening procedure,
- Fig. 3b: the ratio between the number of zeros identified by the joint screening and the standard screening procedures,
- Fig. 3c: the decrease of computational complexity obtained by using joint screening.

More specifically, we considered the following screening procedure to generate Fig. 3c: *i)* we applied joint screening on L regions (Algorithm 1) and discarded all the atoms, say $\mathcal{A}_{\text{scr}}$, belonging to the screened regions; *ii)* we applied standard screening on the remaining atoms $\mathcal{A} \backslash \mathcal{A}_{\text{scr}}$. The screening performance attained by this two-step procedure is therefore identical to that obtained by a brute-force application of standard screening (10) to the overall dictionary $\mathcal{A}$. The computational complexity of these two approaches may nevertheless be quite different since joint screening allows for the removal of sets of atoms via one single test. Fig. 3c thus illustrates the ratio between the number of inner products required by the above two-step procedure and a brute-force implementation of standard screening.

We see in Fig. 3a that, quite logically, the percentage of atoms passing the joint screening test degrades as the value of $\tilde{\tau}$ decreases. We also note a degradation of the performance when the number of regions considered in the joint screening
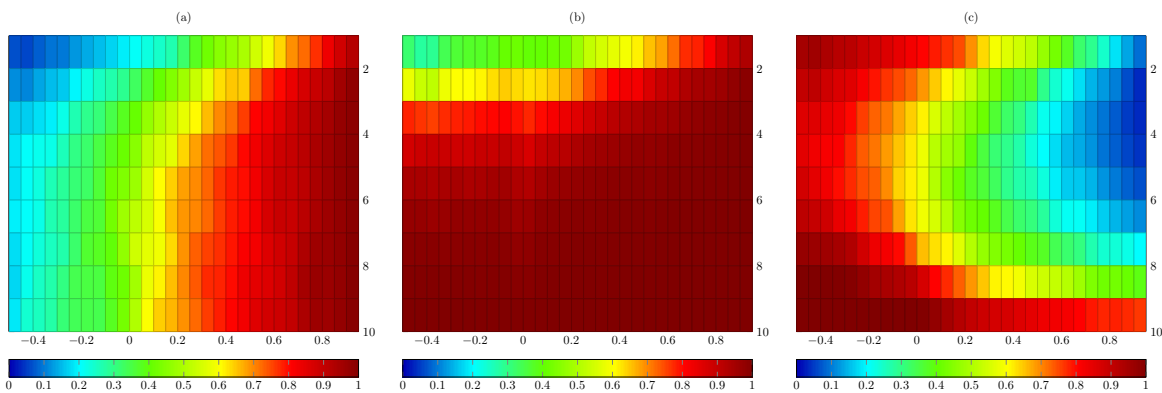
Figure 3. (a) the percentage of zeros of $\mathbf{x}_\lambda^\star$ identified by the screening procedure (b) ratio between the number of zeros identified by the joint screening and the standard screening procedures (c) decrease of the computational complexity obtained by the use of joint screening. The horizontal axis corresponds to the value of $\tilde{\tau}$, the vertical axis represents $\log_2(L)$.

test decreases. The results have to be compared to the last row of Fig. 3a which corresponds to the implementation of the standard screening test ($\log_2(L) = 10$ *i.e.*, the number of regions L is equal to the number of atoms n). We notice that, although the percentage of the screened atoms increases with $\log_2(L)$, the performance of joint screening remains quite similar to that of standard screening when the number of regions is "sufficiently large".

This behavior is further illustrated in Fig. 3b, representing the ratio between the number of zeros identified by the joint screening and the standard screening procedures. We see that the standard and joint screenings identify roughly the same number of zeros on quite a large range of number of regions; the results seem to be all the more favorable as the quality of the safe region (*i.e.*, the value of $\tilde{\tau}$) is high.

The good behavior of joint screening has an impact on the computational complexity required to achieved some given screening performance. This is illustrated in Fig. 3c. This figure represents the ratio between the number of inner products required by the two-step procedure described above and a brute-force implementation of standard screening. We see that, as far as our simulation setup is concerned, it is almost always beneficial to consider the approach based on joint screening. The best improvement seems to occur for $\log_2(L) \in \{4, 5\}$ for which a decrease of the computational complexity by more than 50% is observed for $\tilde{\tau} \geq 0$. When $\tilde{\tau}$ is close to 1, joint screening seems to even lead to computational savings of roughly one order of magnitude.

D. Complexity savings

In this section, we illustrate the overall computational savings induced by the proposed joint screening method. We consider the problem of finding the solution of (1) for the Gaussian deconvolution problem described at the beginning of this section.

The setup of our simulation is as follows. We use a FISTA algorithm [3] and consider a "continuation method" to compute $\mathbf{x}_\lambda^\star$ for some $\lambda = \lambda_{\text{target}}$. More specifically, we solve the problem (1) for a decreasing sequence of λ (down to $\lambda = \lambda_{\text{target}}$) and initialize FISTA with a "warm start" for

each new value of λ. We stop the iterations of FISTA when a dual gap equal to 10^{-4} is attained.

We consider two competitive implementations of this general procedure by applying either the GAP standard screening test or the joint screening procedure described in Sections III and IV. Screening is applied at each iteration of FISTA.

Regarding the joint screening approach, we use the two-step procedure described in Fig. 3c to attain the same screening performance as the standard GAP test with a reduced complexity. Similarly, the two-step procedure used to generate Fig. 2b is implemented to identify a dual feasible point with low computational burden.

The complexity of these two procedures are compared by counting the number of operations needed to compute $\mathbf{x}_{\lambda_{\text{target}}}^\star$. Since multiplications entail a much higher computational burden than additions, we restrict our attention to the former. In Fig. 4, we represent the computational savings induced by joint screening via the following figure of merit:

$$\text{computational savings} = 1 - \frac{N_{\text{joint}}}{N_{\text{std}}}, \qquad (53)$$

where N_{std} (resp. N_{joint}) is the total number of multiplications carried out by the optimization procedure with standard (resp. joint) screening. As far as our simulation setup is concerned, we see that the proposed methodology allows for computational savings irrespective of the number of regions L and the target regularization parameter λ_{target}. The gain induced by our method depends on these parameters: when $L = 64$, we note that up to 45% of computational savings is possible for wide range of λ_{target}. We also note that dome regions usually lead to slightly enhanced performance as compared to sphere regions.

VI. CONCLUSION

In this paper, we proposed several methodologies to improve the tradeoff "complexity/effectiveness" of safe screening tests for LASSO. Our procedures are based on the following general philosophy: we gather subsets of atoms in some regions with "desirable" geometries (*e.g.*, spheres or domes) and exploit the latter to deduce closed-form expressions of the optimization problems encountered in the construction of the safe screening
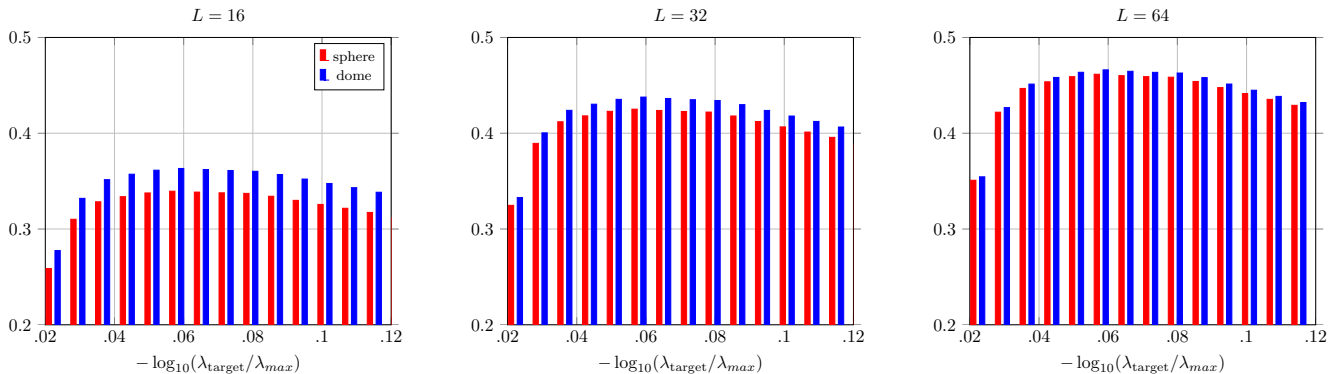
Figure 4. Computational savings (see (53)) induced by joint screening for different number of regions L and values of the regularization parameter.

test. A first procedure relying on this principle is the so-called *"joint screening test"* (Section III) which enables the rejection of *all* the atoms belonging to some region by performing *one* single test. Other procedures grounded on this philosophy are the *"relaxed dual scaling"* and *"dual screening"* procedures proposed in Section IV, which offer some low-complexity alternatives to compute a dual feasible point of the LASSO problem. We showed through numerical simulations that the proposed methodologies lead to significant computational gain in the implementation of safe screening tests.

APPENDIX A
MISCELLANEOUS TECHNICAL LEMMAS

In this appendix we prove two technical lemmas which will be useful in the derivation of our main results. The first lemma (Lemma 3) characterizes the properties of some functions appearing in the dome joint screening procedure. The second lemma (Lemma 4) establishes a relationship between the (norm of the) center $\mathbf{c}$ and the radius τ of any safe sphere.

Lemma 3. *Let*

$$f(\xi) = A\xi + \sqrt{1 - A^2}\sqrt{1 - \xi^2}. \tag{54}$$

The following assertions hold:

(a) if $A \in [-1, 1]$, $f(\xi)$ is concave over $[-1, 1]$
(b) if $A \in (-1, 1)$, $f(\xi)$ is strictly concave over $[-1, 1]$
(c) if $A \in [-1, 1]$, the function $g(\xi) \triangleq \max_{\xi \le \xi' \le 1} f(\xi')$ can be written as

$$g(\xi) = \begin{cases} f(A) & \text{if } \xi < A \\ f(\xi) & \text{otherwise} \end{cases} \tag{55}$$

in the interval $\xi \in [-1, 1]$.
(d) if $A \in [-1, 1]$, $g(\xi)$ is concave and non-increasing over the interval $[-1, 1]$
(e) if $A \in (-1, 1)$, $g(\xi)$ is strictly concave and strictly decreasing over the interval $[A, 1]$.

Proof: ((*a*) and (*b*)) If $A = \pm 1$, $f(\xi)$ is linear and therefore concave. If $A \in (-1, 1)$, $f(\xi)$ is the sum of two terms. The first one is linear and therefore concave. The second corresponds to the equation of a "half-circle" and is thus strictly concave. Hence, $f(\xi)$ is strictly concave for $A \in (-1, 1)$.

(c) If $|A| < 1$, we note that the maximum value of $f(\xi')$ over $[-1, 1]$ is unique and obtained for $\xi' = A$, by virtue of the concavity of $f(\xi')$. Then, for $|A| < 1$ and $\xi \le A$,

$$\max_{\xi \le \xi' \le 1} f(\xi') = f(A) = 1,$$

since $A \in [\xi, 1]$, and for $\xi > A$,

$$\max_{\xi \le \xi' \le 1} f(\xi') = f(\xi),$$

due to the strictly decrreasing behavior of $f(\xi')$ over $[\xi, 1]$. The case $|A| = 1$ leads to trivial solutions. If $A = 1$, we have

$$\max_{\xi \le \xi' \le 1} f(\xi') = \max_{\xi \le \xi' \le 1} \xi' = 1 = f(A) \quad \forall \xi \in [-1, 1].$$

If $A = -1$, we have

$$\max_{\xi \le \xi' \le 1} f(\xi') = \max_{\xi \le \xi' \le 1} -\xi' = -\xi = f(\xi) \quad \forall \xi \in [-1, 1].$$

Eq. (55) merges all the different cases in a synthetic way.

(*d* and *e*) The concavity and the non-increasing (resp. the strict concavity and the strictly decreasing) nature of $g(\xi)$ over $[-1, 1]$ (resp. $[A, 1]$) for $A \in [-1, 1]$ (resp. $A \in (-1, 1)$) follows from the definition of $g(\xi)$ and the concavity (resp. strict concavity) of $f(\xi)$. □

The next lemma relates the center and the radius of any safe sphere test. It is used in the proof of Theorem 1 in Section B.

Lemma 4. *Let $\mathcal{S} = \mathcal{B}(\mathbf{c}, 1 - \tau)$ be a safe sphere for problem (1)-(2) with penalization parameter $\lambda < \lambda_{\max}$ (that is $\mathbf{z}_\lambda^\star \in \mathcal{S}$). Then, we have*

$$\tau \le \|\mathbf{c}\|_2. \tag{56}$$

Proof: If $\mathcal{S} = \{\mathbf{z} : \|\mathbf{z} - \mathbf{c}\|_2 \le 1 - \tau\}$ is a safe region, then

$$\|\mathbf{z}_\lambda^\star - \mathbf{c}\|_2 \le 1 - \tau, \tag{57}$$

which leads, by using a triangle inequality, to

$$\|\mathbf{z}_\lambda^\star\|_2 - 1 + \tau \le \|\mathbf{c}\|_2. \tag{58}$$

Now, if $\lambda < \lambda_{\max}$, we have

$$\|\mathbf{z}_\lambda^\star\|_2 \ge 1. \tag{59}$$

The latter claim follows from the following arguments. If $\lambda < \lambda_{\max}$, we necessarily have $\mathbf{x}_\lambda^\star(i) > 0$ for some

$i \in [1 \dots n]$.[17] From the optimality condition (7), we have for such i: $\langle \mathbf{a}_i, \mathbf{z}_\lambda^\star \rangle = 1$. Hence, we obtain (59) by using the Cauchy-Schwarz inequality:

$$\|\mathbf{z}_\lambda^\star\|_2 \geq \langle \mathbf{a}_i, \mathbf{z}_\lambda^\star \rangle = 1.$$

Finally, we obtain the main result (56) by combining (58) and (59). $\qquad\square$

APPENDIX B
PROOFS OF THEOREMS 1 AND 2

In this appendix we provide a proof to the results stated in Theorems 1 and 2. The proof is divided in two parts: in Section B-A we address the case of a spherical region whereas in Section B-B we focus on a dome region.

A. Sphere region

Let us first notice that the sphere region $\mathcal{B}(\mathbf{t}, \epsilon)$ can be written as

$$\mathcal{B}(\mathbf{t}, \epsilon) = \{ \mathbf{a} = \mathbf{t} + \mathbf{z} : \|\mathbf{z}\|_2 \leq \epsilon \}. \tag{62}$$

Therefore, $\forall \tilde{\mathbf{z}} \in \mathbb{R}^m$ we have

$$\max_{\mathbf{a} \in \mathcal{B}(\mathbf{t}, \epsilon)} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle = \langle \mathbf{t}, \tilde{\mathbf{z}} \rangle + \max_{\|\mathbf{z}\|_2 \leq \epsilon} \langle \mathbf{z}, \tilde{\mathbf{z}} \rangle$$
$$= \langle \mathbf{t}, \tilde{\mathbf{z}} \rangle + \epsilon \|\tilde{\mathbf{z}}\|_2, \tag{63}$$

where the last equality is a consequence of the tightness of the Cauchy-Schwarz inequality. The result stated in Theorem 2 for the sphere region $\mathcal{B}(\mathbf{t}, \epsilon)$ corresponds to (63). Theorem 1 easily follows from (63) by choosing $\tilde{\mathbf{z}} = \mathbf{c}$.

B. Dome region

We assume $\|\mathbf{t}\|_2 = 1$. We first note that the dome region $\mathcal{D}(\mathbf{t}, \delta)$ can be written as

$$\mathcal{D}(\mathbf{t}, \delta) = \{ \mathbf{a} = \alpha \mathbf{t} + \mathbf{z} : \mathbf{z} \in \mathcal{Z}_\alpha, \delta \leq \alpha \leq 1 \}, \tag{64}$$

where

$$\mathcal{Z}_\alpha = \left\{ \mathbf{z} : \mathbf{z} \in (\text{span}[\mathbf{t}])^\perp, \|\mathbf{z}\|_2 \leq \sqrt{1 - \alpha^2} \right\}. \tag{65}$$

Therefore, $\forall \tilde{\mathbf{z}} \in \mathbb{R}^m$ we have

$$\max_{\mathbf{a} \in \mathcal{D}(\mathbf{t}, \delta)} \langle \mathbf{a}, \tilde{\mathbf{z}} \rangle = \max_{\delta \leq \alpha \leq 1} \left(\alpha \langle \mathbf{t}, \tilde{\mathbf{z}} \rangle + \max_{\mathbf{z} \in \mathcal{Z}_\alpha} (\langle \mathbf{z}, \tilde{\mathbf{z}} \rangle) \right)$$
$$= \max_{\delta \leq \alpha \leq 1} \left(\alpha \langle \mathbf{t}, \tilde{\mathbf{z}} \rangle + \max_{\mathbf{z} \in \mathcal{Z}_\alpha} (\langle \mathbf{z}, P_\mathbf{t}^\perp(\tilde{\mathbf{z}}) \rangle) \right)$$
$$= \underbrace{\max_{\delta \leq \alpha \leq 1} \left(\alpha \langle \mathbf{t}, \tilde{\mathbf{z}} \rangle + \sqrt{1 - \alpha^2} \left\| P_\mathbf{t}^\perp(\tilde{\mathbf{z}}) \right\|_2 \right)}_{\triangleq g(\delta)},$$

[17]This statement can be proved as follows. If $\mathbf{x}_\lambda^\star = \mathbf{0}$, then we have $\mathbf{z}_\lambda^\star = \mathbf{y}/\lambda$ from (7). Now, we must have, by dual feasibility:

$$\max_i \langle \mathbf{a}_i, \mathbf{z}_\lambda^\star \rangle \leq 1. \tag{60}$$

Plugging $\mathbf{z}_\lambda^\star = \mathbf{y}/\lambda$ into this inequality, we obtain

$$\lambda_{\max} \triangleq \max_i \langle \mathbf{a}_i, \mathbf{y} \rangle \leq \lambda. \tag{61}$$

Therefore, by contraposition, if $\lambda < \lambda_{\max}$ then $\mathbf{x}_\lambda^\star \neq \mathbf{0}$.

where $P_\mathbf{t}^\perp(\tilde{\mathbf{z}}) = \tilde{\mathbf{z}} - \langle \mathbf{t}, \tilde{\mathbf{z}} \rangle \mathbf{t}$ denotes the orthogonal projector onto $(\text{span}[\mathbf{t}])^\perp$. The last equality is a consequence of the tightness of the Cauchy-Schwarz inequality.

Using Lemma 3(c) in Appendix A with $A = \frac{\langle \mathbf{t}, \tilde{\mathbf{z}} \rangle}{\|\tilde{\mathbf{z}}\|_2}$, we obtain that

$$g(\delta) = \begin{cases} \|\tilde{\mathbf{z}}\|_2 & \text{if } \delta < \frac{\langle \mathbf{t}, \tilde{\mathbf{z}} \rangle}{\|\tilde{\mathbf{z}}\|_2} \\ \delta \langle \mathbf{t}, \tilde{\mathbf{z}} \rangle + \sqrt{1 - \delta^2} \left\| P_\mathbf{t}^\perp(\tilde{\mathbf{z}}) \right\|_2 & \text{otherwise} \end{cases}. \tag{66}$$

This corresponds to the result stated in Theorem 2.

Setting $\tilde{\mathbf{z}} = \mathbf{c}$ in the definition of $g(\delta)$, the joint dome test "$\max_{\mathbf{a} \in \mathcal{D}(\mathbf{t}, \delta)} \langle \mathbf{a}, \mathbf{c} \rangle < \tau$" can thus be simply rewritten as

$$\tau > g(\delta). \tag{67}$$

The expressions (23)-(24) stated in Theorem 1 can then be obtained as follows. First, satisfying (67) necessarily requires that

$$\tau > \min_{\tilde{\delta} \in [-1, 1]} g(\tilde{\delta}) = g(1) = \langle \mathbf{t}, \mathbf{c} \rangle, \tag{68}$$

where the first equality follows from the non-increasing nature of g (Lemma 3(d)) and the second from its definition (66). Inequality (68) corresponds to the condition enforced by (23).

Moreover, we have from Lemma 4 in Appendix A that[18]

$$\tau \leq \|\mathbf{c}\|_2 \tag{69}$$

provided that τ is associated to the radius of a safe sphere. Therefore, if (23) holds, owing to the continuity of g and the fact that it is strictly decreasing over[19] $[\frac{\langle \mathbf{t}, \mathbf{c} \rangle}{\|\mathbf{c}\|_2}, 1]$ (see Lemma 3(e)), there exists $\delta_{\mathbf{t}, \mathbf{c}} \in [\frac{\langle \mathbf{t}, \mathbf{c} \rangle}{\|\mathbf{c}\|_2}, 1]$ such that $g(\delta_{\mathbf{t}, \mathbf{c}}) = \tau$. Using the expression of g over $[\frac{\langle \mathbf{t}, \mathbf{c} \rangle}{\|\mathbf{c}\|_2}, 1]$ in (66), we find

$$\delta_{\mathbf{t}, \mathbf{c}} = \frac{\langle \mathbf{t}, \mathbf{c} \rangle \tau + \sqrt{\|\mathbf{c}\|_2^2 - \langle \mathbf{t}, \mathbf{c} \rangle^2} \sqrt{\|\mathbf{c}\|_2^2 - \tau^2}}{\|\mathbf{c}\|_2^2}. \tag{70}$$

Invoking again the strict decrease of g over $[\frac{\langle \mathbf{t}, \mathbf{c} \rangle}{\|\mathbf{c}\|_2}, 1]$ (see Lemma 3(e)), we have that $\tau = g(\delta_{\mathbf{t}, \mathbf{c}}) > g(\delta)$ if and only if

$$\delta > \delta_{\mathbf{t}, \mathbf{c}}.$$

Combining this condition with the expression of $\delta_{\mathbf{t}, \mathbf{c}}$ in (70), we obtain (24).

APPENDIX C
PROOF OF LEMMA 1

First note that the volume of a dome $\mathcal{D}(\mathbf{t}, \delta)$ is a decreasing function of δ, irrespective of the test vector $\mathbf{t}$. Hence, the smallest dome including all the elements of $\mathcal{U}$ is given by $\mathcal{D}(\mathbf{t}^\star, \delta^\star)$ with (without loss of generality, we assume that the vector $\mathbf{t}$ defining the dome has a unit norm):

$$\mathbf{t}^\star = \arg \max_{\mathbf{t} : \|\mathbf{t}\|_2 = 1} \inf_{\mathbf{a} \in \mathcal{U}} \langle \mathbf{t}, \mathbf{a} \rangle, \tag{71}$$

$$\delta^\star = \min_{\mathbf{a} \in \mathcal{U}} \langle \mathbf{t}^\star, \mathbf{a} \rangle. \tag{72}$$

[18]Note that the assumption "$\lambda < \lambda_{\max}$" in Lemma 4 is satisfied by virtue of our working hypotheses (see Section II-E).

[19]Note that $\mathbf{c} \neq \mathbf{0}$ because of working hypotheses in Section II-E.

On the other hand, the minimum-volume sphere covering all the elements of $\mathcal{U}$ is given by $\mathcal{B}(\tilde{\mathbf{t}}^\star, \epsilon^\star)$ with

$$\tilde{\mathbf{t}}^\star = \arg\min_{\tilde{\mathbf{t}}} \sup_{\mathbf{a}\in\mathcal{U}} \left\| \mathbf{a} - \tilde{\mathbf{t}} \right\|_2, \tag{73}$$

$$\epsilon^\star = \max_{\mathbf{a}\in\mathcal{U}} \left\| \mathbf{a} - \tilde{\mathbf{t}}^\star \right\|_2. \tag{74}$$

We first show that the optimal parameters $(\mathbf{t}^\star, \delta^\star)$ and $(\tilde{\mathbf{t}}^\star, \epsilon^\star)$ are related as follows

$$\tilde{\mathbf{t}}^\star = \bar{\delta}\,\mathbf{t}^\star, \tag{75}$$

$$\epsilon^\star = \sqrt{1 - \bar{\delta}^2}, \tag{76}$$

where

$$\bar{\delta} = \max(0, \delta^\star). \tag{77}$$

Indeed, setting $\tilde{\mathbf{t}} = \beta\mathbf{u}$ with $\beta \geq 0$ and $\|\mathbf{u}\|_2 = 1$, (73) can also be rewritten as

$$(\beta^\star, \mathbf{u}^\star) = \arg\min_{\beta\geq0, \|\mathbf{u}\|_2=1} \sup_{\mathbf{a}\in\mathcal{U}} \left\| \mathbf{a} - \beta\mathbf{u} \right\|_2^2,$$

$$= \arg\min_{\beta\geq0, \|\mathbf{u}\|_2=1} \left(\beta^2 - 2\beta \inf_{\mathbf{a}\in\mathcal{U}} \langle\mathbf{u},\mathbf{a}\rangle \right). \tag{78}$$

From (78), we clearly have that $\mathbf{u}^\star = \arg\max_{\|\mathbf{u}\|_2=1} \inf_{\mathbf{a}\in\mathcal{U}} \langle\mathbf{u},\mathbf{a}\rangle$. In view of (71), we thus have $\mathbf{u}^\star = \mathbf{t}^\star$. Taking this fact into account, we deduce

$$\beta^\star = \arg\min_{\beta\geq0}\left(\beta^2 - 2\beta\delta^\star\right) = \bar{\delta},$$

where $\bar{\delta}$ is defined in (77), and thus $\tilde{\mathbf{t}}^\star = \bar{\delta}\,\mathbf{t}^\star$. Plugging $\tilde{\mathbf{t}}^\star = \bar{\delta}\,\mathbf{t}^\star$ in (74) and using the definition of $\delta^\star$ in (72), we find

$$\epsilon^\star = \sqrt{1 - \bar{\delta}^2}.$$

This shows (75)-(76).

We now prove the result of the lemma, that is

$$\mathcal{D}(\mathbf{t}^\star, \delta^\star) \subseteq \mathcal{B}(\tilde{\mathbf{t}}^\star, \epsilon^\star).$$

This statement can be equivalently rewritten as

$$\forall\mathbf{a}\in\mathcal{D}(\mathbf{t}^\star,\delta^\star): \ \left\|\mathbf{a} - \bar{\delta}\mathbf{t}^\star\right\|_2^2 \leq 1 - \bar{\delta}^2.$$

If $\bar{\delta} = 0$, the inequality is satisfied since $\|\mathbf{a}\|_2 \leq 1$ $\forall\mathbf{a}\in\mathcal{D}(\mathbf{t}^\star,\delta^\star)$. If $\bar{\delta} = \delta^\star$, the inequality is also verified because

$$\|\mathbf{a} - \delta^\star\mathbf{t}^\star\|_2^2 = 1 + (\delta^\star)^2 - 2\delta^\star\langle\mathbf{t}^\star,\mathbf{a}\rangle$$
$$\leq 1 - (\delta^\star)^2,$$

where the last inequality follows from the fact that

$$\forall\mathbf{a}\in\mathcal{D}(\mathbf{t}^\star,\delta^\star): \ \langle\mathbf{t}^\star,\mathbf{a}\rangle \geq \delta^\star.$$

REFERENCES

[1] Mario A. T. Figueiredo, "Teaching a new trick to an old dog: Revisiting the quadratic programming formulation of sparse recovery using ADMM," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on.* May 2014, pp. 1512–1516, IEEE.

[2] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by Basis Pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999.

[3] Amir Beck and Marc Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[4] Neal Parikh and Stephen Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[5] M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *IMA Journal of Numerical Analysis*, vol. 20, no. 3, pp. 389–403, July 2000.

[6] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–451, 2004.

[7] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[8] Martin Jaggi, "Revisiting frank-wolfe: Projection-free sparse convex optimization," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28.* 2013, ICML'13, pp. I–427–I–435, JMLR.org.

[9] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani, "Safe feature elimination in sparse supervised learning," Tech. Rep. UC/EECS-2010-126, EECS Dept., University of California at Berkeley, Sept. 2010.

[10] Zhen J. Xiang, Hao Xu, and Peter J. Ramadge, "Learning sparse representations of high dimensional data on large scale dictionaries," in *Advances in Neural Information Processing Systems 24*, J. Shawe-taylor, Zemel, P. Bartlett, Pereira, and Weinberger, Eds., pp. 900–908. 2011.

[11] Liang Dai and Kristiaan Pelckmans, "An ellipsoid based, two-stage screening test for BPDN," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. Aug. 2012, pp. 654–658, IEEE.

[12] Jie Wang, Peter Wonka, and Jieping Ye, "Lasso screening rules via dual polytope projection," *Journal of Machine Learning Research*, 2015.

[13] Antoine Bonnefoy, Valentin Emiya, Liva Ralaivola, and Remi Gribonval, "Dynamic screening: Accelerating first-order algorithms for the Lasso and group-Lasso," *Signal Processing, IEEE Transactions on*, vol. 63, no. 19, pp. 5121–5132, Oct. 2015.

[14] O. Fercoq, A. Gramfort, and J. Salmon, "Mind the duality gap: safer rules for the Lasso," in *ICML*, 2015.

[15] C. Herzet and A. Malti, "Safe screening tests for LASSO based on firmly non-expansiveness," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 4732–4736.

[16] Zhen J. Xiang and Peter J. Ramadge, "Fast Lasso screening tests based on correlations," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. Mar. 2012, pp. 2137–2140, IEEE.

[17] Zhen J. Xiang, Yun Wang, and Peter J. Ramadge, "Screening tests for Lasso problems," arXiv:1405.4897v1, May 2014.

[18] Yohann de Castro and Fabrice Gamboa, "Exact reconstruction using Beurling minimal extrapolation," *Journal of Mathematical Analysis and Applications*, vol. 395, no. 1, pp. 336 – 354, 2012.

[19] Kristian Bredies and Hanna Katriina Pikkarainen, "Inverse problems in spaces of measures," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 19, no. 1, pp. 190–218, 2013.

[20] Emmanuel J. Candès and Carlos Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Comm. Pure Appl. Math*, vol. 67, no. 6, pp. 906–956, June 2014.

[21] Vincent Duval and Gabriel Peyré, "Exact support recovery for sparse spikes deconvolution," *Foundations of Computational Mathematics*, vol. 15, no. 5, pp. 1315–1355, 2015.

[22] Cédric Herzet and Angélique Drémeau, "Joint screening tests for lasso," in *ICASSP 2018-IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 1–5.

[23] Cassio Fraga Dantas and Rémi Gribonval, "Faster and still safe: combining screening techniques and structured dictionaries to accelerate the lasso," in *ICASSP 2018-IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 1–5.

[24] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing.*, Applied and Numerical Harmonic Analysis. Birkhaüser, 2013.

[25] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, USA, 2003.

[26] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical Recipes in C (2Nd Ed.): The Art of Scientific Computing*, Cambridge University Press, New York, NY, USA, 1992.