Safe Squeezing for Antisparse Coding

Clément Elvira and Cédric Herzet

Abstract—Spreading the information over all coefficients of a representation is a desirable property in many applications such as digital communication or machine learning. This so-called antisparse representation can be obtained by solving a convex program involving an ℓ_∞-norm penalty combined with a quadratic discrepancy. In this paper, we propose a new method-ology, dubbed safe squeezing, to accelerate the computation of antisparse representation. We describe a test that allows to detect saturated entries in the solution of the optimization problem. The contribution of these entries is compacted into a single vector, thus operating a form of dimensionality reduction. We propose two algorithms to solve the resulting lower dimensional problem. Numerical experiments show the effectiveness of the proposed method to detect the saturated components of the solution and illustrates the induced computational gains in the resolution of the antisparse problem.

Index Terms—antisparse coding, safe screening, scaled projected-gradient algorithm, Frank-Wolfe algorithm.

I. INTRODUCTION

IN the last decades, convex optimization has become a central element in the resolution of inverse problems. This success revolves around two main ingredients. First, it has been shown that proper "regularizing" functions enforce desirable properties on the solutions of convex problems. One of the first (and maybe most striking) example of such a behavior is the use of "ℓ_1-norm" regularization, promoting sparsity of the solutions. The second ingredient which sparked the success of convex optimization is the advent of numerical procedures able to solve efficiently (up to some accuracy) problems involving thousands to billions of variables. To name a few, let us mention the Augmented Lagrangian methods [1], the forward-backward splitting [2], or the Alternating direction method of multipliers [3].

Of particular interest in this paper is an acceleration method first proposed by El Ghaoui *et al.* in [4] in the context of sparsity-promoting convex problems, namely "*safe screening*". This procedure leverages two main elements. First, the so-lutions of convex problems involving ℓ_1 regularization are typically sparse, *i.e.*, contains a large number of zeros. For example, the solutions of the well-known "Lasso" problem

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x} \in \mathbb{R}^n} \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1, \qquad (1)$$

where $\mathbf{y} \in \mathbb{R}^m$ is an observation vector, $\mathbf{A} \in \mathbb{R}^{m \times n}$ a representation matrix and λ a penalization parameter, are

known to contain at most m nonzero components, see [5, Theorem 3.1] or [6, Theorem 3]. Second, if the position of (some of) the zeros in $\mathbf{x}_\lambda^\star$ are known, (1) can be transformed into a problem of smaller dimension where the columns of $\mathbf{A}$ associated to the position of known zeros are simply discarded. Solving this reduced problem may then potentially result in huge memory and computational savings. As far as sparsity-promoting problems are concerned, the idea of safe screening thus consists in designing simple tests allowing to detect zeros of the solution.

Since the seminal work by El Ghaoui *et al.*, safe screening has been developed and improved in many contributions of the literature, see *e.g.*, [7–10]. Recently, this strategy has also been extended to other families of regularizers [11–13]. However, to the best of our knowledge, all these contributions have focused so far on the resolution of "sparsity-promoting" problems, where the solutions contain many zeros. In this paper, we show that the principles ruling safe screening may be extended to another family of convex optimization problems. More specifically, we address the following optimization problem

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x} \in \mathbb{R}^n} \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_\infty, \qquad (2 - \mathcal{P}_\lambda^\infty)$$

where the admissible range of the coefficients $\mathbf{x}$ is penalized through an ℓ_∞-norm. Here, the ℓ_∞-norm is advocated for spreading the information over all representation coefficients in the most uniform way. For this reason, the solutions of ($2 - \mathcal{P}_\lambda^\infty$) are sometimes referred to as *"antisparse"* or *"spread"* since, contrary to the Lasso problem, they are known to be dense with many *"saturated"* entries satisfying

$$|\mathbf{x}^\star(i)| = \|\mathbf{x}^\star\|_\infty. \qquad (3)$$

Although less popular than its "sparsity-promoting" ℓ_1 coun-terpart, the "spreading" property of the ℓ_∞ norm has proved to be of practical interest in several applicative domains, *e.g.*, to design robust analog-to-digital conversion schemes [14, 15], reduce the peak-to-average power ratio in multi-carrier transmissions [16–18], perform approximate nearest neighbor search [19], outlier detection [20] or robust beamforming [21]. Besides, when combined with a set of linear equality con-straints, minimizing an ℓ_∞-norm is referred to as the *minimum-effort control problem* in the optimal-control framework [22, 23].

From a numerical point of view, although many generic convex optimization tools may be straightforwardly applied to problem ($2 - \mathcal{P}_\lambda^\infty$), the design of algorithms specifically dedicated to this kind of problems has also focused less attention than its ℓ_1 counterpart. In this contribution, we make one step in that direction by showing that the main principles underlying "safe screening" may be extended to problem ($2 - \mathcal{P}_\lambda^\infty$). More specifically, we emphasize that

$(2 - \mathcal{P}_\lambda^\infty)$ can be reduced to a problem of smaller dimension by identifying "saturated" components of the solution $\mathbf{x}^\star$ and propose simple tests to identify them. We refer to the proposed methodology as *"safe squeezing"* since it allows to compact several components of $\mathbf{x}^\star$ into one single variable. We show numerically that the proposed safe squeezing procedure can significantly reduce the computational complexity needed to solve the antisparse problem $(2 - \mathcal{P}_\lambda^\infty)$ to some accuracy.

The rest of this paper is organized as follows. Section II defines the main notations used in this work. In Section III, we introduce the proposed safe squeezing test and discuss the problem's dimensionality reduction it allows. In Section IV we describe two algorithms to solve the equivalent "reduced" optimization problem. In Section V, we illustrate the performance of the proposed methodology in numerical experiments. Concluding remarks are finally made in Section VI.

II. NOTATIONS

Unless otherwise specified, we will use the following convention throughout the paper. The vectors are denoted by lowercase bold letters (*e.g.,* $\mathbf{x}$) and matrices by uppercase bold letters (*e.g.,* $\mathbf{A}$); $\mathbf{x}(i)$ refers to the ith component of $\mathbf{x}$ and $\mathbf{a}_i$ to the ith column of $\mathbf{A}$. Given $\mathbf{q} \in \mathbb{R}^q$ and $w \in \mathbb{R}$, the shorthand notations "$\mathbf{q} \leq w$" should be understood as "$\mathbf{q}(i) \leq w\ \forall i$". The notations $\ker(\mathbf{A})$ and $\mathrm{kruskal}(\mathbf{A})$ will be used to denote the null space and the Kruskal rank [5, p. 56] of $\mathbf{A}$, respectively.

Caligraphic letters are used to denote sets (*e.g.,* $\mathcal{I}$). $\bar{\mathcal{I}}$ stands for the complementary set of $\mathcal{I}$ and $\mathrm{card}(\mathcal{I})$ refers to the cardinality of $\mathcal{I}$.

Given a vector $\mathbf{x} \in \mathbb{R}^n$ and a set of indices $\mathcal{I} \subseteq \{1, \ldots, n\}$, we let $\mathbf{x}_\mathcal{I}$ be the vector of components of $\mathbf{x}$ with indices in $\mathcal{I}$. Similary, $\mathbf{A}_\mathcal{I}$ denotes the submatrix of $\mathbf{A}$ whose columns have indices in $\mathcal{I}$. Finally, $\sigma_{\min}(\mathbf{A})$ and $\sigma_{\max}(\mathbf{A})$ refer to the lowest and highest eigenvalue of $\mathbf{A}$, respectively.

III. DETECTING AND EXPLOITING SATURATION

A. Working hypotheses

In the rest of this paper, we will assume that the following working hypothesis is verified:

$$0 < \lambda < \left\| \mathbf{A}^\mathrm{T} \mathbf{y} \right\|_1. \tag{4}$$

The left inequality is natural since letting $\lambda = 0$ is tantamount to removing the penalization of problem $(2 - \mathcal{P}_\lambda^\infty)$. On the other hand, the right inequality ensures that the solution of problem $(2 - \mathcal{P}_\lambda^\infty)$ is not the all-zero vector $\mathbf{0}_n$. More specifically, it can be shown that

$$\|\mathbf{x}^\star\|_\infty > 0 \iff \lambda \text{ verifies (4)}. \tag{5}$$

A proof of this result is available in Appendix A. Hypothesis (4) implies in particular that $\mathbf{y} \notin \ker(\mathbf{A}^\mathrm{T})$ since the two inequalities cannot be satisfied simultaneously in the opposite case. In the sequel, we will always assume that (4) holds although not explicitly mentioned in our statements.

We will also assume that $(2 - \mathcal{P}_\lambda^\infty)$ admits a unique minimizer $\mathbf{x}^\star$. Although our subsequent derivations may be quite easily extended to the general case, this working hypothesis

greatly simplifies the exposition of our procedure. We will add different comments all along our presentation to point out how the non-uniqueness of the solutions of $(2 - \mathcal{P}_\lambda^\infty)$ would modify the quantities and results at stake.

B. Dimensionality reduction via saturation detection

In this section, we illustrate how the knowledge of the positions of saturated entries in $\mathbf{x}^\star$ can lead to memory and complexity savings in the resolution of $(2 - \mathcal{P}_\lambda^\infty)$. Let[1]

$$\mathcal{I}_+^\star \triangleq \{i \colon \mathbf{x}^\star(i) = +\|\mathbf{x}^\star\|_\infty\}, \tag{6a}$$
$$\mathcal{I}_-^\star \triangleq \{i \colon \mathbf{x}^\star(i) = -\|\mathbf{x}^\star\|_\infty\}, \tag{6b}$$

be the sets of positive and negative saturated components of $\mathbf{x}^\star$, and $\mathcal{I}^\star = \mathcal{I}_-^\star \cup \mathcal{I}_+^\star$. Then, for any $\mathcal{I} = \mathcal{I}_+ \cup \mathcal{I}_-$ with $\mathcal{I}_+ \subseteq \mathcal{I}_+^\star$ and $\mathcal{I}_- \subseteq \mathcal{I}_-^\star$, problem $(2 - \mathcal{P}_\lambda^\infty)$ can be equivalently rewritten as

$$(w^\star, \mathbf{q}^\star) \in \underset{(w, \mathbf{q}) \in \mathbb{R} \times \mathbb{R}^{n - \mathrm{card}(\mathcal{I})}}{\arg\min} \frac{1}{2}\|\mathbf{y} - \mathbf{A}_{\bar{\mathcal{I}}}\mathbf{q} - \mathbf{s}\,w\|_2^2 + \lambda w$$
$$\text{subject to } \mathbf{q} \leq w, -\mathbf{q} \leq w \qquad (7 - \mathcal{P}_\lambda^{\mathrm{sq}})$$

where

$$\mathbf{s} \triangleq \sum_{i \in \mathcal{I}_+} \mathbf{a}_i - \sum_{i \in \mathcal{I}_-} \mathbf{a}_i. \tag{8}$$

In particular, the following bijection holds between the solutions of $(2 - \mathcal{P}_\lambda^\infty)$ and $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$:

$$\mathbf{x}_{\bar{\mathcal{I}}}^\star = \mathbf{q}^\star \tag{9}$$
$$\mathbf{x}^\star(i) = \begin{cases} -w^\star & \text{if } i \in \mathcal{I}_- \\ +w^\star & \text{if } i \in \mathcal{I}_+. \end{cases} \tag{10}$$

In other words, any minimizer of $(2 - \mathcal{P}_\lambda^\infty)$ defines a minimizer of $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$, and vice-versa.

The rationale behind problem $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$ is as follows. Since any minimizer of $(2 - \mathcal{P}_\lambda^\infty)$ weighs (up to a sign) the elements in $\mathcal{I}$ by the same value (namely $\|\mathbf{x}^\star\|_\infty$), the signed summation of the columns of $\mathbf{A}_\mathcal{I}$ can be done once for all in advance. This is the meaning of vector $\mathbf{s}$ in (8). Variable w plays the role of the largest absolute value of the elements of $\mathbf{x}$. In particular, the constraints "$\mathbf{q} \leq w$" and "$-\mathbf{q} \leq w$" ensure that the absolute value of the coefficients weighting the elements of $\mathbf{A}_{\bar{\mathcal{I}}}$ is no larger than w.

We note that, although defining the same set of solutions, the dimensionality of problems $(2 - \mathcal{P}_\lambda^\infty)$ and $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$ can be quite different: whereas the former manipulates n-dimensional variables, the latter only involves an optimization space of dimension $n - \mathrm{card}(\mathcal{I}) + 1$. As a limit case, when $\mathcal{I} = \mathcal{I}^\star$, the size of problem $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$ drops down to $n - \mathrm{card}(\mathcal{I}^\star) + 1$. The next lemma shows that the dimensionality reduction can be drastic in this case:[2]

Lemma 1. *If* $\mathrm{kruskal}(\mathbf{A}) = m$ *then* $\mathrm{card}(\mathcal{I}^\star) \geq n - m + 1$.

Lemma 1 is a direct consequence of Theorem 1 below. The hypothesis "$\mathrm{kruskal}(\mathbf{A}) = m$" holds as soon as any group

---

[1]If $\mathbf{x}^\star$ is not unique, $\mathcal{I}^\star$ must be understood as the set of indices of the components saturating for *all* the solutions of $(2 - \mathcal{P}_\lambda^\infty)$.

[2]A related result is stated in [24, Sec. 4.1].

of m columns of $\mathbf{A} \in \mathbb{R}^{m \times n}$ is linearly independent [5, p. 56]. This is, for example, a typical setup in machine learning applications where the columns of matrix $\mathbf{A}$ contain randomly generated features.[3] In this case, we see from Lemma 1 that the dimension of problem $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$ is no larger than m if $\mathcal{I} = \mathcal{I}^\star$. In the overcomplete setting, when $m \ll n$, solving $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$ instead of $(2 - \mathcal{P}_\lambda^\infty)$ then leads to dramatic dimensionality reduction.

A decrease of the size of the problem leads de facto to a reduction of the storage and computational costs necessary to evaluate the solution. Tackling problem $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$ instead of $(2 - \mathcal{P}_\lambda^\infty)$ may thus be obviously of interest to save computational resources. The relevance of $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$ is however conditioned to the identification of a (large) subset of $\mathcal{I}^\star$. In the rest of this section, we propose a procedure, dubbed "*safe squeezing*", to perform efficiently this task. The term "*safe*" refers to the fact that all the indices identified by our procedure necessarily belong to $\mathcal{I}^\star$. The term "*squeezing*" relates to the dimensionality reduction allowed by the identification of (some of) the elements of $\mathcal{I}^\star$.

C. Safe squeezing test

In this section, we provide the main ingredient of our safe squeezing procedure. It takes the form of a simple test ensuring that an index i belongs to $\mathcal{I}^\star$, and is grounded on the following result:

Theorem 1. *Let $\mathcal{I} \subseteq \mathcal{I}^\star$ and*

$$\mathbf{u}^\star = \arg\max_{\mathbf{u} \in \mathcal{U}_\mathcal{I}} \tfrac{1}{2}\|\mathbf{y}\|_2^2 - \tfrac{1}{2}\|\mathbf{y} - \mathbf{u}\|_2^2 \qquad (11\text{-}\mathcal{D}_\lambda^\infty)$$

where

$$\mathcal{U}_\mathcal{I} \triangleq \left\{ \mathbf{u} \colon \left\| \mathbf{A}_\mathcal{I}^{\mathrm{T}} \mathbf{u} \right\|_1 + \mathbf{s}^{\mathrm{T}} \mathbf{u} \leq \lambda \right\} \qquad (12)$$

and $\mathbf{s}$ is defined as in (8). Then,

$$\forall i : \left| \mathbf{a}_i^{\mathrm{T}} \mathbf{u}^\star \right| > 0 \Rightarrow i \in \mathcal{I}^\star_{\mathrm{sign}(\mathbf{a}_i^{\mathrm{T}} \mathbf{u}^\star)}. \qquad (13)$$

Moreover, if $\mathrm{kruskal}(\mathbf{A}) = m$, $\mathbf{a}_i^{\mathrm{T}} \mathbf{u}^\star \neq 0$ for at least $n-m+1$ $\mathbf{a}_i$'s.

A proof of this result is available in Appendix B. Theorem 1 provides an appealing way of identifying saturated components in $\mathbf{x}^\star$: we have from (13) that $i \in \mathcal{I}^\star$ as soon as $\mathbf{a}_i$ is not orthogonal to some vector $\mathbf{u}^\star$ defined in $(11\text{-}\mathcal{D}_\lambda^\infty)$. Moreover, if $\mathrm{kruskal}(\mathbf{A}) = m$, then at least $n - m + 1$ saturated components of $\mathbf{x}^\star$ can be identified by verifying that $\left| \mathbf{a}_i^{\mathrm{T}} \mathbf{u}^\star \right| > 0$.

Unfortunately, in practice, finding a maximizer of $(11\text{-}\mathcal{D}_\lambda^\infty)$ turns out to be as difficult as solving our target optimization problem $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$. In particular, if $\mathcal{I} = \emptyset$, addressing $(11\text{-}\mathcal{D}_\lambda^\infty)$ requires the same order of complexity as solving $(2 - \mathcal{P}_\lambda^\infty)$. Hence, the direct use of (13) is of poor interest to save computational resources. Nevertheless, we emphasize below that relaxed versions of (13) can be devised to identify subsets of $\mathcal{I}^\star$ with a low computational burden.

Before giving a detailed description of the proposed method, we make two important remarks about $\mathbf{u}^\star$. First, we see from $(11\text{-}\mathcal{D}_\lambda^\infty)$ that $\mathbf{u}^\star$ is defined as a maximizer of a feasible[4] and strictly-concave problem with a continuous and coercive cost function. Hence, $\mathbf{u}^\star$ always exists [25, Propositions A.8] and is unique [25, Propositions B.10]. This justifies the equality sign used in $(11\text{-}\mathcal{D}_\lambda^\infty)$. In fact, as shown in Appendix B-A, $(11\text{-}\mathcal{D}_\lambda^\infty)$ corresponds to the Lagrangian dual of problem $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$.

Second, the value of $\mathbf{u}^\star$ does not depend on the particular subset $\mathcal{I} \subseteq \mathcal{I}^\star$ considered in $(11\text{-}\mathcal{D}_\lambda^\infty)$. This fact is only stated here for the sake of keeping the discussion at a reasonable level of complexity, but is proved in Appendix B. We note that, although the value of $\mathbf{u}^\star$ is independent of $\mathcal{I}$, there may be some computational advantages in considering $\mathcal{I}$'s with large numbers of elements. In particular, we emphasize in Section III-D below that the computational cost of the "dual scaling" operation, needed to implement the proposed safe squeezing test, evolves linearly with $\mathrm{card}(\bar{\mathcal{I}})$.

We are now ready to expose the main building block of our safe squeezing procedure. Let $\mathcal{S}$ be some subset of $\mathbb{R}^m$ such that

$$\mathbf{u}^\star \in \mathcal{S}. \qquad (14)$$

For now, we only assume that some $\mathcal{S}$ verifying (14) is available. We will see in the next section how to construct such a region. A region satisfying (14) is usually referred to as "safe region" in the screening literature, and the same convention will be used hereafter.

We note that since $\mathbf{u}^\star \in \mathcal{S}$, we have

$$\begin{aligned} \min_{\mathbf{u} \in \mathcal{S}} \mathbf{a}_i^{\mathrm{T}} \mathbf{u} &< \mathbf{a}_i^{\mathrm{T}} \mathbf{u}^\star \\ \max_{\mathbf{u} \in \mathcal{S}} \mathbf{a}_i^{\mathrm{T}} \mathbf{u} &> \mathbf{a}_i^{\mathrm{T}} \mathbf{u}^\star, \end{aligned} \qquad (15)$$

so that

$$\begin{aligned} \min_{\mathbf{u} \in \mathcal{S}} \mathbf{a}_i^{\mathrm{T}} \mathbf{u} > 0 &\Rightarrow \mathbf{a}_i^{\mathrm{T}} \mathbf{u}^\star > 0 \\ \max_{\mathbf{u} \in \mathcal{S}} \mathbf{a}_i^{\mathrm{T}} \mathbf{u} < 0 &\Rightarrow \mathbf{a}_i^{\mathrm{T}} \mathbf{u}^\star < 0. \end{aligned} \qquad (16)$$

Using (13), we thus obtain

$$\begin{aligned} \min_{\mathbf{u} \in \mathcal{S}} \mathbf{a}_i^{\mathrm{T}} \mathbf{u} > 0 &\Rightarrow i \in \mathcal{I}^\star_+ \\ \max_{\mathbf{u} \in \mathcal{S}} \mathbf{a}_i^{\mathrm{T}} \mathbf{u} < 0 &\Rightarrow i \in \mathcal{I}^\star_-. \end{aligned} \qquad (17)$$

Hence, if the maximum and minimum values of $\mathbf{a}_i^{\mathrm{T}} \mathbf{u}$ over $\mathcal{S}$ are easy to evaluate, the left-hand side of (17) provides a simple test to determine whether i belongs to $\mathcal{I}^\star_+$ or $\mathcal{I}^\star_-$. In particular, considering a safe region $\mathcal{S}$ with a spherical geometry, that is

$$\mathcal{S} = \mathcal{B}(\mathbf{c}, r) \triangleq \left\{ \mathbf{u} \in \mathbb{R}^m \colon \|\mathbf{u} - \mathbf{c}\|_2 \leq r \right\}, \qquad (18)$$

for some $\mathbf{c} \in \mathbb{R}^m$ and $r > 0$, leads to

$$\begin{aligned} \min_{\mathbf{u} \in \mathcal{B}(\mathbf{c}, r)} \mathbf{a}_i^{\mathrm{T}} \mathbf{u} &= \mathbf{a}_i^{\mathrm{T}} \mathbf{c} - \|\mathbf{a}_i\|_2 r \\ \max_{\mathbf{u} \in \mathcal{B}(\mathbf{c}, r)} \mathbf{a}_i^{\mathrm{T}} \mathbf{u} &= \mathbf{a}_i^{\mathrm{T}} \mathbf{c} + \|\mathbf{a}_i\|_2 r. \end{aligned} \qquad (19)$$

Plugging these expressions into (17) yields the following result:

---

[3] More particularly, this is the case with probability 1 when the columns of $\mathbf{A}$ are drawn according to any distribution that admits a density with respect to the Lebesgue measure.

[4] We always have $\mathbf{u} = \mathbf{0}_m \in \mathcal{U}_\mathcal{I}$ since $\lambda > 0$ by hypothesis.

Theorem 2 (Safe sphere squeezing test). *If $\mathbf{u}^\star \in \mathcal{B}(\mathbf{c}, r)$, then*

$$\left|\mathbf{a}_i^{\mathrm{T}}\mathbf{c}\right| > r\left\|\mathbf{a}_i\right\|_2 \Rightarrow i \in \mathcal{I}^\star_{\mathrm{sign}(\mathbf{a}_i^{\mathrm{T}}\mathbf{c})}. \tag{20}$$

Squeezing test (20) provides a very practical way of identifying some elements of $\mathcal{I}^\star$. In particular, testing whether $i \in \mathcal{I}^\star$ only requires to evaluate *one* inner product between $\mathbf{a}_i$ and the center $\mathbf{c}$ of a safe sphere. We note however that passing (20) is only a *sufficient* condition for $i \in \mathcal{I}^\star$. In practice, depending on the choice of the center and the radius of the safe region, some elements of $\mathcal{I}^\star$ may not be identified by the proposed safe squeezing test. As a general rule of thumb, "small" safe regions will lead to squeezing tests able to identify more saturated components. We will illustrate this behavior in our numerical results in Section V. We elaborate on the construction of "good" safe sphere regions in the next subsection. For now, let us just notice that if $\mathbf{c} = \mathbf{u}^\star$ and $r = 0$, $\mathcal{S} = \mathcal{B}(\mathbf{c}, r)$ is safe and its volume is equal to zero. In this case, the safe squeezing test (20) boils down to (13). In particular, provided that $\mathrm{kruskal}(\mathbf{A}) = m$, at least $n - m + 1$ saturated components of $\mathbf{x}^\star$ can be identified (see last part of Theorem 1).

As a final remark, let us mention that Theorem 2 corresponds to *one* specific particularization of (17) (*i.e.*, to spherical regions $\mathcal{S}$). As with procedures derived for Lasso screening, it is also possible to devise squeezing tests based on safe regions having more refined geometries (*e.g.*, dome [26] or truncated dome [27]). The extension of our safe squeezing procedure to these geometries is however left for future work and will not be considered hereafter.

D. Construction of safe spheres

A cornerstone of the *safe squeezing* methodology presented in Section III-C is the identification of a "safe" region, that is a region $\mathcal{S}$ verifying (14). We elaborate on this problem hereafter: we emphasize that several safe spheres derived in the context of safe screening for Lasso can be reused for safe squeezing with minor modifications. We focus here on the spheres "ST1" and "GAP" respectively proposed in [4] and [7].

Our reasoning is based on two key observations. First, the Lasso dual problem shares the same cost function as problem (11-$\mathcal{D}_\lambda^\infty$) but with a definition of the dual feasible set different from (12), see *e.g.*, [9, Eq. (2)]. Second, the safe spheres proposed in [4, 7] are based on convex optimality conditions and the knowledge of some dual feasible point $\mathbf{u}$, but never on the specific definition of the dual feasible set. Hence, the center and radius of the "ST1" and "GAP" spheres proposed in [4] and [7] can be reused here with the provisio that the dual variable $\mathbf{u}$ appearing in these expressions is feasible for (11-$\mathcal{D}_\lambda^\infty$) (that is $\mathbf{u} \in \mathcal{U}_\mathcal{I}$).

For example, the ST1 safe sphere derives from the fact that, by definition of $\mathbf{u}^\star$:

$$\forall \mathbf{u} \in \mathcal{U}_\mathcal{I} : \left\|\mathbf{y} - \mathbf{u}\right\|_2 \geq \left\|\mathbf{y} - \mathbf{u}^\star\right\|_2. \tag{21}$$

Hence, the sphere $\mathcal{B}(\mathbf{c}, r)$ with

$$\begin{aligned} \mathbf{c} &= \mathbf{y} \\ r &= \left\|\mathbf{y} - \mathbf{u}\right\|_2 \end{aligned} \tag{22}$$

is safe for problem (11-$\mathcal{D}_\lambda^\infty$). We see that expression of the center and the radius in (22) are the same as those proposed in [4] with the difference that $\mathbf{u}$ must here be feasible for (11-$\mathcal{D}_\lambda^\infty$) rather than for the Lasso dual problem.

Similarly, the expression of the GAP safe sphere can be extended as follows in the context of safe squeezing: for any $\mathcal{I} \subseteq \mathcal{I}^\star$ and $(w, \mathbf{q}, \mathbf{u})$ verifying

$$\mathbf{q} \leq w, \quad -\mathbf{q} \leq w, \quad \mathbf{u} \in \mathcal{U}_\mathcal{I}, \tag{23}$$

the sphere $\mathcal{B}(\mathbf{c}, r)$ with

$$\begin{aligned} \mathbf{c} &= \mathbf{u} \\ r &= \sqrt{2\,\mathrm{gap}(w, \mathbf{q}, \mathbf{u})} \end{aligned} \tag{24}$$

and

$$\begin{aligned} \mathrm{gap}(w, \mathbf{q}, \mathbf{u}) &= \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}_{\bar{\mathcal{I}}}\mathbf{q} - \mathbf{s}\,w\|_2^2 + \lambda w \\ &\quad - (\tfrac{1}{2}\|\mathbf{y}\|_2^2 - \tfrac{1}{2}\|\mathbf{y} - \mathbf{u}\|_2^2) \end{aligned} \tag{25}$$

is safe for problem (11-$\mathcal{D}_\lambda^\infty$). The proof of this result is similar to that provided in [7] and is omitted here. As for the sphere ST1, the expressions of the center and radius remain the same as for the Lasso problem but the dual (resp. primal) variable(s) coming into play must now be feasible for problem (11-$\mathcal{D}_\lambda^\infty$) (resp. $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$).

The only computational difficulty in evaluating (22) and (24) stands in the identification of a dual feasible point. We may nevertheless identify such a point by extending the "dual scaling" procedure proposed in [4, Section 3.3] to the present setup. More precisely, given any $\mathbf{z} \in \mathbb{R}^m$, we clearly have that

$$\mathbf{u} = \mathrm{dualscal}(\mathbf{z}) \tag{26}$$

where

$$\mathrm{dualscal}(\mathbf{z}) = \begin{cases} \mathbf{z} & \text{if } \left\|\mathbf{A}_{\bar{\mathcal{I}}}^{\mathrm{T}}\mathbf{z}\right\|_1 + \mathbf{s}^{\mathrm{T}}\mathbf{z} \leq 0, \\ \frac{\lambda}{\left\|\mathbf{A}_{\bar{\mathcal{I}}}^{\mathrm{T}}\mathbf{z}\right\|_1 + \mathbf{s}^{\mathrm{T}}\mathbf{z}}\mathbf{z} & \text{otherwise.} \end{cases}$$

is dual feasible. We note that the implementation of (26) requires the computation of $\mathrm{card}(\bar{\mathcal{I}}) + 1$ inner products in $\mathbb{R}^m$.

E. Static versus dynamic squeezing

Similarly to screening methods, the proposed safe squeezing procedure can be used either in a *"static"* or a *"dynamic"* way.

Static squeezing refers to the case where (20) is applied once for all (with $\mathcal{I} = \emptyset$) on the columns of matrix $\mathbf{A}$ before the application of a numerical optimization procedure. In such a case, the dual feasible point $\mathbf{u}$ used to build the safe sphere is commonly defined as follows:

$$\mathbf{u} = \mathrm{dualscal}(\mathbf{y}). \tag{27}$$

Static screening may for example be of interest to reduce the problem's dimensionality so that its size fits the computer's memory requirements.

Dynamic squeezing consists in interleaving test (20) with the iterations of an optimization procedure. Here, the goal is to refine the quality of the safe region and increase the cardinality of $\mathcal{I}$ all along the optimization process. The general principle of dynamic squeezing (applied to the resolution of $(2 - \mathcal{P}_\lambda^\infty)$) is described in Algorithm 1. At each iteration,

Algorithm 1: Principle of dynamic squeezing

```
   // Initialization
1  x⁽⁰⁾ = 0ₙ, 𝓘⁽⁰⁾ = ∅
2  u⁽⁰⁾ = dualscal(y)
3  t = 1 // Iteration index
4  repeat
      // Squeezing test
5     (c⁽ᵗ⁾, r⁽ᵗ⁾) = sphere_param(x⁽ᵗ⁻¹⁾, u⁽ᵗ⁻¹⁾, 𝓘⁽ᵗ⁻¹⁾)

6     𝓘⁽ᵗ⁻½⁾ = squeezing_test(c⁽ᵗ⁾, r⁽ᵗ⁾)
7     𝓘⁽ᵗ⁾ = 𝓘⁽ᵗ⁻½⁾ ∪ 𝓘⁽ᵗ⁻¹⁾

      // Iterations of the optimization procedure
8     (x⁽ᵗ⁾, u⁽ᵗ⁾) = optim_update(x⁽ᵗ⁻¹⁾, u⁽ᵗ⁻¹⁾, 𝓘⁽ᵗ⁾)

      // Update iteration index
9     t = t + 1

10 until convergence criterion is met
   Output: x⁽ᵗ⁾, 𝓘⁽ᵗ⁾
```

Algorithm 2: Frank-Wolfe algorithm for problem (29).

```
   Input: y, A_𝓘̄, s, w̄, w₀, q₀
1  w⁽⁰⁾ = w₀, q⁽⁰⁾ = q₀ // Initialization
2  t = 1 // Iteration index
3  repeat
      // Search of a feasible descent direction
4     w⁽ᵗ⁻½⁾ = w̄
5     q⁽ᵗ⁻½⁾ = w̄ sign(A_𝓘̄ᵀ(y − A_𝓘̄q⁽ᵗ⁻¹⁾ − s w⁽ᵗ⁻¹⁾))

      // Convex update
6     w⁽ᵗ⁾ = γw⁽ᵗ⁻¹⁾ + (1 − γ)w⁽ᵗ⁻½⁾
7     q⁽ᵗ⁾ = γq⁽ᵗ⁻¹⁾ + (1 − γ)q⁽ᵗ⁻½⁾

      // Update iteration index
8     t = t + 1

9  until convergence criterion is met
   Output: w⁽ᵗ⁾, q⁽ᵗ⁾
```

the parameters of a safe sphere are evaluated by using a couple a primal-dual feasible points $(\mathbf{x}^{(t-1)}, \mathbf{u}^{(t-1)})$, see line 5. The safe sphere is then exploited in the squeezing test (20) to identify some set $\mathcal{I}^{(t-½)} \subseteq \mathcal{I}^\star$, see line 6. In line 7, the indices identified at the current iteration are merged to those previously identified. Finally, the primal-dual feasible couple is updated in line 8. This operation is typically carried out by running a few iterations of a numerical procedure addressing problem $(7 - \mathcal{P}_\lambda^{\text{sq}})$ with $\mathcal{I} = \mathcal{I}^{(t)}$ and initialized with the previous iterate $(\mathbf{x}^{(t-1)}, \mathbf{u}^{(t-1)})$. In the common case where the optimization method only returns a primal iterate $\mathbf{x}^{(t)}$, a feasible dual point can for example be computed by dual scaling of $\mathbf{z} = \mathbf{y} - \mathbf{A}\mathbf{x}^{(t)}$. By construction, we have

$$\mathcal{I}^{(0)} \subseteq \mathcal{I}^{(1)} \subseteq \ldots \subseteq \mathcal{I}^{(t)} \tag{28}$$

so that the dimension of the reduced problem $(7 - \mathcal{P}_\lambda^{\text{sq}})$ considered in line 8 does not increase through the iterations of Algorithm 1. In particular, if $(\mathbf{x}^{(t)}, \mathbf{u}^{(t)})$ converges to $(\mathbf{x}^\star, \mathbf{u}^\star)$, the parameters of the GAP safe sphere converge to $(\mathbf{c} = \mathbf{u}^\star, r = 0)$ and, from Theorem 1, $\text{card}(\mathcal{I}^{(t)}) \geq n - m + 1$ after a sufficient number of iterations when $\text{kruskal}(\mathbf{A}) = m$.

IV. ALGORITHMIC SOLUTIONS FOR $(7 - \mathcal{P}_\lambda^{\text{sq}})$

A crucial difference between "screening" and "squeezing" is the nature of the optimization problems obtained after reduction. Whereas the reduced problem has the same mathematical form as the original one in the context of screening, the reduced problem $(7 - \mathcal{P}_\lambda^{\text{sq}})$ obtained after squeezing is structurally different from $(2 - \mathcal{P}_\lambda^\infty)$. As a consequence, numerical procedures devised to solve $(2 - \mathcal{P}_\lambda^\infty)$ (*e.g.*, [18]) cannot be used straightforwardly to address $(7 - \mathcal{P}_\lambda^{\text{sq}})$.

In this section, we describe two algorithmic solutions for problem $(7 - \mathcal{P}_\lambda^{\text{sq}})$: in Section IV-A a numerical procedure based on the Frank-Wolfe algorithm [25, Section 2.2] is proposed; in Section IV-B, we present a strategy based on a rescaled projected gradient method [25, Section 2.3]. These

procedures will be used in the next section to assess computational gain allowed by the proposed safe squeezing procedure.

A. Frank-Wolfe algorithm

Let us consider the following equivalent formulation of $(7 - \mathcal{P}_\lambda^{\text{sq}})$:

$$\min_{(w,\mathbf{q}) \in \mathbb{R} \times \mathbb{R}^{n - \text{card}(\mathcal{I})}} \frac{1}{2}\|\mathbf{y} - \mathbf{A}_{\bar{\mathcal{I}}}\mathbf{q} - \mathbf{s}\,w\|_2^2 + \lambda w$$

$$\text{subject to } \begin{cases} \mathbf{q} \leq w, \ -\mathbf{q} \leq w, \\ w \leq \overline{w}, \end{cases} \tag{29}$$

where $\overline{w}$ is some constant such that $w^\star \leq \overline{w}$. We will discuss the identification of $\overline{w}$ at the end of this section. For now, let us notice that the additional constraint "$w \leq \overline{w}$" does not change the minimizers of the problem but makes the feasible set compact. As consequence, the conditional gradient method, also known as Frank-Wolfe algorithm [28], can be applied to find a numerical solution of (29).

The recursions of the Frank-Wolfe algorithm particularized to problem (29) are described in Algorithm 2. One iteration of the procedure consists of two main steps: the identification of a feasible descent direction and the update of the current iterate.

The feasible descent direction is identified as the solution of the following optimization problem:

$$(w^{(t-½)}, \mathbf{q}^{(t-½)}) = \arg\min_{(w,\mathbf{q})} (w \quad \mathbf{q}^\text{T}) \mathbf{g}^{(t-1)} \tag{30}$$

$$\text{subject to } \begin{cases} \mathbf{q} \leq w, \ -\mathbf{q} \leq w \\ w \leq \overline{w}, \end{cases}$$

where $\mathbf{g}^{(t-1)}$ is the gradient of the cost function evaluated at the current iterate $(w^{(t-1)}, \mathbf{q}^{(t-1)})$. Problem (30) is linear and admits the closed-form solution given in lines 4-5 of Algorithm 2.

The update of the current iterate is performed via a convex combination of $(w^{(t-½)}, \mathbf{q}^{(t-½)})$ and $(w^{(t-1)}, \mathbf{q}^{(t-1)})$, see lines 6-7 in Algorithm 2. Parameter γ is chosen to

minimize the cost function over the line segment joining $(\widetilde{w}^{(t-1/2)}, \mathbf{q}^{(t-1/2)})$ and $(w^{(t-1)}, \mathbf{q}^{(t-1)})$. In the particular setup considered here, this amounts to finding the value of γ minimizing a quadratic function over $[0, 1]$, which admits a closed-form solution.

We conclude this section by showing that parameter $\overline{w}$ can be chosen as follows:

$$\overline{w} = \frac{1}{2\lambda}\|\mathbf{y} - \mathbf{A}_{\bar{\mathcal{I}}}\mathbf{q} - \mathbf{s}\,w\|_2^2 + w \qquad (31)$$

where $(w, \mathbf{q})$ is any feasible point of problem $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$. This follows from the following observation:

$$\lambda w^\star \leq \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}_{\bar{\mathcal{I}}}\mathbf{q}^\star - \mathbf{s}\,w^\star\|_2^2 + \lambda w^\star$$
$$\leq \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}_{\bar{\mathcal{I}}}\mathbf{q} - \mathbf{s}\,w\|_2^2 + \lambda w,$$

where the last inequality is valid for any couple $(w, \mathbf{q})$ feasible for $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$. Hence, the definition of $\overline{w}$ in (31) verifies "$w^\star \leq \overline{w}$" as required at the beginning of this section. In particular, $(w, \mathbf{q}) = (0, \mathbf{0}_{\mathrm{card}(\bar{\mathcal{I}})})$ is feasible for $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$ and leads to $\overline{w} = \frac{\|\mathbf{y}\|_2^2}{2\lambda}$. We will consider the latter choice in our simulations in Section V.

B. Rescaled projected gradient method

We consider a rescaled version of $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$:

$$\min_{(\widetilde{w}, \mathbf{q}) \in \mathbb{R} \times \mathbb{R}^{n - \mathrm{card}(\mathcal{I})}} \tfrac{1}{2}\Big\|\mathbf{y} - \mathbf{A}_{\bar{\mathcal{I}}}\mathbf{q} - \tfrac{\mathbf{s}}{\alpha}\widetilde{w}\Big\|_2^2 + \tfrac{\lambda}{\alpha}\widetilde{w} \qquad (32)$$
$$\text{subject to } \alpha\mathbf{q} \leq \widetilde{w}, \ -\alpha\mathbf{q} \leq \widetilde{w}$$

where $\alpha > 0$. The choice of α will be discussed later on in this section and will be made based on convergence rate arguments.

Particularizing the gradient projection algorithm described in [25, Section 2.3] to problem (32) leads to Algorithm 3. Each iteration is divided into three main stages: a "gradient descent" step (lines 4-6), a "projection" operation (line 8) and a "convex update" (line 9-10).

Starting from a current estimate $(\mathbf{q}^{(t-1)}, \widetilde{w}^{(t-1)})$, the gradient step consists in updating the current iterate in the direction of the negative gradient of the cost function (by an amount $\eta > 0$). The step η is chosen to minimize the cost function along the direction of the (negative) gradient. We note that since the cost function is quadratic, the value of η admits a simple closed-form expression.

The projection step consists in solving the optimization problem specified in line 8 of Algorithm 3. Although conceptually simpler than (32), the latter does not admit any closed-form solution. We show nevertheless in Appendix C that the procedure described in Algorithm 4 can compute the unique minimizer of this problem in a *finite* number of steps (upper-bounded by $\mathrm{card}(\bar{\mathcal{I}})$) with a complexity scaling as $\mathcal{O}(\mathrm{card}(\bar{\mathcal{I}})\log(\mathrm{card}(\bar{\mathcal{I}})))$. The interested reader is referred to Appendix C-A for more details on the computational complexity.

Finally, similarly to the Frank-Wolfe algorithm, the update of the current iterate is performed via a convex combination, see lines 9-10 in Algorithm 3. The choice of γ is made to minimize the value of the cost function over the line segment joining the two points.

Algorithm 3: Projected gradient method for problem (32)

Input: $\mathbf{y}$, $\mathbf{A}_{\bar{\mathcal{I}}}$, $\mathbf{s}$, α, η, $\widetilde{w}_0$, $\mathbf{q}_0$

1 $\widetilde{w}^{(0)} = \widetilde{w}_0$, $\mathbf{q}^{(0)} = \mathbf{q}_0$ // `Initialization`
2 $t = 1$ // `Iteration index`
3 **repeat**

    // `Gradient step`

4    $\mathbf{q}^{(t-2/3)} = \mathbf{q}^{(t-1)} + \eta\,\mathbf{A}_{\bar{\mathcal{I}}}^{\mathrm{T}}\mathbf{z}^{(t-1)}$
5    $\widetilde{w}^{(t-2/3)} = \widetilde{w}^{(t-1)} + \frac{\eta}{\alpha}\big(\mathbf{s}^{\mathrm{T}}\mathbf{z}^{(t-1)} - \lambda\big)$
6   where $\mathbf{z}^{(t-1)} = \mathbf{y} - \mathbf{A}_{\bar{\mathcal{I}}}\mathbf{q}^{(t-1)} - \frac{\mathbf{s}}{\alpha}\widetilde{w}^{(t-1)}$

    // `Projection step (solved via Algorithm 4)`

7    $(\widetilde{w}^{(t-1/3)}, \mathbf{q}^{(t-1/3)})$
8    $= \arg\min_{\mathbf{q}, \widetilde{w}} \|\mathbf{q} - \mathbf{q}^{(t-2/3)}\|_2^2 + (\widetilde{w} - \widetilde{w}^{(t-2/3)})^2$
      subject to $\alpha\mathbf{q} \leq \widetilde{w}$, $-\alpha\mathbf{q} \leq \widetilde{w}$

    // `Convex update`

9    $w^{(t)} = \gamma w^{(t-1)} + (1 - \gamma)w^{(t-1/3)}$
10   $\mathbf{q}^{(t)} = \gamma\mathbf{q}^{(t-1)} + (1 - \gamma)\mathbf{q}^{(t-1/3)}$

    // `Update iteration index`

11   $t = t + 1$
12 **until** *convergence criterion is met*

Output: $\mathbf{q}^{(t)}$, $\widetilde{w}^{(t)}$

Before concluding this section, we elaborate on the choice of the parameters α. It can be seen (see [25, Section 2.3.1]), that the speed of convergence of Algorithm 3 to the global minimizer of (32) is a direct function of the conditioning of the Hessian of the cost function[5]

$$\mathbf{H} = \begin{pmatrix} \alpha^{-2}\|\mathbf{s}\|_2^2 & \alpha^{-1}\mathbf{s}^{\mathrm{T}}\mathbf{A}_{\bar{\mathcal{I}}} \\ \alpha^{-1}\mathbf{A}_{\bar{\mathcal{I}}}^{\mathrm{T}}\mathbf{s} & \mathbf{A}_{\bar{\mathcal{I}}}^{\mathrm{T}}\mathbf{A}_{\bar{\mathcal{I}}} \end{pmatrix}. \qquad (33)$$

A proper choice of α may thus help improving the conditioning of this matrix and, consequently, enhance the speed of convergence of the overall procedure. In this paper, we suggest the following rule of thumb:

$$\alpha = \begin{cases} \|\mathbf{s}\|_2 & \text{if } \mathbf{s} \neq \mathbf{0}_m \\ 1 & \text{otherwise.} \end{cases} \qquad (34)$$

This choice is motivated by the ideal case where the columns of $\mathbf{A}_{\bar{\mathcal{I}}}$ and $\mathbf{s}$ are orthogonal. In this case, the Hessian matrix in (33) is diagonal and (because the columns of $\mathbf{A}_{\bar{\mathcal{I}}}$ have unit-norm) we have:

$$\frac{\sigma_{\min}(\mathbf{H})}{\sigma_{\max}(\mathbf{H})} = \min\Big(\alpha^{-1}\|\mathbf{s}\|_2, \alpha\|\mathbf{s}\|_2^{-1}\Big). \qquad (35)$$

Setting α as in (34) ensures that the conditioning of $\mathbf{H}$ is equal to 1 in this case. Although in practice $\mathbf{A}_{\bar{\mathcal{I}}}$ and $\mathbf{s}$ are commonly not orthogonal, we noticed in our numerical experiments that (34) leads to much better speeds of convergence than the trivial choice $\alpha = 1$.

---

[5]That is the ratio between $\sigma_{\max}(\mathbf{H})$ and $\sigma_{\min}(\mathbf{H})$.

Algorithm 4: Projection of $(\widetilde{w}, \mathbf{q})$ onto the feasible set of (32).

Input: $\mathbf{q}, \widetilde{w}, \alpha$

// **Initialization**

1   $\widetilde{w}^{(0)} = \widetilde{w}$

2   $\mathcal{Q}^{(0)} \triangleq \{i : \alpha|\mathbf{q}(i)| \geq \widetilde{w}^{(0)}\}$

3   $t = 0$ // **Iteration index**

    // **Main recursion**

4   **repeat**

5     $t = t + 1$

6     $\widetilde{w}^{(t)} = \dfrac{\alpha^2}{\alpha^2 + \operatorname{card}(\mathcal{Q}^{(t-1)})} \left(\widetilde{w} + \sum_{i \in \mathcal{Q}^{(t-1)}} \dfrac{|\mathbf{q}(i)|}{\alpha} \right)$

7     $\mathcal{Q}^{(t)} \triangleq \{i : \alpha|\mathbf{q}(i)| \geq \widetilde{w}^{(t)}\}$

8   **until** $\mathcal{Q}^{(t)} = \mathcal{Q}^{(t-1)}$

    // **Final assignment**

9   **if** $\widetilde{w}^{(t)} \leq 0$ **then**

10    $\widetilde{w}_{\mathrm{proj}} = 0$ and $\mathbf{q}_{\mathrm{proj}} = \mathbf{0}$

11   **else**

12    $\widetilde{w}_{\mathrm{proj}} = \widetilde{w}^{(t)}$

13    $\mathbf{q}_{\mathrm{proj}}(i) = \begin{cases} \mathbf{q}(i) & \text{if } i \notin \mathcal{Q}^{(t)} \\ \alpha^{-1}\operatorname{sign}(\mathbf{q}(i))\widetilde{w}_{\mathrm{proj}} & \text{otherwise} \end{cases}$

14   **end**

Output: $\widetilde{w}_{\mathrm{proj}}, \mathbf{q}_{\mathrm{proj}}$

V. NUMERICAL EXPERIMENTS

In this section, we report several simulation results illustrating the relevance of the proposed squeezing methodology. In Section V-A, we study the effectiveness of the proposed strategy as a function of the "quality" of the safe sphere used in its construction. In Sections V-B and V-C, we demonstrate that safe squeezing can be advocated to accelerate the resolution of problem $(2 - \mathcal{P}_\lambda^\infty)$: in Section V-B, we investigate the computational savings allowed by "dynamic squeezing"; in Section V-C, we show that for a given computational budget, safe squeezing enables to reach better convergence properties.

In our simulations, new realizations of the dictionary $\mathbf{A}$ and observation vector $\mathbf{y}$ are drawn for each trial according to the following distributions. The observation vector is generated according to a standard normal distribution. The dictionary obeys one of the following distribution: *i)* the entries are i.i.d. realizations of a centered Gaussian; *ii)* the entries are i.i.d. realizations of a uniform law on $[0, 1]$; *iii)* the rows are randomly-sampled rows of a DCT matrix; *iv)* the rows are shifted versions of a Gaussian curve. In the following, these four options will be respectively referred to as "Gaussian", "Uniform", "DCT" and "Toeplitz".

A. Effectiveness of the safe sphere squeezing test

In this section, we assess the effectiveness of the proposed squeezing test. More specifically, we evaluate the proportion of saturated entries of $\mathbf{x}^\star$ which can be identified by the proposed procedure as a function of the "quality" of the safe sphere.
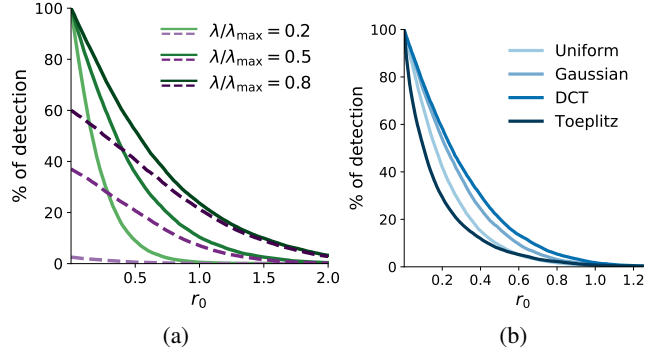

Fig. 1: Percentage of saturated entries identified with the safe sphere squeezing test as a function of the radius for (a) the Gaussian dictionary, 3 values of $\lambda/\lambda_{\max}$ using the GAP sphere (thick line) and the ST1 sphere (dashed line) (b) the four dictionaries and $\lambda/\lambda_{\max} = 0.2$.

Fig. 1 presents our results. They have been obtained by repeating the experiment described below 50 times. For each simulation trial, we draw a realization of $\mathbf{y} \in \mathbb{R}^{200}$ and $\mathbf{A} \in \mathbb{R}^{200 \times 300}$ according to the distributions described at the beginning of Section V. We then compute a "high-accuracy" primal-dual solution $(\mathbf{x}_\mathrm{a}, \mathbf{u}_\mathrm{a})$ of $(2 - \mathcal{P}_\lambda^\infty)$-$(11\text{-}\mathcal{D}_\lambda^\infty)$. $\mathbf{x}_\mathrm{a}$ is evaluated numerically by solving $(2 - \mathcal{P}_\lambda^\infty)$ with Algorithm 3 and a dual gap of 10^{-14} as stopping criterion. $\mathbf{u}_\mathrm{a}$ is obtained by dual scaling of $\mathbf{y} - \mathbf{A}\mathbf{x}_\mathrm{a}$ (see (26)). This high-accuracy couple is used to construct the following safe spheres:

$$\mathbf{c}_{\mathrm{ST1}} = \mathbf{y} \tag{36}$$

$$r_{\mathrm{ST1}} = r_0 + \|\mathbf{y} - \mathbf{u}_\mathrm{a}\|_2 \tag{37}$$

and

$$\mathbf{c}_{\mathrm{GAP}} = \mathbf{u}_\mathrm{a} \tag{38}$$

$$r_{\mathrm{GAP}} = r_0 + \sqrt{2\operatorname{gap}(\|\mathbf{x}_\mathrm{a}\|_\infty, \mathbf{x}_\mathrm{a}, \mathbf{u}_\mathrm{a})} \tag{39}$$

where gap is defined in (25) and $r_0 \geq 0$ is some parameter. If $r_0 = 0$, (36)-(37) and (38)-(39) respectively correspond to the parameters[6] of the ST1 and GAP safe spheres presented in Section III-D. By construction, these expressions thus lead to safe spheres for any value of $r_0 \geq 0$. In our experiments, we consider (36)-(39) with different $r_0 \geq 0$ to simulate different qualities of safe spheres. More specifically, if $r_0 = 0$ and $(\mathbf{x}_\mathrm{a}, \mathbf{u}_\mathrm{a}) = (\mathbf{x}^\star, \mathbf{u}^\star)$, r_{ST1} (resp. r_{GAP}) is the smallest possible radius for an ST1 (resp. GAP) safe sphere. For each value of r_0 and each simulation trial, we apply the proposed safe squeezing test (20) with the spheres defined in (36)-(39). Fig. 1 represents the proportion of saturated entries of $\mathbf{x}^\star$ detected by our squeezing test as a function of parameter r_0. The results are averaged over 50 realizations for each value of r_0.

Fig. 1a shows the percentage of saturated entries detected with a Gaussian dictionary and three different values of $\lambda/\lambda_{\max}$. Although we restrict our attention to the Gaussian dictionary, the following comments are valid for the other dictionaries as well. When $r_0 = 0$ (*i.e.*, the radius of the safe sphere is equal to r_{GAP} or r_{ST1}), one observes that all the

---

[6]Computed with the primal-dual feasible couple $(\mathbf{x}_\mathrm{a}, \mathbf{u}_\mathrm{a})$.

saturated entries are detected with the GAP sphere. However, only a subset of these entries are detected with the ST1 sphere (up to 60% in the best case). This result was expected since the high accuracy estimate of $\mathbf{u}^\star$ leads to $\mathbf{u}_a \simeq \mathbf{u}^\star$ and r_{GAP} is therefore close to zero when $r_0 = 0$ (see Section III-E). On the other side, the radius of the ST1 sphere is usually bounded away from zero even when $\mathbf{u}_a = \mathbf{u}^\star$ and $r_0 = 0$.

For the two tests, the number of saturated entries detected decreases as the volume of the safe sphere (parametrized by r_0 here) increases. The performance also seems to depend on λ: higher values lead to better detection results.

Fig. 1b shows the percentage of saturated entries detected for the four dictionaries when $\lambda/\lambda_{\max} = 0.2$. We observe that some dictionaries seem to be more prone to squeezing. For example, up to 30% additional saturated entries are detected with the DCT dictionary as compared to the Toeplitz dictionary when $r_0 \simeq 0.3$.

B. Complexity savings

In this section, we evaluate the overall computational gain induced by the proposed method in the resolution of problem $(2 - \mathcal{P}_\lambda^\infty)$. More precisely, the total number of operations[7] carried out by several algorithms (with and without squeezing) to solve $(2 - \mathcal{P}_\lambda^\infty)$ to some accuracy are compared. In the case where a sequential implementation of the optimization procedures is considered, this figure of merit can be directly related to CPU time needed to solve $(2 - \mathcal{P}_\lambda^\infty)$; if parallel implementation is envisaged, the number of operations can be interpreted in terms of "energy consumption".

Four numerical procedures are assessed: *(i)* FITRA, an accelerated proximal gradient algorithm proposed in [18]; *(ii)* the Frank-Wolfe procedure described in Algorithm 2 with no squeezing ($\mathcal{I} = \emptyset$); *(iii)* the projected gradient presented in Algorithm 3 with dynamic squeezing (see Algorithm 1); *(iv)* the Frank-Wolfe in Algorithm 2 with dynamic screening. Procedures *(ii)*, *(iii)* and *(iv)* are respectively denoted as "FW", "PGs", "FWs" in the sequel. Finally, empirical evidence suggests that the GAP sphere test is more effective than ST1, and we concentrate therefore on the former in our experiments.

Since this experiment aims at demonstrating the advantage of resorting to safe squeezing, we only compare procedures belonging to the same family of algorithms and therefore with similar rate of convergence. In such a setting, the induced reduction in terms of number of operations is most likely only due to safe squeezing. In particular, we compare FW against FWs and FITRA against PGs since the two couples of procedures belong to the families of Frank-Wolfe and proximal/projection gradient methods, respectively.

The numbers of operations are computed for a decreasing sequence of penalization parameters $\{\lambda_j\}_{j=1}^p$. Moreover, we consider a "warm-start" initialization: for each λ_j, the optimization procedure is initialized with the solution obtained for λ_{j-1}. For λ_1, all algorithms are initialized with the zero vector $\mathbf{0} \in \mathbb{R}^n$. For each run, a stopping criterion in terms of dual gap
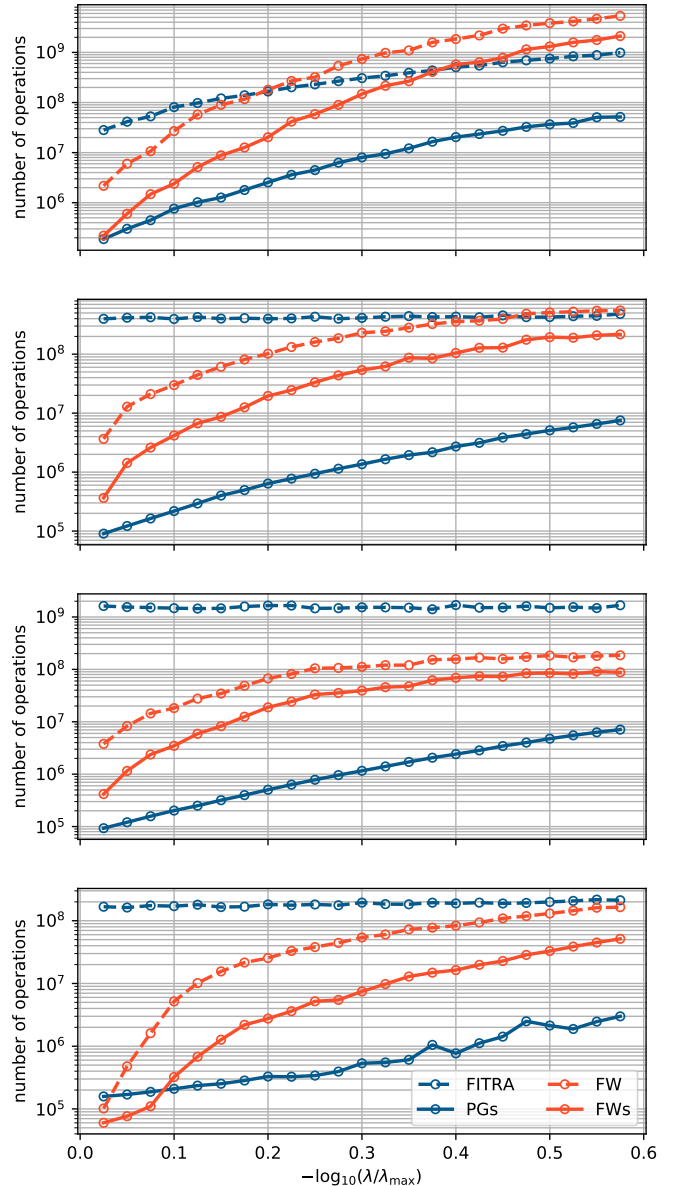

Fig. 2: Number of operations required to achieve convergence for different values of $\lambda/\lambda_{\max}$ and four dictionaries: Uniform (first row), Gaussian (second row), DCT (third row) and Toeplitz (fourth row).

is used. To obtain results with the same order of magnitude[8], two values of the dual gap are used: 10^{-7} for FITRA and PGs and 10^{-4} for the two Frank-Wolfe procedures (FW and FWs). All results are averaged over 50 simulation trials. For each simulation trial, we draw a realization of $\mathbf{y} \in \mathbb{R}^{100}$ and $\mathbf{A} \in \mathbb{R}^{100 \times 150}$ according to the distributions described at the beginning of Section V.

Fig. 2 represents the averaged number of operations needed by FITRA, PGs, FW and FWs to achieve the required convergence accuracy. Each figure corresponds to a different dictionary (Gaussian, Uniform, DCT and Toeplitz). The curves

---

[7]We restrict our attention to multiplications since they entail a much higher computational burden than additions in floating-point arithmetic.

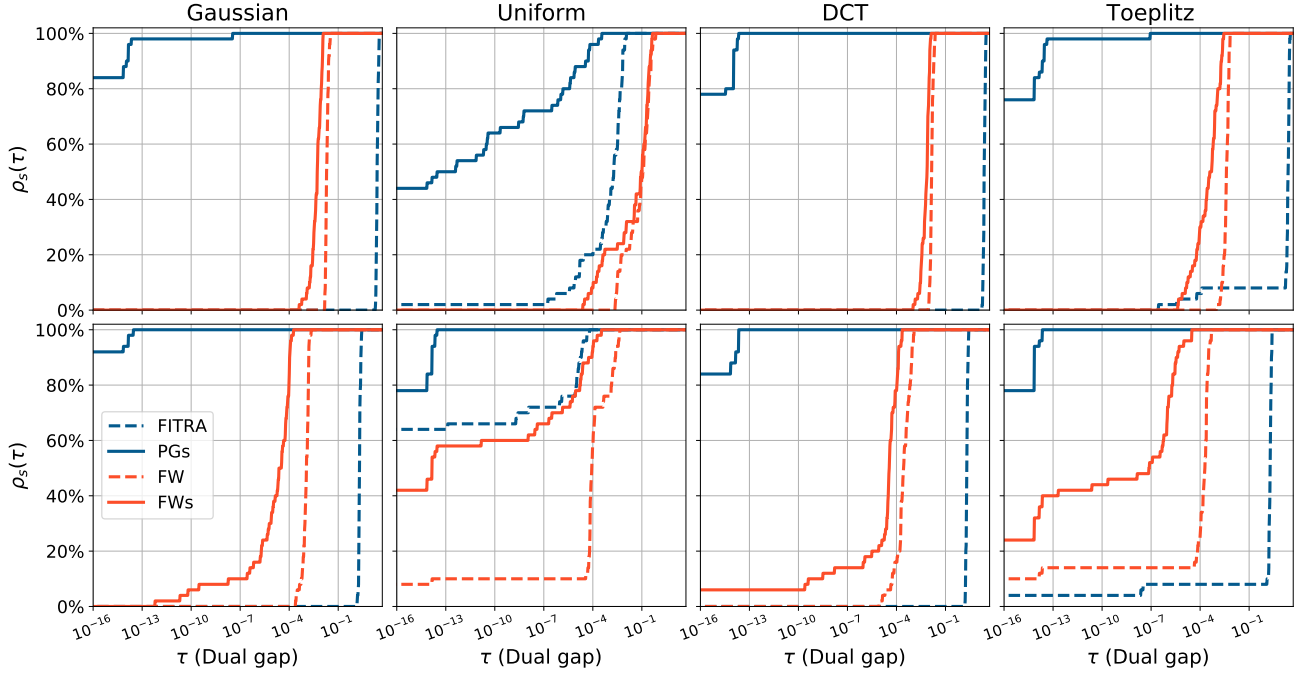[8]Frank-Wolfe procedures are indeed known to converge much more slowly than proximal/projection gradient methods.

Fig. 3: Performance profiles obtained with the Uniform, Gaussian, DCT and Toeplitz dictionaries. First row: $\lambda/\lambda_{\max} = 0.3$; second row: $\lambda/\lambda_{\max} = 0.8$.

are plotted as a function of $-\log_{10}(\lambda/\lambda_{\max})$. We note that, to allow a fair comparison between the different algorithms, we took particular care in the counting of the operations. In particular, each operation which can be reused latter on in an optimization procedure is not counted twice in our simulation.

In all cases, one observes that PGs (straight blue line) achieves convergence with fewer operations than FITRA (dashed blue line). As an example, up to 10^4 times more operations are required by FITRA when $-\log_{10}(\lambda/\lambda_{\max}) = 0.1$ and $\mathbf{A}$ is a DCT dictionary (third row). However, such a "computational" saving tends to decrease as $-\log_{10}(\lambda/\lambda_{\max})$ increases. Such a behavior is in good accordance with typical results of the screening literature where the performance degrades for low values of the penalization parameter.

As far as our simulation setup is concerned, these results show the advantage of considering safe squeezing to reduce the computational burden of computing antisparse representations. The observed reduction of complexity may be explained by analyzing the cost of one iteration of the four algorithms. Indeed, since the squeezing tests are done at few additional costs, the implementations of the four considered algorithms results in a cost per iteration of $\mathcal{O}(mn^{(t)})$ where $n^{(t)} = n - \operatorname{card}(\mathcal{I}^{(t)})$ and $\mathcal{I}^{(t)}$ is the number of saturated entries detected at iteration t. While $n^{(t)}$ is decreasing along iterations for PGs and FWs, it remains constant and equal to n for FITRA and FW. The two mains consequences are *(i)* the computational cost of one iteration is smaller and *(ii)* the induced dimensionality reduction may help improving the conditioning of the optimization problem, resulting potentially in fewer iterations needed to reach convergence.

C. Benchmark profiles

As a final assessment of the performance of the proposed squeezing test, we make use of the Dolan-Moré performance profiles [29]. For each family of dictionaries, we generate a test set consisting of $p = 50$ random dictionaries $\mathbf{A} \in \mathbb{R}^{100 \times 150}$ and observations $\mathbf{y} \in \mathbb{R}^{100}$ according to the distributions described at the beginning of Section V. Each problem is then solved by using the four algorithms mentioned in the previous section, with a budget of 10^8 operations.[9] For each run $j \in \{1 \dots p\}$ and solvers $s \in \{$FITRA, PGs, FW, FWs$\}$, we define

$$d_{j,s} \triangleq \text{dual gap achieved by solver } s \text{ at run } j. \quad (40)$$

In this experiment, we compare the performance profile

$$\rho_s(\tau) \triangleq \frac{100}{p} \operatorname{card}(\{j : d_{j,s} \leq \tau\}) \quad \forall \tau \quad (41)$$

of each algorithm s. In other words, for all possible values of the dual gap τ, $\rho_s(\tau)$ is the (empirical) probability that solver s reaches a dual gap no greater than τ.

Figure 3 presents the performance profiles of the four algorithms with different dictionaries and values of the ratio $\lambda/\lambda_{\max}$. As far as our simulation setup is concerned, these results show that for a given budget of operations and any targeted dual gap smaller that 10^{-4}, resorting to PGs (resp. FWs) allows to solve[10] more problems than FITRA (resp. FW). Moreover, PGs outperforms the other procedures: in all

---

[9]The algorithm stops as soon as 10^8 multiplications have been carried out.
[10]that is, reach a solution with the desired dual gap.

scenarios, at least 45% of the runs reach a dual gap with order of magnitude equal to machine precision ($\tau = 10^{-16}$).

As detailed in the last paragraph of Section V-B, resorting to safe squeezing may reduce the computational burden of one iteration of an optimization procedure and help improving the conditioning of the optimization problem through an easy rescaling. With this conclusion in mind, the performance improvements observed in this experiment can be explained by the combination of two following factors. First, improving the conditioning of the problem results potentially in fewer iterations needed to reach a given dual gap. Second, reducing the computational cost of one iteration allows to perform more iterations for the same budget of operations, reaching possibly a solution with a lower dual gap.

VI. CONCLUSIONS

In this paper, we proposed a new methodology, dubbed *safe squeezing*, to accelerate algorithms that compute a so-called antisparse representation. The proposed procedure consists in a set of tests that detect *saturated* entries in the solution of the optimization problem. These tests rely on the notion of safe region, that is a set that contains the solution of the dual problem. Benefiting from the recent developments in the screening literature, we showed that two existing safe spheres can be extended in the context of safe squeezing. The contribution of the saturated entries is then compacted into one single vector, allowing for a potentially dramatic dimensionality reduction. Computing an antisparse representation becomes equivalent to solving a lower dimensional quadratic problem. When interleaved with the iterations of an optimization procedure, the approach permits to dynamically detect new saturated entries all along the optimization process. Numerical simulations showed that the proposed safe squeezing methodology leads to significant computational gains in several numerical setups.

APPENDIX A
PROOF OF THE RESULT IN (5)

In this section, we prove that "$\lambda < \left\| \mathbf{A}^{\mathrm{T}} \mathbf{y} \right\|_1$" is indeed a necessary and sufficient condition for the solution of ($2 - \mathcal{P}_\lambda^\infty$) to be nonzero. A direct application of the Fermat's rule [30, Th. 16.2] ensures that $\mathbf{x}^\star$ is a minimizer of ($2 - \mathcal{P}_\lambda^\infty$) if and only if

$$\lambda^{-1} \mathbf{A}^{\mathrm{T}} (\mathbf{A} \mathbf{x}^\star - \mathbf{y}) \in \partial \| \mathbf{x}^\star \|_\infty \tag{42}$$

where $\partial \| \mathbf{x}^\star \|_\infty$ denotes the sub-differential of the ℓ_∞-norm evaluated at $\mathbf{x}^\star$. Danskin's theorem [25, Th. B.25] leads to

$$\partial \| \mathbf{x} \|_\infty \triangleq \left\{ \mathbf{d} \in \mathbb{R}^n \colon \| \mathbf{d} \|_1 \leq 1 \text{ and } \mathbf{x}^{\mathrm{T}} \mathbf{d} = \| \mathbf{x} \|_\infty \right\} \tag{43}$$

for all $\mathbf{x} \in \mathbb{R}^n$. Result (5) follows by particularizing (42)-(43) to $\mathbf{x}^\star = \mathbf{0}_n$.

APPENDIX B
PROOF OF THE RESULTS IN SECTION III

In this section, we provide a proof of Theorem 1. We also show that the quantity $\mathbf{u}^\star$ defined in (11-$\mathcal{D}_\lambda^\infty$) is independent of the set $\mathcal{I}$.

In our derivations, we focus on the following generic optimization problem:

$$\min_{\substack{\mathbf{z} \in \mathbb{R}^m, \mathbf{q} \in \mathbb{R}^p \\ w \geq 0}} \tfrac{1}{2} \| \mathbf{y} - \mathbf{z} \|_2^2 + \lambda w \quad \text{s.t.} \begin{cases} \mathbf{q} \leq w \\ -\mathbf{q} \leq w \\ \mathbf{z} = \mathbf{B}\mathbf{q} + \mathbf{s}w. \end{cases} \tag{44}$$

Conclusions about problems ($2 - \mathcal{P}_\lambda^\infty$) and ($7 - \mathcal{P}_\lambda^{\mathrm{sq}}$) are then drawn by considering the following substitutions:

Problem ($2 - \mathcal{P}_\lambda^\infty$): $\mathbf{B} = \mathbf{A}, \ \mathbf{s} = \mathbf{0}_m$ (45)

Problem ($7 - \mathcal{P}_\lambda^{\mathrm{sq}}$): $\mathbf{B} = \mathbf{A}_{\bar{\mathcal{I}}}, \ \mathbf{s} = \sum_{i \in \mathcal{I}_+} \mathbf{a}_i - \sum_{i \in \mathcal{I}_-} \mathbf{a}_i.$ (46)

We organize the presentation of the results as follows. In Section B-A we derive the Lagrangian dual problem of (44). In Section B-B, we emphasize some properties of the solutions of the dual problem. The latter are then exploited in Section B-C to derive (the first part of) the result stated in Theorem 1. In Section B-D, we elaborate on the uniqueness of the maximizer of the dual problem and its independence with respect to the set $\mathcal{I}$. Finally, in Section B-E we prove the last part of Theorem 1.

A. Dual problem of (44)

Let $\mathbf{v}_+ \in \mathbb{R}^m$ and $\mathbf{v}_- \in \mathbb{R}^m$ (resp. $\mathbf{u} \in \mathbb{R}^n$) denote the dual variables associated to the inequality (resp. equality) constraint(s) in (44). The Lagrangian function associated to this problem can then be rewritten as

$$\begin{aligned} \mathscr{L}(\mathbf{q}, \mathbf{z}, w, \mathbf{u}, \mathbf{v}_+, \mathbf{v}_-) =& \tfrac{1}{2} \| \mathbf{y} - \mathbf{z} \|_2^2 + \mathbf{u}^{\mathrm{T}} \mathbf{z} \\ &+ w\left(\lambda - \mathbf{s}^{\mathrm{T}} \mathbf{u} - \mathbf{1}_n^{\mathrm{T}} \mathbf{v}_+ - \mathbf{1}_n^{\mathrm{T}} \mathbf{v}_- \right) \\ &+ \mathbf{q}^{\mathrm{T}} \left(-\mathbf{B}^{\mathrm{T}} \mathbf{u} + \mathbf{v}_+ - \mathbf{v}_- \right). \end{aligned}$$

Moreover, the corresponding dual function is defined as

$$D(\mathbf{u}, \mathbf{v}_+, \mathbf{v}_-) \triangleq \min_{\mathbf{q}, \mathbf{z}, w \geq 0} \mathscr{L}(\mathbf{q}, \mathbf{z}, w, \mathbf{u}, \mathbf{v}_+, \mathbf{v}_-).$$

We note that since $\mathbf{v}_+ \geq 0$ and $\mathbf{v}_- \geq 0$, we have $D(\mathbf{u}, \mathbf{v}_+, \mathbf{v}_-) = -\infty$ as soon as the following conditions are not satisfied:

$$\begin{aligned} \lambda - \mathbf{s}^{\mathrm{T}} \mathbf{u} - \mathbf{1}_n^{\mathrm{T}} \mathbf{v}_+ - \mathbf{1}_n^{\mathrm{T}} \mathbf{v}_- &\geq 0, \\ -\mathbf{B}^{\mathrm{T}} \mathbf{u} + \mathbf{v}_+ - \mathbf{v}_- &= \mathbf{0}_p. \end{aligned} \tag{47}$$

Hence, defining

$$\mathcal{D} \triangleq \{ (\mathbf{u}, \mathbf{v}_+, \mathbf{v}_-) \colon \mathbf{v}_+ \geq 0, \mathbf{v}_- \geq 0 \text{ and } (47) \text{ holds} \},$$

the dual function $D(\mathbf{u}, \mathbf{v}_+, \mathbf{v}_-)$ can be written as

$$D(\mathbf{u}, \mathbf{v}_+, \mathbf{v}_-) = \begin{cases} \tfrac{1}{2} \| \mathbf{y} \|_2^2 - \tfrac{1}{2} \| \mathbf{y} - \mathbf{u} \|_2^2 & \text{if } (\mathbf{u}, \mathbf{v}_+, \mathbf{v}_-) \in \mathcal{D} \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem associated to (44) thus reads

$$\max_{(\mathbf{u}, \mathbf{v}_+, \mathbf{v}_-) \in \mathcal{D}} \tfrac{1}{2} \| \mathbf{y} \|_2^2 - \tfrac{1}{2} \| \mathbf{y} - \mathbf{u} \|_2^2. \tag{48}$$

From [25, Proposition 5.2.1],[11] there is no duality gap between (44) and (48); moreover the dual problem (48) has (at least) one maximizer. In fact, we will show in Section B-D below that (48) admits *one unique* maximizer, denoted $(\mathbf{u}^\star, \mathbf{v}_+^\star, \mathbf{v}_-^\star)$ hereafter.

---

[11]More specifically, (44) is a feasible convex problem with linear constraints and the minimum value of the cost function is upper bounded.

B. Connection between $\mathbf{u}^\star$, $\mathbf{v}_+^\star$ and $\mathbf{v}_-^\star$

In this section, we emphasize that $\mathbf{v}_+^\star$ and $\mathbf{v}_-^\star$ are strongly connected to $\mathbf{u}^\star$. More specifically, we have:

$$\begin{aligned}
\mathbf{v}_+^\star(i) &= \max(0, \mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star) \\
\mathbf{v}_-^\star(i) &= -\min(0, \mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star).
\end{aligned} \tag{49}$$

This connection can be shown as follows. First, since there is no duality gap (see Section B-A), any couple of primal-dual optimal variables, say $(w^\star, \mathbf{q}^\star, \mathbf{z}^\star)$ and $(\mathbf{u}^\star, \mathbf{v}_+^\star, \mathbf{v}_-^\star)$, must verify the well-known Karush-Kuhn-Tucker optimality conditions [25, Proposition 5.1.5]. In particular, we have

$$\begin{aligned}
\mathbf{y} - \mathbf{u}^\star &= \mathbf{z}^\star & \text{(50a)} \\
-\mathbf{B}^{\mathrm{T}}\mathbf{u}^\star + \mathbf{v}_+^\star - \mathbf{v}_-^\star &= \mathbf{0}_p & \text{(50b)} \\
\mathbf{v}_\pm^\star &\geq 0 & \text{(50c)} \\
\forall i: \quad \mathbf{v}_+^\star(i)(\mathbf{q}^\star(i) - w^\star) &= 0 & \text{(50d)} \\
\forall i: \quad \mathbf{v}_-^\star(i)(\mathbf{q}^\star(i) + w^\star) &= 0. & \text{(50e)}
\end{aligned}$$

More precisely, (50a) corresponds to first-order optimality condition while (50b) and (50c) partially describe the set of dual feasible variables (see (47)). One also recognizes in (50d) and (50e) the complementary slackness conditions of the problem.

Using these conditions, we have

$$\mathbf{v}_+^\star(i)\mathbf{v}_-^\star(i) = 0 \quad \forall i \tag{51}$$

since

$$\mathbf{v}_+^\star(i) > 0 \overset{(50d)}{\Rightarrow} \mathbf{q}^\star(i) = w^\star \overset{(4)-(5)}{>} 0 \tag{52}$$

$$\overset{(50e)}{\Rightarrow} \mathbf{v}_-^\star(i) = 0, \tag{53}$$

and

$$\mathbf{v}_-^\star(i) > 0 \overset{(50e)}{\Rightarrow} \mathbf{q}^\star(i) = -w^\star \overset{(4)-(5)}{<} 0 \tag{54}$$

$$\overset{(50d)}{\Rightarrow} \mathbf{v}_+^\star(i) = 0. \tag{55}$$

Therefore, combining (50b), (50c) and (51), we obtain the following implications:

$$\mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star = 0 \Rightarrow \begin{cases} \mathbf{v}_-^\star(i) = 0 \\ \mathbf{v}_+^\star(i) = 0 \end{cases} \tag{56}$$

$$\mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star > 0 \Rightarrow \begin{cases} \mathbf{v}_-^\star(i) = 0 \\ \mathbf{v}_+^\star(i) = \mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star \end{cases} \tag{57}$$

$$\mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star < 0 \Rightarrow \begin{cases} \mathbf{v}_-^\star(i) = -\mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star \\ \mathbf{v}_+^\star(i) = 0. \end{cases} \tag{58}$$

Expression (49) corresponds to a compact reformulation of (56), (57) and (58).

C. Necessary optimality condition

We now exploit (49) together with conditions (50d)-(50e) to prove the necessary optimality condition (13) stated in Theorem 1. Let us first show that $\mathbf{u}^\star$ can be expressed as

$$\mathbf{u}^\star = \arg\max_{\mathbf{u} \in \mathcal{U}} \tfrac{1}{2}\|\mathbf{y}\|_2^2 - \tfrac{1}{2}\|\mathbf{y} - \mathbf{u}\|_2^2 \tag{59}$$

where

$$\mathcal{U} \triangleq \{\mathbf{u}: \|\mathbf{B}^{\mathrm{T}}\mathbf{u}\|_1 + \mathbf{s}^{\mathrm{T}}\mathbf{u} \leq \lambda\}. \tag{60}$$

To this end, let us define

$$\mathcal{F} \triangleq \left\{ (\mathbf{u}, \mathbf{v}_+, \mathbf{v}_-): \begin{array}{l} \mathbf{v}_+(i) = \max(0, \mathbf{b}_i^{\mathrm{T}}\mathbf{u}) \\ \mathbf{v}_-(i) = -\min(0, \mathbf{b}_i^{\mathrm{T}}\mathbf{u}) \end{array} \right\}. \tag{61}$$

Clearly, from (48) and (49) we have that $(\mathbf{u}^\star, \mathbf{v}_+^\star, \mathbf{v}_-^\star) \in \mathcal{D} \cap \mathcal{F}$. Hence,

$$(\mathbf{u}^\star, \mathbf{v}_+^\star, \mathbf{v}_-^\star) \in \arg\max_{(\mathbf{u},\mathbf{v}_+,\mathbf{v}_-) \in \mathcal{D} \cap \mathcal{F}} \tfrac{1}{2}\|\mathbf{y}\|_2^2 - \tfrac{1}{2}\|\mathbf{y} - \mathbf{u}\|_2^2. \tag{62}$$

Moreover, combining (47) and (49) leads to

$$\mathcal{D} \cap \mathcal{F} = \left\{ (\mathbf{u}, \mathbf{v}_+, \mathbf{v}_-): \begin{array}{rl} \lambda \geq & \|\mathbf{B}^{\mathrm{T}}\mathbf{u}\|_1 + \mathbf{s}^{\mathrm{T}}\mathbf{u} \\ \mathbf{v}_+(i) = & \max(0, \mathbf{b}_i^{\mathrm{T}}\mathbf{u}) \\ \mathbf{v}_-(i) = & -\min(0, \mathbf{b}_i^{\mathrm{T}}\mathbf{u}) \end{array} \right\}.$$

Plugging this definition into (62), we obtain that $\mathbf{u}^\star$ is defined by (59)-(60). Moreover, $\mathbf{v}_+^\star$ and $\mathbf{v}_-^\star$ are univocally specified from $\mathbf{u}^\star$ via (49).

We finally show that the following condition holds between the optimal primal and dual variables:

$$(\mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star)(\mathbf{q}^\star(i) - \text{sign}(\mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star)w^\star) = 0. \tag{63}$$

Condition (13) then corresponds to a particular instantiation of (63) when $\mathbf{B}$ and $\mathbf{s}$ are defined as in (46). Relation (63) can be obtained as follows. Taking the difference between conditions (50d)-(50e), we first have

$$\mathbf{q}^\star(i)(\mathbf{v}_+^\star(i) - \mathbf{v}_-^\star(i)) - w^\star(\mathbf{v}_+^\star(i) + \mathbf{v}_-^\star(i)) = 0. \tag{64}$$

Now, using (49) we obtain

$$\begin{aligned}
\mathbf{v}_+^\star(i) + \mathbf{v}_-^\star(i) &= |\mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star| & \text{(65a)} \\
\mathbf{v}_+^\star(i) - \mathbf{v}_-^\star(i) &= \mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star. & \text{(65b)}
\end{aligned}$$

Using (65a)-(65b) and the fact that $|\mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star| = (\mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star)\text{sign}(\mathbf{b}_i^{\mathrm{T}}\mathbf{u}^\star)$, we finally end up with the desired result.

D. Uniqueness of $\mathbf{u}^\star$ and independence with respect to $\mathcal{I}$

Problem (59) admits a unique minimizer since the cost function is continuous, coercive and strictly concave, see [25, Propositions A.8 & B.10]. This shows the uniqueness of $\mathbf{u}^\star$. We note that the uniqueness of $(\mathbf{u}^\star, \mathbf{v}_+^\star, \mathbf{v}_-^\star)$ claimed in Section B-A then straightforwardly follows from property (49).

Setting $\mathbf{B}$ and $\mathbf{s}$ as in (46), we now show the independence of $\mathbf{u}^\star$ with respect to the definition of $\mathcal{I} \subseteq \mathcal{I}^\star$. We note that in this case, problem (44) is equivalent to $(7 - \mathcal{P}_\lambda^{\text{sq}})$. From optimality condition (50a), we have

$$\mathbf{u}^\star = \mathbf{y} - \mathbf{z}^\star. \tag{66}$$

Now, we have by definition that

$$\mathbf{z}^\star = \mathbf{A}_{\bar{\mathcal{I}}}\mathbf{q}^\star + \mathbf{s}w^\star. \tag{67}$$

Moreover, because problems $(2 - \mathcal{P}_\lambda^\infty)$ and $(7 - \mathcal{P}_\lambda^{\text{sq}})$ are equivalent (see Section III), we have from (9)-(10) that

$$\forall \mathcal{I} \subseteq \mathcal{I}^\star: \mathbf{A}_{\bar{\mathcal{I}}}\mathbf{q}^\star + \mathbf{s}w^\star = \mathbf{A}\mathbf{x}^\star, \tag{68}$$

where $\mathbf{x}^\star$ is the unique (see Section III-A) minimizer of $(2 - \mathcal{P}_\lambda^\infty)$. Combining the last three equalities, we obtain that $\mathbf{u}^\star$ does not depend on the choice of $\mathcal{I}$ as long as $\mathcal{I} \subseteq \mathcal{I}^\star$.[12]

---

[12]This property holds true in the more generally setup where $\mathbf{x}^\star$ is not unique although the proof is slightly more involved.

E. Minimum number of saturated components

In this section we prove the last part of Theorem 1. Since we showed in Section B-D that $\mathbf{u}^\star$ is independent of the choice of $\mathcal{I} \subseteq \mathcal{I}^\star$, we focus on the case where $\mathcal{I} = \emptyset$ and $\mathbf{B} = \mathbf{A}$, $\mathbf{s} = \mathbf{0}_m$.

We first note that our working hypothesis (4) prevents us from having $\mathbf{u}^\star = \mathbf{0}_m$. Indeed, considering the optimality conditions of problem (59), it can be seen that

$$\mathbf{y} \notin \ker(\mathbf{A}^{\mathrm{T}}) \Rightarrow \mathbf{u}^\star \neq \mathbf{0}_m. \tag{69}$$

Now, as pointed out in Section III-A, hypothesis (4) implies that $\mathbf{y} \notin \ker(\mathbf{A}^{\mathrm{T}})$ and thus necessarily $\mathbf{u}^\star \neq \mathbf{0}_m$.

Since $\mathbf{u}^\star \neq \mathbf{0}_m$, we observe that "$\mathbf{a}_i^{\mathrm{T}}\mathbf{u}^\star = 0$" can only occur for at most $m-1$ $\mathbf{a}_i$'s when $\mathrm{kruskal}(\mathbf{A}) = m$. Indeed, $\mathrm{kruskal}(\mathbf{A}) = m$ implies that any subset of m columns of $\mathbf{A}$ are linearly independent. On the other hand, the vector space $\mathcal{P} \triangleq \{\mathbf{a} \in \mathbb{R}^m : \mathbf{a}^{\mathrm{T}}\mathbf{u}^\star = 0\}$ is a $(m-1)$-dimensional subspace of $\mathbb{R}^m$. Hence, if more than $m-1$ columns of $\mathbf{A}$ belong to $\mathcal{P}$, they must be linearly dependent, which is in contradiction with $\mathrm{kruskal}(\mathbf{A}) = m$. This proves the last part of Theorem 1.

APPENDIX C
CORRECTNESS OF ALGORITHM 4

In this appendix, we prove that the procedure described in Algorithm 4 solves the following projection problem

$$\min_{\mathbf{q}',w'} (w' - \widetilde{w})^2 + \|\mathbf{q}' - \mathbf{q}\|_2^2$$
$$\text{subject to} \quad \alpha\mathbf{q}' \leq w', \ -\alpha\mathbf{q}' \leq w' \tag{70}$$

in a *finite* number of steps. With a slight abuse of notations,[13] we denote hereafter the unique minimizer of this problem as $(\widetilde{w}^\star, \mathbf{q}^\star)$. The dimension of $\mathbf{q}$ will be denoted q hereafter.

The appendix is organized as follows. In Section C-A, we prove the convergence of Algorithm 4 in a number of steps no greater than q and discuss its computational complexity. In Section C-B we elaborate on necessary and sufficient conditions for $(\widetilde{w}^\star, \mathbf{q}^\star)$ to be a minimizer of (70). Finally, we exploit these optimality conditions in Section C-C to show that the output of Algorithm 4 corresponds to the unique minimizer of (70).

A. Convergence of Algorithm 4 in a finite number of steps

We first note that, by construction, the sequence $\{\widetilde{w}^{(t)}\}$ is non-decreasing. As a matter of fact, by definition of $\mathcal{Q}^{(t)}$ we must have

$$\mathcal{Q}^{(t)} \subseteq \mathcal{Q}^{(t-1)}. \tag{71}$$

Moreover, if equality occurs in (71), then the stopping criterion of Algorithm 4 is satisfied and the procedure terminates. Algorithm 4 carries out t iterations provided that

$$\mathcal{Q}^{(t)} = \mathcal{Q}^{(t-1)} \subset \mathcal{Q}^{(t-2)} \subset \ldots \subset \mathcal{Q}^{(0)} \tag{72}$$

and, in particular,

$$\mathrm{card}(\mathcal{Q}^{(t-1)}) < \mathrm{card}(\mathcal{Q}^{(t-2)}) < \ldots < \mathrm{card}(\mathcal{Q}^{(0)}). \tag{73}$$

---

[13]In particular, $\mathbf{q}^\star$ should not be confused with the minimizer of problem $(7 - \mathcal{P}_\lambda^{\mathrm{sq}})$.

Since $\mathrm{card}(\mathcal{Q}^{(t-1)})$ must be nonnegative and $\mathrm{card}(\mathcal{Q}^{(0)}) \leq q$, we thus have necessarily from (73) that $t \leq q$. This proves the convergence of Algorithm 4 in at most q steps.

In practice, Algorithm 4 is implemented as follows. Vector $\mathbf{q}$ is first sorted by absolute value and its cumulative sum is precomputed. This requires a complexity of $\mathcal{O}(q \log q)$. Then, each of the (at most) q repetitions of lines 5-6-7 scales as $\mathcal{O}(\log(\mathrm{card}(\mathcal{Q}^{(t)})))$ since the complexity is concentrated in the identification of the components of $\mathcal{Q}^{(t)}$ verifying "$\alpha|\mathbf{q}(i)| \geq \widetilde{w}^{(t+1)}$". Inasmuch as $\mathrm{card}(\mathcal{Q}^{(t)}) \leq q$, the overall complexity of the proposed procedure is therefore $\mathcal{O}(q \log q)$.

B. Optimality conditions for (70)

In this section, we derive necessary and sufficient optimality conditions for problem (70). These conditions will be exploited in the next section to show the correctness of Algorithm 4.

We first note that

$$\mathrm{sign}(\mathbf{q}^\star) = \mathrm{sign}(\mathbf{q}). \tag{74}$$

Indeed, $\mathbf{q}^\star$ is also the minimizer of

$$\min_{\mathbf{q}'} \|\mathbf{q}' - \mathbf{q}\|_2^2 \quad \text{subject to} \quad \pm\alpha\mathbf{q}' \leq \widetilde{w}^\star, \tag{75}$$

which admits the following simple analytical expression:

$$\forall i: \quad \mathbf{q}^\star(i) = \mathrm{sign}(\mathbf{q}(i))\min(|\mathbf{q}(i)|, \widetilde{w}^\star). \tag{76}$$

This shows (74).

In the rest of the appendix, we will thus focus, without loss of generality, on the case where

$$\mathbf{q}(i) \geq 0 \quad \forall i. \tag{77}$$

We note that this setup can always be obtained by some simple changes of variables. In this case, we have from (74) that $\mathbf{q}^\star \geq \mathbf{0}_q$ and problem (70) can equivalently be expressed as

$$\min_{\mathbf{q}'\geq\mathbf{0}_q, \widetilde{w}'\geq 0} (\widetilde{w}' - \widetilde{w})^2 + \|\mathbf{q}' - \mathbf{q}\|_2^2 \ \text{s.t.} \ \alpha\mathbf{q}' \leq \widetilde{w}'. \tag{78}$$

Since (78) is convex with linear inequality constraints, any minimizer $(\widetilde{w}^\star, \mathbf{q}^\star)$ must satisfy the following set of necessary and sufficient optimality conditions [25, Propositions 5.2.1 and 5.1.5]:

$$\alpha\mathbf{q}^\star \leq \widetilde{w}^\star, \tag{79}$$

and $\exists \mathbf{v}^\star \geq \mathbf{0}_q$ such that

$$\forall i: \quad 0 = \mathbf{v}^\star(i)(\alpha\mathbf{q}^\star(i) - \widetilde{w}^\star), \tag{80}$$

$$(\widetilde{w}^\star, \mathbf{q}^\star) = \arg\min_{\widetilde{w}'\geq 0, \mathbf{q}'\geq\mathbf{0}_q} \Big\{ (\widetilde{w}' - \widetilde{w})^2 + \|\mathbf{q}' - \mathbf{q}\|_2^2 \tag{81}$$
$$+ \sum_{i=1}^{q} \mathbf{v}^\star(i)(\alpha\mathbf{q}'(i) - \widetilde{w}') \Big\}.$$

In the rest of this section, we derive other optimality conditions, equivalent to (79)-(81), which are more amenable to prove the correctness of Algorithm 4.

We first note that (81) is a separable nonnegative least-square problem and its minimizer admits the following closed-form expression:

$$\widetilde{w}^{\star} = \max\left(0, \widetilde{w} + \sum_{i=1}^{q} \mathbf{v}^{\star}(i)\right) \tag{82}$$

$$\mathbf{q}^{\star}(i) = \max\left(0, \mathbf{q}(i) - \alpha \mathbf{v}^{\star}(i)\right). \tag{83}$$

We then distinguish between the cases "$\widetilde{w}^{\star} = 0$" and "$\widetilde{w}^{\star} > 0$".

From (79), we see that the case "$\widetilde{w}^{\star} = 0$" may occur only if $\mathbf{q}^{\star} = \mathbf{0}_q$. Using (82)-(83), this leads to

$$(\widetilde{w}^{\star}, \mathbf{q}^{\star}) = (0, \mathbf{0}_q) \iff \exists \mathbf{v}^{\star} \geq \mathbf{0}_q \text{ s.t.} \begin{cases} \widetilde{w} + \sum_{i=1}^{q} \mathbf{v}^{\star}(i) \leq 0, \\ \alpha^{-1}\mathbf{q}(i) \leq \mathbf{v}^{\star}(i), \end{cases}$$

or more compactly,

$$(\widetilde{w}^{\star}, \mathbf{q}^{\star}) = (0, \mathbf{0}_q) \iff \widetilde{w} + \alpha^{-1}\sum_{i=1}^{q} \mathbf{q}(i) \leq 0. \tag{84}$$

We next consider the case "$\widetilde{w}^{\star} > 0$". First, this assumption together with (82) leads to

$$\widetilde{w}^{\star} = \widetilde{w} + \sum_{i=1}^{q} \mathbf{v}^{\star}(i). \tag{85}$$

Moreover, from (80) we have that

$$\mathbf{v}^{\star}(i) = 0 \quad \forall i \notin \mathcal{Q}^{\star}, \tag{86}$$

where

$$\mathcal{Q}^{\star} \triangleq \{i \colon \alpha\mathbf{q}^{\star}(i) = \widetilde{w}^{\star}\}. \tag{87}$$

(85) thus further simplifies to

$$\widetilde{w}^{\star} = \widetilde{w} + \sum_{i \in \mathcal{Q}^{\star}} \mathbf{v}^{\star}(i). \tag{88}$$

On the other hand, we have (by definition of $\mathcal{Q}^{\star}$) that

$$\mathbf{q}^{\star}(i) = \alpha^{-1}\widetilde{w}^{\star} \quad \forall i \in \mathcal{Q}^{\star} \tag{89}$$

so that, invoking again our hypothesis "$\widetilde{w}^{\star} > 0$", we obtain from (83)

$$\mathbf{q}^{\star}(i) = \mathbf{q}(i) - \alpha\mathbf{v}^{\star}(i) \quad \forall i \in \mathcal{Q}^{\star}. \tag{90}$$

Finally, combining (88), (89) and (90) leads to

$$\widetilde{w}^{\star} = \frac{\alpha^2}{\alpha^2 + \operatorname{card}(\mathcal{Q}^{\star})}\left(\widetilde{w} + \alpha^{-1}\sum_{i \in \mathcal{Q}^{\star}} \mathbf{q}(i)\right). \tag{91}$$

As a consequence $(\widetilde{w}^{\star}, \mathbf{q}^{\star})$ with $\widetilde{w}^{\star} > 0$ is a solution of (78) if and only if $\exists \mathbf{v}^{\star} \geq \mathbf{0}_q$ such that (88)-(91) are satisfied.

C. Optimality of $(\widetilde{w}_{\mathrm{proj}}, \mathbf{q}_{\mathrm{proj}})$

We finally prove the correctness of Algorithm 4. We assume that the algorithm terminates its main recursion (steps 4-8) after t iterations (that is $\mathcal{Q}^{(t)} = \mathcal{Q}^{(t-1)}$) and show that the assignments defined in steps 9-14 are such that $(\widetilde{w}_{\mathrm{proj}}, \mathbf{q}_{\mathrm{proj}}) = (\widetilde{w}^{\star}, \mathbf{q}^{\star})$.

First, if $\widetilde{w}^{(t)} \leq 0$, we must have (by definition of $\mathcal{Q}^{(t)}$) that

$$\mathcal{Q}^{(t)} = \{1, \ldots, q\} \tag{92}$$

since $\mathbf{q} \geq \mathbf{0}_q$. Combining this equality with the definition of $\widetilde{w}^{(t)}$ in line 6 of Algorithm 4 leads to

$$\widetilde{w}^{(t)} = \frac{\alpha^2}{\alpha^2 + q}\left(\widetilde{w} + \alpha^{-1}\sum_{i=1}^{q} \mathbf{q}(i)\right). \tag{93}$$

Since $\widetilde{w}^{(t)} \leq 0$, the last equality implies

$$\widetilde{w} + \alpha^{-1}\sum_{i=1}^{q} \mathbf{q}(i) \leq 0. \tag{94}$$

In view of (84), we finally have

$$\widetilde{w}^{(t)} \leq 0 \Rightarrow (\widetilde{w}^{\star}, \mathbf{q}^{\star}) = (0, \mathbf{0}_q). \tag{95}$$

This corresponds to the assignment in line 9-10 of Algorithm 4.

If $\widetilde{w}^{(t)} > 0$, it can easily be verified that the optimality conditions (88)-(91) are verified for

$$\mathbf{q}^{\star} = \mathbf{q}_{\mathrm{proj}}$$
$$\widetilde{w}^{\star} = \widetilde{w}_{\mathrm{proj}}$$
$$\mathbf{v}^{\star}(i) = \begin{cases} 0 & \text{if } \alpha\mathbf{q}_{\mathrm{proj}}(i) < \widetilde{w}_{\mathrm{proj}} \\ \alpha^{-2}(\alpha\mathbf{q}(i) - \widetilde{w}_{\mathrm{proj}}) & \text{otherwise,} \end{cases}$$

where $\widetilde{w}_{\mathrm{proj}}$ and $\mathbf{q}_{\mathrm{proj}}$ are defined in lines 12-13 of Algorithm 4. This concludes the proof.

REFERENCES

[1] R. T. Rockafellar, "A dual approach to solving nonlinear programming problems by unconstrained optimization," *Mathematical Programming*, vol. 5, no. 1, pp. 354–373, Dec. 1973.

[2] R. E. Bruck, "An iterative solution of a variational inequality for certain monotone operators in hilbert space," *Bull. Amer. Math. Soc.*, vol. 81, no. 5, pp. 890–892, Sep. 1975.

[3] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17 – 40, 1976.

[4] L. El Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination in sparse supervised learning," EECS Dept., University of California at Berkeley, Tech. Rep. UC/EECS-2010-126, Sep. 2010.

[5] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013.

[6] S. Rosset, J. Zhu, and T. Hastie, "Boosting as a regularized path to a maximum margin classifier," *Journal of Machine Learning Research*, vol. 5, no. Aug, pp. 941–973, 2004.

[7] O. Fercoq, A. Gramfort, and J. Salmon, "Mind the duality gap: safer rules for the Lasso," in *Proc. Int. Conf. Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 37, Lille, France, Jul. 2015, pp. 333–342.

[8] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval, "Dynamic Screening: Accelerating First-Order Algorithms for the

Lasso and Group-Lasso," *IEEE Trans. Signal Process.*, vol. 63, no. 19, p. 20, 2015.

[9] A. Malti and C. Herzet, "Safe screening tests for Lasso based on firmly non-expansiveness," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proces. (ICASSP)*, Mar. 2016, pp. 4732–4736.

[10] C. Herzet, C. Dorffer, and A. Drémeau, "Gather and conquer: Region-based strategies to accelerate safe screening tests," *IEEE Trans. Signal Process.*, vol. 67, no. 12, pp. 3300–3315, Jun. 2019.

[11] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon, "Gap safe screening rules for sparse multi-task and multi-class models," in *Proc. Int. Conf. on Neural Information Processing Systems (NIPS)*. MIT Press, 2015, pp. 811–819.

[12] ——, "Gap safe screening rules for sparsity enforcing penalties," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 4671–4703, Jan. 2017.

[13] G. G. A. Rakotomamonjy and J. Salmon, "Screening rules for Lasso with non-convex sparse regularizers," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019.

[14] Z. Cvetković, "Resilience properties of redundant expansions under additive noise and quantization," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 644–656, Mar. 2003.

[15] A. R. Calderbank and I. Daubechies, "The pros and cons of democracy," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1721–1725, Jun. 2002.

[16] B. Farrell and P. Jung, "A Kashin approach to the capacity of the discrete amplitude constrained Gaussian channel," in *Proc. Int. Conf. Sampling Theory and Applications (SAMPTA)*, Marseille, France, May 2009.

[17] J. Ilic and T. Strohmer, "Papr reductioni in ofdm using kashin's representation," in *IEEE Workshop on Signal Processing Advances in Wireless Communications*, Jun. 2009, pp. 444–448.

[18] C. Studer and E. G. Larsson, "PAR-aware large-scale multi-user MIMO-OFDM downlink," *IEEE J. Sel. Areas Comm.*, vol. 31, no. 2, pp. 303–313, Feb. 2013.

[19] H. Jegou, T. Furon, and J.-J. Fuchs, "Anti-sparse coding for approximate nearest neighbor search," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proces. (ICASSP)*, 2012, pp. 2029–2032.

[20] M. Vural, P. Jung, and S. Stańczak, "A new outlier detection method based on anti-sparse representations," in *Signal Process. and Comm. Appl. Conf. (SIU)*, May 2017, pp. 1–4.

[21] X. Jiang, J. Chen, H. C. So, and X. Liu, "Large-scale robust beamforming via ℓ_∞ -minimization," *IEEE Trans. Signal Process.*, vol. 66, no. 14, pp. 3824–3837, Jul. 2018.

[22] L. W. Neustadt, "Minimum effort control systems," *J. SIAM Control*, vol. 1, no. 1, pp. 16–31, 1962.

[23] J. A. Cadzow, "Algorithm for the minimum-effort problem," *IEEE Trans. on Autom. Control*, vol. 16, no. 1, pp. 60–63, 1971.

[24] J.-J. Fuchs, "Spread representations," in *asilomar*, 2011.

[25] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, 1999.

[26] Z. J. Xiang and P. J. Ramadge, "Fast Lasso screening tests based on correlations," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proces. (ICASSP)*, Mar. 2012, pp. 2137–2140.

[27] Z. J. Xiang, Y. Wang, and P. J. Ramadge, "Screening tests for Lasso problems," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 39, no. 5, pp. 1008–1027, May 2017.

[28] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Research Logistics Quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.

[29] E. D. Dolan and J. J. Moré, "Benchmarking optimization software with performance profiles," *Mathematical Programming*, vol. 91, no. 2, pp. 201–213, Jan. 2002.

[30] H. H. Bauschke and P. L. Combettes, *Subdifferentiability*. New York, NY: Springer New York, 2011, pp. 223–240.