OJMO

Open Journal of Mathematical Optimization

Zhenan Fan, Huang Fang & Michael P. Friedlander

Cardinality-constrained structured data-fitting problems

Volume 5 (2024), article no. 2 (21 pages)

https://doi.org/10.5802/ojmo.27

Article submitted on June 22, 2021, revised on July 19, 2022,
accepted on July 13, 2023.

Open Journal of Mathematical Optimization is a member of the
Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/
e-ISSN: 2777-5860

CENTRE
MERSENNE

Cardinality-constrained structured data-fitting problems

Zhenan Fan
The University of British Columbia, Canada
zhenanf@cs.ubc.ca

Huang Fang
The University of British Columbia, Canada
hgfang@cs.ubc.ca

Michael P. Friedlander
The University of British Columbia, Canada
michael.friedlander@ubc.ca

Abstract

A memory-efficient solution framework is proposed for the cardinality-constrained structured data-fitting problem. Dual-based atom-identification rules reveal the structure of the optimal primal solution from near-optimal dual solutions, which allows for a simple and computationally efficient algorithm that translates any feasible dual solution into a primal solution satisfying the cardinality constraint. Rigorous guarantees bound the quality of a near-optimal primal solution given any dual-based method that generates dual iterates converging to an optimal dual solution. Numerical experiments on real-world datasets support the analysis and demonstrate the efficiency of the proposed approach.

Digital Object Identifier 10.5802/ojmo.27

Keywords convex analysis, sparse optimization, low-rank optimization, primal-retrieval.

Acknowledgments Zhenan Fan and Huang Fang contributed equally to this paper.

1 Introduction

Consider the problem of fitting a model to data by building up model parameters as the superposition of a small set atomic elements taken from a given dictionary. Versions of this cardinality-constrained problem appear in a number of statistical-learning applications in machine learning [5, 49, 57, 62], data mining, and signal processing [15]. In these applications, common atomic dictionaries include the set of one-hot vectors or matrices (i.e., vectors and matrices that contain only a single nonzero element) and rank-1 matrices. The elements chosen from the given dictionary encode a notion of parsimony in the definition of the model parameters.

The cardinality-constrained formulation that we consider aims to

$$\text{find } x \in \mathcal{X} \text{ such that } f(b - Mx) \le \alpha \text{ and } \mathbf{card}_{\mathcal{A}}(x) \le k, \tag{P}$$

where $f : \mathbb{R}^m \to \mathbb{R}$ is an L-smooth and convex function, $M : \mathcal{X} \to \mathbb{R}^m$ is a linear operator, $b \in \mathbb{R}^m$ is the observation vector, and $\mathcal{A} \subseteq \mathcal{X}$ is the atomic dictionary. The cardinality function

$$\mathbf{card}_{\mathcal{A}}(x) := \inf \left\{ \mathbf{nnz}(c) \;\middle|\; x = \sum_{a \in \mathcal{A}} c_a a, \;\; c_a \ge 0 \right\} \tag{1}$$

measures the complexity of x with respect to the dictionary $\mathcal{A}$. When $\mathcal{A} = \{\pm e_1, \pm e_2, \ldots, \pm e_n\}$, for example, is the set of signed canonical unit vectors, the function $\mathbf{card}_{\mathcal{A}}(x)$ simply counts the number of nonzero elements in x. The loss term $f(b - Mx)$ measures the quality of the fit. Typically $k \ll n$, which indicates that we seek a feasible model parameter x with an efficient representation in terms of k atoms from the dictionary $\mathcal{A}$.

For the application areas that we target, the two characteristics of this feasibility problem that pose the biggest challenge to efficient implementation are the combinatorial nature of the cardinality constraint and the high-dimensionality of the parameter space. To address the combinatorial challenge, we follow van den Berg and Friedlander [9, 10] and Chandrasekaran et al. [21], and use the convex gauge function

$$\gamma_{\mathcal{A}}(x) = \inf \left\{ \sum_{a\in\mathcal{A}} c_a \;\middle|\; x = \sum_{a\in\mathcal{A}} c_a a,\ c_a \geq 0 \right\} \tag{2}$$

as a tractable proxy for the cardinality function (1); see Section 3. In tandem with the convexity of the loss function, the gauge function allows us to formulate three alternative relaxed convex optimization problems that, under certain conditions, have approximate solutions satisfying the feasibility problem; see problems (P_1), (P_2), and (P_3) in Section 4.

The high-dimensionality of the parameter space may imply, however, that it is inefficient (and maybe even practically impossible) to solve these convex relaxations because it may be infeasible to directly store the approximations to a feasible solution x. Instead, we wish to develop methods that leverage the efficient representation that low-cardinality solutions have in terms of the atoms in the dictionary $\mathcal{A}$. For example, consider the case in which the dictionary is the set of symmetric $n \times n$ rank-one matrices, and M is the trace linear operator that maps these matrices into m-vectors. Any method that iterates directly on the parameters x requires $\mathcal{O}(n^2 + m)$ storage for the iterates and the data. An alternative is the widely-used conditional gradient method [31], which requires $\mathcal{O}(nt + m)$ storage after t iterations [41], but also often requires a substantial number of iterations t to converge. Instead of storing x directly, however, our approached is based on applying a dual method to one of the convex relaxations (P_1), (P_2), and (P_3) (defined below); first-order dual methods typically require only $\mathcal{O}(m)$ storage, and still allow us to collect information on which atoms in $\mathcal{A}$ participate in the construction of a feasible x. One of the aims of this paper is to describe how to collect and use this information.

1.1 Approach

We propose a unified algorithm-agnostic strategy that uses any sequence of improving dual solutions to one of the convex relaxations. This dual sequence identifies an essential subset of atoms in $\mathcal{A}$ needed to construct an ϵ-infeasible solution x that satisfies the conditions

$$f(b - Mx) \leq \alpha + \epsilon \quad \text{and} \quad \mathbf{card}_{\mathcal{A}}(x) \leq k$$

for any positive tolerance ϵ. These *atomic-identification* rules, described in Section 5, derive from the polar-alignment property and apply to arbitrary dictionaries $\mathcal{A}$ [28]. These atom-identification rules generalize earlier approaches described by El Ghaoui [26] and Hare and Lewis [36]. Once an essential subset of k atoms is identified, an ϵ-feasible solution x can be computed by optimizing over all positive linear combinations of this subset. This relatively small k-variable problem can often be solved efficiently.

We prove that when the atomic dictionary is polyhedral, we can set ϵ to zero and still identify in polynomial time a set of feasible atoms; see Corollary 9. When the atomic dictionary is spectrahedral, we prove that an ϵ-feasible set of atoms can be identified also in polynomial time; see Corollary 15.

We demonstrate via numerical experiments on real-world datasets that this approach is effective in practice.

There are three important elements in our primal-retrieval algorithm. The first element is an atom-identifier function $\mathsf{EssCone}_{\mathcal{A},k}$ that maps the product M^*y, where y is any feasible dual variable, to a cone generated by k atoms that are *essential.* These atoms have the property that

$$\mathsf{EssCone}_{\mathcal{A},k}(M^*y) \subseteq \{x \mid \mathbf{card}_{\mathcal{A}}(x) \leq k\}.$$

The explicit definition of the essential cone depends on the particular dictionary $\mathcal{A}$. In Section 6, we make it explicit for dictionaries that are discrete or polyhedral (Section 6.1) and spectral (Section 6.2).

The second element is an arbitrary function $\texttt{oracle}_{f,\mathcal{A},M,b}$ (such as an appropriate first-order iterative method) that generates dual iterates $y^{(t)}$ converging to the optimal dual variable y^* of any of the dual problem (D_i). It is this oracle that generates the dual estimates subsequently used by $\mathsf{EssCone}_{\mathcal{A},k}$.

The third algorithmic component is the reduced convex optimization problem

$$x^{(t)} \in \operatorname*{arg\,min}_x \left\{ f(b - Mx) \;\middle|\; x \in \mathsf{EssCone}_{\mathcal{A},k}(M^*y^{(t)}) \right\}, \tag{PR}$$

which at each iteration constructs a primal estimate $x^{(t)}$ using the atoms identified through the dual estimate $y^{(t)}$. The detailed algorithm is shown in Algorithm 1. Note that our primal-retrieval strategy does not aim to recover the optimal solutions to (P_1), (P_2), or (P_3), which only serve as guidance for our atom-identification rule. The final output of Algorithm 1 may be different from the solution of these optimization problems.

Algorithm 1: primal-retrieval algorithm

```
Input: data-fitting tolerance $\alpha$, cardinality constraint $k$, dictionary $\mathcal{A}$, loss function $f$, linear operator
        $M$, observation $b$, and tolerance $\epsilon \geq 0$
Initialize dual feasible vector $y^{(0)}$
for $t = 1, 2, \ldots$ do
    $y^{(t)} \leftarrow \texttt{oracle}_{f,\mathcal{A},M,b}(y^{(t-1)})$
    $x^{(t)} \leftarrow$ solution to (PR)
    if $f(b - Mx^{(t)}) \leq \alpha + \epsilon$ then
        break
Return: $x^{(t)}$
```

2 Related work

Many recent approaches for atomic-sparse optimization problems are based on algorithms [23,29]. These methods, however, still need to retrieve at some point a primal solution x, which may require a prohibitive amount of memory for its storage. For example, when the constraint in (P) is a rank constraint, a widely used heuristic applies the truncated singular value decomposition to obtain low-rank solutions, but this heuristic is unreliable in minimizing the model misfit [28, Algorithm 6.4]. Memory-efficient atomic-sparse optimization thus requires efficient and reliable methods to retrieve an atomic-sparse primal solution.

Dual approaches for nuclear- or trace-norm regularized problems are attractive because they enjoy optimal storage, which means that they have space complexity $\mathcal{O}(m)$ instead of $\mathcal{O}(n^2)$ [23,32]. For example, the bundle method for solving the Lagrangian dual formulation of semi-definite programming [37], and the gauge dual formulation of general atomic sparse optimization problem [29], exhibit promising results in practice. Similarly, there are dual approaches for one-norm regularized problems that enjoy better convergence rates than primal approaches [2,43].

A related line of research uses memory-efficient primal-based algorithms based on hard-thresholding. Some examples include gradient hard-thresholding [63], periodical hard-thresholding [1], and many proximal-gradient or ADMM-based hard-thresholding algorithms [40,46,48]. These approaches are primal-based and tangential to our purposes. We do not include them in our discussion.

The theoretical analysis of our primal-retrieval approach is related to optimal atom identification [14,35,36], and especially to recently developed safe-screening rules for various sparse optimization problems [6,7,13,26, 39,45,47,51,54,59–61]. One of our main results, given by Theorem 5, generalizes the gap safe-screening rule developed by Ndiaye et al. [50] to general atomic-sparse problems and to more general problem formulations. Some of the techniques used in our analysis are related to the facial-reduction strategy advocated by Krislock and Wolcowicz [44].

3 Preliminaries

We introduce in this section the main tools of convex analysis used to understand atomic sparsity.

The gauge function (2) is always convex, nonnegative, positively homogeneous, and finite only at points contained within the cone

$$\operatorname{cone}(\mathcal{A}) := \left\{ x = \sum_{a\in\mathcal{A}} c_a a \;\middle|\; c_a \geq 0 \right\}$$

generated from the elements of the set $\mathcal{A}$. The gauge is not necessarily a norm because it may not be symmetric (unless $\mathcal{A}$ is centrosymmetric), may vanish at points other than the origin, and may not be finite valued (unless $\mathcal{A}$ contains the origin in the interior of its convex hull). Throughout, we make the blanket assumption that the dictionary $\mathcal{A} \subseteq \mathcal{X}$ is compact, and that the origin is contained in its convex hull. The assumption on the origin ensures that the gauge function is continuous. The compactness assumption isn't strictly necessary for many of our conclusions, but does considerably simplify the analysis. The set $\mathcal{A}$ may be nonconvex, which is the case, for example, if it consists of a discrete set of two or more items.

Table 1 Commonly used atomic sets and the corresponding gauge and support functions [29]. The indicator function $\delta_{\mathcal{C}}(x)$ is zero if x is in the set $\mathcal{C}$ and $+\infty$ otherwise. The commonly used group-norm is also an atomic norm [28, Example 4.7]. The functions $\|X\|_*$ and $\|X\|_2$, respectively, correspond to the nuclear norm (sum of singular values) and spectral norm (maximum singular value) of a matrix X.

Atomic sparsity	$\mathcal{A}$	$\gamma_{\mathcal{A}}(x)$	$\mathcal{S}_{\mathcal{A}}(x)$	$\sigma_{\mathcal{A}}(z)$
non-negative	$\mathrm{cone}(\{\boldsymbol{e}_1,\dots,\boldsymbol{e}_n\})$	$\delta_{\geq 0}$	$\mathrm{cone}(\{\boldsymbol{e}_i \mid x_i > 0\})$	$\delta_{\leq 0}$
element-wise	$\{\pm\boldsymbol{e}_1,\dots,\pm\boldsymbol{e}_n\}$	$\|\cdot\|_1$	$\{\mathrm{sign}(x_i)\boldsymbol{e}_i \mid x_i \neq 0\}$	$\|\cdot\|_\infty$
element-wise & non-negative	$\{\boldsymbol{e}_1,\dots,\boldsymbol{e}_n\}$	$\sum_j(\,\cdot\,)_j + \delta_{\geq 0}$	$\{\boldsymbol{e}_i \mid x_i > 0\}$	$\|\cdot\|_\infty$
low rank	$\{uv^T \mid \|u\|_2 = \|v\|_2 = 1\}$	$\|\cdot\|_*$	singular vectors of x	$\|\cdot\|_2$
PSD & low rank	$\{uu^T \mid \|u\|_2 = 1\}$	$\mathrm{tr} + \delta_{\succeq 0}$	eigenvectors of x	$\max\{\lambda_{\max}, 0\}$

The definition of the gauge function makes explicit this function's role as a convex penalty for atomic sparsity. The *atomic support* of a vector x to be the collection of atoms $a \in \mathcal{A}$ that contribute positively to the conic decomposition implied by the value $\gamma_{\mathcal{A}}(x)$ [28, Definition 2.1].

▶ **Definition 1** (Atomic support). A subset of atoms $\mathcal{S}_{\mathcal{A}}(x) \subset \mathcal{A}$ is a *support set* for x with respect to $\mathcal{A}$ if $\mathcal{S}_{\mathcal{A}}(x)$ satisfies

$$\gamma_{\mathcal{A}}(x) = \sum_{a \in \mathcal{S}_{\mathcal{A}}(x)} c_a, \qquad x = \sum_{a \in \mathcal{S}_{\mathcal{A}}(x)} c_a a, \text{ and } c_a > 0 \quad \forall\, a \in \mathcal{S}_{\mathcal{A}}(x).$$

For example, the support set $\mathcal{S}_{\mathcal{A}}(x)$ for the atomic set of 1-hot unit vectors $\mathcal{A} = \{e_i \mid i = 1, 2, \dots, n\}$ coincides with the nonzero elements of x with positive entries. The support function to the set $\mathcal{A}$ is given by $\sigma_{\mathcal{A}}(z) := \sup_{a \in \mathcal{A}} \langle a, z \rangle$. Because $\mathcal{A}$ is compact, every direction $z \in \mathcal{X}$ generates a supporting hyperplane to the convex hull of $\mathcal{A}$. The atoms contained in that supporting hyperplane are said the be *exposed* by the direction z. The following definition also includes the notion of atoms that are approximately exposed.

▶ **Definition 2** (Exposed and ϵ-exposed atoms). The exposed atoms and ϵ-exposed atoms, respectively, of a set $\mathcal{A} \subseteq \mathcal{X}$ in the direction $z \in \mathcal{X}$ are defined by the sets

$$\mathcal{E}_{\mathcal{A}}(z) := \{a \in \mathcal{A} \mid \langle a, z \rangle = \sigma_{\mathcal{A}}(z)\} \text{ and } \mathcal{E}_{\mathcal{A}}(z, \epsilon) := \{a \in \mathcal{A} \mid \langle a, z \rangle \geq \sigma_{\mathcal{A}}(z) - \epsilon\},$$

where $\sigma_{\mathcal{A}}(z) := \sup_{a \in \mathcal{A}} \langle a, z \rangle$ is the support function with respect to $\mathcal{A}$.

When $\epsilon = 0$, the ϵ-exposed atoms coincide with the exposed atoms.

We list in Table 1 commonly used atomic sets, their corresponding gauge and support functions, and atomic supports.

4 Atomic-sparse optimization

We introduce in this section convex relaxations to the structured data-fitting problem (P). In particular, we consider the following three related gauge-regularized optimization problems:

$$\underset{x \in \mathcal{X}}{\text{minimize}} \;\; p_1(x) := f(b - Mx) + \lambda\gamma_{\mathcal{A}}(x), \tag{P$_1$}$$

$$\underset{x \in \mathcal{X}}{\text{minimize}} \;\; p_2(x) := f(b - Mx) \;\text{ subject to }\; \gamma_{\mathcal{A}}(x) \leq \tau, \tag{P$_2$}$$

$$\underset{x \in \mathcal{X}}{\text{minimize}} \;\; p_3(x) := \gamma_{\mathcal{A}}(x) \;\text{ subject to }\; f(b - Mx) \leq \alpha. \tag{P$_3$}$$

It is well known that under mild conditions, these three formulations are equivalent for appropriate choices of the positive parameters λ, τ, and α [33]. Practitioners often prefer one of these formulations depending on their application. For example, tasks related to machine learning, including feature selection and recommender systems, typically feature one of the first two formulations [49, 57, 62]. On the other hand, applications in signal processing and related fields, such as compressed sensing and phase retrieval, often use the third formulation [9, 15].

Our primal-retrieval strategy relies on the hypothesis that the atomic-sparse optimization problems (P$_1$), (P$_2$) and (P$_3$) are reasonable convex relaxations to the structured data-fitting problem (P), in the sense that the corresponding optimal solutions are feasible for (P). We formalize this in the following assumption.

▶ **Assumption 3.** *Let x_i^* denote an optimal solution to (P_i), $i = 1, 2, 3$. Then x_i^* is feasible for* (P), *i.e., $f(b - Mx_i^*) \le \alpha$ and there exists a support $\mathcal{S}_{\mathcal{A}}(x_i^*)$ with size less than k.*

Assumption 3 does not hold in general. However, there are many important applications in practice where Assumption 3 does hold, including sparse signal recovery [19, 24, 25], low-rank matrix recovery [17, 18], structured signal decompostion [16, 27], and general structured signal reconstruction [21, 58].

As described in Section 1, the Fenchel–Rockafellar duals for these problems have typically smaller space complexity. These dual problems can be formulated as

$$\operatorname*{minimize}_{y\in\mathbb{R}^m} \ d_1(y) := f^*(y) - \langle b, y\rangle \ \text{ subject to } \ \sigma_{\mathcal{A}}(M^*y) \le \lambda, \tag{D$_1$}$$

$$\operatorname*{minimize}_{y\in\mathbb{R}^m} \ d_2(y) := f^*(y) - \langle b, y\rangle + \tau\sigma_{\mathcal{A}}(M^*y), \tag{D$_2$}$$

$$\operatorname*{minimize}_{y\in\mathbb{R}^m} \ \inf_{\beta>0} d_3(y,\beta) := \beta\left(f^*\left(y/\beta\right) + \alpha\right) - \langle b, y\rangle \ \text{ subject to } \ \sigma_{\mathcal{A}}(M^*y) \le 1, \tag{D$_3$}$$

where $f^*(y) = \sup_w \langle y, w\rangle - f(w)$ is the convex conjugate function of f, and $M^* : \mathbb{R}^m \to \mathbb{R}^n$ is the adjoint operator of M, which satisfies $\langle Mx, y\rangle = \langle x, M^*y\rangle$ for all $x \in \mathcal{X}$ and $y \in \mathbb{R}^m$. The derivation of these dual problems can be found in Appendix A.

5 Atom identification

We demonstrate in this section how an optimal dual solution can be used to identify essential atoms that form the support of a primal solution. In order to develop atomic-identification rules that apply to arbitrary atomic sets $\mathcal{A} \subseteq \mathcal{X}$ (even those that are uncountably infinite) we require generalized notions of active constraint sets. In linear programming, for example, the simplex multipliers give information about the optimal primal support. By analogy, our atomic-identification rules give information about the essential atoms that participate in the support of the primal optimal solutions. In addition, we extend the identification rules to approximate the essential atoms from approximate dual solutions.

We build on the following result, due to Fan et al. [28, Proposition 4.5 and Theorem 5.1].

▶ **Theorem 4** (Atom identification). *Let x^* and y^* be optimal primal-dual solutions for problems (P_i) and (D_i), with $i = 1, 2, 3$. Then*

$$\mathcal{S}_{\mathcal{A}}(x^*) \subseteq \mathcal{E}_{\mathcal{A}}(M^*y^*). \tag{3}$$

The following theorem generalizes this result to show similar atomic support identification properties that also apply to approximate dual solutions. In particular, given a feasible dual variable y close to y^*, the support of x^* is contained in the set of ϵ-exposed atoms that includes $\mathcal{E}_{\mathcal{A}}(M^*y^*)$.

▶ **Theorem 5** (Generalized atom identification). *Let x_i and y_i be feasible primal and dual vectors, respectively for problems (P_i) and (D_i), with $i = 1, 2, 3$. Then*

$$\mathcal{S}_{\mathcal{A}}(x_i^*) \subseteq \mathcal{E}_{\mathcal{A}}(M^*y_i, \epsilon_i), \tag{4}$$

where each ϵ_i is defined for problem i by

a. $\epsilon_1 = \ \|M\|_{\mathcal{A}}\sqrt{2L\left(d_1(y_1) - d_1(y_1^*)\right)}$,
b. $\epsilon_2 = 2\|M\|_{\mathcal{A}}\sqrt{2L\left(d_2(y_2) - d_2(y_2^*)\right)}$,
c. $\epsilon_3 = 2\|M\|_{\mathcal{A}}\sqrt{2\overline{\beta}L(\max\{d_3(y_3,\underline{\beta}), d_3(y_3,\overline{\beta})\} - d_3(y_3^*,\beta^*))}$,

where $\underline{\beta}$ and $\overline{\beta}$ are positive lower and upper bounds, respectively, for β^*, and $\|M\|_{\mathcal{A}} := \max_{a\in\mathcal{A}} \|Ma\|_2$ is the induced atomic operator norm.

Theorem 5 asserts that the underlying atomic support of x_i^* is contained in the set of the ϵ-exposed atoms of M^*y_i. Moreover, when $y_i \to y_i^*$ (and, for problem (D$_3$), the bounds $\underline{\beta} \to \beta^*$ and $\overline{\beta} \to \beta^*$), each $\epsilon_i \to 0$, and thus (4) implies that we have a tighter containment for the optimal atomic support. The proofs for parts a. and b. of Theorem 5 depend on the strong convexity of f^*, which is implied by the Lipschitz smoothness of f [38, Theorem 4.2.1]. This convenient property, however, is not available for part c. because the dual objective of (D$_3$) is the perspective map of $f^* + \alpha$, which is not strongly convex [3]. We resolve this technical difficulty by instead imposing the additional assumption that bounds are available on the dual optimal variable β^*.

Appendix C describes how to obtain these bounds during the course of the level-set method developed by Aravkin et al. [4].

The gap safe-screening rule developed by Ndiaye et al. [51] is a special case of Theorem 5 that applies only to (P_1) for the particular case in which $\gamma_{\mathcal{A}}$ is the one-norm.

6 Primal retrieval

Theorem 5 serves mainly as a technical tool for error bound analysis, in particular because it is impractical to compute or approximate ϵ_i. However, the inclusions (3) and (4), respectively, of Theorems 4 and 5 motivate us to define an atom-identifier function $\mathsf{EssCone}_{\mathcal{A},k}$ that depends on the dual variable y and satisfies the inclusions

$$\operatorname{cone}(\mathcal{E}_{\mathcal{A}}(M^*y)) \subseteq \mathsf{EssCone}_{\mathcal{A},k}(M^*y) \subseteq \operatorname{cone}(\mathcal{E}_{\mathcal{A}}(M^*y,\epsilon)).$$

The next two sections demonstrate how to construct such a function for polyhedral and spectral atomic sets, which are two important examples that appear frequently in practice. With this function we can thus implement the primal-recovery problem required by Step 5 of Algorithm 1. Moreover, we show how to use the error bounds of Theorem 5 to analyze the atomic-identification properties of the resulting algorithm.

6.1 Primal-retrieval for polyhedral atomic sets

We formalize in this section a definition for the function $\mathsf{EssCone}_{\mathcal{A},k}$ for the case in which $\mathcal{A}$ is a finite set of vectors, which implies that the convex hull is polyhedral. Given a feasible dual vector y, consider the top-k atoms in $\mathcal{A}$ with respect to the inner product with the vector M^*y:

$$\mathcal{A}_k := \{a_1,\dots,a_k\} \subseteq \mathcal{A} \text{ such that } \langle M^*y, a_i\rangle \geq \langle M^*y, a\rangle \quad \forall\, i \in [k] \text{ and } a \in \mathcal{A}\setminus\{a_1,\dots,a_k\}. \tag{5}$$

Note that there may be many sets of k atoms that satisfy this property. We then construct the cone of essential atoms as the convex conic hull generated from this set of top-k atoms:

$$\mathsf{EssCone}_{\mathcal{A},k}(M^*y) := \operatorname{cone}\mathcal{A}_k.$$

Thus, the primal-retrieval computation in Step 5 of Algorithm 1 is given by

$$\widehat{x} = \sum_{i=1}^{k} \widehat{c}_i a_i,$$

where

$$\widehat{c} \in \mathop{\arg\min}_{c\in\mathbb{R}^k_+} f_k(c), \text{with} f_k(c) := f\left(b - M\sum_{i=1}^{k} c_i a_i\right), \tag{6}$$

is the k-vector of coefficients obtained by minimizing the reduced objective over a k-dimensional polyhedron defined by the top-k atoms.

▶ Example 6 (Sparse vector recovery). Consider the problem of recovering a sparse vector $x^\natural$ from noisy observations $b := Mx^\natural + \eta$, where $M : \mathbb{R}^n \to \mathbb{R}^m$ is a given measurement matrix and $\eta \in \mathbb{R}^m$ is standard Gaussian noise. For some expected noise level $\alpha > 0$, the sparse recovery problem can be formulated as

$$\text{find } x \in \mathbb{R}^n \text{ such that } \|b - Mx\|_2 \leq \alpha \text{ and } \mathtt{nnz}(x) \leq k,$$

which corresponds to (P) with $f = \|\cdot\|_2$ and with the atomic set

$$\mathcal{A} = \{\pm e_1, \dots, \pm e_n\}, \tag{7}$$

where each e_i is the ith canonical unit vector. The basis pursuit denoising (BPDN) approach approximates this problem by replacing the cardinality constraint with an optimization problem that minimizes the 1-norm of the solution:

$$\mathop{\text{minimize}}_{x\in\mathbb{R}^n} \ \|x\|_1 \ \text{ subject to } \ \|Mx - b\|_2 \leq \alpha; \tag{8}$$

see Chen et al. [22]. This convex relaxation corresponds to problem (P_3).

There are many dual methods that generate iterates $y^{(t)}$ converging to the optimal dual solution to (8), including the level-set method coupled with the dual conditional-gradient suboracle, as described by [4,28]. The resulting primal-retrieval strategy for Step 5 of Algorithm 1 can thus be implemented by executing the following steps:

1. (Top-k atoms) Find the top k indices $\{i_1,\dots,i_k\}\subset[n]$ of the vector $M^*y^{(t)}$ with largest absolute value and gather their corresponding signs $s_i := \mathrm{sign}([M^*y^{(t)}]_i)$ for $i\in\{i_1,\dots,i_k\}$. The top-k atoms are thus $\mathcal{A}_k=\{s_{i_1}e_{i_1},\dots,s_{i_k}e_{i_k}\}$; see (5).
2. (Retrieve coefficients) Solve the reduced problem (6), where in this case,

   $$c^{(t)} \in \operatorname*{arg\,min}_{c\in\mathbb{R}^k_+} f_k(c), \text{where} f_k(c) = \|M[s_{i_1}e_{i_1}\dots s_{i_k}e_{i_k}]c-b\|_2.$$

   This is a nonnegative least-squares problem for which many standard algorithms are available. For example, an accelerated projected gradient descent method requires $\mathcal{O}(mk\log(1/\epsilon))$ iterations when the matrix $M[s_{i_1}e_{i_1}\dots s_{i_k}e_{ik}]$ has full column rank.
3. (Termination) Step 6 of the Algorithm 1 is implemented simply by verifying that $f_k(c^{(t)})\le\alpha$. (As verified by Corollary 9, we may take $\epsilon=0$ in this polyhedral case.) Thus, we can terminate the algorithm and return the primal variable

   $$x^{(t)} = [s_{i_1}e_{i_1}\dots s_{i_k}e_{ik}]c^{(t)},$$

   which is the superposition of the top-k atoms. Otherwise, the algorithm proceeds to the next iteration.

We describe numerical experiments for the sparse vector recovery problem in Section 7.1. ◀

6.1.1 Iteration complexity

In order to guarantee the quality of the recovered solution, we rely on a notion of degeneracy introduced by Nutini et al. [52].

▶ **Definition 7.** Let x^* and y^*, respectively, be optimal primal and dual solutions for problems (P_i) and (D_i), where $\mathcal{A}$ is polyhedral. Let δ be a positive scalar. The problem pair $((\mathrm{P}_i), (\mathrm{D}_i))$ is δ-nondegenerate if for any $a\in\mathcal{A}$, either $a\in\mathcal{S}_{\mathcal{A}}(x^*)$ or $\langle a, M^*y^*\rangle \le \sigma_{\mathcal{A}}(M^*y^*)-\delta$.

The next proposition guarantees a finite-time atom identification property when the atomic set is polyhedral.

▶ **Proposition 8** (Finite-time atom-identification)**.** *For each problem $i=1,2,3$, let $\{y_i^{(t)}\}_{t=1}^{\infty}$ be a sequence that converges to an optimal dual solution y_i^*. If the atomic set $\mathcal{A}$ is polyhedral and the problem pair ((P$_i$), (D$_i$)) is δ-nondegenerate, then there exists $T>0$ such that*

$$\mathsf{EssCone}_{\mathcal{A},k}(M^*y^{(t)}) \supseteq \mathcal{E}_{\mathcal{A}}(M^*y_i^*) \textit{ and } x^{(t)} \textit{ is feasible for } (\mathrm{P}) \quad \forall\, t>T.$$

It follows that Algorithm 1 will terminate in T iterations regardless of the tolerance ϵ.

Proposition 8 ensures that the atom-identification property described by Theorem 5 is guaranteed to discard superfluous atoms in a finite number of iterations as long as we have available an iterative solver that generates dual iterates converging to a solution. Thus, Algorithm 1 is guaranteed to generate a feasible solution to (P). The following corollary characterizes a bound on T in terms of the convergence rate of the dual method.

▶ **Corollary 9.** *For each problem $i=1,2,3$, suppose the dual oracle generates iterates $y^{(t)}$ converging to optimal variable y_i^* with rate*

$$d_i(y_i^{(t)}) - d_i(y_i^*) \in \mathcal{O}\left(t^{-p}\right)$$

for some $p>0$. If the atomic set $\mathcal{A}$ is polyhedral and the problem pair ((P$_i$), (D$_i$)) is δ-nondegenerate, then Algorithm 1 with $\epsilon=0$ terminates in $\mathcal{O}\left(\delta^{-2/p}\right)$ iterations.

6.1.2 Centrosymmetry and unconstrained primal recovery

Further computational savings are possible when the atomic set $\mathcal{A}$ is centrosymmetric, i.e.,

$$a \in \mathcal{A} \iff -a \in \mathcal{A}. \tag{9}$$

Centrosymmetry is a common property, and perhaps the prototypical example is the set of signed canonical unit vectors given by the set (7). Whenever centrosymmetry holds, $\operatorname{cone}\mathcal{A} = \operatorname{span}\mathcal{A}$. This motivates us to replace the function EssCone with the function

$$\mathsf{EssSpan}_{\mathcal{A},k}(M^*y) \coloneqq \operatorname{span}\mathcal{A}_k,$$

where $\mathcal{A}_k$ is the collection of top-k atoms defined by (5). Thus, the primal-retrieval optimization problem (6) reduces to the unconstrained version

$$\widehat{c} \in \operatorname*{arg\,min}_{c\in\mathbb{R}^k} f_k(c).$$

The following corollary simply asserts that the complexity results described in Section 6.1.1 continue to hold for centrosymmetric atomic sets when using the essential span function.

▶ **Corollary 10** (Atom identification under centrosymmetry). *If the atomic set $\mathcal{A}$ is centrosymmetric and polyhedral, then Proposition 8 and Corollary 9 hold with* EssCone *replaced by* EssSpan*.*

6.2 Primal-retrieval for spectral atomic sets

We formalize in this section a definition for the function $\mathsf{EssCone}_{\mathcal{A},k}$ for the case in which $\mathcal{A}$ is a collection of rank-1 matrices, either asymmetric or symmetric, respectively:

$$\mathcal{A} = \{uv^T \mid u \in \mathbb{R}^m,\ v \in \mathbb{R}^n,\ \|u\|_2 = \|v\|_2 = 1\}, \tag{10}$$
$$\mathcal{A} = \{vv^T \mid v \in \mathbb{R}^n,\ \|v\|_2 = 1\}. \tag{11}$$

We mainly focus on the former atomic set of asymmetric matrices because all of our theoretical results easily specialize to the symmetric case. Note that this atomic set is centrosymmetric (cf. (9)), and as we'll see below, the recovery problem is unconstrained. Later in Section 6.2.2 we'll describe the recovery problem for the atomic set of symmetric matrices, which is in fact a not centrosymmetric.

For this section only, we work with the linear operator $M : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$, and replace the vector of observations b with the p-by-q matrix B. In this context, the dual variables for one of the corresponding dual problems is a matrix of the same dimension.

Fix a feasible dual variable Y and define the singular value decomposition (SVD) for its product with the adjoint of M by

$$M^*(Y) = \begin{bmatrix} U_k & U_{-k}\end{bmatrix} \begin{bmatrix} \Sigma_k & \\ & \Sigma_{-k}\end{bmatrix} \begin{bmatrix} V_k^T \\ V_{-k}^T\end{bmatrix}, \tag{12}$$

where Σ_k is the diagonal matrix consisting of top-k singular values of $M^*(Y)$, the matrices U_k and V_k contain the corresponding left and right singular vectors, and the matrices U_{-k}, V_{-k}, and Σ_{-k} contain the remaining singular vectors and values. Then the reduced dictionary implied by U_k and V_k can be expressed as

$$\mathcal{A}_k = \{uv^T \mid u \in \operatorname{range}(U_k),\ v \in \operatorname{range}(V_k),\ \|u\|_2 = \|v\|_2 = 1\} \subset \mathcal{A}.$$

We construct the cone of essential atoms as the convex cone generated from the reduced dictionary $\mathcal{A}_k$, i.e.,

$$\mathsf{EssCone}_{\mathcal{A},k}(M^*(Y)) \coloneqq \operatorname{cone}(\mathcal{A}_k) = \{U_k C V_k^T \mid C \in \mathbb{R}^{k\times k}\}. \tag{13}$$

Thus, the primal-retrieval computation in Step 5 of Algorithm 1 is then given by

$$\widehat{X} = U_k \widehat{C} V_k^T \tag{14}$$

where

$$\widehat{C} \in \operatorname*{arg\,min}_{C\in\mathbb{R}^{k\times k}} f_k(C), \text{with} f_k(C) \coloneqq f\left(B - M(U_k C V_k^T)\right), \tag{15}$$

is a $k \times k$ matrix obtained by solving the reduced problem (PR), which is defined over the cone generated by the essential atoms identified through the top-k singular triples of $M^*(Y)$, described by (13).

▶ Example 11 (Low-rank matrix completion). The low-rank matrix completion (LRMC) problem aims to recover a low-rank matrix from partial observations, which arises in many real applications such as recommender systems [55] and in a convex formulation of the phase retrieval problem [20]. The LRMC problem can be expressed as

$$\text{find } X \in \mathbb{R}^{m\times n} \text{ such that } \sum_{(i,j)\in\Omega} \tfrac{1}{2}\left(X_{i,j} - B_{i,j}\right)^2 \le \alpha \text{ and } \operatorname{rank}(X) \le k, \tag{16}$$

where $\{B_{i,j} \mid (i,j) \in \Omega\}$ is the set of observations over the index set Ω. Problem (16) corresponds to (P) with the objective $f = \frac{1}{2}\|\cdot\|_2^2$, the dictionary $\mathcal{A}$ given by (10), and the linear operator M defined by the mask

$$M(X)_{i,j} = \begin{cases} X_{i,j} & (i,j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

Fazel [30] popularized the convex relaxation of (16) that minimizes the sum of singular values of X:

$$\operatorname*{minimize}_{X\in\mathbb{R}^{m\times n}} \ \|X\|_* \text{subject to } \sum_{(i,j)\in\Omega} \tfrac{1}{2}\left(X_{i,j} - B_{i,j}\right)^2 \le \alpha. \tag{17}$$

This problem corresponds to formulation (P$_3$). As with Example 6, there are many dual methods that can generate dual feasible iterates $Y^{(t)}$ converging to the dual solution of (17), such as a dual bundle method [29]. The resulting primal-retrieval strategy for Step 5 of Algorithm 1 can be implemented by executing the following steps:

1. (Top-k atoms) Compute the leading k singular vectors of the matrix $M^*(Y^{(t)})$, given by $U_k^{(t)} \in \mathbb{R}^{m\times k}$ and $V_k^{(t)} \in \mathbb{R}^{n\times k}$ as defined by the SVD (12).
2. (Retrieve coefficients) Solve the reduced problem (15), where in this case,

   $$C^{(t)} \in \operatorname*{arg\,min}_{C\in\mathbb{R}^{k\times k}} f_k(C), \text{with} f_k(C) := \sum_{(i,j)\in\Omega} \tfrac{1}{2}\left([U_k^{(t)} C (V_k^{(t)})^\intercal]_{i,j} - B_{i,j}\right)^2.$$

   This least-squares problem can be solved to within ϵ-accuracy in $\mathcal{O}((k|\Omega| + (m+n)k + k^3)\epsilon^{-0.5})$ iterations, for example, with the FISTA algorithm [8]. Typically, $k \ll \min\{m,n\}$, and so we expect that this reduced problem is significantly cheaper to solve than the original problem (17).
3. (Termination) Step 6 of Algorithm 1 terminates when the value of the reduced objective satisifes the condition $f_k(C^{(t)}) \le \alpha + \epsilon$, where ϵ is some pre-defined tolerance. In that case, the algorithm returns with the primal estimate constructed from the left and right singular vectors:

   $$X^{(t)} = U_k^{(t)} C^{(t)} (V_k^{(t)})^\intercal.$$

We describe numerical experiments for the low-rank matrix completion problem in Section 7.2. ◀

6.2.1 Iteration complexity

In the polyhedral case, we were able to assert through Proposition 8 that the optimal primal variable's atomic support could be identified in finite time. As we show here, however, finite-time identification is not possible for the spectral case. The following counterexample shows that the partial SVD of $M^*(Y)$, which we used in (12), is not able to give us a safe cover of the essential atoms in $\mathcal{E}_\mathcal{A}(M^*(Y^*))$ even when this set is a singleton and Y arbitrarily close to a dual solution Y^*.

▶ Example 12 (Limitation of Partial SVD). Consider the problem

$$\operatorname*{minimize}_{X\in\mathbb{R}^{n\times n}} \ \tfrac{1}{2}\|X - B\|_F^2 \text{ subject to } \|X\|_* \le 1, \tag{18}$$

where

$$B = U\operatorname{Diag}(2, 0.1, \dots, 0.1)V^T \text{ and } U = V = \begin{bmatrix} \sqrt{1-\epsilon} & 0 & \dots & -\sqrt{\epsilon} \\ 0 & 1 & \dots & \\ \vdots & & \ddots & \\ \sqrt{\epsilon} & 0 & & \sqrt{1-\epsilon} \end{bmatrix}_{n\times n}$$

for some fixed $\epsilon \in (0,1)$. The dual problem is

$$\underset{Y\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \tfrac{1}{2}\|Y-B\|_F^2 - \tfrac{1}{2}\|B\|_F^2 + \|Y\|_2. \tag{19}$$

The solutions for the dual pair (18) and (19) are

$$X^* = U\,\mathrm{Diag}(1,0,\dots,0)V^T \ \text{ and } \ Y^* = B - X^* = U\,\mathrm{Diag}(1,0.1,\dots,0.1)V^T.$$

In this pair of problems, the linear operator M is simply the identity map, and the cone of essential atoms described by (13) depends only on the dual variable Y. Let u_1 and v_1, respectively, be the first columns of U and V. Evidently, the support of X^* coincides with the essential atoms of Y^*, and moreover, the support is a unique singleton. In other words,

$$\mathcal{S}_{\mathcal{A}}(X^*) = \mathcal{E}_{\mathcal{A}}(Y^*) = \{u_1v_1^T\}.$$

We construct the following dual feasible solution

$$\widehat{Y} = \mathrm{Diag}(1,0.1,\dots,0.1).$$

Because $\widehat{Y}$ is diagonal, its left and right singular vectors $\widehat{U}$ and $\widehat{V}$ are given by $\widehat{U} = \widehat{V} = I = [e_1,e_2,\dots,e_n]$. Note also that the top singular vector $u_1 = [\sqrt{1-\epsilon},0,\dots,\sqrt{\epsilon}]^T$ lies in the span of the basis vectors e_1 and e_n that constitute the top and bottom singular vectors of $\widehat{Y}$. Therefore, any top-r SVD of $\widehat{Y}$, with $r<n$, cannot be used to recover exactly a primal solution X^*. Moreover, $\|\widehat{Y}-Y^*\|_F = \mathcal{O}(\sqrt{\epsilon})$ for any $\epsilon\in(0,1)$, which in effect implies that it is impossible to recover exactly the true solution even with an arbitrarily accurate dual approximation $\widehat{Y}$. ◀

This last example motivates our study of the quality with which a partial SVD of a given feasible dual solution $M^*(Y)$ can be used to approximate the support $\mathcal{E}_{\mathcal{A}}(M^*(Y^*))$. The next result measures the difference between $\mathcal{S}_{\mathcal{A}}(x^*)$ and $\mathsf{EssCone}_{\mathcal{A},k}(M^*(Y))$ using one-sided Hausdorff distance.

▶ **Proposition 13** (Error in truncated SVD)**.** *Let Y be feasible for one of the dual problems (D_i) for $i=1,2,3$. Let $\{\sigma_j\}_{j=1}^{\min\{n,m\}}$ be the singular values satisfying $\sigma_1 \ge \cdots \ge \sigma_{\min\{n,m\}}$, where we assume $\sigma_1 > \sigma_{k+1}$. Let $Z := M^*(Y)$ and let $\mathsf{EssCone}_{\mathcal{A},k}(Z)$ denote the cone generated according to equation* (13)*. Then*

$$\begin{aligned}\mathrm{dist}(\mathcal{S}_{\mathcal{A}}(X^*),\ \mathsf{EssCone}_{\mathcal{A},k}(Z)) &\le \mathrm{dist}(\mathcal{E}_{\mathcal{A}}(Z,\epsilon_i),\mathsf{EssCone}_{\mathcal{A},k}(Z))\\ &\le \sqrt{2\min\left\{\frac{\epsilon_i}{\sigma_1-\sigma_{k+1}},1\right\}},\end{aligned}$$

where each ϵ_i is defined in Theorem 5 and the function

$$\mathrm{dist}(\mathcal{A}_1,\mathcal{A}_2) := \sup_{a_1\in\mathcal{A}_1}\ \inf_{a_2\in\mathcal{A}_2}\ \|a_1-a_2\|_F.$$

is the one-sided Hausdorff distance between sets $\mathcal{A}_1$ and $\mathcal{A}_2$.

Oustry [53, Theorem 2.11] developed a related result based on the two-sided Hausdorff distance. Directly applying Oustry's result to our context results in a bound on the order $\mathcal{O}(\sqrt{\epsilon/(\sigma_k-\sigma_{k+1})})$, which is looser than the bound shown in Proposition 13 because $\sigma_1\ge\sigma_k\ge\sigma_{k+1}$.

Finally, we show the error bound for primary recovery.

▶ **Proposition 14.** *Assume that $f\ge 0$ and $f(0)=0$. Let X^* and Y, respectively, be primal optimal and dual feasible for one of the primal-dual pairs (P_i) and (D_i), for $i=1,2,3$. Let $\{\sigma_j\}_{j=1}^{\min\{n,m\}}$ be the singular values satisfying $\sigma_1\ge\cdots\ge\sigma_{\min\{n,m\}}$, where we assume $\sigma_1>\sigma_{k+1}$. Let $\mathsf{EssCone}_{\mathcal{A},k}(M^*(Y))$ denote the cone generated according to equation* (13)*. Let $\widehat{X}$ be the solution recovered via* (14)*. Then*

$$f(B-M(\widehat{X})) \le f(B-M(X^*)) + \mathcal{O}\left(\sqrt{\frac{\epsilon_i}{\sigma_1-\sigma_{k+1}}}\right),$$

where each ϵ_i is defined in Theorem 5.

Proposition 14 characterizes the error bound for our primal-retrieval strategy when $\mathcal{A}$ is spectral. Our next corollary shows that Algorithm 1 can terminate in polynomial time with any tolerance $\epsilon>0$.

▶ **Corollary 15.** *For one of the problems (D_i), $i = 1, 2, 3$, suppose that a dual oracle generates iterates $Y^{(t)}$ converging to optimal variable Y^* with convergence rate*

$$d_i(Y^{(t)}) - d_i(Y^*) \in \mathcal{O}\left(t^{-p}\right)$$

for some $p > 0$. If the atomic set $\mathcal{A}$ is spectral then Algorithm 1 with $\epsilon > 0$ will terminate in $\mathcal{O}\left(\epsilon^{-4/p}\right)$ iterations.

6.2.2 Non-centrosymmetry and constrained primal recovery

We now consider the case in which the atomic set $\mathcal{A}$ given by (11), which is not centrosymmetric. As we show below, the corresponding primal recovery problem is constrained.

Fix a feasible dual variable Y and define the eigenvalue decomposition for its product with the adjoint of M by

$$M^*(Y) = \begin{bmatrix} V_k & V_{-k} \end{bmatrix} \begin{bmatrix} \Sigma_k & \\ & \Sigma_{-k} \end{bmatrix} \begin{bmatrix} V_k^\intercal \\ V_{-k}^\intercal \end{bmatrix},$$

where V_k and the diagonal matrix Σ_k, respectively, contain the top-k eigenvectors and eigenvalues of $M^*(Y)$, and V_{-k} and the diagonal matrix Σ_{-k}, respectively, contain the remaining eigenvectors and eigenvalues. Then the reduced dictionary implied by V_k can be expressed as

$$\mathcal{A}_k = \{vv^\intercal \mid v \in \operatorname{range}(V_k),\ \|v\|_2 = 1\} \subset \mathcal{A}.$$

The convex cone of essential atoms generated from the reduced dictionary $\mathcal{A}_k$ is given by

$$\mathsf{EssCone}_{\mathcal{A},k}(M^*(Y)) := \operatorname{cone}(\mathcal{A}_k) = \{V_k C V_k^\intercal \mid C \in \mathbb{R}^{k\times k},\ C \succeq 0\}.$$

The recovery problem (15) then becomes constrained, i.e.,

$$\widehat{C} \in \operatorname*{arg\,min}_{C\in\mathbb{R}^{k\times k},\ C\succeq 0} f_k(C).$$

7 Numerical experiments

We conduct several numerical experiments on both synthetic and real-world datasets to empirically verify the effectiveness of our proposed primal-retrieval strategy. In Section 7.1, we describe experiments on the basis pursuit denoising problem (Example 6), which shows the performance of our strategy on polyhedral atomic set. In Section 7.2, we apply our primal-retrieval technique to the low-rank matrix completion problem (Example 11) and test the effectiveness of our proposed method on the spectral atomic set. In Section 7.3, we describe experiments on a image preprocessing problem, where the dictionary $\mathcal{A}$ is the sum of a polyhedral atomic set and a spectral atomic set. This shows that our strategy can be applied to more complicated cases. For all experiments, we implement the level-set method proposed by Aravkin et al. [4] where we only store the dual variable y. We implement the level-set method and our primal-retrieval strategy in the Julia language [12] and our code is publicly available at `https://github.com/MPF-Optimization-Laboratory/AtomicOpt.jl`. All the experiments are carried out on a Linux server with 8 CPUs and 64 GB memory.

7.1 Basis pursuit denoising

The experiments in this section include a selection of five relevant basis pursuit problems from the Sparco collection [11] of test problems. The chosen problems are all real-valued and suited to one-norm regularization. Each problem in the collection includes a linear operator $M : \mathbb{R}^n \to \mathbb{R}^m$ and a right-hand-side vector $b \in \mathbb{R}^m$. Table 2 summarizes the selected problems. We compare the results with SPGL1 [9]. In all problems, we set $\alpha = 10^{-3} \cdot \|b\|$. The results are shown in Table 3 where `nMat` denotes the total number of matrix-vector products with M or M^*. As we can observe from Table 3, the level-set algorithm equipped with our primal-retrieval technique can obtain an ϵ-feasible solution within a small number of iterations, which is consistent with the finite-time identification property described by Proposition 8. We also observe that level-set method coupled with the primal-retrieval strategy can converge faster than SPGL1 with its default stopping criterion. This suggests that our primal-retrieval technique is both a memory-efficient method for obtaining approximal primal solutions with provable error bounds, and is also a practical technique that allow the optimization algorithm to stop early.

Table 2 The Sparco test problems used.

Problem	ID	m	n	$\|b\|$	M
`blocksig`	2	1024	1024	7.9e+1	wavelet
`cosspike`	3	1024	2048	1.0e+2	DCT
`gcosspike`	5	300	2048	8.1e+1	Gaussian ensemble + DCT
`sgnspike`	7	600	2560	2.2e+0	Gaussian ensemble
`spiketrn`	903	1024	1024	5.7e+1	1D convolution

Table 3 Basis pursuit denoising comparisons.

Problem	`nnz`(x)	`nMat`(SPGL1)	`nMat`(level-set+PR)
`blocksig`	71	22	5
`cosspike`	113	77	71
`gcosspike`	113	434	141
`sgnspike`	20	44	21
`spiketrn`	35	4761	1888

7.2 Low-rank matrix completion

For this atomic set, we conduct an experiment similar to that carried out by Candés and Plan [17]. We retrieved from the website of National Centers for Environmental Information[1] a 6798-by-366 matrix X whose entries are daily average temperatures at 6798 different weather stations throughout the world in year 2020. The temperature matrix X is approximately low rank in the sense that $\|X - X_5\|_F / \|X\|_F \approx 24\%$, where X_5 is the matrix created by truncating the SVD after the top 5 singular values.

To test the performance of our matrix completion algorithm, we subsampled 50% of X and then recovered an estimate $\widehat{X}$. The solution gives a relative error of $\|X - \widehat{X}\|_F / \|X\|_F \approx 30\%$. The result is shown in Figure 1a. As we can see from Figure 1a, the recovery error exhibits a positive correlation with the duality gap, both the duality gap and the recovery gap decrease as the number of iteration increase. The observation in this experiment is consistent with our theory (Proposition 14).

To further demonstrate the efficiency of our primal-retrieval method, we conducted an experiment to compare the performance of the level-set method with and without primal retrieval on randomly generated low-rank matrices of sizes ranging from 50 × 50 to 500 × 500. For each problem size, we generated observations from $n \log(n)^2$ sampled entries with standard Gaussian noise, and repeated the experiment for 10 times. The error was defined as $\|X^* - X\|_F / \|X\|_F$, where X^* is the recovered matrix and X is the ground truth matrix. The results, as summarized in the Table 4, demonstrate that the level-set method with primal retrieval consistently required fewer iterations than the method without primal retrieval while achieving the same error level. These findings suggest that primal retrieval can significantly improve the efficiency of the level-set method, making it a more effective tool for low-rank matrix recovery tasks.

7.3 Robust principal component analysis

In this section we show that our primal-retrieval strategy can be applied to more complicated atomic sets besides polyhedral and spectral. We conduct a similar experiment as in Candès et al. [16]. Face recognition algorithms are sensitive to shadows on faces, and therefore it is necessary to remove illumination variations and shadows on the face images. We obtained face images from the Yale B face database [34]. We show the original faces in Figure 1b, where each face image was of size 192 × 168 with 64 different lighting conditions. The images were then reshaped into a matrix $M \in \mathbb{R}^{32256 \times 64}$. Because of the similarity between faces and the sparse structure of the shadow, the matrix M can be approximately decomposed into two components, i.e.,

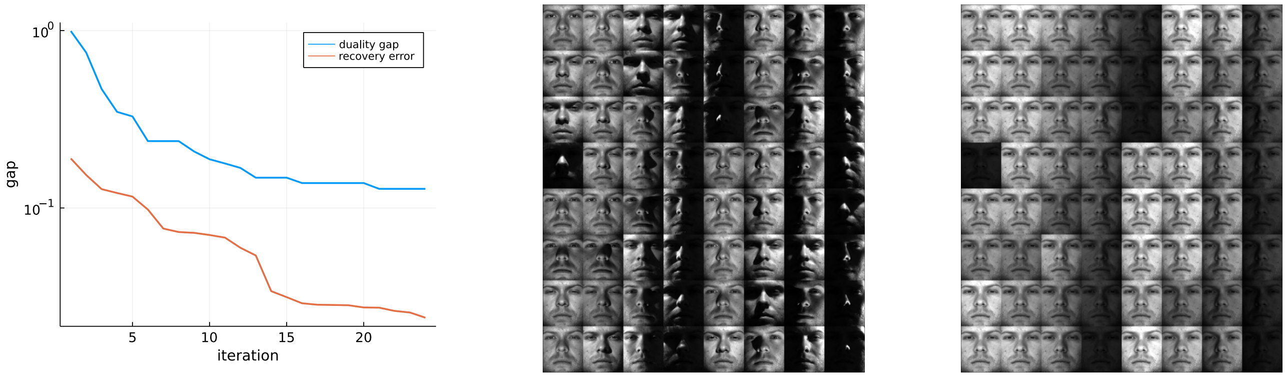
$$M \approx L + S,$$

where L is a low-rank matrix corresponding to the clean faces and S is sparse matrix corresponding to the shadows. Based on the work by Fan et al. [27], we know that such decomposition can be obtained via solving the

[1] `https://www.ncei.noaa.gov`

Table 4 Comparison of level-set method with primal-retrieval (wPR) and level-set method without primal-retrieval (woPR) for different problem sizes. The table shows the means and standard deviations of relative solution errors and number of iterations.

Size	Relative solution error		Number of iterations	
	wPR	woPR	wPR	woPR
50×50	$7 \times 10^{-2} \pm 4 \times 10^{-4}$	$9 \times 10^{-2} \pm 2 \times 10^{-3}$	2 ± 0	2787 ± 375
100×100	$9 \times 10^{-2} \pm 3 \times 10^{-3}$	$9 \times 10^{-2} \pm 1 \times 10^{-3}$	415 ± 675	6839 ± 1022
150×150	$9 \times 10^{-2} \pm 1 \times 10^{-3}$	$9 \times 10^{-2} \pm 1 \times 10^{-3}$	219 ± 175	12889 ± 1652
200×200	$9 \times 10^{-2} \pm 1 \times 10^{-3}$	$9 \times 10^{-2} \pm 1 \times 10^{-3}$	274 ± 375	16217 ± 3489
250×250	$9 \times 10^{-2} \pm 3 \times 10^{-4}$	$9 \times 10^{-2} \pm 2 \times 10^{-4}$	7105 ± 1220	18264 ± 7008
300×300	$9 \times 10^{-2} \pm 4 \times 10^{-4}$	$9 \times 10^{-2} \pm 4 \times 10^{-4}$	8177 ± 3372	22028 ± 4103
350×350	$9 \times 10^{-2} \pm 2 \times 10^{-4}$	$9 \times 10^{-2} \pm 2 \times 10^{-4}$	15829 ± 1186	29652 ± 2073
400×400	$9 \times 10^{-2} \pm 4 \times 10^{-4}$	$9 \times 10^{-2} \pm 4 \times 10^{-4}$	20126 ± 3153	31515 ± 4096
450×450	$8 \times 10^{-2} \pm 5 \times 10^{-4}$	$8 \times 10^{-2} \pm 6 \times 10^{-4}$	23807 ± 5356	37065 ± 10058
500×500	$8 \times 10^{-2} \pm 5 \times 10^{-4}$	$8 \times 10^{-2} \pm 5 \times 10^{-4}$	30069 ± 4111	51373 ± 15287



(a) Recovery error curve. **(b)** Original faces. **(c)** Shadow-removed faces.

Figure 1 The left figure (1a) shows the result of the matrix completion experiment. The middle and right figures (1b, 1c) are for the robust principal component analysis experiment.

following convex optimization problem:

$$\min_{L,S} \max\{\|L\|_*,\ \lambda\|S\|_1\}\ \text{ subject to }\ \|L+S-M\| \le \alpha. \tag{20}$$

By Fan et al. [28, Proposition 7.3], we know that (20) is equivalent to

$$\min_{X} \gamma_{\mathcal{A}}(X)\ \text{ subject to }\ \|X-M\| \le \alpha,$$

with $X = L + S$ and where $\mathcal{A} = \lambda\mathcal{A}_1 + \mathcal{A}_2$, $\mathcal{A}_1 = \{uv^\intercal \mid u \in \mathbb{R}^m,\ v \in \mathbb{R}^n,\ \|u\|_2 = \|v\|_2 = 1\}$ and $\mathcal{A}_2 = \{\pm e_i e_j^\intercal \mid i \in [m], j \in [n]\}$. The recovered low-rank component is shown in Figure 1c. As we can see from the figure, most of the shadow are successfully removed. This experiment suggests that our primal-retrieval technique can potentially be used for more complex atomic set and allow the underlying the dual-algorithm to produce satisfactory result within a reasonable number of iterations.

8 Conclusion

In this work, we proposed a simple primal-retrieval strategy for atomic-sparse optimization. We demonstrate both theoretically and empirically that our proposed strategy can obtain good solutions to the cardinality-constrained problem given a dual-based algorithm converging to the optimum dual solution.

Further research opportunities remain, particularly for designing meaningful primal-retrieval strategies for non-polyhedral and non-spectral atomic sets. The primal-retrieval technique developed in this work is algorithm-agnostic, and it is an open question if it is possible to develop more efficient primal-retrieval approaches tailored to specific optimization algorithms, such as the conditional-gradient method.

Appendix

A Derivation of duals

We derive the dual problems (D_1), (D_2) and (D_3) using the Fenchel–Rockafellar duality framework. We use the following result.

▶ **Theorem 16** ([56, Corollary 31.2.1]). *Let $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^m \to \mathbb{R}$ be two closed proper convex functions and let M be a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$, then*

$$\inf_{x\in\mathbb{R}^n} f_1(x) + f_2(Mx) = \inf_{y\in\mathbb{R}^m} f_1^*(M^*y) + f_2^*(-y).$$

If there exist x in the interior of $\operatorname{dom} f_1$ such that Mx is in the interior of $\operatorname{dom} f_2$, then strong duality holds, namely both infima are attained.

We also need a result that describes the relationship between gauge, support, and indicator functions.

▶ **Proposition 17** ([28, Proposition 3.2]). *Let $C \subset \mathbb{R}^n$ be a closed convex set that contains the origin. Then*

$$\gamma_C = \sigma_{C^\circ} = \delta^*_{C^\circ}.$$

For problem (P_1), let

$$f_1 := \lambda\gamma_{\mathcal{A}} \text{ and } f_2 := f(b - \cdot\,)$$

By the properties of conjugate functions and Proposition 17, we obtain

$$f_1^* = \delta_{(\frac{1}{\lambda}\mathcal{A})^\circ} = \delta_{\{x \mid \sigma_{\mathcal{A}}(x)\le\lambda\}} \text{ and } f_2^* = \langle b, \cdot\,\rangle + f^*(-\,\cdot\,).$$

Then by Theorem 16, we can get the dual problem for (P_1) as

$$\operatorname*{minimize}_{y\in\mathbb{R}^m} \ f^*(y) - \langle b, y\rangle \ \text{subject to} \ \sigma_{\mathcal{A}}(M^*y) \le \lambda.$$

For (P_2),

$$f_1 = \delta_{\{x \mid \gamma_{\mathcal{A}}(x)\le\tau\}} = \delta_{\tau\mathcal{A}} \text{ and } f_2 = f(b - \cdot\,).$$

By the properties of conjugate functions and Proposition 17, we obtain

$$f_1^* = \sigma_{\tau\mathcal{A}} = \tau\sigma_{\mathcal{A}} \text{ and } f_2^* = \langle b, \cdot\,\rangle + f^*(-\,\cdot\,).$$

Then by Theorem 16, it follows that the dual problem for (P_2) is

$$\operatorname*{minimize}_{y\in\mathbb{R}^m} \ f^*(y) - \langle b, y\rangle + \tau\sigma_{\mathcal{A}}(M^*y).$$

For (P_3),

$$f_1 = \gamma_{\mathcal{A}} \text{ and } f_2 = \delta_{\{x \mid f(b-x)\le\alpha\}}.$$

By the properties of conjugate functions and Proposition 17, we can get that

$$f_1^* = \delta_{\{x \mid \sigma_{\mathcal{A}}(x)\le 1\}} \text{ and } f_2^* = \sigma_{\{f(b-x)\le\alpha\}}.$$

Then by [38, Example E.2.5.3], we know that the support function of the sublevel set is

$$f_2^* = \sigma_{\{x \mid f(b-x)\le\alpha\}} = \min_{\beta>0} \beta\left(f^*\left(-\frac{\cdot}{\beta}\right) + \alpha\right) + \langle b, \cdot\,\rangle.$$

Finally, by Theorem 16, we can get the dual problem for (P_3) as

$$\operatorname*{minimize}_{y\in\mathbb{R}^m,\ \beta>0} \ \beta\left(f^*\left(\frac{y}{\beta}\right) + \alpha\right) - \langle b, y\rangle \ \text{subject to} \ \sigma_{\mathcal{A}}(M^*y) \le 1.$$

B Proof of Theorem 5

The proof of this Theorem relies on the following duality property between smoothness and strong convexity.

▶ **Lemma 18** ([42, Theorem 6]). *If f is L-smooth, then f^* is $\frac{1}{L}$-strongly convex.*

Proof of Theorem 5.

a. Let y^* denote the optimal dual variable for D_1. First, we show that $\|y - y^*\|$ can be bounded by the duality gap. Let $g(y) := f^*(y) - \langle b, y\rangle$. By Lemma 18, f^* is $\frac{1}{L}$-strongly convex, and it follows that g is also $\frac{1}{L}$-strongly convex. By the definition of strong convexity,

$$\forall\, s \in \partial g(y^*),\;\; g(y) \geq g(y^*) + \langle s, y - y^*\rangle + \frac{1}{2L}\|y - y^*\|^2.$$

Optimality requires that

$$\exists\, s \in \partial g(y^*),\;\; \langle s, y - y^*\rangle \geq 0 \quad \forall\, y \text{ s.t. } \sigma_{\mathcal{A}}(M^* y) \leq \lambda.$$

Therefore, reordering the inequality gives

$$\|y - y^*\| \leq \sqrt{2L(g(y) - g(y^*))} \quad \forall\, y \in \mathbb{R}^m.$$

Next, we show that $\mathcal{E}_{\mathcal{A}}(M^* y^*) \subseteq \mathcal{E}_{\mathcal{A}}(M^* y, \epsilon_1)$. For any $a \in \mathcal{E}_{\mathcal{A}}(M^* y^*)$,

$$\begin{aligned}
\langle a, M^* y\rangle &= \sigma_{\mathcal{A}}(M^* y^*) + \langle Ma, y - y^*\rangle \\
&\geq \sigma_{\mathcal{A}}(M^* y^*) - \left(\max_{a\in\mathcal{A}} \|Ma\|\right)\|y - y^*\| \\
&\geq \sigma_{\mathcal{A}}(M^* y^*) - \left(\max_{a\in\mathcal{A}} \|Ma\|\right)\sqrt{2L(g(y) - g(y^*))} \\
&\geq \sigma_{\mathcal{A}}(M^* y) - \epsilon_1,
\end{aligned}$$

where the last inequality follows from the definition of ϵ_1 in Theorem 5 and the fact that $\sigma_{\mathcal{A}}(M^* y^*) = \lambda$ and y is feasible for (D_1).

b. Let y^* denote the optimal dual variable for D_2. First, we show that $\|y - y^*\|$ can be bounded by the duality gap. Let $g(y) := f^*(y) - \langle b, y\rangle + \tau\sigma_{\mathcal{A}}(M^* y)$. By Lemma 18, f^* is $\frac{1}{L}$-strongly convex, and it follows that g is also $\frac{1}{L}$-strongly convex. By the definition of strongly convex,

$$\forall\, s \in \partial g(y^*),\;\; g(y) \geq g(y^*) + \langle s, y - y^*\rangle + \frac{1}{2L}\|y - y^*\|^2.$$

By optimality, $0 \in \partial g(y^*)$. Reordering the inequality to deduce that

$$\|y - y^*\|_2 \leq \sqrt{2L(g(y) - g(y^*))}.$$

Next, we show that $\mathcal{E}_{\mathcal{A}}(M^* y^*) \subseteq \mathcal{E}_{\mathcal{A}}(M^* y, \epsilon_2)$. For any $a \in \mathcal{E}_{\mathcal{A}}(M^* y^*)$,

$$\begin{aligned}
\langle a, M^* y\rangle &\geq \sigma_{\mathcal{A}}(M^* y^*) - \left(\max_{a\in\mathcal{A}} \|Ma\|\right)\|y - y^*\| \\
&= \sigma_{\mathcal{A}}(M^* y) - (\sigma_{\mathcal{A}}(M^* y) - \sigma_{\mathcal{A}}(M^* y^*)) - \left(\max_{a\in\mathcal{A}} \|Ma\|\right)\|y - y^*\| \\
&\geq \sigma_{\mathcal{A}}(M^* y) - 2\left(\max_{a\in\mathcal{A}} \|Ma\|\right)\|y - y^*\| \\
&\geq \sigma_{\mathcal{A}}(M^* y) - \epsilon_2,
\end{aligned}$$

where the last inequality follows from the definition of ϵ_2 in Theorem 5.

c. Let (y^*, β^*) denote the optimal dual variables for D_3. First, we show that $\|y - y^*\|$ can be bounded by the duality gap. Let

$$g(y) := \beta^* f^*\left(\frac{y}{\beta^*}\right) + \beta^*\alpha - \langle b, y\rangle.$$

By Lemma 18, f^* is $\frac{1}{L}$-strongly convex, and it is not hard to check that g is $\frac{1}{\beta^* L}$-strongly convex. By the definition of strongly convex,

$$\forall\, s \in \partial g(y^*),\;\; g(y) \geq g(y^*) + \langle s, y - y^* \rangle + \frac{1}{2\beta^* L}\|y - y^*\|^2.$$

By optimality,

$$\exists\, s \in \partial g(y^*),\;\; \langle s, y - y^* \rangle \geq 0 \quad \forall\, y \text{ s.t. } \sigma_{\mathcal{A}}(M^* y) \leq 1.$$

Reorder the inequality to deduce that

$$\|y - y^*\| \leq \sqrt{2\beta^* L(g(y) - g(y^*))}.$$

Because β^* is unknown to us, we will then get an upper bound for $d_3(y, \beta^*)$. Fix y, let $h(\beta) = d_3(y, \beta)$. By the property of perspective function, we know that h is convex. Then it follows that

$$d_3(y, \beta^*) \leq \max\{d_3(y, \underline{\beta}), d_3(y, \overline{\beta})\}.$$

Therefore,

$$\|y - y^*\| \leq \sqrt{2\overline{\beta} L \left(\max\{d_3(y, \underline{\beta}), d_3(y, \overline{\beta})\} - d_3(y^*, \beta^*)\right)}.$$

Finally, we show that $\mathcal{E}_{\mathcal{A}}(M^* y^*) \subseteq \mathcal{E}_{\mathcal{A}}(M^* y, \epsilon_3)$. For any $a \in \mathcal{E}_{\mathcal{A}}(M^* y^*)$,

$$\begin{aligned}
\langle a, M^* y \rangle &\geq \sigma_{\mathcal{A}}(M^* y^*) - \left(\max_{a \in \mathcal{A}} \|Ma\|\right) \|y - y^*\| \\
&= \sigma_{\mathcal{A}}(M^* y) - (\sigma_{\mathcal{A}}(M^* y) - \sigma_{\mathcal{A}}(M^* y^*)) - \left(\max_{a \in \mathcal{A}} \|Ma\|\right) \|y - y^*\| \\
&\geq \sigma_{\mathcal{A}}(M^* y) - 2\left(\max_{a \in \mathcal{A}} \|Ma\|\right) \|y - y^*\| \\
&\geq \sigma_{\mathcal{A}}(M^* y) - \epsilon_3.
\end{aligned}$$

◀

C Upper and lower bound for β^*beta

First, we consider (D$_3$). Let $w = y/\beta$, then (D$_3$) can be equivalently expressed as

$$\underset{w}{\text{minimize}}\;\; \inf_{\beta > 0} \beta(f^*(w) - \langle b, w \rangle + \alpha) \;\text{ subject to }\; \sigma_{\mathcal{A}}(M^* w) \leq \beta.$$

Fix $\beta = \beta^*$, then (D$_3$) can be expressed as

$$\underset{w}{\text{minimize}}\;\; f^*(w) - \langle b, w \rangle \;\text{ subject to }\; \sigma_{\mathcal{A}}(M^* w) \leq \beta^*. \tag{21}$$

Now compare (21) with (D$_1$) to conclude that they are equivalent when $\lambda = \beta^*$. It thus follows that (P$_3$) is equivalent to

$$\underset{x}{\text{minimize}}\;\; f(b - Mx) + \beta^* \gamma_{\mathcal{A}}(x).$$

Next, consider using the level-set method [4] with bisection to solve (P$_3$). There exists $\tau^* > 0$ such that (P$_3$) is equivalent to

$$\underset{x}{\text{minimize}}\;\; f(b - Mx) \;\text{ subject to }\; \gamma_{\mathcal{A}}(x) \leq \tau^*.$$

With the level-set method, we are able to get (x_1, τ_1) and (x_2, τ_2) such that $\tau_1 \leq \tau^* \leq \tau_2$ and x_i is the optimum for

$$\underset{x}{\text{minimize}}\;\; f(b - Mx) \;\text{ subject to }\; \gamma_{\mathcal{A}}(x) \leq \tau_i,$$

for $i = 1, 2$. Then there exits β_1 and β_2 such that $\beta_1 \geq \beta^* \geq \beta_2$ and x_i is optimal for

$$\underset{x}{\text{minimize}}\;\; f(b - Mx) + \beta_i \gamma_{\mathcal{A}}(x),$$

for $i = 1, 2$.

Finally, by [28, Theorem 5.1] we can conclude that

$$\beta_i = \sigma_{\mathcal{A}}(M^*\nabla f(b - Mx_i)) \text{for} i = 1, 2.$$

Therefore, we can get upper and lower bounds for β^* via level-set method with bisection. Moreover, by strong duality and convergence of the bisection method, the gap between β_1 and β_2 will converge to zero.

D Proof of Proposition 8

First, we show that for any y_i such that $\|y_i - y_i^*\| \le \frac{\delta}{4\|M\|_{\mathcal{A}}}$, the condition

$$\mathcal{F}_{\mathcal{A}}(M^*y_i^*) \subseteq \mathsf{EssCone}(\mathcal{A}, M, y_i, k)$$

holds. By Assumption 3 and the definition of δ-nondegeneracy, we know that

$$|\mathcal{F}_{\mathcal{A}}(M^*y_i^*)| = k, \qquad \text{and} \qquad \langle Ma, y_i^*\rangle \le \sigma_{\mathcal{A}}(M^*y_i^*) - \delta \quad \forall\, a \notin \mathcal{F}_{\mathcal{A}}(M^*y_i^*). \tag{22}$$

For any $a \in \mathcal{F}_{\mathcal{A}}(M^*y_i^*)$, we have

$$\begin{aligned}\langle a, M^*y_i\rangle &\ge \langle a, M^*y_i^*\rangle - |\langle Ma, y_i^* - y_i\rangle| \\ &\ge \sigma_{\mathcal{A}}(M^*y_i^*) - \|M\|_{\mathcal{A}}\frac{\delta}{4\|M\|_{\mathcal{A}}} \quad \left(\text{by the condition } \|y_i - y_i^*\| \le \frac{\delta}{4\|M\|_{\mathcal{A}}}\right) \\ &\ge \sigma_{\mathcal{A}}(M^*y_i^*) - \frac{\delta}{4}.\end{aligned}$$

On the other hand, for any $a' \notin \mathcal{F}_{\mathcal{A}}(M^*y_i^*)$, we have

$$\begin{aligned}\langle a', M^*y_i\rangle &\le \langle a', M^*y_i^*\rangle + |\langle Ma', y_i^* - y_i\rangle| \\ &\le \langle a', M^*y_i^*\rangle + \frac{\delta}{4} \\ &\le \sigma_{\mathcal{A}}(M^*y_i^*) - \delta + \frac{\delta}{4} \quad (\text{By } (22)) \\ &= \sigma_{\mathcal{A}}(M^*y_i^*) - \frac{3\delta}{4}.\end{aligned}$$

Therefore,

$$\langle a, M^*y_i\rangle > \langle a', M^*y_i\rangle \quad \forall\, a \in \mathcal{F}_{\mathcal{A}}(M^*y_i^*) \text{ and } a' \notin \mathcal{F}_{\mathcal{A}}(M^*y_i^*).$$

Note that $\mathsf{EssCone}(\mathcal{A}, M, y_i, k)$ contains only the atoms that corresponds to the k largest $\langle a, M^*y_i\rangle$. Therefore $\mathcal{F}_{\mathcal{A}}(M^*y_i^*) \subseteq \mathsf{EssCone}(\mathcal{A}, M, y_i, k)$.

By the assumption $y_i^{(t)} \to y_i^*$. For $i \in \{1, 2, 3\}$, we know there exist $T_i > 0$ such that $\|y_i^{(t)} - y_i^*\| < \frac{\delta}{4\|M\|_{\mathcal{A}}}$ for all $t > T_i$. Therefore $\mathcal{F}_{\mathcal{A}}(M^*y_i^*) \subseteq \mathsf{EssCone}(\mathcal{A}, M, y_i^{(t)}, k)\ \ \forall\, t > T_i$, which completes the proof.

E Proof for Corollary 9

By Theorem 5, we know that

$$\mathcal{S}_{\mathcal{A}}(x_i^*) \subseteq \mathcal{E}_{\mathcal{A}}(M^*y_i, \epsilon_i) \text{with} \epsilon_i \in \mathcal{O}(\sqrt{d_i(y_i^{(t)}) - d_i(y_i^*)}).$$

By the assumption that $d_i(y_i^{(t)}) - d_i(y_i^*) \in \mathcal{O}\left(t^{-p}\right)$ and the problem pair is δ-nondegenerate, it is easy to verify that there exist $T \in \mathcal{O}(\delta^{-2/p})$ such that the algorithm will terminate in T iterations.

F Proof for Proposition 13

First, we derive a monotonicity property of $\operatorname{dist}(\,\cdot\,,\cdot\,)$. By the definition of $\operatorname{dist}(\,\cdot\,,\cdot\,)$, it follows that

$$\operatorname{dist}(A, C) \le \operatorname{dist}(B, C) \quad \forall\, A, B, C \subseteq \mathbb{R}^{n\times m} \text{ such that } A \subseteq B. \tag{23}$$

For any $i \in \{1,2,3\}$, we know that $\mathcal{S}_{\mathcal{A}}(X^*) \subseteq \mathcal{E}_{\mathcal{A}}(Z,\epsilon_i)$ by Theorem 5. Then by (23), we have

$$\operatorname{dist}(\mathcal{S}_{\mathcal{A}}(X^*),\ \mathsf{EssCone}_{\mathcal{A},k}(Z)) \le \operatorname{dist}(\mathcal{E}_{\mathcal{A}}(Z,\epsilon_i),\ \mathsf{EssCone}_{\mathcal{A},k}(Z)).$$

For any $\mathcal{A}_1, \mathcal{A}_2 \subseteq \mathcal{A}$,

$$\rho(\mathcal{A}_1,\mathcal{A}_2) = \sqrt{\sup_{a_1\in\mathcal{A}_1}\ \inf_{a_2\in\mathcal{A}_2} \|a_1 - a_2\|_F^2} = \sqrt{2 - 2\left(\inf_{a_1\in\mathcal{A}_1}\ \sup_{a_2\in\mathcal{A}_2} \langle a_1, a_2\rangle\right)},$$

where the second equality holds since $\|a_1\|_F = \|a_2\|_F = 1$ by the definition of $\mathcal{A}$. Define $\mathcal{A}_1 = \mathcal{E}_{\mathcal{A}}(Z,\epsilon_i)$ and $\mathcal{A}_2 = \mathsf{EssCone}_{\mathcal{A},k}(Z) = \left\{U_r pq^T V_r^T \,|\, \|p\|_2 = \|q\|_2 = 1\right\}$, where U_r, V_r are the top-r singular vectors of M^*y. Let $k := \min\{n,m\}$, $\mathcal{C}_1 = \{(p,q) \,|\, \sum_{i=1}^k \sigma_i p_i q_i \ge \sigma_1 - \epsilon_i,\ \|p\|_2 = \|q\|_2 = 1,\ p,q \in \mathbb{R}^k\}$ and $\mathcal{C}_2 = \{(\widehat{p},\widehat{q}) \,|\, \|\widehat{p}\|_2 = \|\widehat{q}\|_2 = 1,\ \widehat{p},\widehat{q} \in \mathbb{R}^r\}$, then

$$\begin{aligned}
\rho(\mathcal{A}_1,\mathcal{A}_2) &= \sqrt{2 - 2\left(\min_{p,q\in\mathcal{C}_1}\ \max_{\hat{p},\hat{q}\in\mathcal{C}_2} \langle Upq^TV^T, U_r\widehat{p}\widehat{q}^T V_r^T\rangle\right)} \\
&= \sqrt{2 - 2\left(\min_{p,q\in\mathcal{C}_1}\ \max_{\hat{p},\hat{q}\in\mathcal{C}_2} \left(\sum_{i=1}^r p_i\widehat{p}_i\right)\left(\sum_{i=1}^r q_i\widehat{q}_i\right)\right)} \\
&= \sqrt{2 - 2\left(\min_{p,q\in\mathcal{C}_1} \|p_{1:r}\|_2\|q_{1:r}\|_2\right)}. \qquad (24)
\end{aligned}$$

Now we consider the subproblem in (24):

$$\begin{aligned}
&\underset{p,q}{\text{minimize}} \quad \|p_{1:r}\|_2\|q_{1:r}\|_2 && (\mathrm{P_{sub}})\\
&\text{subject to} \quad \sum_{i=1}^k \sigma_i p_i q_i \ge \sigma_1 - \epsilon_i,\ \ \|p\|_2 = \|q\|_2 = 1,\ \ p,q \in \mathbb{R}^k.
\end{aligned}$$

If p^* and q^* is a solution of the problem $(\mathrm{P_{sub}})$, then it is easy to verify that

$$\begin{aligned}
\widetilde{p} &= \left[\|p^*_{1:r}\|_2, 0, \ldots, \|p^*_{r+1:k}\|_2, 0, \ldots, 0\right] \\
\text{and}\quad \widetilde{q} &= \left[\|q^*_{1:r}\|_2, 0, \ldots, \|q^*_{r+1:k}\|_2, 0, \ldots, 0\right]
\end{aligned}$$

is also a valid solution. Therefore there must exist solution p^*, q^* such that $p_i = q_i = 0\ \ \forall\, i \notin \{1, r+1\}$, that is only p_1^*, q_1^* and p_{r+1}^* and q_{r+1}^* are greater or equal than 0. This allow us to further reduce the problem to

$$\begin{aligned}
&\underset{p_1,q_1,p_{r+1},q_{r+1}}{\text{minimize}} \quad p_1q_1 \\
&\text{subject to} \quad \sigma_1p_1q_1 + \sigma_{r+1}p_{r+1}q_{r+1} \ge \sigma_1 - \epsilon_i, \\
&\qquad\qquad\quad p_1^2 + p_{r+1}^2 = q_1^2 + q_{r+1}^2 = 1,\ \ p_1, q_1, p_{r+1}, q_{r+1} \ge 0.
\end{aligned}$$

It is easy to verify that when $\sigma_1 - \sigma_{r+1} \ge \epsilon_i$, the above problem attains solution at

$$p_1 = q_1 = \sqrt{\frac{\sigma_1 - \sigma_{r+1} - \epsilon_i}{\sigma_1 - \sigma_{r+1}}} \qquad \text{and} \qquad p_{r+1} = q_{r+1} = \sqrt{1 - p_1^2}.$$

When $\sigma_1 - \sigma_{r+1} < \epsilon_i$, the solution is simply $p_1 = q_1 = 0, p_{r+1} = q_{r+1} = 1$. Therefore the optimal value of $(\mathrm{P_{sub}})$ is $\max\{1 - \epsilon_i/(\sigma_1 - \sigma_{r+1}), 0\}$, plug this into (24) and we finish the proof.

G Proof of Proposition 14

We describe the following lemma before proceeding to the proof of Proposition 14.

▶ **Lemma 19** (Hausdorff error bound)**.** *Given $\widehat{\mathcal{A}} \subseteq \mathcal{A}$, there exists $X \in \operatorname{cone}(\widehat{\mathcal{A}})$ such that*

$$\|X - X^*\|_F \le \operatorname{dist}(\mathcal{S}_{\mathcal{A}}(X^*), \widehat{\mathcal{A}}) \cdot \sqrt{|\,\mathcal{S}_{\mathcal{A}}(X^*)|} \cdot \|X^*\|_F. \qquad (25)$$

Proof. Let $X^* = \sum_{a\in\mathcal{S}_\mathcal{A}(X^*)} c_a a, c_a > 0$. By the definition of the one-sided Hausdorff distance $\text{dist}(\,\cdot\,,\cdot\,)$, for any $a \in \mathcal{S}_\mathcal{A}(X^*)$, there exist a corresponding $\widehat{a} \in \widehat{\mathcal{A}}$ such that

$$\|\widehat{a} - a\|_F \le \text{dist}(\mathcal{S}_\mathcal{A}(X^*), \widehat{\mathcal{A}}).$$

Let $\widehat{X} = \sum_{a\in\mathcal{S}_\mathcal{A}(X^*)} c_a \widehat{a}$. It is straighforward to verify that $\widehat{X} \in \text{cone}(\widehat{\mathcal{A}})$ and

$$\|X - X^*\|_F \le \text{dist}(\mathcal{S}_\mathcal{A}(X^*), \widehat{\mathcal{A}}) \sum_{a\in\mathcal{S}_\mathcal{A}(X^*)} c_a \overset{(*)}{\le} \text{dist}(\mathcal{S}_\mathcal{A}(X^*), \widehat{\mathcal{A}})\sqrt{|\,\mathcal{S}_\mathcal{A}(X^*)|}\|X^*\|_F,$$

where $(*)$ follows from the orthonormal decomposition $x^* = \sum_{a\in\mathcal{S}_\mathcal{A}(X^*)} c_a a, c_a > 0$ and $\|X^*\|_F^2 = \sum c_a^2$ when our atomic set is the set of rank-one matrices. ◀

Proof of Proposition 14. By Lemma 19, we know that there exist $\widetilde{X}$ that satisfies (25). Then by the L-smoothness of f,

$$\begin{aligned} f(b - M\widetilde{X}) &\le f(b - MX^*) + \langle \nabla f(b - MX^*), M(X^* - \widetilde{X})\rangle + \frac{L}{2}\|M(X^* - \widetilde{X})\|_F^2 \\ &\le f(b - MX^*) + \|\nabla f(b - MX^*)\|_F\|M(X^* - \widetilde{X})\|_F + \frac{L}{2}\|M(X^* - \widetilde{X})\|_F^2. \end{aligned} \tag{26}$$

By the smoothness and convexity of f, we further have

$$\|\nabla f(b - MX^*) - \nabla f(0)\|_F^2 \le 2L(f(b - MX^*) - f(0)).$$

Note we assume $f(0) = 0$ and $\nabla f(0) = 0$, the above reduces to $\|\nabla f(b - MX^*)\|_F \le \sqrt{2L\alpha}$. Combining with (26), we obtain

$$\begin{aligned} f(b - M\widetilde{X}) &\le f(b - MX^*) + \sqrt{2L\alpha}\|M(X^* - \widetilde{X})\|_F + \frac{L}{2}\|M(X^* - \widetilde{X})\|_F^2 \\ &\le f(b - MX^*) + \sqrt{2L\alpha}\|M\|\,\text{dist}(\mathcal{S}_\mathcal{A}(X^*), \widehat{\mathcal{A}}) \cdot \sqrt{|\,\mathcal{S}_\mathcal{A}(X^*)|} \cdot \|X^*\|_F \\ &\qquad + \frac{L\|M\|^2}{2}\,\text{dist}(\mathcal{S}_\mathcal{A}(X^*), \widehat{\mathcal{A}})^2|\,\mathcal{S}_\mathcal{A}(X^*)|\|X^*\|_F^2, \end{aligned}$$

where the last inequality is by Lemma 19. Combining the above with Proposition 13 leads to the desired result. ◀

References

1 Zeyuan Allen-Zhu, Elad Hazan, Wei Hu, and Yuanzhi Li. Linear convergence of a frank-wolfe type algorithm over trace-norm balls. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6191–6200. Curran Associates Inc., 2017.

2 Mariette Annergren. Admm for ℓ_1 Regularized Optimization Problems and Applications Oriented Input Design for MPC, 2012. licentiate thesis, Stockholm, Sweden.

3 Aleksandr Y. Aravkin, James V. Burke, Dmitry Drusvyatskiy, Michael P. Friedlander, and Kellie J. MacPhee. Foundations of gauge and perspective duality. *SIAM J. Optim.*, 28(3):2406–2434, 2018.

4 Aleksandr Y. Aravkin, James V. Burke, Dmitry Drusvyatskiy, Michael P. Friedlander, and Scott Roy. Level-set methods for convex optimization. *Math. Program.*, 174(1-2):359–390, 2018.

5 Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Mach. Learn.*, 73(3):243–272, 2008.

6 Alper Atamtūrk and Andrés Gómez. Safe screening rules for ℓ_0-regression. In *Proceedings of International Conference on Machine Learning*. 2020. `https://arxiv.org/abs/2004.08773`.

7 Runxue Bao, Bin Gu, and Heng Huang. Fast oscar and owl with safe screening rules. In *Proceedings of International Conference on Machine Learning*. 2020. `https://arxiv.org/abs/2006.16433`.

8 Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

9 Ewout van den Berg and Michael P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comput.*, 31(2):890–912, 2008.

10 Ewout van den Berg and Michael P. Friedlander. Sparse optimization with least-squares constraints. *SIAM J. Optim.*, 21(4):1201–1229, 2011.

11 Ewout van den Berg, Michael P. Friedlander, Gilles Hennenfent, Felix J. Herrmann, Rayan Saab, and Özgür Yilmaz. Algorithm 890: Sparco: A testing framework for sparse reconstruction. *ACM Trans. Math. Softw.*, 35(4): article no. 29 (16 pages), 2008.

12 Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A fresh approach to numerical computing. *SIAM Rev.*, 59(1):65–98, 2017.

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

14 James V. Burke and Jorge J. Moré. On the identification of active constraints. *SIAM J. Numer. Anal.*, 25(5):1197–1211, 1988.

15 Emmanuel J. Candès, Yonina C. Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM J. Imaging Sci.*, 6(1):199–225, 2013.

16 Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. Assoc. Comput. Mach.*, 58(3): article no. 11 (37 pages), 2011.

17 Emmanuel J. Candès and Yaniv Plan. Matrix Completion With Noise. *Proc. IEEE*, 98(6):925–936, 2010.

18 Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, 2012.

19 Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

20 Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Commun. Pure Appl. Math.*, 66(8):1241–1274, 2013.

21 Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

22 Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

23 Lijun Ding, Alp Yurtsever, Volkan Cevher, Joel A. Tropp, and Madeleine Udell. An optimal-storage approach to semidefinite programming using approximate complementarity. *SIAM J. Optim.*, 31(4):2695–2725, 2021.

24 David L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

25 David L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm near-solution approximates the sparsest near-solution, 2006. technical report, Department of Statistics, Stanford University, Stanford.

26 Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination in sparse supervised learning. *Pac. J. Optim.*, 8(4):667–698, 2012.

27 Zhenan Fan, Halyun Jeong, Babhru Joshi, and Michael P. Friedlander. Polar Deconvolution of Mixed Signals. *IEEE Trans. Signal Process.*, 70:2713–2727, 2022.

28 Zhenan Fan, Halyun Jeong, Yifan Sun, and Michael P. Friedlander. Atomic decomposition via polar alignment: The geometry of structured optimization. *Found. Trends Optim.*, 3(4):280–366, 2020.

29 Zhenan Fan, Yifan Sun, and Michael P. Friedlander. Bundle methods for dual atomic pursuit. In *Asilomar Conference on Signals, Systems, and Computers (ACSSC 2019)*, pages 264–270. IEEE, 2019.

30 M. Fazel and J. Goodman. Approximations for partially coherent optical imaging systems, 1998. technical report.

31 Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Nav. Res. Logist.*, 3(1-2):95–110, 1956.

32 Michael P. Friedlander and Ives Macêdo. Low-rank spectral optimization via gauge duality. *SIAM J. Sci. Comput.*, 38(3):A1616–A1638, 2016.

33 Michael P. Friedlander and Paul Tseng. Exact regularization of convex programs. *SIAM J. Optim.*, 18(4):1326–1350, 2007.

34 Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):643–660, 2001.

35 Warren L. Hare. Identifying active manifolds in regularization problems. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, volume 49 of *Springer Optimization and Its Applications*, pages 261–271. Springer, 2011.

36 Warren L. Hare and Adrian S. Lewis. Identifying active constraints via partial smoothness and prox-regularity. *J. Convex Anal.*, 11(2):251–266, 2004.

37 Christoph Helmberg and Franz Rendl. A spectral bundle method for semidefinite programming. *SIAM J. Optim.*, 10(3):673–696, 2000.

38 Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Grundlehren. Text Editions. Springer, 2001.

39 Bin Hong, Weizhong Zhang, Wei Liu, Jieping Ye, Deng Cai, Xiaofei He, and Jie Wang. Scaling up sparse support vector machines by simultaneous feature and sample reduction. In *ICML'17: Proceedings of the 34th International*

Conference on Machine Learning, pages 4016–4025. JMLR, 2017.

40 Cho-Jui Hsieh and Peder A. Olsen. Nuclear norm minimization via active subspace selection. In *ICML'14: Proceedings of the 31st International Conference on International Conference on Machine Learning*, pages 575–583. JMLR, 2014.

41 Martin Jaggi. Revisiting Frank–Wolfe: Projection-free sparse convex optimization. In *ICML'13: Proceedings of the 30th International Conference on International Conference on Machine Learning*, pages 427–435. JMLR, 2013.

42 Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. unpublished manuscript, `https://home.ttic.edu/~shai/papers/KakadeShalevTewari09.pdf`, 2009.

43 Kwangmoo Koh, Seung-Jean Kim, and Stephen P. Boyd. A method for large-scale ℓ_1-regularized logistic regression. In *AAAI'07: Proceedings of the 22nd national conference on Artificial intelligence*. AAAI Press, 2007.

44 Nathan Krislock and Henry Wolkowicz. Explicit sensor network localization using semidefinite representations and facial reductions. *SIAM J. Optim.*, 20(5):2679–2708, 2010.

45 Zhaobin Kuang, Sinong Geng, and David Page. A screening rule for ℓ_1-regularized Ising model estimation. *Adv. Neural Inf. Process. Syst.*, 30:720–731, 2017. (NIPS 2017).

46 Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low rank representation. In *NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 612–620. Curran Associates Inc., 2011.

47 Jun Liu, Zheng Zhao, Jie Wang, and Jieping Ye. Safe screening with variational inequalities and its application to lasso. In *ICML'14: Proceedings of the 31st International Conference on International Conference on Machine Learning*, pages 289–297. JMLR, 2014.

48 Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, 11:2287–2322, 2010.

49 Nicolai Meinshausen and Peter Bühlmann. High dimensional graphs and variable selection with the lasso. *Ann. Stat.*, 34(3):1436–1462, 2006.

50 Eugène Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparse-group lasso. In *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 388–396. Curran Associates Inc., 2016.

51 Eugène Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparsity enforcing penalties. *J. Mach. Learn. Res.*, 18: article no. 128 (33 pages), 2017.

52 Julie Nutini, Mark Schmidt, and Warren L. Hare. "active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern? *Optim. Lett.*, 13(4):645–655, 2019.

53 François Oustry. A second-order bundle method to minimize the maximum eigenvalue function. *Math. Program.*, 89(1):1–33, 2000.

54 Anant Raj, Jakob Olbrich, Bernd Gärtner, Bernhard Schölkopf, and Martin Jaggi. Screening rules for convex problems. `https://arxiv.org/abs/1609.07478`, 2015.

55 Jasson D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *ICML'05: Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM Press, 2005.

56 R. Tyrrell Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematical Series*. Princeton University Press, 1970.

57 Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 58(1):267–288, 1996.

58 Joel A. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, pages 67–101. Birkhäuser/Springer, 2015.

59 Jie Wang, Jiayu Zhou, Jun Liu, Peter Wonka, and Jieping Ye. A safe screening rule for sparse logistic regression. In *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 1053–1061. MIT Press, 2014.

60 Jie Wang, Jiayu Zhou, Peter Wonka, and Jieping Ye. Lasso screening rules via dual polytope projection. In *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 1070–1078. Curran Associates Inc., 2013.

61 Zhen James Xiang, Yun Wang, and Peter J. Ramadge. Screening tests for lasso problems. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(5):1008–1027, 2017.

62 Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 68(1):49–67, 2006.

63 Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit. *J. Mach. Learn. Res.*, 18: article no. 166 (43 pages), 2018.