Safe-screening rules for atomic-norm regularization

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
mpf@cs.ubc.ca

Abstract

Safe-screening rules are algorithmic techniques meant to detect and safely discard unneeded variables during the solution process with the aim of accelerating computation. These techniques have been shown to be effective for one-norm regularized problems. This paper generalizes the safe-screening rule proposed by Ndiaye et al. [J. Mach. Learn. Res., 2017] to various optimization problem formulations with atomic-norm regularization. For nuclear-norm regularized problems in particular, it is shown that the proposed generalized safe-screening rule cannot safely reduce the problem size. In that case, approximation guarantees are developed that allows for reducing the problem size.

Keywords convex analysis, sparse optimization, safe screening.

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

1 Introduction

Atomic-sparse optimization problems are characterized by solutions that exhibit a notion of parsimony manifested as sparsity of the solution with respect to a known dictionary or atomic set. Formally, for a given atomic set $\mathcal{A} \subseteq \mathbb{R}^n$, an optimal solution x^* can be decomposed as

$$x^* = \sum_{a \in \mathcal{A}} c_a a, \quad c_a \geq 0, \tag{1}$$

where most of the coefficients c_a are zero. The archetypal example is a sparse vector, which is sparse with respect to the set of signed canonical unit vectors $\mathcal{A} = \{ \pm \boldsymbol{e}_1, \dots, \pm \boldsymbol{e}_n \}$. One-norm regularization is the standard approach to produce sparse solutions with respect to this atomic set. The atoms that participate nontrivially in the decomposition (1) represent latent structure in the solution. The notion of atomic sparsity is prevalent in machine learning [3, 21, 29, 35] and signal processing [7], and has been formalized in the context of inverse problems by Chandrasekaran et al. [8].

Safe-screening rules, originally proposed by El Ghaoui et al. [14], generally refer to approaches that correctly identify atoms that can be safely discarded without changing the optimal solution. These rules are increasingly used in optimization algorithms because of their empirical success for one-norm regularized problems. Generalized screening rules can accommodate a range of atomic sets. Ndiaye et al. [22] propose a screening rule based on monitoring the duality gap that provides an elegant and effective screening framework useful for one- and group-norm regularized problems.

We derive generalized gap-based safe-screening rules that apply to general atomic-norm regularization in various problem formulations. We also present a careful investigation of the gap-based safe-screening rule for the set of rank-one matrices. We show that this safe-screening rule still requires a full singular-value decomposition (SVD)—instead of partial SVD decomposition—even when the problem has a very low-rank solution. This result reveals the limitation of safe-screening rules that depends on the duality gap. As a remedy, we provide approximation guarantees of the gap-based screening rule based on a partial SVD. With minor modification, all of our results can be transferred to the atomic set of symmetric rank-one matrices.

2 Related work

The sequential safe-screening rule proposed by El Ghaoui et al. [14] for one-norm regularization provides a theoretical basis with which to identify variables that are zero at a solution. Subsequent proposals have aimed at new screening rules that work for specific regularization functions, which are typically polyhedral, such as one-norm and box-constrained regularization [4–6, 19, 20, 23, 25, 32–34, 36]. Among these, screening rules that depend on the duality gap have been shown to be effective for one- and group-norm regularized problems such as LASSO and group LASSO [23]. Zhou et al. [37] proposed the only safe-screening rule that applies to nuclear-norm regularized problems. That rule, however, requires strong assumptions on the iterate that may not hold even if the iterate is feasible and arbitrarily close to the optimal solution. More recently, Sun and Bach [28] extended the safe-screening rule to general atomic sets and studied the atom-identification property of the generalized conditional-gradient method. Their work, however, does not offer a safe screening rule that could save computation when applied to the nuclear-norm regularized problems.

Other screening rules in the literature, which are not necessarily safe, include those based on statistical notions [9, 10, 38] and heuristics [16, 30]. These approaches are tangential to our purpose and we do not discuss them further. Some of the techniques used in our analysis are related to the facial reduction strategy from Krislock and Wolcowicz [18].

3 Preliminaries

We introduce in this section the basic tools of convex analysis and atomic sparsity that serve as the cornerstone of our analysis. We make the blanket assumption that the atomic set $\mathcal{A} \subseteq \mathbb{R}^n$ is compact, and that the origin is contained in its convex hull. (We do not assume that $\mathcal{A}$ is convex.) The gauge function to the set $\mathcal{A}$ measure the magnitude of a function relative to that set.

▶ **Definition 1** (Gauge function). The gauge function with respect to $\mathcal{A}$ is defined as

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

The gauge function is always convex, nonnegative, and positively homogeneous. However, it's not necessarily a norm because it may not be symmetric (unless $\mathcal{A}$ is centrosymmetric), may vanish at points other than the origin, and is not necessarily finite valued (unless $\mathcal{A}$ contains the origin in its interior). Definition 1 makes explicit the role of a gauge function as a convex penalty for atomic sparsity. The support of a vector x is the collection of atoms $a \in \mathcal{A}$ that contribute positively in the decomposition described by (2).

▶ **Definition 2** (Atomic support). The atomic support for a point $x \in \mathbb{R}^n$ with respect to the set $\mathcal{A}$ is defined to be the set $\mathcal{S}_{\mathcal{A}}(x)$ that satisfies

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

The atomic set of signed 1-hot unit vectors $\mathcal{A} = \{\pm \boldsymbol{e}_i \mid i = 1, 2, \ldots, n\}$, for example, the support $\mathcal{S}_{\mathcal{A}}(x)$ coincides with the nonzero elements of x with the corresponding sign. The support function, defined below, is dual to the gauge function, and provides a key tool for identifying atoms associated with the support of a vector.

▶ **Definition 3** (Exposed faces and ϵ-exposed faces). The exposed face and ϵ-exposed face, respectively, of a set $\mathcal{A} \subseteq \mathbb{R}^n$ in the direction $z \in \mathbb{R}^n$ are defined by the sets

$$\mathcal{F}_{\mathcal{A}}(z) = \{\, a \in \mathcal{A} \mid \langle a, z \rangle = \sigma_{\mathcal{A}}(z) \,\}, \qquad \mathcal{F}_{\mathcal{A}}(z, \epsilon) = \{\, a \in \mathcal{A} \mid \langle a, z \rangle \geq \sigma_{\mathcal{A}}(z) - \epsilon \,\},$$

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 face coincides with the exposed face. We list in Table 1 commonly used atomic sets, their corresponding gauge and support functions, and atomic supports.

Atomic sparsity	$\mathcal{A}$	$\gamma_{\mathcal{A}}(x)$	$\mathcal{S}_{\mathcal{A}}(x)$	$\sigma_{\mathcal{A}}(z)$
non-negative	$\operatorname{cone}(\{\, e_1, \dots, e_n \,\})$	$\delta_{\geq 0}$	$\operatorname{cone}(\{\, e_i \mid x_i > 0 \,\})$	$\delta_{\leq 0}$
element-wise	$\{\, \pm e_1, \dots, \pm e_n \,\}$	$\|\cdot\|_1$	$\{\, \operatorname{sign}(x_i) e_i \mid x_i \neq 0 \,\}$	$\|\cdot\|_\infty$
low rank	$\{\, uv^T \mid \|u\|_2 = \|v\|_2 = 1 \,\}$	nuclear-norm	singular vectors of x	spectral norm
PSD & low rank	$\{\, uu^T \mid \|u\|_2 = 1 \,\}$	$\operatorname{tr} + \delta_{\succeq 0}$	eigenvectors of x	$\max\{\, \lambda_{\max}, 0 \,\}$

Table 1 Commonly used sets atom sets and the corresponding gauge and support functions [12]. 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; see [11, Example 4.7].

4 Problem setting

Our safe-screening rules apply to these three related atomic-regularized optimization formulations:

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

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

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

where f is an L-smooth and convex function, $M : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator, and $b \in \mathbb{R}^m$ is a known vector. It's well known that under mild conditions, these three formulations are equivalent to each other for appropriate choice of the positive parameters λ, τ, and α [13]. Practitioners often prefer one of these formulations depending on their application. For example, tasks related to machine learning, including feature selection and recommender system, typically feature one of the first two formulations [21, 29, 35]. On the other hand, applications in signal processing and related fields, such as as compressed sensing and phase retrieval, usually use the third formulation [7, 31]. Previous work on safe screening rules usually focus on formulation (P$_1$). The safe screening rules we propose apply equally to all three formulations.

The Fenchel-Rockafellar duals for problems (P$_1$)–(P$_3$) play an important role in our screening rules:

$$\underset{y}{\text{minimize}} \quad d_1(y) := f^*(y) - \langle b, y \rangle \quad \text{s.t.} \quad \sigma_{\mathcal{A}}(M^* y) \leq \lambda, \tag{D$_1$}$$

$$\underset{y}{\text{minimize}} \quad d_2(y) := f^*(y) - \langle b, y \rangle + \tau \sigma_{\mathcal{A}}(M^* y), \tag{D$_2$}$$

$$\underset{y}{\text{minimize}} \quad \inf_{\beta > 0} d_3(y, \beta) := \beta \left(f^*\left(y/\beta \right) + \alpha \right) - \langle b, y \rangle \quad \text{s.t.} \quad \sigma_{\mathcal{A}}(M^* y) \leq 1, \tag{D$_3$}$$

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 $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. For each problem $i \in \{\, 1, 2, 3 \,\}$, denote the duality gap function by $p_i + d_i$, where p_i and d_i, respectively, denote the primal and dual objectives evaluated at feasible primal and dual variables. For the remainder of the paper, we assume strong duality holds for all three pairs of problems—that is, there exist pairs (x^*, y^*) and (x^*, y^*, β^*), in the case of the third formulation—that achieve the optimal primal and dual values. Denote the optimal primal and dual solutions for problems $i \in \{\, 1, 2, 3 \,\}$ by x_i^* and y_i^*, respectively. (Problem (P$_3$) also includes the optimal scalar variable β^*).

5 The gap-based safe-screening rule

The task of developing safe screening rules that apply with full generality to any atomic set $\mathcal{A} \subseteq \mathbb{R}^n$—even those that are uncountably infinite—requires tools that accommodate general notions of "activity", in the same way, for example, that simplex multipliers tell us which primal variables of a linear program are positive or zero. Our screening rules use information about the faces of the atomic sets that are exposed through the dual problem, and shows which atoms support an optimal solution. The following result, due to Fan et al. [11, Proposition 4.5 and Theorem 5.1], makes this precise.

Theorem 4 (Support identification)**.** *Let x^* and y^* be optimal primal-dual solutions for problems (P$_i$) and (D$_i$), with $i = 1, 2, 3$. Then*

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

Our main theorem generalizes this result to show similar atomic support identification properties that also apply to approximate primal and dual solutions. In particular, given a feasible dual variable y close to y^*, the support of x^* is contained in an ϵ-exposed face that includes $\mathcal{F}_{\mathcal{A}}(M^*y^*)$. Our results tie the parameter ϵ to the duality gap. Define the atomic operator norm by $\|M\|_{\mathcal{A}} := \max_{a\in\mathcal{A}} \|Ma\|_2$.

Theorem 5 (Generalized gap-based safe-screening rules)**.** *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*

$$\mathcal{S}_{\mathcal{A}}(x_i^*) \subseteq \mathcal{F}_{\mathcal{A}}(M^*y_i, \epsilon_i), \tag{3}$$

where each ϵ_i is defined for problem i as follows:

a) $\epsilon_1 = \|M\|_{\mathcal{A}}\sqrt{2L\,(p_1(x_1) + d_1(y_1))}$,
b) $\epsilon_2 = 2\|M\|_{\mathcal{A}}\sqrt{2L\,(p_2(x_2) + d_2(y_2))}$,
c) $\epsilon_3 = 2\|M\|_{\mathcal{A}}\sqrt{2\bar{\beta}L(p_3(x_3) + \max\{\, d_3(y_3, \underline{\beta}), d_3(y_3, \bar{\beta})\,\})}$,

where $\underline{\beta}$ and $\bar{\beta}$ are positive lower and upper bounds, respectively, for β^*.

Theorem 5(a) recovers the gap safe-screening rule developed by Ndiaye et al. [23], which applies only to (P_1) with $\gamma_{\mathcal{A}}$ being the one-norm. Note that Theorem 5(a) and Theorem 5(b) also overlap with recent work by Sun and Bach [28, Theorem 2] and we do not claim much novelty for them. But to the best of our knowledge, Theorem 5(c) is a novel result. We also note that the atomic operator norm $\|M\|_{\mathcal{A}}$ may be difficult to compute, and we show in Appendix E how to compute this term for the LASSO, matrix completion, and phase-retrieval problems.

The generalization of the safe-screening rule to (P_3) is more involved than for the first two problems (P_1) and (P_2) because the dual objective contains the perspective map of $f^*(\cdot) + \alpha$ [1]. The proofs for parts (a) and (b) of Theorem 5 depend on the strong convexity of f^*, implied by the Lipschitz smoothness of f. This convenient property, however, does not hold for the perspective map applied to f^*. We resolve this problem by imposing the additional assumption that we have bounds 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. [2].

All three screening rules stated in Theorem 5 are based on having available primal and dual variables. Let $\widehat{x}$ denote an arbitrary feasible primal variable. We can then construct a corresponding feasible dual variable $\widehat{y}$ via some scaling of $\nabla f(b - M\widehat{x})$, which ensures dual feasibility. By this construction, if $\widehat{x}$ converges to a solution of the primal problem, then $\widehat{y}$ also converges to a solution of the dual problem. Our next proposition shows the effectiveness of Theorem 5.

▶ **Proposition 6** (Atomic identification)**.** *For each problem $i = 1, 2, 3$, let $\{x_i^{(t)}\}_{t=1}^{\infty}$ and $\{y_i^{(t)}\}_{t=1}^{\infty}$ be sequences that converge to optimal primal and dual solutions (x_i^*, y_i^*) respectively. For* (D_3)*, let $\{\underline{\beta}^{(t)}\}_{t=1}^{\infty}$ and $\{\bar{\beta}^{(t)}\}_{t=1}^{\infty}$ be positive sequences that satisfy $\beta^* \in (\underline{\beta}^{(t)}, \bar{\beta}^{(t)})$ for all t and $\bar{\beta}^{(t)} - \underline{\beta}^{(t)} \to 0$. Let $\{\epsilon_i^{(t)}\}_{t=1}^{\infty}$ be the gaps defined in Theorem 5, evaluated at $x_i^{(t)}$ and $y_i^{(t)}$ (and $\underline{\beta}^{(t)}, \bar{\beta}^{(t)}$). Then the intersection $\mathcal{A}_i^{(t)} := \cap_{j=1}^{t}\mathcal{F}_{\mathcal{A}}(M^*y_i^{(j)}, \epsilon_i^{(j)})$ has the Painleveé-Kuratowski set limit [27, p. 111]*

$$\lim_{t\to\infty} \mathcal{A}_i^{(t)} = \mathcal{F}_{\mathcal{A}}(M^*y_i^*).$$

Proposition 6 ensures that the screening rule (3) is guaranteed to eventually discard superfluous atoms as long as we have available an iterative solver that can generate primal iterates that converge to a solution. For polyhedral atomic set, e.g., an atomic set with finite elements, it's straightforward to verify that Proposition 6 implies the following finite-time atom identification property: for $i = 1, 2, 3$,

$$\exists\; T > 0 \quad \text{such that} \quad \mathcal{A}_i^{(t)} = \mathcal{F}_{\mathcal{A}}(M^*y_i^*) \quad \forall t > T.$$

The implementation of the gap-based safe-screening rule for polyhedral atomic sets is also straightforward. One can store all atoms in memory during computation, and the gap-based safe-screening rule offers a computable way to discard redundant atoms periodically during the optimization. When $\mathcal{F}_{\mathcal{A}}(M^*y, \epsilon)$ is small enough, let $\widehat{\mathcal{A}} := \mathcal{F}_{\mathcal{A}}(M^*y, \epsilon) := \{\hat{a}_i\}_{i=1}^{r}$, we can solve efficiently the reduced low-dimensional problem

$$\underset{x\in\mathbb{R}^r, x\geq 0}{\text{minimize}} \;\; f\left(b - M\sum_{i=1}^{r}\hat{a}_i x_i\right) \qquad \text{subject to} \qquad \sum_{i=1}^{r} x_i \leq \tau$$

instead of the original high-dimensional problem using an algorithm such as accelerated projected-gradient descent.

A remarkable aspect of the gap-based safe-screening rule is that it depends solely on the duality gap, and hence is algorithm agnostic. As long as we have an algorithm that guarantees duality gap converges to 0, the gap-based safe-screening rule will recover $\mathcal{F}_{\mathcal{A}}(M^*y^*)$ within a finite number of iterations (for finite atomic sets). The gap-based safe-screening rule has been successfully applied to algorithms such as conditional gradient descent and projected coordinate descent to achieve promising performance.

6 Gap-based safe screening rule for nuclear norm

A key question in this work is whether the generalized safe screening rule can give any computational advantage for atomic sets $\mathcal{A}$ with infinite number of atoms. In particular we consider the atomic set to be the set of rank-one matrices, i.e.,

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

In the following proposition, we show that the ϵ-exposed face of M^*y contains all the singular vectors when ϵ is strictly positive.

▶ **Proposition 7** (Limitation of ϵ-Face). *Let $M^*y = U\Sigma V^T$ be the full SVD of M^*y, where*

$$\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{\min\{n,m\}}), \sigma_i \geq 0\ \forall\ 0 < i \leq \min\{n, m\}.$$

For any $\epsilon \geq 0$, the ϵ-face can be explicitly expressed as

$$\mathcal{F}_{\mathcal{A}}(M^*y, \epsilon) = \left\{ (Up)(Vq)^T \;\middle|\; \sum_{i=1}^{\min\{n,m\}} \sigma_i p_i q_i \geq \sigma_1 - \epsilon,\ \|p\|_2 = \|q\|_2 = 1 \right\}.$$

Then for any $\epsilon > 0$, there exist p, q with all entries being nonzero, such that $(Up)(Vq)^T \in \mathcal{F}_{\mathcal{A}}(M^*y, \epsilon)$.

Proposition 7 indicates that the ϵ-exposed face $\mathcal{F}_{\mathcal{A}}(M^*y, \epsilon)$ contains not only the top singular vectors of M^*y but also the bottom singular vectors—even if ϵ is arbitrarily close to 0. This result is unfortunate since the gap-based safe-screening rules stated in Theorem 5 do not allow us to discard any singular vectors of M^*y and thus require a full SVD of M^*y even if the duality gap is arbitrarily close to 0.

6.1 Approximation with partial SVD

The face of $\mathcal{A}$ exposed by the vector M^*y^* is given by

$$\mathcal{F}_{\mathcal{A}}(M^*y^*) = \{\, uv^T \mid u^T(M^*y^*)v = \sigma_1(M^*y^*) \,\},$$

where $\sigma_1(M^*y^*)$ is the largest singular value of M^*y^*. Therefore, when there are few singular vectors associated with the largest singular value, only the top few singular vectors of M^*y^* are actual useful atoms. This property motivates us to use the partial SVD of M^*y to extract the reduced atomic set. This hard-thresholding technique has been widely used as a heuristic. Formally, given a dual feasible solution y with partial singular value decomposition

$$M^*y = U_r\Sigma_r V_r^T \quad U_r \in \mathbb{R}^{n\times r}, \quad V_r \in \mathbb{R}^{m\times r}, \quad r \ll \min\{n, m\},$$

we construct the corresponding reduced atomic set

$$\widehat{\mathcal{A}} = \{\, U_r p q^T V_r^T \mid \|p\|_2 = \|q\|_2 = 1 \,\},$$

and solve the reduced problem over $\widehat{\mathcal{A}}$.

First, we give a concrete example showing that the partial SVD of M^*y is not able to give us a safe cover of $\mathcal{F}_{\mathcal{A}}(M^*y^*)$ even when $\mathcal{F}_{\mathcal{A}}(M^*y^*)$ is a singleton and y arbitrarily close to y^*.

▶ **Example 8** (Limitation of Partial SVD)**.** Consider the problem

$$\underset{X\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \tfrac{1}{2}\|X-Z\|_F^2 \quad \text{subject to} \quad \|X\|_* \le 1, \tag{4}$$

where

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

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

$$\underset{Y\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \tfrac{1}{2}\|Y-Z\|_F^2 - \tfrac{1}{2}\|Z\|_F^2 + \|Y\|_2. \tag{5}$$

The solutions for the dual pair (4) and (5) are

$$X^* = U\,\text{diag}(1, 0, \dots, 0)V^T \quad \text{and} \quad Y^* = Z - X^* = U\,\text{diag}(1, 0.1, \dots, 0.1)V^T.$$

Let u_1 and v_1, respectively, be the first columns of U and V. Then obviously $\mathcal{S}_{\mathcal{A}}(X^*) = \mathcal{F}_{\mathcal{A}}(Y^*) = \{u_1 v_1^T\}$ is a singleton. We construct the following dual feasible solution

$$\widehat{Y} = \text{diag}(1, 0.1, \dots, 0.1).$$

Let $\widehat{U}$ and $\widehat{V}$ be the singular vectors of $\widehat{Y}$, then $\widehat{U} = \widehat{V} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_n]$, where $\boldsymbol{e}_i$ is the n-dimensional vector with 1 at the ith entry and 0 at other entries. In order to cover the leading singular vector $u_1 = [\sqrt{1-\epsilon}, 0, \dots, \sqrt{\epsilon}]^T$, we need both the top singular and bottom singular vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_n$, and therefore any top-r SVD of $\widehat{Y}$ with $r < n$ will end up to be "unsafe". It's also easy to verify that $\|\widehat{Y} - Y^*\|_F = \mathcal{O}(\sqrt{\epsilon})$. Note that our argument holds for any $\epsilon \in (0,1)$. Therefore, $\forall \epsilon \in (0,1)$ there exists $y \in \mathbb{B}(y^*, \epsilon)$ such that only full SVD of M^*y can guarantee a safe coverage of $\mathcal{F}_{\mathcal{A}}(M^*y^*)$. ◀

This result shows that the screening rule with partial SVD is not safe. Therefore, we can only resort to an approximate screening rule. We use the one-sided Hausdorff distance

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

to measure the similarity between any two subsets $\mathcal{A}_1$ and $\mathcal{A}_2$ of the atomic set $\mathcal{A}$. The next result shows that there is a set $\widehat{\mathcal{A}}$ that is close to $\mathcal{S}_{\mathcal{A}}(x^*)$, then there must exist a point in $\widehat{\mathcal{A}}$ that is close to x^*.

▶ **Proposition 9** (Hausdorff error bound)**.** *Given $\widehat{\mathcal{A}} \subseteq \mathcal{A}$, there exists $x \in \text{cone}(\widehat{\mathcal{A}})$ such that*

$$\|x - x^*\|_F \le \rho(\mathcal{S}_{\mathcal{A}}(x^*), \widehat{\mathcal{A}}) \cdot \sqrt{|\,\mathcal{S}_{\mathcal{A}}(x^*)|} \cdot \|x^*\|_F.$$

Next, we study the approximation ability of the partial SVD of a given feasible dual solution M^*y to $\mathcal{F}_{\mathcal{A}}(M^*y^*)$. The next result applies to each one of the dual pairs (P$_i$) and (D$_i$), $i = 1, 2, 3$.

▶ **Proposition 10** (Error in partial SVD)**.** *Let y be a dual feasible vector. Let $M^*y = U_r\Sigma_r V_r^T$, $U_r \in \mathbb{R}^{n\times r}$, and $V_r \in \mathbb{R}^{m\times r}$ be the truncated SVD where $r < \min\{n, m\}$. Let $\{\sigma_i\}_{i=1}^{\min\{n,m\}}$ be the singular values and $\widehat{\mathcal{A}} = \{\, U_r p q^T V_r^T \mid \|p\|_2 = \|q\|_2 = 1\,\}$ be the reduced atomic set. Assume $\sigma_1 > \sigma_{r+1}$. Then*

$$\rho(\mathcal{F}_{\mathcal{A}}(M^*y^*), \widehat{\mathcal{A}}) \le \rho(\mathcal{F}_{\mathcal{A}}(M^*y, \epsilon), \widehat{\mathcal{A}}) = \sqrt{2\min\left\{\frac{\epsilon}{\sigma_1 - \sigma_{r+1}}, 1\right\}},$$

where $\epsilon = \epsilon_i$ for $i = 1, 2, 3$ as defined in Theorem 5 depending on the problem formulation.

Oustry [24, 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 that is $\mathcal{O}(\sqrt{\epsilon/(\sigma_r - \sigma_{r+1})})$, which is looser than the bound shown in Proposition 10 because $\sigma_1 \ge \sigma_r \ge \sigma_{r+1}$.

7 Conclusion

Our extension of gap-based safe-screening rules to the various forms of atomic-norm regularization is based on the convex calculus of sublinear functions. Our proposed screening rules can provide practical computational advantages when the atomic sets are polyhedral. As demonstrated by Example 8, however, there are limitations of the rule when used for non-polyhedral atomic sets. In that case, Proposition 10 provides an error bound based on the truncated SVD.

Further research opportunities remain, particularly for designing meaningful safe-screening rules for non-polyhedral sets. For example, it seems possible to design safe-screening rules for nuclear-norm regularized problems that are particular to the search directions generated by the conditional-gradient method.

References

1 A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and K. MacPhee. Foundations of gauge and perspective duality. *SIAM J. Optim.*, 28(3):2406–2434, 2018.

2 A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and S. Roy. Level-set methods for convex optimization. *Math. Program., Ser. B*, 174(1-2):359–390, December 2018.

3 Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73, 2008.

4 Alper AtamtÃ«rk and Andrés Gómez. Safe screening rules for ℓ_0-regression. In *Proceedings of ICML*, 2020.

5 Runxue Bao, Bin Gu, and Heng Huang. Fast oscar and owl with safe screening rules. In *Proceedings of ICML*, 2020.

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

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

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

9 Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.*, 106(494):544–557, 2011.

10 Jianqing Fan and Rui Song. Sure independence screening in generalized linear models with np-dimensionality. *Annals of Statistics*, 38(6):3567–3604, 2010.

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

12 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.

13 M. P. Friedlander and P. Tseng. Exact regularization of convex programs. Technical Report TR-2006-26, Department of Computer Science, University of British Columbia, Vancouver, BC, Canada, November 2006. To appear in *SIAM J. Optim.*

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

15 J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis.* Springer, New York, NY, 2001.

16 T. Joachims. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.

17 Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09. pdf*, 2(1), 2009.

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

19 Zhaobin Kuang, Sinong Geng, and David Page. A screening rule for ℓ_1-regularized ising model estimation. *Advances in neural information processing systems (NIPS 2017)*, 30:720, 2017.

20 Jun Liu, Zheng Zhao, Jie Wang, and Jieping Ye. Safe screening with variational inequalities and its application to lasso. In *International Conference on Machine Learning*, pages 289–297. PMLR, 2014.

21 Nicolai Meinshausen and Peter BÃ¼hlmann. High dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34, 09 2006.

22 Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparse-group lasso. In *Advances in Neural Information Processing Systems (NIPS 2016)*, pages 388–396, 2016.

23 Eugène Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparsity enforcing penalties. *J. Mach. Learn. Res.*, 18:128:1–128:33, 2017.

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

25 Anant Raj, Jakob Olbrich, Bernd Gärtner, Bernhard Schölkopf, and Martin Jaggi. Screening rules for convex problems. *ArXiv*, abs/1609.07478, 2015.

26 R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

27 R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*, volume 317. Springer, 1998. 3rd printing.

28 Yifan Sun and Francis R. Bach. Safe screening for the generalized conditional gradient method. *ArXiv*, abs/2002.09718, 2020.

29 Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Ser. B*, pages 267–288, 1996.

30 Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.

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

32 Jie Wang, Jiayu Zhou, Jun Liu, Peter Wonka, and Jieping Ye. A safe screening rule for sparse logistic regression. In *Advances in Neural Information Processing Systems (NIPS 2014)*, pages 1053–1061, 2014.

33 Jie Wang, Jiayu Zhou, Peter Wonka, and Jieping Ye. Lasso screening rules via dual polytope projection. In *Advances in Neural Information Processing Systems (NIPS 2013)*, pages 1070–1078. Citeseer, 2013.

34 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.

35 M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Royal Stat. Soc. B.*, 68, 2006.

36 Weizhong Zhang, Bin Hong, Wei Liu, Jieping Ye, Deng Cai, Xiaofei He, and Jie Wang. Scaling up sparse support vector machines by simultaneous feature and sample reduction. In *Proceedings of ICML*, 2017.

37 Qiang Zhou and Qi Zhao. Safe subspace screening for nuclear norm regularized least squares problems. In *Proceedings of ICML*, pages 1103–1112, 2015.

38 Li-Ping Zhu, Lexin Li, Runze Li, and Li-Xing Zhu. Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.*, 106(496):1464–1475, 2011.

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 11 ([26, 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 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 12** ([11, 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

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

By the properties of conjugate functions and Proposition 12, we obtain

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

Then by Theorem 11, we can get the dual problem for (P$_1$) as

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

For (P$_2$),

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

By the properties of conjugate functions and Proposition 12, we obtain

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

Then by Theorem 11, 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$_3$),

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

By the properties of conjugate functions and Proposition 12, we can get that

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

Then by [15, 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 11, we can get the dual problem for (P$_3$) as

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

B Proof of Theorem 5

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

▶ **Lemma 13** ([17, 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 13, 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

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

Therefore, by reordering the inequality,

$$\|y - y^*\| \leq \sqrt{2L(g(y) - g(y^*))} \leq \sqrt{2L\left(p_1(x) + d_1(y)\right)} \quad \forall x \in \mathbb{R}^n, \tag{6}$$

where the last inequality follows from the fact that the duality gap is always an upper bound for $g(y) - g(y^*)$. Next, we show that $\mathcal{F}_{\mathcal{A}}(M^*y^*) \subseteq \mathcal{F}_{\mathcal{A}}(M^*y, \epsilon_1)$. For any $a \in \mathcal{F}_{\mathcal{A}}(M^*y^*)$,

$$\begin{aligned}\langle a, M^*y\rangle &= \sigma_{\mathcal{A}}(M^*y^*) + \langle Ma, y - y^*\rangle\\ &\geq \sigma_{\mathcal{A}}(M^*y^*) - \left(\max_{a\in\mathcal{A}}\|Ma\|\right)\|y - y^*\|\\ &\geq \sigma_{\mathcal{A}}(M^*y^*) - \left(\max_{a\in\mathcal{A}}\|Ma\|\right)\sqrt{2L\left(p_1(x) + d_1(y)\right)}\\ &\geq \sigma_{\mathcal{A}}(M^*y) - \epsilon_1,\end{aligned}$$

where the last inequality follows from the fact that $\sigma_{\mathcal{A}}(M^*y^*) = \lambda$.

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 13, 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, $0 \in \partial g(y^*)$. Reorder the inequality to deduce that

$$\begin{aligned}\|y - y^*\|_2 &\leq \sqrt{2L(g(y) - g(y^*))}\\ &\leq \sqrt{2L\left(p_2(x) + d_2(y)\right)} \quad \forall x \in \tau\mathcal{A}.\end{aligned} \tag{7}$$

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

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

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 13, f^* is $\frac{1}{L}$-strongly convex, and it's 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,

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

Reorder the inequality to deduce that

$$\begin{aligned} \|y - y^*\|_2 &\leq \sqrt{2\beta^* L(g(y) - g(y^*))} \\ &\leq \sqrt{2\beta^* L\left(p_3(x) + d_3(y, \beta^*)\right)} \;\; \forall x \in \mathbb{R}^n \quad \text{s.t.} \quad f(b - Mx) \leq \alpha \\ &\leq \sqrt{2\overline{\beta} L\left(p_3(x) + d_3(y, \beta^*)\right)} \;\; \forall x \in \mathbb{R}^n \quad \text{s.t.} \quad f(b - Mx) \leq \alpha. \end{aligned} \tag{8}$$

Since 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

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

Therefore,

$$\|y - y^*\|_2 \leq \sqrt{2\overline{\beta} L\left(p_3(x) + \max\left\{ d_3(y, \underline{\beta}), d_3(y, \overline{\beta}) \right\}\right)} \;\; \forall x \in \mathbb{R}^n \quad \text{s.t.} \quad f(b - Mx) \leq \alpha.$$

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

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

◀

C Upper and lower bound for β^*

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

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

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

$$\underset{w}{\text{minimize}} \;\; f^*(w) - \langle b, w \rangle \quad \text{subject to} \quad \sigma_{\mathcal{A}}(M^* w) \leq \beta^*. \tag{9}$$

Now compare (9) with (D$_1$) to conclude that they are equivalent when $\lambda = \beta^*$. It thus follows that (P$_3$) is equivalent to

$$\underset{x}{\text{minimize}} \;\; f(b - Mx) + \beta^* \gamma_{\mathcal{A}}(x). \tag{10}$$

Next, consider using the level-set method [2] with bisection to solve (P$_3$). There exists $\tau^* > 0$ such that (P$_3$) is equivalent to

$$\underset{x}{\text{minimize}} \;\; f(b - Mx) \quad \text{subject to} \quad \gamma_{\mathcal{A}}(x) \leq \tau^*. \tag{11}$$

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

$$\underset{x}{\text{minimize}} \quad f(b - Mx) \quad \text{subject to} \quad \gamma_{\mathcal{A}}(x) \le \tau_i, \tag{12}$$

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

$$\underset{x}{\text{minimize}} \quad f(b - Mx) + \beta_i \gamma_{\mathcal{A}}(x), \tag{13}$$

for $i = 1, 2$.

Finally, by [11, Theorem 5.1] we can conclude that

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

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 6

Proof. From the construction of sets $\mathcal{A}^{(t)}$, it's straightforward to see that

$$\mathcal{A}^{(1)} \supseteq \mathcal{A}^{(2)} \supseteq \dots,$$

which shows that $\{\mathcal{A}^{(t)}\}_{t=1}^{\infty}$ is a monotone sequence. By [27, Exercise 4.3], the Painleveé-Kuratowski set limit

$$\mathcal{A}^{(\infty)} = \lim_{t \to \infty} \mathcal{A}^{(t)}$$

is well-defined.

First, we show that $\mathcal{F}_{\mathcal{A}}(M^* y^*) \subseteq \mathcal{A}^{(\infty)}$. By Theorem 5, we know that $\mathcal{F}_{\mathcal{A}}(M^* y^*) \subseteq \mathcal{A}^{(t)}$ for all t. Therefore, it follows that $\mathcal{F}_{\mathcal{A}}(M^* y^*) \subseteq \mathcal{A}^{(\infty)}$.

Next, we show that $\mathcal{A}^{(\infty)} \subseteq \mathcal{F}_{\mathcal{A}}(M^* y^*)$. Consider $a \in \mathcal{A}^{(\infty)}$. Since $\{\mathcal{A}^{(t)}\}_{t=1}^{\infty}$ is a monotone sequence, there exist $T > 0$ such that

$$a \in \mathcal{A}^{(t)}, \quad \forall t \ge T.$$

By the construction of $\mathcal{A}^{(t)}$, we know that $\mathcal{A}^{(t)} \subseteq \mathcal{F}_{\mathcal{A}}(M^* y^{(t)}, \epsilon^{(t)})$ for all t, and thus we can conclude that

$$\langle a, M^* y^{(t)} \rangle \ge \sigma_{\mathcal{A}}(M^* y^{(t)}) - \epsilon^{(t)}, \quad \forall t \ge T.$$

Now by taking limits with respect to t to both sides of the inequality, we can conclude that

$$\langle a, M^* y^* \rangle \ge \sigma_{\mathcal{A}}(M^* y^*),$$

which implies that $a \in \mathcal{F}_{\mathcal{A}}(M^* y^*)$. ◀

E Computing $\|M\|_{\mathcal{A}}$

In general, by the inequality $\|Ma\| \le \|M\|\|a\|$, we can bound the term $\sup_{a \in \mathcal{A}} \|Ma\|$ by $\|M\| \sup_{a \in \mathcal{A}} \|a\|$. Which is computable as long as we can compute $\|M\|$ and $\sup_{a \in \mathcal{A}} \|a\|$. For specific problem, it's possible to compute a tighter bound. We show how to compute the term $\|M\|_{\mathcal{A}}$ for LASSO, matrix completion and phase retrieval.

E.1 LASSO

In this case, $\mathcal{A} = \{\pm \boldsymbol{e}_1, \dots, \pm \boldsymbol{e}_n\}$. Let $M \in \mathbb{R}^{m \times n}$, then

$$\|M\|_{\mathcal{A}} = \max_{i \in [n]} \|M \boldsymbol{e}_i\| = \max_{i \in [n]} \|\boldsymbol{m}_i\|,$$

where $\boldsymbol{m}_i$ denotes the i-th column of the matrix M. Thus $\|M\|_{\mathcal{A}} = 1$ when the matrix M is column-wise normalized.

E.2 Matrix completion

In this case,

$$f(\mathcal{M}X - Y) = \frac{1}{2}\|\Pi_\Omega(X - Y)\|_F^2 = \frac{1}{2}\|\Pi_\Omega(X) - Y\|_F^2,$$

where $X \in \mathbb{R}^{m\times n}$, the second equality is true since Y is a sparse matrix with nonzero entries on Ω. Then $f(\cdot) = 1/2\|\cdot\|_F^2$ and $\mathcal{M}X = \Pi_\Omega X$.

$$\|M\|_{\mathcal{A}} = \sup_{\|\boldsymbol{u}\|=\|\boldsymbol{v}\|=1} \|\Pi_\Omega(\boldsymbol{u}\boldsymbol{v}^\intercal)\|_F \le \sup_{\|\boldsymbol{u}\|=\|\boldsymbol{v}\|=1} \|\boldsymbol{u}\boldsymbol{v}^\intercal\|_F \overset{\text{(i)}}{=} \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(\boldsymbol{u}\boldsymbol{v}^\intercal)} \overset{\text{(ii)}}{=} 1,$$

where $\sigma_i(X)$ denotes the i-th singular values of the matrix X. (i) is from the definition of Frobenius norm. (ii) is true since $\sigma_1(\boldsymbol{u}\boldsymbol{v}^\intercal) = 1$, and $\sigma_i(\boldsymbol{u}\boldsymbol{v}^\intercal) = 0\ \forall i \ge 2$.

E.3 Phase retrieval

In this case,

$$\mathcal{M}X = \left[\langle M_i, X\rangle\right]_{i=1:m},$$

where m is the number of measurements and let $X \in \mathbb{R}^{n\times n}$. Then

$$\begin{aligned}
\|\mathcal{M}\|_{\mathcal{A}} &= \sup_{\|\boldsymbol{u}\|=1} \left\| \left[\langle M_i, \boldsymbol{u}\boldsymbol{u}^\intercal\rangle\right]_{i=1:m} \right\| \\
&= \sup_{\|\boldsymbol{u}\|=1} \left\| \left[\operatorname{vec}(M_i)^\intercal\right]_{i=1:m} \operatorname{vec}(\boldsymbol{u}\boldsymbol{u}^\intercal) \right\| \\
&\le \left\| \left[\operatorname{vec}(M_i)^\intercal\right]_{i=1:m} \right\| \sup_{\|\boldsymbol{u}\|=1} \|\boldsymbol{u}\boldsymbol{u}^\intercal\|_F \\
&\le \left\| \left[\operatorname{vec}(M_i)^\intercal\right]_{i=1:m} \right\|.
\end{aligned}$$

F Proof of Proposition 7

Proof. By the definition of $\mathcal{F}_{\mathcal{A}}(M^*y, \epsilon)$,

$$\begin{aligned}
\mathcal{F}_{\mathcal{A}}(M^*y, \epsilon) &= \{\, uv^T \mid \langle uv^T, M^*y\rangle \ge \sigma_1 - \epsilon,\ \|u\|_2 = \|v\|_2 = 1 \,\} \\
&= \{\, uv^T \mid \langle uv^T, U\Sigma V^T\rangle \ge \sigma_1 - \epsilon,\ \|u\|_2 = \|v\|_2 = 1 \,\}.
\end{aligned}$$

We know that U, V are orthonormal matrices, by setting $u = Up$ and $v = Vp$ for some $p, q \in \mathbb{R}^{\min\{n,m\}}$, $\|p\|_2 = \|q\|_2 = 1$, we obtain

$$\begin{aligned}
\mathcal{F}_{\mathcal{A}}(M^*y, \epsilon) &= \{\, Up(Vq)^T \mid \langle Up(Vq)^T, U\Sigma V^T\rangle \ge \sigma_1 - \epsilon,\ \|p\|_2 = \|q\|_2 = 1 \,\} \\
&= \{\, Up(Vq)^T \mid p^T\Sigma q \ge \sigma_1 - \epsilon,\ \|p\|_2 = \|q\|_2 = 1 \,\} \\
&= \left\{ Up(Vq)^T \;\middle|\; \sum_{i=1}^{\min\{n,m\}} \sigma_i p_i q_i \ge \sigma_1 - \epsilon, \|p\|_2 = \|q\|_2 = 1 \right\}.
\end{aligned}$$

The above finished the proof. ◀

G Proof of Proposition 9

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 ρ, for any $a \in \mathcal{S}_{\mathcal{A}}(x^*)$, there exist a corresponding $\hat{a} \in \widehat{\mathcal{A}}$ such that

$$\|\hat{a} - a\|_F \le \rho(\mathcal{S}_{\mathcal{A}}(x^*), \widehat{\mathcal{A}}).$$

Let $\hat{x} = \sum_{a\in\mathcal{S}_\mathcal{A}(x^*)} c_a \hat{a}$, then it's straighforward to verify that $\hat{x} \in \operatorname{cone}(\widehat{\mathcal{A}})$ and

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

where (i) 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. ◀

H Proof for Proposition 10

Proof. By the definition of $\rho(\cdot,\cdot)$, it follows that

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

We know that $\mathcal{F}_\mathcal{A}(M^*y^*) \subseteq \mathcal{F}_\mathcal{A}(M^*y, \epsilon)$, then obviously we have

$$\rho(\mathcal{F}_\mathcal{A}(M^*y^*), \widehat{\mathcal{A}}) \le \rho(\mathcal{F}_\mathcal{A}(M^*y, \epsilon), \widehat{\mathcal{A}}).$$

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

$$\rho(\mathcal{A}_1, \mathcal{A}_2) = \sqrt{\sup_{a_1\in\mathcal{A}_1} \inf_{a_2\in\mathcal{A}_2} \|a_1 - a_2\|_F^2} = \sqrt{2 - 2\left(\inf_{a_1\in\mathcal{A}_1} \sup_{a_2\in\mathcal{A}_2} \langle a_1, a_2\rangle\right)}, \tag{14}$$

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

$$\begin{aligned}
\rho(\mathcal{A}_1, \mathcal{A}_2) &= \sqrt{2 - 2\left(\min_{p,q\in\mathcal{C}_1} \max_{\hat{p},\hat{q}\in\mathcal{C}_2} \langle Upq^T V^T, U_r \hat{p}\hat{q}^T V_r^T\rangle\right)} \\
&= \sqrt{2 - 2\left(\min_{p,q\in\mathcal{C}_1} \max_{\hat{p},\hat{q}\in\mathcal{C}_2} \left(\sum_{i=1}^r p_i\hat{p}_i\right)\left(\sum_{i=1}^r q_i\hat{q}_i\right)\right)} \\
&= \sqrt{2 - 2\left(\min_{p,q\in\mathcal{C}_1} \|p_{1:r}\|_2\|q_{1:r}\|_2\right)}.
\end{aligned} \tag{15}$$

Now we consider the subproblem in (15):

$$\begin{aligned}
&\min_{p,q} \ \|p_{1:r}\|_2\|q_{1:r}\|_2 && (\text{P}_1) \\
&\text{subject to} \ \sum_{i=1}^k \sigma_i p_i q_i \ge \sigma_1 - \epsilon,\ \|p\|_2 = \|q\|_2 = 1,\ p, q \in \mathbb{R}^k.
\end{aligned}$$

If p^* and q^* is a solution of the problem (P_1), then it's easy to verify that

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

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

$$\begin{aligned}
&\min_{p_1,q_1,p_{r+1},q_{r+1}} \ p_1 q_1 \\
&\text{subject to} \ \sigma_1 p_1 q_1 + \sigma_{r+1} p_{r+1} q_{r+1} \ge \sigma_1 - \epsilon, \\
&\qquad\qquad p_1^2 + p_{r+1}^2 = q_1^2 + q_{r+1}^2 = 1,\ p_1, q_1, p_{r+1}, q_{r+1} \ge 0.
\end{aligned}$$

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

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

When $\sigma_1 - \sigma_{r+1} < \epsilon$, the solution is simply $p_1 = q_1 = 0, p_{r+1} = q_{r+1} = 1$. Therefore the optimal value of (P_1) is $\max\{1 - \epsilon/(\sigma_1 - \sigma_{r+1}), 0\}$. Subsitute this into eq. (15) to obtain the required result. ◀