Safe Element Screening for Submodular Function Minimization

Weizhong Zhang[1], Bin Hong[2], Lin Ma[1], Wei Liu[1], Tong Zhang[1]
[1]Tencent AI Lab, Shenzhen, China
[2]State Key Lab of CAD&CG, College of Computer Science, Zhejiang University

May 25, 2018

Abstract

Submodular functions are discrete analogs of convex functions, which have applications in various fields, including machine learning, computer vision and signal processing. However, in large-scale applications, solving Submodular Function Minimization (SFM) problems remains challenging. In this paper, we make the first attempt to extend the emerging technique named screening in large-scale sparse learning to SFM for accelerating its optimization process. Specifically, we propose a novel safe element screening method—based on a careful studying of the relationships between SFM and the corresponding convex proximal problems, as well as the accurate estimation of the optimum of the proximal problem—to quickly identify the elements that are guaranteed to be included (we refer to them as active) or excluded (inactive) in the final optimal solution of SFM during the optimization process. By removing the inactive elements and fixing the active ones, the problem size can be dramatically reduced, leading to great savings in the computational cost without sacrificing accuracy. To the best of our knowledge, the proposed method is the first screening method in the fields of SFM and even combinatorial optimization, and thus points out a new direction for accelerating SFM algorithms. Experiment results on both synthetic and real datasets demonstrate the significant speedups gained by our screening method.

1 Introduction

Submodular Functions [10] are a special class of set functions, which have rich structures and a lot of links with convex functions. They arise naturally in many domains, such as clustering [18], image segmentation [13, 4], document summarization [14] and social networks [12]. Most of these applications can be finally deduced to a Submodular Function Minimization (SFM) problem, which takes the form of

$$\min_{A \subseteq V} F(A), \tag{SFM}$$

where $F(A)$ is a submodular function defined on a set $V = \{1, 2, ..., p\}$. The problem of SFM has been extensively studied for several decades in the literature [6, 16, 11, 17, 8], in which many algorithms (both exact and approximate) have been developed from the perspectives of combinatorial optimization and convex optimization. The most well known conclusion is that SFM is solvable in strongly polynomial time [11]. Unfortunately, due to the high-degree polynomial dependence, the applications of submodular functions on the large scale problems remain challenging, such as image segmentation [4] and speech analysis [15], which involve huge number of variables.

Algorithm 1 The framework of screening methods

1: Estimate the dual (resp. primal) optimum of the sparse model.
2: Based on the estimation above, infer which components of the primal (resp. dual) optimum are zeros from the KKT conditions.
3: Remove the features (resp. samples) corresponding to the identified components.
4: Train model on the reduced dataset.

Screening [7] is an emerging technique, which has been proved to be effective in accelerating large-scale sparse model training. It is motivated by the well known feature of sparse models that a significant portion of the coefficients in the optimal solutions of them (resp. their dual problems) are zero, that is, the corresponding features (resp. samples) are irrelevant with the final learned models. Screening methods aim to quickly identify these irrelevant features and/or samples and remove them from the datasets before or during the training process. Thus, the problem size can be reduced dramatically, leading to substantial savings in the computational cost. The framework of these methods is given in Algorithm 1. Since screening methods are always independent with the training algorithms, thus can be integrated with all the algorithms flexibly. In the recent few years, specific screening methods for most of the traditional sparse models have been developed, such as lasso [24, 27, 25], sparse logistic regression [26], multi-task learning[19] and SVM [21, 29]. Empirical studies indicate that the speedups they achieved can be orders of magnitudes.

The binary attribute (each element in V must be either in or not in the optimal solution) of SFM motivates us to introduce the key idea of screening into SFM to accelerate its optimization process. The most intuitive approach is to identify the elements that are guaranteed to be included or excluded in the minimizer A^* of SFM prior to or during actually solving it. Then, by fixing the identified active elements and removing the inactive ones, we just need to solve a small-scale problem. However, we note that existing screening methods are all developed for convex models and they can not be applied to SFM directly. The reason is that they all heavily depend on KKT conditions (see Algorithm 1), which do not exist in SFM problems.

In this paper, to improve the efficiency of SFM algorithms, we propose a novel **I**nactive and **A**ctive **E**lement **S**creening (IAES) framework for SFM, which consists of two kinds of screening rules, i.e., **I**nactive **E**lements **S**creening (IES) and **A**ctive **E**lements **S**creening (AES). As we analyze above, the major challenge in developing IAES is the absence of KKT conditions. We bypass this obstacle by carefully studying the relationship between SFM and convex optimization, which can be regarded as another form of KKT conditions. We find that SFM is closely related to a particular convex primal and dual problem pair Q-P and Q-D (see Section 2), that is, the minimizer of SFM can be obtained from the positive components of the optimum of Q-P. Hence, the proposed IAES identifies the active and inactive elements by estimating the lower and upper bounds of the components of the optimum of problem Q-P. Thus, one of our major technical contributions is a novel framework (Section 3)—developed by carefully studying the strong convexity of the corresponding primal and dual objective functions, the structure of the base polyhedra and the optimality conditions of the SFM problem—for deriving accurate estimations of optimum of problem Q-P. We integrate IAES with the solver for problems Q-P and Q-D. As the solver goes on, and the estimation becomes more and more accurate, IAES can identify more and more elements. By fixing the active elements and removing the inactive ones, the problem size can be reduced gradually. IAES is safe in the sense that it would never sacrifice any accuracy on the final output. To the best of our knowledge, IAES is the first screening method in the domain of SFM or even combinatorial optimization. Moreover, compared with the screening methods for sparse models, an outstanding feature of IAES is that it has no theoretical limit in reducing the problem

size. That is we can finally reduce the problem size to zero, leading to substantial saving in computational cost. The reason is that as the optimization proceeds, our estimation will be accurate enough to infer the affiliations of all the elements with the optimizer A^*. While in sparse models, screening methods can never reduce the problem size to zero since the features (resp. samples) with nonzero coefficients in the primal (resp. dual) optimum can never be removed from the dataset. Experiments (see Section 4) on both synthetic and real datasets demonstrate the significant speedups gained by IAES. For the convenience of presentation, we postpone the detailed proofs of theoretical results in the main text to the supplementary materials.

Notations: We consider the set $V = \{1, ..., p\}$, and denote its power set by 2^V, which is composed of the 2^p subsets of V. $|A|$ is the cardinality of a set A. $A \cup B$ and $A \cap B$ are the union and intersection of the sets A and B, respectively. $A \subseteq B$ means that A is a subset of B, potentially equals to B. Moreover, for $\mathbf{w} \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}$, we let $[\mathbf{w}]_k$ be the k-th component of $\mathbf{w}$ and $\{\mathbf{w} \geq \alpha\}$ (resp. $\{\mathbf{w} > \alpha\}$) be the weak (resp. strong) α-sup-level sets of $\mathbf{w}$, which is defined as $\{k : k \in V, [\mathbf{w}]_k \geq \alpha\}$ (resp. $\{k : k \in V, [\mathbf{w}]_k > \alpha\}$). At last, for $\mathbf{s} \in \mathbb{R}^p$, we define a set function by $\mathbf{s}(A) = \sum_{k \in A} [\mathbf{s}]_k$.

2 Basics and Motivations

This section is composed of two parts: a) briefly review some basics of submodular functions, SFM and their relations with convex optimization; b) motivate our screening method IAES.

The followings are the definitions of submodular function, submodular polyhedra and base polyhedra, which play an important role in submodular analysis.

Definition 1. (Submodular Function) *[17]. A set function $F : 2^V \to \mathbb{R}$ is submodular if and only if, for all subsets $A, B \subseteq V$, we have:*

$$F(A) + F(B) \geq F(A \cup B) + F(A \cap B).$$

Definition 2. (Submodular and Base Polyhedra) *[10]. Let F be a submodular function such that $F(\emptyset) = 0$. The submodular polyhedra $P(F)$ and the base polyhedra $B(F)$ are defined as:*

$$\begin{aligned} P(F) &= \{\mathbf{s} \in \mathbb{R}^p : \forall A \subseteq V, \mathbf{s}(A) \leq F(A)\}, \\ B(F) &= \{\mathbf{s} \in \mathbb{R}^p : \mathbf{s}(V) = F(V), \forall A \subseteq V, \mathbf{s}(A) \leq F(A)\}. \end{aligned}$$

Below we give the definition of Lovász extension, which works as the bridge that connects submodular functions and convex functions.

Definition 3. (Lovász Extension) *[10]. Given a set-function F such that $F(\emptyset) = 0$, the Lovász extension $f : \mathbb{R}^p \to \mathbb{R}$ is defined as follows: for $\mathbf{w} \in \mathbb{R}^p$, order the components in decreasing order $[\mathbf{w}]_{j_1} \geq ... \geq [\mathbf{w}]_{j_p}$, and define $f(\mathbf{w})$ through the equation below,*

$$f(\mathbf{w}) = \sum_{k=1}^{p} [\mathbf{w}]_{j_k} \big(F(\{j_1, ..., j_k\}) - F(\{j_1, ..., j_{k-1}\}) \big).$$

Lovász extension $f(\mathbf{w})$ is convex if and only if F is submodular (see [10]).

We focus on the generic submodular function minimization problem SFM defined in Section 1 and denote its minimizer as A^*. To reveal the relationship between SFM and convex optimization and finally motivate our method, we need the following theorems.

Theorem 1. *Let $\psi_1, ..., \psi_p$ be p convex functions on $\mathbb{R}$ and $\psi_1^*, ..., \psi_p^*$ be their Fenchel-conjugates ([3]) and f be the Lovász extension of a submodular function F. Denote the subgradient of $\psi_k(\cdot)$ by $\partial\psi_k(\cdot)$. Then, the followings hold:*
(i) *The problems below are dual of each other:*

$$\min_{\mathbf{w}\in\mathbb{R}^p} f(\mathbf{w}) + \sum_{j=1}^{p} \psi_j([\mathbf{w}]_j), \tag{P}$$

$$\max_{\mathbf{s}\in B(F)} -\sum_{j=1}^{p} \psi_j^*(-[\mathbf{s}]_j). \tag{D}$$

(ii) *The pair $(\mathbf{w}^*, \mathbf{s}^*)$ is optimal for problems (P) and (D) if and only if*

$$\begin{cases} \text{(a): } [\mathbf{s}]_k^* \in -\partial\psi_k([\mathbf{w}]_k^*), \forall k \in V, \\ \text{(b): } \mathbf{w}^* \in N_{B(F)}(\mathbf{s}^*), \end{cases} \tag{Opt}$$

where $N_{B(F)}(\mathbf{s}^*)$ is the normal cone (see Chapter 2 of [3]) of $B(F)$ at $\mathbf{s}^*$.

When $\psi_j(\cdot)$ is differentiable, we consider a sequence of set optimization problems parameterized by $\alpha \in \mathbb{R}$:

$$\min_{A\subseteq V} F(A) + \sum_{j\in A} \nabla\psi_j(\alpha), \tag{SFM'}$$

where $\nabla\psi_j(\cdot)$ is the gradient of $\psi_k(\cdot)$. The problem SFM' has tight connections with the convex optimization problem P (see the theorem below).

Theorem 2. (Submodular function minimization from proximal problem)*[Proposition 8.4 in [1]]. Under the same assumptions in Theorem 1, if $\psi_j(\cdot)$ is differentiable for all $j \in V$ and $\mathbf{w}^*$ is the unique minimizer of problem P, then for all $\alpha \in \mathbb{R}$, the minimal minimizer of problem SFM' is $\{\mathbf{u} > \alpha\}$ and the maximal minimizer is $\{\mathbf{u} \geq \alpha\}$, that is, for any minimizers A_α^*, we have:*

$$\{\mathbf{w}^* > \alpha\} \subseteq A_\alpha^* \subseteq \{\mathbf{w}^* \geq \alpha\}. \tag{1}$$

By choosing $\psi_j(x) = \frac{1}{2}x^2$ and $\alpha = 0$ in SFM', combining Theorems 1 and 2, we can see that SFM can be reduced to the following primal and dual problems, one is quadratic optimization problem and the other is equivalent to finding the minimum norm point in the base polytope $B(F)$:

$$\min_{\mathbf{w}\in\mathbb{R}^p} P(\mathbf{w}) = f(\mathbf{w}) + \frac{1}{2}\|\mathbf{w}\|_2^2, \tag{Q-P}$$

$$\max_{\mathbf{s}\in B(F)} D(\mathbf{s}) = -\frac{1}{2}\|\mathbf{s}\|_2^2. \tag{Q-D}$$

According to Eq. (1), we can define two index sets:

$$\mathcal{E} = \{j \in V : [\mathbf{w}]_j^* > 0\}, \text{ and } \mathcal{G} = \{j \in V : [\mathbf{w}]_j^* < 0\},$$

which imply that

$$\text{(i): } j \in \mathcal{E} \Rightarrow j \in A^*, \tag{R1}$$

$$\text{(ii): } j \in \mathcal{G} \Rightarrow j \notin A^*. \tag{R2}$$

We call the j-th element active if $j \in \mathcal{E}$ and the ones in $\mathcal{G}$ inactive.

Suppose that we are given two subsets of $\mathcal{E}$ and $\mathcal{G}$, by rules R1 and R2, we can see that many affiliations between A^* and the elements of V can be deduced. Thus, we have less unknowns to solve in SFM and its size can be dramatically reduced. We formalize this idea in Lemma 1.

Lemma 1. *Given two subsets $\hat{\mathcal{G}} \subseteq \mathcal{G}$ and $\hat{\mathcal{E}} \subseteq \mathcal{E}$, the followings hold:*
(i): *$\hat{\mathcal{E}} \subseteq A^*$, and for all $j \in \hat{\mathcal{G}}$, we have $j \notin A^*$.*
(ii): *The problem SFM can be reduced to the following scaled problem:*

$$\min_{C \subseteq V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}})} \hat{F}(C) := F(\hat{\mathcal{E}} \cup C) - F(\hat{\mathcal{E}}), \tag{scaled-SFM}$$

which is also a SFM problem.
(iii): *A^* can be recovered by $A^* = \hat{\mathcal{E}} \cup C^*$, where C^* is the minimizer of scaled-SFM.*

Lemma 1 indicates that, if we can identify the active set $\hat{\mathcal{E}}$ and inactive set $\hat{\mathcal{G}}$, we only need to solve a scaled problem scaled-SFM—that may have much smaller size than the original problem SFM—to exactly recover the optimal solution A^* without sacrificing any accuracy.

However, since $\mathbf{w}^*$ is unknown, we cannot directly apply rules R1 and R2 to identify the active set $\hat{\mathcal{E}}$ and inactive set $\hat{\mathcal{G}}$. Inspired by the ideas in the gap safe screening methods ([9, 20, 23]) for convex problems, we can first estimate the region $\mathcal{W}$ that contains $\mathbf{w}^*$ and then relax the rules R1 and R2 to the practicable versions. Specifically, we first denote

$$\hat{\mathcal{E}} := \{j \in V : \min_{\mathbf{w} \in \mathcal{W}} [\mathbf{w}]_j > 0\}, \tag{2}$$

$$\hat{\mathcal{G}} := \{j \in V : \max_{\mathbf{w} \in \mathcal{W}} [\mathbf{w}]_j < 0\}. \tag{3}$$

It is obvious that $\hat{\mathcal{E}} \subseteq \mathcal{E}$ and $\hat{\mathcal{G}} \subseteq \mathcal{G}$. Hence, the rules R1 and R2 can be relaxed as follows:

$$\text{(i): } j \in \hat{\mathcal{E}} \Rightarrow j \in A^*, \tag{R1'}$$

$$\text{(ii): } j \in \hat{\mathcal{G}} \Rightarrow j \notin A^*. \tag{R2'}$$

In view of the rules R1' and R2', we sketch the development of IAES as follows:
Step 1: Derive the estimation $\mathcal{W}$ such that $\mathbf{w}^*(\alpha, \beta) \in \mathcal{W}$.
Step 2: Develop IAES via deriving the detailed screening rules R1' and R2'.

3 The Proposed Element Screening Method

In this section, we first present the accurate optimum estimation by carefully studying the strong convexity of the functions $P(\mathbf{w})$ and $D(\mathbf{s})$, the optimality conditions of SFM and its relationships with the convex problem pair (see Section 3.1). Then, in Section 3.2, we develop our inactive and active element screening rules IES and AES step by step. At last, in Section 3.3, we develop the screening framework IAES by an alternating application of IES and AES.

3.1 Optimum Estimation

Let $\hat{\mathcal{E}}$ and $\hat{\mathcal{G}}$ be the active and inactive sets identified by the previous IAES steps (before applying IAES for the first time, they are $\emptyset$). From Lemma 1, we know that the problem SFM then can be reduced to the following scaled problem:

$$\min_{C \subseteq \hat{V}} \hat{F}(C) := F(\hat{\mathcal{E}} \cup C) - F(\hat{\mathcal{E}}),$$

where $\hat{V} = V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}})$. The second term $-F(\hat{\mathcal{E}})$ at the right side of the equation above is added to make $\hat{F}(\emptyset) = 0$. Thus, the corresponding problems Q-P and Q-D then become:

$$\min_{\hat{\mathbf{w}} \in \mathbb{R}^{\hat{p}}} \hat{P}(\hat{\mathbf{w}}) = \hat{f}(\hat{\mathbf{w}}) + \frac{1}{2}\|\hat{\mathbf{w}}\|_2^2, \tag{Q-P'}$$

$$\max_{\hat{\mathbf{s}} \in B(\hat{F})} \hat{D}(\hat{\mathbf{s}}) = -\frac{1}{2}\|\hat{\mathbf{s}}\|_2^2. \tag{Q-D'}$$

where $\hat{f}(\hat{\mathbf{w}})$ is the Lovász extension of $\hat{F}$ and $\hat{p} = |V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}})|$. Now, we turn to estimate the minimizer $\hat{\mathbf{w}}^*$ of the problem Q-P'. The result is presented in the theorem below.

Theorem 3. *For any $\hat{\mathbf{w}} \in dom\hat{P}(\hat{\mathbf{w}})$, $\hat{\mathbf{s}} \in B(\hat{F})$ and $C \subseteq \hat{V}$, we denote the dual gap as $G(\hat{\mathbf{w}}, \hat{\mathbf{s}}) = \hat{P}(\hat{\mathbf{w}}) - \hat{D}(\hat{\mathbf{s}})$, then we have*

$$\hat{\mathbf{w}}^* \in \mathcal{W} = \mathcal{B} \cap \Omega \cap \mathcal{P},$$

where $\mathcal{B} = \Big\{\mathbf{w} : \|\mathbf{w} - \hat{\mathbf{w}}\| \leq \sqrt{2G(\hat{\mathbf{w}}, \hat{\mathbf{s}})}\Big\}$, $\Omega = \Big\{\mathbf{w} : \hat{F}(\hat{V}) - 2\hat{F}(C) \leq \|\mathbf{w}\|_1 \leq \|\hat{\mathbf{s}}\|_1\Big\}$, and $\mathcal{P} = \Big\{\mathbf{w} : \langle \mathbf{w}, \mathbf{1}\rangle = -\hat{F}(\hat{V})\Big\}$.

From the theorem above, we can see that the estimation $\mathcal{W}$ is the intersection of three sets: the ball $\mathcal{B}$, the ℓ_1-norm equipped spherical Ω and the plane $\mathcal{P}$. As the optimizer goes on, the dual gap $G(\hat{\mathbf{w}}, \hat{\mathbf{s}})$ becomes smaller, and $\hat{F}(\hat{V}) - 2\hat{F}(C)$ and $\|\hat{\mathbf{w}}\|_1$ would converge to $\|\hat{\mathbf{w}}^*\|_1$ (See Chapter 7 of [1]). Thus, the volumes of $\mathcal{B}$ and Ω become smaller and smaller during the optimization process, the estimation $\mathcal{W}$ would be more and more accurate.

3.2 Inactive and Active Element Screening

We now turn to develop the screening rules IES and AES based on the estimation of the optimum $\hat{\mathbf{w}}^*$.

From (2) and (3), we can see that, to develop the screening rules we need to solve two problems: $\min_{\mathbf{w} \in \mathcal{W}} [\mathbf{w}]_j$ and $\max_{\mathbf{w} \in \mathcal{W}} [\mathbf{w}]_j$. However, since $\mathcal{W}$ is highly non-convex and has a complex structure, it is very hard to solve these two problems efficiently. Hence, we rewrite the estimation $\mathcal{W}$ as $\mathcal{W} = (\mathcal{B} \cap \mathcal{P}) \cap (\mathcal{B} \cap \Omega)$, and develop two different screening rules on $\mathcal{B} \cap \mathcal{P}$ and $\mathcal{B} \cap \Omega$, respectively.

3.2.1 Inactive and Active Element Screening based on $\mathcal{B} \cap \mathcal{P}$

Given the estimation $\mathcal{B} \cap \mathcal{P}$, we derive the screening rules by solving the following problems

$$\min_{\mathbf{w} \in \mathcal{B} \cap \mathcal{P}} [\mathbf{w}]_j \text{ and } \max_{\mathbf{w} \in \mathcal{B} \cap \mathcal{P}} [\mathbf{w}]_j.$$

We show that both of the two problems above admit closed form solutions.

Lemma 2. *Given the estimation ball $\mathcal{B}$, the plane $\mathcal{P}$ and the active and inactive sets $\hat{\mathcal{E}}$ and $\hat{\mathcal{G}}$, which are identified in the previous IAES steps, for all $\forall j \in \hat{p}$, we denote*

$$b_j = 2\Big(\sum_{i \neq j} [\hat{\mathbf{w}}]_i + \hat{F}(\hat{V}) - (\hat{p} - 1)[\hat{\mathbf{w}}]_j\Big),$$

$$c_j = \Big(\sum_{i \neq j} [\hat{\mathbf{w}}]_i + \hat{F}(\hat{V})\Big)^2 - (\hat{p} - 1)\Big(2G(\hat{\mathbf{w}}, \hat{\mathbf{s}}) - [\hat{\mathbf{w}}]_j^2\Big),$$

then the followings hold:

$$\text{(i)}: \min_{\mathbf{w}\in\mathcal{B}\cap\mathcal{P}} [\mathbf{w}]_j = [\mathbf{w}]_j^{\min} := \frac{-b_j - \sqrt{b_j^2 - 4\hat{p}c_j}}{2\hat{p}},$$

$$\text{(ii)}: \max_{\mathbf{w}\in\mathcal{B}\cap\mathcal{P}} [\mathbf{w}]_j = [\mathbf{w}]_j^{\max} := \frac{-b_j + \sqrt{b_j^2 - 4\hat{p}c_j}}{2\hat{p}}.$$

We are now ready to present the active and inactive screening rules AES-1 and IES-1.

Theorem 4. *Given the active and inactive sets $\hat{\mathcal{E}}$ and $\hat{\mathcal{G}}$, which are identified in the previous IAES steps, then,*

(i): *The active element screening rule takes the form of*

$$[\mathbf{w}]_j^{\min} > 0 \Rightarrow j \in A^*, \forall j \in V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}}). \tag{AES-1}$$

(ii): *The inactive element screening rule takes the form of*

$$[\mathbf{w}]_j^{\max} < 0 \Rightarrow j \notin A^*, \forall j \in V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}}). \tag{IES-1}$$

(iii): *The active and inactive sets $\hat{\mathcal{E}}$ and $\hat{\mathcal{G}}$ can be updated by*

$$\hat{\mathcal{E}} \leftarrow \hat{\mathcal{E}} \cup \Delta\hat{\mathcal{E}}, \tag{4}$$

$$\hat{\mathcal{F}} \leftarrow \hat{\mathcal{F}} \cup \Delta\hat{\mathcal{F}}, \tag{5}$$

where $\Delta\hat{\mathcal{E}}$ and $\Delta\hat{\mathcal{G}}$ are the newly identified active and inactive sets defined as

$$\Delta\hat{\mathcal{E}} = \{j \in V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}}) : [\mathbf{w}]_j^{\min} > 0\},$$
$$\Delta\hat{\mathcal{G}} = \{j \in V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}}) : [\mathbf{w}]_j^{\max} < 0\}.$$

From the theorem above, we can see that our rules AES-1 and IES-1 are safe in the sense that the detected elements are guaranteed to be include or exclude in A^*.

3.2.2 Inactive and Active Element Screening based on $\mathcal{B} \cap \Omega$

We now derive the second screening rule pair based on the estimation $\mathcal{B} \cap \Omega$.

Due to the high non-convexity and complex structure of $\mathcal{B} \cap \Omega$, directly solving problems $\min_{\mathbf{w}\in\mathcal{B}\cap\Omega}[\mathbf{w}]_j$ and $\max_{\mathbf{w}\in\mathcal{B}\cap\Omega}[\mathbf{w}]_j$ is time consuming. Notice that, to derive IAS and IES, we only need to judge whether the inequalities $\min_{\mathbf{w}\in\mathcal{B}\cap\Omega}[\mathbf{w}]_j > 0$ and $\max_{\mathbf{w}\in\mathcal{B}\cap\Omega}[\mathbf{w}]_j < 0$ are satisfied or not, instead of calculating $\min_{\mathbf{w}\in\mathcal{B}\cap\Omega}[\mathbf{w}]_j$ and $\max_{\mathbf{w}\in\mathcal{B}\cap\Omega}[\mathbf{w}]_j$. Hence, we only need to infer the hypotheses $\left\{\mathbf{w} : \mathbf{w} \in \mathcal{B}, [\mathbf{w}]_j \leq 0\right\} \cap \Omega = \emptyset$ and $\left\{\mathbf{w} : \mathbf{w} \in \mathcal{B}, [\mathbf{w}]_j \geq 0\right\} \cap \Omega = \emptyset$ are true or false. Thus, from the formulation of Ω (see Theorem 3), the problems come down to calculating the minimum and the maximum of $\|\mathbf{w}\|_1$ with $\left\{\mathbf{w} : \mathbf{w} \in \mathcal{B}, [\mathbf{w}]_j \geq 0\right\}$ or $\left\{\mathbf{w} : \mathbf{w} \in \mathcal{B}, [\mathbf{w}]_j \leq 0\right\}$, which admit closed form solutions. The results are presented in the lemma below.

Lemma 3. *Given the estimation ball $\mathcal{B}$ and the active and inactive sets $\hat{\mathcal{E}}$ and $\hat{\mathcal{G}}$, which are identified in the previous IAES steps, then the followings hold:*

(i): *$\forall j \in \hat{p}$, if $|[\hat{\mathbf{w}}]_j| > \sqrt{2G(\hat{\mathbf{w}}, \hat{\mathbf{s}})}$, then the element j can be identified by rule AES-1 or IES-1 to be active or inactive.*

(ii): *$\forall j \in \hat{p}$, if $0 < [\hat{\mathbf{w}}]_j \leq \sqrt{2G(\hat{\mathbf{w}}, \hat{\mathbf{s}})}$, we have*

$$\min_{\mathbf{w}\in\mathcal{B}, [\mathbf{w}]_j \leq 0} \|\mathbf{w}\|_1 < \|\hat{\mathbf{w}}\|_1,$$

$$\max_{\mathbf{w}\in\mathcal{B}, [\mathbf{w}]_j \leq 0} \|\mathbf{w}\|_1 = \begin{cases} \|\hat{\mathbf{w}}\|_1 - 2[\hat{\mathbf{w}}]_j + \sqrt{2\hat{p}G(\hat{\mathbf{w}}, \hat{\mathbf{s}})}, \text{ if } [\hat{\mathbf{w}}]_j - \sqrt{\frac{2G(\hat{\mathbf{w}}, \hat{\mathbf{s}})}{\hat{p}}} < 0, \\ \|\hat{\mathbf{w}}\|_1 - [\hat{\mathbf{w}}]_j + \sqrt{\hat{p} - 1}\sqrt{2G(\hat{\mathbf{w}}, \hat{\mathbf{s}}) - [\hat{\mathbf{w}}]_j^2}, \text{ otherwise.} \end{cases}$$

(iii): $\forall j \in \hat{p}$, *if* $-\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} \leq [\hat{\mathbf{w}}]_j < 0$, *we have*

$$\min_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\geq 0} \|\mathbf{w}\|_1 < \|\hat{\mathbf{w}}\|_1,$$

$$\max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\geq 0} \|\mathbf{w}\|_1 = \begin{cases} \|\hat{\mathbf{w}}\|_1 + 2[\hat{\mathbf{w}}]_j + \sqrt{2\hat{p}G(\hat{\mathbf{w}},\hat{\mathbf{s}})}, \textit{if } [\hat{\mathbf{w}}]_j + \sqrt{\frac{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}{\hat{p}}} > 0, \\ \|\hat{\mathbf{w}}\|_1 + [\hat{\mathbf{w}}]_j + \sqrt{\hat{p}-1}\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - [\hat{\mathbf{w}}]_j^2}, \textit{otherwise.} \end{cases}$$

We are now ready to present the second active and inactive screening rule pair AES-2 and IES-2. From the lemma above, we can see that the element j with $|[\hat{\mathbf{w}}]_j| > \sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}$ can be screened by rules AES-1 and IES-1. So we now only need to consider the cases when $|[\hat{\mathbf{w}}]_j| \leq \sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}$.

Theorem 5. *Given a set* $C \subseteq \hat{V}$ *and the active and inactive sets* $\hat{\mathcal{E}}$ *and* $\hat{\mathcal{G}}$ *identified in the previous IAES steps, then,*

(i): *The active element screening rule takes the form of*

$$\begin{cases} 0 < [\hat{\mathbf{w}}]_j \leq \sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} \\ \max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\leq 0} \|\mathbf{w}\|_1 < \hat{F}(\hat{V}) - 2\hat{F}(C) \end{cases} \Rightarrow j \in A^*, \forall j \in V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}}). \tag{AES-2}$$

(ii): *The inactive element screening rule takes the form of*

$$\begin{cases} -\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} \leq [\hat{\mathbf{w}}]_j < 0 \\ \max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\geq 0} \|\mathbf{w}\|_1 < \hat{F}(\hat{V}) - 2\hat{F}(C) \end{cases} \Rightarrow j \notin A^*, \forall j \in V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}}). \tag{IES-2}$$

(iii): *The active and inactive sets* $\hat{\mathcal{E}}$ *and* $\hat{\mathcal{G}}$ *can be updated by*

$$\hat{\mathcal{E}} \leftarrow \hat{\mathcal{E}} \cup \Delta\hat{\mathcal{E}}, \tag{6}$$

$$\hat{\mathcal{F}} \leftarrow \hat{\mathcal{F}} \cup \Delta\hat{\mathcal{F}}, \tag{7}$$

where $\Delta\hat{\mathcal{E}}$ *and* $\Delta\hat{\mathcal{G}}$ *are the newly identified active and inactive sets defined as*

$$\Delta\hat{\mathcal{E}} = \Big\{ j \in V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}}) : 0 < [\hat{\mathbf{w}}]_j \leq \sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}, \max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\leq 0} \|\mathbf{w}\|_1 < \hat{F}(\hat{V}) - 2\hat{F}(C) \Big\},$$

$$\Delta\hat{\mathcal{G}} = \Big\{ j \in V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}}) : -\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} \leq [\hat{\mathbf{w}}]_j < 0, \max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\geq 0} \|\mathbf{w}\|_1 < \hat{F}(\hat{V}) - 2\hat{F}(C) \Big\}.$$

Theorem 5 verifies the safety of AES-2 and IES-2.

3.3 The Proposed IAES Framework by An Alternating Application of AES and IES

To reinforce the capability of the proposed screening rules, we develop a novel framework IAES in Algorithm 2, which applies the active element screening rules (AES-1 and AES-2) and the inactive element screening rules (IES-1 and IES-2) in an alternating manner during the optimization process. Specifically, we integrate our screening rules AES-1, AES-2, IES-1 and IES-2 with the optimization algorithm $\mathcal{A}$ for the problems Q-P' and Q-D'. During the optimization process, we trigger the screening rules AES-1, AES-2, IES-1 and IES-2 every time when the dual gap is $1-\rho$ times smaller than itself in the last triggering of IAES. As the solver $\mathcal{A}$ goes on, the volumes of Ω and $\mathcal{B}$ would decrease to zero quickly, IAES can thus identify more and more inactive and active elements.

Compared with the existing screening methods for convex sparse models, an appealing feature of IAES is that it has no theoretical limit in identifying the inactive and active elements

Algorithm 2 Inactive and Active Element Screening

Input: an optimization algorithm $\mathcal{A}$ for problems (Q-P') and (Q-D'), $\epsilon > 0, 0 < \rho < 1$.
Initialize: $\hat{\mathcal{E}} = \hat{\mathcal{G}} = \emptyset, C = \emptyset, g = \infty$, choose $\hat{\mathbf{s}} \in B(F)$ and $\hat{\mathbf{w}} = -\hat{\mathbf{s}}$.
repeat
  Run $\mathcal{A}$ on problems (Q-P') and (Q-D') to update $\hat{\mathbf{w}}$, $\hat{\mathbf{s}}$ and C.
  if dual gap $G(\hat{\mathbf{w}}, \hat{\mathbf{s}}) < \rho g$ **then**
    Run the active element screening rule AES-1 and AES-2 based on $(\hat{\mathbf{w}}, \hat{\mathbf{s}})$ and C.
    Update the active set $\hat{\mathcal{E}}$ by (4) and (6).
    Run the inactive element screening rule IES-1 and IES-2 based on $(\hat{\mathbf{w}}, \hat{\mathbf{s}})$ and C.
    Update the inactive set $\hat{\mathcal{G}}$ by (5) and (7).
    if $V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}}) = \emptyset$ **then**
      Return: $\hat{\mathcal{E}}$.
    else
      Update $\hat{F}$, Q-P', Q-D' according to $\hat{\mathcal{E}}$ and $\hat{\mathcal{G}}$.
      Update $\hat{\mathbf{w}}$ and $\hat{\mathbf{s}}$ by:

$$\hat{\mathbf{w}} \leftarrow [\hat{\mathbf{w}}]_{V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}})}, \text{ and } \hat{\mathbf{s}} \leftarrow \arg\max_{\mathbf{s} \in B(\hat{F})} \langle \hat{\mathbf{w}}, \mathbf{s} \rangle.$$

      Update $g \leftarrow G(\hat{\mathbf{w}}, \hat{\mathbf{s}})$.
    end if
  end if
until $G(\hat{\mathbf{w}}, \hat{\mathbf{s}}) < \epsilon$.
Return: $\hat{\mathcal{E}} \cup \{\hat{\mathbf{w}} > 0\}$.

and reducing the problem size. The reason is that, in convex sparse models, screening models can never rule out the features and samples whose corresponding coefficients in the optimal solution are nonzero. While in our case, as the optimizer $\mathcal{A}$ going on, our estimation will be accurate enough for us to infer the affiliation of each element with A^*. Hence, we can finally identify all the inactive and active elements and the problem size can be reduced to zero. This nice feature can lead to significant speedups in the computation times.

Remark 1. *The set C in Algorithm 2 is updated by choosing one of the super-level sets of $\hat{\mathbf{w}}$ with the smallest value $\hat{F}(C)$. It is free to get it. The reason is that most of the existing methods $\mathcal{A}$ for the problems Q-P' and Q-D' need to calculate $\hat{f}(\hat{\mathbf{w}})$ in each iteration, in which they need to calculate the value $\hat{F}$ at all of the super-level sets of $\hat{\mathbf{w}}$ (see the greedy algorithm in [1] for details).*

Remark 2. *The algorithm $\mathcal{A}$ can be all the methods for the problems Q-P' and Q-D', such as minimum-norm point algorithm [28] and conditional gradient descent [5]. Although some algorithms only update $\mathbf{s}$, in IAES, we can update $\mathbf{w}$ in each iteration by letting $\mathbf{w} = -\mathbf{s}$ and refining it by the algorithm named pool adjacent violators [2].*

Remark 3. *Due to Lemma 1 and the safety of AES-1, AES-2, IES-1 and IES-2, we can see that IAES would never sacrifice any accuracy.*

Remark 4. *Although step 14 in Algorithm 2 may increase the dual gap slightly, it is worth it because of the reduced problem size. This is verified by the speedups gained by IAES in the experiments.*

Remark 5. *The parameter ρ in Algorithm 2 controls the frequency how often we trigger IAES. The larger value, the higher frequency to trigger IAES but more computational time consumed by IAES. In our experiment, we set $\rho = 0.5$ and it achieves a good performance.*

4 Experiments

We evaluate IAES through numerical experiments on both synthetic and real datasets by two measurements. The first one is the rejection ratios of IAES over iterations: $\frac{m_i+n_i}{m^*+n^*}$, where m_i and n_i are the numbers of the active and inactive elements identified by IAES after the i-th iteration, and m^* and n^* are the numbers of the active and inactive elements in A^*. We notice that in our experiments $m^* + n^* = p$, so the rejection ratio presents the problem size reduced by IAES. The second measurement is speedup, i.e., the ratio of the running times of the solver without IAES and with IAES. We set the accuracy ϵ to be 10^{-6}.

Recall that, IAES can be integrated with all the solvers for the problems Q-P and Q-D. In this experiment, we use one of the most widely used algorithm minimum-norm point algorithm (MinNorm) [28] as the solver. The function $F(A)$ varies according to the datasets, whose detailed definitions will be given in subsequent sections.

We write the code in Matlab and perform all the computations on a single core of Intel(R) Core(TM) i7-5930K 3.50GHz, 32GB MEM.

4.1 Experiments on Synthetic Datasets

We perform experiments on the synthetic dataset named two-moons with different sample size (see Figure 1 for an example). All the data points are sampled from two different semicircles. Specifically, each point can be presented as $\mathbf{x} = \mathbf{c}_i + \gamma * [\cos(\theta_i), \sin(\theta_i)]$, where $i = 1, 2$ stands for the two semicircles, $\mathbf{c}_1 = [-0.5, 1], \mathbf{c}_2 = [0.5, -1]$, γ is generated from a normal distribution $N(2, 0.5^2)$, θ_1 and θ_2 are sampled from two uniform distributions $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and $[\frac{\pi}{2}, \frac{3\pi}{2}]$, respectively. We first sample p data points from these two semicircles with equal probability. Then, we randomly choose $p_0 = 16$ samples and label each of them as positive if it is from the first semicircle and otherwise label it as negative. We generate five datasets by varying the sample size p in $[200, 400, 600, 800, 1000]$. We perform semi-supervised clustering on each dataset and the objective function $F(A)$ are defined as:

$$F(A) = I(f_A, f_{V/A}) - \sum_{j \in A} \log \eta_j - \sum_{j \in V/A} \log(1 - \eta_j),$$

where $I(f_A, f_{V/A})$ is the mutual information between two Gaussian processes with a Gaussian kernel $k(x, y) = \exp(-\alpha \|\mathbf{x} - \mathbf{y}\|^2), \alpha = 1.5$, $\eta_j \in \{0, 1\}$ if j is labeled and otherwise $\eta_j = \frac{1}{2}$(see Chapter 3 of [1] for more details). The kernel matrix here is dense with the size $p \times p$, leading to a big computational cost when p is large.

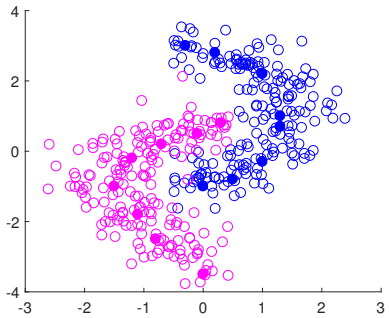
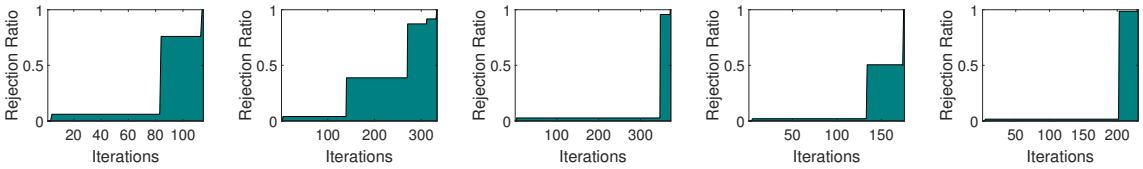


Figure 1: Two-moons dataset with 400 samples. The data points in magenta and blue are sampled from two different distributions. The filled dots are the labeled data points.

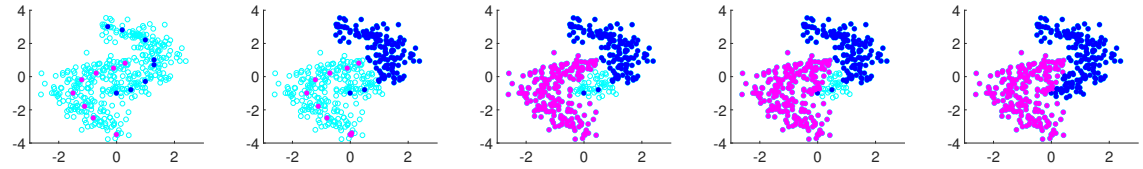
Figure 2 presents the rejection ratios of IAES on two-moons. We can see that IAES can find the active and inactive elements incrementally during the optimization process. It can finally identify almost all of the elements and reduce the problem size to nearly zero in no more than 400 iterations, which is consistent with our theoretical analysis in Section 3.3.



(a) p=200 (b) p=400 (c) p=600 (d) p=800 (e) p=1000

Figure 2: Rejection ratios of IAES over the iterations on two-moons.

Figure 3 visualizes the screening process of IAES on two-moons when $p = 400$. It shows that, during the optimization process, IAES identifies the elements those are easy to be classified first and then identify the rest.



(a) after 4-th iteration (b) after 140-th iteration (c) after 271-th iteration (d) after 312-th iteration (e) after 333-th iteration

Figure 3: The visualization for the screening process of IAES on two-moons with $p = 400$. The filled dots in magenta and blue are the identified active and inactive elements, respectively. The data points in cyan are the unidentified samples.

Table 1 reports the running time of MinNorm without and with AES (AES-1 + AES-2), IES (IES-1 + IES-2) and IAES for solving the problem SFM on two-moons. We can see that the speedup of IAES can be up to 10 times. In all the datasets, IAES is significantly faster than MinNorm, the MinNorm with AES or IES. At last, we can see that the time cost of AES, IES and IAES are negligible.

Table 1: Running time (in seconds) for solving SFM on two-moons.

Data	MinNorm	AES+MinNorm			IES+MniNorm			IAES+MinNorm		
		AES	MinNorm	Speedup	IES	MinNorm	Speedup	IAES	MinNorm	Speedup
$p = 200$	29.1	0.08	12.5	2.3	0.07	12.2	2.4	0.10	4.3	**6.8**
$p = 400$	829.1	0.09	106.5	2.8	0.12	231.7	3.6	0.15	82.7	**10.0**
$p = 600$	2,084.5	0.12	408.7	5.1	0.13	671.1	3.1	0.18	217.1	**9.6**
$p = 800$	2701.1	0.15	534.0	5.1	0.13	998.9	2.7	0.25	400.2	**6.8**
$p = 1000$	5422.9	0.20	1177.4	4.6	0.19	1453.5	3.7	0.30	774.7	**7.0**

4.2 Experiments on Real Datasets

In this experiment, we evaluate the performance of IAES on the image segmentation problem. We use five image segmentation instances (included in the supplemental material)

in [22] to evaluate IAES. The objective function $F(A)$ is the sum of unary potential for each pixel and pairwise potential of 8-neighbor grid graph:

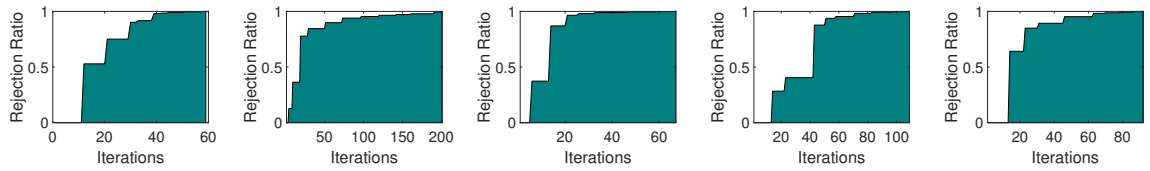
$$F(A) = \mathbf{u}(A) + \sum_{i \in A, j \in V/A} d(i,j),$$

where V presents all the pixels, $\mathbf{u} \in \mathbb{R}^V$ is the unary potential derived from the Gaussian Mixture model [22], $d(i,j) = \exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2\}$ ($\mathbf{x}_i$ and $\mathbf{x}_j$ are the values of two pixels) if i, j are neighbors, otherwise $d(i,j) = 0$. Table 2 provides the statistics of the resulting image segmentation problems, including the numbers of the pixels and the edges in the 8-neighbor grid graph.

Table 2: Statistics of the image segmentation problems.

image	#pixels	#edges
image1	50,246	201,427
image2	26,600	106,733
image3	51,000	204,455
image4	60,000	240,500
image5	45,200	181,226

The rejection ratios in Figure 4 shows that IAES can identify the active and inactive elements during the optimization process incrementally until all of them are identified. This implies that IAES can lead to a significant speedup.



(a) image1 (b) image2 (c) image3 (d) image4 (e) image5

Figure 4: Rejection ratios of IAES on real datasets over the iterations

Table 3 reports the detailed time cost of MinNorm without and with AES, IES and IAES for solving the image segmentation problems. We can see that IAES leads to significant speedups, that is up to 30.7 times. In addition, we notice that the speedup gained by AES is small. The reason is that AES is used to identify the pixels of the foreground, which is a small region in the image, and thus the problem size cannot be reduced dramatically even if all the active elements were identified.

Table 3: Running time (in seconds) for solving SFM on the task of image segmentation.

Data	MinNorm	AES+MinNorm			IES+MniNorm			IAES+MinNorm		
		AES	MinNorm	Speedup	IES	MinNorm	Speedup	IAES	MinNorm	Speedup
image1	1575.6	0.10	1412.8	1.12	0.11	242.7	6.48	0.21	78.3	**20.19**
image2	1780.6	0.21	1201.6	1.48	0.30	616.6	2.89	0.50	130.1	**13.69**
image3	6775.8	0.51	5470.8	1.24	0.53	1080.4	6.27	0.17	220.7	**30.70**
image4	6613.5	0.42	5773.1	1.15	0.43	1286.3	5.14	0.32	553.2	**11.96**
image5	4025.4	0.40	3638.8	1.11	0.17	506.4	7.95	0.51	187.5	**21.50**

At last, from Table 3, we can also see that the speedup we achieve is supper-additive (speedup of AES + speedup of IES $<$ speedup of IAES). This can usually be expected, which comes from the super linear computational complexity of each iteration in MinNorm, leading to a super-additive saving in the computational cost. We notice that the speedup we achieve

on some of the two-moon datasets is not super-additive, the reason is that we cannot identify a lot of elements in the early stage (Figure 2). Thus, the early stage takes up too much time cost.

5 Conclusion

In this paper, we proposed a novel safe element screening method IAES for SFM to accelerate its optimization process by simultaneously identifying the active and inactive elements. Our major contribution is a novel framework for accurately estimating the optimum of the corresponding primal problem of SFM developed by carefully studying the strong convexity of the primal and dual problems, the structure of the base polyhedra and the optimality conditions of SFM. To the best of our knowledge, IAES is the first screening method in the fields of SFM and even combinatorial optimization. Experiment results demonstrate that IAES can achieve significant speedups in solving SFM problems.

A Appendix

In this appendix, we present the detailed proofs of all the theorems in the main text.

A.1 Proof of Theorem 1

Proof. of Theorem 1:

(i) Since $f(\mathbf{w}) = \max_{\mathbf{s}\in B(F)}\langle \mathbf{w}, \mathbf{s}\rangle$, we can have that

$$\min_{\mathbf{w}\in\mathbb{R}^p} f(\mathbf{w}) + \sum_{j=1}^{p}\psi_j([\mathbf{w}]_j) \tag{8}$$

$$= \min_{\mathbf{w}\in\mathbb{R}^p}\max_{\mathbf{s}\in B(F)} \langle \mathbf{w}, \mathbf{s}\rangle + \sum_{j=1}^{p}\psi_j([\mathbf{w}]_j)$$

$$= \max_{\mathbf{s}\in B(F)}\min_{\mathbf{w}\in\mathbb{R}^p} \langle \mathbf{w}, \mathbf{s}\rangle + \sum_{j=1}^{p}\psi_j([\mathbf{w}]_j) \tag{9}$$

$$= \max_{\mathbf{s}\in B(F)} -\sum_{j=1}^{p}\psi_j^*(-[\mathbf{s}]_j), \tag{10}$$

where Eq.(9) holds since the strong duality theorem [3], and Eq.(10) is due to the definitions of the Fenchel conjugate of ψ_j.

(ii) From Eq. (8), we have that

$$\begin{aligned}
&\mathbf{s}^* \in \arg\max_{\mathbf{s}\in B(F)} \langle \mathbf{w}^*, \mathbf{s}\rangle \\
&\Leftrightarrow \langle \mathbf{w}^*, \mathbf{s}^*\rangle \geq \langle \mathbf{w}^*, \mathbf{s}\rangle, \forall \mathbf{s} \in B(F) \\
&\Leftrightarrow \mathbf{w}^* \in N_{B(F)}(\mathbf{s}^*).
\end{aligned}$$

From Eq. (10), we have that

$$\begin{aligned}
&\mathbf{w}^* \in \arg\min_{\mathbf{w}\in\mathbb{R}^p} \langle \mathbf{w}, \mathbf{s}^*\rangle + \sum_{j=1}^{p}\psi_j([\mathbf{w}]_j) \\
&\Leftrightarrow [\mathbf{s}]_k^* \in -\partial\psi_k([\mathbf{w}]_k^*), \forall k \in V.
\end{aligned}$$

The proof is complete. □

A.2 Proof of Lemma 1

Proof. of Lemma 1:

(i) It is the immediate conclusion of Theorem 2.

(ii) Since $\hat{\mathcal{E}} \subseteq A^*$ and $\hat{\mathcal{G}} \subseteq V/A^*$, we can solve the problem SFM by fixing the set $\hat{\mathcal{E}}$ and optimizing over $V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}})$. And the objective function becomes $\hat{F}(C) := F(\hat{\mathcal{E}} \cup C) - F(\hat{\mathcal{E}})$ with $C \subseteq V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}})$. Thus, SFM can be deduced to

$$\min_{C\subseteq V/(\hat{\mathcal{E}}\cup\hat{\mathcal{G}})} \hat{F}(C) := F(\hat{\mathcal{E}} \cup C) - F(\hat{\mathcal{E}}).$$

The second term of the new objective function $\hat{F}(C)$ is added to make $\hat{F}(\emptyset) = 0$, which is essential in submodular function analysis, such as Lovász extension, submodular and base polyhedra.

Below, we argue that $\hat{F}(C)$ is a submodular function.

For all $S \subseteq V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}})$ and $T \subseteq V/(\hat{\mathcal{E}} \cup \hat{\mathcal{G}})$, we have

$$\begin{aligned}
\hat{F}(S) + \hat{F}(T) &= (F(\hat{\mathcal{E}} \cup S) - F(\hat{\mathcal{E}})) + (F(\hat{\mathcal{E}} \cup T) - F(\hat{\mathcal{E}})) \\
&= F(\hat{\mathcal{E}} \cup S) + F(\hat{\mathcal{E}} \cup T) - 2F(\hat{\mathcal{E}}) \\
&\geq F((\hat{\mathcal{E}} \cup S) \cup (\hat{\mathcal{E}} \cup T)) + F((\hat{\mathcal{E}} \cup S) \cap (\hat{\mathcal{E}} \cup T)) - 2F(\hat{\mathcal{E}}) \qquad (11) \\
&= F(\hat{\mathcal{E}} \cup (S \cup T)) + F(\hat{\mathcal{E}} \cup (S \cup T)) - 2F(\hat{\mathcal{E}}) \\
&= (F(\hat{\mathcal{E}} \cup (S \cup T)) - F(\hat{\mathcal{E}})) + (F(\hat{\mathcal{E}} \cup (S \cup T)) - F(\hat{\mathcal{E}})) \\
&= \hat{F}(S \cup T) + \hat{F}(S \cap T).
\end{aligned}$$

The inequality (11) comes from the submoduality of F.

(iii) It is the immediate conclusion of (ii).

The proof is complete. □

A.3 Proof of Theorem 3

To prove Theorem 3, we need the following Lemma.

Lemma 4. *(Dual of minimization of submodular functions)[Proposition 10.3 in [1]] Let F be a submodular function such that $F(\emptyset) = 0$. We have:*

$$\min_{A \subseteq V} F(A) = \max_{\mathbf{s} \in B(F)} \mathbf{s}_-(V) = \frac{1}{2}\Big(F(V) - \min_{\mathbf{s} \in B(F)} \|\mathbf{s}\|_1\Big), \qquad (12)$$

where $[\mathbf{s}_-]_k = \min\{[\mathbf{s}]_k, 0\}$ for $k \in V$.

We now turn to prove Theorem 3.

Proof. of Theorem 3:

Since $\hat{P}(\hat{\mathbf{w}})$ is 1-strongly convex, for any $\hat{\mathbf{w}} \in dom\hat{P}(\hat{\mathbf{w}})$ and $\hat{\mathbf{w}}^* = \arg\min_{\hat{\mathbf{w}} \in \mathbb{R}^{\hat{p}}} \hat{P}(\hat{\mathbf{w}})$, we can have

$$\hat{P}(\hat{\mathbf{w}}) \geq \hat{P}(\mathbf{w}^*) + \langle \hat{\mathbf{g}}, \hat{\mathbf{w}} - \hat{\mathbf{w}}^* \rangle + \frac{1}{2}\|\hat{\mathbf{w}} - \hat{\mathbf{w}}^*\|_2^2.$$

where $\mathbf{g} \in \partial\hat{P}(\hat{\mathbf{w}}^*)$.

Since $dom\hat{P}(\hat{\mathbf{w}}) = \mathbb{R}^{\hat{p}}$, it holds $0 \in \partial\hat{P}(\hat{\mathbf{w}}^*)$. Hence, we can get

$$\frac{1}{2}\|\hat{\mathbf{w}} - \hat{\mathbf{w}}^*\|_2^2 \leq \hat{P}(\hat{\mathbf{w}}) - \hat{P}(\hat{\mathbf{w}}^*).$$

In addition, we notice that $\hat{P}(\mathbf{w}^*) \geq \hat{D}(\hat{\mathbf{s}})$ for all $\hat{\mathbf{s}} \in B(\hat{F})$. By substituting this inequality into the above inequality, we obtain that

$$\frac{1}{2}\|\hat{\mathbf{w}} - \hat{\mathbf{w}}^*\|_2^2 \leq \hat{P}(\hat{\mathbf{w}}) - \hat{P}(\hat{\mathbf{w}}^*) \leq \hat{P}(\hat{\mathbf{w}}) - \hat{D}(\hat{\mathbf{s}}) = G(\hat{\mathbf{w}}, \hat{\mathbf{s}}).$$

Thus

$$\hat{\mathbf{w}}^* \in \mathcal{B} := \Big\{\mathbf{w} : \|\mathbf{w} - \hat{\mathbf{w}}\| \leq \sqrt{2G(\hat{\mathbf{w}}, \hat{\mathbf{s}})}\Big\}. \qquad (13)$$

According to the equation (Opt) in Theorem 1, we have that $-\hat{\mathbf{w}}^*$ is the optimal solution of the problem Q-D'. Therefore, $-\hat{\mathbf{w}}^* \in B(\hat{F})$. From the definition of $B(\hat{F})$, we have that

$$-\langle \hat{\mathbf{w}}^*, \mathbf{1} \rangle = -\hat{\mathbf{w}}^*(\hat{V}) = \hat{F}(\hat{V})$$

Thus,

$$\hat{\mathbf{w}}^* \in \mathcal{P} := \Big\{ \mathbf{w} : \langle \mathbf{w}, \mathbf{1} \rangle = -\hat{F}(\hat{V}) \Big\}. \tag{14}$$

By section 7.3 of [1]), it holds that the unique minimizer of problem Q-D' is also a maximizer of

$$\max_{\mathbf{s} \in B(\hat{F})} \mathbf{s}_-(V).$$

Hence, it holds

$$\|\hat{\mathbf{s}}^*\|_1 \leq \|\hat{\mathbf{s}}\|_1 \text{ for all } \hat{\mathbf{s}} \in B(\hat{F}). \tag{15}$$

From Lemma 4, we have that

$$\begin{aligned} &\hat{F}(C) \geq \frac{1}{2}(\hat{F}(\hat{V}) - \|\hat{\mathbf{s}}\|_1), \text{for all } \hat{\mathbf{s}} \in B(\hat{F}), \\ &\Rightarrow \|\hat{\mathbf{s}}\|_1 \geq \hat{F}(\hat{V}) - 2\hat{F}(C), \text{for all } \hat{\mathbf{s}} \in B(\hat{F}) \end{aligned} \tag{16}$$

By combining (15) and (16), we get that

$$\hat{F}(\hat{V}) - 2\hat{F}(C) \leq \|\hat{\mathbf{s}}^*\|_1 \leq \|\hat{\mathbf{s}}\|_1, \text{for all } \hat{\mathbf{s}} \in B(\hat{F}).$$

Since $\hat{\mathbf{w}}^* = -\hat{\mathbf{s}}^*$, we have

$$\hat{F}(\hat{V}) - 2\hat{F}(C) \leq \|\hat{\mathbf{w}}^*\|_1 \leq \|\hat{\mathbf{s}}\|_1, \text{for all } \hat{\mathbf{s}} \in B(\hat{F}).$$

Thus, we obtain

$$\hat{\mathbf{w}}^* \in \Omega := \Big\{ \mathbf{w} : \hat{F}(\hat{V}) - 2\hat{F}(C) \leq \|\mathbf{w}\|_1 \leq \|\hat{\mathbf{s}}\|_1 \Big\}. \tag{17}$$

From (13), (14) and (17), we have $\hat{\mathbf{w}}^* \in \mathcal{B} \cap \Omega \cap \mathcal{P}$.

The proof is complete. □

A.4 Proof of Lemma 2

Proof. of Lemma 2:

For any $j = 1, ..., \hat{p}$, we have that

$$\sum_{i \neq j} ([\mathbf{w}]_i - [\hat{\mathbf{w}}]_i)^2 \leq 2G(\hat{\mathbf{w}}, \hat{\mathbf{s}}) - ([\mathbf{w}]_j - ([\hat{\mathbf{w}}]_j)^2, \tag{18}$$

$$\sum_{i \neq j} [\mathbf{w}]_i = -\hat{F}(\hat{V}) - [\mathbf{w}]_j, \tag{19}$$

By fixing the component $[\mathbf{w}]_j$, we can see that (18) and (19) are a ball and a plane in $\mathbb{R}^{\hat{p}-1}$, respectively. To make the intersection of (18) and (19) non-empty, we just need to restrict the distance between the center of ball (18) and the plane (19) smaller than the radius, i.e.,

$$\frac{|\sum_{i \neq j} [\hat{\mathbf{w}}]_i + \hat{F}(\hat{V}) + [\mathbf{w}]_j|}{\sqrt{\hat{p} - 1}} \leq \sqrt{2G(\hat{\mathbf{w}}, \hat{\mathbf{s}}) - ([\mathbf{w}]_j - [\hat{\mathbf{w}}]_j)^2},$$

which is equivalent to

$$\hat{p}[\mathbf{w}]_j^2 + b[\mathbf{w}]_j + c \leq 0, \tag{20}$$

where $b = 2\Big(\sum_{i\neq j}[\hat{\mathbf{w}}]_i + \hat{F}(\hat{V}) - (\hat{p}-1)[\hat{\mathbf{w}}]_j\Big), c = \Big(\sum_{i\neq j}[\hat{\mathbf{w}}]_i + \hat{F}(\hat{V})\Big)^2 - 2(\hat{p}-1)G(\hat{\mathbf{w}},\hat{\mathbf{s}}) + (\hat{p}-1)[\hat{\mathbf{w}}]_j^2$. Thus we have,

$$[\mathbf{w}]_j \in [\frac{-b-\sqrt{b^2-4\hat{p}c}}{2\hat{p}}, \frac{-b+\sqrt{b^2-4\hat{p}c}}{2\hat{p}}]$$

At last, we should point out here that since $\hat{\mathbf{w}}^*$ must be in the intersection of the ball (18) and the plane (19). Hence, inequality (20) can be satisfied with $[\hat{\mathbf{w}}^*]_j$, which implies that $b^2 - 4\hat{p}c$ would never be negative.

The proof is complete. □

A.5 Proof of Theorem 4

Proof. of Theorem 4:

(i): According to $\min_{\mathbf{w}\in\mathcal{B}\cap\mathcal{P}}[\mathbf{w}]_j = [\mathbf{w}]_j^{\min} > 0$ and $\hat{\mathbf{w}}^* \in \mathcal{B}\cap\mathcal{P}$, we have that

$$[\hat{\mathbf{w}}^*]_j > 0.$$

From Theorem 2, it holds $j \in \arg\min_{C\subseteq\hat{V}}\hat{F}(C) \subseteq A^*$.

(ii): Since $\max_{\mathbf{w}\in\mathcal{B}\cap\mathcal{P}}[\mathbf{w}]_j = [\mathbf{w}]_j^{\max} < 0$ and $\hat{\mathbf{w}}^* \in \mathcal{B}\cap\mathcal{P}$, we have that

$$[\hat{\mathbf{w}}^*]_j < 0.$$

From Theorem 2, we have $j \notin \arg\min_{C\subseteq\hat{V}}\hat{F}(C)$. Noting that $A^* = \mathcal{E}\cup\arg\min\hat{F}(C)$ and $j\notin\mathcal{E}$. Therefore $j\notin A^*$.

(iii) It is the immediate conclusion from (i) and (ii).

The proof is complete. □

A.6 Proof of Lemma 3

Proof. of Lemma 3:

(i) We just need to prove that

$$\begin{cases}[\mathbf{w}]_j^{\min} > 0 \text{ if } [\hat{\mathbf{w}}]_j > \sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})},\\ [\mathbf{w}]_j^{\max} < 0 \text{ if } [\hat{\mathbf{w}}]_j < -\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}.\end{cases}$$

We divide the proof into two parts. First, when $[\hat{\mathbf{w}}]_j > \sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}$, consider the definition of $[\mathbf{w}]_j^{\min}$, we have

$$[\mathbf{w}]_j^{\min} = \min_{\mathbf{w}\in\mathcal{B}\cap\mathcal{P}}[\mathbf{w}]_j \geq \min_{\mathbf{w}\in\mathcal{B}}[\mathbf{w}]_j = [\hat{\mathbf{w}}]_j - \sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} > 0.$$

In this case, the element j can be screened by rule AES-1.

On the other hand, when $[\hat{\mathbf{w}}]_j < -\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}$, from the definition of $[\mathbf{w}]_j^{\max}$, we have that

$$[\mathbf{w}]_j^{\max} = \max_{\mathbf{w}\in\mathcal{B}\cap\mathcal{P}}[\mathbf{w}]_j \leq \max_{\mathbf{w}\in\mathcal{B}}[\mathbf{w}]_j = [\hat{\mathbf{w}}]_j + \sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} < 0.$$

In this case, the element j can be screened by rule IES-1.

(ii) We note that the point $\mathbf{v}$ with $[\mathbf{v}]_j = 0$ and $[\mathbf{v}]_k = [\hat{\mathbf{w}}]_k$ for all $k \neq j, k = 1, 2, .., \hat{p}$ belongs to the ball $\mathcal{B}$. Thus, we have

$$\min_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\leq 0} \|\mathbf{w}\|_1 \leq \sum_{i\neq j} |[\hat{\mathbf{v}}]_i| = \|\hat{\mathbf{w}}\|_1 - [\hat{\mathbf{w}}]_j < \|\hat{\mathbf{w}}\|_1.$$

Now, we turn to calculate $\max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\leq 0} \|\mathbf{w}\|_1$.

We note that the range of $[\mathbf{w}]_j$ is $[-\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} + [\hat{\mathbf{w}}]_j, \sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} + [\hat{\mathbf{w}}]_j]$ when $\mathbf{w} \in \mathcal{B}$. Hence, the problem $\max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\leq 0} \|\mathbf{w}\|_1$ can be decomposed into

$$\max_{-\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}+[\hat{\mathbf{w}}]_j\leq\alpha\leq 0} \Big\{ \max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j=\alpha} \|\mathbf{w}\|_1 \Big\}.$$

We assume $[\mathbf{w}]_j = \alpha$ with $-\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} + [\hat{\mathbf{w}}]_j \leq \alpha \leq 0$ and first consider the following problem,

$$\max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j=\alpha} \|\mathbf{w}\|_1,$$

which can be rewritten as

$$\begin{aligned} &\max_{[\mathbf{w}]_i, i\neq j} \; -\alpha + \sum_{i\neq j} |[\mathbf{w}]_i| \\ &s.t. \sum_{i\leq\hat{p}, i\neq j} ([\mathbf{w}]_i - [\hat{\mathbf{w}}]_j)^2 \leq 2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - (\alpha - [\hat{\mathbf{w}}]_j)^2. \end{aligned}$$

It is easy to check that the optimal solution of the problem above is

$$[\mathbf{w}]_i = [\hat{\mathbf{w}}]_i + \mathbf{sign}([\hat{\mathbf{w}}]_i)\sqrt{\frac{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - (\alpha - [\hat{\mathbf{w}}]_j)^2}{\hat{p}-1}}.$$

The function $\mathbf{sign}(\cdot) : \mathbb{R} \to \{-1, 1\}$ above takes 1 if the argument is positive, otherwise takes -1. And the corresponded optimal value is

$$\max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j=\alpha} \|\mathbf{w}\|_1 = -\alpha + \sum_{i\neq j} |[\hat{\mathbf{w}}]_i| + \sqrt{\hat{p}-1}\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - (\alpha - [\hat{\mathbf{w}}]_j)^2}.$$

Now, we denote $h(\alpha) := -\alpha + \sum_{i\neq j} |[\hat{\mathbf{w}}]_i| + \sqrt{\hat{p}-1}\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - (\alpha - [\hat{\mathbf{w}}]_j)^2}$ and turn to solve

$$\max_{-\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}+[\hat{\mathbf{w}}]_j\leq\alpha\leq 0} h(\alpha)$$

If $[\hat{\mathbf{w}}]_j - \sqrt{\frac{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}{\hat{p}}} < 0$, then

$$\begin{aligned} \max_{-\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}+[\hat{\mathbf{w}}]_j\leq\alpha\leq 0} h(\alpha) &= h([\hat{\mathbf{w}}]_j - \sqrt{\frac{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}{\hat{p}}}) = -[\hat{\mathbf{w}}]_j + \sum_{i\neq j} |[\hat{\mathbf{w}}]_i| + \sqrt{2\hat{p}G(\hat{\mathbf{w}},\hat{\mathbf{s}})} \\ &= \|\hat{\mathbf{w}}\|_1 - 2[\hat{\mathbf{w}}]_j + \sqrt{2\hat{p}G(\hat{\mathbf{w}},\hat{\mathbf{s}})}. \end{aligned}$$

Else if $[\hat{\mathbf{w}}]_j - \sqrt{\frac{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}{\hat{p}}} \geq 0$, then

$$\begin{aligned}\max_{-\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}+[\hat{\mathbf{w}}]_j \leq \alpha \leq 0} h(\alpha) = h(0) &= \sum_{i\neq j} |[\hat{\mathbf{w}}]_i| + \sqrt{\hat{p}-1}\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - [\hat{\mathbf{w}}]_j^2} \\ &= \|\hat{\mathbf{w}}\|_1 - [\mathbf{w}]_j + \sqrt{\hat{p}-1}\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - [\hat{\mathbf{w}}]_j^2}.\end{aligned}$$

In consequence, we have

$$\max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\leq 0} \|\mathbf{w}\|_1 = \begin{cases} \|\hat{\mathbf{w}}\|_1 - 2[\hat{\mathbf{w}}]_j + \sqrt{2\hat{p}G(\hat{\mathbf{w}},\hat{\mathbf{s}})}, & \text{if } [\hat{\mathbf{w}}]_j - \sqrt{\frac{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}{\hat{p}}} < 0 \\ \|\hat{\mathbf{w}}\|_1 - [\mathbf{w}]_j + \sqrt{\hat{p}-1}\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - [\hat{\mathbf{w}}]_j^2}, & \text{otherwise,} \end{cases}$$

(iii) Recall that the point $\mathbf{v}$ with $[\mathbf{v}]_j = 0$ and $[\mathbf{v}]_k = [\hat{\mathbf{w}}]_k$ for all $k \neq j, k = 1, 2, .., \hat{p}$ lies in the ball $\mathcal{B}$. Thus, we have

$$\min_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\geq 0} \|\mathbf{w}\|_1 \leq \sum_{i\neq j} |[\hat{\mathbf{v}}]_i| = \|\hat{\mathbf{w}}\|_1 - [\hat{\mathbf{w}}]_j < \|\hat{\mathbf{w}}\|_1.$$

Now, we turn to calculate $\max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\geq 0} \|\mathbf{w}\|_1$.

We note that the range of $[\mathbf{w}]_j$ is $[-\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} + [\hat{\mathbf{w}}]_j, \sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} + [\hat{\mathbf{w}}]_j]$ when $\mathbf{w} \in \mathcal{B}$. Hence, we decompose the problem $\max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\geq 0} \|\mathbf{w}\|_1$ into

$$\max_{0\leq\alpha\leq\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}+[\hat{\mathbf{w}}]_j} \Big\{ \max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j=\alpha} \|\mathbf{w}\|_1 \Big\}.$$

We assume $[\hat{\mathbf{w}}]_j = \alpha$ with $0 \leq \alpha \leq \sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} + [\hat{\mathbf{w}}]_j$ and first solve the following problem,

$$\max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j=\alpha} \|\mathbf{w}\|_1,$$

which can be rewritten as

$$\begin{aligned}&\max_{[\mathbf{w}]_i, i\neq j} \alpha + \sum_{i\neq j} |[\mathbf{w}]_i| \\ &s.t. \sum_{i\leq\hat{p}, i\neq j} ([\mathbf{w}]_i - [\hat{\mathbf{w}}]_j)^2 \leq 2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - (\alpha - [\hat{\mathbf{w}}]_j)^2.\end{aligned}$$

It can be verified that the optimal solution of the problem above is

$$[\mathbf{w}]_i = [\hat{\mathbf{w}}]_i + \mathbf{sign}([\hat{\mathbf{w}}]_i)\sqrt{\frac{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - (\alpha - [\hat{\mathbf{w}}]_j)^2}{\hat{p}-1}}.$$

The function $\mathbf{sign}(\cdot) : \mathbb{R} \to \{-1, 1\}$ above takes 1 if the argument is positive, otherwise takes -1. And the corresponding optimal value is

$$\max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j=\alpha} \|\mathbf{w}\|_1 = \alpha + \sum_{i\neq j} |[\hat{\mathbf{w}}]_i| + \sqrt{\hat{p}-1}\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - (\alpha - [\hat{\mathbf{w}}]_j)^2}.$$

Now, we denote $h(\alpha) := \alpha + \sum_{i\neq j} |[\hat{\mathbf{w}}]_i| + \sqrt{\hat{p}-1}\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - (\alpha - [\hat{\mathbf{w}}]_j)^2}$ and turn to solve

$$\max_{0\leq\alpha\leq\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}+[\hat{\mathbf{w}}]_j} h(\alpha)$$

If $[\hat{\mathbf{w}}]_j + \sqrt{\frac{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}{\hat{p}}} > 0$, then

$$\begin{aligned}\max_{0\leq\alpha\leq\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}+[\hat{\mathbf{w}}]_j} h(\alpha) &= h([\hat{\mathbf{w}}]_j + \sqrt{\frac{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}{\hat{p}}}) = [\hat{\mathbf{w}}]_j + \sum_{i\neq j}|[\hat{\mathbf{w}}]_i| + \sqrt{2\hat{p}G(\hat{\mathbf{w}},\hat{\mathbf{s}})} \\ &= \|\hat{\mathbf{w}}\|_1 + 2[\hat{\mathbf{w}}]_j + \sqrt{2\hat{p}G(\hat{\mathbf{w}},\hat{\mathbf{s}})}.\end{aligned}$$

Else if $[\hat{\mathbf{w}}]_j + \sqrt{\frac{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}{\hat{p}}} \leq 0$, then

$$\begin{aligned}\max_{0\leq\alpha\leq\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}+[\hat{\mathbf{w}}]_j} h(\alpha) = h(0) &= \sum_{i\neq j}|[\hat{\mathbf{w}}]_i| + \sqrt{\hat{p}-1}\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - [\hat{\mathbf{w}}]_j^2} \\ &= \|\hat{\mathbf{w}}\|_1 + [\mathbf{w}]_j + \sqrt{\hat{p}-1}\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - [\hat{\mathbf{w}}]_j^2}.\end{aligned}$$

Consequently, we have,

$$\max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\geq 0}\|\mathbf{w}\|_1 = \begin{cases}\|\hat{\mathbf{w}}\|_1 + 2[\hat{\mathbf{w}}]_j + \sqrt{2\hat{p}G(\hat{\mathbf{w}},\hat{\mathbf{s}})}, & \text{if } [\hat{\mathbf{w}}]_j + \sqrt{\frac{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}{\hat{p}}} > 0 \\ \|\hat{\mathbf{w}}\|_1 + [\mathbf{w}]_j + \sqrt{\hat{p}-1}\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}}) - [\hat{\mathbf{w}}]_j^2}, & \text{otherwise,}\end{cases}$$

The proof is complete. □

A.7 Proof of Theorem 5

Proof. of Theorem 5:

(i): Notice that

$$\begin{cases}0 < [\hat{\mathbf{w}}]_j \leq \sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})}, \\ \max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\leq 0}\|\mathbf{w}\|_1 < \hat{F}(\hat{V}) - 2\hat{F}(C),\end{cases}$$

and $\Omega = \left\{\mathbf{w} : \hat{F}(\hat{V}) - 2\hat{F}(C) \leq \|\mathbf{w}\|_1 \leq \|\hat{\mathbf{s}}\|_1\right\}$, we have that

$$\left\{\mathbf{w}, \mathbf{w}\in\mathcal{B}, [\mathbf{w}]_j \leq 0\right\} \cap \Omega = \emptyset. \tag{21}$$

Since $\hat{\mathbf{w}}^* \in \mathcal{B}\cap\Omega$, from (21), we have $[\hat{\mathbf{w}}^*]_j > 0$. Thus, from Theorem 2, we have $j \in \arg\min \hat{F}(C) \subseteq A^*$.

(ii): Since

$$\begin{cases}-\sqrt{2G(\hat{\mathbf{w}},\hat{\mathbf{s}})} \leq [\hat{\mathbf{w}}]_j < 0, \\ \max_{\mathbf{w}\in\mathcal{B},[\mathbf{w}]_j\geq 0}\|\mathbf{w}\|_1 < \hat{F}(\hat{V}) - 2\hat{F}(C),\end{cases}$$

and $\Omega = \left\{\mathbf{w} : \hat{F}(\hat{V}) - 2\hat{F}(C) \leq \|\mathbf{w}\|_1 \leq \|\hat{\mathbf{s}}\|_1\right\}$, we have

$$\left\{\mathbf{w}, \mathbf{w}\in\mathcal{B}, [\mathbf{w}]_j \geq 0\right\} \cap \Omega = \emptyset. \tag{22}$$

Since $\hat{\mathbf{w}}^* \in \mathcal{B} \cap \Omega$, from (22), we have $[\hat{\mathbf{w}}^*]_j < 0$.

From Theorem 2, we have $j \notin \arg\min_{C \subseteq \hat{V}} \hat{F}(C)$. Noting that $A^* = \mathcal{E} \cup \arg\min_{C \subseteq \hat{V}} \hat{F}(C)$ and $j \notin \mathcal{E}$, therefore $j \notin A^*$.

(iii) It is the immediate conclusion of (i) and (ii).

The proof is complete. □

References

[1] Francis Bach et al. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in Machine Learning*, 6(2-3):145–373, 2013.

[2] Michael J Best and Nilotpal Chakravarti. Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 47(1-3):425–439, 1990.

[3] Jonathan Borwein and Adrian S Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

[4] Volkan Cevher, Marco F Duarte, Chinmay Hegde, and Richard Baraniuk. Sparse signal recovery using markov random fields. In *Advances in Neural Information Processing Systems*, pages 257–264, 2009.

[5] Joseph C Dunn and S Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.

[6] Jack Edmonds. Submodular functions, matroids, and certain polyhedra. *Combinatorial structures and their applications*, pages 69–87, 1970.

[7] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization*, 8:667–698, 2012.

[8] Alina Ene, Huy Nguyen, and László A Végh. Decomposable submodular function minimization: discrete and continuous. In *Advances in Neural Information Processing Systems*, pages 2874–2884, 2017.

[9] Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Mind the duality gap: safer rules for the lasso. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 333–342, 2015.

[10] Satoru Fujishige. *Submodular functions and optimization*, volume 58. Elsevier, 2005.

[11] Satoru Iwata, Lisa Fleischer, and Satoru Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM (JACM)*, 48(4):761–777, 2001.

[12] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[13] Vladimir Kolmogorov and Ramin Zabin. What energy functions can be minimized via graph cuts? *IEEE transactions on pattern analysis and machine intelligence*, 26(2):147–159, 2004.

[14] Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520, 2011.

[15] Hui Lin and Jeff Bilmes. Optimal selection of limited vocabulary speech corpora. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[16] László Lovász. Submodular functions and convexity. In *Mathematical Programming The State of the Art*, pages 235–257. Springer, 1983.

[17] S Thomas McCormick. Submodular function minimization. *Handbooks in operations research and management science*, 12:321–391, 2005.

[18] M. Narasimhan and J. Bilmes. Local search for balanced submodular clusterings. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 981–986, 2007.

[19] Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparse multi-task and multi-class models. In *Advances in Neural Information Processing Systems*, pages 811–819, 2015.

[20] Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparse-group lasso. In *Advances in Neural Information Processing Systems 29*, pages 388–396. 2016.

[21] Kohei Ogawa, Yoshiki Suzuki, and Ichiro Takeuchi. Safe screening of non-support vectors in pathwise svm computation. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1382–1390, 2013.

[22] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[23] Atsushi Shibagaki, Masayuki Karasuyama, Kohei Hatano, and Ichiro Takeuchi. Simultaneous safe screening of features and samples in doubly sparse modeling. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1577–1586, 2016.

[24] 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.

[25] Jie Wang and Jieping Ye. Multi-layer feature reduction for tree structured group lasso via hierarchical projection. In *Advances in Neural Information Processing Systems*, pages 1279–1287, 2015.

[26] 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*, pages 1053–1061, 2014.

[27] Jie Wang, Jiayu Zhou, Peter Wonka, and Jieping Ye. Lasso screening rules via dual polytope projection. In *Advances in Neural Information Processing Systems*, pages 1070–1078, 2013.

[28] Philip Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11(1):128–149, 1976.

[29] 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 the 34th International Conference on Machine Learning*, pages 4016–4025, 2017.