Pacific Journal of Optimization
Vol. 21, No. 2, pp. 215–236, 2025

A FAST ALGORITHM FOR STRUCTURED OPTIMIZATION WITH NONCONVEX $\ell_{Q,P}$ REGULARIZATION*

Tiange Li, Xiangyu Yang† and Hao Wang

Abstract: The iteratively reweighted ℓ_1 (IRL1) algorithm is commonly employed for addressing nonconvex optimization problems mainly through solving a sequence of convex subproblems. In this paper, we propose an enhanced IRL1 algorithm tailored for addressing structured optimization problems involving nonconvex $\ell_{q,p}$ regularization. The key to its acceleration lies in a simple yet effective feature screening strategy. The proposed strategy involves a priori screening test capable of identifying potential inactive groups before executing the subproblem solver and also incorporates an efficient posterior Karush-Kuhn-Tucker condition check procedure to ensure an optimal solution even if some screened variables are mistakenly removed. The priori screening procedure primarily exploits the dual subproblem information at current iteration. Furthermore, we establish a theoretical proof that, within a finite number of IRL1 iterations, the screening test correctly removes all inactive variables. Numerical experiments showcase the significant computational advantages of our algorithm over several state-of-the-art algorithms.

Key words: *nonconvex optimization, group sparsity, screening rules*

Mathematics Subject Classification: *65K05, 90C06, 90C26*

1 Introduction

In modern statistics and statistical machine learning [12, 16, 17, 33], many researchers are interested in solving optimization problems that involve empirical risk minimization with an appropriate penalty term, that is,

$$\boldsymbol{x}^* \in \arg\min_{x\in\mathbb{R}^n} \; f(\boldsymbol{x}) + \lambda\mathcal{R}(\boldsymbol{x}), \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, convex function referring to the error or a data fidelity term, while the regularization function $\mathcal{R} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed and possibly nonconvex and nonsmooth. The parameter $\lambda > 0$ controls the trade-off between data fidelity and regularization. This general framework encompasses a variety of optimization models depending on the choice of f and $\mathcal{R}$, such as the least squares regression (with a squared loss function) and logistic regression (with a logistic loss function). Regularization is primarily employed to induce sparsity or low-rank structures in the solution, thereby enhancing generalization performance in high-dimensional learning settings. Notably, the

*The research of Xiangyu Yang was partially supported by National Natural Science Foundation of China No. 12301398

†Corresponding author

widely recognized Lasso (least absolute shrinkage and selection operator) method, introduced by Tibshirani in a seminal paper [24], employs the ℓ_1-norm regularizer to remove irrelevant features of the concerned model, leading to a more interpretable and parsimonious model.

Since the advent of Lasso, sparsity-promoting regularizers have been a central focus in both the statistics and sparse optimization communities owing to their widespread applications [6, 4, 35]. Over the past decade, nonconvex sparsity-promoting regularizers have attracted considerable attention, exhibiting superior performance compared to their convex counterparts, such as the ℓ_1-norm regularizer. Noteworthy examples of nonconvex regularizers include the smoothly clipped absolute deviation (SCAD) [5], the log-sum penalty [4], the Minimax Concave Penalty (MCP) [35] and the ℓ_p norm penalty with $0 < p < 1$ [9, 8]. These regularizers offer improved flexibility and enhanced performance in various sparse optimization scenarios.

Building on this foundation, researchers have naturally extended sparsity-promoting regularization to structured optimization problems, aiming to induce sparsity at the group level. In this setting, variables that form a predefined group structure are either selected or removed simultaneously. A widely used class of group sparsity-promoting regularizers is based on the nonsmooth $\ell_{p,q}$-norm, where different values for $p > 0$ and $q \geq 1$ yield various formulations. Notable examples include convex regularizers such as the $\ell_{2,1}$-norm and $\ell_{\infty,1}$-norm [34, 1], as well as the nonconvex $\ell_{2,0.5}$-norm [13]. This line of research enhances model interpretability and effectiveness by enabling the simultaneous selection or removal of grouped variables, thereby improving performance across a range of applications.

For problems demanding sparse solutions—where only a small subset of variables contributes to the true support—reducing computational burden involves directing the solver towards active variables. This is typically achieved through techniques such as screening strategies and working set methods. Screening rules identify and discard variables that are likely to be zeros in the optimal solution. A rule is considered "safe" if it guarantees the correct identification of inactive variables [11]. In the context of Lasso-type problems, Ghaoui et al. [11] introduced the SAFE screening rule to discard irrelevant features. Tibshirani et al. [25] proposed a heuristic strong rule, building on SAFE, while Lee et al. [15] extended it to general linear models incorporating nonconvex MCP and SCAD regularization. Further advancements include dynamic safe screening rules [2, 19, 21], which exploited the duality gap during optimization to iteratively discard inactive variables within the solver. Collectively, these techniques significantly enhance computational efficiency in sparse optimization problems.

Alternatively, the working set technique begins by heuristically selecting a small subset of variables as the initial working set and gradually expands it until the optimality condition is satisfied. This approach, which has been applied in SVM training for sample filtering [26], has evolved in recent works. [14] introduced the convergent working set method "Blitz" for Lasso, addressing the dual constrained problem. In a subsequent study, Massias et al. [18] proposed an efficient working set technique based on a novel dual extrapolation strategy. Building on these contributions, [20] extended "Blitz" to solve problems with nonconvex regularizers. These developments collectively enhance the efficiency of the working set technique in solving various optimization problems.

Addressing the nonconvex $\ell_{q,p}$-regularized problem [13] with $q \geq 1$ and $0 < p < 1$ is the central focus of this paper, and efficiently solving large-scale instances is our primary motivation. To enhance computational efficiency, feature-reducing methods such as screening rules and working set techniques offer promising strategies. However, adapting these existing methods to handle the nonconvex ℓ_p norm is challenging due to its nonconvex and non-Lipschitz properties. A promising approach is to reformulate the original problem into

a sequence of convex subproblems. In this context, we explore the iteratively reweighted ℓ_1 algorithm, which solves a sequence of tractable weighted ℓ_1-norm regularized subproblems [4, 10]. This method aligns with the approach presented in [21], which incorporates the gap safe rule [19] within a majorization-minimization framework. Moreover, discarding irrelevant features before the solver commences is crucial. In [18], the gap safe rule introduced in [19] was shown to be inefficient at the beginning of the solver due to a large initial duality gap. Recognizing potential inefficiencies, we aim to develop an efficient screening rule for the reweighted ℓ_1 subproblem that prioritizes both accuracy and speed.

In this paper, we propose an enhanced IRL1 framework incorporates a novel screening rule strategy. Our approach consists of two modules: a heuristic screening test that identifies zero variables prior to solving each subproblem, and a posterior module that verifies optimal subproblem solutions. The screening test leverages dual information from the weighted ℓ_1 subproblem to reduce the input data dimension and accelerate each subproblem solution, while the posterior module conducts a simple Karush-Kuhn-Tucker (KKT) check to confirm exact solutions in the reduced space. Notably, our screening rule operates across successive iterations, demonstrating its ability to identify and filter zero variables within a finite number of iterations. Numerical studies demonstrate the substantial computational gains achieved by the IRL1 algorithm with our proposed screening rule, highlighting its effectiveness in real-world applications.

1.1 Notation and Preliminaries

Throughout this paper, we restrict our discussion to the real n-dimensional Euclidean space $\mathbb{R}^n$. We denote by $\mathbb{N}$ the set of natural numbers and use $[n] \subset \mathbb{N}$ to represent the index set $\{1, 2, \ldots, n\}$. For $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we denote by $\boldsymbol{x}_{[\mathcal{S}]}$ the subvector of $\boldsymbol{x}$ indexed by an index set $\mathcal{S} \subset [n]$, and by $\boldsymbol{A}_{[\mathcal{S}]}$ the submatrix of $\boldsymbol{A}$ formed by the columns indexed by $\mathcal{S}$. Let $\mathbf{0}$ denote the zero vector of appropriate size in the given context. We define the active set of a vector $\boldsymbol{x} \in \mathbb{R}^n$ as $\mathcal{A}(\boldsymbol{x}) = \{i \in [n] \mid x_i \neq 0\}$, and its complementary set, the inactive set, as $\mathcal{I}(\boldsymbol{x}) = \{i \in [n] \mid x_i = 0\}$. For a vector with group structure $\boldsymbol{x}_{\mathcal{G}} \in \mathbb{R}^n$, we denote the inactive set by $\mathcal{I}(\boldsymbol{x}) = \{i \in [d] \mid \boldsymbol{x}_{\mathcal{G}_i} = \mathbf{0}\}$, and the active set by $\mathcal{A}(\boldsymbol{x}) = \{i \in [d] \mid \boldsymbol{x}_{\mathcal{G}_i} \neq \mathbf{0}\}$. Finally, we use the notation $\boldsymbol{x} \sim \mathcal{N}(\mu, \sigma^2)$ to indicate that a Gaussian random variable with mean μ and variance σ^2.

The following definition is adapted from [22, Definition 8.3] and is used in our analysis.

Definition 1.1. Consider a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and a point $\bar{\boldsymbol{x}}$ with $f(\bar{\boldsymbol{x}})$ finite. For a vector $\boldsymbol{v} \in \mathbb{R}^n$, one says that

(a) $\boldsymbol{v}$ is a regular subgradient of f at $\bar{\boldsymbol{x}}$, written $\boldsymbol{v} \in \hat{\partial} f(\bar{\boldsymbol{x}})$, if

$$f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}) + \langle \boldsymbol{v}, \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle + o(\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|);$$

(b) $\boldsymbol{v}$ is a (general) subgradient of f at $\bar{\boldsymbol{x}}$, written $\boldsymbol{v} \in \partial f(\bar{\boldsymbol{x}})$, if there are sequences $\boldsymbol{x}^\nu \xrightarrow{f} \bar{\boldsymbol{x}}$ and $\boldsymbol{v}^\nu \to \hat{\partial} f(\boldsymbol{x}^\nu)$ with $\boldsymbol{v}^\nu \to \boldsymbol{v}$.

(c) $\boldsymbol{v}$ is a horizon subgradient of f at $\bar{\boldsymbol{x}}$, written $\boldsymbol{v} \in \partial^\infty f(\bar{\boldsymbol{x}})$, if the same holds as in (b), except that instead of $\boldsymbol{v}^\nu \to \boldsymbol{v}$, one has $\lambda^\nu \boldsymbol{v}^\nu \to \boldsymbol{v}$ for some sequence $\lambda^\nu \searrow 0$.

According to [22, Theorem 8.6 & Eq. 8(5)], the following subgradient relationships hold:

$$\partial f(\bar{\boldsymbol{x}}) = \limsup_{\boldsymbol{x} \xrightarrow{f} \bar{\boldsymbol{x}}} \hat{\partial} f(\bar{\boldsymbol{x}}) \quad \text{and} \quad \hat{\partial} f(\bar{\boldsymbol{x}}) \subset \partial f(\bar{\boldsymbol{x}}),$$

and moreover, $\partial^\infty f(\bar{\boldsymbol{x}})$ is a closed cone. When f is proper and convex, [22, Proposition 8.12] indicates that

$$\partial f(\bar{\boldsymbol{x}}) = \hat{\partial} f(\bar{\boldsymbol{x}}). \tag{1.2}$$

The subdifferential of the convex function $\|\boldsymbol{x}\|$ is given by

$$\partial\|\boldsymbol{x}\| = \{\boldsymbol{v} \in \mathbb{R}^n \mid \langle \boldsymbol{v}, \boldsymbol{x}\rangle = \|\boldsymbol{x}\|,\ \|\boldsymbol{v}\|_* \le 1\}, \tag{1.3}$$

where the dual norm is defined as $\|\boldsymbol{x}\|_* = \sup_{\|\boldsymbol{u}\|\le 1}\langle \boldsymbol{x}, \boldsymbol{u}\rangle$. In particular, the dual of the ℓ_q-norm with $q \ge 1$ is the $\ell_{q'}$-norm, where $q' = q/(q-1)$.

The classical Fermat's rule extends to the nonsmooth setting, as stated in [22, Theorem 10.1].

Theorem 1.2. *If a proper function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ has a local minimum at $\bar{\boldsymbol{x}}$, then $\mathbf{0} \in \partial f(\bar{\boldsymbol{x}})$.*

1.2 Nonconvex Structured Optimization Problem

In this section, we provide a concise overview of the structured optimization model incorporating nonconvex $\ell_{q,p}$ regularization. We then introduce an IRL1 algorithm for solving this problem.

The structured optimization problem under consideration, involving a squared loss term and nonconvex pth power of $\ell_{q,p}$ regularization with $q \ge 1$ and $0 < p < 1$, is formulated as

$$\min_{\boldsymbol{x}\in\mathbb{R}^n}\ \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{x}\|_{q,p}^p, \tag{$\mathcal{P}$}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ ($m < n$ is typically assumed) refers to the feature matrix with column-wise feature $\boldsymbol{a}_j \in \mathbb{R}^m$, $\forall j \in [n]$ and $\boldsymbol{y} \in \mathbb{R}^m$ is a given observation vector. The nonconvex $\ell_{q,p}$ regularization term is defined as $\|\boldsymbol{x}\|_{q,p}^p := \sum_{i=1}^d \|\boldsymbol{x}_{\mathcal{G}_i}\|_q^p$, where $\boldsymbol{x} \in \mathbb{R}^n$ is partitioned into $d > 0$ non-overlapping groups, denoted as $[\boldsymbol{x}_{\mathcal{G}_1}, \ldots, \boldsymbol{x}_{\mathcal{G}_d}]^T$ with $\mathcal{G} = \{\mathcal{G}_i\}_{i=1}^d$ forming a partition of $[n]$. The corresponding grouped features are represented by $\boldsymbol{A}_{\mathcal{G}_i} \in \mathbb{R}^{m\times|\mathcal{G}_i|}$. The derivation of the subdifferential of $\|\boldsymbol{x}\|_{q,p}^p$ is provided in the Appendix. By Theorem 1.2, the first-order necessary optimality condition for problem ($\mathcal{P}$) is given by

$$\mathbf{0} \in \boldsymbol{A}_{\mathcal{G}_j}^T(\boldsymbol{A}_{\mathcal{G}_j}\boldsymbol{x}_{\mathcal{G}_j}^\star - \boldsymbol{y}) + \lambda p(\|\boldsymbol{x}_{\mathcal{G}_j}^\star\|_q)^{p-1}\partial\|\boldsymbol{x}_{\mathcal{G}_j}^\star\|_q, \quad \forall j \in \mathcal{A}(\boldsymbol{x}_{\mathcal{G}}^\star). \tag{1.4}$$

In this work, we focus on employing the IRL1 algorithm, as proposed in [28, 29, 30, 31], to solve ($\mathcal{P}$). The IRL1 algorithm is a specific instance of the majorization-minimization framework. To overcome the nonsmoothness of the $\ell_{q,p}$-norm, a perturbation $\boldsymbol{\epsilon} \in \mathbb{R}_{++}^d$ is added at each iteration, yielding a continuously differentiable approximation. At the kth iteration, it holds that

$$\sum_{i=1}^d(\|\boldsymbol{x}_{\mathcal{G}_i}\|_q+\epsilon_i^k)^p \le \sum_{i=1}^d(\|\boldsymbol{x}_{\mathcal{G}_i}^k\|_q + \epsilon_i^k)^p + p(\|\boldsymbol{x}_{\mathcal{G}_i}^k\|_q + \epsilon_i^k)^{p-1}(\|\boldsymbol{x}_{\mathcal{G}_i}\|_q - \|\boldsymbol{x}_{\mathcal{G}_i}^k\|_q), \tag{1.5}$$

where the inequality follows from the concavity of $(\cdot)^p$ over $\mathbb{R}_+$. Consequently, the regularization term in the objective function is replaced by the right-hand side of (1.5), leading to the following subproblem at the kth iteration for updating $\boldsymbol{x}^{k+1}$. That is,

$$\boldsymbol{x}^{k+1} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n}\left\{\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|_2^2+\lambda\sum_{i=1}^d w_i^k\|\boldsymbol{x}_{\mathcal{G}_i}\|_q\right\}, \tag{1.6}$$

where $w_i^k = p(\|\boldsymbol{x}_{\mathcal{G}_i}^k\|_q + \epsilon_i^k)^{p-1}$. As the algorithm proceeds, the perturbation $\boldsymbol{\epsilon}$ is driven to $\mathbf{0}$ to ensure global convergence. For completeness, we summarize the IRL1 algorithm in Algorithm 1.

Algorithm 1 An Iteratively Reweighted ℓ_1 Algorithm for Solving $(\mathcal{P})$

Require: $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{y} \in \mathbb{R}^m$, $\lambda \in \mathbb{R}_{++}$, $\mu \in (0,1)$, $\boldsymbol{\epsilon}^0 \in \mathbb{R}^d_{++}$ and $\boldsymbol{x}^0 \in \mathbb{R}^n$.
1: Set $k = 0$.
2: **repeat**
3: Compute $w_i^k = p(\|\boldsymbol{x}_{\mathcal{G}_i}^k\|_q + \epsilon_i^k)^{p-1}$, $\forall i \in [d]$.
4: Solve (1.6) for $\boldsymbol{x}^{k+1}$.
5: Set $\boldsymbol{\epsilon}^{k+1} \leq \mu\boldsymbol{\epsilon}^k$ and set $k \leftarrow k+1$.
6: **until** convergence

2 Proposed Screening Rule

In this section, we develop a novel screening rule designed to identify and filter inactive groups in the optimal solution of the subproblem solver. By applying this screening strategy, the subproblem can be solved in a reduced space, leading to a significant acceleration of the overall computational process.

2.1 A Priori Screening Test Procedure

The proposed heuristic screening rule is motivated by exploiting the dual information of the subproblem associated with (1.6). Specifically, dropping the superscript k, the kth primal subproblem (1.6) can be rewritten in the compact form

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; P(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \sum_{i=1}^{d} \lambda_i \|\boldsymbol{x}_{\mathcal{G}_i}\|_q, \tag{2.1}$$

where $\lambda_i = \lambda w_i > 0$ is the group-wise regularization parameter for each $i \in [d]$. Let $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}$, problem (2.1) can be equivalently reformulated as

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{z}\in\mathbb{R}^m} \; & \frac{1}{2}\|\boldsymbol{z}\|_2^2 + \langle \boldsymbol{\lambda}, \|\boldsymbol{x}_{\mathcal{G}}\|_q \rangle \\ \text{s.t.} \quad & \boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}, \end{aligned} \tag{2.2}$$

where $\|\boldsymbol{x}_{\mathcal{G}}\|_q = [\|\boldsymbol{x}_{\mathcal{G}_1}\|_q, \ldots, \|\boldsymbol{x}_{\mathcal{G}_d}\|_q]^T$ and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_d]^T$. The Lagrangian associated with (2.2) reads

$$\mathcal{L}_{\boldsymbol{\lambda}}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{z}^T\boldsymbol{z} + \langle \boldsymbol{\lambda}, \|\boldsymbol{x}_{\mathcal{G}}\|_q \rangle + \boldsymbol{\theta}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} - \boldsymbol{z}),$$

where $\boldsymbol{\theta} \in \mathbb{R}^m$ is the Lagrange multiplier associated with (2.2). Decomposing the Lagrangian into two terms,

$$\mathcal{L}_1(\boldsymbol{z}) = \frac{1}{2}\boldsymbol{z}^T\boldsymbol{z} - \boldsymbol{\theta}^T\boldsymbol{z} - \boldsymbol{\theta}^T\boldsymbol{y}, \quad \mathcal{L}_2(\boldsymbol{x}) = \langle \boldsymbol{\lambda}, \|\boldsymbol{x}_{\mathcal{G}}\|_q \rangle + \boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{x}.$$

The Lagrange dual function $G : \mathbb{R}^m \to \mathbb{R}$ is

$$G(\boldsymbol{\theta}) = \inf \mathcal{L}_1(\boldsymbol{z}) + \inf \mathcal{L}_2(\boldsymbol{x}).$$

Since $\mathcal{L}_1(\boldsymbol{z})$ is a convex quadratic function with respect to $\boldsymbol{z}$, its infimum is attained at

$$\inf \mathcal{L}_1(\boldsymbol{z}) = -\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} - \boldsymbol{\theta}^T\boldsymbol{y} = -\frac{1}{2}\|\boldsymbol{\theta}+\boldsymbol{y}\|_2^2 + \frac{1}{2}\|\boldsymbol{y}\|_2^2.$$

On the other hand, for $\mathcal{L}_2(\boldsymbol{x})$, we have

$$\begin{aligned}\mathcal{L}_2(\boldsymbol{x}) &= \sum_{i=1}^{d}(\lambda_i\|\boldsymbol{x}_{\mathcal{G}_i}\|_q + (\boldsymbol{A}_{\mathcal{G}_i}^T\boldsymbol{\theta})^T\boldsymbol{x}_{\mathcal{G}_i}) \geq \sum_{i=1}^{d}(\lambda_i\|\boldsymbol{x}_{\mathcal{G}_i}\|_q - |(\boldsymbol{A}_{\mathcal{G}_i}^T\boldsymbol{\theta})^T\boldsymbol{x}_{\mathcal{G}_i}|)\\ &\geq \sum_{i=1}^{d}(\lambda_i\|\boldsymbol{x}_{\mathcal{G}_i}\|_q - \|\boldsymbol{A}_{\mathcal{G}_i}^T\boldsymbol{\theta}\|_{q'}\|\boldsymbol{x}_{\mathcal{G}_i}\|_q) = \sum_{i=1}^{d}\|\boldsymbol{x}_{\mathcal{G}_i}\|_q(\lambda_i - \|\boldsymbol{A}_{\mathcal{G}_i}^T\boldsymbol{\theta}\|_{q'}),\end{aligned}$$

where the second inequality holds by Hölder's inequality, and $q' = q/(q-1)$ is the Hölder conjugate of q. For the dual function, we concentrate on the case where $\lambda_i \geq \|\boldsymbol{A}_{\mathcal{G}_i}^T\boldsymbol{\theta}\|_{q'}, \forall i \in [d]$, since otherwise, $\inf_{\boldsymbol{x}} \mathcal{L}_2(\boldsymbol{x})$ would be unbounded below. Thus, the Lagrange dual problem of (2.1) is formulated as

$$\begin{aligned}\max_{\boldsymbol{\theta}\in\mathbb{R}^m} \quad & G(\boldsymbol{\theta}) = -\frac{1}{2}\|\boldsymbol{\theta}+\boldsymbol{y}\|_2^2 + \frac{1}{2}\|\boldsymbol{y}\|_2^2\\ \text{s.t.} \quad & \lambda_i \geq \|\boldsymbol{A}_{\mathcal{G}_i}^T\boldsymbol{\theta}\|_{q'}, \quad \forall i \in [d].\end{aligned} \tag{2.3}$$

We introduce the proposed screening rule by first considering an extreme case where the optimal solution $\boldsymbol{x}^\star$ of (2.1) is $\mathbf{0}$. Let $\boldsymbol{\lambda}^0$ denote the corresponding regularization parameter. By the strong duality theorem [3, §5.2.3], at $(\boldsymbol{x}^\star, \boldsymbol{\theta}^\star)$, we have $\boldsymbol{\theta}^\star = -\boldsymbol{y}$ since the primal objective in (2.1) equals the dual objective in (2.3) at optimality. Meanwhile, the dual feasibility condition for $\boldsymbol{\theta}^\star$ gives

$$\|\boldsymbol{A}_{\mathcal{G}_i}^T\boldsymbol{\theta}^\star\|_{q'} = \|\boldsymbol{A}_{\mathcal{G}_i}^T\boldsymbol{y}\|_{q'} \leq \lambda_i^0, \ \forall i \in [d]. \tag{2.4}$$

This leads to the following lemma.

Lemma 2.1. *For problem* (2.1)*, the solution is* $\mathbf{0}$ *if and only if*

$$\mathbf{0} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} P(\boldsymbol{x}) \iff \lambda_i \geq \|\boldsymbol{A}_{\mathcal{G}i}^T\boldsymbol{y}\|_{q'}, \ \forall i \in [d]. \tag{2.5}$$

Proof. (sufficiency). By strong duality, at $(\boldsymbol{x}^\star, \boldsymbol{\theta}^\star)$, we have $\boldsymbol{\theta}^\star = -\boldsymbol{y}$. The feasibility of $\boldsymbol{\theta}^\star$ then directly implies the desired result.

(Necessity). It follows from the optimality condition of (2.1) and [22, Eq. 10(6)] that there exists $\boldsymbol{\beta}_i \in \partial\|\boldsymbol{x}_{\mathcal{G}_i}^\star\|_q$ such that

$$\mathbf{0} = \boldsymbol{A}_{\mathcal{G}_i}^T(\boldsymbol{A}_{\mathcal{G}_i}\boldsymbol{x}_{\mathcal{G}_i}^\star - \boldsymbol{y}) + \lambda_i\boldsymbol{\beta}_i, \ \forall i \in [d]. \tag{2.6}$$

At $\boldsymbol{x}^\star = \mathbf{0}$, condition (2.6) simplifies to $\mathbf{0} = -\boldsymbol{A}_{\mathcal{G}_i}^T\boldsymbol{y} + \lambda_i\boldsymbol{\beta}_i, \ \forall i \in [d]$. Next, we have that

$$\|\boldsymbol{\beta}_i\|_{q'} = \frac{1}{\lambda_i}\|\boldsymbol{A}_{\mathcal{G}_i}^T\boldsymbol{y}\|_{q'} \leq 1, \forall i \in [d],$$

where the inequality holds by the assumption $\lambda_i \geq \|\boldsymbol{A}_{\mathcal{G}_i}^T\boldsymbol{y}\|_{q'}$. From (1.3), it holds that $\boldsymbol{\beta}_i \in \partial\|\mathbf{0}\|_q, \forall i \in [d]$. Therefore, we can write

$$\mathbf{0} \in -\boldsymbol{A}_{\mathcal{G}_i}^T\boldsymbol{y} + \lambda_i\partial\|\mathbf{0}\|_q, \forall i \in [d].$$

Finally, since the first-order optimality condition (2.6) is satisfied, it follows that $\mathbf{0} \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} P(\boldsymbol{x})$. This completes the proof. □

Let $\tilde{\boldsymbol{\lambda}}^{\max} = \|\boldsymbol{A}_{\mathcal{G}_i}^T \boldsymbol{y}\|_{q'}, \forall i \in [d]$. By Lemma 2.1, we know $\boldsymbol{x}^\star = \mathbf{0}$ in the primal problem (2.1) if $\lambda_i \geq \tilde{\lambda}_i^{\max}, \forall i \in [d]$. This leads to the following ideal screening rule:

$$\forall i \in [d],\ \lambda_i \geq \tilde{\lambda}_i^{\max} \implies \boldsymbol{x}^\star = \mathbf{0}. \tag{NR}$$

However, this ideal screening rule (NR) is often too stringent, as the conditions it requires are difficult to implement in practice. To address this, we propose a more practical screening condition, referred to as the weak rule (WR), for filtering out inactive groups. Specifically, for any $i \in [d]$, we use the following condition:

$$\lambda_i \geq \tilde{\lambda}_i^{\max} \implies \boldsymbol{x}_{\mathcal{G}_i}^\star = \mathbf{0}. \tag{WR}$$

In a pioneering work [25], the authors proposed a strong screening rule for Lasso-type problems. For the problem in (1.6), the strong rule suggests discarding variables $\boldsymbol{x}_{\mathcal{G}_i}$ if $\|\boldsymbol{A}_{\mathcal{G}_i}^T \boldsymbol{y}\|_{q'} < w_i(2\lambda - \lambda_{\max})$, where λ is the tuning parameter in (1.6), and $\lambda_{\max}$ is the smallest tuning parameter that yields the solution $\mathbf{0}$. Specifically, their rule can be equivalently written as:

$$\|\boldsymbol{A}_{\mathcal{G}_i}^T \boldsymbol{y}\|_{q'} < 2\lambda_i - w_i \max_j \left\{ \frac{\|\boldsymbol{A}_{\mathcal{G}_j}^T \boldsymbol{y}\|_{q'}}{w_j} \right\}. \tag{2.7}$$

It is clear that our proposed screening rule (WR) is notably simpler compared to the strong rule in (2.7). This simplicity stems from the fact that (WR) does not require the computation of $\max_j \left\{ (\|\boldsymbol{A}_{\mathcal{G}_j}^T \boldsymbol{y}\|_{q'})/w_j \right\}$. In contrast, the right-hand side of the strong rule, as given by (2.7), explicitly necessitates the regularization parameter λ satisfy $\lambda > \frac{1}{2} \max_j \left\{ (\|\boldsymbol{A}_{\mathcal{G}_j}^T \boldsymbol{y}\|_{q'})/w_j \right\}$ to ensure non-negativity. This condition can, in practice, limit the applicability of the strong rule, as it requires a specific range for the regularization parameter. On the other hand, our proposed rule in (WR) imposes no such restriction on λ, allowing for more flexible screening of potentially inactive feature groups.

We should highlight that the proposed screening rule (WR) is practical yet efficient, as confirmed by our numerical studies. Additionally, the heuristic embedded in (WR) requires an optimality check to ensure that all variables are correctly discarded in the optimal solution. To guarantee the reliability of our screening strategy, we employ the KKT conditions of (2.1) to ensure the safety of the proposed screening strategy. This will be discussed in the next subsection.

2.2 A Posterior KKT Check Procedure

To prevent mistakenly discarding inactive feature groups, we follow a similar spirit in [25] by incorporating an optimality check. In particular, note that the KKT optimality condition for (2.1) is

$$\mathbf{0} \in \boldsymbol{A}_{\mathcal{G}_i}^T \left(\sum_{i=1}^d \boldsymbol{A}_{\mathcal{G}_i} \boldsymbol{x}_{\mathcal{G}_i} - \boldsymbol{y} \right) + \lambda_i \partial \|\boldsymbol{x}_{\mathcal{G}_i}\|_q, \forall i \in [d]. \tag{2.8}$$

Suppose we set $\boldsymbol{x}_{\mathcal{G}_i} = \mathbf{0}$. Then there exists $\boldsymbol{\beta} \in \partial \|\mathbf{0}\|_q$ such that

$$-\boldsymbol{A}_{\mathcal{G}_i}^T \left(\sum_{i=1}^d \boldsymbol{A}_{\mathcal{G}_i} \boldsymbol{x}_{\mathcal{G}_i} - \boldsymbol{y} \right) = \lambda_i \boldsymbol{\beta}. \tag{2.9}$$

Thus, we can exploit $\boldsymbol{\beta}$ to check whether the condition (2.9) is satisfied. By (1.3), we can check whether

$$\left\| \frac{-\boldsymbol{A}_{\mathcal{G}_i}^T (\sum_{i=1}^d \boldsymbol{A}_{\mathcal{G}_i} \boldsymbol{x}_{\mathcal{G}_i} - \boldsymbol{y})}{\lambda_i} \right\|_{q'} \leq 1, \forall i \in \mathcal{I}(\boldsymbol{x}).$$

Equivalently, this condition can be written as

$$\left\| \boldsymbol{A}_{\mathcal{G}_i}^T \left(\sum_{i=1}^d \boldsymbol{A}_{\mathcal{G}_i} \boldsymbol{x}_{\mathcal{G}_i} - \boldsymbol{y} \right) \right\|_{q'} \leq \lambda_i, \forall i \in \mathcal{I}(\boldsymbol{x}). \tag{2.10}$$

We use this KKT check, denoted by (2.10), to detect any incorrectly discarded variables after applying the proposed screening rule (WR). Specifically, if (2.10) is violated for a particular group, then—as suggested in [25]—that group is added back, and the problem (2.1) is resolved until the condition (2.10) holds. In practice, rather than running multiple rounds of full optimization, a more efficient strategy is to warm-start the original problem using the solution obtained in the reduced space, requiring at most one additional optimization cycle. This is an improvement over methods such as that in [19], which rely on two complete optimization cycles when using heuristic screening rules. Moreover, since the computations of $\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{A}^T\boldsymbol{y}$ can be precomputed, the overall computational burden of the KKT checking procedure in (2.10) is further reduced.

Overall, the IRL1 algorithm equipped with the proposed screening strategy is summarized in Algorithm 2.

In Algorithm 2, we develop an efficient screening rule to accelerate the solution of the weighted ℓ_1 subproblem. Unlike most existing dynamic screening rules [2, 7, 19, 21], our screening rule is applied as a preprocessing step before initiating the solver for (2.1) at each iteration. This approach enables us to combine our rule with dynamic screening techniques. Meanwhile, the working set strategy acts as a meta-algorithm that iteratively solves reduced-dimensional subproblems. Consequently, one can first use the proposed screening rule to trim features from the input matrix, and then apply an established working set method—augmented by an efficient solver and dynamic screening—to tackle the resulting lower-dimensional problem. This synergistic combination enhances the overall efficiency of the solution process.

3 IRL1 with Proposed Screening Strategy

In this section, we present the theoretical analysis for the proposed screening rule integrated into the IRL1 framework. We first show that groups filtered out by the screening rule are fully identified in the next iteration. Next, we establish that the rule detects all inactive groups within a finite number of iterations. Additionally, we extend the working set strategy from [18] to the weighted ℓ_1 subproblem (1.6) and present a streamlined framework that serves as an effective guide.

3.1 Screening Between Iterations

We first state the IRL1 algorithm incorporating our proposed screening rule in Algorithm 3.

In Step 5 of Algorithm 3, we employ the proximal gradient method introduced in [32] to solve (2.1), which admits efficient soft-thresholding operations. Moreover, we adopt the

Algorithm 2 Proposed Screening Strategy

Require: $\boldsymbol{\lambda}_{\mathcal{G}}$, $\boldsymbol{A}$, $\boldsymbol{y}$, an index set $list \subset [d]$ representing the currently active groups, and a screening set $scrlist \subset [d]$.

Screening Condition:
for $j \in list$ **do**
  if the condition (WR) holds for group j **then**
    $list \leftarrow list \setminus \{j\}$ and $scrlist \leftarrow scrlist \cup \{j\}$.
  end if
end for
Subproblem Solution:
Initialize $\boldsymbol{x} \leftarrow \boldsymbol{0}$.
Solve (2.1) with $\boldsymbol{A}_{[list]}$ to obtain $\hat{\boldsymbol{x}}$ and set $\boldsymbol{x}_{[list]} \leftarrow \hat{\boldsymbol{x}}$.
KKT Check:
Initialize an empty error set: $errlist \leftarrow \emptyset$.
for $i \in scrlist$ **do**
  if the condition (2.10) is not satisfied for group i **then**
    $errlist \leftarrow errlist \cup \{i\}$.
  end if
end for
if $errlist \neq \emptyset$ **then**
  Warm-start by solving (2.1) over all groups with $\boldsymbol{A}$ to obtain an updated solution $\boldsymbol{x}$.
  for $i \in scrlist$ **do**
    if $\boldsymbol{x}_{\mathcal{G}_i} \neq \boldsymbol{0}$ **then**
      $scrlist \leftarrow scrlist \setminus \{i\}$ and $list \leftarrow list \cup \{i\}$
    end if
  end for
end if
Output: $\boldsymbol{x}$, *list*, and *scrlist*.

warm-start technique described in [32], using the solution from the previous iteration as the initialization for the subsequent subproblem to accelerate convergence.

The following lemma states that once a group is added to the screened list in the current iteration, it is guaranteed to remain screened (i.e., be detected as inactive) in the next iteration. As a result, in the subsequent subproblem, the screening procedure only needs to verify those groups that were previously considered active.

Lemma 3.1. *Let $\mathcal{G}_{\mathcal{S}}$ with $\mathcal{S} \subset [d]$, be the set of groups that have been screened by Algorithm 2 at the kth subproblem. Then, for each $i \in \mathcal{S}$, it holds that*

$$\|\boldsymbol{A}_{\mathcal{G}_i}^T \boldsymbol{y}\|_{q'} < \lambda_i^{k+1}.$$

Proof. At the kth subproblem, for any $i \in \mathcal{S}$, the screening rule ensures that $\boldsymbol{x}_{\mathcal{G}_i}^{k+1} = \boldsymbol{0}$. Consequently, we have

$$\|\boldsymbol{x}_{\mathcal{G}_i}^{k+1}\|_q + \epsilon_i^{k+1} = \epsilon_i^{k+1} < \|\boldsymbol{x}_{\mathcal{G}_i}^{k}\|_q + \epsilon_i^{k},$$

where the inequality holds since the perturbation parameter $\boldsymbol{\epsilon}$ decreases at each iteration. Hence,

$$\lambda_i^{k+1} = \lambda p(\|\boldsymbol{x}_{\mathcal{G}_i}^{k+1}\|_q + \epsilon_i^{k+1})^{p-1} > \lambda p(\|\boldsymbol{x}_{\mathcal{G}_i}^{k}\|_q + \epsilon_i^{k})^{p-1} = \lambda_i^{k},$$

Algorithm 3 IRL1 with Proposed Screening Strategy

Require: $\mu \in (0,1)$, $\lambda > 0$, $\boldsymbol{x}^0 \in \mathbb{R}^n_{++}$, $\boldsymbol{\epsilon}^0 \in \mathbb{R}^d_{++}$, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{y} \in \mathbb{R}^m$.
1: Set $k = 0$, *list* $\leftarrow [d]$ (all groups active), *scrlist* $\leftarrow \emptyset$ (no groups screened).
2: **repeat**
3: Compute $w^k_{\mathcal{G}_i} = p(\|\boldsymbol{x}^k_{\mathcal{G}_i}\|_q + \epsilon^k_i)^{p-1}$, $\forall i \in [d]$.
4: Set $\lambda^k_i \leftarrow \lambda w^k_{\mathcal{G}_i}$, $\forall i \in [d]$.
5: Call Algorithm 2 to obtain $\boldsymbol{x}^{k+1}$ and *list* and *scrlist*.
6: Set $\boldsymbol{\epsilon}^{k+1} \leq \mu\boldsymbol{\epsilon}^k$ and set $k \leftarrow k+1$.
7: **until** convergence

where the inequality holds since $0 < p < 1$ and $(\cdot)^{p-1}$ monotonically decreases on $\mathbb{R}_{++}$. Moreover, the screening rule applied at iteration k guarantees that $\|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{y}\|_{q'} \leq \lambda^k_i$. Thus, we have

$$\|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{y}\|_{q'} \leq \lambda^k_i < \lambda^{k+1}_i,$$

which completes the proof. □

3.2 Analysis

The global convergence of the basic IRL1 algorithm for general nonconvex problems with convex constraints has been established in [31], and these convergence properties extend to $(\mathcal{P})$. Importantly, the proposed screening strategy does not affect the convergence of the IRL1 algorithm; rather, it accelerates the solution of the subproblems by reducing the data dimension while preserving the optimality of each subproblem solution. In other words, the screening technique is applied solely within the subproblem solver and is guaranteed to yield the optimal solution for that subproblem. It is also noteworthy that the global convergence result in [31] relies on the assumption that the sequence $\{\boldsymbol{x}^k\}$ is contained within a bounded level set. This assumption is also integral to our analysis.

Assumption 3.1. The level set $L(\boldsymbol{x}^0, \boldsymbol{\epsilon}^0) := \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|^p_{q,p} \leq f(\boldsymbol{x}^0) + \lambda\sum_{i=1}^d(\|\boldsymbol{x}^0_{\mathcal{G}_i}\|_q + \epsilon^0_i)^p\}$ is bounded.

Under this assumption, the convergence of the IRL1 algorithm is maintained even with the integration of the proposed screening strategy. Next, we establish that the proposed screening rule can discard all inactive group features in a finite number of iterations in the following lemma.

Lemma 3.2. *Let $\{\boldsymbol{x}^k\}$ be the sequence generated by Algorithm 3. Then, there exists a constant $C > 0$ such that*

$$\left\|\boldsymbol{A}^T_{\mathcal{G}_i}\left(\sum_{j=1}^d \boldsymbol{A}_{\mathcal{G}_j}\boldsymbol{x}^k_{\mathcal{G}_j} - \boldsymbol{y}\right)\right\|_{q'} < C, \ \forall i \in [d], \forall k \in \mathbb{N}. \tag{3.1}$$

Proof. By Assumption 3.1, Algorithm 3 generates a bounded sequence $\{\boldsymbol{x}^k\}$ that remains within the bounded level set $L(\boldsymbol{x}^0, \boldsymbol{\epsilon}^0)$. The boundedness of $\{\boldsymbol{x}^k\}$ implies that there exists a constant $C > 0$ such that (3.1) holds. This completes the proof. □

An important property of the IRL1 algorithm is that the support of the iterates remains stable after a finite number of iterations.

Lemma 3.3. *Let $\{\boldsymbol{x}^k\}$ be the sequence generated by Algorithm 3, and let the constant C be as defined in Lemma 3.2. If there exists $\tilde{k} \in \mathbb{N}$ such that $\lambda_i^{\tilde{k}} \geq C$, then it follows $\boldsymbol{x}_{\mathcal{G}_i}^k \equiv \mathbf{0}$ for all $k \geq \tilde{k}$.*

Proof. By (2.8), we have $-\boldsymbol{A}_{\mathcal{G}_i}^T \left(\sum_{j=1}^d \boldsymbol{A}_{\mathcal{G}_j} \boldsymbol{x}_{\mathcal{G}_j}^{\tilde{k}} - \boldsymbol{y}\right) = \lambda_i^{\tilde{k}} \boldsymbol{\beta}$, where $\boldsymbol{\beta} \in \partial \|\boldsymbol{x}_{\mathcal{G}_i}^{\tilde{k}}\|_q$. Taking the norm on both sides, we obtain

$$\left\| \boldsymbol{A}_{\mathcal{G}_i}^T \left(\sum_{j=1}^d \boldsymbol{A}_{\mathcal{G}_j} \boldsymbol{x}_{\mathcal{G}_j}^{\tilde{k}} - \boldsymbol{y} \right) \right\|_{q'} = \lambda_i^{\tilde{k}} \|\boldsymbol{\beta}\|_{q'}.$$

If $\boldsymbol{x}_{\mathcal{G}_i}^{\tilde{k}} \neq \mathbf{0}$, then by the properties of the subdifferential, we have $\|\boldsymbol{\beta}\|_{q'} = 1$. Substituting this into the previous equation gives

$$1 = \frac{1}{\lambda_i^{\tilde{k}}} \left\| \boldsymbol{A}_{\mathcal{G}_i}^T \left(\sum_{j=1}^d \boldsymbol{A}_{\mathcal{G}_j} \boldsymbol{x}_{\mathcal{G}_j}^{\tilde{k}} - \boldsymbol{y} \right) \right\|_{q'}.$$

However, this contradicts the assumption that $\lambda_i^{\tilde{k}} \geq C > \left\| \boldsymbol{A}_{\mathcal{G}_i}^T \left(\sum_{j=1}^d \boldsymbol{A}_{\mathcal{G}_j} \boldsymbol{x}_{\mathcal{G}_j}^{\tilde{k}} - \boldsymbol{y}\right) \right\|_{q'}$. Therefore, $\boldsymbol{x}_{\mathcal{G}_i}^{\tilde{k}} = \mathbf{0}$. By induction, it follows that $\boldsymbol{x}_{\mathcal{G}_i}^k \equiv \mathbf{0}$ for any $k \geq \tilde{k}$. This completes the proof. □

For completeness, we include the proof from [30], which establishes that the active and inactive groups remain stable after a sufficiently large number of iterations.

Proposition 3.4. *Let $\{\boldsymbol{x}^k\}$ be the sequence generated by Algorithm 3, and let C be the constant defined in Lemma 3.2. Then, there exist an index set $\mathcal{I}^\star \subset [n]$ and an iteration index $\bar{k} > 0$ such that, $\forall k > \bar{k}$, the inactive set remains fixed, i.e., $\mathcal{I}(\boldsymbol{x}^k) \equiv \mathcal{I}^\star$. Furthermore, for all $i \in [d] \setminus \mathcal{I}^\star$, there exists a positive constant $\varepsilon > 0$ such that $\left\|\boldsymbol{x}_{\mathcal{G}_i}^k\right\|_q \geq \varepsilon > 0$ for sufficiently large k. Consequently, for any cluster point $\boldsymbol{x}^\star$ of $\{\boldsymbol{x}^k\}$, it holds that $\mathcal{I}(\boldsymbol{x}^\star) = \mathcal{I}^\star$.*

Proof. We prove the proposition by contradiction. Suppose, to the contrary, that the inactive set does not stabilize. Then, there exists one index $j \in [d]$ such that the sequence $\left\{ \left\|\boldsymbol{x}_{\mathcal{G}_j}^k\right\|_q \right\}$ takes both zero and nonzero values infinitely often. That is, there exist two disjoint infinite subsequences $\mathcal{S}_1, \mathcal{S}_2 \subset \mathbb{N}$ such that $|\mathcal{S}_1| = \infty$ and $|\mathcal{S}_2| = \infty$, and that

$$\boldsymbol{x}_{\mathcal{G}_j}^k = \mathbf{0}, \forall k \in \mathcal{S}_1 \text{ and } \boldsymbol{x}_{\mathcal{G}_j}^k \neq \mathbf{0}, \forall k \in \mathcal{S}_2.$$

Since $\{\boldsymbol{\epsilon}^k\}$ is monotonically decreased to $\mathbf{0}$, there exists $\tilde{k} > 0$ such that for all $k \geq \tilde{k}$,

$$\lambda_j^k = \lambda p \left(\left\|\boldsymbol{x}_{\mathcal{G}_j}^k\right\|_q + \epsilon_j^k \right)^{p-1} > \lambda p \left(\epsilon_j^k\right)^{p-1} > C.$$

By Lemma 3.3, this implies that $\boldsymbol{x}_{\mathcal{G}_j}^k \equiv \mathbf{0}$ for any $k > \tilde{k}$. In particular, the set $\{\tilde{k}+1, \tilde{k}+2, \ldots\} \subset \mathcal{S}_1$, meaning $|\mathcal{S}_2| < \infty$. However, this contradicts the assumption that $|\mathcal{S}_2| = \infty$.

Now suppose by contradiction that there exists some $j \in [d] \setminus \mathcal{I}^\star$ (i.e., an active group) and a subsequence $\mathcal{S}_3 \subset \mathbb{N}$ such that $\left\|\boldsymbol{x}_{\mathcal{G}_i}^k\right\|_q \to 0, k \in \mathcal{S}_3$. Since $\boldsymbol{\epsilon}^k \to \mathbf{0}$, we obtain for sufficiently large k,

$$\lambda_j^k = \lambda p (\|\boldsymbol{x}_{\mathcal{G}_j}^k\|_q + \epsilon_j^k)^{p-1} > C.$$

By Lemma 3.3, this implies that $\boldsymbol{x}^k_{\mathcal{G}_j} \equiv \mathbf{0}$ for all sufficiently large k, contradicting the assumption that $j \in [d] \setminus \mathcal{I}^\star$. Thus, all active groups remain bounded away from zero after a sufficiently large number of iterations. Since any cluster point $\boldsymbol{x}^\star$ of $\{\boldsymbol{x}^k\}$ satisfies $\mathcal{I}(\boldsymbol{x}^\star) = \mathcal{I}^\star$, the proof is complete. □

Using Proposition 3.4, we show that for sufficiently large k, all components within the inactive set of the optimal solution satisfy the proposed criterion (WR). This implies that they are correctly identified by (WR) within a finite number of iterations. Once this identification occurs, Algorithm 3 reverts to the conventional iteratively reweighted algorithm, operating in a lower-dimensional space restricted to the active groups.

Theorem 3.5. *Let $\{\boldsymbol{x}^k\}$ be the sequence generated by Algorithm 3. Then there exists $\bar{k} \in \mathbb{N}$ such that for any $k \geq \bar{k}$,*

$$\left\|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{y}\right\|_{q'} \leq \lambda^k_i, \quad \forall i \in \mathcal{I}^\star.$$

Consequently, Algorithm 3 reverts to the traditional iteratively reweighted algorithm described in Algorithm 1.

Proof. By Proposition 3.4, the inactive set stabilizes, i.e., $\mathcal{I}(\boldsymbol{x}^k) \equiv \mathcal{I}^\star$ for sufficiently large k. Since $\boldsymbol{x}^k_{\mathcal{G}_i} = \mathbf{0}$ for all $i \in \mathcal{I}^\star$, it follows from $\boldsymbol{\epsilon}^k \to \mathbf{0}$ that

$$\lambda^k_i = \lambda p \left(\epsilon^k_i\right)^{p-1} \to \infty, \quad \forall i \in \mathcal{I}^\star.$$

This ensures that the screening condition (WR) holds, completing the proof. □

3.3 Working Set Strategy Based on Duality Gap

In conjunction with the proposed efficient screening rule, we extend the working set strategy introduced in [18] to address the weighted ℓ_1-norm subproblem. Specifically, we propose a tailored working set strategy for the case $q = 2$. This strategy initializes with a single working variable, selecting only one group. At each iteration, it expands the working set by sequentially incorporating selected groups until an optimal solution is reached. First, recall that

$$\mathcal{L}_2(\boldsymbol{x}) \geq \sum_{i=1}^{d} \|\boldsymbol{x}_{\mathcal{G}_i}\|_{q'} \left(\lambda_i - \left\|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{\theta}\right\|_{q'}\right).$$

If $\boldsymbol{x}^\star_{\mathcal{G}_i} \neq \mathbf{0}$, then the optimal dual variable $\boldsymbol{\theta}^\star$ should satisfy $\|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{\theta}^\star\|_{q'} = \lambda_i$. Thus, from the dual formulation, whenever $\|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{\theta}^\star\|_{q'} < \lambda_i$, it follows that $\boldsymbol{x}^\star_{\mathcal{G}_i} = \mathbf{0}$. This condition aligns precisely with the safe screening rule for the weighted ℓ_1-norm subproblem:

$$\forall i \in [d],\ \|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{\theta}^\star\|_{q'} < \lambda_i \implies \boldsymbol{x}^\star_{\mathcal{G}_i} = \mathbf{0}.$$

Notably, when $q = 2$, the dual norm remains the ℓ_2-norm, effectively transforming the subproblem into a group Lasso problem [34]. Building on the derivations in [19], we establish a safe screening rule specifically for the weighted ℓ_1-norm subproblem (1.6).

To begin, using the triangle inequality and the Cauchy–Schwarz inequality, we obtain for any $\boldsymbol{\theta}$:

$$\|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{\theta}^\star\|_2 = \|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{\theta}+\boldsymbol{A}^T_{\mathcal{G}_i}(\boldsymbol{\theta}^\star-\boldsymbol{\theta})\|_2 \leq \|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{\theta}\|_2+\|\boldsymbol{A}^T_{\mathcal{G}_i}(\boldsymbol{\theta}^\star-\boldsymbol{\theta})\|_2 \leq \|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{\theta}\|_2+\|\boldsymbol{A}_{\mathcal{G}_i}\|_2\|\boldsymbol{\theta}^\star-\boldsymbol{\theta}\|_2.$$

Next, we derive an upper bound for $\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}\|_2$. Since the objective function of the dual problem (2.3) is strongly concave, for any dual variables $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, we have

$$G(\boldsymbol{\theta}_2) \leq G(\boldsymbol{\theta}_1) + \langle \nabla G(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \rangle - \frac{1}{2}\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2^2.$$

Setting $\boldsymbol{\theta}_1 = \boldsymbol{\theta}^\star$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}$, we have

$$G(\boldsymbol{\theta}) \leq G(\boldsymbol{\theta}^\star) + \langle \nabla G(\boldsymbol{\theta}^\star), \boldsymbol{\theta} - \boldsymbol{\theta}^\star \rangle - \frac{1}{2}\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}\|_2^2.$$

Since $\boldsymbol{\theta}^\star$ maximizes $G(\boldsymbol{\theta})$, we know that $\langle \nabla G(\boldsymbol{\theta}^\star), \boldsymbol{\theta} - \boldsymbol{\theta}^\star \rangle \leq 0$. Thus, we obtain

$$\begin{aligned}
& G(\boldsymbol{\theta}) \leq G(\boldsymbol{\theta}^\star) - \frac{1}{2}\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}\|_2^2 \\
\Longleftrightarrow\ & \frac{1}{2}\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}\|_2^2 \leq G(\boldsymbol{\theta}^\star) - G(\boldsymbol{\theta}) \\
\Longleftrightarrow\ & \frac{1}{2}\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}\|_2^2 \leq P(\boldsymbol{x}) - G(\boldsymbol{\theta}) \\
\Longleftrightarrow\ & \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}\|_2 \leq \sqrt{2\mathrm{Gap}(\boldsymbol{x}, \boldsymbol{\theta})},
\end{aligned}$$

where $\mathrm{Gap}(\boldsymbol{x}, \boldsymbol{\theta}) := P(\boldsymbol{x}) - G(\boldsymbol{\theta})$. Thus, for the subproblem of interest, we obtain the gap-based screening rule:

$$\|\boldsymbol{A}_{\mathcal{G}_i}^T \boldsymbol{\theta}\|_2 + \|\boldsymbol{A}_{\mathcal{G}_i}\|_2 \sqrt{2\mathrm{Gap}(\boldsymbol{x}, \boldsymbol{\theta})} < \lambda_i \implies \boldsymbol{x}_{\mathcal{G}_i}^\star = \mathbf{0}.$$

Given that the dual gap is independent of the variable index, we can isolate it on one side of the equation, resulting in the following inequality:

$$\sqrt{2\mathrm{Gap}(\boldsymbol{x}, \boldsymbol{\theta})} < \frac{\lambda_i - \|\boldsymbol{A}_{\mathcal{G}_i}^T \boldsymbol{\theta}\|_2}{\|\boldsymbol{A}_{\mathcal{G}_i}\|_2} \implies \boldsymbol{x}_{\mathcal{G}_i}^\star = \mathbf{0},$$

where the dual variable can be obtained using the dual mapping, as detailed in [19, 18]. Assuming that $\boldsymbol{A}_{\mathcal{G}_i} \neq \mathbf{0}$ for all $i \in [d]$, and following the approach in [18], we define

$$d_{\mathcal{G}_i}(\boldsymbol{\theta}) = \frac{\lambda_i - \|\boldsymbol{A}_{\mathcal{G}_i}^T \boldsymbol{\theta}\|_2}{\|\boldsymbol{A}_{\mathcal{G}_i}\|_2}.$$

We then sort all group variables in ascending order of $d_{\mathcal{G}_i}(\boldsymbol{\theta})$. Groups with larger values of $d_{\mathcal{G}_i}(\boldsymbol{\theta})$ are more likely to correspond to zeros in the optimal solution. Thus, in constructing the working set, we select the group with the smallest $d_{\mathcal{G}_i}(\boldsymbol{\theta})$ at each iteration. Based on this criterion, we propose a suitable working set strategy for solving the weighted ℓ_1-norm subproblem.

It's important to note that the Algorithm 4 is essentially an extension of an existing working set strategy for group-sparsity problems. As a result, we do not delve into a detailed exploration of this working set strategy in the numerical experiments section.

4 Numerical Experiments

In this section, we conduct extensive experiments on both synthetic data and real-world datasets to demonstrate the substantial gains in computational efficiency achieved by the proposed screening rule strategy. All numerical experiments are implemented in Matlab

Algorithm 4 Working Set Strategy Based on Safe Screening Rules

Require: $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{\lambda} \in \mathbb{R}^d_{++}$ and $\boldsymbol{x}^0 \in \mathbb{R}^n$.
1: Initialize $c_0 = \max\{100, |\mathcal{A}(\boldsymbol{x}^0)|\}$ and $\mathcal{W}_0 = \{\mathcal{G}_1, \ldots, \mathcal{G}_{c_0}\}$. Set $t = 1$.
2: **repeat**
3: Compute the dual variable $\boldsymbol{\theta}^t$ based on $\boldsymbol{x}^{t-1}$ and compute $\mathrm{Gap}(\boldsymbol{x}^{t-1}, \boldsymbol{\theta}^t)$.
4: For each group $\mathcal{G}_i$, compute $d^t_{\mathcal{G}_i} = \frac{\lambda_i - \|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{\theta}^t\|_2}{\|\boldsymbol{A}_{\mathcal{G}_i}\|_2}$.
5: Assign $d^t_{\mathcal{G}_i} = -1$ for groups already in the working set $\mathcal{W}_{t-1}$.
6: Update $c_t = \min(2c_{t-1}, d)$.
7: Update $\mathcal{W}_t = \{\mathcal{G}_i : d^t_{\mathcal{G}_i}$ is among the first c_t smallest values$\}$.
8: Solve for $\boldsymbol{x}^t$ based on $\boldsymbol{A}_{[\mathcal{W}_t]}$.
9: **until** $\mathrm{Gap}(\boldsymbol{x}^{t-1}, \boldsymbol{\theta})$ falls below a predefined threshold.
10: **Output:** $\boldsymbol{x}^t$.

R2020b and executed on Macbook Air equipped with an Intel Core i7 processor (1.2 GHz) and 16 GB of RAM.

On the experiment setup, we initialize $\epsilon^0_i = (\frac{\lambda_{\max}}{2p\lambda})^{\frac{1}{p-1}}$ for each $i \in [d]$, where $\lambda_{\max} = \max_{i\in[d]} \|\boldsymbol{A}^T_{\mathcal{G}_i}\boldsymbol{y}\|_{q'}$, and set $\mu = 0.9$ for the IRL1 algorithm. The initial point $\boldsymbol{x}^0$ is obtained by solving the $\ell_{q,1}$ regularization problem with early stopping. To determine the weighting parameter λ, we perform a grid search over $\{\lambda_t\} = \{10^{-(1+\frac{2t}{Q-1})}\lambda_{\max}\}$, where $t \in \{0, 1, \ldots, Q-1\}$. All algorithms are terminated when the relative change in successive iterates satisfies $\left(\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|_2 / \|\boldsymbol{x}^{k+1}\|_2\right) \le 10^{-6}$.

In our comparisons, we consider the following benchmark algorithms: the proximal gradient method for group sparse optimization (PGM-GSO) proposed in [13], the original IRL1 algorithm (ori-IRL1) without screening, and the Safe-Scr-IRL1 method presented in [19], which employs heuristic screening rules by using solutions of lower-dimensional problems to warm-start the optimization. In particular, the proposed IRL1 algorithm enhanced with our screening rule strategy is abbreviated as scr-IRL1.

4.1 Experiments on Synthetic Data

In this subsection, we conduct a sparse signal recovery experiment to demonstrate the effectiveness and efficiency of the proposed screening rule. Following the setup in [13], we generate a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ such that $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$, with each entry drawn from a standard Gaussian distribution. The ground-truth vector $\boldsymbol{x}_{\text{true}}$, which we aim to estimate, has a fixed group size $|\mathcal{G}_i|$ for each $i \in [d]$, and its nonzero entries are also sampled from a standard Gaussian distribution. The observed signal is generated according to the model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_{\text{true}} + \boldsymbol{\zeta}$, where each $\zeta_i \sim \mathcal{N}(0, 10^{-4})$ represents Gaussian noise.

We begin by evaluating the performance of the proposed IRL1 algorithm. In this experiment, we generate the matrix $\boldsymbol{A} \in \mathbb{R}^{256\times 1024}$ and set the group size $|\mathcal{G}_i| = 8$ for each $i \in [d]$. Then, we vary the number of nonzero groups in $\boldsymbol{x}_{\text{true}}$ from 2 to 36 in increments of 2. We fix $q = 2$ and $p = 0.5$ for this test. A signal $\boldsymbol{x}$ is considered successfully recovered if the relative error satisfies $(\|\boldsymbol{x} - \boldsymbol{x}_{\text{true}}\|_2 / \|\boldsymbol{x}_{\text{true}}\|_2) < 5 \times 10^{-3}$. With a fine-tuned weighting parameter λ in $(\mathcal{P})$, we compare the successful recovery rate between scr-IRL1 and PGM-GSO. Here, the successful recovery rate is defined as the number of successful recoveries divided by the total number of runs at different levels of group sparsity in $\boldsymbol{x}_{\text{true}}$. The performance comparison, showing the relationship between sparsity levels in $\boldsymbol{x}_{\text{true}}$ and the successful recovery rate, is presented in Figure 1.

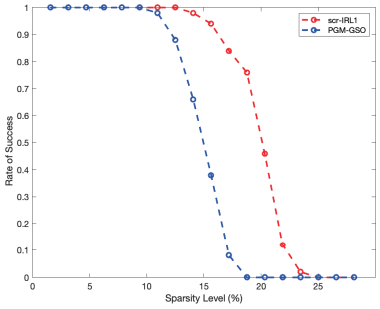
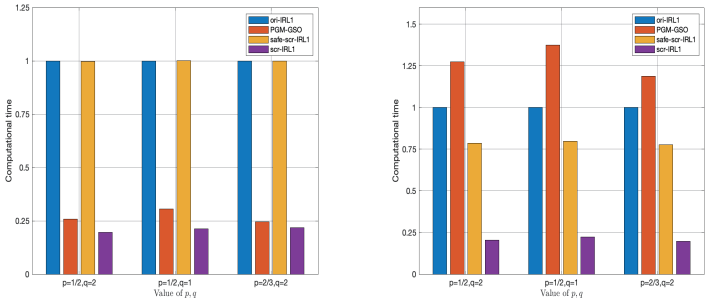


Figure 1: The successful recovery rate across different sparsity levels in $\boldsymbol{x}_{\text{true}}$. The presented results represent the average over 50 random trials.



(a) Low-dimension data (b) High-dimension data

Figure 2: Comparison of computation time across different methods and screening rule strategies. The computation time of ori-IRL1 is normalized to 1 for reference.

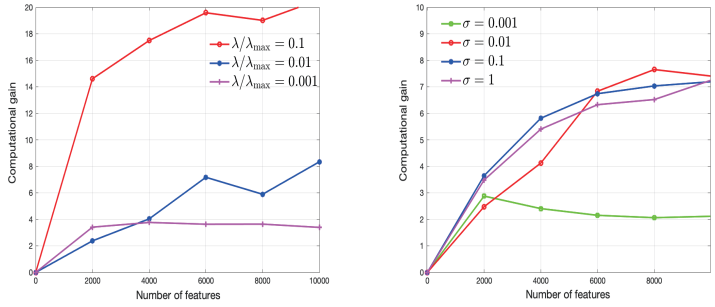
From Figure 1, we can see that IRL1 outperforms PGM-GSO, which demonstrates its competitiveness in solving $(\mathcal{P})$.

We also evaluate the efficiency of the proposed IRL1 algorithm when equipped with the screening strategy. We set $Q = 20$ and fix the number of active groups at $k = 10$ with the group size of $|\mathcal{G}_i| = 5$. We consider two problem sizes: $(m, n) = (500, 2000)$ and $(m, n) = (500, 10000)$. Additionally, we compare different pairs of (p, q), specifically $(p, q) = (\frac{1}{2}, 2)$, $(p, q) = (\frac{1}{2}, 1)$, and $(p, q) = (\frac{2}{3}, 2)$.

From Figure 2, we observe that the proposed scr-IRL1 consistently outperforms all other methods in terms of computation time across all scenarios. Notably, equipping ori-IRL1 with the proposed screening rule reduces computation time by at least a factor of three. Moreover, our screening rule demonstrates significantly higher efficiency compared to the safe method proposed in [19]. Furthermore, the effectiveness of our screening strategy becomes increasingly pronounced in high-dimensional settings.

Next, we investigate the relationship between the computational gain and both the regularization parameter λ and noise level σ. Here, the computational gain is defined as the ratio of the computation time ori-IRL1 to that of scr-IRL1. We consider $(p, q) = (\frac{1}{2}, 2)$ and fix $k = 10$, while varying the number of features n in increments of 2000. The results are presented in Figure 3.

As observed in Figure 3 (a), a larger λ generally leads to greater computational gain. From Figure 3 (b), we observe that $\sigma = 0.01$ results in the highest computational gain



(a) Influence of λ (b) Influence of σ

Figure 3: Influence of λ and σ on computational efficiency. Left: Varying λ with σ fixed at $= 0.01$. Right: Varying σ with λ fixed at $0.01\lambda_{\max}$.

among the tested scenarios.

4.2 Experiments on Real-World Datasets

In this subsection, we conduct experiments on real-world datasets to evaluate the efficiency of scr-IRL1. Specifically, we test datasets from the LIBSVM repository[1], Kaggle[2], and the UCI machine learning repository[3].

Following [25], we evaluate both regression and classification models on these datasets. Specifically, in the regression setting, the observation vector $\boldsymbol{y}$ remains unchanged, unchanged, whereas in the classification setting, $\boldsymbol{y}$ is mapped into $\{0, 1\}$. For further details, refer to [25]. Additionally, we adopt the strategy from [23] to construct a group structure for the data, expanding its dimensions via polynomial feature mapping. Throughout our experiments, we set $Q = 20$ and consider $(p, q) = (\frac{1}{2}, 2)$.

We first examine the computational gains achieved by the proposed screening strategy. For a more comprehensive comparison, this test includes the PGM-GSO algorithm as a benchmark. Next, we verify the correctness of the screening strategy by comparing the predictive performance of both regression and classification models. As shown in Table 1(a), applying the screening rule to IRL1 significantly reduces computation time. Moreover, as demonstrated in Table 1(b), the prediction accuracy remains nearly identical for both IRL1 variants, indicating that the proposed screening rule has a negligible impact on solution quality.

Next, we validate the efficiency of the proposed screening rule, including both the screening procedure and the KKT check procedure, on the **breastcancer** dataset. In this experiment, we fix $(p, q) = (\frac{1}{2}, 2)$ while varying λ. During the first 20 iterations, we record three quantities: the number of screened groups, the number of incorrectly screened groups (as identified by the KKT check), and the number of inactive groups in the solution obtained without using the screening rule. These quantities are reported as ratios in Figure 4. Specifically, the ratio of the number of screened groups to the number of inactive groups (from the solution without screening) is denoted by RSN, and the ratio of the number of wrongly screened groups (detected via the KKT check) to the number of inactive groups is denoted

[1]bodyfat and ionosphere(https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/)
[2]mobileprice(https://www.kaggle.com/iabhishekofficial/mobile-price-classification)
[3]breastcancer(https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic))

Table 1: Performance evaluation on real-world datasets (for each dataset, 70% of the samples are used for training, while the remaining 30% are are used for testing). Mean squared error (MSE) is reported for regression tasks, and classification accuracy is used to assess classification performance.

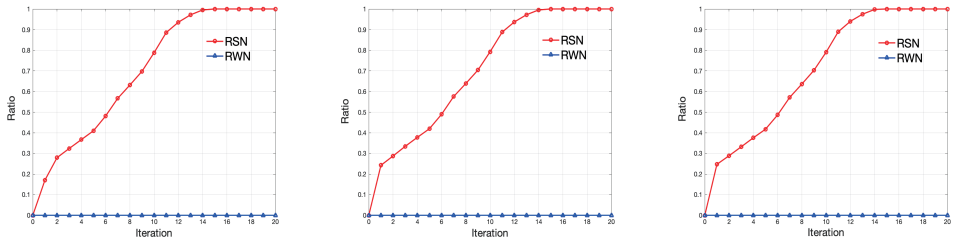
(a) Computation time (s)

Datasets	Dimension	Model	ori-IRL1	scr-IRL1	PGM-GSO
bodyfat	252×455	Regression	56.2361	23.1422	1.7801×10^3
		Classification	101.2709	68.4473	3.3692×10^3
mobileprice	2000×950	Regression	847.4916	263.3385	8.9154×10^3
		Classification	970.0614	535.0475	1.0068×10^4
ionosphere	351×2805	Regression	1.4965×10^3	508.2043	1.5738×10^4
		Classification	971.1735	323.9606	5.6018×10^3
breastcancer	569×2175	Regression	1.6621×10^3	695.7239	1.5424×10^4
		Classification	3.9659×10^3	1.4383×10^3	1.4630×10^4

(b) Prediction performance

Datasets	Model	ori-IRL1	scr-IRL1
bodyfat	Regression	3.2164×10^{-4}	3.2178×10^{-4}
	Classification	100%	100%
mobileprice	Regression	0.1190	0.1190
	Classification	93.17%	93.17%
ionosphere	Regression	0.2947	0.2951
	Classification	98.10%	98.10%
breastcancer	Regression	0.3296	0.3296
	Classification	88.30%	88.30%

by RWN. As shown in Figure 4, the proposed screening strategy identifies all inactive groups within a finite number of iterations while making almost no mistakes.



(a) $\lambda = 0.001\lambda_{\max}$ (b) $\lambda = 0.01\lambda_{\max}$ (c) $\lambda = 0.1\lambda_{\max}$

Figure 4: Evaluation of the efficiency and accuracy of the proposed screening strategy on the **breastcancer** dataset.

5 Conclusion

In this paper, we have proposed a screening rule strategy for a structured optimization problem with a nonconvex $\ell_{q,p}$ regularizer. The proposed screening rule is designed to op-

erate as a prepossessing step before starting the subproblem solver, efficiently identifying inactive group features in the optimal solution. After solving the reduced-dimensional subproblem, a straightforward KKT check verifies the optimality of the solution. Our analysis have demonstrated that inactive group features can be safely identified and removed within a finite number of iterations. The empirical performance of the proposed screening rule is validated through numerical experiments conducted on both synthetic and real-world datasets.

References

[1] F. Bach, R. Jenatton, J. Mairal and G. Obozinski, Optimization with sparsity-inducing penalties, *Found. Trends Mach. Learn.* 4 (2012) 1–106.

[2] A. Bonnefoy, V. Emiya, L. Ralaivola and R. Gribonval, A dynamic screening principle for the Lasso, in: *22nd Eur. Signal Process. Conf. (EUSIPCO), IEEE*, 2014, pp. 6–10.

[3] S.P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[4] E.J. Candès, M.B. Wakin and S.P. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization, *J. Fourier Anal. Appl.* 14 (2008) 877–905.

[5] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.* 96 (2001) 1348–1360.

[6] M. Fazel, H. Hindi and S.P. Boyd, Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices, *in Proc. Amer. Control Conf, IEEE* 3 (2003) 2156–2162.

[7] O. Fercoq, A. Gramfort and J. Salmon, Mind the duality gap: safer rules for the Lasso, in: *Proc. Mach. Learn. Res. PMLR*, 2015, pp. 333–342.

[8] M.A. Figueiredo, J.M. Bioucas-Dias and R.D. Nowak, Majorization–minimization algorithms for wavelet-based image restoration, *IEEE Trans. Image Process.* 16 (2007) 2980–2991.

[9] M.A. Figueiredo and R.D. Nowak, A bound optimization approach to wavelet-based image deconvolution, in: *Proc. IEEE Int. Conf. Image Process., IEEE* 2, 2005, II–782.

[10] G. Gasso, A. Rakotomamonjy and S. Canu, Recovering sparse signals with a certain family of nonconvex penalties and dc programming, *IEEE Trans. Signal Process.* 57 (2009) 4686–4698.

[11] L.E. Ghaoui, V. Viallon and T. Rabbani, Safe feature elimination for the lasso and sparse supervised learning problems, *arXiv preprint* arXiv:1009.4219, (2010).

[12] T. Hastie, R. Tibshirani and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC press, 2015.

[13] Y. Hu, C. Li, K. Meng, J. Qin and X. Yang, Group sparse optimization via $\ell_{p,q}$ regularization, *J. Mach. Learn. Res.* 18 (2017) 960–1011.

[14] T. Johnson and C. Guestrin, Blitz: A principled meta-algorithm for scaling sparse optimization, *Int. Conf. Mach. Learn., PMLR* (2015) 1171–1179.

[15] S. Lee and P. Breheny, Strong rules for nonconvex penalties and their implications for efficient algorithms in high-dimensional regression, *J. Comput. Graph. Statist.* 24 (2015) 1074–1091.

[16] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang and H. Liu, Feature selection: A data perspective, *ACM Comput. Surv. (CSUR)* 50 (2017) 1–45.

[17] Z. Liu, F. Jiang, G. Tian, S. Wang, F. Sato, S.J. Meltzer and M. Tan, Sparse logistic regression with ℓ_p penalty for biomarker identification, *Stat. Appl. Genet. Mol. Biology* 6 (2007).

[18] M. Massias, A. Gramfort and J. Salmon, Celer: a fast solver for the Lasso with dual extrapolation, *Int. Conf. Mach. Learn., PMLR* (2018) 3315–3324.

[19] E. Ndiaye, O. Fercoq and J. Salmon, Gap safe screening rules for sparsity enforcing penalties, *J. Mach. Learn. Res.* 18 (2017) 4671–4703.

[20] A. Rakotomamonjy, R. Flamary, J. Salmon and G. Gasso, Convergent working set algorithm for Lasso with non-convex sparse regularizers, in: *Int. Conf. Artif. Intell. Stat., PMLR*, 2022, pp. 5196–5211.

[21] A. Rakotomamonjy, G. Gasso and J. Salmon, Screening rules for Lasso with non-convex sparse regularizers, in: *Int. Conf. Mach. Learn., PMLR*, 2019, pp. 5341–5350.

[22] R.T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer Science & Business Media, vol. 317, 2009.

[23] V. Roth and B. Fischer, The group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms, in: *Int. Conf. Mach. Learn., PMLR* 2008, pp. 848–855.

[24] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 58 (1996) 267–288.

[25] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor and R.J. Tibshirani, Strong rules for discarding predictors in lasso-type problems, *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 74 (2012) 245–266.

[26] S. Vishwanathan, A.J. Smola and M.N. Murty, SimpleSVM, in: *Int. Conf. Mach. Learn., PMLR*, 2003, pp. 760–767.

[27] H. Wang, Y. Gao, J. Wang and H. Liu, Constrained optimization involving non-convex ℓ_p norms: Optimality conditions, algorithm and convergence, *arXiv preprint* arXiv:2110.14127 (2021).

[28] H. Wang, H. Zeng and J. Wang, An extrapolated iteratively reweighted ℓ_1 method with complexity analysis, *Comput. Optim. Appl.* 83 (2022) 967–997.

[29] H. Wang, H. Zeng and J. Wang, Convergence rate analysis of proximal iteratively reweighted ℓ_1 methods for ℓ_p regularization problems, *Optim. Lett.* 17 (2023) 413–435.

[30] H. Wang, H. Zeng, J. Wang and Q. Wu, Relating ℓ_p regularization and reweighted ℓ_1 regularization, *Optim. Lett.* 15 (2021) 2639–2660.

[31] H. Wang, F. Zhang, Y. Shi and Y. Hu, Nonconvex and nonsmooth sparse optimization via adaptively iterative reweighted methods, *J. Global Optim.* 81 (2021) 717–748.

[32] S.J. Wright, R.D. Nowak and M.A.T. Figueiredo, and Sparse reconstruction by separable approximation, *IEEE Trans. Signal Process.* 57 (2009) 2479–2493.

[33] P. Yin, Y. Lou, Q. He and J. Xin, Minimization of ℓ_{1-2} for compressed sensing, *SIAM J. Sci. Comput.* 37 (2015) A536–A563.

[34] M. Yuan and Y. Lin, Model selection and estimation in regression with grouped variables, *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 68 (2006) 49–67.

[35] C. Zhang, Nearly unbiased variable selection under minimax concave penalty, *Ann. Stat.* 38 (2010) 894–942.

Appendix

Calculation of Subgradients for $\|\boldsymbol{x}_{\mathcal{G}_i}\|_{q,p}$

The extended chain rule for subgradients, as presented in [22, Theorem 10.49], is adapted and applied below.

Theorem 5.1. *Let $f(\boldsymbol{x}) = g(F(\boldsymbol{x}))$ for a proper, lower semicontinuous function $g : \mathbb{R}^m \to \bar{\mathbb{R}}$ and a strictly continuous vector-valued function $F : \mathbb{R}^n \to \mathbb{R}^m$. Let $\bar{\boldsymbol{x}}$ be a point where f is finite. Then*

$$\hat{\partial} f(\bar{\boldsymbol{x}}) \supset \hat{D}^* F(\bar{\boldsymbol{x}}) \left[\hat{\partial} g(F(\bar{\boldsymbol{x}}))\right] = \bigcup \left\{\hat{\partial}(\boldsymbol{y}F)(\bar{\boldsymbol{x}}) \mid \boldsymbol{y} \in \hat{\partial} g(F(\bar{\boldsymbol{x}}))\right\}.$$

If the only vector $\boldsymbol{y} \in \partial^\infty g(F(\bar{\boldsymbol{x}}))$ with $\boldsymbol{0} \in \partial(\boldsymbol{y}F)(\bar{\boldsymbol{x}})$ is $\boldsymbol{y} = \boldsymbol{0}$, one also has

$$\partial f(\bar{\boldsymbol{x}}) \subset D^* F(\bar{\boldsymbol{x}})[\partial g(F(\bar{\boldsymbol{x}}))] = \bigcup \{\partial(yF)(\bar{\boldsymbol{x}}) \mid y \in \partial g(F(\bar{\boldsymbol{x}}))\}.$$

Recall that $\|\boldsymbol{x}\|_{q,p} = \sum_{i=1}^m \|x_i\|_q^p$, where $x_i \in \mathbb{R}^{n_i}, n = \sum_{i=1}^m n_i, q \geq 1$, and $p \in (0,1)$. Using Theorem 5.1, we can compute the subgradients of $\|\boldsymbol{x}\|_{q,p}$.

Consider the function $f(\boldsymbol{x}) = f_1(x_1) + \cdots + f_m(x_m)$, where $f_i : \mathbb{R}^{n_i} \to \bar{\mathbb{R}}$ are lower semicontinuous functions. From [22, Proposition 10.5], for any $\bar{\boldsymbol{x}} = [\bar{x}_1, \ldots, \bar{x}_m]^T$ with $f(\bar{\boldsymbol{x}})$ finite and $df_i(\bar{x}_i)(0) = 0$, one has

$$\partial f(\bar{\boldsymbol{x}}) = \ \partial f_i(\bar{x}_1) \times \ldots \times \partial f_m(\bar{x}_m),$$

and

$$\hat{\partial} f(\bar{\boldsymbol{x}}) = \ \hat{\partial} f_i(\bar{x}_1) \times \ldots \times \hat{\partial} f_m(\bar{x}_m).$$

Therefore, we only need to compute the subgradient of $\|x_i\|_q^p$. From (1.2), we have

$$\hat{\partial}\|x_i\|_q = \partial\|x_i\|_q = \{\boldsymbol{u} \mid \langle \boldsymbol{u}, x_i\rangle = \|x_i\|_q, \|\boldsymbol{u}\|_* \leq 1\},$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|_q$. By [27, Theorem 2.1], we have

$$\partial(\cdot)^p(\|x_i\|_q) = \hat{\partial}(\cdot)^p(\|x_i\|_q) = \begin{cases} \mathbb{R} & \text{if } \|x_i\|_q = 0, \\ p\|x_i\|_q^{p-1} & \text{otherwise.} \end{cases}$$

First, consider the case where $\|x_i\|_q = 0$. For any $y \in \mathbb{R}$, we have $\hat{\partial}(y\|\cdot\|_q)(x_i) = \{y\boldsymbol{u} \mid \|\boldsymbol{u}\|_* \leq 1\}$, which implies

$$\hat{\partial}\|x_i\|_q^p \supset \left\{\hat{\partial}(y\|\cdot\|_q)(x_i) \mid y \in \mathbb{R}\right\} = \mathbb{R}^{n_i}.$$

Hence, in this case, $\mathbb{R}^{n_i} \subset \hat{\partial}\|x_i\|_q^p \subset \partial\|x_i\|_q^p \subset \mathbb{R}^{n_i}$, which indicates $\hat{\partial}\|x_i\|_q^p = \partial\|x_i\|_q^p = \mathbb{R}^{n_i}$.

Next, consider the case where $\|x_i\|_q \neq 0$. Let $y = p\|x_i\|_q^{p-1}$, we obtain

$$\hat{\partial}(y\|\cdot\|_q)(x_i) = \{y\boldsymbol{u} \mid \langle \boldsymbol{u}, x_i\rangle = \|x_i\|_q, \|\boldsymbol{u}\|_* \leq 1\} = y\partial\|x_i\|_q.$$

This implies

$$\hat{\partial}\|x_i\|_q^p \supset \left\{\hat{\partial}(y\|\cdot\|_q)(x_i) \mid y \in \hat{\partial}(\cdot)^p(x_i)\right\} = y\partial\|x_i\|_q.$$

On the other hand, since $\|x_i\|_q \neq 0$, it is clear that $\mathbf{0} \notin \partial\|x_i\|_q$. The only value of y satisfying $y \in \partial^\infty(\cdot)^p(\|x_i\|_q)$ with $\mathbf{0} \in \partial(y\|\cdot\|_q)(x_i)$ is $y = 0$. Therefore, by Theorem 5.1,

$$\partial\|x_i\|_q^p \subset \left\{\partial(y\|\cdot\|_q)(x_i) \mid y = p\|x_i\|_q^{p-1}\right\} = y\partial\|x_i\|_q.$$

Overall, we have established the chain of inclusions

$$y\partial\|x_i\|_q \subset \hat{\partial}(y\|\cdot\|_q)(x_i) \subset \partial(y\|\cdot\|_q)(x_i) = y\partial\|x_i\|_q,$$

which confirms that

$$\hat{\partial}(y\|\cdot\|_q)(x_i) = \partial(y\|\cdot\|_q)(x_i) = y\partial\|x_i\|_q.$$

Using [22, Proposition 10.5], the subgradients of $\|x_i\|_{q,p}$ are given by

$$\partial\|x_i\|_{q,p} = \hat{\partial}\|x_i\|_{q,p} = C_1 \times \ldots \times C_m,$$

where

$$C_i = \begin{cases} p\|x_i\|_q^{p-1}\partial\|x_i\|_q & \text{if } x_i \neq 0, \\ \mathbb{R}^{n_i} & \text{otherwise.} \end{cases}$$

Manuscript received 4 March 2023
revised 23 November 2023
accepted for publication 2 December 2023

Tiange Li
School of Information Science and Technology
ShanghaiTech University
Shanghai, 201210, China
E-mail address: litg@shanghaitech.edu.cn

Xiangyu Yang
School of Mathematics and Statistics, Henan University
Kaifeng, 475000, China; and
Center for Applied Mathematics of Henan Province
Henan University, Zhengzhou, 450046, China
E-mail address: yangxy@henu.edu.cn

Hao Wang
School of Information Science and Technology
ShanghaiTech University, Shanghai, 201210, China
E-mail address: haw309@gmail.com