Journal of Global Optimization (2021) 81:717–748
https://doi.org/10.1007/s10898-021-01093-0

Nonconvex and Nonsmooth Sparse Optimization via Adaptively Iterative Reweighted Methods

Hao Wang[1] · **Fan Zhang**[1,3,4] · **Yuanming Shi**[1] · **Yaohua Hu**[2]

Received: 31 August 2019 / Accepted: 5 September 2021 / Published online: 5 October 2021

Abstract
We propose a general formulation of nonconvex and nonsmooth sparse optimization problems with convex set constraint, which can take into account most existing types of nonconvex sparsity-inducing terms, bringing strong applicability to a wide range of applications. We design a general algorithmic framework of iteratively reweighted algorithms for solving the proposed nonconvex and nonsmooth sparse optimization problems, which solves a sequence of weighted convex regularization problems with adaptively updated weights. First-order optimality condition is derived and global convergence results are provided under loose assumptions, making our theoretical results a practical tool for analyzing a family of various reweighted algorithms. The effectiveness and efficiency of our proposed formulation and the algorithms are demonstrated in numerical experiments on various sparse optimization problems.

Keywords Nonconvex and nonsmooth sparse optimization · Iteratively reweighted methods

1 Introduction

Nonconvex and nonsmooth sparse optimization problems have been becoming a prevalent research topic in many disci plines of applied mathematics and engineering. Indeed, there has been a tremendous increase in a number of application areas in which nonconvex sparsity-inducing techniques have been employed, such as machine learning [5,42], telecommunications [36,38], image reconstruction [23], sparse recovery [11,30] and signal processing [8,31]. This is mainly because of their superior ability to reduce the complexity

✉ Hao Wang
wanghao1@shanghaitech.edu.cn

1 School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, People's Republic of China

2 Shenzhen Key Laboratory of Advanced Machine Learning and Applications, College of Mathematics and Statistics, Shenzhen University, Shenzhen 518061, People's Republic of China

3 University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

4 Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, People's Republic of China

of a system, improve the generalization of the prediction performance, and/or enhance the robustness of the solution, compared with traditional convex sparsity-inducing techniques.

Despite their wide application, nonconvex and nonsmooth sparse optimization problems are computationally challenging to solve due to the nonconvex and nonsmooth nature of the sparsity-inducing terms. A popular method for handling the convex/nonconvex regularization problems is the iteratively reweighted algorithm, which approximates the nonconvex and nonsmooth problem by a sequence of trackable convex subproblems. There have been some iteratively reweighted algorithms proposed for special cases of the nonconvex and nonsmooth problems. For example, in [9,24,44] Yin et al. have designed an iteratively reweighted algorithm for solving the unconstrained nonconvex ℓ_p norm model and in [28] Lu has analyzed the global convergence of a class of reweighted algorithms for the unconstrained nonconvex ℓ_p regularized problem. The constrained ℓ_p-regularization problem is studied to improve the image restoration using a priori information and the optimality condition of this problem is given in [41]. The critical technique of this type of algorithms is to add relaxation parameters to transform the nonconvex and nonsmooth sparsity-inducing terms into smooth approximate functions and then use linearization to obtain convex subproblems [10,22]. It should be noticed that the relaxation parameter should be driven to 0 in order to obtain the solution of the original unrelaxed problem. Two most popular variants of iteratively reweighted type of algorithms are the iteratively reweighted ℓ_1 minimization and the iteratively reweighted squared ℓ_2-norm minimization. The former has convex but nonsmooth subproblem, while the latter leads to convex and smooth subproblems. It has been reported by E. Candes et al. in [8] that the reweighted ℓ_1 minimization can significantly enhance sparsity of the solution. It has been demonstrated that Iteratively reweighted least-squares have greatly promoted the computation and correctness of robust regression estimation [17,39].

However, iteratively reweighted algorithms are generally difficult to track and analyze. This is mainly because most nonconvex functions are non-Lipschitz continuous, especially around sparse solutions that we are particularly interested in. The major issue caused by this situation is that the optimal solution cannot be characterized by common optimality conditions. For some special cases where the sparsity-inducing term is ℓ_p-norm and no constraint is involved, the first-order and second-order sufficient optimality conditions have been studied in [32,41]. Chen et al have derived a first-order necessary optimality condition for local minimizers and define the generalized stationary point of the constrained optimization problems with nonconvex regularization [4]. These results are used by Lu to derive the global convergence of a class of iteratively reweighted ℓ_1 and ℓ_2 methods for unconstrained ℓ_p regularization problems [28]. For the sum of a convex function and a (nonconvex) nondecreasing function applied to another convex function, the convergence to a critical point of the iteratively reweighted ℓ_1 algorithm is provided when the objective function must be coercive [32]. As far as we know, the convergence results of iteratively reweighted ℓ_1 and ℓ_2 methods for Non-Lipschitz nonconvex and nonsmooth problems with general convex-set constraint have not been provided.

As for more general cases, the analysis in current work has many limitations due to this obstacle for theoretical analysis. First, instead of driving the relaxation parameter to 0, many existing methods [6,9] aim to show the convergence to the optimal solution of the relaxed sparse optimization problems. An iteratively reweighted least squares algorithm for the relaxed problem ℓ_p problem in sparse signal recovery has been discussed in [1] with local convergence rate analysis. A critical aspect of any implementation of such an approach is the selection of the relaxation parameters which prevents the weights from becoming overwhelmingly large. As been explained in [9], large relaxation parameters will smooth out many local minimizers, whereas small values can cause the subproblems difficult to solve and the algorithm too quickly get trapped into local minimizers. For this purpose, updating strategies of the relaxation parameter have been studied in [6], but it is only designed for constrained convex problems.

Second, some methods assume Lipschitz continuity of the objective in their analysis, which only holds true for few nonconvex sparsity-inducing terms such as log-sum regularization [28,32]. Though this assumption is not explicitly required by some other researchers, they need another assumption that the negative of the sparsity-inducing term—which is convex—is subdifferentiable everywhere [32]. However, it must be noticed that this is a quite strong assumption and generally not suitable for most sparsity-inducing terms e.g. ℓ_p-norm, consequently limiting their applicability to various cases.

This situation may become even worse when a general convex set constraint is added to the problem. To the best of our knowledge, only simple cases such as linearly constrained cases have been studies by current work. To circumvent the obstacle for analysis, current methods either focus on the relaxed problem as explained above, or unconstrained reformulations where the constraint violation is penalized in the objective [9,28,41]. The latter approach then arises the issue of how to select the proper regularization parameter value.

Moreover, some work [27] assumes the coercivity of the objective to guarantee that the iterates generated by the algorithms must have clustering points. This sets another limitation as many sparsity-inducing terms is bounded above, e.g., arctan function. This assumption therefore requires the rest part of the objective must be coercive, which is generally not the case.

Overall, for cases involving more general sparsity-inducing terms and convex-set constraints, the analysis of the behavior of iteratively reweighted algorithms remains an open question.

Despite of the iteratively reweighted algorithms, the Difference of Convex (DC) algorithm is also used for tackling some specific-form sparse optimization problems. As Pang et al. mentioned in [47,48], a great deal of existing nonconvex regularizations can be represented as DC functions, and then can be solved by the DC algorithm. Unfortunately, these works need the assumption that the nonconvex regularization term is Lipschitz differentiable, which limits the applicability to the cases like ℓ_p-norm regularization. To address this issue, Liu et al. have proposed a Successive Difference of Convex Approximation Method (SDCAM) in [45], which makes use of the Moreau envelope as a smoothing technique for the nonsmooth terms. However, they need the proximal mapping of each nonsmooth function is easy to compute. This requirement obstacles the use of lots of nonconvex regularizations, of which the proximal operator can not be solved effectively. For example, when we use the ℓ_p norm $(0 < p < 1)$ as the regularization term, only the cases $p = 1/2$ and $p = 2/3$ have an analytical solution of the associated proximal mapping as shown in [46]. This fact limits the SDCAM to select other values of p. And as shown in experimental results, the ℓ_p regularization with different values of p perform differently. Usually, smaller p induces more sparse solution.

In this paper, we consider a unified formulation of the convex set constrained nonconvex and nonsmooth sparse optimization problems. A general algorithmic framework of Adaptively Iterative Reweighted (AIR) algorithm is presented for solving these problems. We derive the first-order condition to characterize the optimal solutions and analyze global convergence of the proposed method to the first-order optimal solutions. The most related research work mainly includes the iteratively reweighted algorithms proposed by [6] for solving general constrained convex problems, the reweighted methods by [28,29] for solving unconstrained ℓ_p regularization problems, and the algorithmic framework proposed in [32] for solving the unconstrained nonsmooth and nonconvex optimization problems. However, we emphasize again that our focus is on dealing with cases with general nonconvex and nonsmooth sparsity-inducing terms and general convex set constraints—a stark contrast to the situations considered by most existing methods.

The contributions of this paper can be summarized as follows.

- Our unified problem formulation can take into account most existing types of nonconvex sparsity-inducing functions which also allows for group structure as well as general convex set constraint. A general algorithmic framework of adaptively iteratively reweighted algorithms is developed by solving a sequence of trackable convex subproblems. A unified first-order necessary conditions is derived to characterize the optimal solutions by using Fréchet subdifferentials, and global convergence of the proposed algorithm is provided.
- For reweighted ℓ_1 and ℓ_2 minimizations, our algorithm allows for vanishing relaxation parameters, which can avoid the issue of selecting appropriate value of relaxation parameter. The global convergence analysis for both iteratively reweighted ℓ_1 and ℓ_2 algorithms are provided. We show that every limit point generated by the algorithms must satisfy the first-order necessary optimality condition for the original unrelaxed problem, instead of the relaxed problem—a novel result that most current work does not possess. Candes et al. have put forward several open questions in [8] about the reweighted ℓ_1 algorithm, including under what conditions the reweighted ℓ_1 algorithm will converge, when the iterates have limit point, and the ways of updating ϵ as the algorithm progresses towards a solution. Our work provides answers to these questions.
- We have proven that the existence of the cluster point to the algorithm can be guaranteed by understanding conditions under which the iterates generated is bounded. The conditions guaranteeing the boundedness is also given. Conditions for selecting the starting point and the initial relaxation parameters are also provided to guarantee the global convergence. This makes our methods also apply for cases where the objective is not coercive.

1.1 Organization

In the remainder of this section, we outline our notation and introduce various concepts that will be employed throughout the paper. In Sect. 2, we describe our problem of interests and explain its connection to various existing types of sparsity-inducing techniques. In Sect. 3, we describe the details of our proposed AIR method and apply it to different types of nonconvex sparsity-inducing terms. The optimality condition and the global convergence of the proposed algorithm in different situations are provided in Sect. 4. We discuss implementations of our methods and the results of numerical experiments in in Sect. 5. Concluding remarks are provided in Sect. 6.

1.2 Notation and preliminaries

Much of the notation that we use is standard, and when it is not, a definition is provided. For convenience, we review some of this notation and preliminaries here.

We use $\mathbf{0}$ to represent the vector filled with all 0s of appropriate dimension. Let $\mathbb{R}^n$ be the space of real n-vectors, $\mathbb{R}^n_+$ be the nonnegative orthant of $\mathbb{R}^n$, $\mathbb{R}^n_+ = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq \mathbf{0}\}$ and the nonpositive orthant $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \leq \mathbf{0}\}$. Moreover, let $\mathbb{R}^n_{++}$ be its interior $\mathbb{R}^n_{++} := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} > \mathbf{0}\}$. The set of $m \times n$ real matrices is denoted by $\mathbb{R}^{m\times n}$. For a pair of vectors $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^n \times \mathbb{R}^n$, their inner product is written as $\langle \mathbf{u}, \mathbf{v}\rangle$. The set of nonnegative integers is denoted by $\mathbb{N}$. Suppose $\mathbb{R}^n$ be the product space of subspaces $\mathbb{R}^{n_i}, i = 1, \ldots, m$ with $\sum_{i=1}^m n_i = n$, i.e., it takes decomposition $\mathbb{R}^n = \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_m}$. Given a closed convex set $X \subset \mathbb{R}^n$, the normal cone to X at a point $\bar{\mathbf{x}} \in X$ is given by

$$N(\bar{\mathbf{x}}|X) := \{\mathbf{z}|\langle \mathbf{z}, \mathbf{x} - \bar{\mathbf{x}}\rangle \le 0,\ \forall \mathbf{x} \in X\}.$$

The characteristic function of X is defined as

$$\delta(\mathbf{x}|X) = \begin{cases} 0 & \text{if } \mathbf{x} \in X, \\ +\infty & \text{otherwise.} \end{cases}$$

The indicator operator $\mathbb{I}(\cdot)$ is an indicator function that takes a value of 1 if the statement is true and 0 otherwise.

For a given $\alpha \in \mathbb{R}$, denote the level set of f as

$$L(\alpha; f) := \{\mathbf{x} \in \mathbb{R}^n | f(\mathbf{x}) \le \alpha\}.$$

In particular, we are interested in level set with an upper bound reachable for f:

$$L(f(\hat{\mathbf{x}}); f) := \{\mathbf{x} \in \mathbb{R}^n | f(\mathbf{x}) \le f(\hat{\mathbf{x}})\}.$$

The subdifferential of a convex function f at $\mathbf{x}$ is a set defined by

$$\partial f(\mathbf{x}) = \{\mathbf{z} \in \mathbb{R}^n | f(\mathbf{y}) - f(\mathbf{x}) \ge \langle \mathbf{z}, \mathbf{y} - \mathbf{x}\rangle, \forall \mathbf{y} \in \mathbb{R}^n\}.$$

Every element $\mathbf{z} \in \partial f(\mathbf{x})$ is referred to as a subgradient. To characterize the optimality conditions for nonsmooth problems, we need to introduce the concepts of Fréchet subdifferentiation. In fact, there are a variety of subdifferentials known by now including limiting subdifferentials, approximate subdifferentials and Clarke's generalized gradient, many of which can be used here for deriving the optimality conditions. The major tool we choose in this paper is the Fréchet subdifferentials, which were introduced in [3,20] and discussed in [21].

Definition 1 (Fréchet subdifferential) Let f be a function from a real Banach space into an extended real line $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$, finite at $\mathbf{x}$. The Fréchet subdifferential of f at $\mathbf{x}$, denoted as $\partial_F f(\mathbf{x})$, is the set

$$\partial_F f(\mathbf{x}) = \left\{\mathbf{x}^* \in \mathbb{R}^n : \liminf_{\mathbf{u} \to \mathbf{x}} \frac{f(\mathbf{u}) - f(\mathbf{x}) - \langle \mathbf{x}^*, \mathbf{u} - \mathbf{x}\rangle}{\|\mathbf{u} - \mathbf{x}\|} \ge 0\right\}.$$

Its elements are referred to as Fréchet subgradients.

For a composite function $r \circ c(\mathbf{x})$, where $c : \mathbb{R}^n \to \mathbb{R}$ and $r : \mathbb{R} \to \mathbb{R}$, denote $\partial_F r(c(\mathbf{x}))$ (or simply $\partial_F r(c)$) as the Fréchet subdifferential of r with respect to c, and $r'(c(\mathbf{x}))$ (or simply $r'(c)$) as the derivative of r with respect to $c(\mathbf{x})$ if r is differentiable at $c(\mathbf{x})$.

2 Problem statement and its applications

In this section, we propose a unified formulation of the constrained nonconvex and nonsmooth sparse optimization problem, and list the instances in some prominent applications.

2.1 Problem statements

We consider the following nonconvex and nonsmooth sparse optimization problem

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) + \Phi(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in X, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is smooth and convex and $X \subset \mathbb{R}^n$ is a closed convex set. This type of problem is a staple for many applications in signal processing [7,25], wireless communications [34,37] and machine learning [2,16]. For example, in signal processing, f may be the mean-squared error for signal recovery, X may be a nonnegative constraint for signal [19]; in wireless communications, f may represent the system performance such as transmit power consumption, X may be the transmit power constraints and quality of service constraints [38]; in machine learning, f can represent the convex loss function, such as the cross-entropy loss for logistic regression [15].

In a large amount of applications, the being recovered vector $\mathbf{x}$ is expected to have some sparse property in a structured manner. To handle this type of structured sparsity, various types of group-based Φ has been studied in [40]. Consider a collection of groups $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_m\}$ with $|\mathcal{G}_i| = n_i$. The union over all groups covers the full index set and $\sum_{i=1}^m n_i = n$. The structured vector $\mathbf{x}$ can be written as

$$\mathbf{x} = [\underbrace{x_1, x_2, \cdots, x_{n_1}}_{\mathbf{x}_{\mathcal{G}_1}^T}, \cdots, \underbrace{x_{n-n_m+1}, \cdots, x_n}_{\mathbf{x}_{\mathcal{G}_m}^T}]^T.$$

With these ingredients, the associated group-based function Φ takes the form

$$\Phi(\mathbf{x}) = \sum_{i=1}^m r_i(c_i(\mathbf{x}_{\mathcal{G}_i})),$$

where $c_i : \mathbb{R}^{n_i} \to \mathbb{R}$ is convex and $r_i : \mathbb{R} \to \mathbb{R}$ is concave for each i. Throughout this paper, we make the following assumptions about f, r_i, c_i and X.

Assumption 1 The functions $f, r_i, c_i, i = 1, \ldots, m$, and set X are such that

(i) X is closed and convex.
(ii) f is smooth, convex and bounded below by $\underline{f}$ on X.
(iii) r_i is smooth on $\mathbb{R} \setminus \{0\}$, concave and strictly increasing on $\mathbb{R}_+$ with $r_i(-c) = r_i(c)$ and $r_i(0) = 0$, and is Fréchet subdifferentiable at 0.
(iv) c_i is convex and coercive with $c_i(\mathbf{x}_i) \geq 0$, $\forall \mathbf{x} \in X$ where the equality holds if and only if $\mathbf{x}_i = \mathbf{0}$.
(v) The composite function $\phi_i = r_i \circ c_i(x) = r_i(c_i(\mathbf{x}_i))$ is concave on regions $\{x \mid c_i(\mathbf{x}_i) \geq 0\}$ and $\{x \mid c_i(\mathbf{x}_i) \leq 0\}$.

Remark 1 The symmetry of r_i is not a requirement, since $c_i(\mathbf{x}) \geq 0$ is assumed always true; the purpose of this assumption is to simplify the analysis.

Most existing sparse optimization problems can be reverted to (1). In next subsection, we describe the important applications of problem (1) and explain the specific forms of the functions f, r_i, c_i in the example. Based on different formulations of the composite function $\Phi(\mathbf{x})$, there are a great deal of nonconex sparsity-inducing techniques to promote sparse solutions, such as the approximations of the ℓ_0 norm of $\mathbf{x}$.

The problems considered here allow the sparse-inducing terms to be non-Lipschitz, and it also allows the presence of a general convex set constraint. If some components of the stationary point tend to zero, it is possible that both the Fréchet subdifferentials of the non-Lipschitz term and the normal direction of the constraint tend to infinity. This leads to the difficulty of analyzing the convergence of their linear combination (i.e., the optimality condition residual of the problem).

2.2 Sparsity-inducing Functions

Many applications including signal processing, wireless communications and machine learning involve the minimization of the ℓ_0-norm of the variables $\|\mathbf{x}\|_0$, i.e., the number of nonzero components in $\mathbf{x}$. However, this is regarded as an NP-hard problem, thus various approximations of ℓ_0 norm have been proposed. By different choice of the formulation r_i and c_i, there exist many approximations to ℓ_0 norm, so that a smooth approximate problem of (1) is derived with

$$\|\mathbf{x}\|_0 \approx \Phi(\mathbf{x}) = \sum_{i=1}^{n} r_i(c_i(x_i)).$$

In the following discussion, we only provide the expression of r_i on $\mathbb{R}_+$, since by Assumption 1, r_i can be defined accordingly on $\mathbb{R}_-$.

2.2.1 The EXP approximation

The first instance is the feature selection algorithm via concave minimization proposed by Bradley and Mangasarian [5] with approximation

$$\|\mathbf{x}\|_0 \approx \sum_{i=1}^{n} 1 - e^{-p|x_i|} \;\text{ with }\; p > 0, \tag{EXP}$$

where p is chosen to be sufficiently large to promote sparse solutions. The concavity of this function leads to a finitely terminating algorithm and a more accurate representation of the feature selection algorithm. It is reported that the algorithms with this formulation obtained a reduction in error with selected features fewer in number and they are faster compared to traditional convex feature selection algorithms. For example, we can choose

$$c_i(x_i) = |x_i|,\; r_i = 1 - e^{-pc_i} \text{ or } c_i = x_i^2,\; r_i = 1 - e^{-p\sqrt{c_i}},$$

so that this approximation can be viewed as a specific formulation of Φ.

2.2.2 The LPN approximation

The second instance, which is widely used in many applications currently, is to approximate the ℓ_0 norm by ℓ_p quasi-norm [13]

$$\|\mathbf{x}\|_0 \approx \sum_{i=1}^{n} |x_i|^p \;\text{ with } p \in (0, 1) \tag{LPN}$$

and p is chosen close to 0 to enforce sparsity in the solutions. Based on this approximation, numerous applications and algorithms have emerged. Here we can choose

$$c_i(x_i) = |x_i|, r_i(c_i) = c_i^p \text{ or } c_i(x_i) = x_i^2, r_i(c_i) = c_i^{p/2}$$

in the formulation of Φ.

2.2.3 The LOG approximation

Another option for approximating ℓ_0 norm, proposed in [26], is to use the log-sum approximation

$$\|\mathbf{x}\|_0 \approx \sum_{i=1}^{n} \log(1 + p|x_i|) \text{ with } p > 0, \tag{LOG}$$

and setting p sufficiently large leads to sparse solutions. We can choose

$$c_i(x_i) = |x_i|, r_i(c_i) = \log(1 + pc_i),$$

or

$$c_i(x_i) = x_i^2, r_i(c_i) = \log(1 + p\sqrt{c_i}).$$

2.2.4 The FRA approximation

The approximation technique proposed in [13] suggests

$$\|\mathbf{x}\|_0 \approx \sum_{i=1}^{n} \frac{|x_i|}{|x_i| + p}, \text{ with } p > 0, \tag{FRA}$$

and p is required to be sufficiently small to promote sparsity. One can use

$$c_i(x_i) = |x_i|, \ r_i(c_i) = \frac{c_i}{c_i + p},$$

or

$$c_i x_i = x_i^2, \ r_i(c_i) = \frac{\sqrt{c_i}}{\sqrt{c_i} + p}.$$

2.2.5 The TAN approximation

Candès et al. propose an approximation to the ℓ_0 norm in [8]

$$\|\mathbf{x}\|_0 \approx \sum_{i=1}^{n} \arctan(p|x_i|), \text{ with } p > 0, \tag{TAN}$$

and sufficiently small p can cause sparsity in the solution. The function arctan is bounded above and ℓ_0-like. It is reported that this approximation tends to work well and often better than the log-sum (LOG). In this case, we can choose

$$c_i(x_i) = |x_i|, r_i(c_i) = \arctan(pc_i),$$

or

$$c_i(x_i) = x_i^2, r_i(c_i) = \arctan(p\sqrt{c_i}).$$

2.2.6 The SCAD and MCP approximation

Another nonconvex sparsity-inducing technique needs to be mentioned is the SCAD regularization proposed in [12], which require the derivative of ϕ_i to satisfy

$$c_i(x_i) = |x_i|, \phi_i'(c_i) = \lambda\{\mathbb{I}(c_i \le \lambda) + \frac{(a\lambda - c_i)_+}{(a-1)\lambda}\mathbb{I}(c_i > \lambda)\}, \tag{SCAD}$$

for some $a > 2$, where often $a = 3.7$ is used. Alternatively, the MCP [43] regularization uses

$$c_i(x_i) = |x_i|, \phi_i'(c_i) = (a\lambda - c_i)_+/a \text{ for some } a \ge 1. \tag{MCP}$$

2.2.7 The group structure

These sparsity-inducing functions can also take into account group structures. For example, $\ell_{p,q}$-norm with $p \ge 1$ and $0 < q < 1$ [18] is defined as

$$\|\mathbf{x}\|_{p,q} = \left(\sum_{i=1}^{m} \|\mathbf{x}_{\mathcal{G}_i}\|_p^q\right)^{1/q}.$$

Therefore, we can choose

$$\Phi(\mathbf{x}) = \|\mathbf{x}\|_{p,q}^q, \quad \text{with } c_i(\mathbf{x}_{\mathcal{G}_i}) = \|\mathbf{x}_{\mathcal{G}_i}\|_p \text{ and } r_i(c_i) = c_i^q.$$

2.3 Problem analysis

There have been various literatures for solving the nonconvex and nonsmooth sparse optimization problems. In [9,24] Wotao Yin et al. have considered solve the sparse signal recovery problem by using the unconstrained nonconvex ℓ_p norm model, proposed the associated iterative reweighted unconstrained ℓ_p algorithm and provided the convergence analysis for the reweighted ℓ_2 case. In [28] Zhaosong Lu have provied the first-order optimality condition for the unconstrained nonconvex ℓ_p norm problem, and convergence analysis for both ℓ_1 and ℓ_2 types reweighted algorithm. However, it is not clear for analyzing the first-order optimalizty condition for the constrained nonconvex and nonsmooth sparse optimization problem (1). In order to address this issue, we propose the AIR algorithm in Sect. 3, provide the first-order optimality condition for (1) and the convergence analysis for the AIR algorithm in §4.

3 Adaptively iterative reweighted algorithm

In this section, we present the adaptively iterative reweighted algorithm for minimizing the nonconvex and nonsmooth sparse optimization problem (1).

3.1 Smoothing method

In this subsection, we show how we deal with the nonsmoothness. Before proceeding, we define the following functions for $\mathbf{x} \in X$. Problem (1) can be rewritten as

$$\min_{\mathbf{x}} J_0(\mathbf{x}) := f(\mathbf{x}) + \sum_{i\in\mathcal{G}} r_i(c_i(\mathbf{x}_i)) + \delta(\mathbf{x}|X). \tag{2}$$

Adding relaxation parameter $\boldsymbol{\epsilon} \in \mathbb{R}^m_+$ to smooth the (possibly) nondifferentiable r_i, we propose the relaxed problem as

$$\min_{\mathbf{x}} \ J(\mathbf{x}; \boldsymbol{\epsilon}) := f(\mathbf{x}) + \sum_{i \in \mathcal{G}} r_i(c_i(\mathbf{x}_i) + \epsilon_i) + \delta(\mathbf{x}|X), \tag{3}$$

and in particular, $J(\mathbf{x}; \mathbf{0}) = J_0(\mathbf{x})$. Here we extend the notation of ϕ_i and use $\phi_i(\mathbf{x}_i; \epsilon_i)$ to denote the relaxed sparsity-inducing function, so that

$$\begin{aligned} \phi_i(\mathbf{x}_i; \epsilon_i) &:= r_i(c_i(\mathbf{x}_i) + \epsilon_i), \\ \Phi(\mathbf{x}; \boldsymbol{\epsilon}) &:= \sum_{i \in \mathcal{G}} \phi_i(\mathbf{x}_i; \epsilon_i) \ \text{ and } \ \phi_i(\mathbf{x}_i) = \phi_i(\mathbf{x}_i; \mathbf{0}). \end{aligned}$$

The following theorem shows that the pointwise convergence of $J(\mathbf{x}; \boldsymbol{\epsilon})$ to $J_0(\mathbf{x})$ on X as $\boldsymbol{\epsilon} \to \mathbf{0}$.

Theorem 1 *For any $\mathbf{x} \in X$ and $\epsilon \in \mathbb{R}_{++}$, it holds true that*

$$J_0(\mathbf{x}) \le J(\mathbf{x}; \boldsymbol{\epsilon}) \le J_0(\mathbf{x}) + \sum_{c_i(\mathbf{x}_i)=0} r_i(\epsilon_i) + \sum_{c_i(\mathbf{x}_i)>0} r'(c_i(\mathbf{x}_i))\epsilon_i.$$

This implies that $J(\mathbf{x}; \boldsymbol{\epsilon})$ pointwise convergence to $J_0(\mathbf{x})$ on X as $\boldsymbol{\epsilon} \to \mathbf{0}$.

Proof The first inequality is trivial, so we only have to show the second inequality. Since $r(\cdot)$ is concave on $\mathbb{R}_+$, we have

$$r_i(z) \le r_i(z_0) + r_i'(z_0)(z - z_0) \quad \text{for any } z, z_0 \in \mathbb{R}_+, \tag{4}$$

Therefore,

$$\begin{aligned} J(\mathbf{x}; \boldsymbol{\epsilon}) =& f(\mathbf{x}) + \sum_{i \in \mathcal{G}} r_i(c_i(\mathbf{x}_i) + \epsilon_i) \\ =& f(\mathbf{x}) + \sum_{c_i(\mathbf{x}_i)=0} r_i(\epsilon_i) + \sum_{c_i(\mathbf{x}_i)>0} r_i(c_i(\mathbf{x}_i) + \epsilon_i) \\ \le& f(\mathbf{x}) + \sum_{c_i(\mathbf{x}_i)=0} r_i(\epsilon_i) + \sum_{c_i(\mathbf{x}_i)>0} r_i(c_i(\mathbf{x}_i)) + \sum_{c_i(\mathbf{x}_i)>0} r'(c_i(\mathbf{x}_i))\epsilon_i \\ =& J_0(\mathbf{x}) + \sum_{c_i(\mathbf{x}_i)=0} r_i(\epsilon_i) + \sum_{c_i(\mathbf{x}_i)>0} r'(c_i(\mathbf{x}_i))\epsilon_i, \end{aligned}$$

where the inequality follows by (4). This completes the first statement.

On the other hand, since

$$\lim_{\boldsymbol{\epsilon} \to 0} \sum_{c_i(\mathbf{x}_i)=0} r_i(\epsilon_i) + \sum_{c_i(\mathbf{x}_i)>0} r'(c_i(\mathbf{x}_i))\epsilon_i = 0,$$

it holds

$$\lim_{\boldsymbol{\epsilon} \to \mathbf{0}} J(\mathbf{x}; \boldsymbol{\epsilon}) = J_0(\mathbf{x}), \quad \mathbf{x} \in X.$$

□

3.2 Adaptively iterative reweighted algorithm

A convex and smooth function $G_{(\tilde{\mathbf{x}},\tilde{\boldsymbol{\epsilon}})}(\mathbf{x})$ can be derived as an approximation of $J(\tilde{\mathbf{x}},\tilde{\boldsymbol{\epsilon}})$ at $\tilde{\mathbf{x}}$ by linearizing r_i at $c_i(\tilde{\mathbf{x}}_i)+\tilde{\epsilon}_i$ to have the subproblem

$$G_{(\tilde{\mathbf{x}},\tilde{\boldsymbol{\epsilon}})}(\mathbf{x}) := f(\mathbf{x}) + \sum_{i\in\mathcal{G}} w_i(\tilde{\mathbf{x}}_i,\tilde{\epsilon}_i)c_i(\mathbf{x}_i) + \delta(\mathbf{x}|X), \tag{5}$$

where the weights are given by

$$w_i(\mathbf{x},\epsilon_i) = r_i'(c_i(\mathbf{x}_i)+\epsilon_i), \quad i\in\mathcal{G}.$$

Note that the relaxation parameter can be simply chosen as $\boldsymbol{\epsilon}=\mathbf{0}$ if r is smooth at 0.

At iterate $\mathbf{x}^k$, the new iterate is obtained by

$$\mathbf{x}^{k+1} \in \arg\min_{\mathbf{x}} G_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}).$$

Therefore, $\mathbf{x}^{k+1}$ satisfies optimality condition

$$0 \in \partial G_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^{k+1}).$$

The relaxation parameter is selected such that $\boldsymbol{\epsilon}^{k+1}\le\boldsymbol{\epsilon}^k$ and possibly driven to 0 as the algorithm proceeds.

Our proposed *A*daptively *I*terative *R*eweighted algorithm for nonconvex and nonsmooth sparse optimization problems is stated in Algorithm 1.

Algorithm 1 AIR: Adaptively Iterative Reweighted

1: (Initialization) Choose $\mathbf{x}^0\in X$ and $\boldsymbol{\epsilon}^0\in\mathbb{R}^n_{++}$. Set $k=0$.
2: (Subproblem Solution) Compute new iterate

$$\mathbf{x}^{k+1} \in \arg\min G_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}).$$

3: (Reweighting) Choose $\boldsymbol{\epsilon}^{k+1}\in(\mathbf{0},\boldsymbol{\epsilon}^k]$.
4: Set $k\leftarrow k+1$. Go to Step 2.

3.3 ℓ_1-Algorithm & ℓ_2-Algorithm

In this subsection, we describe the details of how to construct $G_{(\tilde{\mathbf{x}},\tilde{\boldsymbol{\epsilon}})}(\mathbf{x})$ for the nonconvex and nonsmooth sparsity-inducing functions (EXP)–(MCP) in Sect. 2. Notice that the relaxation parameter ϵ_i could set as 0 if $\lim\limits_{c_i\to 0+} r_i'(c_i)<+\infty$. For simplicity, denote $\tilde{w}_i = w_i(\tilde{\mathbf{x}}_i,\tilde{\epsilon}_i)$. In Table 1, we provide the explicit forms of the weights $\tilde{w}_i$ at $(\tilde{\mathbf{x}}_i,\tilde{\epsilon}_i)$ when choosing $c_i(\mathbf{x}_i)=\|\mathbf{x}_i\|_1$ and $c_i(\mathbf{x}_i)=\|\mathbf{x}_i\|_2^2$ for each case, so that the corresponding subproblem is an ℓ_1-norm sparsity-inducing problem and an ℓ_2-norm sparsity-inducing problem

$$G_{(\tilde{\mathbf{x}},\tilde{\boldsymbol{\epsilon}})}(\mathbf{x}) = f(\mathbf{x}) + \sum_{i\in\mathcal{G}} \tilde{w}_i\|\mathbf{x}_i\|_1 + \delta(\mathbf{x}|X) \quad \text{and}$$

$$G_{(\tilde{\mathbf{x}},\tilde{\boldsymbol{\epsilon}})}(\mathbf{x}) = f(\mathbf{x}) + \sum_{i\in\mathcal{G}} \tilde{w}_i\|\mathbf{x}_i\|_2^2 + \delta(\mathbf{x}|X).$$

Table 1 Different AIR weights based on different choice of r_i and c_i

ϕ_i	$r_i(c_i)$	$\tilde{w}_i$	$r_i(\infty)$	$r_i'(0+)$
(EXP)	$1-e^{-pc_i}$	$pe^{-p(\|\tilde{\mathbf{x}}_i\|_1)}$	$<\infty$	$<\infty$
	$1-e^{-p\sqrt{c_i}}$	$\frac{pe^{-p\tilde{\mathbf{x}}_i}}{2\tilde{\mathbf{x}}_i}$	$<\infty$	$<\infty$
(LPN)	c_i^p	$p(\|\tilde{\mathbf{x}}_i\|_1+\tilde{\epsilon}_i)^{p-1}$	$+\infty$	$+\infty$
	$c_i^{p/2}$	$\frac{p}{2}(\|\tilde{\mathbf{x}}_i\|_2^2+\epsilon_i)^{\frac{p}{2}-1}$	$+\infty$	$+\infty$
(LOG)	$\log(1+pc_i)$	$\frac{p}{1+p\|\tilde{\mathbf{x}}_i\|_1}$	$+\infty$	$<\infty$
	$\log(1+p\sqrt{c_i})$	$\frac{p}{2\sqrt{\|\tilde{\mathbf{x}}_i\|_2^2+\tilde{\epsilon}_i}(1+p\sqrt{\|\tilde{\mathbf{x}}_i\|_2^2+\tilde{\epsilon}_i})}$	$+\infty$	$+\infty$
(FRA)	$\frac{c_i}{c_i+p}$	$\frac{p}{(\|\tilde{\mathbf{x}}_i\|_1+p)^2}$	$<\infty$	$<\infty$
	$\frac{\sqrt{c_i}}{\sqrt{c_i}+p}$	$\frac{p}{2\sqrt{\|\tilde{\mathbf{x}}_i\|_2^2+\epsilon_i}(\sqrt{\|\tilde{\mathbf{x}}_i\|_2^2+\epsilon_i}+p)^2}$	$<\infty$	$+\infty$
(TAN)	$\frac{c_i}{c_i+p}$	$\frac{p}{1+p^2(\|\tilde{\mathbf{x}}_i\|_1)^2}$	$<\infty$	$<\infty$
	$c_i^{p/2}$	$\frac{p}{2\sqrt{\|\tilde{\mathbf{x}}_i\|_2^2+\epsilon_i}(1+p^2(\|\tilde{\mathbf{x}}_i\|_2^2+\epsilon_i))}$	$<\infty$	$+\infty$

For each sparsity-inducing function, we consider $c_i(\mathbf{x}_i) = \|\mathbf{x}_i\|_1$ in the first row and $c_i(\mathbf{x}_i) = \|\mathbf{x}_i\|_2^2$ in the second row. We also list the properties of the r_i with $c_i \to \infty$ and its side-derivative of r_i at 0 in the fourth and fifth columns. This is because these properties can lead to different behaviors of each AIR as shown in the theoretical analysis.

As for SCAD and MCP, the explicit forms of r_i are not necessary to be known, but it can be easily verified using r_i' that Assumption (1) still holds true. The reweighted ℓ_1 subproblem for SCAD has weights

$$\tilde{w}_i = \lambda\{\mathbb{I}(|\tilde{x}_i| + \tilde{\epsilon}_i \le \lambda) + \frac{(a\lambda - \|\tilde{\mathbf{x}}_i\|_1 - \tilde{\epsilon}_i)_+}{(a-1)\lambda}\mathbb{I}(\|\tilde{\mathbf{x}}_i\|_1 + \tilde{\epsilon}_i > \lambda)\}.$$

The weights of reweighted ℓ_2 subproblem for SCAD are

$$\begin{aligned}\tilde{w}_i =& \frac{\lambda}{2\sqrt{\|\tilde{\mathbf{x}}_i\|_2^2+\epsilon_i)}}\{\mathbb{I}(\sqrt{\|\tilde{\mathbf{x}}_i\|_2^2+\epsilon_i)} \le \lambda) \\ &+ \frac{(a\lambda - \sqrt{\|\tilde{\mathbf{x}}_i\|_2^2+\epsilon_i)})_+}{(a-1)\lambda}\mathbb{I}(\sqrt{\|\tilde{\mathbf{x}}_i\|_2^2+\epsilon_i)} > \lambda)\}.\end{aligned}$$

As for MCP, the reweighted ℓ_1 subproblem has weights

$$\tilde{w}_i = (a\lambda - \|\tilde{\mathbf{x}}_i\|_1 - \tilde{\epsilon}_i)_+/a,$$

and the weights for reweighted ℓ_2 subproblem are

$$\tilde{w}_i = (a\lambda - \sqrt{\|\tilde{\mathbf{x}}_i\|_2^2+\epsilon_i})_+/a.$$

4 Convergence analysis

In this section, we analyze the global convergence of our proposed AIR. First we provide a unified first-order optimality condition for the constrained nonconvex and nonsmooth sparse optimization problem (1). Then we establish the global convergence anlaysis followed by the existence of cluster points.

For simplicity, denote $w_i^k = w_i(\mathbf{x}^k, \epsilon_i^k)$, $\mathbf{w}_i^k = w_i^k \mathbf{e}_{n_i}$, $\mathbf{w}^k = [\mathbf{w}_1^k; \mathbf{w}_2^k; \ldots; \mathbf{w}_m^k]$, and $\mathbf{W}^k = \text{diag}(\mathbf{w}^k)$, and so forth.

4.1 First-order optimality condition

In this subsection, we derive conditions to characterize the optimal solution of (1). Due to the nonconvex and nonsmooth nature of the sparsity-inducing function, we use Fréchet subdifferentials as the major tool in our analysis. Some important properties of Fréchet subdifferentials derived in [21] that will be used in this paper are summarized below. Part (i)-(iv) are Proposition 1.1, 1.2, 1.10, 1.13 and 1.18 in [21], respectively.

Proposition 1 *The following statements about Fréchet subdifferentials is true.*

(i) If f is differentiable at $\mathbf{x}$ with gradient $\nabla f(\mathbf{x})$, then $\partial_F f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.
(ii) If f is convex, then $\partial_F f(\mathbf{x}) = \partial f(\mathbf{x})$.
(iii) If f is Fréchet subdifferential at $\mathbf{x}$ and attains local minimum at $\mathbf{x}$, then

$$\mathbf{0} \in \partial_F f(\mathbf{x}).$$

(iv) Let $r(\cdot)$ be Fréchet subdifferentiable at $c^* = c(\mathbf{x}^*)$ with $c(\mathbf{x})$ being convex, then $r \circ c(\mathbf{x})$ is Fréchet subdifferentiable at $\mathbf{x}^*$ and that

$$y^* \partial c(\mathbf{x}^*) \subset \partial_F r \circ c(\mathbf{x}^*)$$

for any $y^* \in \partial_F r(c^*)$.
(v) $N(\mathbf{x}|X) = \partial_F \delta(\mathbf{x}|X)$ if X is closed and convex.

The properties of Fréchet subdifferentials in Proposition 1 can be used to characterize the optimal solution of (1). The following theorem is straightforward from Proposition 1, which describes the necessary optimality condition of problem (1).

Theorem 2 *If (3) attains a local minimum at $\mathbf{x}$, then it holds true that*

$$\mathbf{0} \in \partial_F J(\mathbf{x}; \boldsymbol{\epsilon}) = \nabla f(\mathbf{x}) + \partial_F \Phi(\mathbf{x}; \boldsymbol{\epsilon}) + N(\mathbf{x}|X). \tag{6}$$

Next we shall further investigate the properties of $\partial_F \phi(\mathbf{x}; \boldsymbol{\epsilon})$.

Lemma 1 *It holds that*

$$\nabla f(\mathbf{x}) + \prod_{i \in \mathcal{G}} y_i \partial c_i(\mathbf{x}_i) + N(\mathbf{x}|X) \subset \partial_F J(\mathbf{x}; \boldsymbol{\epsilon})$$

for any $y_i \in \partial_F r_i(c_i(\mathbf{x}_i) + \epsilon_i)$.

Proof Note that $\phi(\mathbf{x}; \boldsymbol{\epsilon})$ takes structure

$$\Phi(\mathbf{x}; \boldsymbol{\epsilon}) = \sum_{i \in \mathcal{G}} \phi_i(\mathbf{x}_i; \epsilon_i) \quad \text{with } \phi_i(\mathbf{x}_i; \epsilon_i) = r_i(c_i(\mathbf{x}_i) + \epsilon_i).$$

Thus we can write the Fréchet subdifferentials of Φ

$$\begin{aligned}\partial_F \Phi(\mathbf{x};\boldsymbol{\epsilon}) &= \prod_{i\in\mathcal{G}} \partial_F \phi_i(\mathbf{x}_i;\epsilon_i)\\ &= \partial_F \phi_1(\mathbf{x}_1;\epsilon_1)\times\ldots\times\partial_F\phi_m(\mathbf{x}_m;\epsilon_m),\end{aligned}$$

meaning that

$$\partial_F J(\mathbf{x};\boldsymbol{\epsilon}) = \nabla f(\mathbf{x}) + \prod_{i\in\mathcal{G}} \partial_F \phi_i(\mathbf{x}_i;\epsilon_i) + N(\mathbf{x}|X).$$

On the other hand, every c_i is assumed to be convex. From Proposition 1, we know that

$$y_i\partial c_i(\mathbf{x}_i) \subset \partial_F\phi_i(\mathbf{x}_i;\epsilon_i),\quad \forall y_i\in\partial_F r_i(c_i(\mathbf{x}_i)+\epsilon_i),$$

completing the proof. □

If $c_i(\mathbf{x}_i) > 0$ or $\epsilon_i > 0$, r_i is differentiable at $c_i+\epsilon_i$ so that $\partial_F\phi_i(\mathbf{x}_i;\epsilon_i) = r_i'(c_i(\mathbf{x}_i^*)+\epsilon_i)\partial c_i(\mathbf{x}^*)$ by Proposition 1. Of particular interests are the properties of $\partial_F r_i(0)$. Notice that r_i' is decreasing on $\mathbb{R}_{++}$. We investigate $\partial_F\phi_i(\mathbf{x}_i;\epsilon_i)$ bases on the limits (possibly infinite) in the lemma below.

Lemma 2 *Let* $y_i^* := \lim_{c_i\to 0^+} r_i'(c_i) \geq 0$. *It holds true that*

$$\begin{cases}\partial_F r_i(c_i) = r_i'(c_i) & \textit{if } c_i > 0\\ \partial_F r_i(0) = [-y_i^*, y_i^*], & \textit{if } y_i^* < +\infty,\\ \partial_F r_i(0) = \mathbb{R}, & \textit{if } y_i^* = +\infty,\end{cases}$$

so that

1. If $c_i(\mathbf{x}_i^*)+\epsilon_i > 0$,

$$\partial_F\phi_i(\mathbf{x}_i;\epsilon_i) = r_i'(c_i(\mathbf{x}_i^*)+\epsilon_i)\partial c_i(\mathbf{x}_i^*);$$

2. If $c_i(\mathbf{x}_i^*)+\epsilon_i = 0$, $y_i^* < +\infty$,

$$y_i\partial c_i(\mathbf{x}_i^*) \subset \partial_F\phi_i(\mathbf{x}_i;\epsilon_i),\ \forall y_i\in[-y_i^*, y_i^*];$$

3. If $c_i(\mathbf{x}_i^*)+\epsilon_i = 0$, $y_i^* = +\infty$,

$$y_i\partial c_i(\mathbf{x}_i^*) \subset \partial_F\phi_i(\mathbf{x}_i;\epsilon_i),\ \forall y_i\in\mathbb{R}.$$

Proof The statement about the case that $c_i(\mathbf{x}_i^*) > 0$ is obviously true. We only need consider the case that $c_i(\mathbf{x}_i^*) = 0$. Notice that

$$\liminf_{c_i\to 0^+}\frac{r_i(c_i)-r_i(0)}{c_i} = \liminf_{\substack{0<\tilde{c}_i<c_i\\ c_i\to 0^+}} r_i'(\tilde{c}_i) = r_i'(0+) = y_i^* \geq 0$$

by Assumption 1(iii). It can be easily verified by [21, Proposition 1.17] that

$$\partial_F r_i(0) = \begin{cases}[-y_i^*, y_i^*] & \text{if } y_i^* < +\infty,\\ \mathbb{R} & \text{if } y_i^* = +\infty.\end{cases}$$

It then follows from Proposition 1(iv) that

$$\begin{cases} y_i\partial c_i(\mathbf{x}_i^*) \subset \partial_F\phi_i(\mathbf{x}_i;\epsilon_i), \forall y_i\in[-y_i^*, y_i^*], & \text{if } y_i^* < +\infty,\\ y_i\partial c_i(\mathbf{x}_i^*) \subset \partial_F\phi_i(\mathbf{x}_i;\epsilon_i), \forall y_i\in\mathbb{R}, & \text{if } y_i^* = +\infty.\end{cases}$$

□

Note that we only require $\epsilon \in \mathbb{R}_+$. If $\epsilon = \mathbf{0}$, all the results we have derived for $J(\cdot\,;\epsilon)$ in this subsection also hold for J_0.

4.2 Global convergence of the AIR algorithm

In this subsection, we analyze the global convergence of AIR under Assumption 1. First of all, we need to show that the subproblem always has a solution. For $\hat{\epsilon} \in \mathbb{R}_{++}$, the subproblem is obviously well-defined on X since the weights $w_i^k = r_i'(\mathbf{x}_i^k + \epsilon_i^k) < +\infty$. To guarantee the proposed AIR is well defined, we must show the existence of the subproblem solution.

We have the following lemma about the solvability of the subproblems.

Lemma 3 *For $\epsilon^k \in \mathbb{R}_{++}$, $\arg\min_{\mathbf{x}} G_{(\mathbf{x}^k,\epsilon^k)}(\mathbf{x})$ is nonempty, so that $\mathbf{x}^{k+1}$ is well-defined.*

Proof Pick $\tilde{\mathbf{x}} \in X$ and let $\alpha := G_{(\mathbf{x}^k,\epsilon^k)}(\tilde{\mathbf{x}})$. The level set

$$\{\mathbf{x} \in X | G_{(\mathbf{x}^k,\epsilon^k)}(\mathbf{x}) \le G_{(\mathbf{x}^k,\epsilon^k)}(\tilde{\mathbf{x}})\}$$

must be nonempty since it contains $\tilde{\mathbf{x}}$, and bounded due to the coercivity of $w_i^k c_i$, $i \in \mathcal{G}$ and the lower boundedness of f on X. This completes the proof by [33, Theorem 4.3.1]. □

We have the following key facts about solutions to (5), which implies that the new iterate $\mathbf{x}^{k+1}$ causes a decrease in the model $J(\mathbf{x}, \epsilon^k)$.

Lemma 4 *Let $\tilde{\mathbf{x}} \in X$, $\hat{\epsilon}, \tilde{\epsilon} \in \mathbb{R}^m_{++}$ with $\hat{\epsilon} \le \tilde{\epsilon}$ and $\tilde{w}_i = w_i(\tilde{\mathbf{x}}_i, \tilde{\epsilon}_i)$ for $i \in \mathcal{G}$. Suppose that $\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in X} G_{(\tilde{\mathbf{x}},\tilde{\epsilon})}(\mathbf{x})$. Then, for any k, it holds true that*

$$J(\hat{\mathbf{x}}, \hat{\epsilon}) - J(\tilde{\mathbf{x}}, \tilde{\epsilon}) \le G_{(\tilde{\mathbf{x}},\tilde{\epsilon})}(\hat{\mathbf{x}}) - G_{(\tilde{\mathbf{x}},\tilde{\epsilon})}(\tilde{\mathbf{x}}) \le 0.$$

Proof First of all, $\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}} G_{(\tilde{\mathbf{x}},\tilde{\epsilon})}(\mathbf{x})$, so that $G_{(\tilde{\mathbf{x}},\tilde{\epsilon})}(\hat{\mathbf{x}}) - G_{(\tilde{\mathbf{x}},\tilde{\epsilon})}(\tilde{\mathbf{x}}) \le 0$. Hence

$$\begin{aligned} J(\hat{\mathbf{x}}; \hat{\epsilon}) &\le J(\hat{\mathbf{x}}; \tilde{\epsilon}) = f(\hat{\mathbf{x}}) + \sum_{i\in\mathcal{G}} r_i(c_i(\hat{\mathbf{x}}_i) + \tilde{\epsilon}_i) \\ &\le f(\tilde{\mathbf{x}}) + f(\hat{\mathbf{x}}) - f(\tilde{\mathbf{x}}) + \sum_{i\in\mathcal{G}} r_i(c_i(\tilde{\mathbf{x}}) + \tilde{\epsilon}_i) + \sum_{i\in\mathcal{G}} \tilde{w}_i(c_i(\hat{\mathbf{x}}) - c_i(\tilde{\mathbf{x}})) \\ &= J(\tilde{\mathbf{x}}; \tilde{\epsilon}) + [G_{(\tilde{\mathbf{x}},\tilde{\epsilon})}(\hat{\mathbf{x}}) - G_{(\tilde{\mathbf{x}},\tilde{\epsilon})}(\tilde{\mathbf{x}})], \end{aligned}$$

where the second inequality follows from (4). □

Lemma 4 indicates $J(\mathbf{x}; \epsilon)$ is monotonically decreasing for any $\mathbf{x}^0 \in X$, $\epsilon^0 \in \mathbb{R}^m$. Define the model reduction

$$\Delta G_{(\mathbf{x}^k,\epsilon^k)}(\mathbf{x}^{k+1}) = G_{(\mathbf{x}^k,\epsilon^k)}(\mathbf{x}^k) - G_{(\mathbf{x}^k,\epsilon^k)}(\mathbf{x}^{k+1}).$$

The next lemma indicates this model reduction converges to zero, which naturally follows from Lemma 4.

Lemma 5 *Suppose $\mathbf{x}^0 \in X$, $\epsilon^0 \in \mathbb{R}^m_{++}$, and $\{\mathbf{x}^k\}$ are generated by AIR. The following statements hold true*

(i) The sequence $\{\mathbf{x}^k\} \subset L(J(\mathbf{x}^0; \epsilon^0); J_0)$.

(ii) $\lim_{k\to\infty} \Delta G_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^{k+1}) \to 0$.

Proof Part (i) follows naturally from the fact that

$$J_0(\mathbf{x}^k) \le J(\mathbf{x}^k, \boldsymbol{\epsilon}^k) \le J(\mathbf{x}^0, \boldsymbol{\epsilon}^0),$$

for all $k \in \mathbb{N}$ by Lemma 4.

For part (ii), by Assumption 1, $\tilde{J} := \inf_k J(\mathbf{x}^k; \boldsymbol{\epsilon}^k) > -\infty$. It follows from Lemma 4, that

$$J(\mathbf{x}^{k+1}, \boldsymbol{\epsilon}^{k+1}) \le J(\mathbf{x}^k, \boldsymbol{\epsilon}^k) - \Delta G_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^{k+1}).$$

Summing up both sides of the above inequality from 0 to t, we have

$$\begin{aligned} 0 &\le \sum_{k=1}^{t} \Delta G_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^{k+1}) \\ &\le J(\mathbf{x}^0, \boldsymbol{\epsilon}^0) - J(\mathbf{x}^{t+1}, \boldsymbol{\epsilon}^{t+1}) \le J(\mathbf{x}^0, \boldsymbol{\epsilon}^0) - \tilde{J}. \end{aligned}$$

Letting $t \to \infty$, we know part (ii) holds true. □

4.2.1 Convergence analysis for bounded weights

We first analyze the convergence when $\boldsymbol{\epsilon}^k \to \boldsymbol{\epsilon}^* \in \mathbb{R}_{++}$ or $\lim_{c_i\to 0^+} r_i'(c_i) < +\infty, i \in \mathcal{G}$. In this case, $w_i^k \to w_i^* < +\infty$ if $\mathbf{x}_i^k \to \mathbf{0}$. The "limit subproblem" takes form

$$\min_{\mathbf{x}} \quad \widetilde{G}_{(\tilde{\mathbf{x}},\tilde{\boldsymbol{\epsilon}})}(\mathbf{x}) := f(\mathbf{x}) + \sum_{i\in\mathcal{G}} \tilde{w}_i c_i(\mathbf{x}_i) + \delta(\mathbf{x}|X). \tag{7}$$

The existence of the solution to (7) is shown in the next lemma.

Lemma 6 *For $\tilde{\boldsymbol{\epsilon}} \in \mathbb{R}_{++}$, the optimal solution set of (7) is nonempty. Furthermore, if $\tilde{\mathbf{x}}$ is an optimal solution of (7), then $\tilde{\mathbf{x}}$ also satisfies the first-order optimality condition of (3).*

Proof Notice that $\tilde{\mathbf{x}}$ is feasible for (7) by the definition of $\widetilde{G}$. The level set

$$\{\mathbf{x} \in X \mid \widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}) \le \widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\tilde{\mathbf{x}})\}$$

must be nonempty since it contains $\tilde{\mathbf{x}}$ and bounded due to the coercivity of $\tilde{w}_i c_i$, $i \in \mathcal{G}$ and the lower boundedness of f on X. This completes the proof by [33, Theorem 4.3.1].

Therefore, any optimal solution $\mathbf{x}$ must satisfies

$$\mathbf{0} = \nabla f(\mathbf{x})_i + \mathbf{z}_i + \boldsymbol{\nu}_i, i \in \mathcal{G}$$

where $\boldsymbol{\nu} \in N(\mathbf{x}|X)$, $\mathbf{z}_i = \tilde{w}_i \boldsymbol{\xi}_i$ with

$$\tilde{w}_i \in \partial_F r_i(c_i(\tilde{\mathbf{x}}_i) + \tilde{\boldsymbol{\epsilon}}_i),\ \boldsymbol{\xi}_i \in \partial c_i(\mathbf{x}_i),\ i \in \mathcal{G}.$$

The KKT conditions thus can be rewritten as following by Lemma 2

$$\begin{aligned} \mathbf{0} &= \nabla f(\mathbf{x})_i + \tilde{w}_i \boldsymbol{\xi}_i + \boldsymbol{\nu}_i, \\ \tilde{w}_i &\in \partial_F r_i(c_i(\tilde{\mathbf{x}}_i) + \tilde{\epsilon}_i),\ \boldsymbol{\xi}_i \in \partial c_i(\mathbf{x}_i), \end{aligned}$$

where $i \in \mathcal{G}$. If $\tilde{\mathbf{x}}$ is an optimal solution, then we have

$$\mathbf{0} \in \nabla f(\tilde{\mathbf{x}}) + \partial_F \Phi(\tilde{\mathbf{x}}; \tilde{\boldsymbol{\epsilon}}) + N(\tilde{\mathbf{x}}|X),$$

implying $\tilde{\mathbf{x}}$ is optimal for $J(\,\cdot\,; \tilde{\boldsymbol{\epsilon}})$. □

Now we are ready to prove our main result in this section.

Theorem 3 *Suppose* $\{\mathbf{x}^k\}_{k=0}^{\infty}$ *is generated by AIR with initial point* $\mathbf{x}^0 \in X$ *and relaxation vector* $\boldsymbol{\epsilon}^0 \in \mathbb{R}^m_{++}$ *with* $\boldsymbol{\epsilon}^k \to \boldsymbol{\epsilon}^*$. *Assume either*

$$\epsilon_i^* > 0 \text{ or } r'(0+) < +\infty, \ i \in \mathcal{G}$$

is true. Then if $\{\mathbf{x}^k\}$ *has any cluster point, it satisfies the optimality condition (6) for* $J(\mathbf{x}; \boldsymbol{\epsilon}^*)$.

Proof Let $\mathbf{x}^*$ be a cluster point of $\{\mathbf{x}^k\}$. From Lemma 6, it suffices to show that $\mathbf{x}^* \in \arg\min_{\mathbf{x}} \widetilde{G}_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\mathbf{x})$. We prove this by contradiction. Assume that there exists a point $\bar{\mathbf{x}}$ such that $\varepsilon := \widetilde{G}_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\mathbf{x}^*) - \widetilde{G}_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\bar{\mathbf{x}}) > 0$. Suppose $\{\mathbf{x}^k\}_{\mathcal{S}} \to \mathbf{x}^*$, $\mathcal{S} \subset \mathbb{N}$. Based on Lemma 5$(ii)$, there exists $k_1 > 0$, such that for all $k > k_1$

$$\widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^k) - \widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^{k+1}) \le \varepsilon/4. \tag{8}$$

To derive a contradiction, notice that $\mathbf{x}_i^k \overset{\mathcal{S}}{\to} \mathbf{x}_i^*$ and $w_i^k \overset{\mathcal{S}}{\to} w_i^*$. There exists k_2 such that for all $k > k_2$, $k \in \mathcal{S}$,

$$\begin{aligned} \sum_{i\in\mathcal{G}} (w_i^* - w_i^k) c_i(\bar{\mathbf{x}}_i) &\ge -\varepsilon/12, \\ \sum_{i\in\mathcal{G}} (w_i^k c_i(\mathbf{x}_i^k) - w_i^* c_i(\mathbf{x}_i^*)) &\ge -\varepsilon/12, \\ f(\mathbf{x}^k) - f(\mathbf{x}^*) &\ge -\varepsilon/12. \end{aligned}$$

Therefore, for all $k > k_2$, $k \in \mathcal{S}$,

$$\begin{aligned} &\widetilde{G}_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\mathbf{x}^*) - \widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\bar{\mathbf{x}}) \\ &\quad = [f(\mathbf{x}^*) + \sum_{i\in\mathcal{G}} w_i^* c_i(\mathbf{x}_i^*)] - [f(\bar{\mathbf{x}}) + \sum_{i\in\mathcal{G}} [w_i^* - (w_i^* - w_i^k)] c_i(\bar{\mathbf{x}}_i)) \\ &\quad = [\widetilde{G}_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\mathbf{x}^*) - \widetilde{G}_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\bar{\mathbf{x}})] + \sum_{i\in\mathcal{G}} (w_i^* - w_i^k) c_i(\bar{\mathbf{x}}_i), \\ &\quad \ge [\widetilde{G}_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\mathbf{x}^*) - \widetilde{G}_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\bar{\mathbf{x}})] - \varepsilon/12 \\ &\quad = \varepsilon - \varepsilon/12 = 11\varepsilon/12, \end{aligned}$$

and that

$$\begin{aligned} &\widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^k) - \widetilde{G}_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\mathbf{x}^*) \\ &\quad = [f(\mathbf{x}^k) + \sum_{i\in\mathcal{G}} w_i^k c_i(\mathbf{x}_i^k)] - [f(\mathbf{x}^*) + \sum_{i\in\mathcal{G}} w_i^* c_i(\mathbf{x}_i^*)] \\ &\quad \ge -\varepsilon/6 \end{aligned}$$

Hence, for all $k > \max(k_1, k_2)$, $k \in \mathcal{S}$, it holds that

$$\begin{aligned} &\widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^k) - \widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\bar{\mathbf{x}}) \\ &\quad = \widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^k) - \widetilde{G}_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\mathbf{x}^*) + \widetilde{G}_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\mathbf{x}^*) - \widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\bar{\mathbf{x}}) \\ &\quad = 11\varepsilon/12 - \varepsilon/6 = 3\varepsilon/4, \end{aligned}$$

contradicting with (8). Therefore, $\mathbf{x}^* \in \arg\min_{\mathbf{x}} \widetilde{G}_{(\mathbf{x}^*,\mathbf{0})}(\mathbf{x})$. By Lemma 6, $\mathbf{x}^*$ satisfies the first-order optimality for (3). □

Remark 2 Note that there are several choices of ϕ_i satisfy $r'(0+) < +\infty,\ i \in \mathcal{G}$ as shown in Table 1. For example, ϕ_i takes the (EXP) form with both $c_i(\boldsymbol{x}_i) = \|\boldsymbol{x}_i\|_1$ and $c_i(\boldsymbol{x}_i) = \|\boldsymbol{x}_i\|_2^2$, ϕ_i takes the (LOG) form with $c_i(\boldsymbol{x}_i) = \|\boldsymbol{x}_i\|_1$ and ϕ_i takes the (TAN) form with $c_i(\boldsymbol{x}_i) = \|\boldsymbol{x}_i\|_1$.

Remark 3 The convexity of f is not necessary if $\mathbf{x}^{k+1}$ is found as the global minimizer of (3). In this case, the global convergence we have derived so far can be modified accordingly, and in the statement of Lemma 6, a global minimizer $\tilde{\mathbf{x}}$ of (7) implies its optimality of (3).

4.2.2 Convergence analysis for reweighed ℓ_1 and ℓ_2 with vanishing $\boldsymbol{\epsilon}$

We have shown the convergence of AIR with fixed $\boldsymbol{\epsilon}$. By Theorem 1, we can choose sufficiently small $\boldsymbol{\epsilon}$ and minimize $J(\cdot;\boldsymbol{\epsilon})$ instead of J_0 to obtain an approximate solution. However, as also shown by Theorem 1, $J(\cdot;\boldsymbol{\epsilon})$ converges to J_0 only pointwisely. It then may be difficult to assert that the minimizer of $J(\cdot;\boldsymbol{\epsilon})$ is sufficiently close to the minimizer of J_0 for given $\boldsymbol{\epsilon}$. Therefore, we consider to minimize a sequence of $J(\cdot;\boldsymbol{\epsilon})$ with $\boldsymbol{\epsilon}$ driven to $\mathbf{0}$.

As the algorithm proceeds, of particular interest is the properties of the "limit subproblem" as the (sub)sequence of iterates converges. Let $\mathcal{L} := \{i \mid r_i'(0+) = +\infty\}$. Notice that it may happen $w_i^k \to \infty$ if $\mathbf{x}_i^k \to 0$ and $\epsilon_i^k \to 0$, so that G may be not well-defined. Therefore we consider an alternative form of the "limit subproblem" for $\tilde{\boldsymbol{\epsilon}} = \mathbf{0}$

$$\begin{aligned} \min_{\mathbf{x}}\quad & \widetilde{G}_{(\tilde{\mathbf{x}},\mathbf{0})}(\mathbf{x}) := f(\mathbf{x}) + \sum_{i\in\mathcal{N}(\tilde{\mathbf{x}},\mathbf{0})} \tilde{w}_i c_i(\mathbf{x}_i) + \delta(\mathbf{x}|X), \\ \text{s.t.}\quad & \mathbf{x}_i = \mathbf{0},\ i \in \mathcal{A}(\tilde{\mathbf{x}},\mathbf{0}), \end{aligned} \tag{9}$$

where $\mathcal{A}(\tilde{\mathbf{x}},\mathbf{0}) := \{i \mid \tilde{\mathbf{x}}_i = \mathbf{0}, \tilde{\epsilon}_i = 0\} \cap \mathcal{L}$ and $\mathcal{N}(\tilde{\mathbf{x}},\mathbf{0}) := \mathcal{G} \setminus \mathcal{A}(\tilde{\mathbf{x}},\mathbf{0})$. The existence of the solution to (9) is shown in the next lemma.

Lemma 7 *For $\tilde{\boldsymbol{\epsilon}} = \mathbf{0}$, the optimal solution set of (9) is nonempty. Furthermore, if $\tilde{\mathbf{x}}$ is an optimal solution of (9), then $\tilde{\mathbf{x}}$ also satisfies the first-order optimality condition of (2).*

Proof Notice that $\tilde{\mathbf{x}}$ is feasible for (general.sub.alter) by the definition of $\widetilde{G}$. The level set

$$\{\mathbf{x} \in X \mid \widetilde{G}_{(\tilde{\mathbf{x}},\mathbf{0})}(\mathbf{x}) \le \widetilde{G}_{(\tilde{\mathbf{x}},\mathbf{0})}(\tilde{\mathbf{x}});\ \mathbf{x}_i = \mathbf{0}, i \in \mathcal{A}(\tilde{\mathbf{x}},\mathbf{0})\}$$

must be nonempty since it contains $\tilde{\mathbf{x}}$ and bounded due to the coercivity of $\tilde{w}_i c_i$, $i \in \mathcal{G}$ and the lower boundedness of f on X. This completes the proof by [33, Theorem 4.3.1].

Obviously Slater's condition holds at any feasible point of (general.sub.alter). Therefore, any optimal solution $\mathbf{x}$ must satisfies the KKT conditions

$$\mathbf{0} = \nabla f(\mathbf{x})_i + \mathbf{z}_i + \boldsymbol{\nu}_i,\ i \in \mathcal{G}$$

with $\boldsymbol{\nu} \in N(\mathbf{x}|X)$, $\mathbf{z}_i = \tilde{y}_i \boldsymbol{\xi}_i$. For the case $c_i(\mathbf{x}_i) = \|\mathbf{x}_i\|_1$, let $\tilde{y}_i := \tilde{w}_i = r_i'(c_i(\tilde{\mathbf{x}}_i))$, $\boldsymbol{\xi}_i \in \partial c_i(\mathbf{x}_i)$, $i \in \mathcal{N}(\tilde{\mathbf{x}},\mathbf{0})$. Now for $i \in \mathcal{A}(\tilde{\mathbf{x}},\mathbf{0})$, let $\tilde{y}_i = \|\mathbf{z}_i\|_\infty$ and $\boldsymbol{\xi}_i = \mathbf{z}_i/\|\mathbf{z}_i\|_\infty$ so that $\boldsymbol{\xi}_i \in \partial c_i(0) = \partial c_i(\tilde{\mathbf{x}}_i)$. For the case $c_i(\mathbf{x}_i) = \|\mathbf{x}_i\|_2^2$, let $\tilde{y}_i := \tilde{w}_i = r_i'(c_i(\tilde{\mathbf{x}}_i))$, $\boldsymbol{\xi}_i = 2\tilde{\mathbf{x}}_i$, $i \in \mathcal{N}(\tilde{\mathbf{x}},\mathbf{0})$. Now for $i \in \mathcal{A}(\tilde{\mathbf{x}},\mathbf{0})$, let $\tilde{y}_i = \|\mathbf{z}_i\|_\infty$ and $\boldsymbol{\xi}_i = \mathbf{z}_i/\|\mathbf{z}_i\|_\infty$ so that $\boldsymbol{\xi}_i \in \partial\|\mathbf{0}\|_1$. The KKT conditions can be rewritten as

$$\begin{aligned} &\mathbf{0} = \nabla f(\mathbf{x})_i + \tilde{y}_i \boldsymbol{\xi}_i + \boldsymbol{\nu}_i, \\ &\text{when } i \in \mathcal{N}(\tilde{\mathbf{x}},\mathbf{0}) : \begin{cases} y_i \in \partial_F r_i(c_i(\tilde{\mathbf{x}}_i)), \\ \boldsymbol{\xi}_i \in \partial c_i(\mathbf{x}_i) \end{cases} \end{aligned}$$

$$\text{when } i \in \mathcal{A}(\tilde{\mathbf{x}}, \mathbf{0}) : \begin{cases} y_i \in \partial_F r_i(\sqrt{c_i(\tilde{\mathbf{x}}_i)}), \\ \boldsymbol{\xi}_i \in \partial\sqrt{c_i(\mathbf{x}_i)} \end{cases}$$

For both cases, the KKT conditions can be rewritten as

$$\begin{aligned} \mathbf{0} &= \nabla f(\mathbf{x})_i + \tilde{y}_i \boldsymbol{\xi}_i + \boldsymbol{\nu}_i, \\ y_i &\in \partial_F r_i(c_i(\tilde{\mathbf{x}}_i)), \\ \boldsymbol{\xi}_i &\in \partial c_i(\mathbf{x}_i), \, i \in \mathcal{G} \end{aligned}$$

by Lemma subdiffr. If $\tilde{\mathbf{x}}$ is an optimal solution, then we have

$$\mathbf{0} \in f(\tilde{\mathbf{x}}) + \partial_F \phi(\tilde{\mathbf{x}}; \mathbf{0}) + N(\tilde{\mathbf{x}}|X),$$

implying $\tilde{\mathbf{x}}$ is optimal for $J_0(\cdot)$. □

Now we are ready to prove our main result in this section.

Theorem 4 *Suppose sequence $\{\mathbf{x}^k\}_{k=0}^{\infty}$ is generated by AIR ℓ_1-algorithm with initial point $\mathbf{x}^0 \in X$ and relaxation vector $\boldsymbol{\epsilon}^0 \in \mathbb{R}^m_{++}$. If $\{\mathbf{x}^k\}$ has any cluster point $\mathbf{x}^*$, then it satisfies the optimality condition for $J_0(\mathbf{x})$.*

Proof Let $\mathbf{x}^*$ be a cluster point of $\{\mathbf{x}^k\}$ and $\lim_{k\to\infty} \boldsymbol{\epsilon}^k = \mathbf{0}$. There exists $k_0 > 0$ such that for all $k > k_0$, $\boldsymbol{\epsilon}^k = \mathbf{0}$. From Lemma 7, it suffices to show that $\mathbf{x}^* \in \arg\min_{\mathbf{x}} \widetilde{G}_{(\mathbf{x}^*,\mathbf{0})}(\mathbf{x})$. We prove this by contradiction. Assume that there exists a point $\bar{\mathbf{x}}$ such that $c_i(\bar{\mathbf{x}}_i) = 0$ for all $i \in \mathcal{A}(\mathbf{x}^*, \mathbf{0})$ and $G_{(\mathbf{x}^*,\mathbf{0})}(\mathbf{x}^*) - G_{(\mathbf{x}^*,\mathbf{0})}(\bar{\mathbf{x}}) > \varepsilon > 0$. Suppose $\{\mathbf{x}^k\}_{\mathcal{S}}$, $\mathcal{S} \subset \mathbb{N}$. Based on Lemma 5$(ii)$, there exists $k_1 > 0$, such that for all $k > k_1$

$$\widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^k) - \widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^{k+1}) \le \varepsilon/4. \tag{10}$$

Notice that $\mathbf{x}_i^k \xrightarrow{\mathcal{S}} \mathbf{x}_i^*$ and $w_i^k \xrightarrow{\mathcal{S}} w_i^*$. To derive a contradiction, there exists $k_2 > k_0$ such that for all $k > k_2$, $k \in \mathcal{S}$,

$$\begin{aligned} \sum_{i\in\mathcal{N}(\mathbf{x}^*,\mathbf{0})} (w_i^* - w_i^k) c_i(\bar{\mathbf{x}}_i) &> -\varepsilon/12, \\ \sum_{i\in\mathcal{N}(\mathbf{x}^*,\mathbf{0})} (w_i^k c_i(\mathbf{x}_i^k) - w_i^* c_i(\mathbf{x}_i^*)) &> -\varepsilon/12, \\ f(\mathbf{x}^k) - f(\mathbf{x}^*) &> -\varepsilon/12. \end{aligned}$$

Therefore, for all $k > k_2$, $k \in \mathcal{S}$,

$$\begin{aligned} &\widetilde{G}_{(\mathbf{x}^*,\mathbf{0})}(\mathbf{x}^*) - \widetilde{G}_{(\mathbf{x}^k,\mathbf{0})}(\bar{\mathbf{x}}) \\ &= [f(\mathbf{x}^*) + \sum_{i\in\mathcal{N}(\mathbf{x}^*,\mathbf{0})} w_i^* c_i(\mathbf{x}_i^*)] - [f(\bar{\mathbf{x}}) + \sum_{i\in\mathcal{N}(\mathbf{x}^*,\mathbf{0})} [w_i^* - (w_i^* - w_i^k)] c_i(\bar{\mathbf{x}}_i)] \\ &= [\widetilde{G}_{(\mathbf{x}^*,\mathbf{0})}(\mathbf{x}^*) - \widetilde{G}_{(\mathbf{x}^*,\mathbf{0})}(\bar{\mathbf{x}})] + \sum_{i\in\mathcal{N}(\mathbf{x}^*,\mathbf{0})} (w_i^* - w_i^k) c_i(\bar{\mathbf{x}}_i), \\ &\ge [\widetilde{G}_{(\mathbf{x}^*,\mathbf{0})}(\mathbf{x}^*) - \widetilde{G}_{(\mathbf{x}^*,\mathbf{0})}(\bar{\mathbf{x}})] - \varepsilon/12 \\ &\ge \varepsilon - \varepsilon/12 = 11\varepsilon/12, \end{aligned}$$

and that

$$\begin{aligned}
&\widetilde{G}_{(\mathbf{x}^k,\mathbf{0})}(\mathbf{x}^k) - \widetilde{G}_{(\mathbf{x}^*,\mathbf{0})}(\mathbf{x}^*) \\
&= [f(\mathbf{x}^k) + \sum_{i\in\mathcal{A}(\mathbf{x}^*,\mathbf{0})} w_i^k c_i(\mathbf{x}_i^k) + \sum_{i\in\mathcal{N}(\mathbf{x}^*,\mathbf{0})} w_i^k c_i(\mathbf{x}_i^k)] \\
&\quad - [f(\mathbf{x}^*) + \sum_{i\in\mathcal{N}(\mathbf{x}^*,\mathbf{0})} w_i^* c_i(\mathbf{x}_i^*)] \\
&\geq [f(\mathbf{x}^k) + \sum_{i\in\mathcal{N}(\mathbf{x}^*,\mathbf{0})} w_i^k c_i(\mathbf{x}_i^k)] - [f(\mathbf{x}^*) + \sum_{i\in\mathcal{N}(\mathbf{x}^*,\mathbf{0})} w_i^* c_i(\mathbf{x}_i^*)] \\
&\geq -\varepsilon/6
\end{aligned}$$

Hence, for all $k > \max(k_1, k_2)$, $k \in \mathcal{S}$, it holds that

$$\begin{aligned}
&\widetilde{G}_{(\mathbf{x}^k,\mathbf{0})}(\mathbf{x}^k) - \widetilde{G}_{(\mathbf{x}^k,\mathbf{0})}(\mathbf{x}^{k+1}) \\
&= \widetilde{G}_{(\mathbf{x}^k,\mathbf{0})}(\mathbf{x}^k) - \widetilde{G}_{(\mathbf{x}^*,\mathbf{0})}(\mathbf{x}^*) + \widetilde{G}_{(\mathbf{x}^*,\mathbf{0})}(\mathbf{x}^*) - \widetilde{G}_{(\mathbf{x}^k,\mathbf{0})}(\bar{\mathbf{x}}) \\
&= 11\varepsilon/12 - \varepsilon/6 = 3\varepsilon/4,
\end{aligned}$$

contradicting with (10). Therefore, $\mathbf{x}^* \in \arg\min_{\mathbf{x}} \widetilde{G}_{(\mathbf{x}^*,\mathbf{0})}(\mathbf{x})$. By Lemma 7, $\mathbf{x}^*$ satisfies the first-order optimality for (3). □

4.3 Existence of cluster points

We will show that our proposed algorithm AIR is a descent method for the function $J(\mathbf{x}, \boldsymbol{\epsilon})$. Consequently, both the existence of solutions to (1) as well as the existence of the cluster point to AIR can be guaranteed by understanding conditions under which the iterates generated by AIR is bounded. For this purpose, we need to investigate the asymptotic geometry of J and X. In the following a series of results, we discuss the conditions guaranteeing the boundedness of $L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)$. The concept of horizon cone is a useful tool to characterize the boundedness of a set, which is defined as follows.

Definition 2 [35, Definition 3.3] Given $Y \subset \mathbb{R}^n$, the horizon cone of Y is

$$Y^\infty := \{\mathbf{z} \mid \exists t^k \downarrow 0, \{\mathbf{y}^k\} \subset Y \text{ such that } t^k\mathbf{y}^k \to \mathbf{z}\}.$$

We have the basic properties about horizon cones given in the following proposition, where the first case is trivial to show and others are from [35].

Proposition 2 *The following hold:*

(i) If $X \subset Y \subset \mathbb{R}^n$, then $X^\infty \subset Y^\infty$.
(ii) [35, Theorem 3.5] The set $Y \subset \mathbb{R}^n$ is bounded if and only if $Y^\infty = \{0\}$.
(iii) [35, Exercise 3.11] Given $Y_i \subset \mathbb{R}^{n_i}$ for $i \in \mathcal{G}$, we have $(Y_1 \times \ldots \times Y_m)^\infty = Y_1^\infty \times \ldots \times Y_m^\infty$.
(iv) [35, Theorem 3.6] If $C \subset \mathbb{R}^n$ is non-empty, closed, and convex, then

$$C^\infty = \{\mathbf{z} \mid C + \mathbf{z} \subset C\}.$$

Next we investigate the boundedness of $L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0), J_0)$, and provide upper and lower estimates of $L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0), J_0)$. For this purpose, define

$$\begin{aligned}
H(\mathbf{x}^0, \boldsymbol{\epsilon}^0) := \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}} \in X^\infty, \bar{\mathbf{x}} \in L(f(\mathbf{x}^0); f)^\infty, \\
\bar{\mathbf{x}}_i \in L(c_i(\mathbf{x}_i^0) + \epsilon_i^0; c_i)^\infty, i \in \mathcal{G}\}, \textit{ and}
\end{aligned}$$

$$\tilde{H}(\mathbf{x}^0, \boldsymbol{\epsilon}^0) := X^\infty \cap L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); f)^\infty$$
$$\cap (\prod_{i \in \mathcal{G}} L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0) - \underline{f}; r_i \circ c_i)^\infty).$$

We now prove the following result about the lower level sets of $L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0), J_0)$.

Theorem 5 *Let* $\mathbf{x}^0 \in X$ *and* $\boldsymbol{\epsilon}^0 \in \mathbb{R}^m_{++}$. *Then*

$$L(r_i(c_i(\mathbf{x}_i^0) + \epsilon_i^0); r_i \circ c_i) = L(c_i(\mathbf{x}_i^0) + \epsilon_i^0; c_i)$$

for $i \in \mathcal{G}$. *Moreover, it holds that*

$$\hat{H}(\mathbf{x}^0, \boldsymbol{\epsilon}^0) \subset L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)^\infty. \tag{11}$$

Furthermore, suppose $\underline{f} := \inf_{\mathbf{x} \in X} f(\mathbf{x}) > -\infty$. *Then*

$$L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)^\infty \subset \tilde{H}(\mathbf{x}^0, \boldsymbol{\epsilon}^0). \tag{12}$$

Proof The convexity of $L(\mathbf{x}_i^0; r_i(c_i(\cdot) + \epsilon_i^0))$ is by the fact that

$$\begin{aligned}
\mathbf{x}_i &\in L(r_i(c_i(\mathbf{x}_i^0) + \epsilon_i^0); r_i \circ c_i) \\
&\iff r_i(c_i(\mathbf{x}_i)) \le r_i(c_i(\mathbf{x}_i^0) + \epsilon_i^0) \\
&\iff c_i(\mathbf{x}_i) \le c_i(\mathbf{x}_i^0) + \epsilon_i^0 \\
&\iff \mathbf{x}_i \in L(c_i(\mathbf{x}_i^0) + \epsilon_i^0; c_i),
\end{aligned}$$

where the second equivalence is from the monotonic increasing property of r_i. Notice that $L(c_i(\mathbf{x}_i^0) + \epsilon_i^0; c_i)$ is convex.

Now we prove (11). Let $\mathbf{x} \in L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)$ and $\bar{\mathbf{x}}$ be an element of $\hat{H}(\mathbf{x}^0, \boldsymbol{\epsilon}^0)$.

$$\mathbf{x} + \lambda\bar{\mathbf{x}} \in X, \;\; \mathbf{x} + \lambda\bar{\mathbf{x}} \in L(f(\mathbf{x}^0); f)^\infty,$$

and

$$\mathbf{x}_i + \lambda\bar{\mathbf{x}}_i \in L(c_i(\mathbf{x}_i^0) + \epsilon_i^0; c_i)^\infty.$$

Therefore, it holds that

$$\begin{aligned}
J_0(\mathbf{x} + \lambda\bar{\mathbf{x}}) &= f(\mathbf{x} + \lambda\bar{\mathbf{x}}) + \sum_{i \in \mathcal{G}} r_i(c_i(\mathbf{x}_i + \lambda\bar{\mathbf{x}}_i)) \\
&\le f(\mathbf{x}^0) + \sum_{i \in \mathcal{G}} r_i(c_i(\mathbf{x}_i^0) + \epsilon_i^0) \\
&= J(\mathbf{x}^0; \boldsymbol{\epsilon}^0).
\end{aligned}$$

Consequently, $\bar{\mathbf{x}} \in L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)$, proving (11).

For (12), let $\bar{\mathbf{x}} \in L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)^\infty$. We need to show that $\bar{\mathbf{x}}$ is an element of $\tilde{H}(\mathbf{x}^0, \boldsymbol{\epsilon}^0)$. For this, we may as well assume that $\bar{\mathbf{x}} \neq \mathbf{0}$. By the fact that $L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)^\infty$, there exists $t^k \downarrow 0$ and $\{\mathbf{x}^k\} \subset X$ such that $J_0(\mathbf{x}^k) \le J(\mathbf{x}^0; \boldsymbol{\epsilon}^0)$ and $t^k\mathbf{x}^k \to \bar{\mathbf{x}}$. Consequently, $\bar{\mathbf{x}} \in X^\infty$. Hence

$$L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)^\infty \subset X^\infty. \tag{13}$$

On the other hand, let $\tilde{\mathbf{x}} \in L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)$. It then follows that

$$f(\tilde{\mathbf{x}}) = J_0(\tilde{\mathbf{x}}) - \sum_{i \in \mathcal{G}} r_i(c_i(\tilde{\mathbf{x}}_i)) \le J_0(\tilde{\mathbf{x}}) \le J(\mathbf{x}^0; \boldsymbol{\epsilon}^0),$$

where the first inequality is by the fact that $r_i \geq 0$. Consequently, $\tilde{\mathbf{x}} \in L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); f)$, implying $L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0) \subset L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); f)$. Hence

$$L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)^\infty \subset L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); f)^\infty. \tag{14}$$

Now consider c_i. We have for $i \in \mathcal{G}$

$$r_i \circ c_i(\tilde{\mathbf{x}}_i) = J_0(\tilde{\mathbf{x}}) - f(\tilde{\mathbf{x}}) - \sum_{j \in \mathcal{G}, j \neq i} r_i(c_i(\tilde{\mathbf{x}}_i)) \leq J(\mathbf{x}^0; \boldsymbol{\epsilon}^0) - \underline{f},$$

implying $\tilde{\mathbf{x}}_i \in L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); r_i \circ c_i)$, $i \in \mathcal{G}$. Therefore,

$$L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0) \subset \prod_{i \in \mathcal{G}} L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0) - \underline{f}; r_i \circ c_i),$$

This implies that

$$\begin{aligned} L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)^\infty &\subset (\prod_{i \in \mathcal{G}} L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0) - \underline{f}; r_i \circ c_i))^\infty \\ &= \prod_{i \in \mathcal{G}} L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0) - \underline{f}; r_i \circ c_i)^\infty, \end{aligned}$$

which, combined with (13) and (14), yields (12). □

The following results follow directly from Theorem 5.

Corollary 1 *If there exists* $\bar{\mathbf{x}} \neq 0$ *such that*

$$\bar{\mathbf{x}} \in X^\infty, \ \bar{\mathbf{x}} \in L(f(\mathbf{x}^0); f)^\infty, \ \bar{\mathbf{x}}_i \in L(c_i(\mathbf{x}_i^0) + \epsilon_i^0; c_i)^\infty, i \in \mathcal{G},$$

then $L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)$ *is unbounded. Conversely, if one of the sets*

$$X^\infty, \ L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); f)^\infty, \ \textit{and} \ (\prod_{i \in \mathcal{G}} L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0) - \underline{f}; r_i \circ c_i)^\infty)$$

is empty, then $L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)$ *is bounded.*

Based on Corollary 1, we provide specific cases in the following proposition that can guarantee the boundedness of $L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); J_0)$.

Proposition 3 *Suppose* $\mathbf{x}^0 \in X$ *and relaxation vector* $\boldsymbol{\epsilon}^0 \in \mathbb{R}^m_{++}$. *Then the level set* $L(J(\mathbf{x}^0, \boldsymbol{\epsilon}^0), J_0)$ *is bounded, if one of the following conditions holds true*

(i) X *is compact.*
(ii) f *is coercive.*
(iii) $r_i \circ c_i$, $i \in \mathcal{G}$ *are all coercive.*
(iv) Assume

$$\gamma_i := \sup_{\|\mathbf{x}_i\| \to \infty} r_i(c_i(\mathbf{x}_i)) < +\infty, \ i \in \mathcal{G}.$$

Suppose $(\mathbf{x}^0, \boldsymbol{\epsilon}^0)$ *is selected to satisfy* $\sum_{i \in \mathcal{G}} r_i(c_i(\mathbf{x}_i^0) + \epsilon_i^0) \leq \underline{f} + \min_i \gamma_i$.

Proof Part (i)–(iii) are trivial true by Corollary 1. We only prove part (iv).

Assume by contradiction that $L(J(\mathbf{x}^0;\boldsymbol{\epsilon}^0);J_0)$ is unbounded, then there exists $\bar{\mathbf{x}} \in L(J(\mathbf{x}^0;\boldsymbol{\epsilon}^0);J_0)^\infty$ with $\bar{\mathbf{x}} \neq 0$. By the definition of horizon cone, there exists $\{t^k\} \subset \mathbb{R}$ and $\{\mathbf{x}^k\} \subset X$ such that

$$t^k \downarrow 0,\ J_0(\mathbf{x}^k) \le J(\mathbf{x}^0;\boldsymbol{\epsilon}^0),\ \text{ and } t^k\mathbf{x}^k \to \bar{\mathbf{x}}.$$

Therefore, there must be an $\bar{i} \in \mathcal{G}$, such that $\|\mathbf{x}^k_{\bar{i}}\|_2 \to \infty$, implying $r_i \circ c_i(\mathbf{x}^k_{\bar{i}}) \to \gamma_{\bar{i}}$. This means,

$$\begin{aligned} J(\mathbf{x}^0;\boldsymbol{\epsilon}^0) &\ge \lim_{k\to\infty} J_0(\mathbf{x}^k) \ge \underline{f} + \lim_{k\to\infty} r_i \circ c_i(\mathbf{x}^k_{\bar{i}}) \\ &= \underline{f} + \gamma_{\bar{i}} \ge \underline{f} + \min_{i\in\mathcal{G}} \gamma_i, \end{aligned}$$

a contradiction. Therefore, $L(J(\mathbf{x}^0,\boldsymbol{\epsilon}^0),J_0)$ is bounded. □

Proposition 3(iv) indicates that the initial iterate $\mathbf{x}^0$ and $\boldsymbol{\epsilon}^0$ may need to be chosen sufficiently close to 0 to enforce convergence if ϕ_i is not coercive such as (FRA).

5 Numerical experiments

In this section, we test our proposed AIR algorithm for nonconvex and nonsmooth sparse optimization problems in two numerical experiments and exhibit its performance. In both experiments, the test problems have $f(\mathbf{x}) \equiv 0$. The algorithm is implemented in Matlab with the subproblems solved by the CVX solver [14]. We consider two ways of choosing r_i and c_i, $c_i(\mathbf{x}_i) = \|\mathbf{x}_i\|_1$ and $c_i = \|\mathbf{x}_i\|_2^2$, as described in Table 1, so that they can be referred as ℓ_1-algorithm and ℓ_2-algorithm, respectively. In the subproblem, we use identical value for each component of the relaxation parameter $\boldsymbol{\epsilon}^k$, i.e., $\boldsymbol{\epsilon}^k = \epsilon^k\mathbf{e}$. In following two experiments, we define *sparsity* as the nonzeros of the vectors.

5.1 Sparse signal recovery

In this subsection, we consider a sparse signal recovery problem [8], which aims to recover sparse vectors from linear measurements. This problem can be formulated as

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n}\ & \|\mathbf{x}\|_p^p \\ \text{s.t.}\ & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{aligned} \tag{15}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the measurement matrix, $\mathbf{b} \in \mathbb{R}^{m\times 1}$ is the measurement vector and $p \in (0,1)$.

In the numerical experiments, we fix $n = 256$ and the measurement numbers $m = 100$. Draw the measurement matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with entries normally distributed. Denote s as the number of nonzero entries of $\mathbf{x}_0$, and set $s = 40$. We repeat the following procedure 100 times:

(i) Construct $\mathbf{x}_0 \in \mathbb{R}^{n\times 1}$ with randomly zeroing $n - s$ components. Each nonzero entries is chosen randomly from a zero-mean unit-variance Gaussian distribution.
(ii) Form $\mathbf{b}_0 = \mathbf{A}\mathbf{x}_0$.
(iii) Solve the problem for $\hat{\mathbf{x}}_0$.

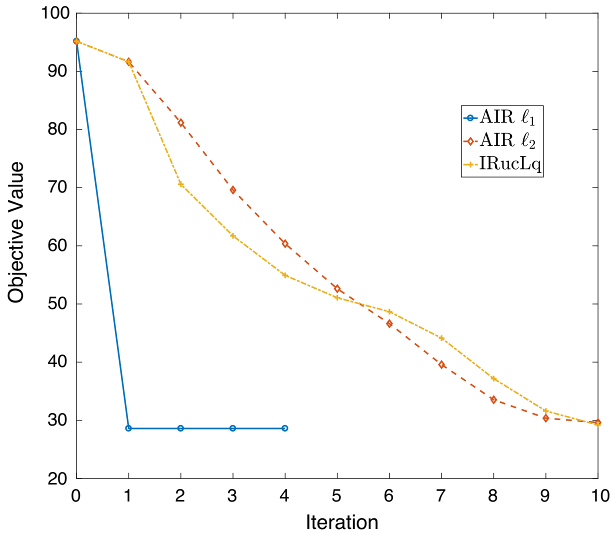

Fig. 1 Objective value versus iteration

We compare our AIR algorithms with the iterative reweighted unconstrained ℓ_q for sparse vector recovery (IRucLq-v) algorithm [44]. The IRucLq-v algorithm penalizes the linear constraint with a fixed parameter λ, yielding an unconstrained problem. Then, it uses a reweighted least square method to approximately solve the unconstrained problem, of which the subproblem can be addressed by solving a linear system. We set $\lambda = 10^{-6}$, initialize $\epsilon^0 = 1$ and $\mathbf{x}^0 = \mathbf{0}$. Update $\epsilon^{k+1} = \min\{\epsilon^k, \alpha \cdot r(\mathbf{x}^{k+1})_{s+1}\}$, where $r(\mathbf{x}^{k+1})_{s+1}$ denotes the $s+1$ largest (in absolute value) component of $\mathbf{x}^{k+1}$. Set $s = \lfloor m/2 \rfloor$ and $\alpha = 0.9$.

For our AIR algorithms, at each iteration, the subproblem of AIR ℓ_1-algorithm can be equivalently formulated as a Linear Programming (LP) problem; the subproblem of AIR ℓ_2-algorithm is a Quadratic Programming (QP) problem. We initialize $\epsilon^0 = 1$ and $\mathbf{x}^0 = \mathbf{0}$. Update $\epsilon^{k+1} = 0.7\epsilon^k$ for AIR ℓ_1-algorithm, and $\epsilon^{k+1} = 0.9\epsilon^k$ for AIR ℓ_2-algorithm. We terminate our AIR ℓ_1-algorithm whenever $\frac{\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_2}{\|\mathbf{x}^k\|_2} \le 10^{-8}$ and $\epsilon^{k+1} \le 10^{-3}$ are satisfied and record the final objective value as $f(\mathbf{x}_{\ell_1})$. The AIR ℓ_2-algorithm and the IRucLq-v algorithm are terminated when $\frac{|f(\mathbf{x}^k)-f(\mathbf{x}_{\ell_1})|}{f(\mathbf{x}_{\ell_1})} \le 10^{-3}$ or $k \ge 500$.

We first use one typical realization of the simulation to examine the convergence of all algorithms. We solve the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ to get the same feasible initial point for all algorithms. The experimental result is shown in Fig. 1. From the result, we observe that the AIR ℓ_1-algorithm converges faster than the other algorithms, and AIR ℓ_2-algorithm and IRucLq-v algorithm own similar convergence rates.

Then, we further investigate the properties of the solutions generated by all these algorithms. We select the threshold in different levels as

$$\nu \in \{10^{-3}, 10^{-6}, 10^{-9}, 10^{-12}\},$$

and set $x_i = 0$ if $|x_i| <= \nu,\ i = 1, \cdots, n$ for each ν. The box plots in Fig. 2 demonstrate the statistical properties of the number of nonzeros of all algorithms versus different thresholds.

The corresponding computation time and constraint violation results are shown in Table 2.

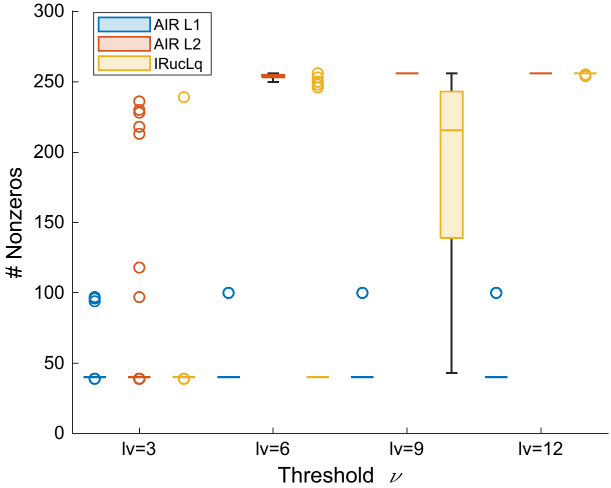

Fig. 2 Empirical success probability versus sparsity

Table 2 Average runtime and constraint violation results with associated standard deviation

	Method		
	AIR ℓ_1	AIR ℓ_2	IRucLq-v
Time (s)	0.289 ± 0.035	1.672 ± 0.187	0.013 ± 0.022
$\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\|_2$	$(9.18 \pm 7.31) \times$ 1e-13	$(5.09 \pm 0.79) \times$ 1e-14	$(1.76 \pm 0.67) \times$ 1e-5

We have the following observations from Fig. 2 and Table 2.

(i) From Fig. 2, we see that the AIR ℓ_1-algorithm outputs the most sparse solution. Furthermore, it has the most robust solution with respect to different thresholds. The IRucLq-v algorithm outperforms the AIR ℓ_2-algorithm both in sparsity and robustness.

(ii) From Table 2, it shows that the IRucLq-v algorithm is more efficient than the AIR algorithms, since it only needs to solve a linear system for each subproblem rather than a LP or QP problem. However, the cost of efficiency is sacrificing the feasibility, which is observed by the constraint violation results.

5.2 Group Sparse Optimization

In the second experiment, we consider the cloud radio access network power consumption problem [38]. In order to solve this problem, a three-stage group sparse beamforming method (GSBF) is proposed in [38]. The GSBF method solves a group sparse problem in the first stage to induce the group sparsity for the beamformers to guide the remote radio head (RRH) selection. This group sparse problem is addressed by minimizing the mixed ℓ_1/ℓ_2-norm. For further promoting the group sparsity, we replace the ℓ_1/ℓ_2-norm by the ℓ_p/ℓ_2 quasi-norm (LPN) with $p \in (0, 1)$ [36], yielding the following problem

$$\begin{aligned} \min_{\mathbf{v}} \quad & \sum_{l=1}^{L} \sqrt{\frac{\rho_l}{\eta_l}} \|\tilde{\mathbf{v}}_l\|_2^p \\ \text{s.t.} \quad & \sqrt{\sum_{i \neq k} \|h_k^H \mathbf{v}_i\|_2^2 + \sigma_k^2} \leq \frac{1}{\gamma_k} \Re(h_k^H \mathbf{v}_k) \\ & \|\tilde{\mathbf{v}}_l\|_2 \leq \sqrt{P_l},\ l = 1, \cdots, L, k = 1, \ldots, K. \end{aligned} \tag{16}$$

We consider the Cloud-RAN architecture with L remote radio heads (RRHs) and K single-antenna Mobile Users (MUs), where the l-th RRH is equipped with N_l antennas. $\mathbf{v}_{lk} \in \mathbb{C}^{N_l}$ is the transmit beamforming vector from the l-th RRH to the k-th user with the group structure of transmit vectors $\tilde{\mathbf{v}}_l = [\mathbf{v}_{l1}^T, \cdots, \mathbf{v}_{lK}^T]^T \in \mathbb{C}^{KN_l \times 1}$. Denote the relative fronthaul link power consumption by ρ_l, and the inefficient of drain efficiency of the radio frequency power amplifier by η_l. The channel propagation between user k and RRH l is denoted as $\mathbf{h}_{lk} \in \mathbb{C}^{N_l}$. P_l is the maximum transmit power of the l-th RRH. σ_k is the noise at MU k. $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_K)^T$ is the target signal-to-interference-plus-noise ratio (SINR).

5.2.1 Comparison with the mixed ℓ_1/ℓ_2 algorithm

In this experiment, we compare our AIR ℓ_1- and ℓ_2-algorithms with the mixed ℓ_1/ℓ_2 algorithm [38]. We consider a network with $L = 10$ 2-antenna RRHs (i.e., $N_l = 2$), and $K = 6$ single-antenna MUs uniformly and independently distributed in the square region $[-1000,\ 1000] \times [-1000,\ 1000]$ meters. We set $P_l = 1$, $\rho_l = 13$, $\eta_l = \frac{1}{4}$ for $l \in \{1, \cdots, L\}$, $\sigma_k = 1$ for $k \in \{1, \ldots, K\}$. For each quality of service (QoS) q in $\{0, 2, 4, 6\}$, we set the target SINR $\gamma_k = 10^{q/10}$ for $k = 1, \ldots, K$. Repeat the following procedure 50 times:

(i) Randomly generated network realizations for the channel propagation $\mathbf{h}_{lk}$, $l \in \{1, \cdots, L\}$, $k \in \{1, \cdots, K\}$.
(ii) Adopt AIR ℓ_1- and ℓ_2-algorithm to solve (16) for $\tilde{\mathbf{v}}^*$.
(iii) Regard the l-th component of $\tilde{\mathbf{v}}^*$ as zero, if $\|\tilde{\mathbf{v}}_l^*\| \leq 10^{-3}$ for $l \in \{1, \cdots, L\}$.

We set the maximum number of iterations as $T = 500$, $\epsilon_i^0 = 1$ for AIR and update by $\epsilon^{k+1} = 0.7\epsilon^k$ at each iteration with minimum threshold 10^{-6}. Set $p = 0.1$. The algorithm is terminated whenever $|f(\mathbf{v}^{k+1}) - f(\mathbf{v}^k)| \leq 10^{-4}$ is satisfied or $k \geq T$.

In Fig. 3, we depict the sparsity of the final solution returned by mixed ℓ_1/ℓ_2 algorithm, AIR ℓ_1- and ℓ_2-algorithms for problems with different SINR. The proposed AIR ℓ_1- and ℓ_2-algorithms outperform the mixed ℓ_1/ℓ_2 algorithm in promoting the group sparsity. And it is witnessed again that the AIR ℓ_1-algorithm outperforms AIR ℓ_2-algorithm in the ability of accurately recovering sparse solutions.

5.2.2 Comparison with the difference of convex algorithm

We consider the same group sparse optimization problem (16) in this section, and compare our AIR ℓ_1- and ℓ_2-algorithms with the SDCAM [45]. We defer the details of the SDCAM for solving problem (16) to the appendix.

In this experiment, we consider a larger-sized network with $L = 20$ 2-antenna RRHs (i.e., $N_l = 2$), and $K = 15$ single-antenna MUs uniformly and independently distributed in the square region $[-2000,\ 2000] \times [-2000,\ 2000]$ meters. We set $P_l = 1$, $\rho_l = 20$,

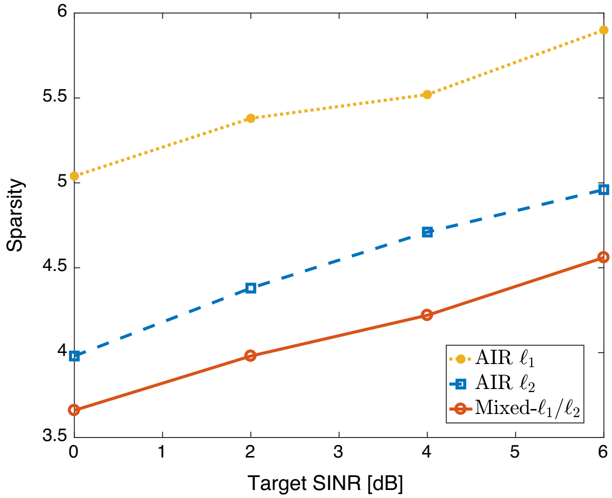

Fig. 3 Average sparsity versus target SINR

$\eta_l = \frac{1}{4}$ for $l \in \{1, \cdots, L\}$, $\sigma_k = 1$ for $k \in \{1, \ldots, K\}$. For each quality of service (QoS) q in $\{0, 2, 4, 6\}$, we set the target SINR $\gamma_k = 10^{q/10}$ for $k = 1, \ldots, K$. Randomly generated network realizations for the channel propagation $\mathbf{h}_{lk}$, $l \in \{1, \cdots, L\}$, $k \in \{1, \cdots, K\}$.

For the AIR algorithms, we set the maximum number of iterations as $T = 500$, $\epsilon_i^0 = 100$ for AIR and update by $\epsilon^{k+1} = 0.5\epsilon^k$ at each iteration with minimum threshold 10^{-6}. Set $p = 0.1$ (and $p = 0.5$). For the case $p = 0.1$, the AIR algorithms are terminated whenever both $\frac{\|\mathbf{v}^{k+1} - \mathbf{v}^k\|_2}{\|\mathbf{v}^k\|_2} \leq 10^{-5}$ and $\epsilon^{k+1} \leq 10^{-3}$ are satisfied or $k \geq T$. For the case $p = 0.5$, the AIR ℓ_1-algorithm is terminated whenever $\frac{\|\mathbf{v}^{k+1} - \mathbf{v}^k\|_2}{\|\mathbf{v}^k\|_2} \leq 10^{-5}$ and $\epsilon^{k+1} \leq 10^{-3}$ are satisfied or $k \geq T$. We denote the solution as $\mathbf{v}_{\ell_1}$ and record the final objective value as $f(\mathbf{v}_{\ell_1})$. We terminate the AIR ℓ_2-algorithm whenever $f(\mathbf{v}^{k+1}) \leq f(\mathbf{v}_{\ell_1})$ or $k \geq T$.

The SDCAM applies the Moreau envelope to approximate problem (16), which yields a sequence of DC subproblems. They solves the DC subproblems by the Nonmonotone Proximal Gradient method with majorization ($\text{NPG}_{\text{major}}$). In SDCAM, we set $\lambda_k = \max\{1/10^{k+1}, 10^{-10}\}$ and $\mathbf{v}^{\text{feas}}$ to be the vector of all ones. In $\text{NPG}_{\text{major}}$, we set $M = 4$, $L_{\max} = 10^8$, $L_{\min} = 10^{-8}$, $\tau = 2$, $c = 10^{-5}$, $L_{k,0} = 1$ and for $t \geq 1$,

$$L_{k,t} = \max\left\{\min\left\{\frac{{\mathbf{s}^t}^T \mathbf{y}^t}{\|\mathbf{s}^t\|_2^2}, L_{\max}\right\}, L_{\min}\right\},$$

l where $\mathbf{s}^t = \mathbf{v}^t - \mathbf{v}^{t-1}$, $\mathbf{y}^t = \nabla h(\mathbf{v}^t) - \nabla h(\mathbf{v}^{t-1})$. We terminate $\text{NPG}_{\text{major}}$ when

$$\frac{\|\mathbf{s}^t\|_2}{\max(\|\mathbf{v}^t\|_2, 1)} < \epsilon_k \text{ or } \frac{\|F_{\lambda_k}(\mathbf{v}^t) - F_{\lambda_k}(\mathbf{v}^{t-1})\|_2}{\max(|F_{\lambda_k}(\mathbf{v}^t)|, 1)} < 10^{-6},$$

where $\epsilon_0 = 10^{-3}$ and $\epsilon_k = \max\{\epsilon_{k-1}/1.5, 10^{-5}\}$. We terminate the SDCAM whenever $f(\mathbf{v}^{k+1}) \leq f(\mathbf{v}_{\ell_1})$ or $k \geq 1000$.

First, we explore the convergence rates of SDCAM and our AIR algorithms with $p = 0.5$. We demonstrate the results of objective values versus CPU time for all algorithms in Fig. 4,

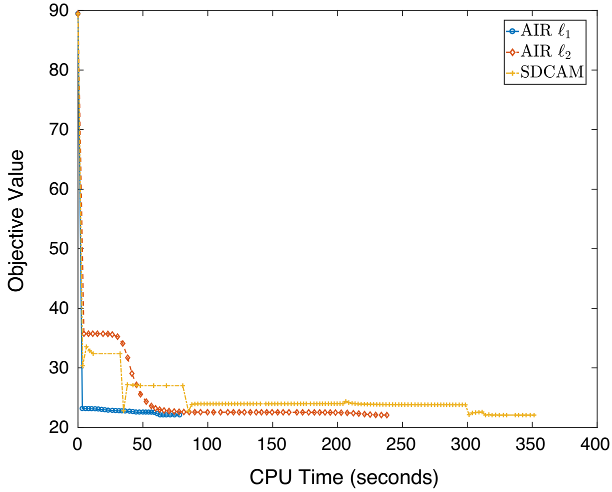

Fig. 4 Objective value versus CPU time

where we only report one typical channel realization. The results show that AIR algorithms converge faster than SDCAM, especially the AIR ℓ_1-algorithm.

Then, we further investigate the properties of the solutions generated by all these algorithms. For each quality of service (QoS) q in $\{0, 2, 4, 6\}$, we randomly generate the network realizations 10 times. We select the threshold in different levels as

$$\nu \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\},$$

and set $\|\tilde{\mathbf{v}}_l\|_2 = 0$ if $\|\tilde{\mathbf{v}}_l\|_2 <= \nu,\ l = 1, \cdots, L$. The box plots in Fig. 5 demonstrate the distributions of the number of nonzeros of all algorithms versus different Target SINR. The corresponding computation time results are shown in Table 3.

We have the following observations from Fig. 5 and Table 3.

(i) From Fi. 5a, we see that the AIR algorithms with $p = 0.1$, especially the AIR ℓ_1-algorithm, has more sparse solution than the SDCAM. Furthermore, it shows that the solutions generated by our AIR algorithms are more robust with respect to different sparse levels.

(ii) From Fig. 5b, we see that the AIR ℓ_1-algorithm with $p = 0.5$ and the SDCAM perform similarly in this case. The AIR ℓ_1-algorithm with $p = 0.5$ has the best performance both in terms of the sparsity of the solution as well as the robustness of the solution in different levels.

(iii) From Table 3, it shows that AIR ℓ_1-algorithm converges faster than the AIR ℓ_2-algorithm and the SDCAM. Moreover, the SDCAM fails to achieve the target objective twice for the cases $q = 0$, $q = 2$ and $q = 4$.

6 Conclusions

In this paper, we consider solving a general formulation for nonconvex and nonsmooth sparse optimization problem, which can take into account different sparsity-inducing terms. An iteratively reweighted algorithmic framework is proposed by solving a sequence of weighted

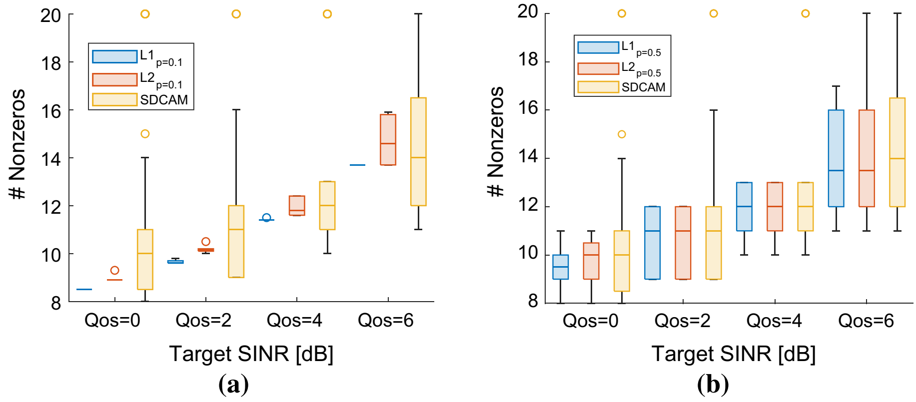

Fig. 5 Comparing SDCAM with AIR algorithms by choosing (5a) $p = 0.1$ and (5b) $p = 0.5$

Table 3 Average runtime with associated standard deviation for SDCAM and AIR algorithms (with $p = 0.5$). The number in parentheses is the number of the cases that the corresponding algorithms fail to achieve the target objective. Note that for the case SDCAM fails to achieve the target objective, we do not take the corresponding time into the statistical results

Qos	Method		
	AIR ℓ_1	AIR ℓ_2	SDCAM
0	106.651±64.525 (0)	227.949±82.590 (0)	219.306±78.987 (2)
2	73.070±51.722 (0)	124.946±41.412 (0)	168.794±70.370 (2)
4	74.113±40.362 (0)	206.876±116.152 (0)	214.730±81.465 (2)
6	52.764±11.141 (0)	149.554±70.900 (0)	150.455±109.333 (0)

convex regularization subproblems. We have also derived the optimality condition for the nonconvex and nonsmooth sparse optimization problem and provided the global convergence analysis for the proposed iteratively reweighted methods.

Two variants of our proposed algorithm, the ℓ_1-algorithm and the ℓ_2-algorithm, are implemented and tested. Numerical results exhibits their ability of recovering sparse signals. It is also witnessed that the iteratively ℓ_1-algorithm is generally faster than the ℓ_2-algorithm because much fewer iterations are needed for ℓ_1-algorithm. Overall, our investigation leads to a variety of interesting research directions:

- A thorough comparison, through either theoretical analysis or numerical experiments, of the existing nonconvex and nonsmooth sparse optimization problems using AIR would be interesting to see. This should be helpful in providing the guidance for the users to select sparsity-inducing functions.
- Our implementation reduces the relaxation parameter ϵ by a fraction each time. It would be useful if a dynamic updating strategy can be derived to reduce the efforts of parameter tuning as well as the sensitivity of the algorithm to ϵ.
- It would be meaningful to have a (local) complexity analysis for the reweigh-ted algorithms.

Acknowledgements Hao Wang's work was supported by National Natural Science Foundation of China [12001367] and the Natural Science Foundation of Shanghai [21ZR1442800]. Yaohua Hu's work was supported in part by the National Natural Science Foundation of China [12071306, 11871347], Natural Science

Foundation of Guangdong Province of China [2019A1515011917, 2020B1515310008, 2020A1515010372], and Natural Science Foundation of Shenzhen [JCYJ20190808173603590].

A implementation details of SDCAM

In this section, we provide the details of solving problem (16) by using SDCAM. By denoting

$$\Omega := \left\{ \mathbf{v} : \sqrt{\sum_{i \neq k} \|\mathbf{h}_k^H \mathbf{v}_i\|_2^2 + \sigma_k^2} \leq \frac{1}{\gamma_k} \Re(\mathbf{h}_k^H \mathbf{v}_k),\ \|\tilde{\mathbf{v}}_l\|_2 \leq \sqrt{P_l},\ l = 1, \cdots, L,\ k = 1, \cdots, K \right\},$$

and

$$P_1(\mathbf{v}) := \sum_{l=1}^{L} \sqrt{\frac{\rho_l}{\eta_l}} \|\tilde{\mathbf{v}}_l\|_2^p \quad \text{and} \quad P_0(\mathbf{v}) := \delta(\mathbf{v}|\Omega),$$

we can reformulate problem (16) as

$$\min_{\mathbf{v}} \quad F(\mathbf{v}) := P_0(\mathbf{v}) + P_1(\mathbf{v}).$$

Then this problem can be solved by the SDCAM [45]. They approximate F by its Moreau envelope at each iteration. More specifically, at iteration k, they solve the following approximate problem

$$\min_{\mathbf{v}} \quad F_{\lambda_k}(\mathbf{v}) := P_0(\mathbf{v}) + e_{\lambda_k} P_1(\mathbf{v}),$$

where $e_{\lambda_k} P_1(\mathbf{v})$ is the Moreau envelope of $P_1(\mathbf{v})$ with parameter λ_k, which takes the form

$$e_{\lambda_k} P_1(\mathbf{v}) =:= \inf_{\mathbf{x}} \{ \frac{1}{2\lambda_k} \|\mathbf{v} - \mathbf{y}\|_2^2 + P_1(\mathbf{y}) \}.$$

The SDCAM drives the parameter λ_k to 0 and solves each corresponding subproblem F_{λ_k} iteratively. By taking advantage of the equivalently formulation of $e_{\lambda_k} P_1(\mathbf{v})$, i.e.,

$$e_{\lambda_k} P_1(\mathbf{v}) = \frac{1}{2\lambda} \|\mathbf{v}\|_2^2 - \underbrace{\sup_{\mathbf{y} \in \text{dom } P_1} \{ \frac{1}{\lambda} \mathbf{v}^T \mathbf{y} - \frac{1}{2\lambda} \|\mathbf{y}\|_2^2 - P_1(\mathbf{y}) \}}_{D_{\lambda, P_1}(\mathbf{v})},$$

we can reformulate F_{λ_k} as a DC problem, which can be solved by the $\text{NPG}_{\text{major}}$ method. The $\text{NPG}_{\text{major}}$ method solves the DC subproblem by combing the proximal gradient method with the nonmonotone line search technique, and terminates when the first-order optimality condition is satisfied. Note that, for the $\text{NPG}_{\text{major}}$ method, we need to calculate the subgradient of $D_{\lambda, P_1}(\mathbf{v})$, which is proved equal to $\frac{1}{\mu} \text{prox}_{\lambda_k, P_1}(\mathbf{v})$. The proximity operator of ℓ_p norm with $p = 1/2$ (or $p = 2/3$) has an analytic solution, which is provided in [46].

References

1. Ba, D., Babadi, B., Purdon, P.L., Brown, E.N.: Convergence and stability of iteratively re-weighted least squares algorithms. IEEE Trans. Signal Process. **62**(1), 183–195 (2014)
2. Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al.: Optimization with sparsity-inducing penalties. Foundations and Trends®. Mach. Learn. **4**(1), 1–106 (2012)

3. Bazaraa, M.S., Goode, J., Nashed, M.Z.: On the cones of tangents with applications to mathematical programming. J. Optim. Theory Appl. **13**(4), 389–426 (1974)
4. Bian, W., Chen, X.: Optimality and complexity for constrained optimization problems with nonconvex regularization. Math. Oper. Res. **42**(4), 1063–1084 (2017)
5. Bradley, P.S., Mangasarian, O.L., Street, W.N.: Feature selection via mathematical programming. INFORMS J. Comput. **10**(2), 209–217 (1998)
6. Burke, J.V., Curtis, F.E., Wang, H., Wang, J.: Iterative reweighted linear least squares for exact penalty subproblems on product sets. SIAM J. Optim. **25**(1), 261–294 (2015)
7. Candès, E.J., Wakin, M.B.: An introduction to compressive sampling. IEEE Signal Process. Mag. **25**(2), 21–30 (2008)
8. Candes, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted 1 minimization. J. Fourier Anal. Appl. **14**(5–6), 877–905 (2008)
9. Chartrand, R., Yin, W.: Iteratively reweighted algorithms for compressive sensing. In: Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on, pp. 3869–3872. IEEE (2008)
10. Chen, X., Zhou, W.: Convergence of Reweighted l1 Minimization Algorithms and Unique Solution of Truncated lp Minimization. The Hong Kong Polytechnic University, Department of Applied Mathematics (2010)
11. Daubechies, I., DeVore, R., Fornasier, M., Güntürk, C.S.: Iteratively reweighted least squares minimization for sparse recovery. Commun. Pure Appl. Math. **63**(1), 1–38 (2010)
12. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
13. Fazel, M., Hindi, H., Boyd, S.P.: Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In: American control conference, 2003. Proceedings of the 2003, vol. 3, pp. 2156–2162. IEEE (2003)
14. Grant, M., Boyd, S., Ye, Y.: CVX: Matlab software for disciplined convex programming (2008). (Web page and software.) http://stanford.edu/~boyd/cvx (2015)
15. Harrell, F.E.: Ordinal logistic regression. In: Regression modeling strategies, pp. 311–325. Springer (2015)
16. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton (2015)
17. Holland, P.W., Welsch, R.E.: Robust regression using iteratively reweighted least-squares. Commun. Stat. Theory Methods **6**(9), 813–827 (1977)
18. Hu, Y., Li, C., Meng, K., Qin, J., Yang, X.: Group sparse optimization via $\ell_{p,q}$ regularization. J. Mach. Learn. Res. **18**(1), 960–1011 (2017)
19. Khajehnejad, M.A., Dimakis, A.G., Xu, W., Hassibi, B.: Sparse recovery of nonnegative signals with minimal expansion. IEEE Trans. Signal Process. **59**(1), 196–208 (2011)
20. Kruger, A.Y.: ε-semidifferentials and ε-normal elements. Depon. VINITI **1331**,(1981)
21. Kruger, A.Y.: On fréchet subdifferentials. J. Math. Sci. **116**(3), 3325–3358 (2003)
22. Lai, M.J., Wang, J.: An unconstrained ℓ_q minimization with $0 \leq q \leq 1$ for sparse solution of underdetermined linear systems. SIAM J. Optim. **21**(1), 82–101 (2011)
23. Lanza, A., Morigi, S., Sgallari, F.: Convex image denoising via non-convex regularization. In: International Conference on Scale Space and Variational Methods in Computer Vision. Springer , pp. 666–677 (2015)
24. Ling, Q., Wen, Z., Yin, W.: Decentralized jointly sparse optimization by reweighted ℓ_q minimization. IEEE Trans. Signal Process. **61**(5), 1165–1170 (2013)
25. Liu, L., Larsson, E.G., Yu, W., Popovski, P., Stefanovic, C., de Carvalho, E.: Sparse signal processing for grant-free massive connectivity: a future paradigm for random access protocols in the internet of things. IEEE Signal Process. Mag. **35**(5), 88–99 (2018)
26. Lobo, M.S., Fazel, M., Boyd, S.: Portfolio optimization with linear and fixed transaction costs. Ann. Oper. Res. **152**(1), 341–365 (2007)
27. Lu, C., Wei, Y., Lin, Z., Yan, S.: Proximal iteratively reweighted algorithm with multiple splitting for nonconvex sparsity optimization. In: AAAI, pp. 1251–1257 (2014)
28. Lu, Z.: Iterative reweighted minimization methods for ℓ_p regularized unconstrained nonlinear programming. Math. Program. **147**(1–2), 277–307 (2014)
29. Lu, Z., Zhang, Y., Lu, J.: ℓ_p regularized low-rank approximation via iterative reweighted singular value minimization. Comput. Optim. Appl. **68**(3), 619–642 (2017)
30. Malek-Mohammadi, M., Babaie-Zadeh, M., Skoglund, M.: Iterative concave rank approximation for recovering low-rank matrices. IEEE Trans. Signal Process. **62**(20), 5213–5226 (2014)
31. Miosso, C.J., von Borries, R., Argaez, M., Velázquez, L., Quintero, C., Potes, C.: Compressive sensing reconstruction with prior information by iteratively reweighted least-squares. IEEE Trans. Signal Process. **57**(6), 2424–2431 (2009)

32. Ochs, P., Dosovitskiy, A., Brox, T., Pock, T.: On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. SIAM J. Imaging Sci. **8**(1), 331–372 (2015)
33. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables. Siam, New Delhi (1970)
34. Qin, Z., Fan, J., Liu, Y., Gao, Y., Li, G.Y.: Sparse representation for wireless communications: a compressive sensing approach. IEEE Signal Process. Mag. **35**(3), 40–58 (2018)
35. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis, vol. 317. Springer Science & Business Media, New York (2009)
36. Shi, Y., Cheng, J., Zhang, J., Bai, B., Chen, W., Letaief, K.B.: Smoothed l_p-minimization for green cloud-ran with user admission control. IEEE J. Sel. Areas Commun. **34**(4), 1022–1036 (2016)
37. Shi, Y., Zhang, J., Chen, W., Letaief, K.B.: Generalized sparse and low-rank optimization for ultra-dense networks. IEEE Commun. Mag. **56**(6), 42–48 (2018)
38. Shi, Y., Zhang, J., Letaief, K.B.: Group sparse beamforming for green cloud-ran. IEEE Trans. Wirel. Commun. **13**(5), 2809–2823 (2014)
39. Street, J.O., Carroll, R.J., Ruppert, D.: A note on computing robust regression estimates via iteratively reweighted least squares. Am. Stat. **42**(2), 152–154 (1988)
40. Wainwright, M.J.: Structured regularizers for high-dimensional problems: statistical and computational issues. Annu. Rev. Stat. Appl. **1**, 233–253 (2014)
41. Wang, H., Li, D.H., Zhang, X.J., Wu, L.: Optimality conditions for the constrained ℓ_p-regularization. Optimization **64**(10), 2183–2197 (2015)
42. Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M.: Use of the zero-norm with linear models and kernel methods. J. Mach. Learn. Res. **3**, 1439–1461 (2003)
43. Zhang, C.H.: Penalized linear unbiased selection. In: Department of Statistics and Bioinformatics. Rutgers University, New Jersey (2007)
44. Lai, M.J., Xu, Y.Y., Yin, W.T.: Improved iteratively reweighted least squares for unconstrained smoothed *ell_q* minimization. SIAM J. Numer. Anal. **51**(2), 927–957 (2013)
45. Liu, T., Pong, T.K., Takeda, A.: A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems. Math. Program. **176**(1), 339–367 (2019)
46. Cao, W.F., Sun, J., Xu, Z.B.: Fast image deconvolution using closed-form thresholding formulas of L_q $(q = 1/2, 2/3)$ regularization. J. Vis. Commun. Image Represent. **24**(1), 31–41 (2013)
47. Ahn, M.J., Pang, J.S., Xin, J.: Difference-of-convex learning: directional stationarity, optimality, and sparsity. SIAM J. Optim. **27**(3), 1637–1665 (2017)
48. Cao, S.S., Huo, X.M., Pang, J.S.: A unifying framework of high-dimensional sparse estimation with difference-of-convex (DC). Regularizations **1812**(07130), 31–41 (2018). arXiv:1812.07130