# Nonconvex and Nonsmooth Sparse Optimization via Adaptively Iterative Reweighted Methods

Hao Wang Fan Zhang Qiong Wu Yaohua Hu Yuanming Shi

*Abstract*—**We propose a general formulation of nonconvex regularization problems with convex set constraint, which can take into account most existing types of nonconvex regularization terms, bringing strong applicability to a wide range of applications. We design an algorithmic framework of iteratively reweighted algorithms for solving the proposed nonconvex regularization problems, which solves a sequence of weighted convex regularization problems with iteratively updated weights. We also provide global convergence under loose assumptions. This makes our method a tool for a family of various reweighted algorithms. The effectiveness and efficiency of our proposed formulation and the algorithms are demonstrated in numerical experiments for various regularization problems.**

*Index Terms*—**nonconvex regularization, nonsmooth regularization, iteratively reweighted methods, nonconvex regularization**

## I. Introduction

The central focus of this paper is the solution of a wide class of nonconvex regularized optimization problems, which have been becoming a prevalent research topic in many disciplines of applied mathematics and engineering. Indeed, there has been a tremendous increase in the number of application areas in which nonconvex regularization algorithms have been employed, such as machine learning [1], [2], telecommunications [3], image reconstruction [4] and signal processing [5]. This is mainly because of its superior ability to reduce the complexity of a system, improve the generalization of the prediction performance, or enhance the robustness of the solution, compared with traditional convex regularization techniques.

Despite their wide application, nonconvex regularization problems are computationally difficult to solve in most cases due to the nonconvex and nonsmooth nature of the regularization terms. Iteratively reweighted method (IRWA) is one of the most popular methods for handling the convex/nonconvex regularization problems, which approximates the nonconvex regularization problem by a sequence of convex subproblems. Another issue caused by the nonconvex and nonsmooth regularization is the difficulty in characterizing the optimality condition as well as the convergence analysis, since the traditional analysis for smooth problems cannot be directly used. Therefore, many existing reweighted algorithms focus on showing the convergence to the optimal solution of the relaxed regularization problems. A critical aspect of any implementation of such an approach is the selection of the smoothing parameters. As been explained in [6], large relaxation parameters will smooth out many local minimizers, whereas small values can make the subproblems difficult to solve and the algorithm too quickly get trapped into local minimizers. Therefore, warm-start techniques are used to solve a sequence of subproblems with relaxation parameters driven from relatively large value to zero. Some other work focuses on developing dynamic relaxation parameter updating strategy [7] for convex regularization problems.

In this paper, we propose a general framework of Adaptively Iterative Reweighted (AIR) algorithm for solving the nonconvex and nonsmooth sparse optimization problems, along with an analysis of the convergence to the optimal solution of the original problem. The most related paper to our work may be the algorithm proposed in [7] for convex problems on convex set and [8], [9] for unconstrained $\ell_p$ regularization problems. This makes our work different from that in [7].

Overall, the contributions in this paper can be summarized as the following.

- To propose a general problem formulation that can take into account most existing types of nonconvex regularization terms. This formulation allows for different common regularization terms, group structure, as well as a general convex set constraint, leading to strong applicability to a wide range of applications.
- To develop a general algorithmic framework of iteratively reweighted algorithms, so that the nonconvex regularization problem can be attacked by solving a sequence of simple subproblems.
- First-order optimality condition for the regularization problem and convergence analysis of the proposed algorithms are also provided under loose assumptions, making our method a tool for a family of various reweighted algorithms.

### A. Organization

In the remainder of this section, we outline our notation and introduce various concepts that will be employed throughout the paper. In § II, we describe our problem of interests and explain its connection to various existing types of regularization techniques. In § III, we describe the detail of our proposed AIR and apply it to different types of nonconvex regularizers.

Sch. of Inf. Sci. and Tech., ShanghaiTech University; (wanghao1@shanghaitech.edu.cn)
Sch. of Inf. Sci. and Tech., ShanghaiTech University; (zhangfan4@shanghaitech.edu.cn)
Sch. of Inf. Sci. and Tech., ShanghaiTech University; (wuqiong@shanghaitech.edu.cn)
College of Mathematics and Statistics, Shenzhen University; (mayhhu@szu.edu.cn)
Sch. of Inf. Sci. and Tech., ShanghaiTech University; (shiym@shanghaitech.edu.cn)

The optimality condition and the global convergence of the proposed algorithm in different situations are provided in § IV. We discuss implementations of our methods and the results of numerical experiments in in § V. Concluding remarks are provided in § VI.

### *B. 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.

Let $\mathbb{R}^n$ be the space of real $n$-vectors, $\mathbb{R}^n_+$ be the nonnegative orthant of $\mathbb{R}^n$, $\mathbb{R}^n_+ = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0\}$ and the nonpositive orthant $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \leq 0\}$. Moreover, let $\mathbb{R}^n_{++}$ be its interior $\mathbb{R}^n_{++} := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} > 0\}$. 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

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

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

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

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

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

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

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

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 [10], [11], [12] and discussed in [13].

**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*

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

*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 $\partial_F r \circ c(\mathbf{x})$ as the Fréchet subdifferential of $r$ with respect to $\mathbf{x}$, $\partial_F r(c(\mathbf{x}))$ (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 $c(\mathbf{x})$.

## II. 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.

### *A. 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. Here $\Phi = r \circ c(x) = r(c(\mathbf{x}))$ is a nonconvex and nonsmooth composite function with $c$ convex and $r$ nonconvex. This type of problem is a staple for many applications in signal processing [14], [15], wireless communications [16], [17] and machine learning [18], [19]. For example, in signal processing, $f$ may be the mean-squared error for signal recovery, $X$ may be a nonnegative constraint for signal [20]; 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 [21]; in machine learning, $f$ can represent the convex loss function, such as the cross-entropy loss for logistic regression [22].

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 $\Phi$ has been studied in [23]. 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

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

With these ingredients, the associated group-based function $\Phi$ 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 and convex 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 with* $c_i(\mathbf{x}_i) \geq 0, \forall \mathbf{x} \in X$ *where the equality holds if and only if* $\mathbf{x}_i = 0$*.*

**Remark 1.** *The 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 $\ell_0$ norm of $\mathbf{x}$.

### *B. Sparsity-inducing Functions*

Many applications including signal processing, wireless communications and machine learning involve the minimization of the $\ell_0$-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 $\ell_0$ norm have been proposed. By different choice of the formulation $r_i$ and $c_i$, there exist many approximations to $\ell_0$ 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 $\mathbb{R}_-$.

The first instance is the feature selection algorithm via concave minimization proposed by Bradley and Mangasarian [24] with approximation

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

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 $\Phi$.

The second instance, which is widely used in many applications currently, is to approximate the $\ell_0$ norm by $\ell_p$ quasi-norm [25]

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

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 $\Phi$.

Another option for approximating $\ell_0$ norm, proposed in [26], is to use the log-sum approximation

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

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

$$c_i(x_i) = |x_i|, r_i(c_i) = \log\left(1 + c_i^p\right),$$

or

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

The approximation technique proposed in [25] 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

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

Candès et al. propose an approximation to the $\ell_0$ norm in [5]

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

and sufficiently small $p$ can cause sparsity in the solution. The function $\arctan$ is bounded above and $\ell_0$-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

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

or

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

Another nonconvex regularization technique needs to be mentioned is the SCAD penalty proposed in [27], which require the derivative of $\phi_i$ to satisfy

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

for some $a > 2$, where often $a = 3.7$ is used. Alternatively, the MCP [28] penalty uses

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

**Remark 2.** *These sparsity-inducing functions can also take into account group structures. For example, $\ell_{p,q}$-norm with $p \geq 1$ and $0 < q < 1$ 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*

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

### C. Problem Analysis

There have been various literatures for solving the nonconvex and nonsmooth sparse optimization problems. In [29], [30] Wotao Yin *et al*. have considered solve the sparse signal recovery problem by using the unconstrained nonconvex $\ell_p$ norm model, proposed the associated iterative reweighted unconstrained $\ell_p$ algorithm and provided the convergence analysis for the reweighted $\ell_2$ case. In [8] Zhaosong Lu have provied the first-order optimality condition for the unconstrained nonconvex $\ell_p$ norm problem, and convergence analysis for both $\ell_1$ and $\ell_2$ 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 § III, provide the first-order optimality condition for (1) and the convergence analysis for the AIR algorithm in § IV.

## III. Adaptively Iterative Reweighted Algorithm

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

### A. 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, $J(\mathbf{x}; 0) = J_0(\mathbf{x})$. Here we extend the notation of $\phi_i$ and use $\phi_i(\mathbf{x}_i; \epsilon_i)$ to denote the relaxed regularization function, so that

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

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

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

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

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

*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

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

Therefore,

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

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

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

□

### B. 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 $\boldsymbol{\epsilon} = 0$ 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 $\boldsymbol{\epsilon}^{k+1} \leq \boldsymbol{\epsilon}^k$ 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
       $\mathbf{x}^{k+1} \in \arg\min_{\mathbf{x} \in X} G_{(\mathbf{x}^k, \boldsymbol{\epsilon}^k)}(\mathbf{x})$.
3: (Reweighting) Choose $\boldsymbol{\epsilon}^{k+1} \in (0, \boldsymbol{\epsilon}^k]$.
4: Set $k \leftarrow k + 1$. Go to Step 2.
```

### C. Iterative Reweighter $\ell_1$ Algorithm & $\ell_2$ 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 sparse-inducing functions (EXP)–(MCP) in § II. Notice that the relaxation parameter $\epsilon_i$ could set as 0 if $\lim_{c_i \to 0+} r_i'(c_i) < +\infty$. For simplicity, denote $\tilde{w_i} = w_i(\tilde{x}_i, \tilde{\epsilon}_i)$. In Table I, we provide the explicit forms of the weights $\tilde{w}_i$ at $(\tilde{x}_i, \tilde{\epsilon}_i)$ when choosing $c_i(x_i) = |x_i|_1$ and $c_i(x_i) = x_i^2$ for each case, so that the corresponding subproblem is an $\ell_1$-norm sparse-inducing problem and an $\ell_2$-norm sparse-inducing problem

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

For each regularizer, we consider $c_i(x_i) = |x_i|$ in the first row and $c_i(x_i) = x_i^2$ 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.

TABLE I: Different AIR weights based on different choice of $r_i$ and $c_i$.

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

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 $\ell_1$ subproblem for SCAD has weights

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

The weights of reweighted $\ell_2$ subproblem for SCAD are

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

As for MCP, the reweighted $\ell_1$ subproblem has weights

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

and the weights for reweighted $\ell_2$ subproblem are

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

## IV. 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 $\mathbf{W}^k = \operatorname{diag}(\mathbf{w}^k)$, and so forth.

### A. 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 regularizer, we use Fréchet subdifferentials as the major tool in our analysis. Some important properties of Fréchet subdifferentials derived in [13] 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 [13], 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*

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

*(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*

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

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

**Lemma 1.** *Suppose Assumption 1 is satisfied. Then it holds that*

$$\nabla f(\mathbf{x}) + \prod_{i\in\mathcal{G}} y_i \partial c_i(\mathbf{x}_i) + N(\mathbf{x}|X) \subset \partial_F J_0(\mathbf{x})$$

*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 $\phi$

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

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.** *Suppose Assumption 1 is satisfied. Let* $y_i^* := \lim\limits_{c_i\to 0^+} r_i'(c_i) \geq 0$. *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* $c_i(\mathbf{x}^*) + \epsilon_i > 0$,

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

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

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

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

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

*Proof.* The statement about the case that $c_i(\mathbf{x}^*) > 0$ is obviously true. We only need consider the case that $c_i(\mathbf{x}^*) = 0$. 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$(ii)$. It can be easily verified by [13, 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

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

□

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

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

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

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

must be nonempty since it contains $\tilde{\mathbf{x}}$, and bounded due to the coercivity of $w_i^k c_i$, $i \in \mathcal{G}$ and the lower boundedness of $f$ on $X$. This completes the proof by [31, 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 $J(\mathbf{x},\boldsymbol{\epsilon}^k)$.

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

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

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

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

where the second inequality follows from (4). □

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

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

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

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

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

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

*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, $\tilde{J} := \inf\limits_{k} J(\mathbf{x}^k;\boldsymbol{\epsilon}^k) > -\infty$. 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. □

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

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

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 [31, Theorem 4.3.1].

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

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

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

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

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

$$0 = \nabla f(\mathbf{x})_i + \tilde{w}_i \boldsymbol{\xi}_i + \boldsymbol{\nu}_i,$$

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

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

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

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 7, 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 $\varepsilon := G_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\mathbf{x}^*) - G_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\bar{\mathbf{x}}) > 0$. 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 $k > k_1$

$$G_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^k) - G_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^{k+1}) \le \varepsilon/4. \qquad (8)$$

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

$$\sum_{i\in\mathcal{G}} (w_i^* - w_i^k) c_i(\bar{\mathbf{x}}_i) > -\varepsilon/12,$$

$$\sum_{i\in\mathcal{G}} (w_i^k c_i(\mathbf{x}_i^k) - w_i^* c_i(\mathbf{x}_i^*)) > -\varepsilon/12,$$

$$f(\mathbf{x}^k) - f(\mathbf{x}^*) > -\varepsilon/12.$$

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

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

and that

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

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

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

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

**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 7, a global minimizer $\tilde{\mathbf{x}}$ of* (9) *implies its optimality of* (3).

*2) Convergence Analysis for Degenerated Weights:* We have shown the convergence of AIR with fixed $\epsilon$. By Theorem 1, we can choose sufficiently small $\epsilon$ 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 0. We analyze the global convergence of AIR in this case with $c_i(\mathbf{x}_i) = \|\mathbf{x}_i\|_1$. Notice that

$$\partial c_i(0) = \{\boldsymbol{\xi}_i \in \mathbb{R}^{n_i} \mid \|\boldsymbol{\xi}_i\|_\infty \leq 1\}.$$

As the algorithm proceeds, of particular interest is the properties of the "limit subproblem" as the (sub)sequence of iterates converges. 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 $\tilde{\boldsymbol{\epsilon}} \in \mathbb{R}^m_+$

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

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

**Lemma 7.** *For $\tilde{\boldsymbol{\epsilon}} \in \mathbb{R}_+$, 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* (3)*.*

*Proof.* Notice that $\tilde{\mathbf{x}}$ is feasible for (9) by the definition of $\widetilde{G}$. The level set

$$\{\mathbf{x} \in X \mid \widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}) \leq \widetilde{G}_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\tilde{\mathbf{x}});\ \mathbf{x}_i = 0, i \in \mathcal{A}(\tilde{\mathbf{x}},\tilde{\boldsymbol{\epsilon}})\}$$

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 [31, Theorem 4.3.1].

Obviously Slater's condition holds at any feasible point of (9). Therefore, any optimal solution $\mathbf{x}$ must satisfies the KKT conditions

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

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

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

by Lemma 2. If $\tilde{\mathbf{x}}$ is an optimal solution, then we have

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

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

*Proof.* Let $\mathbf{x}^*$ be a cluster point of $\{\mathbf{x}^k\}$ and $\boldsymbol{\epsilon}^* = \lim_{k\to\infty} \boldsymbol{\epsilon}^k$. From Lemma 7, 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 $c_i(\bar{\mathbf{x}}_i) = 0$ for all $i \in \mathcal{A}(\mathbf{x}^*,\boldsymbol{\epsilon}^*)$ and $G_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\mathbf{x}^*) - G_{(\mathbf{x}^*,\boldsymbol{\epsilon}^*)}(\bar{\mathbf{x}}) > \varepsilon > 0$. 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 $k > k_1$

$$G_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^k) - G_{(\mathbf{x}^k,\boldsymbol{\epsilon}^k)}(\mathbf{x}^{k+1}) \leq \varepsilon/4. \tag{10}$$

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

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

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

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

and that

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

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

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

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

**Remark 4.** *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 7, a global minimizer $\tilde{\mathbf{x}}$ of* (9) *implies its optimality of* (3)*.*

### C. 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.** *[32, 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 [32].

**Proposition 2.** *The following hold:*

*(i) If $X \subset Y \subset \mathbb{R}^n$, then $X^\infty \subset Y^\infty$.*

*(ii) [32, Theorem 3.5] The set $Y \subset \mathbb{R}^n$ is bounded if and only if $Y^\infty = \{0\}$.*

*(iii) [32, 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) [32, 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

$$\begin{aligned} H(\mathbf{x}^0, \boldsymbol{\epsilon}^0) := \{\bar{\mathbf{x}} \mid \; & \bar{\mathbf{x}} \in X^\infty, \bar{\mathbf{x}} \in L(f(\mathbf{x}^0); f)^\infty, \\ & \bar{\mathbf{x}}_i \in L(c_i(\mathbf{x}_i^0) + \epsilon_i^0; c_i)^\infty, i \in \mathcal{G}\}, \textit{and} \\ \tilde{H}(\mathbf{x}^0, \boldsymbol{\epsilon}^0) := \; & X^\infty \cap L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0); f)^\infty \\ & \cap (\prod_{i \in \mathcal{G}} L(J(\mathbf{x}^0; \boldsymbol{\epsilon}^0) - \underline{f}; r_i \circ c_i)^\infty). \end{aligned}$$

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 $\bar{\mathbf{x}} \neq 0$. 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 $r_i \ge 0$. 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 $i \in \mathcal{G}$

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

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 6.** *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*

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

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

Based on Corollary 6, 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) $f$ is bounded below on $X$ and $r_i \circ c_i$, $i \in \mathcal{G}$ are all coercive.*
*(iv) Assume $\underline{f} := \inf_{\mathbf{x} \in X} f(\mathbf{x}) > -\infty$ and*

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

*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 6. 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

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

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

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

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 $\phi_i$ is not coercive such as (FRA).

## V. Numerical Experiments

In this section, we test our proposed AIR algorithm as a sparsity-promoting tool in two numerical experiments and exhibit its performance. In both experiment, the test problems have $f(\mathbf{x}) \equiv 0$. The algorithm is mplemented in Matlab with the subproblems solved by the CVX solver [33]. We consider two ways of choosing $r$ and $c$, $c_i(x_i) = |x_i|$ and $c_i = x_i^2$, as described in Table I, so that they can be referred as $\ell_1$-AIR and $\ell_2$-AIR, respectively. In the subproblem, we use the same value $\epsilon$ for each relaxation parameter $\epsilon_i = \epsilon$.

### A. Nonnegative Sparse Optimization

In this experiment, we use AIR to solve the nonnegative sparse optimization (NSO) problem introduced in [34] for a nonnegative sparse signal and compare its performance on sparsity-promoting with the contemporary algorithm proposed in [34]. In particular, by relaxing $\ell_0$ norm with $\ell_p$ norm $(0 < p < 1)$, the NSO problem can be formulated as

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n} \quad & \|\mathbf{x}\|_p^p \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & x_i \geq 0,\ i = 1, \ldots, n, \end{aligned}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the sensing matrix, $\mathbf{b} \in \mathbb{R}^m$ is the measurement vector, and $\mathbf{x} \in \mathbb{R}^n$ is the nonnegative sparse signal to be recovered. This $\ell_p$ norm problem can be solved by our proposed AIR algorithm.

In the numerical experiments, the simulation data are generated by the standard process of compressive sensing. Specifically, we randomly generate an i.i.d. Gaussian ensemble $\mathbf{A} \in \mathbb{R}^{m \times n}$ satisfying $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ and a nonnegative sparse signal $\bar{\mathbf{x}} \in \mathbb{R}$ via randomly setting $n_z$ components drawn from the standard uniform distribution on $[0, 1]$ for a given number of nonzeros $n_z$, while the remaining components are all zeros. Then we generate the observation data $\mathbf{b}$ by

$$\mathbf{b} = \mathbf{A}\bar{\mathbf{x}} + \sigma\boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ is the standard Gaussian noise and $\sigma = 10^{-6}$ is the corresponding deviation. The problem size is set as $n = 1024$ and $m = 256$.

We choose $p = 0.1$ and initial point $\mathbf{x}^0 = \mathbf{0}$ with maximum number of iterations $T = 500$. The relaxation parameter is chosen as $\epsilon_i^0$ and updated by $\epsilon^{k+1} = 0.7\epsilon^k$ for each iteration with minimum threshold $10^{-5}$. The algorithm is terminated and deemed to find an optimal solution when

$$\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2 \leq 10^{-4}.$$

Our results are the average of 10 experiments where $\mathbf{A}, \bar{\mathbf{x}}$ and $\boldsymbol{\varepsilon}$ are regenerated according to the same rules.

The performance of AIR is demonstrated in two aspects including the quality of recovered signal and the computational efficiency. Figure 1 depicts the rate of success achieved by AIR and IPTA proposed in [35] for signals with different number of nonzero components $n_z$. Here the signal is considered as "successfully recovered" if the relative error between the final solution returned by the algorithm and the original signal is smaller than $0.5\%$, i.e.,

$$\|\mathbf{x}^k - \bar{\mathbf{x}}\|_2 / \|\bar{\mathbf{x}}\|_2 \leq 0.5\%.$$

Since the results for $\ell_1$-AIR and $\ell_2$-AIR are the same for each experiment, we use the same curve for the two versions of AIR.

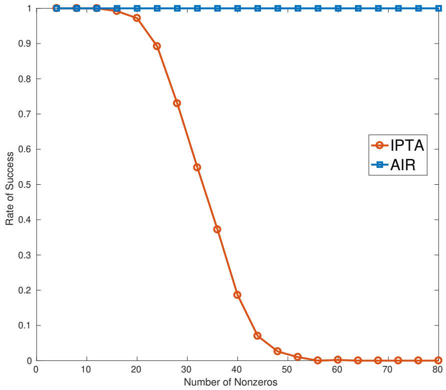

Fig. 1: Rate of success v.s. number of nonzeros for AIR and IPTA.

From Figure 1, it is witnessed that AIR maintains 100% rate of success for all test examples. As for IPTA, it can successfully solve all the examples for signals with less than 15 nonzero components. However, its rate of success quickly reduces to 0 as the number of nonzeros exceeds 20, and remains 0 for signals with more than 55 nonzeros. Overall, we can see that AIR overperforms IPTA in recovering accurate sparse signals. It should be noticed that AIR does not require the (estimated) number of nonzeros as a prior, as compared to IPTA.

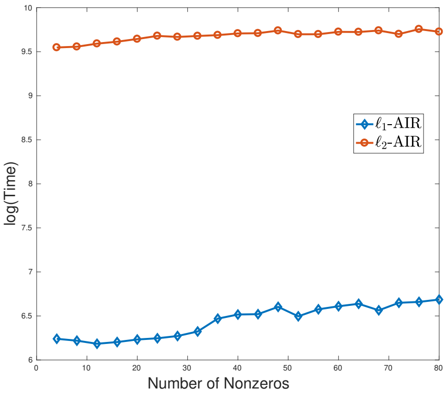

Fig. 2: CPU times v.s. number of nonzeros for AIR and IPTA.

It would be interesting to compare the performance efficiency of $\ell_1$-AIR and $\ell_2$-AIR, since the $\ell_1$-reweighted subproblems are nonsmooth compared with the smooth $\ell_2$-reweighted subproblems. Figure 2 depicts the logarithm of the CPU time required by $\ell_1$-AIR and $\ell_2$-AIR for signals with different number of nonzeros. One can see from Figure 2 that $\ell_1$-AIR is computationally much faster than $\ell_2$-AIR. This indicates that $\ell_2$-AIR needs more iterations than $\ell_1$-AIR, since $\ell_2$-AIR has simpler subproblems to solve. This indication is supported by the numerical results shown in Table II, which summarizes the number of iterations needed by $\ell_1$-AIR and $\ell_2$-AIR in Table II on average. We can see that for cases with fewer nonzeros, $\ell_1$-AIR only needs 1-2 iterations, and as for $\ell_2$-AIR, 10-15 iterations are needed. This implies that the $\ell_p$ regularization can be easily approximated by the weighted $\ell_1$ regularization, yet by a sequence of weighted $\ell_2$ regularization.

TABLE II: The average number of nonzeros

| | Number of Nonzeros | | | |
|---|---|---|---|---|
| | 4-20 | 24-40 | 44-60 | 64-80 |
| $\ell_1$-AIR | 1.02 | 1.44 | 1.98 | 2.20 |
| $\ell_2$-AIR | 12.50 | 13.54 | 14.06 | 14.44 |

## B. Group Sparse Optimization

In the second experiment, we consider an optimization model for cloud radio access network (Cloud-RAN) power consumption problem with group sparse structure [36], which can be formulated as a mixed-integer nonlinear programming (MINLP) problem. In order to solve this MINLP problem, a three-stage group sparse beamforming (GSBF) method is proposed in [36], which solves a group sparse problem in the first stage to induce the group sparsity for the beamformers.

We consider the Cloud-RAN architecture with $L$ remote radio heads (RRHs) and $K$ single-antenna mobile users, where the $l$-th RRH is equipped with $N_l$ antennas. To promote sparsity in the transmit vector, the group sparse problem aims to minimize the $\ell_0$ norm of each transmit vector. By relaxing $\ell_0$ norm with $\ell_p$ norm $(0 < p < 1)$, the group sparse problem can be formulated as

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

where $\mathbf{v}_{lk} \in \mathbb{C}^{N_l}$ is the transmit beamforming vector from the $l$-th RRH to the $k$-th user, and $\tilde{\boldsymbol{v}}_l = [\mathbf{v}_{l1}^T, ..., \mathbf{v}_{lK}^T]^T \in \mathbb{C}^{KN_l \times 1}$ is the group structure of transmit vectors. $\rho_l$ is the weight for the beamforming coefficients group $\tilde{\mathbf{v}}_l$ at RRH $l$. 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. $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_K)^T$ is the target signal-to-interference-plus-noise ratio (SINR). The SINR constraint for user $k$ is formulated as a second-order cone constraint [21]. This Cloud-RAN $\ell_p$ norm problem can be solved by our AIR algorithm.

In our experiment, we consider a network with $L = 10$, $K = 6$, 2-antenna RRHs and single-antenna MUs uniformly and independently distributed in the square region $[-1000, \ 1000] \times [-1000, \ 1000]$ meters. Each point of the simulation results is averaged over 50 randomly generated network realizations.

We set the maximum number of iterations as $T = 500$, $\epsilon^0 = 100$ for AIR and update by $\epsilon^{k+1} = 0.7\epsilon^k$ at each iteration with minimum threshold $10^{-6}$. The algorithm is terminated whenever

$$\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2 \leq 10^{-5}$$

is satisfied.

In Figure 3, we depicts the number of nonzero components of the final solution returned by $\ell_1$-AIR and $\ell_2$-AIR for problems with different SINR. It is witnessed again that the $\ell_1$-AIR outperforms $\ell_2$-AIR in the ability of accurately recovering sparse solutions.

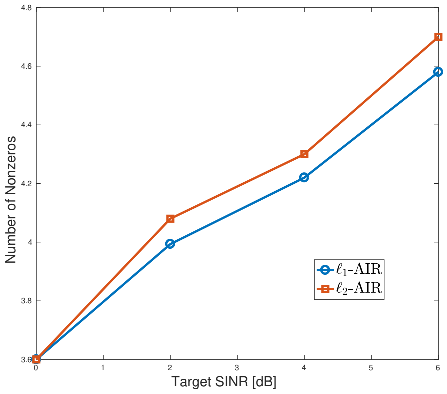

Fig. 3: Average sparsity versus target SINR.

To further investigate the behaviors of two variants of AIR, we depicts the final solution found by $\ell_1$-AIR and $\ell_2$-AIR in Figure 4. Here, the $x$-axis is the element index $l = 1, \ldots, 10$ and the each point in Figure 4 is the corresponding $\log(\|\tilde{\mathbf{v}}_l\|_2)$. We can see that for $\ell_1$-AIR, elements $l = 3, \ldots, 7$ have already been driven to 0 (as small as $10^{-20}$), whereas these elements are still relatively large for $\ell_2$-AIR, between $10^{-5}$ and $10^{-2}$. This may be largely due to the derivative of the squared $\ell_2$ norm diminishes as the variable tends to 0, resulting in tiny step and hence slow progress.

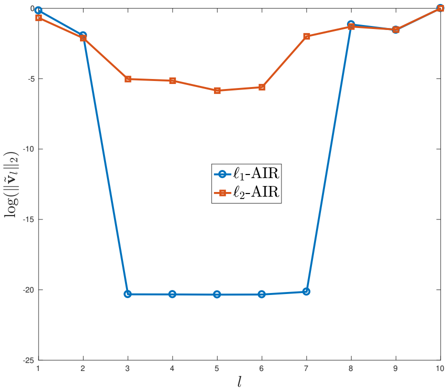

Fig. 4: Final solution returned by $\ell_1$-AIR and $\ell_2$-AIR.

## VI. Conclusions

In this paper, we have proposed a general formulation for nonconvex regularization problem, which can take into account different regularization terms. An iteratively reweighted algorithm is proposed by solving a sequence of weighted convex regularization subproblems. We have also derived the optimality condition for the nonconvex regularization problem and provided the global convergence analysis for the proposed iteratively reweighted methods.

Two variants of our proposed algorithm, the reweighted $\ell_1$ method and the reweighted $\ell_2$ method, are implemented and tested. Numerical results exhibits their ability of recovering sparse signals. It is also witnessed that the iteratively reweighted $\ell_1$ method is generally faster than the reweighted $\ell_2$ method because much fewer iterations are needed for reweighted $\ell_1$. Overall, our investigation leads to a variety of interesting research directions:

- For $\ell_2$-AIR with $\epsilon$ driven to zero, existing global convergence results is provided only for the $\ell_p$ norm regularization without constraints. For general constrained regularized problem, the global convergence of $\ell_1$-AIR is still an open question.
- A thorough comparison, through either theoretical analysis or numerical experiments, of the existing nonconvex regularizations using AIR would be interesting to see. This should be helpful in providing the guidance for the users to select regularizers.
- Our implementation reduces the relaxation parameter $\epsilon$ 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 $\epsilon$.
- It would be meaningful to have a (local) complexity analysis for the reweighted algorithms.

## References

[1] P. S. Bradley, O. L. Mangasarian, and W. N. Street, "Feature selection via mathematical programming," *INFORMS Journal on Computing*, vol. 10, no. 2, pp. 209–217, 1998.

[2] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero-norm with linear models and kernel methods," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1439–1461, 2003.

[3] Y. Shi, J. Cheng, J. Zhang, B. Bai, W. Chen, and K. B. Letaief, "Smoothed $\ell_p$-minimization for green cloud-ran with user admission control," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 1022–1036, 2016.

[4] A. Lanza, S. Morigi, and F. Sgallari, "Convex image denoising via nonconvex regularization," in *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 666–677, Springer, 2015.

[5] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted 1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[6] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on*, pp. 3869–3872, IEEE, 2008.

[7] J. V. Burke, F. E. Curtis, H. Wang, and J. Wang, "Iterative reweighted linear least squares for exact penalty subproblems on product sets," *SIAM Journal on Optimization*, vol. 25, no. 1, pp. 261–294, 2015.

[8] Z. Lu, "Iterative reweighted minimization methods for $\ell_p$ regularized unconstrained nonlinear programming," *Mathematical Programming*, vol. 147, no. 1-2, pp. 277–307, 2014.

[9] Z. Lu, Y. Zhang, and J. Lu, "$\ell_p$ regularized low-rank approximation via iterative reweighted singular value minimization," *Computational Optimization and Applications*, vol. 68, no. 3, pp. 619–642, 2017.

[10] M. S. Bazaraa, J. Goode, and M. Z. Nashed, "On the cones of tangents with applications to mathematical programming," *Journal of Optimization Theory and Applications*, vol. 13, no. 4, pp. 389–426, 1974.

[11] A. Y. Kruger, "Subdifferentials of nonconvex functions and generalized directional derivatives," *Mimeographied notes, VINITI Moscow*, pp. 2661–77, 1977.

[12] A. Y. Kruger, "$\varepsilon$-semidifferentials and $\varepsilon$-normal elements," *Depon. VINITI*, vol. 1331, 1981.

[13] A. Y. Kruger, “On fréchet subdifferentials,” *Journal of Mathematical Sciences*, vol. 116, no. 3, pp. 3325–3358, 2003.

[14] E. J. Candès and M. B. Wakin, “An introduction to compressive sampling,” *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[15] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, “Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things,” *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 88–99, 2018.

[16] Z. Qin, J. Fan, Y. Liu, Y. Gao, and G. Y. Li, “Sparse representation for wireless communications: A compressive sensing approach,” *IEEE Signal Processing Magazine*, vol. 35, no. 3, pp. 40–58, 2018.

[17] Y. Shi, J. Zhang, W. Chen, and K. B. Letaief, “Generalized sparse and low-rank optimization for ultra-dense networks,” *IEEE Communications Magazine*, vol. 56, no. 6, pp. 42–48, 2018.

[18] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, *et al.*, “Optimization with sparsity-inducing penalties,” *Foundations and Trends® in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[19] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[20] M. A. Khajehnejad, A. G. Dimakis, W. Xu, and B. Hassibi, “Sparse recovery of nonnegative signals with minimal expansion,” *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 196–208, 2011.

[21] Y. Shi, J. Zhang, and K. B. Letaief, “Group sparse beamforming for green cloud-ran,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, 2014.

[22] F. E. Harrell, “Ordinal logistic regression,” in *Regression modeling strategies*, pp. 311–325, Springer, 2015.

[23] M. J. Wainwright, “Structured regularizers for high-dimensional problems: Statistical and computational issues,” *Annual Review of Statistics and Its Application*, vol. 1, pp. 233–253, 2014.

[24] P. S. Bradley and O. L. Mangasarian, “Feature selection via concave minimization and support vector machines,” in *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML ’98, (San Francisco, CA, USA), pp. 82–90, Morgan Kaufmann Publishers Inc., 1998.

[25] M. Fazel, H. Hindi, and S. P. Boyd, “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.

[26] M. S. Lobo, M. Fazel, and S. Boyd, “Portfolio optimization with linear and fixed transaction costs,” *Annals of Operations Research*, vol. 152, no. 1, pp. 341–365, 2007.

[27] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[28] C. H. Zhang, “Penalized linear unbiased selection,” *Department of Statistics and Bioinformatics, Rutgers University*, vol. 3, 2007.

[29] R. Chartrand and W. Yin, “Iterative reweighted algorithms for compressive sensing,” tech. rep., 2008.

[30] Q. Ling, Z. Wen, and W. Yin, “Decentralized jointly sparse optimization by reweighted $\ell_q$ minimization,” *IEEE Transactions on Signal Processing*, vol. 61, no. 5, pp. 1165–1170, 2013.

[31] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, vol. 30. Siam, 1970.

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

[33] M. Grant, S. Boyd, and Y. Ye, “Cvx: Matlab software for disciplined convex programming (2008),” *(Web page and software.) http://stanford.edu/ boyd/cvx*, 2015.

[34] D. L. Donoho and J. Tanner, “Sparse nonnegative solution of underdetermined linear equations by linear programming,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9446–9451, 2005.

[35] L. Zhang, Y. Hu, C. K. W. Yu, and J. Wang, “Iterative positive thresholding algorithm for non-negative sparse optimization,” *Optimization*, pp. 1–19, 2018.

[36] Y. Shi, J. Zhang, and K. B. Letaief, “Group sparse beamforming for green cloud-ran,” *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2809–2823, 2014.