COMMUN. MATH. SCI.
Vol. 18, No. 1, pp. 227–258

AN EFFICIENT AND GLOBALLY CONVERGENT ALGORITHM FOR $\ell_{p,q}$-ℓ_r MODEL IN GROUP SPARSE OPTIMIZATION*

YUNHUA XUE†, YANFEI FENG‡, AND CHUNLIN WU§

Abstract. Group sparsity has lots of applications in various data science related problems. It combines the underlying sparsity and group structure of the variables. A general and important model for group sparsity is the $\ell_{p,q}$-ℓ_r optimization model with $p \geq 1$, $0<q<1$, $1 \leq r \leq \infty$, which is applicable to different types of measurement noises. It includes not only the non-smooth composition of ℓ_q $(0<q<1)$ and ℓ_p $(p \geq 1)$, but also the non-smooth ℓ_1/ℓ_∞ fidelity term. In this paper, we present a nontrivial extension of our recent work to solve this general group sparse minimization model. By a motivating proposition, our algorithm is naturally designed to shrink the group support and eliminate the variables gradually. It is thus very fast, especially for large-scale problems. Combined with a proximal linearization, it allows an inexact inner loop implemented by scaled alternating direction method of multipliers (ADMM), and still has global convergence. The algorithm gives a unified framework for the full parameters. Many numerical experiments are presented for various combinations of the parameters p,q,r. The comparisons show the advantages of our algorithm over others in the existing works.

Keywords. group sparse; $\ell_{p,q}$-ℓ_r model; non-Lipschitz optimization; Laplace noise; Gaussian noise; uniform distribution noise; lower bound theory; Kurdyka-Łojasiewicz (KL) property.

AMS subject classifications. 49M05; 65K10; 90C26; 90C30.

1. Introduction

In Big Data era, data variables used to describe the structures, segments and features usually have group property. Namely, they have a natural grouping of their components. Sparsity allows us to reconstruct high-dimensional data with only a small number of sample variables, leading to better recovery performance. By combining them, group sparse recovery or reconstruction has a wide variety of applications, such as signal recovery [17, 22], image processing [38], compressed sensing [37], model selection in birth weight prediction [42], sparse learning [8, 40], variable selection in gene finding [30] and so on. This topic is enhanced to be an active research topic in recent years.

In this paper, we consider the following $\ell_{p,q}$-ℓ_r group sparse optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^N} \mathcal{E}(\mathbf{x}) := \|\mathbf{x}\|_{p,q}^q + F_r(\mathbf{x}), \tag{1.1}$$

where $p \in [1,\infty)$, $q \in (0,1)$ and

$$F_r(\mathbf{x}) = \begin{cases} \frac{1}{r\alpha}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_r^r, \; r \geq 1, \\ \frac{1}{\alpha}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_\infty, \; r=\infty, \end{cases}$$

with $r \in [1,\infty], \alpha \in (0,\infty)$, $\mathbf{A} \in \mathbb{R}^{M\times N}$, $\mathbf{y} \in \mathbb{R}^M$. The $\ell_{p,q}$ regularization term, measuring

*Received: July 13, 2019; Accepted (in revised form): September 17, 2019. Communicated by Wotao Yin.

†School of Mathematical Sciences and Key Laboratory of Pure Mathematics and Combinatorics of Ministry of Education (LPMC), Nankai University, Tianjin 300071, China (yhxue@nankai.edu.cn).

‡School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China (fengyanfei@mail.nankai.edu.cn).

§Corresponding author. School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China (wucl@nankai.edu.cn).

the group sparse structure of $\mathbf{x}$, is a quasi-norm and defined by

$$\|\mathbf{x}\|_{p,q} = \left(\sum_{\mathsf{i}=1}^{\mathsf{g}} \|\mathbf{x}_\mathsf{i}\|_p^q\right)^{1/q},$$

where $\mathbf{x}_\mathsf{i}, \mathsf{i}=1,\cdots,\mathsf{g}$ are the group members defined in Section 2 and $\|\cdot\|_p$ is the standard ℓ_p norm for vectors. The minimization model (1.1) is a quite general non-convex extension of the group lasso [42] and covers a wide range of applications. Firstly, the $\ell_{p,q}$ regularization term is a general form with different choices of p,q. Here $p\in[1,\infty)$ reflects the group structure of the variables and $q\in(0,1)$ is used to describe the sparsity. When $p=1$, it actually allows the sparsity to exist within a group, according to the theory of compressive sensing. Secondly, the data fidelity term $F_r(\mathbf{x})$ covers extensive applications. It is derived from the statistical property of the random measurement noise $\mathbf{n}\in\mathbb{R}^M$ in the observed data

$$\mathbf{y} = \mathbf{A}\mathbf{x}_{or} + \mathbf{n},$$

where $\mathbf{x}_{or}$ is the ground truth or the original signal. Nowadays it is clear from MAP to use squared ℓ_2 fidelity term ($r=2$) for Gaussian noise, the ℓ_1 fidelity term ($r=1$) for Laplace noise and heavy-tailed noise, and the ℓ_∞ fidelity term ($r=\infty$) for uniformly distributed noise and quantization error.

The goal of this paper is to design an efficient and unified algorithm with a global convergence guarantee to solve the general non-Lipshitz minimization model (1.1). Before this, we review some related algorithms.

We first summarize several classes of algorithms for a special case of (1.1), i.e., the non-group sparse optimization model in which the number of groups g equals N. In this case, the group structure vanishes and the $\ell_{p,q}$ term in (1.1) is degenerated to ℓ_q $(0<q<1)$ regularization one. The first class of methods is the smoothing approximate one [5,12–14,25]. By a sophisticated smoothing function $\varphi(x,\theta)$, the non-Lipschitz property of the objective function can be removed. The second class of method is the general iterative shrinkage-thresholding algorithm (GISA) for ℓ_q-ℓ_2 problem [9,39,46], which was inspired by the soft thresholding and iterative shrinkage-thresholding algorithm (ISTA) [3,16] for convex ℓ_1-ℓ_2 problem. The third class of method is the iterative reweighted minimization method, like the very successful IRL1 and IRLS for ℓ_q-ℓ_2 minimization problem; see, e.g., [10,15,24,28]. Actually the reweighted methods reformulate the original non-Lipschitz ℓ_q-ℓ_2 to Lipschitz ones by a de-singularizing parameter. Very recently, a strategy of progressively shrinking the support of the variables to overcome non-Lipschitz property was derived and presented for different problems respectively in [26,43,44], where [26] considered the non-group case with $r\neq\infty$ and [43,44] focused on the image restoration with $r=1,2$. To the best of our knowledge, we note that none of these references considered all the $1\le r\le\infty$, even in the non-group case.

As for the group sparse optimization problem (1.1), the authors in [22] proposed an GISA extension with interesting convergence results for the case of $r=2$. This is the only existing algorithm for the general model (1.1) we can find in the literature. Note that with $r=2$, the data fidelity term is a smooth one. For the general problem (1.1) with more difficult non-smooth fidelity terms, there isn't any approach. On the other hand, the general model (1.1) with $1\le r\le\infty$ has various applications, as mentioned before. We therefore try to design a unified algorithm for solving (1.1). Because of the diversity of p,q,r and the non-smooth $\ell_{p,q}$ (especially $\ell_{1,q}$) in the objective function $\mathcal{E}$, we present a non-trivial extension of the recent iterative support shrinking strategy, especially [26], for this general problem (1.1).

We firstly establish a motivating proposition by developing subdifferential lemmas in group variables. This gives us the rationality to apply a unified iterative support shrinking technique over group support set with variable elimination for various p,q,r. To make the technique more practical and easily implementable, we linearize the objective and present an inexact iterative support shrinking algorithm with a proximal linearization for group sparse optimization (InISSAPL-GSO). Although the algorithm allows an inexact inner loop, we prove its global convergence from a new lower bound theory for the ℓ_p norm of the nonzero groups of iteration sequence. The algorithm implementation by scaled ADMM is also discussed where, especially for the case of $r=\infty$, we give an analytical derivation of the sub-solvers. Numerical experiments show that the algorithm is robust to the diversity of p,q,r. Compared with others in group sparse optimization with respect to relative errors, successful rates and running time, our algorithm outperforms them. The main characters of InISSAPL-GSO for model (1.1) are presented as follows:

(1) The algorithm provides a unified framework for the full parameters p,q,r. It can particularly deal with the case of the addition of non-smooth $\ell_{1,q}$ regularization term and non-smooth ℓ_1/ℓ_∞ fidelity term.

(2) The computation is implemented only on the shrinking group support set of $\mathbf{x}$ at each iteration. Naturally our algorithm is efficient, especially for large-scale sparse recovery problems.

(3) The key step is to overcome the non-Lipschitz property of the objective function and construct an appropriate subdifferential formula, when using KL property to prove the global convergence of the algorithm. It is solved by developing a lower bound theory of the nonzero groups of the iterative sequence and a technical construction of the subdifferential; see Section 4 for details.

The rest of the paper is outlined as follows. In Section 2, we give some basic notations and preliminaries. In Section 3, we give the motivating proposition and propose the corresponding algorithms. In Section 4, we establish the global convergence theorem for the proposed algorithms. In Section 5, we describe the implementation of the algorithm by scaled ADMM. Numerical experiments and comparisons are showed in Section 6. Section 7 concludes the paper.

2. Notation and preliminary

Suppose that $\mathbf{A}$ is an $M\times N$ matrix and $\mathbf{x}$ is a column vector with N entries. $I=\{1,2,\ldots,M\}$ denotes the row index set of $\mathbf{A}$. To be specialized, we use another kind of upright font to express the group index such as $\mathsf{G},\mathsf{i},\mathsf{g}$. Let $\mathbf{x}:=\left(\mathbf{x}_1^T,\mathbf{x}_2^T,\cdots,\mathbf{x}_\mathsf{g}^T\right)^T$ represent the group structure of $\mathbf{x}$. $\mathsf{G}=\{1,2,\ldots,\mathsf{g}\}$ denotes the group index set of $\mathbf{x}$. For each group member $\mathbf{x}_\mathsf{i}$, we denote by $J_\mathsf{i}=\{1,2,\cdots,N_\mathsf{i}\}$ the index set. Therefore $N=N_1+\cdots+N_\mathsf{g}$. We also refer to $\mathbf{x}_{\mathsf{i},j}$ as the j-th entry of $\mathbf{x}_\mathsf{i}$ and denote the group support set of $\mathbf{x}$ by

$$\mathrm{supp}_\mathsf{G}(\mathbf{x}):=\{\mathsf{i}\in\mathsf{G}:\mathbf{x}_\mathsf{i}\neq\mathbf{0}\},$$

where $\mathbf{x}_\mathsf{i}\neq\mathbf{0}$ means that $\mathbf{x}_{\mathsf{i},j}\neq 0$ for some $j\in J_\mathsf{i}$. Furthermore, we use $\mathbf{x}_\mathsf{i}=\mathbf{0}$ when $\mathbf{x}_{\mathsf{i},j}=0$ for all $j\in J_\mathsf{i}$. The support of group member $\mathbf{x}_\mathsf{i}$ is defined by

$$\mathrm{supp}(\mathbf{x}_\mathsf{i})=\{j\in J_\mathsf{i}:\mathbf{x}_{\mathsf{i},j}\neq 0\}.$$

Let S be a subset of G. We denote by $\mathbf{x}_\mathsf{S}$ the group vectors of $\mathbf{x}$ indexed by S, which consist of the nonzero group members of $\mathbf{x}$ with $\mathsf{S}=\mathrm{supp}_\mathsf{G}(\mathbf{x})$.

For a matrix $\mathbf{A}\in\mathbb{R}^{M\times N}$, we partition it into submatrices $A_{k,\mathsf{i}}, k\in I, \mathsf{i}\in\mathsf{G}$, which is the i-th block of the k-th row of $\mathbf{A}$ associating to the group structure of $\mathbf{x}$, i.e.,

$$\mathbf{A}=\begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,\mathsf{g}} \\ \cdots & \cdots & \cdots & \cdots \\ A_{M,1} & A_{M,2} & \cdots & A_{M,\mathsf{g}} \end{bmatrix}.$$

Because $A_{k,\mathsf{i}}, k\in I, \mathsf{i}\in\mathsf{G}$ are row vectors, we denote by $(A_{k,\mathsf{i}})_j$ its j-th entry. In a similar way with $\mathbf{x}_\mathsf{S}$, we denote by $\mathbf{A}_\mathsf{S}$ the column sub-matrix of $\mathbf{A}$ consisting of the columns indexed by S.

Define $\phi:[0,\infty)\to[0,\infty)$ by $\phi(x)=x^q (0<q<1)$. We state some useful properties for $\phi(\cdot)$.

PROPOSITION 2.1 ([26]). *The function $\phi(\cdot)$ has the following properties:*

(1) $\phi(0)=0$ and $\phi'(x)=qx^{q-1}>0$ on $(0,\infty)$.

(2) $\phi(x)$ is concave and the following inequality holds,

$$\phi(y)\le\phi(x)+\phi'(x)(y-x),\ \forall x\in(0,\infty), y\in[0,\infty). \tag{2.1}$$

(3) For any $c>0$, $\phi'(x)$ is L_c-Lipschitz continuous on $[c,\infty)$, i.e., there exists a constant $L_c>0$ determined by c, such that $\forall x,y\in[c,\infty)$,

$$|\phi'(x)-\phi'(y)|\le L_c|x-y|. \tag{2.2}$$

LEMMA 2.1 ([22]). *Let $\mathbf{y}\in\mathbb{R}^m$ be the m-dimensional vector. The following inequality holds:*

$$\|\mathbf{y}\|_{\gamma_2}\le\|\mathbf{y}\|_{\gamma_1}, 0<\gamma_1\le\gamma_2. \tag{2.3}$$

LEMMA 2.2. *Let $s>0, \mathbf{y}\in\mathbb{R}^m$. Then there exists a constant $\widetilde{C}_s>0$ such that,*

$$\|\mathbf{y}\|_s\le\widetilde{C}_s\|\mathbf{y}\|_{s+1}. \tag{2.4}$$

Proof. For $s\ge1$, there exists $C>0$ which satisfies the above inequality by the norm equivalence. For $0<s<1$, from [22, Lemma 1], there exists $C_s>0$ such that

$$\|\mathbf{y}\|_s\le C_s\|\mathbf{y}\|_2.$$

We use the norm equivalence once again to have

$$\|\mathbf{y}\|_s\le C_sC\|\mathbf{y}\|_{s+1}.$$

Hence we can choose $\widetilde{C}_s=\max\{C, C_sC\}$ to finish the proof. □

3. Motivation and the proposed algorithm

3.1. Subdifferentials and regularity. By the definition of $\phi(\cdot)$, we have $\|\mathbf{x}\|_{p,q}^q=\sum_{\mathsf{i}\in\mathsf{G}}\phi(\|\mathbf{x}_\mathsf{i}\|_p)$. We also define the norm function $g(\mathbf{y})=\|\mathbf{y}\|_p$ for a vector $\mathbf{y}$. In order to calculate the subdifferential (see Definition A.1 in the Appendix A) of the objective function $\mathcal{E}(\mathbf{x})$ in (1.1), we firstly give two lemmas. Although they are straightforward results, we put their proofs here for the sake of completeness.

LEMMA 3.1 (Subdifferential). *Let $\mathbf{y}\in\mathbb{R}^m$ be an m-dimensional vector. We have the following results:*

(1) For $\mathbf{y}=\mathbf{0}$ *and* $p\geq 1$*, the subdifferential is,*

$$\partial(\phi\circ g)(\mathbf{y})=\prod_{j=1}^{m} S_j,$$

where $S_j=(-\infty,\infty),\forall j=1,2,\cdots,m$ *and* Π *means the Cartesian product of sets;*

(2) For $\mathbf{y}\neq\mathbf{0}$*, the subdifferential is*

$$\partial(\phi\circ g)(\mathbf{y})=\prod_{j=1}^{m} S_j,$$

where

$$S_j=\begin{cases}\phi'(\|\mathbf{y}\|_p)\,\|\mathbf{y}\|_p^{1-p}\,|\mathbf{y}_j|^{p-1}\,\mathrm{sgn}(\mathbf{y}_j), & p>1,\\ \phi'(\|\mathbf{y}\|_1)\,\mathrm{sgn}(\mathbf{y}_j), & j\in\mathrm{supp}(\mathbf{y})\ \textit{and}\ p=1,\\ [-\phi'(\|\mathbf{y}\|_1),\phi'(\|\mathbf{y}\|_1)], & j\notin\mathrm{supp}(\mathbf{y})\ \textit{and}\ p=1.\end{cases}$$

Proof. For brevity, we denote the set $\prod_{j=1}^{m} S_j$ by S.

(i). We take a $\mathbf{u}\in\widehat{\partial}(\phi\circ g)(\mathbf{y})$ with $\mathbf{y}=\mathbf{0}$. By the definition,

$$\liminf_{\substack{\mathbf{z}\to\mathbf{0}\\ \mathbf{z}\neq\mathbf{0}}}\frac{\|\mathbf{z}\|_p^q-<\mathbf{u},\mathbf{z}-\mathbf{0}>}{\|\mathbf{z}-\mathbf{0}\|_2}\geq 0.$$

From the equivalence of norms when $p\geq 1$, we have

$$\|\mathbf{z}\|_p\geq C\,\|\mathbf{z}\|_2,$$

where $C>0$ is a constant. It is sufficient to have

$$\frac{\|\mathbf{z}\|_p^q-<\mathbf{u},\mathbf{z}-\mathbf{0}>}{\|\mathbf{z}-\mathbf{0}\|_2}\geq\frac{C^q\,\|\mathbf{z}\|_2^q-<\mathbf{u},\mathbf{z}>}{\|\mathbf{z}\|_2}\geq 0,\quad \mathbf{z}\to\mathbf{0}.$$

This is true for any $\mathbf{u}\in S$ due to $0<q<1$. Then the proof is done by the fact $\widehat{\partial}(\phi\circ g)(\mathbf{y})\subseteq\partial(\phi\circ g)(\mathbf{y})$.

(ii). For $p>1$, the function $(\phi\circ g)(\mathbf{y})$ is continuously differentiable at $\mathbf{y}\neq\mathbf{0}$, so the subdifferential is the gradient in this case. For $p=1$, we show that $S=\widehat{\partial}(\phi\circ g)(\mathbf{y})$ firstly. On the one hand, letting $\mathbf{u}\in\widehat{\partial}(\phi\circ g)(\mathbf{y})$ with $\mathbf{y}\neq\mathbf{0}$, we know that the limit inferior holds along a special direction, i.e.,

$$\liminf_{\substack{\mathbf{z}_k=\mathbf{y}_k,k\neq j\\ \mathbf{z}_j\to\mathbf{y}_j\\ \mathbf{z}\neq\mathbf{y}}}\frac{\|\mathbf{z}\|_1^q-\|\mathbf{y}\|_1^q-<\mathbf{u},\mathbf{z}-\mathbf{y}>}{\|\mathbf{z}-\mathbf{y}\|_2}\geq 0.$$

Then we have

$$\begin{cases}(\mathbf{u})_j=\phi'(\|\mathbf{y}\|_1)\cdot\mathrm{sgn}(\mathbf{y}_j), & j\in\mathrm{supp}(\mathbf{y}),\\ |(\mathbf{u})_j|\leq\phi'(\|\mathbf{y}\|_1), & j\notin\mathrm{supp}(\mathbf{y}),\end{cases}$$

by the differential mean value theorem. So $\widehat{\partial}(\phi\circ g)(\mathbf{y})\subseteq S$.

On the other hand, we construct a function $h(\mathbf{z})$ in the neighbourhood of $\mathbf{y}$ ($\mathbf{y}\neq 0$):

$$h(\mathbf{z})=\left(\sum_{j\in\operatorname{supp}(\mathbf{y})}|\mathbf{z}_j|+\sum_{j\notin\operatorname{supp}(\mathbf{y})}k_j\mathbf{z}_j\right)^q,$$

where k_j can be any value in $[-1,1]$. Then h satisfies the condition of [35, Proposition 8.5] and we have $\nabla h(\mathbf{y})\in\widehat{\partial}(\phi\circ g)(\mathbf{y})$. Since

$$(\nabla h(\mathbf{y}))_j=\begin{cases}\phi'(\|\mathbf{y}\|_1)\cdot\operatorname{sgn}(\mathbf{y}_j), & j\in\operatorname{supp}(\mathbf{y}),\\ \phi'(\|\mathbf{y}\|_1)\cdot k_j, & j\notin\operatorname{supp}(\mathbf{y}),\end{cases}$$

we obtain $S\subseteq\widehat{\partial}(\phi\circ g)(\mathbf{y})$. Hence $S=\widehat{\partial}(\phi\circ g)(\mathbf{y})$.

The only thing left to show is $\partial(\phi\circ g)(\mathbf{y})\subseteq\widehat{\partial}(\phi\circ g)(\mathbf{y})$, since the inclusion relationship in the other direction holds naturally. In fact, for $\mathbf{u}\in\partial(\phi\circ g)(\mathbf{y})$ with $\mathbf{y}\neq\mathbf{0}$, there exists $\mathbf{z}^{(k)}\to\mathbf{y},\phi(\|\mathbf{z}^{(k)}\|_1)\to\phi(\|\mathbf{y}\|_1)$ and $\mathbf{u}^{(k)}\in\widehat{\partial}(\phi\circ g)(\mathbf{z}^{(k)}),\mathbf{u}^{(k)}\to\mathbf{u}$. It is clear that $\operatorname{supp}(\mathbf{y})\subseteq\operatorname{supp}(\mathbf{z}^{(k)})$ when k is sufficiently large. Together with the fact

$$\begin{cases}(\mathbf{u}^{(k)})_j=\phi'(\|\mathbf{z}^{(k)}\|_1)\cdot\operatorname{sgn}(\mathbf{z}_j^{(k)}), & j\in\operatorname{supp}(\mathbf{z}^{(k)}),\\ |(\mathbf{u}^{(k)})_j|\leq\phi'(\|\mathbf{z}^{(k)}\|_1), & j\notin\operatorname{supp}(\mathbf{z}^{(k)}),\end{cases}$$

we therefore obtain $\mathbf{u}\in\widehat{\partial}(\phi\circ g)(\mathbf{y})$ by a limiting procedure. □

The regularity property of functions is essential for subdifferential calculus of the addition of two non-smooth terms, e.g., $\ell_{1,q}$ term and ℓ_1/ℓ_∞ fidelity term. The following lemma is very useful in this work.

LEMMA 3.2 (Regularity). *Let $\mathbf{y}\in\mathbb{R}^m$ be an m-dimensional vector. Then $(\phi\circ g)(\mathbf{y})$ is regular at $\mathbf{y}$ for $p\geq 1$.*

Proof. By [35, Corollary 8.11], $(\phi\circ g)(\mathbf{y})$ is regular at $\mathbf{y}$ if and only if

$$\partial(\phi\circ g)(\mathbf{y})=\widehat{\partial}(\phi\circ g)(\mathbf{y}),\quad \partial^\infty(\phi\circ g)(\mathbf{y})=(\widehat{\partial}(\phi\circ g)(\mathbf{y}))^\infty. \tag{3.1}$$

In the proof of Lemma 3.1, we know that the first equality in (3.1) holds. The only thing left is to verify the second equality.

For $\mathbf{y}=\mathbf{0}$, we have

$$\widehat{\partial}(\phi\circ g)(\mathbf{y})=(-\infty,\infty)^m.$$

Some simple calculations using Definition A.2 in the Appendix A show that the horizon cone $(\widehat{\partial}(\phi\circ g)(\mathbf{y}))^\infty$ is the same set $(-\infty,\infty)^m$. We can also conclude $\partial^\infty(\phi\circ g)(\mathbf{y})=(-\infty,\infty)^m$ by the same trick.

For $\mathbf{y}\neq\mathbf{0}$, we have

$$\partial^\infty(\phi\circ g)(\mathbf{y})=(\widehat{\partial}(\phi\circ g)(\mathbf{y}))^\infty=\{\mathbf{0}\},$$

by the boundedness of $\widehat{\partial}(\phi\circ g)(\mathbf{y})$. □

REMARK 3.1. From [35, Proposition 10.5] for separable functions, the sum function

$$\|\mathbf{x}\|_{p,q}^q=\sum_{\mathsf{i}\in\mathsf{G}}\phi(\|\mathbf{x}_\mathsf{i}\|_p)$$

is also regular at $\mathbf{x}$ and

$$\partial\|\mathbf{x}\|_{p,q}^q=\prod_{\mathsf{i}\in\mathsf{G}}\partial(\phi\circ g)(\mathbf{x}_\mathsf{i}).$$

The objective function $\mathcal{E}$ in (1.1) reads

$$\mathcal{E}(\mathbf{x})=\sum_{\mathsf{i}\in\mathsf{G}}\phi(\|\mathbf{x}_\mathsf{i}\|_p)+F_r(\mathbf{x}),\ p\geq 1,\ 1\leq r\leq\infty, \tag{3.2}$$

which is bounded below, coercive, and continuous. It has at least one minimizer.

Now, we give the subdifferential of $\mathcal{E}$ at $\mathbf{x}$. From the Remark of Lemma 3.2 and the convexity of $F_r(\mathbf{x})$ with $1\leq r\leq\infty$, we clearly see

$$\begin{aligned}\partial\mathcal{E}(\mathbf{x})&=\partial\sum_{\mathsf{i}\in\mathsf{G}}\phi(\|\mathbf{x}_\mathsf{i}\|_p)+\partial F_r(\mathbf{x})\\&=\prod_{\mathsf{i}\in\mathsf{G}}\partial(\phi\circ g)(\mathbf{x}_\mathsf{i})+\begin{cases}\frac{1}{\alpha r}\mathbf{A}^T\partial\|\mathbf{v}\|_r^r|_{\mathbf{v}=\mathbf{Ax}-\mathbf{y}},\ r\geq 1,\\ \frac{1}{\alpha}\mathbf{A}^T\partial\|\mathbf{v}\|_\infty|_{\mathbf{v}=\mathbf{Ax}-\mathbf{y}},\ r=\infty.\end{cases}\end{aligned} \tag{3.3}$$

Therein the factors can be calculated by Lemma 3.1. The subdifferential of the infinity norm is calculated by the Danskin-Bertsekas Theorem in [4, Proposition A.22] as follows

$$\partial\|\mathbf{h}\|_\infty=\{\mathbf{u}\in\mathbb{R}^M\ |\ \|\mathbf{u}\|_1\leq 1,\mathbf{h}^T\mathbf{u}=\|\mathbf{h}\|_\infty\}. \tag{3.4}$$

For any given $\mathbf{x}$, it is obvious that $\partial F_r(\mathbf{x})$ is a bounded set.

We call $\mathbf{x}^*$ a stationary point of (1.1) if and only if

$$\mathbf{0}\in\partial\mathcal{E}(\mathbf{x}^*). \tag{3.5}$$

3.2. A motivating proposition. The following proposition presents a motivation to design the algorithm in the next section.

PROPOSITION 3.1. *Suppose $\mathbf{x}\in\mathbb{R}^N$ be with the group structure $\mathbf{x}:=\left(\mathbf{x}_1^T,\mathbf{x}_2^T,\cdots,\mathbf{x}_\mathsf{g}^T\right)^T$. For any local minimizer (or stationary point) $\mathbf{x}^*$ of (1.1) sufficiently close to $\mathbf{x}$, it holds that*

$$\mathbf{x}_\mathsf{i}^*=\mathbf{0},\ \forall\mathsf{i}\in\mathsf{G}\setminus\mathrm{supp}_\mathsf{G}(\mathbf{x}). \tag{3.6}$$

Proof. As $\mathbf{x}^*$ is a local minimizer (or a stationary point) of $\mathcal{E}$, we have $\mathbf{0}\in\partial\mathcal{E}(\mathbf{x}^*)$.

We prove (3.6) by contradiction. Assume that there exists $\mathsf{i}'\in\mathsf{G}\setminus\mathrm{supp}_\mathsf{G}(\mathbf{x})$ such that $\mathbf{x}_{\mathsf{i}'}^*\neq\mathbf{0}$. That is, $\mathbf{x}_{\mathsf{i}',j}^*\neq 0$ for some $j\in J_{\mathsf{i}'}$.

From (3.5), we have

$$0=\phi'(\|\mathbf{x}_{\mathsf{i}'}^*\|_p)\,\|\mathbf{x}_{\mathsf{i}'}^*\|_p^{1-p}\,|\mathbf{x}_{\mathsf{i}',j}^*|^{p-1}\,\mathrm{sgn}(\mathbf{x}_{\mathsf{i}',j}^*)+\boldsymbol{\eta}_{\mathsf{i}',j}(\mathbf{x}^*), \tag{3.7}$$

where $\boldsymbol{\eta}_{\mathsf{i}',j}(\mathbf{x}^*)$ is the entry of a specific subgradient $\boldsymbol{\eta}(\mathbf{x}^*)\in\partial F_r(\mathbf{x}^*)$. It follows that

$$q\|\mathbf{x}_{\mathsf{i}'}^*\|_p^{q-1}\leq q\|\mathbf{x}_{\mathsf{i}'}^*\|_p^{q-p}\,\|\mathbf{x}_{\mathsf{i}'}^*\|_{p-1}^{p-1}=\sum_{j\in\mathrm{supp}(\mathbf{x}_{\mathsf{i}'}^*)}\left|\boldsymbol{\eta}_{\mathsf{i}',j}(\mathbf{x}^*)\right|. \tag{3.8}$$

Here the left inequality is due to Lemma 2.1 for $p>1$ and $\|\mathbf{x}_{\mathsf{i}'}^*\|_{p-1}^{p-1}=\#\{\text{nonzero entries of } \mathbf{x}_{\mathsf{i}'}^*\}$ for $p=1$. The right-hand side of (3.8) is uniformly bounded in any predefined neighborhood of $\mathbf{x}$, by (3.3) and (3.4). Since $\mathbf{x}_{\mathsf{i}'}^*$ can be sufficiently close to $\mathbf{x}_{\mathsf{i}'}=0, \mathsf{i}'\in\mathsf{G}\setminus\mathrm{supp}_{\mathsf{G}}(\mathbf{x})$, it contradicts (3.8) by $0<q<1$. ∎

REMARK 3.2. For the special case $r=2$ in the fidelity term, [14, 22] established the lower bound theory, which can also inspire our proposition.

3.3. Algorithm. Motivated by Proposition 3.1, we propose to solve the problem (1.1) by an iterative process, which generates a sequence with a shrinking group support set. Suppose that $\mathbf{x}^{(l)}$ is an approximate solution in the l-th iteration. In the next iteration, we minimize the objective function only on the group support set $\mathsf{S}^{(l)}$ of $\mathbf{x}^{(l)}$, with the remaining variables being zero. This yields the iterative support shrinking algorithm (ISSA). Given $\mathbf{x}^{(l)}$, we compute $\mathbf{x}^{(l+1)}$ by solving

$$\min_{\mathbf{x}_{\mathsf{S}^{(l)}}} \mathcal{E}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}) := \sum_{\mathsf{i}\in\mathsf{S}^{(l)}} \phi(\|\mathbf{x}_{\mathsf{i}}\|_p) + F_r^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}), \qquad (\mathcal{P}^{(l)})$$

and setting

$$\mathbf{x}_{\mathsf{i}}^{(l+1)} = \mathbf{0}, \text{ for } \mathsf{i}\in\mathsf{G}\setminus\mathsf{S}^{(l)},$$

where $F_r^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}})$ reads as follows,

$$F_r^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}) = \begin{cases} \dfrac{1}{r\alpha}\displaystyle\sum_{k\in I}\left|\sum_{\mathsf{i}\in\mathsf{S}^{(l)}} A_{k,\mathsf{i}}\mathbf{x}_{\mathsf{i}} - y_k\right|^r, & r\geq 1, \\ \dfrac{1}{\alpha}\displaystyle\max_{k\in I}\left|\sum_{\mathsf{i}\in\mathsf{S}^{(l)}} A_{k,\mathsf{i}}\mathbf{x}_{\mathsf{i}} - y_k\right|, & r=\infty. \end{cases}$$

The following proposition tells us that if we can find a local minimizer of $(\mathcal{P}^{(l)})$ by ISSA, then we can construct a local minimizer of the original problem (1.1). This provides the rationality for ISSA in some sense.

PROPOSITION 3.2. *Suppose that $\mathbf{x}_{\mathsf{S}^{(l)}}^*$ is a local minimizer of $(\mathcal{P}^{(l)})$. Then its zero padding $\mathbf{z}^*$ by*

$$\mathbf{z}_{\mathsf{i}}^* = \begin{cases} \mathbf{x}_{\mathsf{i}}^*, & \mathsf{i}\in\mathsf{S}^{(l)}, \\ \mathbf{0}, & \mathsf{i}\in\mathsf{G}\setminus\mathsf{S}^{(l)}, \end{cases}$$

is a local minimizer of (1.1).

Proof. There exists $\delta_1>0$, such that

$$\mathcal{E}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}) \geq \mathcal{E}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^*), \quad \forall \mathbf{x}_{\mathsf{S}^{(l)}} \in B(\mathbf{x}_{\mathsf{S}^{(l)}}^*, \delta_1). \tag{3.9}$$

Let $\mathbf{x}\in B(\mathbf{z}^*,\delta_2)$ with $\delta_2<\delta_1$. From (3.9), we have

$$\begin{aligned} \mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{z}^*) &= \sum_{\mathsf{i}\in\mathsf{S}^{(l)}} \phi(\|\mathbf{x}_{\mathsf{i}}\|_p) + \sum_{\mathsf{i}\in\mathsf{G}\setminus\mathsf{S}^{(l)}} \phi(\|\mathbf{x}_{\mathsf{i}}\|_p) + F_r(\mathbf{x}) - \mathcal{E}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^*) \\ &\geq \sum_{\mathsf{i}\in\mathsf{G}\setminus\mathsf{S}^{(l)}} \phi(\|\mathbf{x}_{\mathsf{i}}\|_p) + F_r(\mathbf{x}) - F_r^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}). \end{aligned} \tag{3.10}$$

Let $\mathbf{z}$ be the zero padding of $\mathbf{x}_{\mathsf{S}^{(l)}}$. From the convexity of $F_r(\mathbf{x})$, we can obtain

$$\begin{aligned} F_r(\mathbf{x}) - F_r^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}) = F_r(\mathbf{x}) - F_r(\mathbf{z}) &\geq \langle \boldsymbol{\eta}, \mathbf{x} - \mathbf{z} \rangle \\ &\geq -M \sum_{\mathsf{i} \in \mathsf{G} \setminus \mathsf{S}^{(l)}} \|\mathbf{x}_{\mathsf{i}}\|_2 \\ &\geq -CM \sum_{\mathsf{i} \in \mathsf{G} \setminus \mathsf{S}^{(l)}} \|\mathbf{x}_{\mathsf{i}}\|_p, \; p \geq 1, \end{aligned}$$

where $\boldsymbol{\eta} \in \partial F_r(\mathbf{z})$ and $\|\boldsymbol{\eta}\|_2 \leq M$ uniformly since $\mathbf{z} \in B(\mathbf{z}^*, \delta_2)$. $C > 0$ is the constant for norm equivalence. Hence, using the concavity of ϕ in Proposition 2.1, we obtain

$$\mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{z}^*) \geq \sum_{\mathsf{i} \in \mathsf{G} \setminus \mathsf{S}^{(l)}} [\phi(\|\mathbf{x}_{\mathsf{i}}\|_p) - CM\|\mathbf{x}_{\mathsf{i}}\|_p] \geq \sum_{\mathsf{i} \in \mathsf{G} \setminus \mathsf{S}^{(l)} \cap \mathrm{supp}_{\mathsf{G}}(\mathbf{x})} [\phi'(\|\mathbf{x}_{\mathsf{i}}\|_p) - CM]\|\mathbf{x}_{\mathsf{i}}\|_p.$$

There exists $\delta_3 > 0$ such that $\phi'(\|\mathbf{x}_{\mathsf{i}}\|_p) \geq CM$ for all $\mathsf{i} \in \mathsf{G} \setminus \mathsf{S}^{(l)} \cap \mathrm{supp}_{\mathsf{G}}(\mathbf{x})$ when $\mathbf{x} \in B(\mathbf{z}^*, \delta_3)$.

Let $\delta = \min\{\delta_1, \delta_2, \delta_3\}$. When $\mathbf{x} \in B(\mathbf{z}^*, \delta)$, $\mathcal{E}(\mathbf{x}) \geq \mathcal{E}(\mathbf{z}^*)$. ∎

To make the problem ($\mathcal{P}^{(l)}$) more practical, we linearize $\phi(\|\mathbf{x}_{\mathsf{i}}\|_p), \mathsf{i} \in \mathsf{S}^{(l)}$ at $\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p \neq 0$ in the objective function. With an additional proximal term, we introduce the following energy functional:

$$\mathcal{H}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}) = \sum_{\mathsf{i} \in \mathsf{S}^{(l)}} \left[\phi(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p) + \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p) \left(\|\mathbf{x}_{\mathsf{i}}\|_p - \|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p \right) \right] + F_r^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}) + \frac{\beta}{2} \|\mathbf{x}_{\mathsf{S}^{(l)}} - \mathbf{x}_{\mathsf{S}^{(l)}}^{(l)}\|_2^2, \tag{3.11}$$

where $\beta \geq 0$. We now present an inexact iterative support shrinking algorithm with proximal linearization to solve (1.1).

InISSAPL-GSO: Inexact Iterative Support Shrinking Algorithm with Proximal Linearization for Group Sparse Optimization

Initialization: Select $\mathbf{x}^{(0)} = c\mathbb{1}$ with $c \neq 0$ or randomly, where $\mathbb{1}$ is the all-one vector.
Iteration: For $l = 0, 1, \ldots$ until convergence:

(1) Set $\mathsf{S}^{(l)} = \mathrm{supp}_{\mathsf{G}}(\mathbf{x}^{(l)})$. Set $\beta = 0$ for $l = 0$ and $\beta > 0$ fixed for $l \geq 1$.

(2) Compute $\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)}$ by approximately solving

$$\min_{\mathbf{x}_{\mathsf{S}^{(l)}}} \mathcal{H}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}) \tag{$\mathcal{P}_x^{(l)}$}$$

such that

$$\mathbf{u}_{\mathsf{S}^{(l)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)}) \in \partial \mathcal{H}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)}), \|\mathbf{u}_{\mathsf{S}^{(l)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)})\|_2 \leq \frac{\beta}{2}\nu \|\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)} - \mathbf{x}_{\mathsf{S}^{(l)}}^{(l)}\|_2. \tag{3.12}$$

with the tolerance error $\nu < 1$.

(3) Set

$$\mathbf{x}_{\mathsf{i}}^{(l+1)} = \mathbf{0}, \text{ for } \mathsf{i} \in \mathsf{G} \setminus \mathsf{S}^{(l)}.$$

REMARK 3.3. The condition (3.12) in InISSAPL-GSO is motivated by [2, 26]. It corresponds to an inexact inner loop and a guide to select the approximate solution for $(\mathcal{P}_x^{(l)})$. Due to the strong convexity of the problem $(\mathcal{P}_x^{(l)})$, it can be solved to any given accuracy. Therefore, the condition (3.12) in InISSAPL-GSO can be satisfied, as long as the problem $(\mathcal{P}_x^{(l)})$ is solved sufficiently accurately.

REMARK 3.4. By the motivating Proposition 3.1, $\mathbf{x}^{(0)}$ is required to be with as large a group support as possible. There are two strategies to choose this starting point. One is to set $\mathbf{x}^{(0)}$ by a nonzero scalar multiplication of the all-one vector, which yields a group lasso for $p=2$ in the first step. The other is to set $\mathbf{x}^{(0)}$ by randomly generating data of i.i.d. Gaussian (with zero probability to obtain zero group member), indicating a weighted group lasso when $p=2$. Due to the fact that $\mathbf{x}^{(0)}$ is not the proximal solution, we also set $\beta=0$ in the first step of the algorithm. The results of experiments with these two kinds of starting points are given in Section 6.1.

REMARK 3.5. This support shrinking strategy is related to but different from the active set methods [31].

REMARK 3.6. In fact, our algorithm can be regarded as a variant of iteratively reweighted ℓ_1 minimization. We mention that, a recent parallel manuscript [20] presents to use iteratively reweighted least square (IRLS) algorithms for the special group sparse $\ell_{2,q}-\ell_2$ model, after studying its sparse recovery property. Obviously the approach in [20] cannot be applied to our general model (1.1).

For the convenience of later description, we represent the subgradient in (3.12) as follows,

$$\mathbf{u}_{\mathsf{S}^{(l)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}})=\boldsymbol{\zeta}_{\mathsf{S}^{(l)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}})+\boldsymbol{\eta}_{\mathsf{S}^{(l)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}})+\beta(\mathbf{x}_{\mathsf{S}^{(l)}}-\mathbf{x}_{\mathsf{S}^{(l)}}^{(l)}), \tag{3.13}$$

where

$$\boldsymbol{\zeta}_{\mathsf{S}^{(l)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}})\in\prod_{\mathsf{i}\in\mathsf{S}^{(l)}}\phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)\partial\|\mathbf{x}_{\mathsf{i}}\|_p,\ \boldsymbol{\eta}_{\mathsf{S}^{(l)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}})\in\partial F_r^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}).$$

Each entry of $\boldsymbol{\zeta}_{\mathsf{S}^{(l)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}})$ has the form by

$$\zeta_{\mathsf{i},j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}})=\begin{cases}\phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)\|\mathbf{x}_{\mathsf{i}}\|_p^{1-p}|\mathbf{x}_{\mathsf{i},j}|^{p-1}\operatorname{sgn}(\mathbf{x}_{\mathsf{i},j}), & p>1,\\ \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_1)\operatorname{sgn}(\mathbf{x}_{\mathsf{i},j}), & p=1,\ \text{and}\ j\in\operatorname{supp}(\mathbf{x}_{\mathsf{i}}),\\ \in[-\phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_1),\phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_1)], & p=1,\ \text{and}\ j\notin\operatorname{supp}(\mathbf{x}_{\mathsf{i}}).\end{cases} \tag{3.14}$$

Since $\mathbf{x}^{(l+1)}$ has group support set $\mathsf{S}^{(l)}$, we also clarify the relation of the subgradients in $\partial F_r^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)})$ and $\partial F_r(\mathbf{x}^{(l+1)})$. Indeed, for any $\boldsymbol{\eta}_{\mathsf{S}^{(l)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)})\in\partial F_r^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)})$, there exists $\mathbf{w}$ in $\frac{1}{\alpha r}\partial\|\mathbf{v}\|_r^r$ or $\frac{1}{\alpha}\partial\|\mathbf{v}\|_\infty$ with $\mathbf{v}=\mathbf{A}_{\mathsf{S}^{(l)}}\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)}-\mathbf{y}$ such that

$$\boldsymbol{\eta}_{\mathsf{S}^{(l)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)})=\mathbf{A}_{\mathsf{S}^{(l)}}^T\mathbf{w}. \tag{3.15}$$

Noting that $\mathbf{A}\mathbf{x}^{(l+1)}-\mathbf{y}=\mathbf{A}_{\mathsf{S}^{(l)}}\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)}-\mathbf{y}$, there exists a subgradient $\boldsymbol{\eta}(\mathbf{x}^{(l+1)})\in\partial F_r(\mathbf{x}^{(l+1)})$ with

$$\boldsymbol{\eta}(\mathbf{x}^{(l+1)})=\mathbf{A}^T\mathbf{w}. \tag{3.16}$$

4. Convergence analysis

In this section, we establish the global convergence result of the sequence generated by InISSAPL-GSO. Nowadays, a celebrating theoretical framework developed in [2] for convergence analysis of descent methods, has been extensively applied in non-convex optimization [29, 32–34, 36, 45]. We state the main result by Theorem B.2 in the Appendix B. It is required that the iterative sequence satisfies *(H1)-(H3)* in this theorem. To verify the relative error condition *(H2)*, we face two difficulties. The first one is the non-Lipschitz property of $\mathcal{E}(\mathbf{x})$, which is solved by proving a lower bound theory on the sequence. The second one is the highly non-smooth composition of ℓ_q $(0<q<1)$ and ℓ_1, which requires a technical construction of the subgradient in $\partial\mathcal{E}(\mathbf{x})$. See the details in Lemma 4.3 and Lemma 4.4.

We firstly prove the sufficient decrease property about $\mathcal{E}(\mathbf{x}^{(l)})$ in Lemma 4.1. To this end, we introduce, for each index l, the following intermediate energy functional on $\mathbf{x}\in\mathbb{R}^N$:

$$\widetilde{\mathcal{H}}^{(l)}(\mathbf{x})=\sum_{\mathsf{i}\in\mathsf{S}^{(l)}}\left[\phi(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)+\phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)\left(\|\mathbf{x}_{\mathsf{i}}\|_p-\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p\right)\right]+F_r(\mathbf{x})+\frac{\beta}{2}\|\mathbf{x}-\mathbf{x}^{(l)}\|_2^2. \tag{4.1}$$

LEMMA 4.1. *Let* $\{\mathbf{x}^{(l)}\}$ *be a sequence generated by* InISSAPL-GSO. *Then*
(1) The sequence $\{\mathcal{E}(\mathbf{x}^{(l)})\}$ *is nonincreasing and satisfies*

$$\mathcal{E}(\mathbf{x}^{(l+1)})+\frac{\beta}{2}(1-\nu)\|\mathbf{x}^{(l+1)}-\mathbf{x}^{(l)}\|_2^2\leq\mathcal{E}(\mathbf{x}^{(l)}), \tag{4.2}$$

where $\beta>0$ *and* $0\leq\nu<1$.
(2) The sequence $\{\mathbf{x}^{(l)}\}$ *is bounded and satisfies* $\lim_{l\to\infty}\|\mathbf{x}^{(l+1)}-\mathbf{x}^{(l)}\|_2=0$.

Proof. Since $\mathbf{x}^{(l)}$ has the group support set $\mathsf{S}^{(l)}$, we have

$$\widetilde{\mathcal{H}}^{(l)}(\mathbf{x}^{(l)})=\mathcal{E}(\mathbf{x}^{(l)}). \tag{4.3}$$

At the $(l+1)$-th iteration, we obviously obtain

$$\begin{aligned}\widetilde{\mathcal{H}}^{(l)}(\mathbf{x}^{(l+1)})&=\sum_{\mathsf{i}\in\mathsf{S}^{(l)}}\left[\phi(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)+\phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)\left(\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_p-\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p\right)\right]\\&\quad+F_r(\mathbf{x}^{(l+1)})+\frac{\beta}{2}\|\mathbf{x}^{(l+1)}-\mathbf{x}^{(l)}\|_2^2\\ [\text{ by (2.1) }]\ &\geq\sum_{\mathsf{i}\in\mathsf{S}^{(l)}}\phi(\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_p)+F_r(\mathbf{x}^{(l+1)})+\frac{\beta}{2}\|\mathbf{x}^{(l+1)}-\mathbf{x}^{(l)}\|_2^2\\&=\mathcal{E}(\mathbf{x}^{(l+1)})+\frac{\beta}{2}\|\mathbf{x}^{(l+1)}-\mathbf{x}^{(l)}\|_2^2.\end{aligned} \tag{4.4}$$

By (3.15)-(3.16), for any element $\boldsymbol{\eta}_{\mathsf{S}^{(l)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)})\in\partial F_r^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)})$, there exists a subgradient $\boldsymbol{\eta}(\mathbf{x}^{(l+1)})\in\partial F_r(\mathbf{x}^{(l+1)})$ with

$$\boldsymbol{\eta}_{\mathsf{i}}(\mathbf{x}^{(l+1)})=\boldsymbol{\eta}_{\mathsf{i}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)}),\ \mathsf{i}\in\mathsf{S}^{(l)}.$$

Now we define $\widetilde{\mathbf{u}}^{(l)}(\mathbf{x}^{(l+1)})$ by

$$\widetilde{\mathbf{u}}_{\mathsf{i}}^{(l)}(\mathbf{x}^{(l+1)})=\begin{cases}\mathbf{u}_{\mathsf{i}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)}), & \mathsf{i}\in\mathsf{S}^{(l)},\\ \boldsymbol{\eta}_{\mathsf{i}}(\mathbf{x}^{(l+1)}), & \mathsf{i}\in\mathsf{G}\setminus\mathsf{S}^{(l)},\end{cases} \tag{4.5}$$

where $\mathbf{u}_{\mathsf{i}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)})$ is given in (3.13). Then $\widetilde{\mathbf{u}}^{(l)}(\mathbf{x}^{(l+1)}) \in \partial\widetilde{\mathcal{H}}^{(l)}(\mathbf{x}^{(l+1)})$.

Since for any $\mathsf{i} \in \mathsf{G} \setminus \mathsf{S}^{(l)}$, $\mathbf{x}_{\mathsf{i}}^{(l+1)} = \mathbf{x}_{\mathsf{i}}^{(l)} = \mathbf{0}$, we have

$$\begin{aligned}
\langle \widetilde{\mathbf{u}}^{(l)}(\mathbf{x}^{(l+1)}), \mathbf{x}^{(l)} - \mathbf{x}^{(l+1)} \rangle &= \sum_{\mathsf{i}\in\mathsf{S}^{(l)}} \sum_{j\in J_{\mathsf{i}}} \mathbf{u}_{\mathsf{i},j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)})(\mathbf{x}_{\mathsf{i},j}^{(l)} - \mathbf{x}_{\mathsf{i},j}^{(l+1)}) \\
&\geq -\|\mathbf{u}_{\mathsf{S}^{(l)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)})\|_2 \cdot \|\mathbf{x}_{\mathsf{S}^{(l)}}^{(l)} - \mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)}\|_2 \\
[\text{ by (3.12) }] \quad &\geq -\frac{\beta}{2}\nu \|\mathbf{x}_{\mathsf{S}^{(l)}}^{(l)} - \mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)}\|_2^2 \\
&= -\frac{\beta}{2}\nu \|\mathbf{x}^{(l)} - \mathbf{x}^{(l+1)}\|_2^2. \qquad (4.6)
\end{aligned}$$

Putting (4.3), (4.4) and (4.6) together, we obtain

$$\begin{aligned}
\mathcal{E}(\mathbf{x}^{(l)}) = \widetilde{\mathcal{H}}^{(l)}(\mathbf{x}^{(l)}) &\geq \widetilde{\mathcal{H}}^{(l)}(\mathbf{x}^{(l+1)}) + \langle \widetilde{\mathbf{u}}^{(l)}(\mathbf{x}^{(l+1)}), \mathbf{x}^{(l)} - \mathbf{x}^{(l+1)} \rangle \\
&\geq \widetilde{\mathcal{H}}^{(l)}(\mathbf{x}^{(l+1)}) - \frac{\beta}{2}\nu \|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\|_2^2 \\
&\geq \mathcal{E}(\mathbf{x}^{(l+1)}) + \frac{\beta}{2}(1-\nu)\|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\|_2^2.
\end{aligned}$$

With the fact that $\mathcal{E}(\mathbf{x})$ is bounded from below and $\frac{\beta}{2}(1-\nu) > 0$, it follows that $\{\mathcal{E}(\mathbf{x}^{(l)})\}$ is nonincreasing and converges to a finite value. Thus

$$\lim_{l\to\infty} \|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\|_2 = 0.$$

Because $\mathcal{E}(\mathbf{x})$ is coercive, we know that $\{\mathbf{x}^{(l)}\}$ is bounded. ◻

The following lemma gives the finite convergence property of the group support set, from which we can deduce a lower bound theory on the nonzero groups of the iteration sequence.

LEMMA 4.2. *The sequence $\{\mathsf{S}^{(l)}\}$ converges in a finite number of iterations, i.e., there exists an integer $L > 0$ such that when $l \geq L$, $\mathsf{S}^{(l)} \equiv \mathsf{S}^{(L)}$.*

Proof. The statement is straightforward by the finiteness of G and $\mathsf{G} \supseteq \mathsf{S}^{(0)} \supseteq \cdots \supseteq \mathsf{S}^{(l)} \supseteq \cdots$, like those in [26, 43, 44]. ◻

REMARK 4.1. From this Lemma, we know that the sequence $\{\mathbf{x}^{(l)}\}$ by InISSAPL-GSO has the fixed group support set $\mathsf{S}^{(L)}$ when $l \geq L$.

LEMMA 4.3. *There exist constants $C > c > 0$ such that*

$$\text{either} \quad \mathbf{x}_{\mathsf{i}}^{(l)} = 0 \quad \text{or} \quad c \leq \|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p \leq C, \forall \mathsf{i} \in \mathsf{G}, \forall l \geq L, \qquad (4.7)$$

where L is defined in Lemma 4.2.

Proof. From Lemma 4.2, for any $\mathsf{i} \in \mathsf{S}^{(L)}$ and $l \geq L$, $\mathbf{x}_{\mathsf{i}}^{(l)} \neq \mathbf{0}$. The sequence has an upper bound from Lemma 4.1,

$$\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p \leq C.$$

We now prove by contradiction that $\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p$ has a positive lower bound for any $\mathsf{i} \in \mathsf{S}^{(L)}, l \geq L$.

Suppose that there exists $\mathsf{i}' \in \mathsf{S}^{(L)}$ for some subsequence $\mathbf{x}^{(l_k)}$, still denoted by $\mathbf{x}^{(l)}$, such that

$$\mathbf{x}_{\mathsf{i}'}^{(l)} \neq \mathbf{0} \text{ and } \lim_{l\to\infty} \mathbf{x}_{\mathsf{i}'}^{(l)} = \mathbf{0}.$$

By the subdifferential expression (3.13), we have, for $j \in \operatorname{supp}(\mathbf{x}_{\mathsf{i}'}^{(l+1)})$,

$$\left|\mathbf{u}_{\mathsf{i}',j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)})\right| + \left|\boldsymbol{\eta}_{\mathsf{i}',j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)})\right| + \beta\left|\mathbf{x}_{\mathsf{i}',j}^{(l+1)} - \mathbf{x}_{\mathsf{i}',j}^{(l)}\right| \geq \left|\boldsymbol{\zeta}_{\mathsf{i}',j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)})\right| \tag{4.8}$$

with the right-hand side satisfying,

$$|\boldsymbol{\zeta}_{\mathsf{i}',j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)})| \geq \phi'(\|\mathbf{x}_{\mathsf{i}'}^{(l)}\|_p)\cdot\|\mathbf{x}_{\mathsf{i}'}^{(l+1)}\|_p^{1-p}\cdot|\mathbf{x}_{\mathsf{i}',j}^{(l+1)}|^{p-1}.$$

Summing up all the terms for $j \in \operatorname{supp}(\mathbf{x}_{\mathsf{i}'}^{(l+1)})$, we have

$$\begin{aligned}
&\sum_{j\in\operatorname{supp}(\mathbf{x}_{\mathsf{i}'}^{(l+1)})} \left|\mathbf{u}_{\mathsf{i}',j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)})\right| + \left|\boldsymbol{\eta}_{\mathsf{i}',j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)})\right| + \beta\left|\mathbf{x}_{\mathsf{i}',j}^{(l+1)} - \mathbf{x}_{\mathsf{i}',j}^{(l)}\right| \\
&\geq \phi'(\|\mathbf{x}_{\mathsf{i}'}^{(l)}\|_p)\cdot\|\mathbf{x}_{\mathsf{i}'}^{(l+1)}\|_p^{1-p}\cdot\|\mathbf{x}_{\mathsf{i}'}^{(l+1)}\|_{p-1}^{p-1} \\
&\geq \phi'(\|\mathbf{x}_{\mathsf{i}'}^{(l)}\|_p) \\
&= q\|\mathbf{x}_{\mathsf{i}'}^{(l)}\|_p^{q-1},
\end{aligned}$$

where the second inequality holds from the same reason as in (3.8). It follows from the boundedness of $\{\mathbf{x}^{(l)}\}$ that $\left|\boldsymbol{\eta}_{\mathsf{i}',j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)})\right| + \beta\left|\mathbf{x}_{\mathsf{i}',j}^{(l+1)} - \mathbf{x}_{\mathsf{i}',j}^{(l)}\right|$ is bounded. The condition (3.12) implies that $\left|\mathbf{u}_{\mathsf{i}',j}^{(l)}(\mathbf{x}^{(l+1)})\right|$ is also bounded. Thus the Equation (4.8) is impossible to hold when $l \to \infty$ because of $0 < q < 1$. ◻

REMARK 4.2. By combining Lemma 4.3 and Proposition 2.1, we can obtain the Lipschitz property of ϕ' over the group support:

$$\begin{aligned}
\left|\phi'(\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_p) - \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)\right| &\leq L_c\left|\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_p - \|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p\right| \\
&\leq L_c\|\mathbf{x}_{\mathsf{i}}^{(l+1)} - \mathbf{x}_{\mathsf{i}}^{(l)}\|_p,\ \mathsf{i}\in\mathsf{S}^{(L)},\ l \geq L.
\end{aligned} \tag{4.9}$$

Using this property, we can prove the relative error condition in Lemma 4.4 through a sophisticated construction of $\mathbf{v}^{(l+1)} \in \partial\mathcal{E}(\mathbf{x}^{(l+1)})$.

LEMMA 4.4. *For each $l \geq L$, there exists $\mathbf{v}^{(l+1)} \in \partial\mathcal{E}(\mathbf{x}^{(l+1)})$ such that*

$$\|\mathbf{v}^{(l+1)}\|_2 \leq \bar{C}\|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\|_2, \tag{4.10}$$

where the constant $\bar{C} > 0$ is independent of l.

Proof. For $l \geq L$, by (3.13), the subgradient $\mathbf{u}_{\mathsf{S}^{(L)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)})$ in $\partial\mathcal{H}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)})$ can be written as

$$\mathbf{u}_{\mathsf{S}^{(L)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}) = \boldsymbol{\zeta}_{\mathsf{S}^{(L)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}) + \boldsymbol{\eta}_{\mathsf{S}^{(L)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}) + \beta(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)} - \mathbf{x}_{\mathsf{S}^{(L)}}^{(l)}).$$

For the above $\boldsymbol{\eta}_{\mathsf{S}^{(L)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)})$, by (3.15)-(3.16), there exists a subgradient $\boldsymbol{\eta}(\mathbf{x}^{(l+1)})$ in $\partial F_r(\mathbf{x}^{(l+1)})$ which satisfies

$$\boldsymbol{\eta}_{\mathsf{i}}(\mathbf{x}^{(l+1)}) = \boldsymbol{\eta}_{\mathsf{i}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}),\ \mathsf{i}\in\mathsf{S}^{(L)}.$$

We introduce the intermediate variable $\widetilde{\mathbf{v}}^{(l+1)}$ by

$$\widetilde{\mathbf{v}}_{\mathsf{i}}^{(l+1)} = \begin{cases} \boldsymbol{\zeta}_{\mathsf{i}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}) + \boldsymbol{\eta}_{\mathsf{i}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}), & \mathsf{i} \in \mathsf{S}^{(L)}, \\ \mathbf{0}, & \mathsf{i} \in \mathsf{G} \setminus \mathsf{S}^{(L)}. \end{cases} \tag{4.11}$$

It can be measured by the iterative error,

$$\begin{aligned} \|\widetilde{\mathbf{v}}^{(l+1)}\|_2 &= \|\mathbf{u}_{\mathsf{S}^{(L)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}) - \beta(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)} - \mathbf{x}_{\mathsf{S}^{(L)}}^{(l)})\|_2 \\ &\le \|\mathbf{u}_{\mathsf{S}^{(L)}}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)})\|_2 + \beta\|\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)} - \mathbf{x}_{\mathsf{S}^{(L)}}^{(l)}\|_2 \\ [\text{ by } (3.12)\ ] \quad &\le \frac{\beta}{2}(\nu+2)\|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\|_2. \end{aligned} \tag{4.12}$$

The objective is to construct $\mathbf{v}^{(l+1)} \in \partial\mathcal{E}(\mathbf{x}^{(l+1)})$ satisfying the theorem statement. We can choose

$$\mathbf{v}^{(l+1)} = \boldsymbol{\zeta}(\mathbf{x}^{(l+1)}) + \boldsymbol{\eta}(\mathbf{x}^{(l+1)}), \tag{4.13}$$

where $\boldsymbol{\zeta}(\mathbf{x}^{(l+1)}) \in \prod_{\mathsf{i}\in\mathsf{G}} \partial(\phi\circ g)(\mathbf{x}_{\mathsf{i}}^{(l+1)})$ is to be determined in the following.

By Lemma 3.1, for $\mathsf{i} \in \mathsf{S}^{(L)}$, each entry of $\boldsymbol{\zeta}_{\mathsf{i}}(\mathbf{x}^{(l+1)}) \in \partial(\phi\circ g)(\mathbf{x}_{\mathsf{i}}^{(l+1)})$ is given by

$$\begin{cases} \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_p)\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_p^{1-p}|\mathbf{x}_{\mathsf{i},j}^{(l+1)}|^{p-1}\cdot \operatorname{sgn}(\mathbf{x}_{\mathsf{i},j}^{(l+1)}), & p>1, \\ \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_1)\operatorname{sgn}(\mathbf{x}_{\mathsf{i},j}^{(l+1)}), & p=1, j\in \operatorname{supp}(\mathbf{x}_{\mathsf{i}}^{(l+1)}), \\ \psi_{\mathsf{i},j}, & p=1, j\notin \operatorname{supp}(\mathbf{x}_{\mathsf{i}}^{(l+1)}), \end{cases} \tag{4.14}$$

where $\psi_{\mathsf{i},j}$ satisfies

$$\psi_{\mathsf{i},j} \in I^{(l+1)} := [-q\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_1^{q-1}, q\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_1^{q-1}].$$

The choice of $\psi_{\mathsf{i},j}$ is technical. In order to estimate the ℓ_1 error of $\mathbf{v}^{(l+1)}$ and $\widetilde{\mathbf{v}}^{(l+1)}$ in (4.17), the key idea of designing $\psi_{\mathsf{i},j}$ is to find the nearest point in $I^{(l+1)}$ from the point $\boldsymbol{\zeta}_{\mathsf{i},j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}^{(l+1)})$ in the set $[-q\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_1^{q-1}, q\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_1^{q-1}]$ by (4.11). That is,

$$\psi_{\mathsf{i},j} = \begin{cases} \boldsymbol{\zeta}_{\mathsf{i},j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}), & \text{if } \boldsymbol{\zeta}_{\mathsf{i},j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}) \in I^{(l+1)}; \\ -q\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_1^{q-1}, & \text{if } \boldsymbol{\zeta}_{\mathsf{i},j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}) \in \left[-q\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_1^{q-1}, -q\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_1^{q-1}\right); \\ q\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_1^{q-1}, & \text{if } \boldsymbol{\zeta}_{\mathsf{i},j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}) \in \left(q\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_1^{q-1}, q\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_1^{q-1}\right]. \end{cases} \tag{4.15}$$

For $\mathsf{i} \in \mathsf{G}\setminus\mathsf{S}^{(L)}$, since $\partial(\phi\circ g)(\mathbf{x}_{\mathsf{i}}^{(l+1)}) = \Pi_{j\in J_{\mathsf{i}}}(-\infty,+\infty)$, we can choose $\boldsymbol{\zeta}_{\mathsf{i}}(\mathbf{x}^{(l+1)}) = -\boldsymbol{\eta}_{\mathsf{i}}(\mathbf{x}^{(l+1)})$. Naturally, $\mathbf{v}_{\mathsf{i}}^{(l+1)} = \mathbf{0}$ when $\mathsf{i} \in \mathsf{G}\setminus\mathsf{S}^{(L)}$.

Now we measure the difference between $\mathbf{v}^{(l+1)}$ and $\widetilde{\mathbf{v}}^{(l+1)}$. We divide this into two cases: $p>1$ and $p=1$. For $p>1$, the ℓ_1 norm of the difference can be bounded as follows

$$\begin{aligned} \|\mathbf{v}^{(l+1)} - \widetilde{\mathbf{v}}^{(l+1)}\|_1 &= \sum_{\mathsf{i}\in\mathsf{S}^{(L)}} \sum_{j\in J_{\mathsf{i}}} \left|\boldsymbol{\zeta}_{\mathsf{i},j}(\mathbf{x}^{(l+1)}) - \boldsymbol{\zeta}_{\mathsf{i},j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)})\right| \\ &= \sum_{\mathsf{i}\in\mathsf{S}^{(L)}} \sum_{j\in J_{\mathsf{i}}} \left|\phi'(\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_p) - \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)\right| \cdot \|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_p^{1-p} \cdot \left|\mathbf{x}_{\mathsf{i},j}^{(l+1)}\right|^{p-1} \\ &= \sum_{\mathsf{i}\in\mathsf{S}^{(L)}} \left|\phi'(\|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_p) - \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)\right| \cdot \|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_p^{1-p} \cdot \|\mathbf{x}_{\mathsf{i}}^{(l+1)}\|_{p-1}^{p-1} \end{aligned}$$

$$
\begin{aligned}
[\text{ by (4.9), (2.4) }] \leq & L_c \cdot \widetilde{C}_{p-1} \sum_{\mathsf{i}\in \mathsf{S}^{(L)}} \|\mathbf{x}_\mathsf{i}^{(l+1)} - \mathbf{x}_\mathsf{i}^{(l)}\|_p \cdot \|\mathbf{x}_\mathsf{i}^{(l+1)}\|_p^{1-p} \cdot \|\mathbf{x}_\mathsf{i}^{(l+1)}\|_p^{p-1} \\
\leq & L_c \cdot \widetilde{C}_{p-1} \cdot C \|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\|_2,
\end{aligned} \tag{4.16}
$$

where C is the coefficient of norm equivalence. For $p=1$, it follows,

$$
\begin{aligned}
\|\mathbf{v}^{(l+1)} - \widetilde{\mathbf{v}}^{(l+1)}\|_1 = & \sum_{\mathsf{i}\in \mathsf{S}^{(L)}} \sum_{j \in \operatorname{supp}(\mathbf{x}_\mathsf{i}^{(l+1)})} \left| \zeta_{\mathsf{i},j}(\mathbf{x}^{(l+1)}) - \zeta_{\mathsf{i},j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}) \right| \\
& + \sum_{\mathsf{i}\in \mathsf{S}^{(L)}} \sum_{j \notin \operatorname{supp}(\mathbf{x}_\mathsf{i}^{(l+1)})} \left| \psi_{\mathsf{i},j} - \zeta_{\mathsf{i},j}^{(l)}(\mathbf{x}_{\mathsf{S}^{(L)}}^{(l+1)}) \right| \\
\leq & \sum_{\mathsf{i}\in \mathsf{S}^{(L)}} \sum_{j \in \operatorname{supp}(\mathbf{x}_\mathsf{i}^{(l+1)})} \left| \phi'(\|\mathbf{x}_\mathsf{i}^{(l+1)}\|_1) - \phi'(\|\mathbf{x}_\mathsf{i}^{(l)}\|_1) \right| \cdot \left| \operatorname{sgn}(\mathbf{x}_{\mathsf{i},j}^{(l+1)}) \right| \\
& + \sum_{\mathsf{i}\in \mathsf{S}^{(L)}} \sum_{j \notin \operatorname{supp}(\mathbf{x}_\mathsf{i}^{(l+1)})} \left| q\|\mathbf{x}_\mathsf{i}^{(l+1)}\|_1^{q-1} - q\|\mathbf{x}_\mathsf{i}^{(l)}\|_1^{q-1} \right| \\
= & \sum_{\mathsf{i}\in \mathsf{S}^{(L)}} \sum_{j\in J_\mathsf{i}} \left| \phi'(\|\mathbf{x}_\mathsf{i}^{(l+1)}\|_1) - \phi'(\|\mathbf{x}_\mathsf{i}^{(l)}\|_1) \right| \\
\leq & L_c \|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\|_1 \\
\leq & L_c \cdot C \|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\|_2.
\end{aligned} \tag{4.17}
$$

where the first inequality comes from (4.14), (4.15).

Putting (4.12), (4.16) and (4.17) together, we obtain,

$$
\begin{aligned}
\|\mathbf{v}^{(l+1)}\|_2 \leq \|\mathbf{v}^{(l+1)}\|_1 \leq & \|\mathbf{v}^{(l+1)} - \widetilde{\mathbf{v}}^{(l+1)}\|_1 + \sqrt{N}\|\widetilde{\mathbf{v}}^{(l+1)}\|_2 \\
\leq & \bar{C}\|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\|_2,
\end{aligned}
$$

where $\bar{C} = \max\{L_c\widetilde{C}_{p-1}C, L_cC\} + \sqrt{N}\beta(2+\nu)/2$. □

Finally, we can establish the following convergence result.

THEOREM 4.1. *The iterative sequence* $\{\mathbf{x}^{(l)}\}$ *generated by* InISSAPL-GSO *converges globally to the limit point* $\mathbf{x}^*$*, which is a stationary point of problem* (1.1).

Proof. From Appendix B, we know that $\mathcal{E}(\mathbf{x})$ satisfies KL property. Since $\{\mathbf{x}^{(l)}\}$ is bounded and $\mathcal{E}(\mathbf{x})$ is continuous, there exists a subsequence $(\mathbf{x}^{(l_k)})$ and $\mathbf{x}^*$ such that

$$
\mathbf{x}^{(l_k)} \to \mathbf{x}^* \text{ and } \mathcal{E}(\mathbf{x}^{(l_k)}) \to \mathcal{E}(\mathbf{x}^*), \text{as } k \to \infty. \tag{4.18}
$$

By Lemma 4.1, Lemma 4.4, and Theorem B.2 in the Appendix B, the sequence $\{\mathbf{x}^{(l)}\}$ converges globally to the limit point $\mathbf{x}^*$, which is a stationary point of $\mathcal{E}(\mathbf{x})$. □

5. Algorithm implementation

At each iteration step in InISSAPL-GSO, we actually solve a weighted $\ell_{p,1} - \ell_r(\, p \geq 1,\ r \geq 1)$ minimization problem. It is convex and an inexact inner loop is allowed in implementation. Some standard methods like ADMM [8,17], split Bregman method [21, 41] and primal-dual algorithm [11,19] can be used to efficiently solve it. Here we adopt scaled ADMM.

5.1. Scaled ADMM. The problem $(\mathcal{P}_x^{(l)})$ in InISSAPL-GSO is equivalent to

$$\min_{\mathbf{x}_{\mathsf{S}^{(l)}}} \sum_{\mathsf{i}\in\mathsf{S}^{(l)}} \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)\|\mathbf{x}_{\mathsf{i}}\|_p + F_r^{(l)}(\mathbf{x}_{\mathsf{S}^{(l)}}) + \frac{\beta}{2}\|\mathbf{x}_{\mathsf{S}^{(l)}} - \mathbf{x}_{\mathsf{S}^{(l)}}^{(l)}\|_2^2. \tag{5.1}$$

Since it is solved only on the group support, we omit the subscripts $\mathsf{S}^{(l)}$ of the variables for the brevity in the following. With some auxiliary variables, (5.1) is written as

$$\begin{aligned} &\min_{\mathbf{z}} \sum_{\mathsf{i}\in\mathsf{S}^{(l)}} \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)\,\|\mathbf{z}_{\mathsf{i}}\|_p + f_r(\mathbf{s}) + \frac{\beta}{2}\|\mathbf{x}-\mathbf{x}^{(l)}\|_2^2 \\ &\text{s.t. } \mathbf{z}=\mathbf{x},\ \mathbf{s}=\mathbf{A}_{\mathsf{S}^{(l)}}\mathbf{x}-\mathbf{y}, \end{aligned} \tag{5.2}$$

where

$$f_r(\mathbf{s}) = \begin{cases} \frac{1}{r\alpha}\|\mathbf{s}\|_r^r, & r\geq 1, \\ \frac{1}{\alpha}\|\mathbf{s}\|_\infty, & r=\infty. \end{cases}$$

Let $\rho_1,\rho_2>0$ (denoted by $\rho=(\rho_1,\rho_2)$) be the penalty parameters and $\boldsymbol{\lambda},\boldsymbol{\mu}$ be the Lagrangian multipliers. The scaled augmented Lagrangian functional for the weighted problem (5.2) at l-th step is,

$$\begin{aligned} \mathcal{L}_\rho^{(l)}(\mathbf{x},\mathbf{z},\mathbf{s};\boldsymbol{\lambda},\boldsymbol{\mu}) = &\sum_{\mathsf{i}\in\mathsf{T}^{(l)}} \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)\,\|\mathbf{z}_{\mathsf{i}}\|_p + f_r(\mathbf{s}) + \frac{\rho_1}{2}\left(\|\mathbf{A}_{\mathsf{S}^{(l)}}\mathbf{x}-\mathbf{y}-\mathbf{s}+\boldsymbol{\lambda}\|_2^2 - \|\boldsymbol{\lambda}\|_2^2\right) \\ &+\frac{\rho_2}{2}\left(\|\mathbf{x}-\mathbf{z}+\boldsymbol{\mu}\|_2^2 - \|\boldsymbol{\mu}\|_2^2\right) + \frac{\beta}{2}\|\mathbf{x}-\mathbf{x}^{(l)}\|_2^2. \end{aligned}$$

The scaled ADMM for solving (5.2) is described as follows, where we use t as the iteration index.

Scaled ADMM: Scaled Alternating Direction Method of Multipliers for Solving (5.2)

Initialization: Start with $\mathbf{x}^{(l,0)}=\mathbf{x}^{(l)},\boldsymbol{\lambda}^{(l,0)}=\mathbf{0},\boldsymbol{\mu}^{(l,0)}=\mathbf{0}$.
Iteration: For $t=0,1,\ldots,$MAXit and a stopping criterion is not met,
(1) Compute

$$(\mathbf{z}^{(l,t+1)},\mathbf{s}^{(l,t+1)}) = \arg\min_{\mathbf{z},\mathbf{s}} \mathcal{L}_\rho^{(l)}(\mathbf{x}^{(l,t)},\mathbf{z},\mathbf{s};\boldsymbol{\lambda}^{(l,t)},\boldsymbol{\mu}^{(l,t)}). \tag{5.3}$$

(2) Compute

$$\mathbf{x}^{(l,t+1)} = \arg\min_{\mathbf{x}} \mathcal{L}_\rho^{(l)}(\mathbf{x},\mathbf{z}^{(l,t+1)},\mathbf{s}^{(l,t+1)};\boldsymbol{\lambda}^{(l,t)},\boldsymbol{\mu}^{(l,t)}). \tag{5.4}$$

(3) Update

$$\boldsymbol{\lambda}^{(l,t+1)} = \boldsymbol{\lambda}^{(l,t)} + \mathbf{A}_{\mathsf{S}^{(l)}}\mathbf{x}^{(l,t+1)} - \mathbf{y} - \mathbf{s}^{(l,t+1)}, \tag{5.5}$$

$$\boldsymbol{\mu}^{(l,t+1)} = \boldsymbol{\mu}^{(l,t)} + \mathbf{x}^{(l,t+1)} - \mathbf{z}^{(l,t+1)}. \tag{5.6}$$

Output: $\mathbf{x}^{(l+1)}=\mathbf{x}^{(l,t)}$.

REMARK 5.1. There are different choices for the stopping criterions in the scaled ADMM, like (3.12) and the condition given in Section 6. As explained in Remark 3.3, the inner problem is a strongly convex problem and ADMM converges to its global minimizer. We can check (3.12) by using $\mathbf{x}^{(l,t+1)}$ at each iteration in the inner loop. As long as the inner iteration index t is large enough, (3.12) can be guaranteed. However, in real computation, we do not check this condition, since it involves subdifferential calculus of non-convex non-smooth functions, which is time consuming. We instead use the condition given in Section 6. Experiments in it show that this strategy works quite well.

5.2. Solving (5.3) and (5.4). The subproblems (5.3) and (5.4) can be efficiently solved.

(1) The minimization subproblem in (5.3) is equivalent to

$$\min_{\mathbf{z},\mathbf{s}} \sum_{\mathsf{i}\in \mathsf{S}^{(l)}} \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)\|\mathbf{z}_{\mathsf{i}}\|_p + f_r(\mathbf{s}) + \frac{\rho_1}{2}\left\|\mathbf{A}_{\mathsf{S}^{(l)}}\mathbf{x}^{(l,t)} - \mathbf{y} - \mathbf{s} + \boldsymbol{\lambda}^{(l,t)}\right\|_2^2 + \frac{\rho_2}{2}\left\|\mathbf{x}^{(l,t)} - \mathbf{z} + \boldsymbol{\mu}^{(l,t)}\right\|_2^2,$$

which can be separated into two independent subproblems.

(a) $\mathbf{z}$-minimization problem:

$$\min_{\mathbf{z}} \sum_{\mathsf{i}\in \mathsf{S}^{(l)}} \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_p)\|\mathbf{z}_{\mathsf{i}}\|_p + \frac{\rho_2}{2}\left\|\mathbf{x}^{(l,t)} - \mathbf{z} + \boldsymbol{\mu}^{(l,t)}\right\|_2^2.$$

This is a strongly convex problem. There are many well-developed methods to solve it. In the two special cases with $p=1,2$, we actually have explicit solutions shown in the following. These two cases are also the most interesting cases, because, as explained in the introduction, $p=1$ allows within-group sparsity, while $p>1$ does not. For $p>1$, the typical choice is $p=2$, like [22].
Specifically, for $p=1$, we have the unique explicit solution by the shrinkage as

$$\mathbf{z}_{\mathsf{i}}^{(l,t+1)} = \mathcal{S}(\mathbf{x}_{\mathsf{i}}^{(l,t)} + \boldsymbol{\mu}_{\mathsf{i}}^{(l,t)}, \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|)/\rho_2),$$

where

$$\mathcal{S}(\mathbf{y},\gamma) = \operatorname{sgn}(\mathbf{y}) \odot \max\{|\mathbf{y}| - \gamma, 0\}.$$

For $p=2$, its minimizer can be obtained by the multi-dimensional shrinkage, i.e.,

$$\mathbf{z}_{\mathsf{i}}(\mathbf{v}_{\mathsf{i}}) = \max\{\|\mathbf{v}_{\mathsf{i}}\|_2 - \phi'(\|\mathbf{x}_{\mathsf{i}}^{(l)}\|_2)/\rho_2, 0\}\frac{\mathbf{v}_{\mathsf{i}}}{\|\mathbf{v}_{\mathsf{i}}\|_2}, \quad \mathbf{v}_{\mathsf{i}} = \mathbf{x}_{\mathsf{i}}^{(l,t)} + \boldsymbol{\mu}_{\mathsf{i}}^{(l,t)}.$$

(b) $\mathbf{s}$-minimization problem:

$$\min_{\mathbf{s}} f_r(\mathbf{s}) + \frac{\rho_1}{2}\left\|\mathbf{A}_{\mathsf{S}^{(l)}}\mathbf{x}^{(l,t)} - \mathbf{y} - \mathbf{s} + \boldsymbol{\lambda}^{(l,t)}\right\|_2^2.$$

Note that it is separable to some 1D minimization problems.
For $r=1$, it is the same problem as the $\mathbf{z}$-minimization one for $p=1$, and we omit the details here.
For $r=2$, the solution is clearly as follows,

$$\mathbf{s} = \alpha\rho_1/(1+\alpha\rho_1)\left(\mathbf{A}_{\mathsf{S}^{(l)}}\mathbf{x}^{(l,t)} - \mathbf{y} + \boldsymbol{\lambda}^{(l,t)}\right).$$

For $r \in (1,\infty)\backslash\{2\}$, it is strongly convex and can be efficiently solved by some standard numerical methods.

For $r=\infty$, the **s**-minimization problem reads

$$\min_{\mathbf{s}} \frac{1}{\alpha}\|\mathbf{s}\|_\infty + \frac{\rho_1}{2}\|\mathbf{s}-\mathbf{v}\|_2^2,$$

where $\mathbf{v}=\mathbf{A}_{\mathsf{S}^{(l)}}\mathbf{x}^{(l,t)}-\mathbf{y}+\boldsymbol{\lambda}^{(l,t)}$. Let $\widetilde{\mathbf{s}},\widetilde{\mathbf{v}}$ be sorted from $\mathbf{s},\mathbf{v}$ by the absolute values of elements of the known vector $\mathbf{v}$ in ascending order. Then the **s**-problem is equivalent to

$$\min_{\widetilde{\mathbf{s}}} \|\widetilde{\mathbf{s}}\|_\infty + \alpha\rho_1/2\,\|\widetilde{\mathbf{s}}-\widetilde{\mathbf{v}}\|_2^2. \tag{5.7}$$

It is strongly convex and its solution can be given by Theorem 5.1 in the next subsection.

(2) The minimization problem in (5.4) is equivalent to

$$\min_{\mathbf{x}} \frac{\rho_1}{2}\left\|\mathbf{A}_{\mathsf{S}^{(l)}}\mathbf{x}-\mathbf{y}-\mathbf{s}^{(l,t+1)}+\boldsymbol{\lambda}^{(l,t)}\right\|_2^2+\frac{\rho_2}{2}\left\|\mathbf{x}-\mathbf{z}^{(l,t+1)}+\boldsymbol{\mu}^{(l,t)}\right\|_2^2+\frac{\beta}{2}\|\mathbf{x}-\mathbf{x}^{(l)}\|_2^2.$$

The optimality condition is the following linear system

$$\begin{aligned}(\rho_1\mathbf{A}_{\mathsf{S}^{(l)}}^T\mathbf{A}_{\mathsf{S}^{(l)}}+(\rho_2+\beta)\mathbf{I})\mathbf{x}=&\rho_1\mathbf{A}_{\mathsf{S}^{(l)}}^T(\mathbf{y}+\mathbf{s}^{(l,t+1)}-\boldsymbol{\lambda}^{(l,t)})\\&+\rho_2(\mathbf{z}^{(l,t+1)}-\boldsymbol{\mu}^{(l,t)})+\beta\mathbf{x}^{(l)}.\end{aligned}$$

Since $\#\mathsf{S}^{(l)}$ becomes smaller and smaller, it can be solved efficiently by the inverse of a symmetric positive-definite matrix.

REMARK 5.2. In fact, when $r=2$, it is unnecessary to introduce the variable **s**. The scaled ADMM can be simplified in this case.

5.3. The analytical solution for the s-problem with infinity norm.

THEOREM 5.1. *Suppose that the entries of $\widetilde{\mathbf{v}}\in\mathbb{R}^n$ are in ascending order by $|\widetilde{\mathbf{v}}_1|\le|\widetilde{\mathbf{v}}_2|\cdots\le|\widetilde{\mathbf{v}}_n|$. Then the minimization problem*

$$\min_{\widetilde{\mathbf{s}}} \|\widetilde{\mathbf{s}}\|_\infty + \frac{\beta}{2}\|\widetilde{\mathbf{s}}-\widetilde{\mathbf{v}}\|_2^2 \tag{5.8}$$

has the explicit optimal solution $\widetilde{\mathbf{s}}^*$ by

$$\widetilde{\mathbf{s}}_i^* = \begin{cases}\widetilde{\mathbf{v}}_i, & i<i^*,\\ \operatorname{sgn}(\widetilde{\mathbf{v}}_i)t_{i^*}, & i\ge i^*,\end{cases} \tag{5.9}$$

where $i^*\in\{0,1,\cdots,n-1\}$ such that

$$t_{i^*}=\frac{1}{n-i^*}\left(\sum_{j=i^*+1}^{n}|\widetilde{\mathbf{v}}_j|-\frac{1}{\beta}\right) \quad \text{and} \quad t_{i^*}\in[|\widetilde{\mathbf{v}}_{i^*}|,|\widetilde{\mathbf{v}}_{i^*+1}|] \tag{5.10}$$

with $\widetilde{\mathbf{v}}_0=0$.

Proof. Suppose $t=\|\widetilde{\mathbf{s}}\|_\infty$. The minimization problem (5.8) can be rewritten as

$$\min_t f(t)=t+\frac{\beta}{2}\sum_{|\widetilde{\mathbf{v}}_i|>t}(|\widetilde{\mathbf{v}}_i|-t)^2. \tag{5.11}$$

Clearly, $t \in [0, |\tilde{\mathbf{v}}_n|]$. In fact, the objective functional $f(t)$ is a piecewise continuous function. We have

$$f(t) = t + \frac{\beta}{2} \sum_{j=i+1}^{n} (|\tilde{\mathbf{v}}_j| - t)^2, \quad t \in [|\tilde{\mathbf{v}}_i|, |\tilde{\mathbf{v}}_{i+1}|], \quad i = 0, \cdots, n-1,$$

and

$$f'(t) = 1 + \beta \sum_{j=i+1}^{n} (t - |\tilde{\mathbf{v}}_j|), \quad t \in (|\tilde{\mathbf{v}}_i|, |\tilde{\mathbf{v}}_{i+1}|), \quad i = 0, \cdots, n-1. \tag{5.12}$$

We can check that $f'(t)$ is also continuous at $t = |\tilde{\mathbf{v}}_i|$. Therefore $f(t)$ is continuously differentiable.

Moreover, from (5.12), we know that $f'(t)$ is monotonically increasing. Hence $f(t)$ is convex and $f'(t) = 0$ gives us the optimal solution of (5.11), i.e., there exists $i^* \in \{0, 1, \cdots, n-1\}$ such that

$$t_{i^*} = \frac{1}{n - i^*} \left(\sum_{j=i^*+1}^{n} |\tilde{\mathbf{v}}_j| - \frac{1}{\beta} \right) \quad \text{and} \quad t_{i^*} \in [|\tilde{\mathbf{v}}_{i^*}|, |\tilde{\mathbf{v}}_{i^*+1}|].$$

The argument is then straightforward. □

6. Numerical experiments

Numerical experiments are reported in this section to show the efficiency of the InISSAPL-GSO. All of them are implemented on a Laptop (Intel(R) Core(TM) Duo i5-7200u @2.50GHz 2.70GHz, 4.00GB RAM) using Matlab (License ID:1108635).

We consider the application in group sparse signal recovery. Let $\mathbf{x}_{or}$ denote the original signal, which is generated by randomly splitting its components into $\mathtt{g}$ groups. For each nonzero group member, its entries are randomly generated by an i.i.d. Gaussian, unless otherwise specified. The matrix $\mathbf{B} \in \mathbb{R}^{M \times N}$ is also randomly generated by an i.i.d. Gaussian. We let $\mathbf{A}$ be the row orthogonalized matrix of $\mathbf{B}$ by $\mathbf{A} = (orth(\mathbf{B}'))'$ in Matlab code. Then the measurement $\mathbf{y}$ is obtained by

$$\mathbf{y} = \mathbf{A} * \mathbf{x}_{or} + \sigma * noise,$$

where σ is the noise level and *noise* represents the three popular ones, Laplace noise, Gaussian noise or uniform noise.

We denote by s the number of nonzero groups of the original signal $\mathbf{x}_{or}$. Then the sparsity level k_s is defined by $k_s = s/\mathtt{g}$. Without loss of genrality, we consider the uniform group partitions which have the same group size n. Define the relative error ϵ by

$$\epsilon = \frac{\|\mathbf{x} - \mathbf{x}_{or}\|_2}{\|\mathbf{x}_{or}\|_2}.$$

In the numerical experiments, we set $M = 256, N = 1024$ for the size of problem, $\sigma = 0.001$ for the noise level and $n = 8$ for the group size, unless otherwise noted. The model parameter α depends naturally on the noise level. It increases for higher levels of noise. The α values for $\sigma = 0.001$ noise level are shown in Table 6.1. For every predefined triple (p, q, r) in the problem setting, we use a uniform (β, ρ_1, ρ_2) for the algorithmic parameters in all of our experiments; see also Table 6.1.

$(q=0.1,0.3,0.5,0.7,0.9)$	α	β	$\rho_1=\rho_2$
$p=2,r=2$	$5\cdot 10^{-5}\|\mathbf{A}^T\mathbf{y}\|_\infty$	0.01	1
$p=1,r=2$	$5\cdot 10^{-7}\|\mathbf{A}^T\mathbf{y}\|_\infty$	1	10
$p=2,r=1$	$5\cdot 10^{-5}\|\mathbf{A}^T\mathbf{y}\|_\infty$	0.001	1
$p=1,r=1$	$5\cdot 10^{-7}\|\mathbf{A}^T\mathbf{y}\|_\infty$	0.1	0.5
$p=2,r=\infty$	$5\cdot 10^{-10}\|\mathbf{A}^T\mathbf{y}\|_\infty$	0.1	1
$p=1,r=\infty$	$5\cdot 10^{-12}\|\mathbf{A}^T\mathbf{y}\|_\infty$	1	5

Table 6.1: *The parameter settings in the scaled ADMM for the InISSAPL-GSO.*

The recovery is recognized as a success when the relative error ϵ is less than 1%. The stopping criterion of the inner loop of the scaled ADMM is the same as in [8]. It is required to satisfy

$$\|\widehat{\boldsymbol{r}}^{(i+1)}\|_2 \le \sqrt{M}\epsilon^{\text{abs}} + \epsilon^{\text{rel}} \max\left\{\|\widehat{\mathbf{A}}\widehat{\mathbf{x}}^{(i+1)}\|_2, \|\widehat{\mathbf{y}}\|_2, \|\widehat{\boldsymbol{s}}^{(i+1)}\|_2\right\},$$
$$\|\widehat{\boldsymbol{\rho}}\widehat{\mathbf{A}}(\widehat{\mathbf{x}}^{(l,t+1)} - \widehat{\mathbf{x}}^{(l,t)})\|_2 \le \sqrt{N}\epsilon^{\text{abs}} + \epsilon^{\text{rel}}\|\widehat{\boldsymbol{\rho}}\widehat{\boldsymbol{\lambda}}^{(i+1)}\|_2,$$

where

$$\epsilon^{\text{abs}} = \epsilon^{\text{rel}} = 10^{-3},\ \widehat{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \\ & \mathbf{I} \end{bmatrix}, \widehat{\boldsymbol{\rho}} = \begin{bmatrix} \rho_1 & \\ & \rho_2 \end{bmatrix},$$

$$\widehat{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}, \widehat{\boldsymbol{s}} = \begin{bmatrix} \boldsymbol{s} \\ \boldsymbol{z} \end{bmatrix}, \widehat{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ \mathbf{x} \end{bmatrix}, \widehat{\boldsymbol{\lambda}} = \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\mu} \end{bmatrix}, \widehat{\boldsymbol{r}}^{(l,t+1)} = \widehat{\mathbf{A}}\widehat{\mathbf{x}}^{(l,t+1)} - \widehat{\mathbf{y}} - \widehat{\boldsymbol{s}}^{(l,t+1)}.$$

The stopping criterion of outer iteration is $\|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\|_2 / \|\mathbf{x}^{(l)}\|_2 \le 10^{-3}$. The maximal iteration number is set to MAXit=1000 in the inner loop and MAX=100 in the outer iteration.

6.1. Experiments on the initialization of the InISSAPL-GSO. We report the results of experiments when the different types of starting points are used in InISSAPL-GSO. The first kind of starting point is $c\mathbb{1}$ with $c \neq 0$. We choose $c=1$ in the tests. By setting $p=2, q=0.5, r=2$ for Gaussian noise, we compute the relative errors ϵ. The second kind of starting point is randomly generated. We compute the average relative error $\bar{\epsilon}$ over 1000 trials for the same problem setting as in the first kind.

The experiments are performed for different signal recovery problems with three sensing matrices $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ and three sparsity cases $s=8, s=16, s=24$. The comparisons are displayed in Table 6.2.

It shows that the InISSAPL-GSO is effective and not sensitive to the choice of the starting points, even for the less sparsity case $s=24$. Based on this conclusion, we will use all-one vector as the starting point in the following experiments.

		$\mathbf{A}_1$	$\mathbf{A}_2$	$\mathbf{A}_3$
$s=8$	ϵ	0.0042	0.0036	0.0041
	$\bar{\epsilon}$	0.0042	0.0036	0.0041
$s=16$	ϵ	0.0059	0.0063	0.0058
	$\bar{\epsilon}$	0.0059	0.0063	0.0058
$s=24$	ϵ	0.4107	0.0930	0.0084
	$\bar{\epsilon}$	0.4013	0.1016	0.0095

Table 6.2: *Relative errors of the reconstruction by InISSAPL-GSO with two kinds of starting points.*

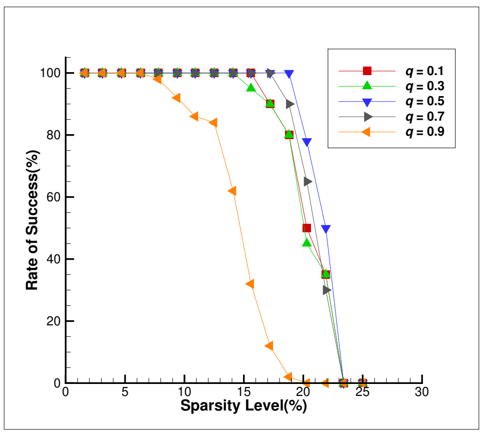


Fig. 6.1: *The comparisons of InISSAPL-GSO on rates of success for different q with $p=r=2$.*

6.2. Accessible to the diversity of noise. Our algorithm is applicable to different types of noises. Here we fix the model parameters $q=1/2, p=2$ and use different r values in the data term for each type of noise to demonstrate the fidelity form derived from the noise statistics. To clearly see those differences, we increase the noise level to $\sigma=0.01$. Accordingly, the model parameter α is enlarged (compared to the values in Table 6.1) and carefully tuned so that each model case achieves its best recovery error for every type of noise, respectively. Meanwhile, the values for the algorithmic parameters β, ρ_1, ρ_2 are the same as in Table 6.1.

For a specific type of noise, we record the relative errors in Table 6.3 when the fidelity term uses different ℓ_r $(r=1,2,\infty)$ norms. It clearly shows that $r=1$ is best for Laplace noise, $r=2$ is best for Gaussian noise and $r=\infty$ is best for uniform noise. We will use them appropriately in the following experiments.

6.3. Choice of p and q. We numerically discuss the effect of the model parameters p, q in the InISSAPL-GSO. Firstly, letting $p=r=2$, we test the algorithm when q varies among $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The rates of success versus the sparsity level are illustrated in Figure 6.1. It shows that the algorithm performs best when $q=1/2$.

Laplace noise	$\epsilon(r=1)$	$\epsilon(r=2)$	$\epsilon(r=\infty)$
$s=4$	**0.0227**	0.0612	0.0794
$s=8$	**0.0308**	0.0502	0.0658
$s=12$	**0.0332**	0.0427	0.0582
$s=16$	**0.0329**	0.0455	0.0584
Gaussian noise	$\epsilon(r=1)$	$\epsilon(r=2)$	$\epsilon(r=\infty)$
$s=4$	0.0656	**0.0362**	0.0679
$s=8$	0.0458	**0.0267**	0.0448
$s=12$	0.0395	**0.0216**	0.0396
$s=16$	0.0557	**0.0334**	0.0556
uniform noise	$\epsilon(r=1)$	$\epsilon(r=2)$	$\epsilon(r=\infty)$
$s=4$	0.0401	0.0387	**0.0303**
$s=8$	0.0367	0.0387	**0.0296**
$s=12$	0.0279	0.0332	**0.0230**
$s=16$	0.0411	0.0517	**0.0292**

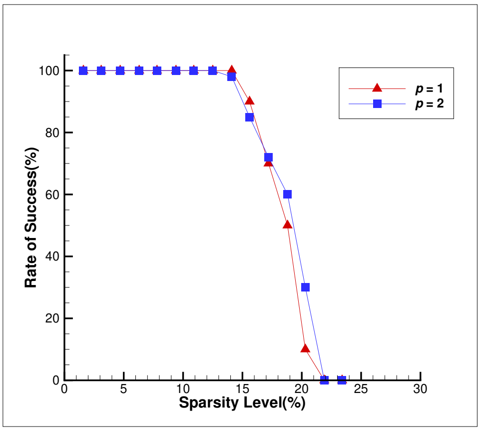
Table 6.3: *Relative errors ϵ of InISSAPL-GSO over r for Laplace noise (top), Gaussian noise (middle), uniform noise (bottom) with $p=2, q=0.5, \sigma=0.01$.*

This fact is consistent with the numerical results in [22, 39]. Hence, we always choose $q=1/2$ in the experiments.

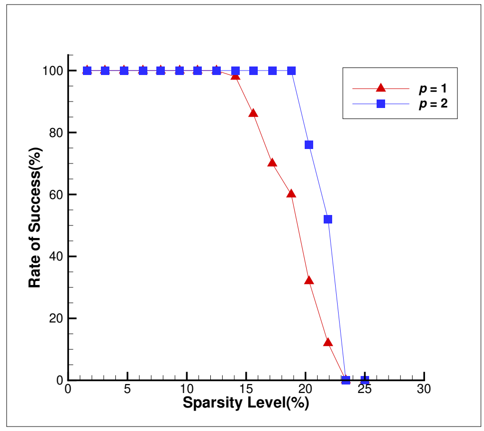
Secondly, we examine the algorithm on commonly used $p=1$ and $p=2$ for the three types of noises. We compare the rates of success versus the sparsity level in Figure 6.2. It can be observed that the rates of success have no essential numerical difference between $p=1$ and $p=2$ for the different types of noises. The reason is that there is zero probability to get zero entries in the nonzero group member generated by the i.i.d. Gaussian.

6.4. Sensitivity analysis. In this subsection, we study the sensitivity of our algorithm on group size n. The rates of success versus the sparsity level are illustrated in Figure 6.3 for group size $n=4,8,16,32$. It shows that the larger the group size, the higher the rate of success. This fact is true because more information is included for larger group size.

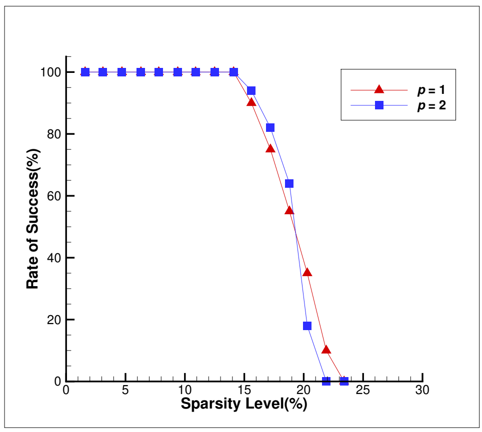
6.5. Comparison with some state-of-the-art algorithms. We compare the InISSAPL-GSO with others in the existing works for the group sparse model. The algorithms are typically PGM-GSO [22] and the convex optimization group lasso [8]. In the code of PGM-GSO algorithm (available online https://CRAN.R-project.org/package=GSparO), there is a predefined group sparsity number s for regularization parameter update. Since it is hard to know this exact s beforehand in



(a) Laplace noise($r=1, q=0.5$)



(b) Gaussian noise($r=2, q=0.5$)

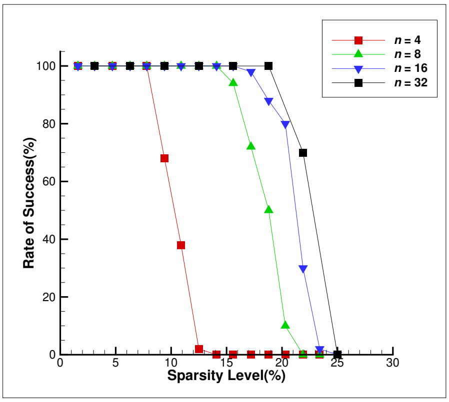


(c) uniform noise($r=\infty, q=0.5$)

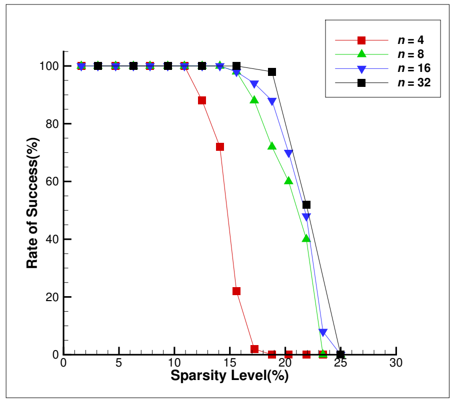
Fig. 6.2: *Comparisons of InISSAPL-GSO on rates of success for Laplace noise (a), Gaussian noise (b) and uniform noise (c) between* $p=1$ *and* $p=2$.

applications, we also use an estimated value s_e (close to the true value s) with $s_e = s+2$ in the experiments for more tests. The PGM-GSO with estimated s_e is named e-PGM-GSO. The comparisons on rates of success are displayed in Figure 6.4 by setting the parameters $p=2, q=1/2, r=2, n=8$ for Gaussian noise. We can see that the rates of success of PGM-GSO (with exact s of the number of nonzero groups of the ground truth) and our InISSAPL-GSO are similar, which are considerably higher than e-PGM-GSO and group lasso. Note that our InISSAPL-GSO does not require to input any number of the nonzero groups.

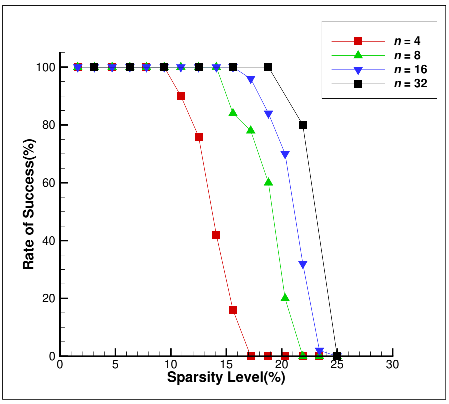
For the competitive algorithms, InISSAPL-GSO, PGM-GSO, and e-PGM-GSO, we also compare the running time and relative errors for different sized problems in Table 6.4. It is illustrated again that, the recovery accuracies of InISSAPL-GSO and PGM-GSO are similar, and higher than e-PGM-GSO. Meanwhile, InISSAPL-GSO is more efficient than PGM-GSOers, especially for larger scale problems. The reason is that the computation is implemented only on the shrinking group support set.



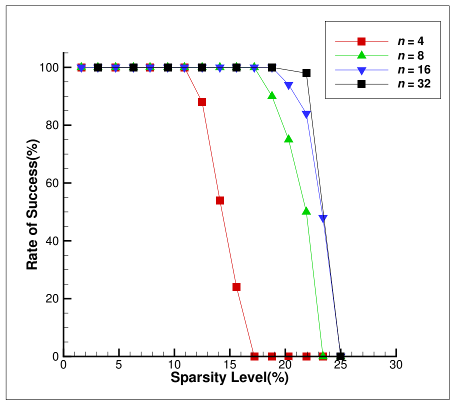
(a) Laplace noise with $p=1$



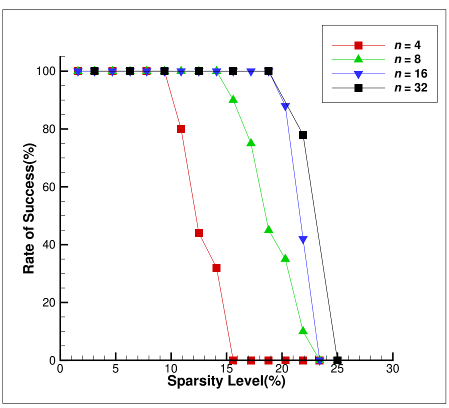
(b) Laplace noise with $p=2$



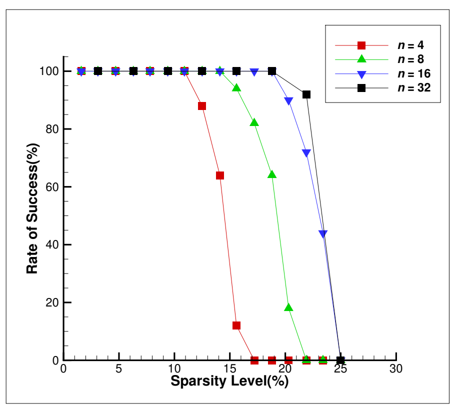
(c) Gaussian noise with $p=1$



(d) Gaussian noise with $p=2$



(e) uniform noise with $p=1$



(f) uniform noise with $p=2$

Fig. 6.3: *Sensitivity analysis of InISSAPL-GSO over group size for Laplace noise (a) and (b), Gaussian noise (c) and (d), uniform noise (e) and (f).*

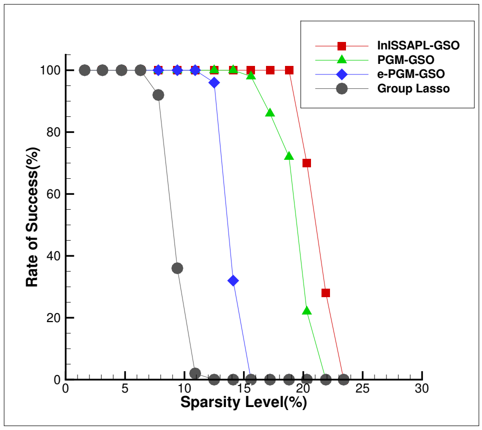
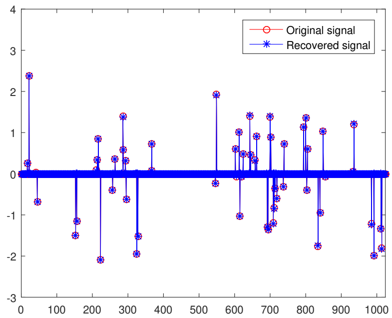


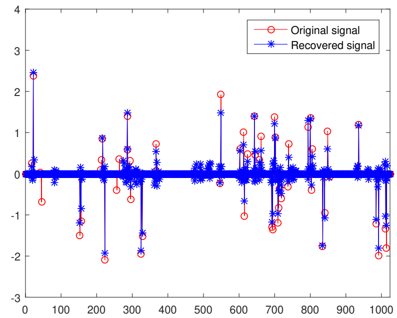
Fig. 6.4: *Comparisons on rate of success for InISSAPL-GSO, PGM-GSO (with true value of the number of nonzero groups s), e-PGM-GSO (with estimated value of the number of nonzero groups $s_e = s+2$) and Group Lasso algorithms.*

$M = 256$						
$N = 1024$	PGM-GSO		e-PGM-GSO		InISSAPL-GSO	
s	Time(s)	ϵ	Time(s)	ϵ	Time(s)	ϵ
4	0.56	0.0024	0.59	0.0031	**0.46**	**0.0023**
8	0.58	**0.0025**	0.59	0.0033	**0.49**	0.0027
12	0.58	**0.0030**	0.60	0.0032	**0.50**	**0.0030**
16	0.59	0.0033	0.81	0.0040	**0.52**	**0.0031**
$M = 1024$						
$N = 4096$	PGM-GSO		e-PGM-GSO		InISSAPL-GSO	
s	Time(s)	ϵ	Time(s)	ϵ	Time(s)	ϵ
16	17.08	0.0022	17.75	0.0023	**4.39**	**0.0021**
32	17.89	0.0020	17.76	0.0024	**5.02**	**0.0019**
48	18.10	**0.0025**	18.04	0.0028	**5.44**	**0.0025**
64	18.95	**0.0027**	18.36	0.0029	**7.68**	0.0028

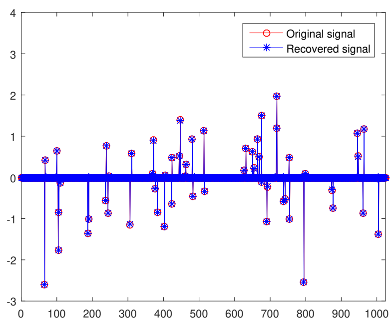
Table 6.4: *Comparisons on running time and relative errors ϵ for PGM-GSO, e-PGM-GSO, InISSAPL-GSO algorithms for two different sized problems. It can be seen that the advantages of our algorithm become greater for larger scale problems.*



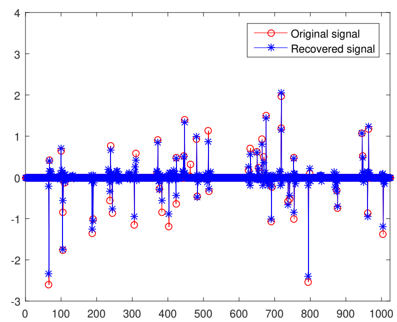
(a) Laplace noise with $p=1, r=1$



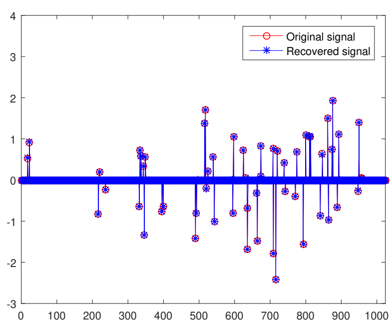
(b) Laplace noise with $p=2, r=1$



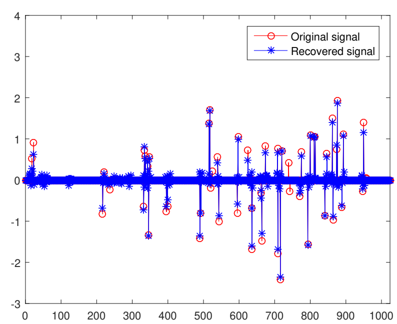
(c) Gaussian noise with $p=1, r=2$



(d) Gaussian noise with $p=2, r=2$



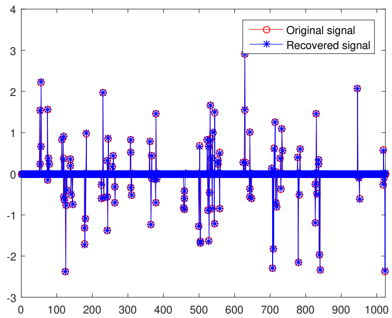
(e) uniform noise with $p=1, r=\infty$



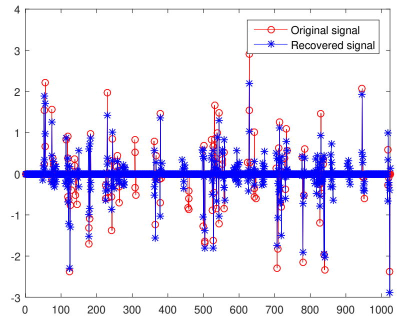
(f) uniform noise with $p=2, r=\infty$

Fig. 6.5: *Signal recovery by InISSAPL-GSO in the case of $g_s=2$, i.e., the 25% intra-group sparsity level.*

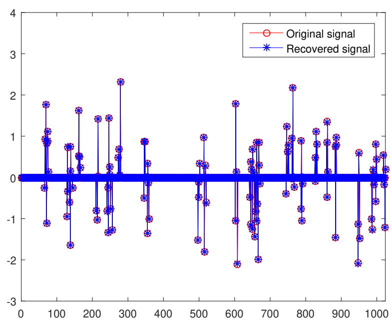
6.6. Experiments on the intra-group sparse signal recovery. It is natural that the signal may also be sparse within some groups. We do the experiments on the signal recovery of this situation. Let the number of nonzero groups $s=28$ and the group size $n=8$. In each group, the number of nonzero elements is denoted by g_s. We consider $g_s=2$, 4 in these tests. Figure 6.5 and Figure 6.6 display the recovery results with $p=1$ and $p=2$ for different types of noises. We can observe that it works well with $p=1$ while it fails with $p=2$.



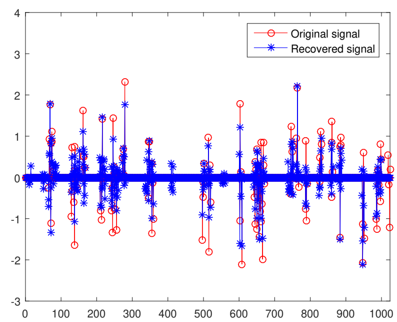
(a) Laplace noise with $p=1, r=1$



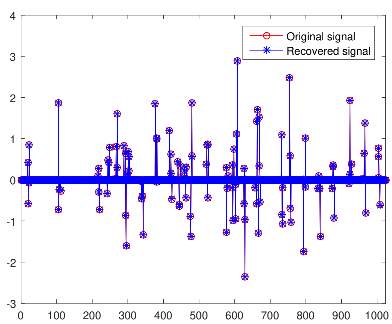
(b) Laplace noise with $p=2, r=1$



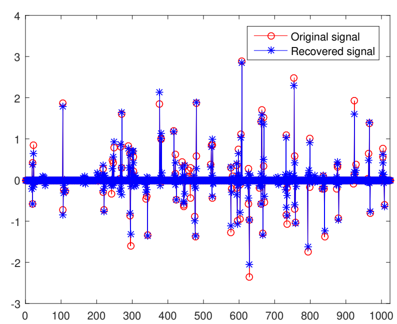
(c) Gaussian noise with $p=1, r=2$



(d) Gaussian noise with $p=2, r=2$



(e) uniform noise with $p=1, r=\infty$



(f) uniform noise with $p=2, r=\infty$

Fig. 6.6: *Signal recovery by InISSAPL-GSO in the case of $g_s=4$, i.e., the 50% intra-group sparsity level.*

7. Conclusions

The group sparse $\ell_{p,q}$-ℓ_r model is very useful in many applications. The InISSAPL-GSO provides a unified framework to deal with all the cases of parameters $p \geq 1, 0 < q < 1, 1 \leq r \leq \infty$. When proving the global convergence of algorithm with KL property, we develop a lower bound theory for the nonzero groups of the iterative sequence to avoid the non-Lipschitz property and construct a sophisticated subgradient. Along iterations, the unknowns become fewer and fewer and can be calculated by the scaled ADMM in the inner loop. Therefore it is specially efficient for large-scale problems. Lots of numerical experiments and comparisons demonstrate the good performance of our algorithm.

In our future work, the model and algorithm can be extended to other applications with overlapping groups structure such as the gene expression data and the patch patterns in image processing.

Acknowledgements. We greatly appreciate helpful discussions with Xue Feng, and thank the authors of [22] for making their code available online. This work is supported in part by the Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin (C. Wu,Y. Xue), NSFTJ-17JCYBJC15800 (Y. Xue), NSFC 11871035 (C. Wu), NSFC 11531013 (C. Wu) and Recruitment Program of Global Young Experts (C. Wu).

Appendix A. Subdifferential. We firstly recall the basic definitions of subdifferential and horizon cone from the reference [35].

DEFINITION A.1 (Subdifferentials). *Let $h:\mathbb{R}^N\to\mathbb{R}\cup\{+\infty\}$ be a proper, lower semi-continuous function.*

(1) The regular subdifferential *of h at $\bar{\mathbf{x}}\in\operatorname{dom}h=\{\mathbf{x}\in\mathbb{R}^N: h(\mathbf{x})<+\infty\}$ is defined as*

$$\widehat{\partial}h(\bar{\mathbf{x}}):=\left\{\mathbf{v}\in\mathbb{R}^N:\liminf_{\substack{\mathbf{x}\to\bar{\mathbf{x}}\\ \mathbf{x}\neq\bar{\mathbf{x}}}}\frac{h(\mathbf{x})-h(\bar{\mathbf{x}})-\langle\mathbf{v},\mathbf{x}-\bar{\mathbf{x}}\rangle}{\|\mathbf{x}-\bar{\mathbf{x}}\|}\geq 0\right\};$$

(2) The (limiting) subdifferential *of h at $\bar{\mathbf{x}}\in\operatorname{dom}h$ is defined as*

$$\partial h(\bar{\mathbf{x}}):=\left\{\mathbf{v}\in\mathbb{R}^N:\exists\mathbf{x}^{(k)}\to\bar{\mathbf{x}},h(\mathbf{x}^{(k)})\to h(\bar{\mathbf{x}}),\mathbf{v}^{(k)}\in\widehat{\partial}h(\mathbf{x}^{(k)}),\mathbf{v}^{(k)}\to\mathbf{v}\right\};$$

(3) The horizon subdifferential *of h at $\bar{\mathbf{x}}\in\operatorname{dom}h$ is defined as*

$$\partial^\infty h(\bar{\mathbf{x}})$$
$$:=\left\{\mathbf{v}\in\mathbb{R}^N:\exists\mathbf{x}^{(k)}\to\bar{\mathbf{x}},h(\mathbf{x}^{(k)})\to h(\bar{\mathbf{x}}),\mathbf{v}^{(k)}\in\widehat{\partial}h(\mathbf{x}^{(k)}),\lambda^{(k)}\mathbf{v}^{(k)}\to\mathbf{v}\text{for some sequence }\lambda^{(k)}\searrow 0\right\}.$$

REMARK A.1. As summarized in [26], the following properties hold:

(1) For any $\bar{\mathbf{x}}\in\operatorname{dom}h$, $\widehat{\partial}h(\bar{\mathbf{x}})\subseteq\partial h(\bar{\mathbf{x}})$. If h is continuously differentiable at $\bar{\mathbf{x}}$, then $\widehat{\partial}h(\bar{\mathbf{x}})=\partial h(\bar{\mathbf{x}})=\{\nabla h(\bar{\mathbf{x}})\}$;

(2) For any $\bar{\mathbf{x}}\in\operatorname{dom}h$, the subdifferential set $\partial h(\bar{\mathbf{x}})$ is closed, i.e,

$$\left\{\mathbf{v}\in\mathbb{R}^N:\exists\mathbf{x}^{(k)}\to\bar{\mathbf{x}},h(\mathbf{x}^{(k)})\to h(\bar{\mathbf{x}}),\mathbf{v}^{(k)}\in\partial h(\mathbf{x}^{(k)}),\mathbf{v}^{(k)}\to\mathbf{v}\right\}\subset\partial h(\bar{\mathbf{x}}).$$

DEFINITION A.2 (Horizon cone). *For a set $C\subset\mathbb{R}^N$, the* horizon cone *is the closed cone C^∞ given by*

$$C^\infty=\begin{cases}\{\mathbf{v}\mid\exists\ \mathbf{v}^{(k)}\in C,\lambda^{(k)}\searrow 0,\lambda^{(k)}\mathbf{v}^{(k)}\to\mathbf{v}\} & \text{when } C\neq\emptyset,\\ \{\mathbf{0}\} & \text{when } C=\emptyset.\end{cases}$$

REMARK A.2. A set $C\subset\mathbb{R}^N$ is bounded if and only if its horizon cone is just the zero cone: $C^\infty=\{\mathbf{0}\}$.

Appendix B. KL property and convergence theorem. The Kurdyka-Łojasiewicz (KL) property [1,2,7,23,27] is a useful tool for establishing the convergence of bounded sequence.

DEFINITION B.1. *A proper function h is said to have the* Kurdyka-Łojasiewicz property *at $\bar{\mathbf{x}}\in\mathrm{dom}\,\partial h=\{\mathbf{x}\in\mathbb{R}^{\mathsf{N}}:\partial h(\mathbf{x})\neq\emptyset\}$ if there exist $\zeta\in(0,+\infty]$, a neighborhood U of $\bar{\mathbf{x}}$, and a continuous concave function $\varphi:[0,\zeta)\to\mathbb{R}_+$ such that*

(1) $\varphi(0)=0$;

(2) $\varphi(0)$ is C^1 on $(0,\zeta)$;

(3) for all $s\in(0,\zeta)$, $\varphi'(s)>0$;

(4) for all $\mathbf{x}\in U$ satisfying $h(\bar{\mathbf{x}})<h(\mathbf{x})<h(\bar{\mathbf{x}})+\zeta$, the Kurdyka-Łojasiewicz inequality holds:

$$\varphi'(h(\mathbf{x})-h(\bar{\mathbf{x}}))\,\mathrm{dist}(0,\partial h(\mathbf{x}))\geq 1.$$

where $\mathrm{dist}(0,\partial h(\mathbf{x}))=\min\{\|\mathbf{v}\|:\mathbf{v}\in\partial h(\mathbf{x})\}$,

A proper, lower semicontinuous function h satisfying the KL property at all points in $\mathrm{dom}\,\partial h$ is called a *KL function.* One can refer to [2,7] for examples of KL functions and the application of KL property in optimization theory.

Recently, the KL property has been extended to the definable functions in an o-minimal structure for the nonsmooth version; see [1,6,18,23] and the references therein. The following definitions and Theorem B.1 are from them.

DEFINITION B.2. *Let $\mathcal{O}=\{\mathcal{O}_n\}_{n\in\mathbb{N}}$ be such that each $\mathcal{O}_n$ is a collection of subsets of $\mathbb{R}^n$. The family $\mathcal{O}$ is an* o-minimal structure *over $\mathbb{R}$, if it satisfies the following axioms:*

(1) Each $\mathcal{O}_n$ is a Boolean algebra. Namely $\emptyset\in\mathcal{O}_n$ and for each $A,B\in\mathcal{O}_n$, $A\cup B, A\cap B$, and $\mathbb{R}^n\backslash A$ belong to $\mathcal{O}_n$.

(2) For all $A\in\mathcal{O}_n$, $A\times\mathbb{R}$ and $\mathbb{R}\times A$ belong to $\mathcal{O}_{n+1}$.

(3) For all $A\in\mathcal{O}_{n+1}$, $\prod(A):=\{(x_1,\cdots,x_n)\in\mathbb{R}^n|(x_1,\cdots,x_n,x_{n+1})\in A\}$ belongs to $\mathcal{O}_n$.

(4) For all $i\neq j$ in $\{1,2,\cdots,n\}$, $\{(x_1,\cdots,x_n)\in\mathbb{R}^n|x_i=x_j\}$ belong to $\mathcal{O}_n$.

(5) The set $\{(x_1,x_2)\in\mathbb{R}^2|x_1<x_2\}$ belongs to $\mathcal{O}_2$.

(6) The elements of $\mathcal{O}_1$ are exactly finite unions of intervals.

DEFINITION B.3. *Given an o-minimal structure $\mathcal{O}$ over $\mathbb{R}$. A set C is said to be* definable *(in $\mathcal{O}$) if C belongs to $\mathcal{O}$. A function $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ is said to be* definable *in $\mathcal{O}$ if its graph belongs to $\mathcal{O}_{n+1}$.*

Then the definable function has the following property:

- finite sums of definable functions are definable;
- compositions of definable functions are definable;
- function of $f(y)=\sup_{x\in C}g(x,y)$ is definable if $g(x,y)$ and the set C are definable.

As an example [1, 18], there exists an o-minimal structure containing the graph of $x^r:\mathbb{R}\to\mathbb{R}, r\in\mathbb{R}$, which is given by

$$a\mapsto\begin{cases}a^r, & a>0\\ 0, & a\leq 0.\end{cases}\tag{B.1}$$

THEOREM B.1. *Any proper lower semicontinuous function $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ that is definable in an o-minimal structure $\mathcal{O}$ has the Kurdyka-Łojasiewicz property at each point of* $\mathrm{dom}\,\partial f$.

From this theorem and Definition B.3, the objective function $\mathcal{E}$ in this paper is the compositions of definable functions. So it satisfies the KL property.

The following theorem gives a general and important theoretical framework for the sequence convergence.

THEOREM B.2 ([2]). *Let $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ be a proper lower semicontinous function. a and b are fixed positive constants. Consider a sequence $\{\mathbf{x}^{(l)}\}$ that satisfies*

(H1). (Sufficient decrease condition). For each l,

$$f(\mathbf{x}^{(l+1)})+a\|\mathbf{x}^{(l+1)}-\mathbf{x}^{(l)}\|^2\le f(\mathbf{x}^{(l)});$$

(H2). (Relative error condition). For each l, there exists $w^{(l+1)}\in\partial f(\mathbf{x}^{(l+1)})$ such that

$$\|w^{(l+1)}\|\le b\|\mathbf{x}^{(l+1)}-\mathbf{x}^{(l)}\|;$$

(H3). (Continuity condition). There exists a subsequence $\{\mathbf{x}^{(l_k)}\}$ and $\widetilde{\mathbf{x}}$ such that

$$\mathbf{x}^{(l_k)}\to\widetilde{\mathbf{x}} \text{ and } f(\mathbf{x}^{(l_k)})\to f(\widetilde{\mathbf{x}}),\quad \text{as } k\to\infty.$$

If f has the KL property at the cluster point $\widetilde{\mathbf{x}}$ specified in (H3), then the sequence $\{\mathbf{x}^{(l)}\}$ converges to $\bar{\mathbf{x}}=\tilde{\mathbf{x}}$ as $l\to\infty$ and $\bar{\mathbf{x}}$ is a critical point.

REFERENCES

[1] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality*, Math. Oper. Res., 35(2):257–512, 2010. B, B, B

[2] H. Attouch, J. Bolte, and B.F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Math. Program., 137(1-2):91–129, 2013. 3.3, 4, B, B, B.2

[3] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2(1):183–202, 2009. 1

[4] D.P. Bertsekas, *Control of uncertain systems with a set-membership description of the uncertainty*, PhD Thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology. 1971. 3.1

[5] W. Bian and X. Chen, *Worst-case complexity of smoothing quadratic regularization methods for non-Lipschitzian optimization*, SIAM J. Optim., 23(3):1718–1741, 2013. 1

[6] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, *Clarke subgradients of stratifiable functions*, SIAM J. Optim., 18(2):556–572, 2007. B

[7] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program., 146(1-2):459–494, 2014. B, B

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3(1):1–122, 2011. 1, 5, 6, 6.5

[9] K. Bredies, D.A. Lorenz, and S. Reiterer, *Minimization of non-smooth, non-convex functionals by iterative thresholding*, J. Optim. Theory Appl., 165(1):78–112, 2015. 1

[10] E.J. Candès, M.B. Wakin, and S.P. Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, J. Fourier Anal. Appl., 14(5):877–905, 2008. 1

[11] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40(1):120–145, 2011. 5

[12] R.H. Chan and H.-X. Liang, *Half-quadratic algorithm for ℓ_p-ℓ_q problems with applications to TV-ℓ_1 image restoration and compressive sensing*, in A. Bruhn, T. Pock, and X.-C. Tai, (eds.), Efficient Algorithms for Global Optimization Methods, Computer Vision, Berlin, Heidelberg, 78–103, 2014. 1

[13] X. Chen, L. Niu, and Y. Yuan, *Optimality conditions and a smoothing trust region Newton method for non Lipschitz optimization*, SIAM J. Optim., 23(3):1528–1552, 2013. 1

[14] X. Chen, F. Xu, and Y. Ye, *Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization*, SIAM J. Sci. Comput., 32(5):2832–2852, 2010. 1, 3.2

[15] X. Chen and W. Zhou, *Convergence of the reweighted ℓ_1 minimization algorithm for ℓ_2-ℓ_p minimization*, Comput. Optim. Appl., 59(1):47–61, 2014. 1

[16] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math., 57(11):1413–1457, 2004. 1

[17] W. Deng, W. Yin, and Y. Zhang, *Group sparse optimization by alternating direction method*, Proc. SPIE8858, Wavelets and Sparsity XV, 88580R, 2013. 1, 5

[18] L. van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Math. J., 84(2):497–540, 1996. B, B

[19] E. Esser, X. Zhang, and T.F. Chan, *A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science*, SIAM J. Imaging Sci., 3(4):1015–1046, 2010. 5

[20] X. Feng, S. Yan, and C. Wu, *The $\ell_{2,q}$ regularized group sparse optimization: lower bound theory, recovery bound and algorithms*, submitted. 3.6

[21] T. Goldstein and S. Osher, *The split Bregman method for L_1-regularized problems*, SIAM J. Imaging Sci., 2(2):323–343, 2009. 5

[22] Y. Hu, C. Li, K. Meng, J. Qin, and X. Yang, *Group sparse optimization via $L_{p,q}$ regularization*, J. Mach. Learn. Res., 18:1–52, 2017. 1, 1, 2.1, 2, 3.2, 1a, 6.3, 6.5, 7

[23] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, Ann. Inst. Fourier (Grenoble), 48(3):769–783, 1998. B, B

[24] M.-J. Lai, Y. Xu, and W. Yin, *Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization*, SIAM J. Numer. Anal., 51(2):927–957, 2013. 1

[25] A. Lanza, S. Morigi, L. Reichel, and F. Sgallari, *A generalized Krylov subspace method for ℓ_p-ℓ_q minimization*, SIAM J. Sci. Comput., 37(5):S30–S50, 2015. 1

[26] Z. Liu, Y. Zhao, and C. Wu, *A new globally convergent algorithm for non-Lipschitz ℓ_p-ℓ_q minimization*, Adv. Comput. Math., 45(3):1369–1399, 2019. 1, 2.1, 3.3, 4, A.1

[27] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, In Les Équations aux Dérivées Partielles (Paris, 1962), Éditions du Centre National de la Recherche Scientifique, Paris, 87–89, 1963. B

[28] Z. Lu, *Iterative reweighted minimization methods for ℓ_p regularized unconstrained nonlinear programming*, Math. Program., 147(1):277–307, 2014. 1

[29] J.-J. Mei, Y. Dong, T.-Z. Huang, and W. Yin, *Cauchy noise removal by nonconvex ADMM with convergence guarantees*, J. Sci. Comput., 74(2):743–766, 2018. 4

[30] L. Meier, S.A. van de Geer, and P. Bühlmann, *The group Lasso for logistic regression*, J.R. Statist. Soc. B, 70:53–71, 2008. 1

[31] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, 2006. 3.5

[32] P. Ochs, Y. Chen, T. Brox, and T. Pock, *iPiano: Inertial Proximal Algorithm for Nonconvex Optimization*, SIAM J. Imaging Sci., 7(2):1388–1419, 2014. 4

[33] P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock, *On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision*, SIAM J. Imaging Sci., 8(1):331–372, 2015. 4

[34] T. Pock and S. Sabach, *Inertial Proximal Alternating Linearized Minimization (iPALM) for nonconvex and nonsmooth problems*, SIAM J. Imaging Sci., 9(4):1756–1787, 2016. 4

[35] R.T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Grundlehren der Mathematischen Wissenschaften, Springer-Verlag Berlin Heidelberg, 317, 2009. 3.1, 3.1, 3.1, A

[36] L. Shen, Y. Xu, and X. Zeng, *Wavelet inpainting with the ℓ_0 sparse regularization*, Appl. Comput. Harmon. Anal., 41(1):26–53, 2016. 4

[37] M. Usman, C. Prieto, T. Schaeffter, and P.G. Batchelor, *$k-t$ group sparse: A method for accelerating dynamic MRI*, Magn. Reson. Med., 66(4):1163–1176, 2011. 1

[38] Q. Wang, X. Zhang, Y. Wu, L. Tang, and Z. Zha, *Nonconvex weighted ℓ_p minimization based group sparse representation framework for image denoising*, IEEE Signal Process. Lett., 24(11):1686–1690, 2017. 1

[39] Z. Xu, X. Chang, F. Xu, and H. Zhang, *$L_{1/2}$ regularization: A thresholding representation theory and a fast solver*, IEEE Trans. Neural Netw. Learn. Syst., 23(7):1013–1027, 2012. 1, 6.3

[40] H. Yang, Z. Xu, I. King, and M.R. Lyu, *Online learning for group Lasso*, Proceedings of the 27th International Conference on Machine Learning (ICML), 1191–1198, 2010. 1

[41] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, *Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing*, SIAM J. Imaging Sci., 1(1):143–168, 2008. 5

[42] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. R. Statist. Soc. B, 68:49–67, 2006. 1, 1

[43] C. Zeng, R. Jia, and C. Wu, *An iterative support shrinking algorithm for Non-Lipschitz optimization in image restoration*, J. Math. Imaging Vis., 61(1):122–139, 2019. 1, 4

[44] C. Zeng, C. Wu, and R. Jia, *Non-Lipschitz models for image restoration with impulse noise*

removal, SIAM J. Imaging Sci., 12(1):420–458, 2019. 1, 4

[45] X. Zhang, M. Bai, and M.K. Ng, *Nonconvex-TV based image restoration with impulse noise removal*, SIAM J. Imaging Sci., 10(3):1627–1667, 2017. 4

[46] W. Zuo, D. Meng, L. Zhang, X. Feng, and D. Zhang, *A generalized iterated shrinkage algorithm for non-convex sparse coding*, Proc. IEEE Int. Conf. Comput. Vis., 217–224, 2013.

1