Robust Sparse Recovery via Weakly Convex Regularization

Chengzhu Yang, Xinyue Shen, Hongbing Ma, Badong Chen, Yuantao Gu, and Hing Cheung So

Abstract—Robust sparse signal recovery against impulsive noise is a core issue in many applications. Numerous methods have been proposed to recover the sparse signal from measurements corrupted by various impulsive noises, but most of them either lack theoretical guarantee for robust sparse recovery or are not efficient enough for large-scale problems. In this work, a general optimization problem for robust sparse signal recovery, which includes many existing works as concrete instances, is analyzed by a freshly defined Double Null Space Property (DNSP), and its solution is proved to be able to robustly reconstruct the sparse signal under mild conditions. Moreover, for computational tractability, weakly convex sparsity-inducing penalties are applied to the general problem, and properties of the solution to the resultant non-convex problem are further analyzed. The experimental results prove that the sparse signal can be precisely reconstructed by RPGG from compressive measurements with sparse noise or robustly recovered from those with impulsive noise. Meanwhile, simulations demonstrate that RPGG with tuned parameters outperforms other robust sparse recovery algorithms.

Index Terms—Compressed sensing, Weakly convex, Generalized gradient, Robust sparse recovery, Impulsive noise

I. INTRODUCTION

We consider the problem of recovering a sparse signal $x \in \mathbb{R}^N$ from a measurement vector $y \in \mathbb{R}^M$ corrupted by the noise $e$, i.e.,

$$y = Ax + e,$$

where $A \in \mathbb{R}^{M \times N}$ is a known sensing matrix. Given that $e$ is a zero-mean Gaussian noise vector, to reconstruct the sparse $x$, we can solve the following optimization problem [1], [2]

$$\min_{\tilde{x}} \lambda F(\tilde{x}) + \|y - A\tilde{x}\|_2^2,$$

where $F(\cdot)$ is a sparsity-inducing penalty (for instance, $\ell_1$ norm) to promote the sparseness of $\tilde{x}$ [1], [3], the $\ell_2$ norm is a denoising penalty, and $\lambda$ is a regularization parameter to balance the two penalties. Theoretical analyses of the robust reconstruction performance of problem (1) can be found in [1], [4], [5]. However, if $e$ is non-Gaussian noise, then problem (1) cannot recover $x$ robustly [6]–[8]. The non-Gaussian noise that we consider includes outliers and more generally impulsive noise. Throughout this work, the outliers form a sparse vector, and impulsive noise is sum of a vector of outliers and a dense vector with small bounded entries [8]–[10], so a vector of outliers is also a kind of impulsive noise.

Impulsive noises appear frequently in sparse recovery applications. Wireless Body Area Network detects health condition via analyzing the electrocardiogram (ECG) signals, but they are always corrupted by electromyographic noise, which is impulsive [8]. Nonintrusive load monitoring system detecting the usages of house appliances acquires the power signal by Compressive Sensing (CS) methods, but the power signal is corrupted by impulsive noise due to the switching of the appliances [8]. In image processing, the desired original image to be recovered from an image is contaminated by the salt and pepper noise. In super-resolution and inpainting, only partial entries of the desired image are available, and the signal at hand can be modeled as measurements corrupted by outliers [11]–[13]. In wireless sensor networks, nodes may be power off or broken, so the collected signal can also be modeled as polluted by outliers. Besides, non-Gaussian noises appear in real-world systems in wireless and power line communications, telettraffic, hydrology, geology, and economics. The interested readers is referred to [8], [14]–[16] and the references therein.

As a consequence, apart from classical methods based on $\ell_2$ norm regularization, robust sparse recovery methods based on other regularizations aiming at reducing impulsive noises have attracted much attention in recent years, and we will review some of them in the following. In this work, robust sparse recovery refers to recovery of sparse signal with non-Gaussian noise, which will be denoted by robust I sparse recovery hereafter. To avoid ambiguity, robust II is adopted at the rest of the paper if robust means that the recovered signal is close to the original one. Besides, robustly I, robustly II, robustness I, and robustness II are defined in a similar manner.

A. Prior Work

By exploiting the fact that a vector of outliers is sparse, and impulsive noise is approximately sparse, i.e., only a few elements are large in absolute amplitude and most of them are...
close to zero [8], [17], [18], sparsity-inducing penalties have been used to penalize impulsive noise including outliers.

As the convex relaxation of $\ell_0$ pseudo norm, $\ell_1$ norm has been used for sparse recovery against outliers in the following $\ell_1$-$\ell_1$ problem [19], [20]

$$\min_{\tilde{x}} \lambda \|\tilde{x}\|_1 + \|y - A\tilde{x}\|_1.$$  \hspace{1cm} (2)

Because of the convexity of problem (2), its global optimum can be readily obtained, and numerous algorithms have been proposed to solve (2) and its variants [9], [10], [18], [19], [21].

Though $\ell_1$ norm has been widely used, it is not as good at inducing sparsity as $\ell_p$ pseudo norm ($0 < p < 1$) [22], [23]. Consequently, the following non-convex $\ell_1$-$\ell_p$ and $\ell_p$-$\ell_p$ problems have been proposed in [24] and [25], respectively,

$$\min_{\tilde{x}} \lambda \|\tilde{x}\|_1 + \|y - A\tilde{x}\|_p^p,$$  \hspace{1cm} (3)

$$\min_{\tilde{x}} \lambda \|\tilde{x}\|_p^p + \|y - A\tilde{x}\|_p^p.$$  \hspace{1cm} (4)

However, existing algorithms for (3) and (4) are not guaranteed to be able to recover the sparse signal robustly II.

Assuming that the impulsive noise follows Cauchy distribution, the Lorentzian regularization problem has been proposed [7]:

$$\min_{\tilde{x}} \lambda \|\tilde{x}\|_{\tilde{1},\tilde{2},\tilde{p}} + \|y - A\tilde{x}\|_{\tilde{1},\tilde{2},\tilde{p}},$$  \hspace{1cm} (5)

where $\|\tilde{x}\|_{\tilde{1},\tilde{2},\tilde{p}} = \sum_i \log(1 + \frac{\tilde{x}_i^2}{\tilde{p}^2})$. However, its solving algorithms are not theoretically justified to be able to robustly II recover the sparse signal.

Very recently, a general problem with non-convex sparsity-inducing regularization has been studied [26]:

$$\min_{\tilde{x}} \lambda \|\tilde{x}\|_p + G(y - A\tilde{x}),$$  \hspace{1cm} (6)

where $G(\cdot)$ is a sparsity-inducing penalty. An Alternating Direction Method of Multipliers (ADMM) algorithm with convergence analysis is provided, but there is no analysis on the recovery condition or any guarantee that the method can find a sparse local minimum.

There have been many works considering the following sparsity constrained problem

$$\min_{\tilde{x}} \lambda \|\tilde{x}\|_p \text{ s.t. } \|\tilde{x}\|_0 \leq k_1,$$  \hspace{1cm} (6)

where $G(\cdot)$ is a non-convex penalty. The $\ell_0$-Least Absolute Deviations ($\ell_0$-LAD) algorithm is proposed to solve (6), when $G(\cdot)$ is $\ell_1$ norm [27]. The $\ell_p$-Orthogonal Matching Pursuit ($\ell_p$-OMP) algorithm is proposed to solve this problem, where $G(\cdot)$ is substituted by $\ell_p$ pseudo norm [6]. Lorentzian function is introduced to this problem, and the Lorentzian Iterative Hard Thresholding (LIHT) is proposed in [28]. Subsequently, a more general case is analyzed in [29], and Huber Iterative Hard Thresholding (HIHT) algorithm is proposed. However, these algorithms require the sparsity of the signal $x$ as a prior.

In [30], an elegant analysis proves that a class of weakly convex sparsity-inducing penalty (WCSIP) can be used for sparse recovery, then the weakly convex regularized sparse recovery problem as well as its solving algorithm Projected Generalized Gradient (PGG) are proposed and analyzed. PGG is computationally efficient, and the convergence analysis states that under mild conditions, PGG can reconstruct the sparse signal. Recently, Lasso and Square-Root Lasso using WCSIP to replace $\ell_1$ norm are investigated in [31], [32], where better performances are demonstrated compared with their counterparts using $\ell_1$ norm. However, WCSIP has not been applied to robust I sparse recovery problems.

According to the above survey, on the one hand, impulsive noises are frequently encountered in sparse recovery applications. Sparsity-inducing penalties are effective to promote sparsity of signal and reduce the impulsive noise, and optimization problems based on the sparsity-inducing penalties have been proposed to robustly I recover the sparse signal. However, either most of these problems lack theoretical guarantee to robustly II recover the sparse signal, or the solving algorithms are computationally demanding. On the other hand, problems using WCSIP have been proposed with theoretical guarantees and efficient solvers for sparse recovery, but WCSIP has not been studied in robust I sparse recovery. Motivated by the two aspects above, this work mainly investigates the problem of robust sparse recovery via weakly convex regularization.

**B. Main Contribution**

The main contributions of this work are three-fold.

1) Firstly, the following general robust I signal sparse recovery problem is proposed

$$P_1 : \min_{\tilde{x},\tilde{e}} \lambda F(\tilde{x}) + G(\tilde{e}) \text{ s.t. } y = A\tilde{x} + \tilde{e},$$  \hspace{1cm} (7)

where $F(\cdot)$ and $G(\cdot)$ are sparsity-inducing penalties. We define Double Null Space Property (DNISP), a variant of NSP, for robust sparse recovery, use it to analyze $P_1$, and state that under mild conditions the sparse signal can be robustly II reconstructed via $P_1$.

2) Secondly, WCSIP is introduced to $P_1$, and the following non-convex problem is proposed

$$P_2 : \min_{\tilde{x},\tilde{e}} \lambda F_{\eta_1}(\tilde{x}) + G_{\eta_2}(\tilde{e}) \text{ s.t. } y = A\tilde{x} + \tilde{e},$$  \hspace{1cm} (8)

where $F_{\eta_1}(\cdot)$ and $G_{\eta_2}(\cdot)$ are WCSIPs with weakly convex parameters $\eta_1$ and $\eta_2$. Theoretical analyses of $P_2$ are provided, which reveal the uniqueness of optimum of $P_2$ and property of the generalized gradient.

3) Thirdly, an algorithm named Robust Projected Generalized Gradient (RPGG) is proposed to solve $P_2$. Theoretical convergence analysis of RPGG is provided, which states that if the measurement matrix satisfies DNISP with proper parameters, RPGG can reconstruct the sparse signal robustly II.

It is worth noticing that RPGG can recover sparse signal and detect outliers at the same time. In fact, by using RPGG, detecting outliers does not bring any additional burden, but contributes to sparse signal reconstruction.

**C. Organization**

In Section II, definitions and properties of WCSIP and NSP as well as PGG algorithm are reviewed. In Section III,
DNSP and Double Null Space Constant (DNSC) are defined and discussed. Then we use DNSP to analyze $P_1$ and reveal conditions of $P_1$ for robust II sparse recovery. In Section IV, the uniqueness of the optimum of $P_2$ is investigated, and a property of generalized gradient in a neighborhood of the optimum is analyzed. In Section V, RPGG is proposed with theoretical convergence guarantee. In Section VI, simulations showing performances of RPGG with different parameters are provided, and the performances of RPGG against the sparse noise and Generalized Gaussian Distribution (GGD) noise are also tested and compared with other robust I sparse recovery algorithms. Finally, Section VII concludes this work.

II. Preliminaries

In this section we define notations, review the definitions and properties of WCSIP and NSP, and concisely introduce the PGG algorithm.

A. Notations

The set of integers from 1 to $N$ is denoted by $[N]$, $T \subset [N]$ denotes its subset, and $T^C$ and $|T|$ are the complement and the cardinality of $T$, respectively. $[z]$ denotes the smallest integer no less than $z \in \mathbb{R}$. Let $z \in \mathbb{R}^N$ denote a vector, and its support set is denoted as $\text{supp}(z)$. Define $z_T$ as the vector keeping the entries of $z$ in $T$ and setting the other entries to 0. For a positive integer $k$, let $z_k$ denote the vector that keeps the $k$ largest entries (in magnitude) and sets the others to 0, and $z_{-k} = z - z_k$. The stack of vectors $u \in \mathbb{R}^N$ and $v \in \mathbb{R}^M$ is denoted by $[u;v] \in \mathbb{R}^{N+M}$. For a matrix $A \in \mathbb{R}^{M \times N}$, $A^T$ and $A^\dagger$ denote the transpose and pseudo inverse of $A$, respectively, and its largest and smallest singular values are denoted by $\sigma_{\text{max}}(A)$ and $\sigma_{\text{min}}(A)$, respectively. The null space of $A$ is denoted by $N(A)$, and $N(A)^\perp$ denotes the orthogonal complement of $N(A)$. Identity matrix is denoted by $I \in \mathbb{R}^{M \times M}$, and $[A \ I] \in \mathbb{R}^{M \times (N+M)}$ denotes the concatenation of $A$ and $I$. The indicator function of a set $S$ is denoted as

$$1_S(z) = \begin{cases} 1 & z \in S \\ 0 & z \notin S. \end{cases}$$

Throughout this paper, $x \in \mathbb{R}^N$ and $e \in \mathbb{R}^M$ denote the original signal and noise respectively, and the measurement signal is obtained as $y = Ax + e$.

B. Weakly Convex Sparsity-Inducing Penalty

In this subsection, the definitions of weakly convex function and sparsity-inducing penalty are introduced, then the WCSIP is defined and its properties are presented.

The $\rho$-convex function is defined as follows, which provides the definition of weakly convex function.

Definition 1 ([30], [33]) ($\rho$-convex function) If for any $u$, $z \in \mathbb{R}^N$ and $\lambda \in [0,1]$

$$F(\lambda u + (1-\lambda)z) \leq \lambda F(u) + (1-\lambda)F(z) - \lambda(1-\lambda)\rho\|u - z\|_2^2$$

holds, then function $F(\cdot)$ is $\rho$-convex. If $\rho < 0$, $F(\cdot)$ is said to be weakly convex; if $\rho > 0$, $F(\cdot)$ is said to be strongly convex.

The generalized gradient of a $\rho$-convex function is defined as below.

Definition 2 ([34]) If $F(z)$ is $\rho$-convex, $z \in \mathbb{R}^N$, then its generalized gradient is given by:

$$\partial F(z) = \{ F'(z) : \langle F'(z), u \rangle \leq D_F(z; u) \ \forall u \in \mathbb{R}^N \},$$

where $D_F(z; u) = \lim_{\theta \rightarrow 0^+} \frac{F(z + \theta u) - F(z)}{\theta}$.

If $F(\cdot)$ is convex, $\partial F(\cdot)$ is the subgradient of $F(\cdot)$, the generalized gradient is a generalization of subgradient. A simple method to evaluate the generalized gradient is presented as follows.

Proposition 1 ([33]) If $F(z)$ is $\rho$-convex, and $F(z) = H(z) + \rho|z|^p_2$, where $H(\cdot)$ is convex, then

$$\partial F(z) = \partial H(z) + 2\rho z.$$ 

The sparsity-inducing penalty is defined as below.

Definition 3 ([23]) The function $F(z) : \mathbb{R}^N \rightarrow \mathbb{R}$ is a sparsity-inducing penalty if

$$F(z) = \sum_{i=1}^N f(z_i),$$

where $f(z) : \mathbb{R} \rightarrow \mathbb{R}$ is a sparseness measure, which satisfies

1) $f(\cdot)$ is even, not identically zero, and $f(0) = 0$;
2) $f(\cdot)$ is non-decreasing on $[0, +\infty)$;
3) $\frac{f(\alpha z)}{z^p}$ is non-increasing on $(0, +\infty)$.

It can be verified that $\|\cdot\|_p$ with $0 \leq p \leq 1$ is a sparsity-inducing penalty.

The Weakly Convex Sparsity-Inducing Penalty (WCSIP) is defined as below.

Definition 4 ([30]) $F(z) = \sum_{i=1}^N f(z_i)$ is a weakly convex sparsity-inducing penalty, if

1) $f(\cdot)$ is a sparsity-inducing penalty;
2) $f(\cdot)$ is weakly convex.

Useful properties of the WCSIP are introduced as follows.

Property 1 ([23], [30]) If $F(z) = \sum_{i=1}^N f(z_i)$ is a sparsity-inducing penalty, then for any $a, b \in \mathbb{R}$ and $a, b \in \mathbb{R}^N$,

1) if $|a| \leq |b|$, then $f(a) \leq f(b)$,
2) $|f(a) - f(b)| \leq f(a + b) - f(a) + f(b), |F(a) - F(b)| \leq F(a + b) - F(a) + F(b)$;
3) if $f(\cdot)$ is weakly convex, then there exists $\alpha, \rho < 0$ such that $\forall x \in \mathbb{R}$,

- $\alpha = \lim_{z \rightarrow 0^+} \frac{f(z)}{z}$,
- $\rho f(\cdot)$ is $\rho$-convex and $\rho = \max \rho$ s.t. $f(z) - \rho z^2$ is convex,
- $f(z) - \alpha z - \rho z^2 \geq 0, f(z) \leq \alpha |z|$.

From Property 1, the weakly convex parameter $\eta$ of $F(\cdot)$, is defined as

$$\eta = -\frac{\rho}{\alpha}.$$
Given a sparsity-inducing penalty notion in the sparse recovery literature, is defined as follows. C. Null Space Property

The Null Space Property (NSP), which has been a key notion in the sparse recovery literature, is defined as follows.

**Definition 5** ([13]) Given a sparsity-inducing penalty \( F(\cdot) \), a matrix \( \mathbf{A} \) is said to satisfy the Null Space Property (NSP) with Null Space Constant (NSC) \( 0 < \gamma(\mathbf{A}, F, k) < \infty \), if for any \( \mathbf{z} \in \mathcal{N}(\mathbf{A}) \) and \( S \subset [N], |S| < k \), the following holds

\[
F(\mathbf{z}_S) < \gamma(\mathbf{A}, F, k) F(\mathbf{z}_{SC}).
\]

D. Projected Generalized Gradient Algorithm

The PGG algorithm is proposed to solve the following weakly convex sparse recovery problem [30]

\[
\min_{\mathbf{x}} F_\eta(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x}.
\]

With a proper initial point \( \tilde{x}^0 \), at the \( (t+1) \)th iteration, based on the current iterate \( \tilde{x}^t \), PGG takes a generalized gradient descent step

\[
\tilde{x}^{t+1} = \tilde{x}^t - \kappa \nabla F_\eta(\tilde{x}^t),
\]

where \( \kappa \) is the stepsize, and then takes a projection step to update the iterate

\[
\tilde{x}^{t+1} = \tilde{x}^{t+1} + \mathbf{A}^\dagger \left( \mathbf{y} - \mathbf{A}\tilde{x}^{t+1} \right).
\]

Theoretical analysis shows that if the NSC of \( \mathbf{A} \) is less than a constant determined by weakly convexity parameter \( \eta \), then after \( O(1/\kappa) \) iterations, PGG can recover the sparse signal with reconstruction error \( O(\kappa) \).

### III. THEORETICAL ANALYSIS ON GENERAL ROBUST I RECOVERY PROBLEM \( P_1 \)

In this section, DNSTS as well as Double Null Space Constant (DNSTC) are firstly defined, and the superiority of DNSTS compared with NSP is discussed. Then the robustness II of \( P_1 \) defined in (7) is studied in a theorem, of which the significance is emphasized by being compared with related theoretical results.

#### A. Double Null Space Property

For the problem studied in this work, the sparsity of both the signal and noise is considered, so we generalize the null space property to the double null space property.

**Definition 6** (DNSTC) Suppose that \( F(\cdot) \) and \( G(\cdot) \) are sparsity-inducing penalties. Any matrix \( \mathbf{A} \in \mathbb{R}^{M \times (N+M)} \) is said to satisfy DNSTS with DNSTS \( 0 < \tilde{\gamma}(\mathbf{A}, F, G, k_1, k_2) < +\infty \), such that

\[
F(\mathbf{u}_{S_1}) + G(\mathbf{v}_{S_2}) < \tilde{\gamma}(\mathbf{A}, F, G, k_1, k_2) \left( F \left( \mathbf{u}_{SC} \right) + G \left( \mathbf{v}_{SC} \right) \right)
\]

holds for any support sets \( S_1 \subset [N], |S_1| \leq k_1, S_2 \subset [M], |S_2| \leq k_2 \) and

\[
[\mathbf{u}; \mathbf{v}] \in \{[\mathbf{u}; \mathbf{v}] : \mathbf{u} \in \mathbb{R}^N, \mathbf{v} \in \mathbb{R}^M, \mathbf{A}[u; v] = 0 \}.
\]

In the most general setting, the penalties to promote the sparsity of signal and to penalize the impulsive noise are different in robust I sparse recovery problems. NSP cannot be utilized to analyze these robust I sparse recovery problems, whereas DNSTS can be suitably applied, and more detailed discussion on this issue will be provided in Section III-B and IV. If the penalties for the signal and noise are the same, we have the following property.

**Property 2** For any matrix \( \widehat{\mathbf{A}} \in \mathbb{R}^{M \times (N+M)} \) and any sparsity-inducing penalty \( F(\cdot) \),

\[
\tilde{\gamma}(\widehat{\mathbf{A}}, F, F, k_1, k_2) \leq \gamma(\widehat{\mathbf{A}}, F, k_1 + k_2)
\]

holds.

**Proof:** Let \( \tilde{\gamma} = \tilde{\gamma}(\widehat{\mathbf{A}}, F, F, k_1, k_2) \), \( \gamma = \gamma(\widehat{\mathbf{A}}, F, k_1 + k_2) \). From Definitions 5 and 6,

\[
F(\mathbf{u}_{S_1}; \mathbf{v}_{S_2}) = F(\mathbf{u}_{S_1}) + F(\mathbf{v}_{S_2})
\]

\[
\leq \tilde{\gamma} \left( F \left( \mathbf{u}_{SC} \right) + F \left( \mathbf{v}_{SC} \right) \right) = \tilde{\gamma} F \left( \mathbf{u}_{SC}; \mathbf{v}_{SC} \right),
\]

which concludes Property 2.

To visualize Property 2, a concrete example is shown here. Suppose \( \widehat{\mathbf{A}} = [\mathbf{A} \mathbf{I}] \), the spark of \( \mathbf{A} \) is \( s \), i.e., \( s = \min_x \|\mathbf{x}\|_0 \text{ s.t. } \mathbf{A}\mathbf{x} = \mathbf{0} \). Assume there exists an integer \( t \geq 1 \), such that a union set of any \( s \) columns of \( \mathbf{A} \) and any \( t \) columns of \( \mathbf{I} \) is linearly independent. One may readily examine that for any \( k = k_1 + k_2 \leq s \), \( \tilde{\gamma}(\mathbf{A}, \mathbf{I}, \|\cdot\|_0, \|\cdot\|_{k_1}, \|\cdot\|_{k_2}) \leq \tilde{\gamma}(\mathbf{A}, \mathbf{I}, \|\cdot\|_0, k) = \frac{k}{s+k} \) for \( k < s \) and \( \tilde{\gamma}(\mathbf{A}, \mathbf{I}, \|\cdot\|_0, s) = +\infty \) for \( k = s \).

Property 2 supports the claim that though NSP can be applied to analyze a robust I sparse signal problem using the same penalty function for the signal and noise, it is not as effective as DNSTS. More specifically, if \( \lambda = 1 \), and the penalties for signal and impulsive noise are both \( F(\cdot) \), then the general robust I sparse recovery problem \( P_1 \) defined in (7) can be rewritten as a classical sparse recovery problem

\[
\min_{\tilde{x}, \tilde{e}} F(\tilde{x}; \tilde{e}) \quad \text{s.t.} \quad \tilde{y} = \mathbf{A}[\tilde{x}; \tilde{e}],
\]
where $\hat{A} = [A \ I]$. According to [3], assume that $x$ and $e$ are $k_1$- and $k_2$-sparse, respectively (which means $e$ is an outlier vector), then if $\gamma(\hat{A}, F, k_1 + k_2) < 1$, (9) can reconstruct the sparse signal $x$. If $\gamma(\hat{A}, F, k_1 + k_2) \geq 1$, the property of the solution of (9) cannot be analyzed by NSP. However, if $\gamma(\hat{A}, F, F, k_1, k_2) < 1 \leq \gamma(\hat{A}, F, k_1 + k_2)$, according to the coming result from Section III-B, (9) can reconstruct $x^\#$ exactly. In conclusion, the condition on $\hat{A}$ for robust I sparse recovery based on DNSP is more easily satisfied than the condition based on NSP.

**Property 3** For any matrix $\hat{A} \in \mathbb{R}^{M \times (N+M)}$ and any sparsity-inducing penalties $F(\cdot)$ and $G(\cdot)$,

$$\gamma(\hat{A}, \| \cdot \|_0, \| \cdot \|_1, k_1, k_2) \leq \gamma(\hat{A}, F, F, G, k_1, k_2) \leq \gamma(\hat{A}, \| \cdot \|_1, \| \cdot \|_2, k_1, k_2).$$

Property 3 discusses the relations of DNSCs of matrix $\hat{A}$ with different sparsity-inducing penalties, including $\ell_1$ norm, $\ell_0$ pseudo norm, and other sparsity-inducing penalties. According to this property, with $k_1$ and $k_2$ fixed, the DNSC of $\hat{A}$ with $\ell_0$ pseudo norm is the smallest among the three, while the DNSC of $\hat{A}$ with $\ell_1$ norm is the largest. This property reveals the difference of the performances of $P_1$ defined (7) using different sparsity-inducing penalties to promote the sparsity of signals and reduce impulsive noise, and more detailed analysis will be presented in Section III-B.

**Property 4** Suppose $\hat{A} = [A \ I]$, where $A \in \mathbb{R}^{M \times N}$ is a Gaussian matrix with its i.i.d. entries drawn from $\mathcal{N}(0, \frac{1}{M})$. Assume

$$M \geq c_1 \left( \max(k_1 \ln \frac{N}{k_1} + k_2) \right).$$

(10)

Then for any given sparsity-promoting functions $F$ and $G$, $\gamma(\hat{A}, F, G, k_1, k_2) < 1$ holds with probability at least $1 - e^{-c_2 M}$, where $c_1$ and $c_2$ are numerical constants.

**Proof:** See Appendix A.

This property shows if $\hat{A}$ consists of a Gaussian matrix $A$ and an identity matrix $I$ with $M = O\left( \max(k_1 \ln \frac{N}{k_1} + k_2) \right)$, then the DNSP of $\hat{A}$ is satisfied with overwhelming probability. It is remarkable that according to Corollary 9.34 of [3], $\gamma(A, F, k_1 + k_2) \leq \gamma(\hat{A}, F, k_1 + k_2) < 1$ at least requires $M = O\left( (k_1 + k_2) \log \frac{N}{k_1 + k_2} \right)$, which is larger than (10) in terms of the order.

**B. Robustness II of $P_1$**

In this subsection, based on DNSP, we study the condition for $P_1$ defined in (7) to robustly II reconstruct the sparse signal. Remind that $P_1$ is a general problem which includes problems (2), (3), (4), and (5) as special cases.

In the following analysis, $\hat{\gamma}(\hat{A}, F, G, k_1, k_2)$ is denoted as $\hat{\gamma}$ for short, where $\hat{A} = [A \ I]$.

**Theorem 1** Suppose that $(x^*, e^*)$ is the global optimum of model $P_1$ defined in (7). If

$$\hat{\gamma} < \min \left( \lambda, \frac{1}{\lambda} \right),$$

(11)

then for signal $x$ and noise $e$,

$$F(x - x^*) + G(e - e^*) \leq \frac{2(1 + \hat{\gamma})}{\beta} \left( F(x - k_1) + G(e - k_2) \right)$$

holds, where $\lambda$ is the regularization parameter of $P_1$, and $\beta = \min(\lambda, \frac{1}{\lambda}) - \hat{\gamma}$.

**Proof:** See Appendix B.

According to the definition of sparsity-inducing penalty and the second term in Property 1, for any $a, b, c \in \mathbb{R}^N$, the following holds.

- $F(a - b) \geq 0$, and $F(a - b) = 0 \iff a = b$
- $F(a - b) = F(b - a)$
- $F(a - b) \leq F(a - c) + F(c - b)$

Thus, $F(\cdot)$ and $G(\cdot)$ are distance functions. Theorem 1 reveals that if (11) holds, the bound of the sum of distances, defined by $F(\cdot)$ and $G(\cdot)$, between the estimated signal and the original signal, and between the estimated noise and the original noise is proportional to $F(x - k_1) + G(e - k_2)$. If $F(\cdot)$ and $G(\cdot)$ are continuous, small $\|x - k_1\|_2$ and $\|e - k_2\|_2$ lead to small $F(x - k_1) + G(e - k_2)$, then according to Theorem 1 $F(x - x^*) + G(e - e^*)$ is small, so the estimated signal $x^*$ is close to the original signal $x$, i.e., the reconstruction error of $P_1$ defined in (7) is small. Specifically, if $x$ is $k_1$-sparse, and $e$ is an outlier vector with sparsity $k_2$, then $P_1$ can exactly recover $x$ and detect $e$.

For some concrete instances of $F(\cdot)$ and $G(\cdot)$, the reconstruction error $\|x - x^*\|_2$ has been studied. Table II lists the reconstruction error bounds in the literature, and compares them with the result deduced from Theorem 1. For the first and the fourth cases in Table II, the orders of the bounds from [21] and [25] are the same as the orders of the bounds deduced from Theorem 1. For the second case, because $\|e - k_2\|_p < \|e\|_p$, the error bound deduced from Theorem 1 is better than the corresponding result in [24]. For the other cases in Table II, to the best of our knowledge, no result is provided in the existing literature, whereas we can use Theorem 1 to guarantee their reconstruction error bounds.

Theorem 1 states that the condition for $P_1$ defined in (7) to robustly II reconstruct the sparse signal is that the DNSP of $\hat{A}$ satisfies (11). According to Property 3 and Theorem 1, if $\ell_0$ pseudo norm is used to promote the sparsity of signal and impulsive noise, it is relatively easy for $\hat{A}$ to satisfy (11), i.e., the condition for $P_1$ to robustly II recover the sparse vector $x$, and if $\ell_1$ norm is used, then it is relatively hard for $\hat{A}$ to satisfy (11). If other sparsity-inducing penalties are used, then (11) is easier to be satisfied than that corresponding to $\ell_1$ norm, but harder than that corresponding to $\ell_0$ pseudo norm. Such analysis further motivates the usage of non-convex sparsity-inducing functions such as weakly convex functions instead of $\ell_1$ norm.

**IV. Uniqueness of Optimum of Weakly Convex Robust I Sparse Recovery Problem $P_2$**

If the sparsity-inducing penalties in $P_1$ defined in (7) are weakly convex, $P_1$ becomes $P_2$ defined in (8). Because $F_{\eta_1}(\cdot)$ and $G_{\eta_2}(\cdot)$ in $P_2$ are continuous and weakly convex, there are
Suppose that \( \hat{\lambda} \) and \( \rho \) are defined in (7), which are k1- and k2-sparse, respectively; \( \lambda \) is a regularization parameter to balance \( F_{\eta_1} \) and \( G_{\eta_2} \). In this section, \( \gamma_{\eta_1, \eta_2}(A, F_{\eta_1}, G_{\eta_2}, k_1, k_2) \) is denoted as \( \tilde{\gamma}_{\eta_1, \eta_2} \) for short, where \( A = [A \ I] \).

**Lemma 1** Suppose that \( x \) and \( e \) are k1- and k2-sparse, respectively. If
\[
\tilde{\gamma}_{\eta_1, \eta_2} < \min \left( \lambda, \frac{1}{\lambda} \right),
\]
then for any
\[
[\tilde{x}; \tilde{e}] \in U = \{ [x; e] : \|u\|_2 \leq r_1, \|v\|_2 \leq r_2 \},
\]
we have
\[
J(\tilde{x}, \tilde{e}) - J(x, e) \geq C_{f, r_1} \|u\|_2 + C_{g, r_2} \|v\|_2 - C_3 \|Au + v\|_2,
\]
where
\[
C_{f, r_1} = \frac{f(r_1)}{r_1} \min(1, \lambda) - \frac{1}{1 + \gamma_{\eta_1, \eta_2}},
\]
\[
C_{g, r_2} = \frac{g(r_2)}{r_2} \min(1, \lambda) - \frac{1}{1 + \gamma_{\eta_1, \eta_2}},
\]
\[
C_3 = \frac{\max(1, \lambda) \alpha \sqrt{M + N + \sqrt{2} \max(C_{f, r_1}, C_{g, r_2})}}{\sigma_{\min}(A)}.
\]
with \( \alpha = \max(\alpha_1, \alpha_2) \). Moreover, if \( y = Ax + \tilde{e} \), then
\[
J(\tilde{x}, \tilde{e}) - J(x, e) \geq C_{f, r_1} \|u\|_2 + C_{g, r_2} \|v\|_2,
\]
where \( \lambda \) is the regularization parameter defined in (8), \( J(x; e) \) is defined in (12), and \( \alpha_1 \) and \( \alpha_2 \) are defined in (13) and the descriptions therein.

**Proof:** See Appendix C.

As \( P_2 \) defined in (8) is a special case of \( P_1 \) defined in (7), according to Theorem 1, if (15) holds and \( x \) and \( e \) are k1- and k2-sparse, respectively, then \( y = Ax + \tilde{e} \) is the global minimum of \( P_2 \). Lemma 1 further reveals that for any \( [\tilde{x}; \tilde{e}] \in \{ [x; e] : y = Ax + \tilde{e} \} \),
\[
J(\tilde{x}, \tilde{e}) - J(x, e) \geq C_{f, r_1} \|u\|_2 + C_{g, r_2} \|v\|_2 > 0
\]
holds, which means that \( y = Ax + \tilde{e} \) is the unique global minimum of \( P_2 \). What's more, \( J(\tilde{x}, \tilde{e}) \) minus \( J(x, e) \) is \( \min(\{C_{f, r_1}, C_{g, r_2}\}) \) times greater than the Euclidean distance between \( [\tilde{x}; \tilde{e}] \) and \( [x; e] \), where \( [\tilde{x}; \tilde{e}] \in U \cup \{ [x; e] : y = Ax + \tilde{e} \} \).

**Theorem 2** Suppose that \( x \) and \( e \) are k1- and k2-sparse, respectively. If (15) holds, \( r_1, r_2 > 0 \) and the weakly convex parameters \( \eta_1, \eta_2 \) satisfy
\[
\eta_1 \leq \frac{1}{r_1} \frac{\beta_{\eta_1, \eta_2} \max(\frac{1}{\lambda}, 1)}{\beta_{\eta_1, \eta_2} \max(\frac{1}{\lambda}, 1) + (1 + \gamma_{\eta_1, \eta_2})},
\]
\[
\eta_2 \leq \frac{1}{r_2} \frac{\beta_{\eta_1, \eta_2} \max(\lambda, 1)}{\beta_{\eta_1, \eta_2} \max(\lambda, 1) + (1 + \gamma_{\eta_1, \eta_2})},
\]
then \( [x; e] \) is the unique local minimum of \( P_2 \) in \( U = \{ [x; e] : \|u\|_2 < r_1, \|v\|_2 < r_2 \} \), where \( \beta_{\eta_1, \eta_2} = \min(\lambda, \frac{1}{\lambda}) - \tilde{\gamma}_{\eta_1, \eta_2} \), and \( \lambda \) is the regularization parameter of \( P_2 \) defined in (8).

**Proof:** See Appendix D.
Theorem 2 reveals that if (15) holds, and \( x \) and \( e \) are \( k_1 \)- and \( k_2 \)-sparse, respectively, then \( [x; e] \) is the only local minimum of \( P_2 \) defined in (8) in a neighborhood of \([x; e]\), and the size of \( U \) is determined by the weakly convex parameters.

The following lemma reveals the property of generalized gradient in \( U \), which contributes to proving Theorem 2.

**Lemma 2** Suppose that \( x \) and \( e \) are \( k_1 \)- and \( k_2 \)-sparse, respectively. If (15) holds, \( r_1, r_2 > 0 \) and the weakly convex parameters satisfy (17) and (18), then for any \( \nabla F_{\eta_1}(\hat{x}) \in \partial F_{\eta_1}(\hat{x}) \), \( \nabla G_{\eta_2}(\hat{e}) \in \partial G_{\eta_2}(\hat{e}) \) and

\[
[x; \hat{e}] \in \{ [\hat{x}; \hat{e}] : 2C_3\|A\hat{x} + \hat{e} - y\|_2 \leq C_{f,r_1}\|u\|_2 + C_{g,r_2}\|v\|_2 \} \cap U,
\]

we have

\[
\lambda u^T \nabla F_{\eta_1}(\hat{x}) + v^T \nabla G_{\eta_2}(\hat{e}) \geq \frac{C_{f,r_1}}{4}\|u\|_2 + \frac{C_{g,r_2}}{4}\|v\|_2,
\]

where \( U \), \( C_{f,r_1}, C_{g,r_2} \), and \( \beta_{\eta_1, \eta_2} \) are defined in Lemma 1, and \( \partial F_{\eta_1}(\hat{x}) \) and \( \partial G_{\eta_2}(\hat{e}) \) are generalized gradients of \( F_{\eta_1}(\hat{x}) \) and \( G_{\eta_2}(\hat{e}) \), respectively.

**Proof:** See Appendix E.

This lemma indicates that if (15) holds, the weakly convex parameters satisfy (17) and (18), \( \kappa > 0 \) is small enough, and \([\hat{x}; \hat{e}] \in U \) is close to the affine subspace \( y = A\hat{x} + \hat{e} \), then \([\hat{x}; \hat{e}] - \kappa[\nabla F_{\eta_1}(\hat{x}); \nabla G_{\eta_2}(\hat{e})] \) is closer to \([x; e]\) than \([\hat{x}; \hat{e}] \). Specifically, we have the following remark.

**Remark 1** Lemma 2 implies that if \( A\hat{x} + \hat{e} = y \), \( \|u\|_2 \leq r_1 \), \( \|v\|_2 \leq r_2 \), (15), (17) and (18) are satisfied, then

\[
\lambda u^T \nabla F_{\eta_1}(\hat{x}) + v^T \nabla G_{\eta_2}(\hat{e}) \geq \frac{C_{f,r_1}}{4}\|u\|_2 + \frac{C_{g,r_2}}{4}\|v\|_2 \geq \frac{\min(C_{f,r_1}, C_{g,r_2})}{4}
\]

holds. Thus, in a neighborhood of the local optimum, the inner product of the generalized gradient and the unit vector along the direction pointing from \((\hat{x}, \hat{e})\) towards \((x, e)\) is positive, namely, the negative generalized gradient is a descend direction of the problem \( P_2 \) in a neighborhood of the local optimum.

**V. ROBUST PROJECTED GENERALIZED GRADIENT**

In this section, the RPGG algorithm is proposed to solve \( P_2 \) defined in (8), and its convergence analysis is provided.

From Remark 1 in Section IV, with a proper initial point, if we take a generalized gradient descent step and a projection step to update the variables iteratively, then the iterate approaches the original sparse signal and noise gradually. Following such observation, RPGG is proposed in Algorithm 1.

At the beginning of RPGG, an initial point \([\hat{x}^0; \hat{e}^0]\) needs to be provided; a popular initialization is \([\hat{x}^0; \hat{e}^0] = A^\dagger y\). Then at \((t + 1)\)-th iteration of the algorithm, the estimated signal and noise take the generalized gradient descent steps (19) and (20), respectively. \( \nabla F_{\eta_1}(\hat{x}^t) \in \partial F_{\eta_1}(\hat{x}^t) \) and \( \nabla G_{\eta_2}(\hat{e}^t) \in \partial G_{\eta_2}(\hat{e}^t) \) are generalized gradients of \( F_{\eta_1}(\hat{x}^t) \) and \( G_{\eta_2}(\hat{e}^t) \), respectively. \( \kappa\lambda \) and \( \kappa \) are the corresponding stepsizes, which should be small to reduce the reconstruction error. One may refer to Theorem 3 and the corresponding discussions for more details. After the generalized gradient steps, the cost function of \( P_2 \) defined in (8) tends to decrease. Then a step of projection (21) is taken to guarantee that the linear constraint of \( P_2 \) is satisfied. The algorithm terminates when the iteration number \( t + 1 \) exceeds a certain number.

Though both PGG and RPGG are dominated by the generalized gradient descent and projection step, there are significant differences between them. In PGG, only the estimated signal \( \hat{x}^t \) is updated and the projection is taken to guarantee \( y = A\hat{x} + \hat{e} \), while in RPGG, the estimated signal and noise are both updated, and the projection aims to guarantee \( y = A\hat{x} + \hat{e} \). Moreover, the PGG algorithm is proposed to solve a weakly convex sparse recovery problem, where the WCSIP is only used to promote the sparsity of the signal, while RPGG aims to solve \( P_2 \) which adopts WCSIPs to not only promote the sparseness of the signal but also contribute to reduce the impulsive noise. Thus, as a generalization of PGG, RPGG is specifically designed for robust I sparse recovery.

The following theorem reveals a relation between the reconstruction error and the required number of iterations in RPGG.

**Theorem 3** Define \( r_1 = \|\hat{x}^0 - x_{k_1}\|_2, r_2 = \|\hat{e}^0 - e_{k_2}\|_2 \), \( \hat{A} = [A, I] \). If (15) holds, and the weakly convex parameters satisfy (17) and (18), then after at most

\[
\left\lfloor 8 \sqrt{r_1^2 + r_2^2} \frac{\kappa}{\min(C_{f,r_1}, C_{g,r_2})} \right\rfloor \text{iterations, the estimated point of RPGG satisfies}
\]

\[
\sqrt{\|\hat{x} - \hat{x}'\|_2^2 + \|\hat{e} - \hat{e}'\|_2^2} \leq \left( 4\sqrt{2} \left( \frac{M\alpha_1^2 + N\lambda^2\alpha_1^2}{\min(C_{f,r_1}, C_{g,r_2})} + \sqrt{N\lambda^2\alpha_1^2 + M\alpha_2^2} \right) \kappa + \frac{8\sqrt{2}C_3\sigma_{\max}(\hat{A})}{\min(C_{f,r_1}, C_{g,r_2})} + 1 \right) \|x_{-k_1}\|_2 + \|e_{-k_2}\|_2 \right),
\]

where \( C_{f,r_1} \) and \( C_{g,r_2} \) are defined in Lemma 1, \( \alpha_1 \) and \( \alpha_2 \) are defined in (13) and the descriptions therein, and \( \kappa \) is the stepsize defined in Algorithm 1.
A. Performance of RPGG with Different Parameters

In this part, the performance of the RPGG is examined.

\textbf{Proof:} See Appendix F.

According to Theorem 3, if (15) holds and the related parameters are chosen properly, RPGG can robustly reconstruct the signal $x$ and detect the noise $e$, with error $\sqrt{\|x-x^*\|^2 + \|e-e^*\|^2}$ bounded by $O(\kappa) + O(\|x-k_1\|_2) + O(\|e-k_2\|_2)$. If the signal and noise are $k_1$- and $k_2$-sparse (which means $e$ is an outlier vector), respectively, then the upper bound of the error is proportional to $\kappa$. Thus, smaller stepsize $\kappa$ leads to smaller error bound, whereas more iterations are required according to (22).

Weakly convex parameters $\eta_1$ and $\eta_2$ control not only how mild the condition is that the matrix $A$ has to satisfy to guarantee robust II recovery, but also how difficult it is for RPGG to solve $P_2$ defined in (8). More concretely, on the one hand, as discussed in the previous section, $\eta_1$ and $\eta_2$ characterize the weakly convexity of the WCSIPs $F_{\eta_1}(\cdot)$ and $G_{\eta_2}(\cdot)$, which reflects the similarity between $\ell_0$ pseudo norm and these WCSIPs. If $\eta_1$ and $\eta_2$ are large, the WCSIPs are similar to $\ell_0$ pseudo norm. According to Property 3, Theorem 1 and Lemma 1, to guarantee the robustness of $P_2$, the condition that $A$ has to satisfy is relatively mild. On the other hand, Theorems 2 and 3 imply that if RPGG can recover the signal and detect the outliers robustly II, $r_1$ and $r_2$ are required to be inversely proportional to $\eta_1$ and $\eta_2$, respectively. If $\eta_1$ and $\eta_2$ are large, $r_1$ and $r_2$ are restricted to be small, then it is relatively difficult to choose an initial point for RPGG properly, but if $\eta_1$ and $\eta_2$ are small, then $r_1$ and $r_2$ are tolerated to be large, so it is relatively easy to choose an initial point.

VI. NUMERICAL SIMULATIONS

In Section VI-A we demonstrate the performances of RPGG with different parameters including stepsize $\kappa$, weakly convex parameters $\eta_1$ and $\eta_2$, and regularization parameter $\lambda$. In Section VI-B the performance of RPGG will be compared with other state-of-the-art robust I sparse recovery methods.

Throughout this section, $A \in \mathbb{R}^{M \times N}$ is a Gaussian matrix with entries following distribution $\mathcal{N}(0, \frac{1}{M})$ independently, where $N = 100$, $M = 50$. The signal $x$ is $k_1$-sparse, its support set is chosen randomly, and the nonzero elements are independently drawn from the standard Gaussian distribution. In this section, we adopt $F_{\eta_1}(\cdot)$ and $G_{\eta_2}(\cdot)$ as the following sparsity inducing function [30], [32].

$$f_{\eta}(z) = (|z| - \sigma z^2){\bf 1}_{|z| \leq \frac{1}{\sigma}}(z) + \frac{1}{4\sigma}{\bf 1}_{|z| > \frac{1}{\sigma}}(z),$$

because one can easily control the weakly convexity $\eta = \sigma$ and calculate the generalized gradient by

$$(\text{sign}(z) - 2\sigma){\bf 1}_{|z| \leq \frac{1}{\sigma}}(z) + \frac{1}{2\sigma}{\bf 1}_{|z| > \frac{1}{\sigma}}(z) \in \partial f_{\eta}(z).$$

The initial point of RPGG is chosen as $[\hat{x}^0; \hat{e}^0] = [0; 0]$.

A. Performance of RPGG with Different Parameters

In this part, the performance of the RPGG is examined.

1) Stepsize $\kappa$: At first we examine the ability of RPGG with different stepsize $\kappa$ to reconstruct $x$ from $y = Ax + e$, where $e = e_{sp} + e_{de}$, $e_{sp}$ is a vector of outliers with sparsity $k_2$, and $e_{de}$ denotes dense noise. We set $t_{max} = 10^6$, $\eta_1 = \eta_2 = 10^{-0.2} = 0.6310$, $k_1 = k_2 = 5$, and $\lambda = 1$. The nonzero elements of $e_{sp}$ are i.i.d. drawn from $\mathcal{N}(2, 1)$. The elements of $e_{de}$ are i.i.d. drawn from $\mathcal{N}(0, \sigma^2)$, with $\sigma^2$ changing from 0 to $10^{-2}$. Thus, one can easily check that $E\|e_{sp}\|_2 \gg E\|e_{de}\|_2$. For each $\sigma^2$, $\kappa$ is varied from $10^{-5}$ to $10^{-1}$. We conduct RPGG 100 times for each pair of $\sigma^2$ and $\kappa$. Because RPGG can recover signal and detect outlier, here we define the original SNR as

$$\text{SNR}_{org} = 10\log_{10} \frac{E\|x\|_2}{E\|e_{de}\|_2},$$

and reconstructed SNR as

$$\text{SNR}_{rec} = 10\log_{10} \frac{E\|x\|_2}{E\|x - x^*\|_2}.$$
According to Fig. 2, for fixed \( k_1 \) and \( k_2 \), if \( \eta_1 \) and \( \eta_2 \) are too large or too small, the exact reconstruction probability of RPGG tends to be small, and the largest sparsity of signal or noise that RPGG can exactly recover with one hundred percents also gets small. Thus, we can see that Fig. 2 is coincident with the analysis in Section V that a small \( \eta \) parameter respectively.

Fig. 2. Probability of exact reconstruction of RPGG. From left to right, the corresponding \( \eta_1 \) for each column of subfigures is \( 0, 10^{-0.6}, 10^{-0.2}, 10^0, 10^0.6 \), respectively; from top to bottom, the corresponding \( \eta_2 \) for each row of subfigures is \( 0, 10^{-0.6}, 10^{-0.2}, 10^0, 10^0.6 \), respectively.

3) Regularization parameter \( \lambda \): Now we test the performance of RPGG with different values of the regularization parameter \( \lambda \). The settings of \( \mathbf{e}^\#, \kappa, t_{\text{max}}, k_1 \) and \( k_2 \) are the same as the previous part. We set \( \eta_1 = \eta_2 = 0.6310 \), and \( \lambda \) is varied from \( 10^{-0.4} \) to \( 10^{0.4} \). For each set of \( (k_1, k_2, \lambda) \), RPGG is conducted 100 times. For different \( \lambda \), we plot the exact reconstruction probability of RPGG in Fig. 3.

According to Fig. 3, if \( \lambda \) is large, the largest \( k_1 \) that RPGG can recover is relatively large, but the largest \( k_2 \) that RPGG can detect is relatively small, while if \( \lambda \) is small, the largest recoverable \( k_1 \) becomes small, but the largest detectable \( k_2 \) is relatively large. When \( \lambda = 1 \), the performance of RPGG is almost the best.

B. Performances Comparison with Reference Methods

In this subsection, we test the performances of RPGG against outliers and impulsive noise with Generalized Gaussian Distribution (GGD), and compare them with the results of several state-of-the-art sparse recovery algorithms.

1) Outliers: In this experiment, we test the ability of RPGG to recover the sparse signal \( x \) from \( y = Ax + e \), where \( e \) is a vector of outliers with sparsity \( k_2 \) and nonzero elements \( i.i.d. \) drawn from \( N(2, 1) \), and compare RPGG with other robust I sparse recovery algorithms, including \( \ell_1-\ell_1 \) (solved by ADMM), Robust Lasso (RLasso), \( \ell_1-\ell_\max \) (solved by ADMM), \( \ell_0-\text{LAD} \) [27], \( \ell_1-\text{OMP} \) [6], LIHT [28] and HIHT [29]. Among them, \( \ell_0-\text{LAD}, \ell_1-\text{OMP}, \text{LIHT} \) and HIHT require that the (estimated) sparsity is known. For RPGG, we set \( \kappa = 10^{-4}, t_{\text{max}} = 2 \times 10^5 \), and \( \lambda = 1 \). We also consider two pairs of weakly convex parameters, \( \eta_1 = \eta_2 = 0.6310 \) and \( \eta_1 = \eta_2 = 0 \), which are denoted by RPGG and \( \ell_1-\ell_1^{(\text{RPGG})} \), respectively. The parameters of the other algorithms are set optimally or sub-optimally according to the corresponding references or cross validation.

Two experiments are provided here to test the performance of RPGG with different sparsities and measurement numbers. In the first experiment, we set \( M = 50, N = 100 \), and \( k_1 \) and \( k_2 \) are varied from 1 to 10. We conduct RPGG and other methods 100 times for each set of \( (k_1, k_2) \), respectively. For each algorithm, the corresponding exact reconstruction probability is plotted in Fig. 4, from which we can see that the performance of the proposed algorithm is the best in terms of recoverable sparsity levels.

In the second experiment, we set \( N = 100, k_1 = 10, k_2 = 6 \) and \( M \) varies from 20 to 100. RPGG and other algorithms are conducted 100 times for each \( M \). The exact reconstruction probabilities of each algorithm are evaluated. As shown in Fig. 5, the results illustrate the superiority of RPGG over the other algorithms: the required measurement numbers of RPGG for 100% successful reconstruction is the least.

2) GGD noise with \( p \leq 1 \): Because GGD can character many varieties of impulsive noises, we test the performance of RPGG to recover the sparse signal \( x \) from \( y = Ax + e \), and the elements of \( e \) are \( i.i.d. \) drawn from \( \text{GGD}(0, \delta, p) \), where 0 refers to the position parameter, \( \delta \) is shape parameter, and \( p \) is the order. We set \( k_1 = 10, \delta = 0.01, \) and \( p = 0.5 \) or \( p = 1 \). For RPGG, we set \( \kappa = 10^{-4} \), \( t_{\text{max}} = 2 \times 10^5 \). When \( p = 0.5 \), we take \( \eta_1 = 10^{0.2} \approx 1.5849, \eta_2 = 10^{0.1} \approx 1.2589 \), and when \( p = 1 \) we take \( \eta_1 = 10^{0.3} \approx 1.9953, \eta_2 = 10^{0.2} \approx 1.5849 \).
The $\lambda$ is varied from $10^{-1}$ to $10^3$. The parameters of the other algorithms are tuned to be optimal. For each set of $(p, \lambda)$, we run RPGG 100 times, and evaluate the reconstruction SNR. The results for $p = 1$ and $p = 0.5$ are plotted in Fig. 6 and Fig. 7.

According to the figures, RPGG with proper parameters outperforms the other reference methods: Though the performance of LIHT is similar to that of RPGG, LIHT requires the sparsity of signal as a prior, while RPGG does not need such information.

VII. CONCLUSION

In this paper, aiming at sparse recovery against impulsive noise, we propose a general optimization problem, of which the sparsity-inducing functions are adopted to not only promote the sparseness of the signal, but also contribute to reduce the impulsive noise. The Double Null Space Property (an extension of Null Space Property) is defined, and we use it to analyze the robustness II of the general problem and deduce sufficient conditions to recover the sparse signal. As a class of sparsity-inducing functions with good properties, the weakly convex sparsity-inducing function is firstly introduced to robust I sparse recovery. We analyze the uniqueness of both global and local optima of the weakly convex problem and reveal a property of the generalized gradient in a neighborhood of the optimum. In addition, Robust Projection Generalized Gradient (RPGG) algorithm is proposed to solve the weakly convex problem. We prove that, under mild conditions, if the signal is sparse and the impulsive noise consists of sparse vector of outliers and small bounded noise, then RPGG with tuned parameters can reconstruct the sparse signal with small error. Simulations demonstrate that with properly chosen parameters, RPGG outperforms other robust I sparse recovery algorithms.

APPENDIX A

PROOF OF PROPERTY 4

According to [20], $\hat{\mathbf{A}}$ with $M \geq c_1 \max (k_1 \log \frac{N}{k_1} , k_2)$ satisfies a generalization of Restricted Isometry Property (generalized RIP) with probability at least $1 - e^{-c_2 M}$, where $c_1$
and $c_2$ are constants determined by the generalized Restricted Isometry Constant $\delta$. Thus, what we need is to prove that if the generalized RIP is satisfied, then $\gamma(\hat{A}, F, G, k_1, k_2) < 1$. In this subsection, we first provide the definition of generalized RIP, and then investigate the relation between DNSP and generalized RIP. At last, some discussions are given to complete the proof.

**Definition 7 (20)** For any matrix $\hat{A} \in \mathbb{R}^{M \times (N+M)}$, define the generalized RIP-constant $\delta(\hat{A}, k_1, k_2)$ by the infimum value of $\delta$ such that

$$(1 - \delta) \left( \|z\|^2_f + \|w\|^2_f \right) \leq \|\hat{A}[z; w]\|^2_f \leq (1 + \delta) \left( \|z\|^2_f + \|w\|^2_f \right)$$

holds for any $k_1$-sparse $z \in \mathbb{R}^N$ and $k_2$-sparse $w \in \mathbb{R}^M$.

The following lemma gives some insight about $\delta(\hat{A}, k_1, k_2)$.

**Lemma 3** For any $u^1, u^2 \in \mathbb{R}^N$ and $v^1, v^2 \in \mathbb{R}^M$, Suppose that $|\text{supp}(u^1) \cup \text{supp}(u^2)| \leq k_1$, $\text{supp}(u^1) \cap \text{supp}(u^2) = \emptyset$, and $|\text{supp}(v^1) \cup \text{supp}(v^2)| \leq k_2$, $\text{supp}(v^1) \cap \text{supp}(v_2) = \emptyset$, then

$$\|\hat{A}[u^1; v^1]\|_2^2 = (1 + t)\|u^1; v^1\|_2^2$$

and

$$\left| \langle \hat{A}[u^1; v^1], \hat{A}[u^2; v^2] \rangle \right| \leq \sqrt{\delta(\hat{A}, 2k_1, 2k_2) - t^2} \|u^1; v^1\|_2 \|u^2; v^2\|_2$$

holds for some $|t| \leq \delta(\hat{A}, 2k_1, 2k_2)$.

**Proof:** The proof is a direct combination of Lemma 2.2 of [20] and (6.22) of [3]. The following lemma reveals the relation between generalized RIP and DNSP corresponding to $\ell_1$ norm.

**Lemma 4** Assume $R \geq 1$, $\sqrt{\frac{k_1}{k_2}} \in [\frac{1}{2}, R]$. If $\delta(\hat{A}, 2k_1, 2k_2) \leq \delta < \frac{8}{2(1+2R^2)}$, then

$$\gamma(\hat{A}, \cdot \cdot \cdot ) \leq \delta \leq \frac{8}{2(1+2R^2)}$$

holds.

**Proof:** Suppose $u \in \mathbb{R}^N$ and $v \in \mathbb{R}^M$ satisfy $\hat{A}[u; v] = 0$. We first decompose $[u; v]$ by its index sets, which are constructed as follows.

- $T_{u_0}$ denotes the index set of the entries with $k_1$ largest absolute values in $u$;
- $T_{u_i}$ ($i \geq 1$) denotes the index set of the entries with $k_1$ largest absolute values in $u_{T_{u_0} \cup \ldots \cup T_{u_i}}$.

Following a similar way, $T_{v_j}$ ($j = 0, 1, 2, \ldots$) is defined. According to Lemma 3 and the fact that $0 = \hat{A}[u; v] = \hat{A}[u_{T_{u_0}}; v_{T_{v_0}}] + \sum_{i \geq 1} \hat{A}[u_{T_{u_0} \cup \ldots \cup T_{u_i}}; v_{T_{v_0}}]$, we have

$$(1 + t)\|u_{T_{u_0}}; v_{T_{v_0}}\|_2^2 \leq \|\hat{A}[u_{T_{u_0}}; v_{T_{v_0}}]\|_2^2$$

and

$$\left| \langle \hat{A}[u_{T_{u_0}}; v_{T_{v_0}}], \hat{A}[u_{T_{u_0}}; v_{T_{v_0}}] \rangle \right| \leq \sqrt{\delta^2 - t^2} \|u_{T_{u_0}}; v_{T_{v_0}}\|_2 \|u_{T_{u_0}}; v_{T_{v_0}}\|_2$$

According to square root lifting inequality (Lemma 6.14 of [3]), one can obtain that

$$\sum_{i \geq 1} \|u_{T_{u_0} \cup \ldots \cup T_{u_i}}; v_{T_{v_0}}\|_2^2 \leq \sum_{i \geq 1} \|u_{T_{u_0} \cup \ldots \cup T_{u_i}}\|_2^2 + \|v_{T_{v_0}}\|_2^2$$

$$\leq \frac{1}{\sqrt{k_1}} \|u_{T_{u_0}}\|_1 + \frac{1}{4} \|u_{T_{u_0}}\|_2 + \frac{1}{\sqrt{k_2}} \|v_{T_{v_0}}\|_1 + \frac{1}{4} \|v_{T_{v_0}}\|_2.$$

Plugging the above inequality and

$$\|u_{T_{u_0}}\|_2 + \|v_{T_{v_0}}\|_2 \leq \sqrt{2} \|u_{T_{u_0}}; v_{T_{v_0}}\|_2$$

into (25) yields

$$\|u_{T_{u_0}}\|_2 + \|v_{T_{v_0}}\|_2 \leq \frac{\sqrt{2}\delta}{\sqrt{1 - \delta^2}} \|u_{T_{u_0}}\|_1 + \frac{\delta}{2\sqrt{2\sqrt{1 - \delta^2}}} \|v_{T_{v_0}}\|_2$$

$$+ \frac{\sqrt{2}\delta}{\sqrt{1 - \delta^2}} \|v_{T_{v_0}}\|_1 + \frac{\delta}{2\sqrt{2\sqrt{1 - \delta^2}}} \|u_{T_{u_0}}\|_2.$$  

where the last inequality of (26) follows the fact $\frac{\sqrt{2}\delta}{\sqrt{1 - \delta^2}} \leq \frac{\delta}{\sqrt{1 - \delta^2}}$. On the one hand, by the well-known Holder inequality, we have

$$\frac{1}{\sqrt{k_1}} \|u_{T_{u_0}}\|_1 + \frac{1}{\sqrt{k_2}} \|v_{T_{v_0}}\|_1 \leq \|u_{T_{u_0}}\|_2 + \|v_{T_{v_0}}\|_2.$$  

On the other hand, according to $\sqrt{\frac{k_1}{k_2}} \in [\frac{1}{2}, R]$, one can easily obtain that for any $a_1, a_2, b_1, b_2 \geq 0$, if

$$\frac{1}{\sqrt{k_1}} \|a_1\| + \frac{1}{\sqrt{k_2}} \|a_2\| \leq \frac{1}{\sqrt{k_1}} b_1 + \frac{1}{\sqrt{k_2}} b_2,$$

then

$$a_1 + a_2 \leq R(b_1 + b_2).$$  

Rearranging (26) and inserting (27) and (28) into it, we have

$$\|u_{T_{u_0}}\|_1 + \|v_{T_{v_0}}\|_1 \leq \frac{4R\delta}{2\sqrt{2\sqrt{1 - \delta^2}}} (\|u_{T_{u_0}}\|_1 + \|v_{T_{v_0}}\|_1).$$

Combining the above inequality with the definition of DNSP, we have Lemma 4.

**Appendix B**

**Proof of Theorem 1**

Before the main body of the proof, we prove a useful lemma first.

**Lemma 5** Given that $S \subseteq [N]$, $u, z \in \mathbb{R}^N$, and $F(\cdot)$ is a sparsity-inducing penalty, we have

$$F((u - z)_S) \leq F(z) - F(u) + F((u - z)_S) + 2F(u_S).$$

**Proof:** According to the second term in Property 1, we have

$$F(u) = F(u_S) + F(u_{\bar{S}}) \leq F(u_S) + F((u - z)_S) + F(z_S),$$

$$F((u - z)_S) \leq F(u_S) + F(z_S).$$
By summing up the two inequalities above, the proof is completed.

Now we start the main body of the proof. Let \( u = x - x^* \), \( v = e - e^* \), and \( S_1 \) \((S_2)\) denote the indices set of \( k_1 \) \((k_2)\) largest magnitudes in \( x \) \((e)\). We have \( Au + v = 0 \) and
\[
F(u_{S_1}) + G(v_{S_1}) \leq \gamma \left( F(u_{S_1}) + G(v_{S_1}) \right).
\] (29)

By Lemma 5 we have
\[
F(u_{S_1}) \leq F(x^*) - F(x) + 2F(x_{-k_1}),
\]
\[
G(v_{S_1}) \leq G(e^*) - G(e) + 2G(e_{-k_2}).
\]

Combining the above two inequalities, and using the fact that
\[
\lambda(F(x^*) - F(x)) + (G(e^*) - G(e)) \leq 0,
\]
we have
\[
\lambda F(u_{S_1}) + G(v_{S_1}) \leq \lambda F(u_{S_1}) + 2F(x_{-k_1}) + G(v_{S_1}) + 2G(e_{-k_2}).
\] (30)

Reorganizing (30) results in
\[
\min(1, \lambda) \left( F(u_{S_1}) + G(v_{S_1}) \right) \leq \max(1, \lambda) \left( F(u_{S_1}) + G(v_{S_1}) \right)
\]
\[
+ 2\lambda F(x_{-k_1}) + 2G(e_{-k_2}).
\] (31)

Plugging (29) into the first item of right-hand side (RHS) of (31) yields
\[
F(u_{S_1}) + G(v_{S_1}) \leq \frac{2\lambda F(x_{-k_1}) + 2G(e_{-k_2})}{\beta \max(1, \lambda)},
\] (32)

where \( \beta = \min(\lambda, \frac{1}{\lambda}) - \gamma. \) Plugging (29) into (32) once again we have
\[
F(u) + G(v) \leq (1 + \frac{\gamma}{\beta \max(1, \lambda)}) \left( F(u_{S_1}) + G(v_{S_1}) \right)
\]
\[
\leq \frac{1 + \gamma}{\beta \max(1, \lambda)} \left( 2\lambda F(x_{-k_1}) + 2G(e_{-k_2}) \right)
\]
\[
\leq \frac{2(1 + \gamma)}{\beta} \left( F(x_{-k_1}) + G(e_{-k_2}) \right),
\]
which completes the proof.

APPENDIX C
PROOF OF LEMMA 1

Based on the definition of \( u = \tilde{x} - x \), \( v = \tilde{e} - e \), we define
\[
u = s + s^+,\]
\[
v = t + t^+,
\]
where \( [s; t] \in N(\hat{A}) \) and \( [s^+; t^+] \in N(\hat{A}) \). Suppose \( \tau_x = \text{supp}(x) \) and \( \tau_e = \text{supp}(e) \). For simplicity, we use \( F(\cdot) \) and \( G(\cdot) \) to denote \( F_{\tau_x}(\cdot) \) and \( G_{\tau_e}(\cdot) \) in this part.

Beginning with the left-hand side (LHS) of (16), we have
\[
J(\tilde{x}, \tilde{e}) - J(x, e) = \lambda(F(\tilde{x}) - F(x)) + G(\tilde{e}) - G(e).
\] (33)

Considering the first item in the RHS of (33) and recalling that \( x \) is \( k_1 \)-sparse, we have
\[
F(\tilde{x}) - F(x) = F(x + u) - F(x)
\]
\[
= F(x_{\tau_x} + u_{\tau_x}) + F(u_{\tau_x^c}) - F(x).
\] (34)

According to the second term in Property 1, we have
\[
F(x_{\tau_x} + u_{\tau_x}) \geq F(x_{\tau_x}) - F(u_{\tau_x}),
\]
\[
F(u_{\tau_x^c}) \geq F(s_{\tau_x^c}) - F(s_{\tau_x^c}^+),
\]
\[
-F(u_{\tau_x}) \geq -F(s_{\tau_x}) - F(s_{\tau_x}^+).
\]

Inserting the above three inequalities into the RHS of (34) yields
\[
F(\tilde{x}) - F(x) \geq F(s_{\tau_x^c}) - F(s_{\tau_x}) - F(s^+).
\] (35)

In a similar manner, we have
\[
G(\tilde{e}) - G(e) \geq G(t_{\tau_x^c}) - G(t_{\tau_x}) - G(t^+).
\] (36)

Combining (35) and (36) in (33) results in
\[
J(\tilde{x}, \tilde{e}) - J(x, e)
\]
\[
\geq Q_1(F(s) + G(t)) - (\lambda F(s^+) + G(t^+)),
\] (40)

where
\[
Q_1 = \min(1, \lambda) - \gamma_{1,2} \max(1, \lambda).
\]

For the first item in the RHS of (40), according to the first term in Property 1 and the fact that \( \| \cdot \|_2 \leq \| \cdot \|_1 \), we have
\[
F(s) + G(t) \geq \frac{f(r_1)}{r_1} \|s\|_1 + \frac{g(r_2)}{r_2} \|t\|_1
\]
\[
\geq \frac{f(r_1)}{r_1} \|s\|_2 + \frac{g(r_2)}{r_2} \|t\|_2.
\] (41)

For the second item in the RHS of (40), according to the third term in Property 1, we have
\[
\lambda F(s^+) + G(t^+) \leq \alpha_1 \lambda \|s^+\|_1 + \alpha_2 \|t^+\|_1
\]
\[
\leq \max(1, \lambda) \alpha \sqrt{M + N} \|s^+; t^+\|_2,
\] (42)

where \( \alpha = \max(\alpha_1, \alpha_2) \).

Recalling that
\[
\|s^+; t^+\|_2 \leq \frac{\|Au + v\|_2}{\sigma_{\min}(\hat{A})}
\]
and the fact \( \|u\|_2 \leq \|s\|_2 + \|s^+\|_2, \|v\|_2 \leq \|w\|_2 + \|w^+\|_2 \), and according to (40), (41) and (42), we draw the conclusion of Lemma 1.
Following a similar way, it is easy to get the assumption of Lemma 2, we have

\[\lambda(x^o - x)^T\nabla F_{\eta_1}(x^o) + (e^o - e)^T\nabla G_{\eta_2}(e^o) \geq \frac{C_{f,r_1}}{4}\|x^o - x\|^2 + \frac{C_{g,r_2}}{4}\|e^o - e\|^2. \] 

(43)

On the other hand by Generalized Kuhn-Tucker (GKT) conditions [35], there exists \( \mu \in \mathbb{R}^M \) such that \( 0 \in \partial F_{\eta_1}(x^o) - A^T \mu \) and \( 0 \in \partial G_{\eta_2}(e^o) - \mu \). Let \( \nabla F_{\eta_1}(x^o) = A^T \mu \) and \( \nabla G_{\eta_2}(e^o) = \mu \), we have

\[\lambda(x^o - x)^T\nabla F_{\eta_1}(x^o) + (e^o - e)^T\nabla G_{\eta_2}(e^o)
= (\lambda A(x^o - x) + (e^o - e))^T \mu = 0. \] 

(44)

By (43) and (44), we conclude that \([x^o; e^o] = [x; e]\), which completes the proof.

### APPENDIX E

#### PROOF OF Lemma 2

Because of the definition of the generalized gradient [30] [33], we have

\[\lambda u^T \nabla F_{\eta_1}(\tilde{x}) + v^T \nabla G_{\eta_2}(\tilde{e}) \geq J(\tilde{x}, \tilde{e}) - J(x, e) + (\rho_1 \lambda \|u\|^2 + \rho_2 \|v\|^2), \]

where \( u = \tilde{x} - x \) and \( v = \tilde{e} - e \). By Lemma 1 and the assumption of Lemma 2, we have

\[J(\tilde{x}, \tilde{e}) - J(x, e) \geq \frac{C_{f,r_1} \|u\|^2}{2} + \frac{C_{g,r_2} \|v\|^2}{2}, \]

so

\[\lambda u^T \nabla F_{\eta_1}(\tilde{x}) + v^T \nabla G_{\eta_2}(\tilde{e}) \geq \frac{C_{f,r_1}}{2}\|u\|^2 + \frac{C_{g,r_2}}{2}\|v\|^2 + (\rho_1 \lambda \|u\|^2 + \rho_2 \|v\|^2). \] 

(45)

Considering the definition of \( C_{f,r_1} \), we have an equation

\[\frac{C_{f,r_1}}{4\rho_1 \lambda} = \frac{f(r_1)}{4\lambda \rho_1 r_1} \beta_{\eta_1, m_2} \max(1, \lambda). \] 

(46)

According to the third term in Property 1 and the definition of \( \eta_1 \), we get

\[-\frac{f(r_1)}{\rho_1 r_1} \geq -\frac{\alpha_1 r_1 + \rho_1 r_1^2}{\rho_1 r_1} \geq \frac{1}{\eta_1} - r_1. \] 

(47)

Inserting (47), (17) and (18) into (46), we obtain

\[-\frac{C_{f,r_1}}{4\rho_1 \lambda} \geq \frac{\beta_{\eta_1, m_2} \max(1, \lambda)}{4\lambda (1 + \gamma_1)} \left( \frac{1}{\eta_1} - r_1 \right) \geq r_1 \geq \|u\|^2. \] 

(48)

Following a similar way, it is easy to get

\[-\frac{C_{g,r_2}}{4\rho_2} \geq \|v\|^2. \] 

(49)

Combining (48) and (49) into (45) we obtain Lemma 2.

### APPENDIX F

#### PROOF OF Lemma 3

Define \( x^* = x_{k_1}, e^* = e_{k_2}, e = Ax_{k_1} + e_{k_2}, u^t = \tilde{x}^t - x_{k_1}, v^t = \tilde{e}^t - e_{k_2} \), then \( y = Ax^* + e^* + e \). Before the main body of the proof, a useful lemma is provided, and then a crucial proposition is stated and proved.

#### Lemma 6 (Lemma 2 of [30])

If \( f(z) : \mathbb{R} \to \mathbb{R} \) is a weakly convex sparseness measure, and \( \alpha = \lim_{z \to 0+} \frac{f'(z)}{z} \), then \( f'(z) \) satisfies \( |f'(z)| \leq \alpha \).

#### Proposition 2

If \( \|u^t\|^2 \leq r_1, \|e^t\|^2 \leq r_2 \), (15) holds, weakly convexity parameters \( \eta_1, \eta_2 \) satisfy (17) and (18), and

\[2q \left( M \alpha_2^2 + N \lambda^2 \alpha_1^2 \right) \kappa + 4C_3 \|e\|^2 \|g_r\|^2 \leq C_{f,r_1} \|u^t\|^2 + C_{g,r_2} \|v^t\|^2, \] 

(50)

then

\[2\|u^{t+1}\|^2 + 2\|v^{t+1}\|^2 \leq \|u^t\|^2 + \|v^t\|^2 - (q - 1) \left( M \alpha_2^2 + N \lambda^2 \alpha_1^2 \right) \kappa^2 \] 

(51)

holds, where \( C_{f,r_1}, C_{g,r_2} \) and \( C_3 \) are defined in Lemma 1, \( \kappa \) is a stepsize defined in Algorithm 1, and \( q > 1 \) is an adjustable parameter.

#### Proof:

We first define

\[\nabla J(\tilde{x}^t, \tilde{e}^t) = [\lambda \nabla F_{\eta_1}(\tilde{x}^t); \nabla G_{\eta_2}(\tilde{e}^t)]. \]

According to the \((t+1)\)th iteration of RPGG, we have

\[\|u^{t+1}\|^2 \leq r_1, \|v^{t+1}\|^2 \leq r_2 \]

so

\[\|u^{t+1}\|^2 + \|v^{t+1}\|^2 \leq \|u^t\|^2 + \|v^t\|^2 - \kappa^2 \left( \lambda^2 \alpha^2 \right) \nabla J(\tilde{x}^t, \tilde{e}^t). \] 

(52)

On the one hand, according to Lemma 6 and the definition of \( \nabla J(\tilde{x}^t, \tilde{e}^t) \), we have

\[\|\nabla J(\tilde{x}^t, \tilde{e}^t)\|^2 \leq \sqrt{N \lambda^2 \alpha_1^2 + M \alpha_2^2}. \] 

(53)

where \( \alpha_1 \) and \( \alpha_2 \) are defined in (13) and the descriptions therein. Combining (53), the third term of RHS of (52), and the fact that

\[\|I - \hat{A}^T \hat{A}\| \leq 1, \]

we have

\[\kappa^2 \|I - \hat{A}^T \hat{A}\| \nabla J(\tilde{x}^t, \tilde{e}^t)\|^2 \leq \left( N \lambda^2 \alpha_1^2 + M \alpha_2^2 \right) \kappa^2. \] 

(54)

On the other hand, we consider the fourth term of RHS of (52),

\[-2\kappa \|u^t; v^t\|^T \left( I - \hat{A}^T \hat{A} \right) \nabla J(\tilde{x}^t, \tilde{e}^t) \]

\[-2\kappa \|u^t; v^t\|^T \nabla J(\tilde{x}^t, \tilde{e}^t) + 2\kappa \|u^t; v^t\|^T \hat{A}^T \hat{A} \nabla J(\tilde{x}^t, \tilde{e}^t). \] 

(55)
By Lemma 2 we have
\[ -2\kappa[u^t;v^t]^T \nabla J(\hat{x}^t, \hat{e}^t) \]
\[ \leq -\frac{\kappa}{2} (C_{f,r_1}\|u^t\|^2 + C_{g,r_2}\|v^t\|^2). \] (56)

Inserting (53) and the following three inequalities
\[ \hat{A}[u^t;v^t] = \epsilon, \]
\[ \| (\hat{A}^T \hat{A})^{-1} \hat{A} \|_2 \leq \frac{1}{\delta_{\min}(A)}, \]
\[ \sqrt{N\lambda^2a_1^2 + M^2} \leq \sqrt{M + N} \max(1, \lambda) \max(\alpha_1, \alpha_2) \]
to the second term of RHS of (55) and using the definition of $C_3$, we get
\[ 2\kappa[u^t;v^t]^T \hat{A}^T \hat{A} \nabla J(\hat{x}^t, \hat{e}^t) \leq 2C_3\kappa\|\epsilon\|_2. \] (57)

Combining (52), (54), (56) and (57), we have
\[ \|u^{t+1}\|^2 + \|v^{t+1}\|^2 \leq \|u^t\|^2 + \|v^t\|^2 + (N\lambda^2a_1^2 + M^2)\kappa^2 \]
\[ -\frac{\kappa}{2} (C_{f,r_1}\|u^t\|^2 + C_{g,r_2}\|v^t\|^2) + 2C_3\kappa\|\epsilon\|^2. \] (58)

With the use of (50) and (58), we obtain (51), which completes the proof. ■

The proposition above states that if the current iterate $[\hat{x}^t;\hat{e}^t]$ is not close enough to the optimum, after the next iteration of RPPG, the new iterate $[\bar{x}^{t+1};\bar{e}^{t+1}]$ is closer to the optimum.

Now we begin to prove Theorem 3. Define
\[ D_1 = (M\alpha_2^2 + N\lambda^2\alpha_1^2)\kappa, \]
\[ D_2 = \frac{D_1}{\min(C_{f,r_1}, C_{g,r_2})}, \]
\[ D_3 = \frac{C_3\|\epsilon\|_2}{\min(C_{f,r_1}, C_{g,r_2})}, \]
\[ a^t = \sqrt{\|u^t\|^2 + \|v^t\|^2}, \]
and suppose that
\[ 4\sqrt{2}D_2 + 8\sqrt{2}D_3 \leq \|u^t\|^2 + \|v^t\|^2 \leq \sqrt{2}a^t \] (59)
holds. By Proposition 2, we set
\[ q = \frac{a^t - 4D_3}{2D_2} > 1, \] (60)
and we have
\[ (a^{t+1})^2 \leq (a^t)^2 - (q - 1)D_1\kappa. \] (61)

Substituting (60) into (61) yields
\[ (a^{t+1})^2 \leq (a^t)^2 - \frac{D_1\kappa}{2D_2} (a^t - 2D_2 - 4D_3). \] (62)

Inserting (59) into (62), we obtain
\[ (a^{t+1})^2 \leq (a^t)^2 - \frac{D_1\kappa}{4D_2} a^t \leq \left( a^t - \frac{D_1\kappa}{8D_2} \right)^2, \]
which means
\[ a^{t+1} \leq a^t - \frac{D_1\kappa}{8D_2}. \] (63)

According to Proposition 2 and (59)-(63), if
\[ 4\sqrt{2}D_2 + 8\sqrt{2}D_3 \leq \|u^t\|^2 + \|v^t\|^2 \leq \sqrt{2}a^t \] (64)
then the following holds:
\[ a^{t+1} \leq a^t - \frac{D_1\kappa}{8D_2}. \] (65)

Thus, there exists an integer
\[ n \leq \left[ \frac{D_2 8\sqrt{r_1^2 + r_2^2}}{D_1 \kappa} \right], \]
such that after $n$ iterations,
\[ a^n \leq \|u^n\|^2 + \|v^n\|^2 < 4\sqrt{2}D_2 + 8\sqrt{2}D_3. \] (66)

On the other hand, because (64) is not satisfied when replacing $t$ with $n$, (65) is not guaranteed. On the other hand, by the relation of $a^{n+1}$ and $a^n$, (53) and (66), we have
\[ a^{n+1} \leq a^n + \kappa\| (1 - \hat{A}^T\hat{A}) \nabla J(x^n, e^n) \|_2 \]
\[ \leq 4\sqrt{2}D_2 + 8\sqrt{2}D_3 + \kappa \sqrt{N\lambda^2\alpha_1^2 + M\alpha_2^2}. \] (67)

Then
\[ \bullet \] if (66) holds when replacing $n$ with $n + 1$, following a similar way of (67), we have $a^{n+2} \leq 4\sqrt{2}D_2 + 8\sqrt{2}D_3 + \kappa \sqrt{N\lambda^2\alpha_1^2 + M\alpha_2^2}$;
\[ \bullet \] else, (64) holds with $t = n + 1$, then $a^{n+2} \leq a^{n+1} - \frac{D_1\kappa}{8D_2} \leq 4\sqrt{2}D_2 + 8\sqrt{2}D_3 + \kappa \sqrt{N\lambda^2\alpha_1^2 + M\alpha_2^2}$. Through the discussion above, we conclude that for any $t \geq n$,
\[ a^{t+1} \leq 4\sqrt{2}D_2 + 8\sqrt{2}D_3 + \kappa \sqrt{N\lambda^2\alpha_1^2 + M\alpha_2^2}. \] (68)

Plugging
\[ \sqrt{\|u^t - x_{-k_1}\|^2 + \|v^t - e_{-k_2}\|^2} \]
\[ = \sqrt{\|u^t - x_{-k_1}\|^2 + \|v^t - e_{-k_2}\|^2} \]
\[ \leq \sqrt{\|u^t\|^2 + \|v^t\|^2 + \|x_{-k_1}\|^2 + \|e_{-k_2}\|^2} \]
\[ \leq a^t + \|x_{-k_1}\|_2 + \|e_{-k_2}\|_2 \]
into (68), we complete the proof.

REFERENCES


1Here we assume that $a^t \geq \frac{D_1\kappa}{8D_2}$, or the estimated signal is close enough to the original one.


