Robust Sparse Recovery in Impulsive Noise via ℓ_p-ℓ_1 Optimization

Fei Wen, *Member, IEEE*, Peilin Liu, *Member, IEEE*, Yipeng Liu, Robert C. Qiu, *Fellow, IEEE*, and Wenxian Yu

***Abstract*—This paper addresses the issue of robust sparse recovery in compressive sensing (CS) in the presence of impulsive measurement noise. Recently, robust data-fitting models, such as ℓ_1-norm, Lorentzian-norm, and Huber penalty function, have been employed to replace the popular ℓ_2-norm loss model to gain more robust performance. In this paper, we propose a robust formulation for sparse recovery using the generalized ℓ_p-norm with $0 \le p < 2$ as the metric for the residual error. To solve this formulation efficiently, we develop an alternating direction method (ADM) via incorporating the proximity operator of ℓ_p-norm functions into the framework of augmented Lagrangian methods. Furthermore, to derive a convergent method for the nonconvex case of $p < 1$, a smoothing strategy has been employed. The convergence conditions of the proposed algorithm have been analyzed for both the convex and nonconvex cases. The new algorithm has been compared with some state-of-the-art robust algorithms via numerical simulations to show its improved performance in highly impulsive noise.**

***Index Terms*—Alternating direction method (ADM), augmented Lagrangian methods, compressive sensing (CS), ℓ_p-norm data-fitting, robust sparse recovery.**

I. Introduction

COMPRESSIVE SENSING (CS) is a paradigm to acquire sparse, or compressible, signals at a rate significantly lower than that of the classical Nyquist sampling, which has attracted much attention in recent years [1], [2]. Basically, the CS theory states that if a signal $\mathbf{x} \in \mathbb{R}^n$ is sparse, only a small number of linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$ $(m < n)$ of the signal suffice to accurately reconstruct it, $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the sensing matrix (also called the sampling or measurement matrix). In most practical applications, the measurements are inevitably contaminated by some noise. In this situation, the compressed measurements can be typically modeled as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$$

where $\mathbf{n} \in \mathbb{R}^m$ is additive measurement noise.

In the CS setting, the recovery of $\mathbf{x}$ from the compressed measurement $\mathbf{y}$ is generally ill-posed because of $m < n$. However, provided that $\mathbf{x}$ is sparse and the sensing matrix $\mathbf{A}$ satisfies some stable embedding conditions [3], $\mathbf{x}$ can be reliably recovered with an error upper bounded by the noise strength. To reconstruct $\mathbf{x}$ such that it is of the sparsest structure leads to the following optimization problem

$$\underset{\mathbf{x}}{\text{minimize}} \ \|\mathbf{x}\|_0 \ \ \text{subject to} \ \ \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \le \epsilon \qquad (1)$$

where $\|\mathbf{x}\|_0$, formally called ℓ_0-norm, counts the number of nonzero components in the vector $\mathbf{x}$, $\epsilon > 0$ bounds the ℓ_2-norm of the residual error and is pre-determined by the noise level. In general, solving the nonconvex problems (1) is known to be NP-hard. Thus, convex relaxation methods are often considered, such as basis-pursuit (BP) [4], [5] or LASSO [6], which relax the ℓ_0-norm minimization into the ℓ_1-norm minimization, e.g., BP denoising (BPDN),

$$\underset{\mathbf{x}}{\text{minimize}} \ \|\mathbf{x}\|_1 \ \ \text{subject to} \ \ \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \le \epsilon. \qquad (2)$$

This constrained optimization problem can be converted into an alternative unconstrained form (called LASSO)

$$\underset{\mathbf{x}}{\text{minimize}} \left\{ \frac{1}{\mu} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \|\mathbf{x}\|_1 \right\} \qquad (3)$$

where $\mu > 0$ is a regularization parameter that controls a tradeoff between the residual error term and the regularization term. Under certain conditions, the solution of the ℓ_1-norm minimization problem coincides with that of ℓ_0-norm minimization problem [7], [8]. The ℓ_1-norm minimization problems are more tractable due to their convexity and hence most widely used in sparse reconstruction.

This work mainly focuses on the issue of robust denoising models in CS [9]. As in (1)–(3) and many other variants, the ℓ_2-norm data-fitting model, which is optimal for Gaussian noise in the maximum likelihood sense, is the most widely used one. However, in practical applications, the measurement noise may be of different kinds or combinations. Impulsive noise is a typical case which can model large errors in observations and has been widely studied in robust statistics [10]. Impulsive corruption in measurements may come from missing data in the measurement process, transmission problems [11]–[13], faulty memory locations [14], buffer overflow [15], and has been raised in many image and video processing works [16]–[19]. In these cases, the ℓ_2-norm data-fitting model is inefficient as

Manuscript received June 4, 2016; accepted July 29, 2016. Date of publication August 4, 2016; date of current version October 28, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wenwu Wang. This work was supported in part by the National Natural Science Foundation of China under Grants 61401501, 61304225, 61573242, and 61472442.

F. Wen is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China and also with the Air Control and Navigation Institution, Air Force Engineering University, Xian 710000, China (e-mail: wenfei@sjtu.edu.cn).

P. Liu, R. C. Qiu, and W. Yu are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: liupeilin@sjtu.edu.cn; rcqiu@sjtu.edu.cn; wxyu@sjtu.edu.cn).

Y. Liu is with the School of Electronic Engineering/Center for Robotics/Center for Information in Medicine, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: yipengliu@uestc.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2598316

it is well-known that least-squares based estimators are highly sensitive to outliers in the observations.

Recently, various robust formulations have been proposed for CS to suppress the outliers in measurements. In [20]–[22], the Lorentzian-norm has been employed as the metric for the residual error, and a geometric optimization problem has been introduced for sparse signal recovery. In [23], the ℓ_1-norm has been used as the data-fitting model to obtain a robust formulation as

$$\underset{\mathbf{x}}{\text{minimize}} \left\{ \frac{1}{\mu} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_1 + \|\mathbf{x}\|_1 \right\}. \tag{4}$$

It has been shown in [23] that, when the measurements contain large errors or impulsive noise, the ℓ_1-norm loss function can result in dramatically better performance compared with the ℓ_2-one. Subsequently, more efficient alternating direction methods (ADM) for the problem (4) have been proposed in [24]. Meanwhile, an extension to ℓ_1-norm constrained ℓ_1-minimization problem has been given in [25]. In [26], the Huber penalty function has been used to design robust formulation for sparse recovery. Subsequently, efficient first-order algorithms have been proposed in [27] to solve the Huber-loss based formulation and its several variants. It has been shown in [27] that, the ℓ_1-norm loss based formulation (4) offers considerable gain over the Huber and Lorentzian-norm loss based ones. Notably, the ℓ_1-norm loss function has also been employed in sparse representation based face recognition [28], channel estimation [29], and signal separation [30], [31] to achieve robustness. Moreover, the ℓ_0-norm loss has been used in [60]–[63] for robust restoration of images corrupted by salt-and-pepper impulsive noise.

There also exists Bayesian robust algorithm [32], [59], which extend the Bayesian sparse recovery method [56] and model the impulsive noise as Student-t and Gaussian mixture distributions, respectively. Moreover, robust recovery in the presence of saturation error in practical quantization has been addressed in [33].

In this paper, we use the generalized ℓ_p-norm, $0 \le p < 2$, as the loss function for the residual error to propose the following robust formulation

$$\underset{\mathbf{x}}{\text{minimize}} \left\{ \frac{1}{\mu} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_p^p + \|\mathbf{x}\|_1 \right\}. \tag{5}$$

When $0 < p < 1$, $\|\cdot\|_p^p$ is the ℓ_p quasi-norm defined in a similar manner as the case of $p \ge 1$, i.e., $\|\mathbf{v}\|_p^p = \sum_{i=1}^m |v_i|^p$. Note that, the formulations (3) and (4) can be viewed as two special cases of (5) with $p = 2$ and $p = 1$, respectively.

The intuition behind utilizing ℓ_p-norm loss function is that, compared with the quadratic function, it is a less rapidly increasing function when $p < 2$, and, accordingly, is less sensitive to large outliers, especially when p is small. Notably, the ℓ_p-norm cost function has been widely used in various signal processing applications for developing robust algorithms in impulsive noise, such as array beamforming [34], [35], direction-of-arrival estimation [36], time delay estimation [37], and spectrum sensing [38]. In these works it is commonly restricted to the case of $p \ge 1$ because $p < 1$ leads to intractable nonconvex problems. In this work, the nonconvex case is also considered for robust CS.

Except for the special case of $p = 1$, the problem (5) has still not been well addressed. When $1 < p < 2$, it can be solved by traditional convex optimization methods such as interior-point methods. Specifically, this problem can be converted into [39]

$$\underset{\mathbf{x}}{\text{minimize}} \, \|\mathbf{x}\|_1 \;\; \text{subject to} \;\; \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_p \le \epsilon \tag{6}$$

where $\epsilon > 0$ bounds the ℓ_p-norm of the residual error. A semi-definite program (SDP) conversion method has been proposed in [40] to handle the problem (6). It decomposes the ℓ_p-norm inequality constraint into a number of linear matrix inequalities and recast the problem into a SDP. However, this approach is generally inefficient and impractical for large-scale problems. Moreover, when $1 < p < 2$, the ℓ_p-norm is smooth and convex but its gradient is not Lipschitz continuous (not bounded), thus, traditional proximal gradient methods cannot be directly applied.

When $0 \le p < 1$, the problem (5) is more difficult to solve since in addition to the nonconvexity of the loss term, both the loss and regularization terms are nonsmooth. This case has still not been reported in the open literatures. The main contributions of this work are as follows.

A. Contributions

First, we provide some analysis on the proposed formulation. We show that, even when the noise is highly impulsive with infinite variance, this formulation using an appropriate choice of p has the potential to stably recover the desired signal with a finite recovery error, e.g., using $p < \alpha$ in α-stable noise.

Second, we propose an efficient ADM, termed Lp-ADM, for the optimization problem (5). The new algorithm is derived via incorporating the proximity operator of ℓ_p-norm functions into the framework of augmented Lagrangian methods, which facilitates solving (5) efficiently in a unified framework for both the convex and nonconvex cases. Furthermore, for the nonconvex case of $p < 1$, a smoothing strategy has been employed to derive a convergent algorithm.

Third, the convergence condition of the new algorithm has been analyzed for both the convex and nonconvex cases. Finally, we have compared the new algorithm with some recently proposed robust algorithms via simulations. The results demonstrated that, with an appropriate choice of p, e.g., $p < 1$, it has the capability to achieve the state-of-the-art robust performance in highly impulsive noise.

B. Outline and Notations

The rest of this paper is organized as follows. Section II provides some analysis on the proposed formulation. In Section III, we introduce the proximity operator for ℓ_p-norm functions, which is employed in the proposed algorithm. In Section IV, the new algorithm is presented. Section V contains convergence analysis and Section VI provides experimental results. Finally, Section VII ends the paper with concluding remarks.

Notations: $\lambda_{\max}(\cdot)$ denotes the maximal eigenvalue of a matrix. $E(\cdot)$, $\langle \cdot, \cdot \rangle$ and $(\cdot)^T$ stand for the expectation, inner product and transpose, respectively. $\nabla f(\cdot)$ and $\partial f(\cdot)$ stand for the gradient and subdifferential of the function f. $\text{sign}(\cdot)$ denotes the sign of a quantity with $\text{sign}(0) = 0$. $\mathbf{I}_n$ stands for an $n \times n$ identity matrix. $\|\cdot\|_p$ denotes the ℓ_p-norm. $\text{dist}(\mathbf{x}, S) := \inf\{\|\mathbf{y} - \mathbf{x}\| :$

$\mathbf{y} \in S\}$ denotes the distance from a point $\mathbf{x} \in \mathbb{R}^n$ to a subset $S \subset \mathbb{R}^n$. $\mathbf{X} \succeq \mathbf{0}$ means that $\mathbf{X}$ is positive semidefinite.

II. On the Proposed Robust Formulation

A. Bayesian Sparse Recovery in Impulsive GGD Noise

Impulsive noise can be well modeled as symmetric α-stable ($S\alpha S$) process [9], which can be conveniently described by the characteristic function

$$\varphi(\omega) = \exp\left(ja\omega - \gamma^\alpha |\omega|^\alpha\right)$$

where $0 < \alpha \leq 2$ is the characteristic exponent which measures the thickness of the tail of the distribution, a is the location parameter, and $\gamma > 0$ is the scale parameter. The smaller the value of α, the thicker the tail of the $S\alpha S$ distributions and hence the more impulsive the noise is.

Except for a few special cases, there are no closed-form expressions for the probability density function (PDF) of the $S\alpha S$ distributions. Thus, it is difficult to obtain the MAP estimate of $\mathbf{x}$ when the noise is modeled as an $S\alpha S$ distribution. The generalized Gaussian distribution (GGD) is an alternative that can also be used to model impulsive noise. The PDF of a zero-mean GGD variable x is given by

$$f(x) = \frac{v}{2\sigma\Gamma(\frac{1}{v})} \exp\left(-\frac{|x|^v}{\sigma^v}\right) \tag{7}$$

where $\Gamma(\cdot)$ is the gamma function, $v > 0$ denotes a shape parameter which controls the distribution shape, $\sigma > 0$ is the scale parameter. The flexible parametric form of the GGD (7) adapts to a large family of symmetric distributions, from super-Gaussian ($v < 2$) to sub-Gaussian ($v > 2$), including specific distributions such as Laplacian ($v = 1$) and Gaussian ($v = 2$). When $v < 2$, the GGD shows a heavy tail and hence is suitable for modeling impulsive noise.

Assume that the (impulsive) noise samples are independently and identically distributed (i.i.d.) GGD with zero-mean, the PDF of $\mathbf{n}$ is

$$f(\mathbf{n}) = \frac{v^N}{\left[2\sigma_n\Gamma(\frac{1}{v})\right]^N} \exp\left(-\frac{\|\mathbf{n}\|_v^v}{\sigma_n^v}\right).$$

Then, the conditional PDF $f(\mathbf{y}|\mathbf{x})$ also follows an i.i.d. GGD as

$$f(\mathbf{y}|\mathbf{x}) = \frac{v^N}{\left[2\sigma_n\Gamma(\frac{1}{v})\right]^N} \exp\left(-\frac{\|\mathbf{Ax}-\mathbf{y}\|_v^v}{\sigma_n^v}\right).$$

For sparse signals, the i.i.d. zero-mean Laplacian distribution *prior* is of particular interest to the Bayesian community for dimensionality reduction problems, e.g.,

$$f(\mathbf{x}) = \frac{1}{\left(\sqrt{2}\sigma_x\right)^N} \exp\left(-\frac{\sqrt{2}\|\mathbf{x}\|_1}{\sigma_x}\right).$$

From the Bayes formula, the *a posteriori* PDF of $\mathbf{x}$ can be expressed as

$$\begin{aligned} f(\mathbf{x}|\mathbf{y}) &\propto f(\mathbf{y}|\mathbf{x})f(\mathbf{x}) \\ &= C\exp\left(-\frac{1}{\sigma_n^v}\|\mathbf{Ax}-\mathbf{y}\|_v^v - \frac{\sqrt{2}}{\sigma_x}\|\mathbf{x}\|_1\right) \end{aligned}$$

where C is a constant. Consequently, the MAP estimate of $\mathbf{x}$ given $\mathbf{y}$ can be obtained via minimizing (5) with $\mu = \sqrt{2}\sigma_n^p/\sigma_x$.

For highly impulsive noise with very heavily tailed distributions, it may be required that $p < 1$. In this case, the loss function in problem (5) is nonconvex, which leads to a NP-hard optimization problem. Moreover, the MAP formula sheds some light on the optimal choice of the regularization parameter μ, which is related with the statistical information of the noise and the true signal.

B. Analysis on ℓ_p-Norm Data-Fitting Model

In this section, we show that with an appropriately choice of p, the proposed formulation has the capability to successfully recover the desired signal with a finite ℓ_2-norm error when the noise is highly impulsive with infinite variance.

A well-known condition of the sensing matrix $\mathbf{A}$ ensuring the satisfactory recovery of $\mathbf{x}$ is called restricted isometry property (RIP) [41]. For each integer $s = 1, 2, \ldots$, define the s-restricted isometry constant δ_s of $\mathbf{A}$, which is the smallest positive number such that

$$(1-\delta_s)\|\mathbf{z}\|_2^2 \leq \|\mathbf{Az}\|_2^2 \leq (1+\delta_s)\|\mathbf{z}\|_2^2$$

holds for all s-sparse vectors. It has been shown in [41] that if $\|\mathbf{n}\|_2 \leq \epsilon$ and $\delta_{2s} < \sqrt{2}-1$, the solution to the BPDN problem (2), denoted by $\hat{\mathbf{x}}$, obeys

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq C\epsilon \tag{8}$$

where C is a constant depends on δ_{2s}. This result indicates that, when the noise variance is finite, BPDN can stably recover the desired signal with an error bounded by the noise strength. However, for highly impulsive noise with infinite variance, BPDN (also LASSO) is no longer stable in statistics. In this case, the formulations (5) and (6) are superior.

Theorem 1: Suppose that the sensing matrix $\mathbf{A}$ satisfies the RIP of order $2s$ with $\delta_{2s} < \sqrt{2}-1$. Then for any signal $\mathbf{x}$ supported on T_0 with $|T_0| \leq s$, and any measurement noise $\mathbf{n}$ with $\|\mathbf{n}\|_p \leq \epsilon$, $0 \leq p < 2$, the solution to (6), denoted by $\hat{\mathbf{x}}$, obeys

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq C_s\epsilon \tag{9}$$

where C_s is a constant depends on δ_{2s}.

Proof: See Appendix A. ■

Remark: In Theorem 1, the condition of the noise is relaxed to $\|\mathbf{n}\|_p \leq \epsilon$, $0 \leq p < 2$, while that of the BPDN problem is $\|\mathbf{n}\|_2 \leq \epsilon$. This implicitly relaxes the condition of the noise for stable recovery (in statistics) from that, its variance is finite, to that, its pth-order *fractional-lower-order moment* is finite. That means the proposed formulation has the capability to stably recover $\mathbf{x}$ when the noise is highly impulsive with infinite variance, e.g., α-stable processes, while BPDN (also LASSO) is unstable in this case. More specifically, for an $S\alpha S$ random variable x with $0 < \alpha < 2$, zero location parameter, and scale parameter γ, its pth-order moment is finite when $p < \alpha$ but infinite when $p \geq \alpha$ [42]

$$\begin{cases} E\{|x|^p\} = C(p,\alpha)\gamma^p, & 0 < p < \alpha \\ E\{|x|^p\} = +\infty, & p \geq \alpha \end{cases} \tag{10}$$

with $C(p,\alpha) = 2^{p+1}\Gamma(\frac{p+1}{2})\Gamma(-\frac{p}{\alpha})[\alpha\sqrt{\pi}\Gamma(-\frac{p}{2})]^{-1}$. Accordingly, assume that the noise samples are i.i.d. $S\alpha S$ variables, we have $E\{\|\mathbf{n}\|_p^p\} < +\infty$ if $0 < p < \alpha$ and $E\{\|\mathbf{n}\|_p^p\} = +\infty$ if $p \geq \alpha$. For such heavily tailed impulsive noise, stable reconstruction by the proposed robust formulation is guaranteed if the value of p is chosen to be less than α.

III. Proximity Operator for ℓ_p-Norm Functions

Recall the proximity operator of a function $g(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^m$ with penalty η [43]

$$\operatorname{prox}_{g,\eta}(\mathbf{t}) = \arg\min_{\mathbf{x}} \left\{ g(\mathbf{x}) + \frac{\eta}{2}\|\mathbf{x} - \mathbf{t}\|_2^2 \right\}. \tag{11}$$

For the case of $g(\mathbf{x}) = a\|\mathbf{x}\|_p^p$ with $0 \leq p < 2$ and $a > 0$, $\operatorname{prox}_{g,\eta}$ reduces to solving m univariate minimization problems and thus is easy to compute.

Case 1: $p = 0$. In this case, the proximity operator becomes the well-known hard-thresholding operator

$$\operatorname{prox}_{g,\eta}(\mathbf{t})_i = \begin{cases} 0, & |t_i| \leq \sqrt{2a/\eta} \\ t_i, & \text{otherwise} \end{cases}, \quad i = 1, \ldots, m \tag{12}$$

where t_i is the i-th element of the vector $\mathbf{t}$.

Case 2: $0 < p < 1$. The proximity operator in this case can be computed as [44], [45]

$$\operatorname{prox}_{g,\eta}(\mathbf{t})_i = \begin{cases} 0, & |t_i| < \tau \\ \{0, \operatorname{sign}(t_i)\beta\}, & |t_i| = \tau \\ \operatorname{sign}(t_i) z_i, & |t_i| > \tau \end{cases}, \quad i = 1, \ldots, m \tag{13}$$

where $\beta = [2a(1-p)/\eta]^{\frac{1}{2-p}}$, $\tau = \beta + ap\beta^{p-1}/\eta$, z_i is the solution of $h_1(z) = apz^{p-1} + \eta z - \eta|t_i| = 0$ over the region $(\beta, |t_i|)$. Since $h_1(z)$ is convex, when $|t_i| > \tau$, z_i can be efficiently solved using a Newton's method. For the special cases of $p = \frac{1}{2}$ or $p = \frac{2}{3}$, the proximal mapping can be explicitly expressed as the solution of a cubic or quartic equation [46].

Case 3: $p = 1$. In this case, the proximity operator has a closed-form expression as [47]

$$\operatorname{prox}_{g,\eta}(\mathbf{t})_i = S_{a/\eta}(\mathbf{t})_i = \operatorname{sign}(t_i)\max\{|t_i| - a/\eta, 0\}$$

for $i = 1, \ldots, m$, where $S_a : \mathbb{R}^m \to \mathbb{R}^m$ is the well-known soft-thresholding or shrinkage operator.

Case 4: $1 < p < 2$. In this case, $g(\mathbf{x})$ is convex and smooth, and the proximity operator satisfies [43]

$$\operatorname{prox}_{g,\eta}(\mathbf{t})_i = \operatorname{sign}(t_i) z_i \tag{14}$$

where z_i is the solution of the equality

$$h_2(z) = paz^{p-1} + \eta z - \eta|t_i| = 0, \quad z \geq 0. \tag{15}$$

Note that, $h_2(z)$ is an increasing and concave function for $z \geq 0$, with $h_2(0) < 0$ and $h_2(|t_i|) > 0$ when $t_i \neq 0$. Thus, when $t_i \neq 0$, the solution of (15) satisfies $0 < z_i < |t_i|$ and can be computed by a Newton's method. The starting point can be chosen to be a positive lower bound of the solution as (see Appendix B)

$$z_i^0 = \begin{cases} \phi^{\frac{1}{p-1}}, & \phi < 1 \\ \phi, & \phi \geq 1 \end{cases} \tag{16}$$

with $\phi = \eta|t_i|/(pa + \eta)$. In practical implementation, $\phi^{\frac{1}{p-1}}$ may be very small when $\phi < 1$ and $p \to 1^+$. To address this problem, we preset a small constant $\delta > 0$ (e.g., $\delta = 10^{-10}$), which is used as the starting point $z_i^0 = \delta$ if $h_2(\delta) \leq 0$ and, otherwise, we directly set $z_i = 0$ since the true solution is very small and less than δ when $h_2(\delta) > 0$.

IV. Proposed Algorithm

ADM is a powerful optimization framework that is suitable for large-scale problems arising in machine learning and signal processing, which has been developed long ago and reviewed recently in [48]. In the following we propose a computationally efficient algorithm for the ℓ_p-ℓ_1 minimization problem (5) based on ADM, with the use of the proximity operator introduced in Section III. In the nonconvex case of $p < 1$, since both the loss and regularization terms in (5) are nonsmooth and the loss term is nonconvex, the directly extended ADM algorithm is not guaranteed to converge. To derive a convergent algorithm for the nonconvex case, we use a smoothing strategy and develop a proximal ADM algorithm, which is guaranteed to converge if the penalty parameter is chosen sufficiently large.

A. *Lp-ADM Algorithm Without Smoothing*

In the ADM framework, the ℓ_p-norm loss term and the nonsmooth ℓ_1-regularization term are naturally separated. It decouples the variables and makes the problem easy to tackle. More specifically, using an auxiliary variable $\mathbf{v} \in \mathbb{R}^m$, the problem (5) can be equivalently reformulated as

$$\min_{\mathbf{x},\mathbf{v}} \left\{ \frac{1}{\mu}\|\mathbf{v}\|_p^p + \|\mathbf{x}\|_1 \right\} \quad \text{subject to} \quad \mathbf{A}\mathbf{x} - \mathbf{y} = \mathbf{v}. \tag{17}$$

The corresponding augmented Lagrangian function is given by

$$\mathcal{L}_\rho(\mathbf{v},\mathbf{x},\mathbf{w}) = \frac{1}{\mu}\|\mathbf{v}\|_p^p + \|\mathbf{x}\|_1 - \langle \mathbf{w}, \mathbf{A}\mathbf{x} - \mathbf{y} - \mathbf{v} \rangle + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y} - \mathbf{v}\|_2^2$$

where $\mathbf{w} \in \mathbb{R}^m$ is the dual variable, $\rho > 0$ is a penalty parameter associated with the augmentation. Then, ADM applied to (17) consists of the following iterations

$$\mathbf{v}^{k+1} = \arg\min_{\mathbf{v}} \left(\frac{1}{\mu}\|\mathbf{v}\|_p^p + \frac{\rho}{2}\|\mathbf{A}\mathbf{x}^k - \mathbf{y} - \mathbf{v} - \frac{\mathbf{w}^k}{\rho}\|_2^2 \right) \tag{18}$$

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} \left(\|\mathbf{x}\|_1 + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y} - \mathbf{v}^{k+1} - \frac{\mathbf{w}^k}{\rho}\|_2^2 \right) \tag{19}$$

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{y} - \mathbf{v}^{k+1}). \tag{20}$$

The $\mathbf{x}$-subproblem (19) itself is an ℓ_2-ℓ_1 minimization problem as (3). Since ADM would converge even when the inner steps are not carried out exactly [49], we can approximately solve this subproblem by linearizing the quadratic term of its objective function. More precisely, at a given point $\mathbf{x}^k$ we have

$$\frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{u}^k\|_2^2 \approx \frac{1}{2}\|\mathbf{A}\mathbf{x}^k-\mathbf{u}^k\|_2^2 + \langle \mathbf{x}-\mathbf{x}^k, d(\mathbf{x}^k)\rangle + \frac{L_1}{2}\|\mathbf{x}-\mathbf{x}^k\|_2^2$$

where $d(\mathbf{x}^k)=\mathbf{A}^T(\mathbf{A}\mathbf{x}^k-\mathbf{u}^k)$, $\mathbf{u}^k=\mathbf{y}+\mathbf{v}^{k+1}+\mathbf{w}^k/\rho$, $L_1>0$ is a proximal parameter. With this linearization, the $\mathbf{x}$-subproblem degenerates to the soft-thresholding operator

$$\mathbf{x}^{k+1}=S_{1/(\rho L_1)}\left(\mathbf{x}^k-\frac{1}{L_1}\mathbf{A}^T(\mathbf{A}\mathbf{x}^k-\mathbf{u}^k)\right). \tag{21}$$

The $\mathbf{v}$-subproblem (18) is a form of the proximity operator (11), which can be efficiently solved as

$$\mathbf{v}^{k+1}=\operatorname{prox}_{\frac{1}{\mu}\|\mathbf{v}\|_p^p,\rho}(\mathbf{b}^k)=\begin{cases}\text{solved as (12)}, & p=0\\ \text{solved as (13)}, & 0<p<1\\ S_{1/(\mu\rho)}(\mathbf{b}^k), & p=1\\ \text{solved as (14)}, & 1<p<2\\ \rho\mathbf{b}^k/(\rho+2/\mu), & p=2\end{cases}$$

where $\mathbf{b}^k=\mathbf{A}\mathbf{x}^k-\mathbf{y}-\mathbf{w}^k/\rho$.

B. Lp-ADM Algorithm Using Smoothed ℓ_1-Regularization For the Nonconvex Case

As will be shown in Section V, in the convex case of $p\geq 1$, the above ADM algorithm with an appropriate choice of L_1 is guaranteed to converge. However, in the nonconvex case of $p<1$, this algorithm is not guaranteed to converge. To address this problem, we propose to solve a smoothed version of the problem (5) when $p<1$. Specifically, the ℓ_1-norm regularization in (5) is smoothed as

$$\|\mathbf{x}\|_{1,\varepsilon}=\sum_i\left(x_i^2+\varepsilon^2\right)^{\frac{1}{2}}.$$

$\varepsilon>0$ is an approximation parameter and we have

$$\lim_{\varepsilon\to 0}\|\mathbf{x}\|_{1,\varepsilon}=\|\mathbf{x}\|_1$$

which means that with a small ε, $\|\mathbf{x}\|_{1,\varepsilon}$ accurately approximates the ℓ_1-norm of $\mathbf{x}$. More importantly, with $\varepsilon>0$, the gradient of $\|\mathbf{x}\|_{1,\varepsilon}$ is Lipschitz continuous. In this case, the derived algorithm is guaranteed to converge if the penalty parameter is chosen sufficiently large such that $\rho>\frac{C}{\varepsilon}$, where C is a constant depends on $\mathbf{A}$ and a proximal parameter in the $\mathbf{x}$-subproblem (see Section V).

Using $\|\mathbf{x}\|_{1,\varepsilon}$ as the regularization, the problem becomes

$$\min_{\mathbf{x},\mathbf{v}}\left\{\frac{1}{\mu}\|\mathbf{v}\|_p^p+\|\mathbf{x}\|_{1,\varepsilon}\right\}\ \text{subject to}\ \mathbf{A}\mathbf{x}-\mathbf{y}=\mathbf{v}. \tag{22}$$

The corresponding augmented Lagrangian function is

$$\mathcal{L}_{\rho,\varepsilon}(\mathbf{v},\mathbf{x},\mathbf{w})=\frac{1}{\mu}\|\mathbf{v}\|_p^p+\|\mathbf{x}\|_{1,\varepsilon}-\langle\mathbf{w},\mathbf{A}\mathbf{x}-\mathbf{y}-\mathbf{v}\rangle + \frac{\rho}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}-\mathbf{v}\|_2^2.$$

The $\mathbf{x}$-subproblem becomes

$$\mathbf{x}^{k+1}=\arg\min_{\mathbf{x}}\left(\|\mathbf{x}\|_{1,\varepsilon}+\frac{\rho}{2}\|\mathbf{A}\mathbf{x}-\mathbf{u}^k\|_2^2\right). \tag{23}$$

In this case, the objective function in (23) is smooth, thus general iterative methods can be used to solve this subproblem. However, to gain overall efficiency of the algorithm, we use the standard trick for ADM again to solve (23) approximately. Specifically, we linearize the term $\|\mathbf{x}\|_{1,\varepsilon}$ at a given point $\mathbf{x}^k$ as

$$\|\mathbf{x}\|_{1,\varepsilon}\approx\|\mathbf{x}^k\|_{1,\varepsilon}+\langle\mathbf{x}-\mathbf{x}^k,d_2(\mathbf{x}^k)\rangle+\frac{L_2}{2}\|\mathbf{x}-\mathbf{x}^k\|_2^2$$

which results in the following closed-form solution

$$\mathbf{x}^{k+1}=(L_2\mathbf{I}_n+\rho\mathbf{A}^T\mathbf{A})^{-1}[L_2\mathbf{x}^k-d_2(\mathbf{x}^k)+\rho\mathbf{A}^T\mathbf{u}^k] \tag{24}$$

where $d_2(\mathbf{x}^k)=\nabla\|\mathbf{x}^k\|_{1,\varepsilon}$ with $d_2(\mathbf{x}^k)_i=x_i(x_i^2+\varepsilon^2)^{-\frac{1}{2}}$, $L_2>0$ is a proximal parameter. Note that, we do not linearize the quadratic term in the objective as the previous case since it does not yield a closed-form solution when $\varepsilon>0$.

In computing (24), Cholesky decomposition can be used to reduce the computational complexity (see [48] for detail). Moreover, when the problem size is large, it may be worth using an iterative method to solve (23) rather than the approximate direct method (24). Specifically, for large-scale problems, the direct method may not work due to the requirement of too much memory. In this case, the minimization (23) can be more efficiently carried out by any first-order iterative method, such as the gradient method, conjugate gradient method, and the quasi-Newton methods.

Furthermore, when the sensing matrix $\mathbf{A}$ is orthonormal, i.e., $\mathbf{A}\mathbf{A}^T=\mathbf{I}_m$, the inversion in (24) can be avoided. Specifically, using the matrix inversion lemma we have

$$(L_2\mathbf{I}_n+\rho\mathbf{A}^T\mathbf{A})^{-1}=\frac{1}{L_2}\mathbf{I}_n-\frac{\rho}{L_2(L_2+\rho)}\mathbf{A}^T\mathbf{A}.$$

In some applications with high-dimensional problems (e.g., $n=10^6$), the sensing matrix $\mathbf{A}$ is hardly explicitly available and instead implicit representations are usually used. In this case, the $\mathbf{x}$-subproblem can be computed as

$$\mathbf{x}^{k+1}=\frac{1}{L_2}\mathbf{z}^k-\frac{\rho}{L_2(L_2+\rho)}\mathbf{A}^T(\mathbf{A}\mathbf{z}^k)$$

with $\mathbf{z}^k=L_2\mathbf{x}^k-d_2(\mathbf{x}^k)+\rho\mathbf{A}^T\mathbf{u}^k$. This formulation facilitates the fast computation of $\mathbf{x}^{k+1}$ when the multiplication of $\mathbf{A}$ (and $\mathbf{A}^T$) with a vector can be rapidly obtained, e.g., for $\mathbf{A}$ be a partial DCT matrix.

C. Regularization Path for Robust Recovery

As well as other unconstrained formulations for sparse recovery such as (3) and (4), the performance of the proposed formulation is closely related to the selection of regularization parameter μ. In these problems, μ balances the fidelity and sparsity of the solution. A larger value of μ tends to give a sparser solution, but would result in larger residual error. In general, the optimal value is dependent on the noise, the true signal, and the value of p. A popular and useful approach is to compute the recovery along the regularization path, and select the optimal value based on the statistical information of the noise. More specifically, given an estimated noise variance σ_n^2, the optimal μ is selected as the maximum value of μ such that the bound constraint on the residual is met, e.g., $\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_2^2 \leq \epsilon$ with $\epsilon = m\sigma_n^2$ for the LASSO problem (3).

To extend this selection approach to the proposed robust formulation, we consider a generalized constraint on the residual, $\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_p^p \leq \epsilon$, where ϵ in this case bounds the p-th order moment of the noise. The p-th order moments for three typical impulsive noise are given as follows.

1) *GGD noise.* For i.i.d GGD noise, the p-th order moment is given by

$$E\{\|\mathbf{n}\|_p^p\} = \frac{m\sigma^p \Gamma(\frac{p+1}{v})}{\Gamma(\frac{1}{v})}. \tag{25}$$

2) *Gaussian mixture noise.* Consider a two-term Gaussian mixture model

$$(1-\xi)\mathcal{N}(0,\sigma^2) + \xi\mathcal{N}(0,\kappa\sigma^2) \tag{26}$$

where $0 \leq \xi < 1$ denotes the portion of outliers in the noise and $\kappa > 1$ indicates the strength of outliers. This model is an approximation to Middleton's Class A noise model, where the background noise is represented by the first term while the impulsive property of the noise is captured by the second term. The total noise variance is $(1 - \xi + \kappa\xi)\sigma^2$. The p-th order moment of such a noise process is given by

$$E\{\|\mathbf{n}\|_p^p\} = \frac{m 2^{\frac{p}{2}} \sigma^p \Gamma(\frac{p+1}{2})(1 - \xi + \kappa^{\frac{p}{2}}\xi)}{\sqrt{\pi}}. \tag{27}$$

3) *SαS noise.* For i.i.d. $S\alpha S$ noise with zero location parameter and scale parameter γ, it follows from (10) that

$$\begin{cases} E\{\|\mathbf{n}\|_p^p\} = mC(p,\alpha)\gamma^p, & 0 < p < \alpha \\ E\{\|\mathbf{n}\|_p^p\} = +\infty, & p \geq \alpha \end{cases}. \tag{28}$$

Such statistical information of the noise is essential in constructing the bound on the residual. In practical applications, such information can be estimated via incorporating (25)–(28) with standard statistical methods, e.g., robust parameter estimation for α-stable distributions [9], maximum likelihood parameter estimation for GGD [50].

V. Convergence Analysis

This section gives two convergence conditions of Lp-ADM when the $\mathbf{x}$-subproblem is updated via (21) for $p \geq 1$ and (24) for $p \geq 0$, respectively. While the convergence properties of ADM have been extensively studied for the convex case, there have been only a few studies for the nonconvex case. The convergence condition for the convex case is derived following straightforwardly from [23], whilst the condition for the nonconvex case is derived by extending the approaches proposed very recently in [51]–[53].

First, we give the convergence condition of Lp-ADM for arbitrary $p \geq 1$ when the $\mathbf{x}$-subproblem is updated via (21). The convergence for $p = 2$ and $p = 1$ has been analyzed in [23] and [24], respectively.

Theorem 2: For any $\rho > 0$, $p \geq 1$, and arbitrary starting point $(\mathbf{x}^0, \mathbf{w}^0)$, the sequence $\{(\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k)\}$ generated by Lp-ADM via (18), (21) and (20) with $L_1 > \lambda_{\max}(\mathbf{A}^T\mathbf{A})$ converges to $(\mathbf{v}^*, \mathbf{x}^*, \mathbf{w}^*)$, where $(\mathbf{v}^*, \mathbf{x}^*, \mathbf{w}^*)$ is a solution of (17).

Proof: See Appendix C. ■

Remark: The convergence condition given in Theorem 2 is the same as that for $p = 1$ [24, Theorem 1] and $p = 2$ [23, Theorem 2.1] (with a special choice of $\gamma = 1$). That is reasonable since in each iteration the $\mathbf{v}$-subproblem updated via (18) descends for any $p \geq 1$ at arbitrary $\mathbf{x}^k$ and $\mathbf{w}^k$, while the linearized $\mathbf{x}$-subproblem updated via (21) is guaranteed to descend at arbitrary $\mathbf{v}^{k+1}$ and $\mathbf{w}^k$ if the proximal parameter is chosen to be a Lipschitz constant of $d_1(\mathbf{x})$, i.e., $L_1 > \lambda_{\max}(\mathbf{A}^T\mathbf{A})$. From ADM theory, an ADM algorithm would converge even when the subproblems are not solved exactly, provided that certain suboptimality measures in the minimizations are summable [49].

Next, we give a sufficient condition for the convergence of Lp-ADM for the generalized case of $p \geq 0$ when the $\mathbf{x}$-subproblem is updated via (24).

Theorem 3: Suppose that $\varepsilon > 0$ and $\mathbf{A}\mathbf{A}^T \succeq \mu_A \mathbf{I}_m$ with some $\mu_A > 0$, then, for any $p \geq 0$ if $L_2 = \frac{\alpha}{\varepsilon} > \frac{1}{2\varepsilon}$ (i.e., $\alpha > \frac{1}{2}$) and

$$\rho > \frac{C}{\varepsilon} \quad \text{with} \quad C = \frac{4(2\alpha^2 + 2\alpha + 1)}{\mu_A(2\alpha - 1)}, \tag{29}$$

the sequence $\{(\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k)\}$ generated via (18), (24) and (20) converges to a stationary point of the problem (22).

Proof: See Appendix D. ■

Remark: The sufficient condition given in Theorem 3 is especially useful for the nonconvex case of $p < 1$. The assumption there exists some $\mu_A > 0$ requires that $\mathbf{A}$ be full-row rank, which is easily fulfilled in CS setting with $m < n$. For example, for an orthonormal sensing matrix, we have $\mu_A = 1$. When $\varepsilon \to 0$, the problem (22) reduces to the problem (17) and the error between the optimal solutions of these two problems vanishes. However, in this case the sufficient condition (29) requires $\rho \to \infty$. In general, an ADM tends to be very slow when the penalty parameter ρ gets very large. Thus, a tradeoff should be made between the approximating accuracy and the algorithm convergent rate. A warm-start strategy to speed up the algorithm is to use a properly small starting value of ρ and gradually increase it by iteration until reaching the target value. Theorem 3 still applies if ρ becomes fixed after a finite number of iterations. Moreover, in practical applications, an initialization is usually used for the nonconvex case, which is advantageous for improving the convergence of Lp-ADM. There also exist acceleration schemes proposed recently for ADM in [54], which can be used to effectively accelerate the new algorithm.

In the nonconvex case, a good initialization is crucial for Lp-ADM to achieve satisfactory performance. Intensive numerical studies show that, when the impulsive noise has finite variance, e.g., GGD and Gaussian mixture noise, Lp-ADM can achieve satisfactory performance with an initialization by a standard CS method, such as BPDN or LASSO. As shown in (8), the recovery error of such a method is finite and upper bounded by the noise strength in this case. However, when the impulsive noise has infinite variance, e.g., $S\alpha S$ noise with $\alpha < 2$, such an initialization may lead to poor performance of Lp-ADM, since a standard CS method is unstable in this case. For example, as will be seen in Section VI, the Homotopy solver, which solves the LASSO problem (3), performs quite poorly in highly impulsive $S\alpha S$ noise. Thus, it is recommended to employ a robust (convex) method for initialization, e.g., Lp-ADM with $p = 1$.

VI. Numerical Experiments

This section illustrates the robustness of the new method via numerical simulations, compared with a standard reconstruction algorithm, Homotopy [55], an ℓ_q-regularized algorithm, Lq-min [57], and four robust algorithms, Huber-fast iterative shrinkage/thresholding algorithm (FISTA) [27], YALL1 [23], BP-JP [31], and BP-SEP [58]. Homotopy solves the standard CS problem (3) and is generally faster than the interior-point algorithm. Lq-min solves a linear constrained ℓ_q-minimization problem with $0 < q \leq 1$. Huber-FISTA solves a robust formulation, which employs the Huber penalty function as the data-fitting model, using the FISTA. YALL1 solves the ℓ_1-ℓ_1 problem (4) using an ADM scheme. We also use an ADM procedure to solve BP-JP and BP-SEP by firstly converting the constrained formulations into unconstrained formulations. Matlab code for the proposed algorithm is available at https://github.com/FWen/Lp-Robust-CS.git.

We use a simulated K-sparse (with $K = 30$) signal of length $n = 512$ in the experiments, which is constructed as follows. First, the positions of the K nonzeros are uniformly randomly chosen. Then, the amplitude of each nonzero entry is generated according to the Gaussian distribution. An $m \times n$ orthonormal Gaussian random matrix is used as the sensing matrix $\mathbf{A}$. The number of random measurements is set to $m = 200$ unless otherwise specified. Each provided experimental result is an average over 200 independent runs, except for Figs. 1 and 7.

The algorithms are tested in three types of impulsive noise: GGD, Gaussian mixture, and $S\alpha S$. The noise is appropriately scaled and added to generate noisy measurements with desired noise levels. For the first two types of noise, the desired signal-to-noise ratio (SNR), measured in decibel (dB), is defined by

$$\mathrm{SNR} = 20\log_{10}\left(\frac{\|\mathbf{A}\mathbf{x}^o - E\{\mathbf{A}\mathbf{x}^o\}\|_2}{\|\mathbf{n}\|_2}\right)$$

where $\mathbf{x}^o$ stands for the true signal. As the variance of an $S\alpha S$ noise process is infinite for $\alpha < 2$, we use the scale parameter γ to quantify the strength of $S\alpha S$ impulsive noise. For GGD and Gaussian mixture noise, we assume the p-th order moments (25) and (27) are known in computing the regularization path for Lp-ADM ($0 \leq p \leq 2$), Homotopy ($p = 2$), and YALL1 ($p = 1$). For $S\alpha S$ noise with $\alpha < 2$, the true noise strength $\|\mathbf{n}\|_p^p$ is used in computing the regularization path for these algorithms, since the p-th order moment is infinite when $p \geq \alpha$.

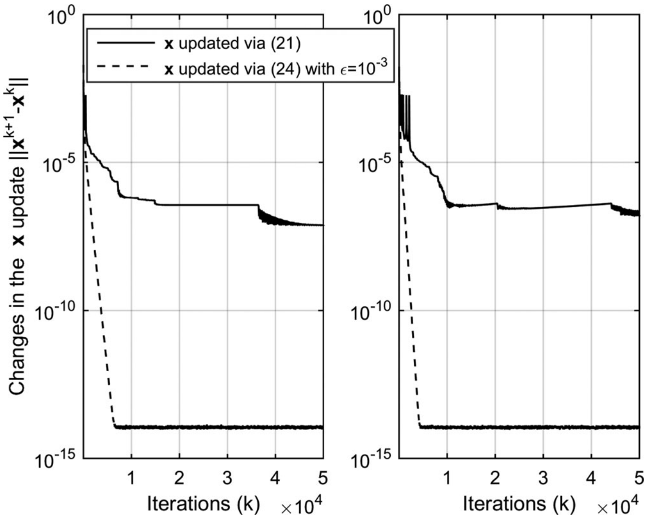

Fig. 1. Convergence behavior of Lp-ADM in the nonconvex case with $p = 0.5$, $\rho = 2 \times 10^4$, $\mu = 1$ and SNR $= 40$ dB. *Left:* Gaussian noise. *Right:* Impulsive Gaussian mixture noise with $\xi = 0.1$, $\kappa = 1000$.

When $p \geq 1$, Lp-ADM is run with $\rho = 10^2$ and the $\mathbf{x}$-subproblem is updated via (21) with $L_1 = 2$. When $p < 1$, Lp-ADM is run with $\rho = 2 \times 10^4$ and the $\mathbf{x}$-subproblem is updated via (24) with $\varepsilon = 10^{-3}$ and $L_2 = \frac{1}{\varepsilon}$. In implementing Lp-ADM in the nonconvex case, we firstly run it with $p = 1$ to obtain a starting point. Moreover, we set a stopping tolerance parameter of 10^{-6} (for both the primal and dual residuals) and a maximal iteration number of 2000 for it.

With the above settings, Lp-ADM is guaranteed to converge in the nonconvex case. Fig. 1 shows the typical convergence behavior of Lp-ADM in the nonconvex case in two conditions with the $\mathbf{x}$-subproblem is solved by (21) and (24), respectively. In both conditions, we set $\rho = 2 \times 10^4$. It can be seen that, Lp-ADM does not converge when using (21).

A. Performance of Lp-ADM in Different Impulsive Noise

In the first group of experiments, we evaluate the new algorithm in various noise conditions with different noise levels and types. The value of p is varied in the interval [0 2]. The performance is evaluated in terms of relative error of recovery defined as $\|\hat{\mathbf{x}} - \mathbf{x}^o\|_2/\|\mathbf{x}^o\|_2$, where $\hat{\mathbf{x}}$ is the recovered version of the true signal $\mathbf{x}^o$.

1) Generalized Gaussian Noise: Fig. 2 displays the averaged relative error of recovery of Lp-ADM versus p in GGD noise. Different values of the shape parameter are considered to generate GGD noise with different impulsive properties, e.g., $v = 2$ (Gaussian noise), $v = 1$ (Laplace noise), $v = 0.5$, and $v = 0.2$. It can be clearly seen from Fig. 2 that, in the Gaussian and Laplace noise conditions, using $p = 2$ generally yields better recovery performance than using $p < 2$. Only in the case of $v = 0.2$, the most impulsive case among the four, using $p < 2$ can gain distinct advantage over $p = 2$. The results indicate that, for GGD noise, the ℓ_2-norm loss function performs sufficiently well in Gaussian and slightly impulsive conditions, e.g., $v \leq 0.5$, but is inefficient in more impulsive conditions, e.g., $v = 0.2$.

Another interesting observation is that, when the noise level is fixed, using $p = 2$ yields approximately the same

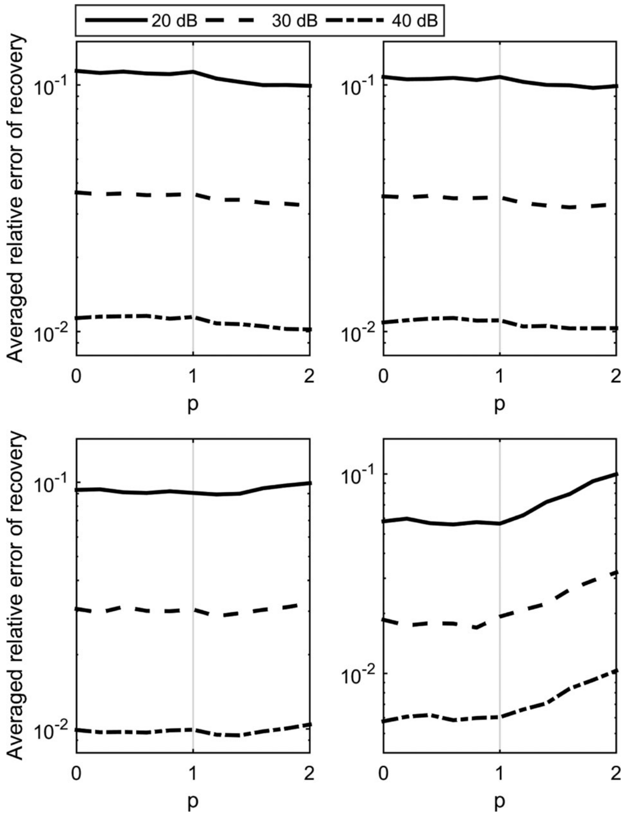

Fig. 2. Recovery performance of Lp-ADM versus p in generalized Gaussian noise. *Top Left:* $v = 2$ (Gaussian noise). *Top Right:* $v = 1$ (Laplace noise). *Bottom left:* $v = 0.5$. *Bottom right:* $v = 0.2$.

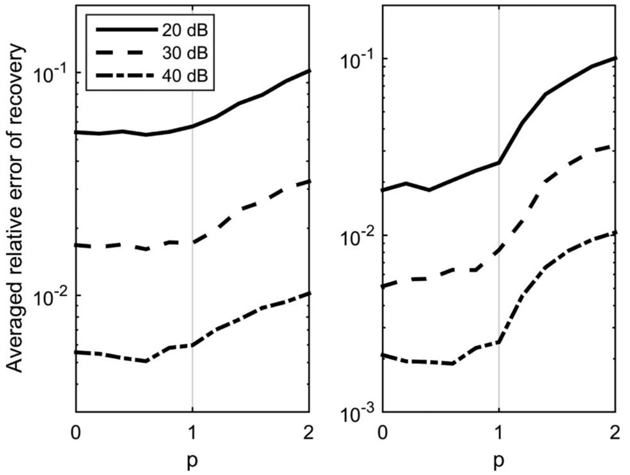

Fig. 3. Recovery performance of Lp-ADM versus p in Gaussian mixture noise with $\xi = 0.1$. *Left:* $\kappa = 100$. *Right:* $\kappa = 1000$.

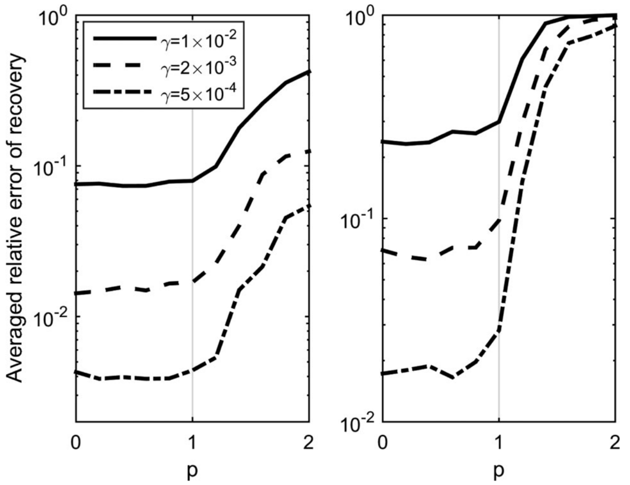

Fig. 4. Recovery performance of Lp-ADM versus p in $S\alpha S$ noise. *Left:* $\alpha = 1$ (Cauchy noise). *Right:* $\alpha = 0.5$.

recovery performance in the four conditions with different impulsiveness. This is reasonable since the recovery error of the ℓ_2-ℓ_1 formulation is bounded by the noise variance (see (8)) which does not change when the shape parameter v varies from 2 to 0 in the case of fixed noise power. However, the p-th ($p < 2$) order moment of the noise decreases significantly when v varies from 2 to 0 (see (25)). Thus, from Theorem 1, performance gain can be expected via using $p < 2$ when the noise gets more impulsive (i.e., v gets small). For example, in the case of SNR = 40 dB, the recovery errors of Lp-ADM using p = 0.6 are 1.16 × 10^{-2}, 1.13 × 10^{-2}, 0.97 × 10^{-2}, and 0.58 × 10^{-2}, respectively, in the four conditions. Note that, the discussion here holds only when the noise has finite variance, e.g., GGD and Gaussian mixture noise, and it breaks down when the noise has infinite variance, e.g., $S\alpha S$ impulsive noise as shown in Fig. 4.

2) Gaussian Mixture Noise: Fig. 3 plots the recovery performance of Lp-ADM versus p in Gaussian mixture noise with $\xi = 0.1$. Two impulsive conditions with $\kappa = 100$ and $\kappa = 1000$ are considered. It is clear that, in both conditions, the averaged recovery error is approximately a monotonically increasing function of p when $p > 0.5$. Using an appropriate $p < 1$ has the potential to achieve distinctly better performance compared with $p \geq 1$. This advantage is more significant in the more impulsive condition with $\kappa = 1000$. Again, it can be observed that, when the SNR is fixed, using $p = 2$ generally yields the same performance in the two conditions, but using $p < 2$ (with a fixed value of p) can result in more accurate recovery in the more impulsive condition. For example, in the case of SNR = 40 dB, the recovery errors of Lp-ADM using $p = 0.6$ are 5.0 × 10^{-3} and 1.9 × 10^{-4} respectively in the two conditions.

3) $S\alpha S$ Noise: Fig. 4 shows the recovery performance of Lp-ADM versus p in $S\alpha S$ noise. Two impulsive conditions, with characteristic exponents $\alpha = 1$ (Cauchy noise) and $\alpha = 0.5$, and three noise levels, with scale parameters of $\gamma \in \{10^{-2}, 2 \times 10^{-3}, 5 \times 10^{-4}\}$, are considered. Unlike the cases of GGD and Gaussian mixture noise, the variance of such $S\alpha S$ noise is infinite. In this case, the ℓ_2-norm loss formulation is unstable in statistics. Accordingly, as shown in Fig. 4, the performance corresponds to $p = 2$ deteriorates drastically when the noise gets more impulsive. Meanwhile, using an appropriately smaller value of p can yield significantly better performance, especially in the more impulsive condition with $\alpha = 0.5$.

From the results across Figs. 2 to 4, selecting a value $p < 1$ in the interval [0 0.8] is recommended. Such a choice has the potential to yield distinctly more accurate recovery than $p \geq 1$ in highly impulsive noise. Even in the conditions with slightly or non-impulsive noise, e.g., white Gaussian noise, it does not lead to significant performance loss compared with $p = 2$.

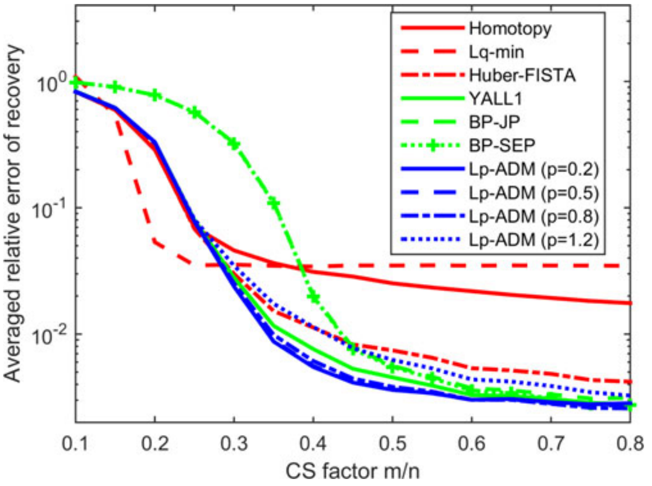

Fig. 5. Recovery performance versus CS factor m/n for the compared algorithms in Gaussian mixture noise with $\xi = 0.1$, $\kappa = 1000$, SNR = 30 dB.

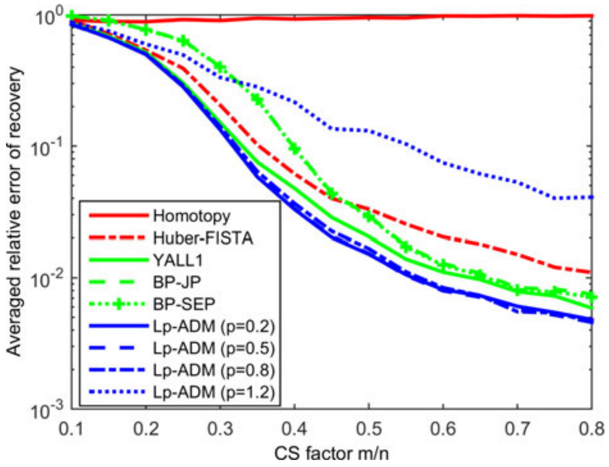

Fig. 6. Recovery performance versus CS factor m/n for the compared algorithms in $S\alpha S$ noise with $\alpha = 0.5$ and $\gamma = 1 \times 10^{-3}$.

B. Comparison With Existing Methods

In the second group of experiments, we compare the proposed algorithm with the Homotopy, Lq-min, Huber-FISTA, BP-JP, BP-SEP, and YALL1 algorithms. Figs. 5 and 6 show the performance of the compared algorithms versus CS factor m/n respectively for the two noise conditions: Gaussian mixture noise with $\xi = 0.1$, $\kappa = 1000$, and SNR = 30 dB, and $S\alpha S$ noise with $\alpha = 0.5$ and $\gamma = 1 \times 10^{-3}$. We set $K = 30$ and $n = 512$. Four typical values of p, $p \in \{0.2,\ 0.5,\ 0.8,\ 1.2\}$, are examined for the new algorithm.

It can be seen from Figs. 5 and 6 that, Huber-FISTA, YALL1 and Lp-ADM distinctly outperform Homotopy when $m/n > 0.25$ in the case of Gaussian mixture noise or when $m/n > 0.1$ in the case of $S\alpha S$ noise. As the CS factor increases, the recovery accuracy of each robust algorithm improves significantly in both conditions, but that of Homotopy does not improve distinctly in the $S\alpha S$ noise condition. This is due to the fact that the considered $S\alpha S$ noise is highly impulsive, and the ℓ_2-norm loss function is very sensitive to extremely large outliers. When $m/n > 0.3$ in the case of Gaussian mixture noise or when $m/n > 0.25$ in the case of $S\alpha S$ noise, Lp-ADM with $p < 1$ achieves better performance than Huber-FISTA and YALL1. That advantage is more significant in the more impulsive case of $S\alpha S$ noise.

In Gaussian mixture noise, Lq-min has better performance than the other algorithms when m/n is relatively small, which is due to the fact that ℓ_q-regularized methods require fewer measurements to achieve reliable reconstruction than ℓ_1-regularized methods. However, Lq-min breaks down in the case of α-stable noise. BP-JP is outperformed by YALL1 and Lp-ADM with $p < 1$ in most cases. It is reasonable since BP-JP in fact solves the ℓ_1-ℓ_1 problem (4) with $\mu = 1$ and thus is a special case of YALL1. Since YALL1 often attains the best performance at a value $\mu \neq 1$, it outperforms BP-JP in most cases. Moreover, BP-SEP performs comparably as BP-JP in the considered conditions.

Next, we consider a practical condition that the measurements are contaminated by bit errors like corruption, which causes potentially unbounded errors in the measurements. Such arbitrary large error can model the corruption caused by bit errors in transmission, malfunctioning pixels, and faulty memory locations. The test signal with length 256 has 15 non-zero values. The measurement matrix is a 100 × 256 orthonormal Gaussian random matrix. 10% of the measurements are randomly set to be ±1000, which models the arbitrary unbounded errors. Moreover, zero-mean Gaussian noise with variance 10^{-4} is added to model small background noise.

In this condition, both Homotopy and Lq-min break down. Fig. 7 shows the recovery performance of the compared robust algorithms, including the recovered signal and the corresponding relative error of recovery (RelErr) for each algorithm. It can be seen that, each robust algorithm achieves a RelErr less than 20%. Moreover, Lp-ADM with $p \in \{0.5,\ 0.8\}$ significantly outperforms BP-JP, YALL1 and Huber-FISTA. In the two cases with $p \in \{0.5,\ 0.8\}$, the RelErr of Lp-ADM are approximately 64%, 52%, and 29% that of YALL1, Huber-FISTA, and BP-JP, respectively.

On the whole, Huber-FISTA is more robust than Homotopy but less robust than YALL1. That is due the nature that Huber function fitting lies in between the least-squares and least-absolute-deviations. Along with the results in [27] that the least-absolute based algorithm gives the best performance compared with the Lorentzian-BP [20], Huber-FISTA and its several variants, Lp-ADM with $p < 1$ can achieve state-of-the-art robust performance in highly impulsive noise.

Finally, we compare the computational complexity of the robust algorithms. For BP-JP and BP-SEP solved via ADM, the dominant computational load in each iteration is matrix-vector multiplication (involving an $m \times (m + n)$ matrix) with complexity $O(mn + m^2)$, which is the same as that of YALL1. When the $\mathbf{x}$-subproblem of Lp-ADM is solved by (21) (e.g., for $p \geq 1$), Lp-ADM costs $O(mn)$ flops in each iteration, which is the same as that of Huber-FISTA. Lp-ADM with the $\mathbf{x}$-subproblem updated via (24) requires the inversion of $n \times n$ matrices. Using matrix inversion lemma and Cholesky decomposition, we only need to factor an $m \times m$ matrix once and can use cheaper back-solve in computing the

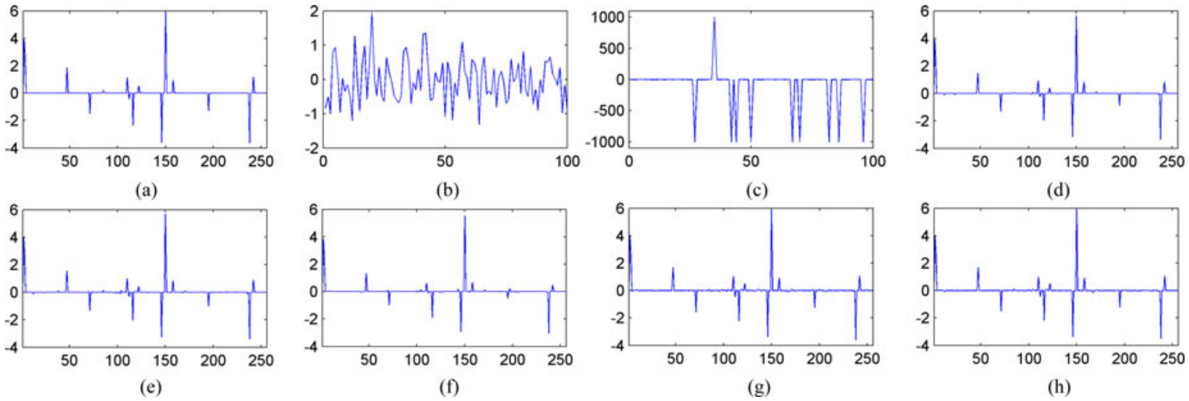

Fig. 7. Recovery performance of the compared algorithms in the presence of bit errors like corruption. (a) Test signal. (b) Measurements without noise. (c) Corrupted measurements. (d) Huber-FISTA, RelErr $= 0.110$. (e) YALL1, RelErr $= 0.090$. (f) BP-JP, RelErr $= 0.197$. (g) Lp-ADM ($p = 0.5$), RelErr $= 0.058$. (h) Lp-ADM ($p = 0.8$), RelErr $= 0.057$.

inverse based updates [48]. In such a manner, Lp-ADM cost $O(m^2 n)$ flops in the first iteration and $O(mn)$ flops in each of the subsequent iterations. As discussed in Section VI-B, when $\mathbf{A}$ is orthonormal, the inversion in (24) can be avoided. For the setting in the last experiment and on a desktop PC with an Intel Core i5-4670 CPU at 3.4 GHz with 8 GB RAM, the average runtime of YALL1 and Huber-FISTA are approximately 0.6 and 0.4 seconds, respectively, while that of Lp-ADM for different $p < 1$ ranges from 2.6 to more than three seconds.

VII. Conclusion

This work introduced a robust formulation for sparse recovery, which employs the ℓ_p-norm with $0 \leq p < 2$ as the metric for the residual error. An efficient algorithm has been proposed to solve this formulation via incorporating the generalized proximity operator for ℓ_p-norm functions into the framework of augmented Lagrangian methods. In such a manner, both the convex ($p \geq 1$) and nonconvex ($p < 1$) cases of the introduced formulation have been casted into a unified framework. Moreover, we have analyzed the convergence condition of the new algorithm for both the convex and nonconvex cases. Simulation results showed that, in the presence of highly impulsive measurement noise, the new algorithm with an appropriate choice of p ($p < 1$) has the capability to achieve distinctly better recovery accuracy compared with existing robust algorithms.

Appendix A
Proof of Theorem 1

The derivation of (9) is similar to that of (8) in [5], [41]. Briefly, since $\hat{\mathbf{x}}$ is a feasible point of the optimization problem (6) and the noise obeys $\|\mathbf{n}\|_p \leq \varepsilon$, it yields $\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_p \leq \varepsilon$ and $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_p \leq \epsilon$. Then, let $\mathbf{h} = \hat{\mathbf{x}} - \mathbf{x}$, we have

$$\begin{aligned} \|\mathbf{A}\mathbf{h}\|_2 &\leq \|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_2 + \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \\ &\leq \|\mathbf{A}\hat{\mathbf{x}} - \mathbf{y}\|_p + \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_p \\ &\leq 2\epsilon \end{aligned} \tag{30}$$

where the second inequality follows from that $\|\mathbf{z}\|_2 \leq \|\mathbf{z}\|_p$ holds for arbitrary $\mathbf{z} \in \mathbb{R}^m$ if $0 < p \leq 2$. It has been shown in [41] that

$$\|\mathbf{h}\|_2 \leq \frac{\sqrt{2 + 2\delta_{2s}}\|\mathbf{A}\mathbf{h}\|_2}{1 - \delta_{2s} - \sqrt{2}\delta_{2s}} \tag{31}$$

holds if $\delta_{2s} < \sqrt{2} - 1$. Consequently, plugging (30) into (31) yields (9) with $C_s = 2\sqrt{2 + 2\delta_{2s}}(1 - \delta_{2s} - \sqrt{2}\delta_{2s})^{-1}$.

Appendix B
The Derivation of (16)

Since $h_2(0) < 0$ when $t_i \neq 0$, $z^0 = 0$ is a lower bound for the solution. However, we cannot use a starting point $z^0 = 0$ since $h_2'(z) \rightarrow \infty$ as $z \rightarrow 0$. In the following, we find a positive lower bound for the solution.

Since $h_2(z)$ is a monotonously increasing function of $z > 0$, the solution of $h_2(z) = 0$, denoted by z_0, satisfies that

$$\begin{cases} 0 < z_0 < 1, & \text{if } h_2(1) > 0 \\ z_0 \geq 1, & \text{if } h_2(1) \leq 0 \end{cases}. \tag{32}$$

Further, for $1 < p < 2$, it is easy to observe that

$$\begin{cases} paz_0^{p-1} + \eta z_0^{p-1} - \eta|t_i| > 0, & \text{if } 0 < z_0 < 1 \\ paz_0 + \eta z_0 - \eta|t_i| \geq 0, & \text{if } z_0 \geq 1 \end{cases}. \tag{33}$$

Then, it follows from (32) and (33) that

$$\begin{cases} [\eta|t_i|/(pa + \eta)]^{1/(p-1)} < z_0 < 1, & \text{if } h_2(1) > 0 \\ 1 \leq \eta|t_i|/(pa + \eta) \leq z_0, & \text{if } h_2(1) \leq 0 \end{cases}$$

which finally results in (16).

Appendix C
Proof of Theorem 2

Let $(\mathbf{v}^*, \mathbf{x}^*)$ be any solution of (17), accordingly, there exists $\mathbf{w}^* \in \mathbb{R}^m$ such that the following optimality conditions are satisfied

$$\frac{1}{\mu}\nabla\|\mathbf{v}^*\|_p^p = -\mathbf{w}^*, \ \mathbf{A}^T\mathbf{w}^* \in \partial\|\mathbf{x}^*\|_1, \text{ and } \mathbf{A}\mathbf{x}^* - \mathbf{y} = \mathbf{v}^*.$$

For $\mathbf{x} = \mathbf{x}^k$ and $\mathbf{w} = \mathbf{w}^k$ fixed, the minimizer $\mathbf{v}^{k+1}$ of (18) satisfies

$$\frac{1}{\mu}\nabla\|\mathbf{v}^{k+1}\|_p^p - \rho(\mathbf{A}\mathbf{x}^k - \mathbf{y} - \mathbf{v}^{k+1} - \mathbf{w}^k/\rho) = 0. \qquad (34)$$

Plugging (20) into (34) yields

$$\frac{1}{\mu}\nabla\|\mathbf{v}^{k+1}\|_p^p + \mathbf{w}^{k+1} - \rho\mathbf{A}(\mathbf{x}^k - \mathbf{x}^{k+1}) = 0. \qquad (35)$$

Then, it follows from (35) and $\frac{1}{\mu}\nabla\|\mathbf{v}^*\|_p^p = -\mathbf{w}^*$ that

$$\mathbf{w}^* - \mathbf{w}^{k+1} + \rho\mathbf{A}(\mathbf{x}^k - \mathbf{x}^{k+1}) = \frac{1}{\mu}(\nabla\|\mathbf{v}^{k+1}\|_p^p - \nabla\|\mathbf{v}^*\|_p^p)$$

which further results in

$$\begin{aligned}&(\mathbf{v}^* - \mathbf{v}^{k+1})^T(\mathbf{w}^{k+1} - \mathbf{w}^* - \rho\mathbf{A}(\mathbf{x}^k - \mathbf{x}^{k+1}))\\ &= \frac{1}{\mu}(\mathbf{v}^{k+1} - \mathbf{v}^*)^T\left(\nabla\|\mathbf{v}^{k+1}\|_p^p - \nabla\|\mathbf{v}^*\|_p^p\right)\\ &\geq 0. \qquad (36)\end{aligned}$$

The inequality in (36) is due to the convexity of $\|\cdot\|_p^p$, which holds for arbitrary $p \geq 1$. Inequation (36) is just the inequation (A.2) in [23, Theorem 2.1], and the rest of the proof follows similarly the proof of Theorem 2.1 in [23], which is omitted here for succinctness.

Appendix D
Proof of Theorem 3

We first prove the following lemmas in the proof of Theorem 3.

Lemma 1: Let $h(\mathbf{x}) = \frac{\rho}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y} - \mathbf{v}^{k+1} - \frac{1}{\rho}\mathbf{w}^k\|_2^2$. For any $L_2 > \frac{1}{2\varepsilon}$ and $\mathbf{x}^k \in \mathbb{R}^n$, the minimizer $\mathbf{x}^{k+1}$ given by (24) satisfies

$$\begin{aligned}&\|\mathbf{x}^{k+1}\|_{1,\varepsilon} + h(\mathbf{x}^{k+1})\\ &\leq \|\mathbf{x}^k\|_{1,\varepsilon} + h(\mathbf{x}^k) - \left(L_2 - \frac{1}{2\varepsilon}\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2.\end{aligned}$$

Lemma 2: Suppose that $\varepsilon > 0$, $L_2 > \frac{1}{2\varepsilon}$, $\mathbf{A}\mathbf{A}^T \succeq \mu_A\mathbf{I}_m$ with some $\mu_A > 0$ and (29) holds, then

$$\begin{aligned}\hat{\mathcal{L}}(\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k, \mathbf{x}^{k-1}) &\geq \hat{\mathcal{L}}(\mathbf{v}^{k+1}, \mathbf{x}^{k+1}, \mathbf{w}^{k+1}, \mathbf{x}^k)\\ &\quad + c_1\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2\end{aligned}$$

where $\hat{\mathcal{L}}(\mathbf{v}, \mathbf{x}, \mathbf{w}, \hat{\mathbf{x}}) := \mathcal{L}_{\rho,\varepsilon}(\mathbf{v}, \mathbf{x}, \mathbf{w}) + c_0\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$ with $c_0, c_1 > 0$ are given by

$$c_0 = \frac{2(L_2 + \frac{1}{\varepsilon})^2}{\rho\mu_A}, \; c_1 = L_2 - \frac{1}{2\varepsilon} - \frac{2L_2^2 + 2(L_2 + \frac{1}{\varepsilon})^2}{\rho\mu_A}.$$

Lemma 3: Let $\mathbf{z}^k := (\mathbf{v}^k, \mathbf{x}^k, \mathbf{w}^k)$. Suppose that $\varepsilon > 0$, $L_2 > \frac{1}{2\varepsilon}$, $\mathbf{A}\mathbf{A}^T \succeq \mu_A\mathbf{I}_m$ with some $\mu_A > 0$ and (29) holds, then

$$\lim_{k\to\infty}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_2^2 = 0.$$

Moreover, any cluster point of $\mathbf{z}^k$ is a stationary point of $\mathcal{L}_{\rho,\varepsilon}$.

Proof of Lemma 1: First, the Hessian of $\|\mathbf{x}\|_{1,\varepsilon}$ is

$$\nabla^2\|\mathbf{x}\|_{1,\varepsilon} = \varepsilon^2\mathrm{diag}\{(x_1^2+\varepsilon^2)^{-\frac{3}{2}}, \ldots, (x_N^2+\varepsilon^2)^{-\frac{3}{2}}\} \preceq \frac{1}{\varepsilon}\mathbf{I}_n \qquad (37)$$

which implies that the gradient of $\|\mathbf{x}\|_{1,\varepsilon}$ is $\frac{1}{\varepsilon}$-Lipschitz continuous, thus, for any $\mathbf{x}^k, \mathbf{x}^{k+1} \in \mathbb{R}^n$ we have

$$\begin{aligned}\|\mathbf{x}^{k+1}\|_{1,\varepsilon} &\leq \|\mathbf{x}^k\|_{1,\varepsilon} + \langle\mathbf{x}^{k+1} - \mathbf{x}^k, \nabla\|\mathbf{x}^k\|_{1,\varepsilon}\rangle\\ &\quad + \frac{1}{2\varepsilon}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2. \qquad (38)\end{aligned}$$

Moreover, the $\mathbf{x}$-subproblem actually minimizes the following approximate objective

$$Q_{\mathbf{x}^k}(\mathbf{x}) = \langle\mathbf{x}-\mathbf{x}^k, \nabla\|\mathbf{x}^k\|_{1,\varepsilon}\rangle + \frac{L_2}{2}\|\mathbf{x}-\mathbf{x}^k\|_2^2 + h(\mathbf{x}). \quad (39)$$

Since $Q_{\mathbf{x}^k}(\mathbf{x})$ is L_2-strongly convex, for any $\mathbf{x}^k \in \mathbb{R}^n$ we have

$$\begin{aligned}Q_{\mathbf{x}^k}(\mathbf{x}^k) &\geq Q_{\mathbf{x}^k}(\mathbf{x}^{k+1}) + \langle\mathbf{x}^k - \mathbf{x}^{k+1}, \nabla Q_{\mathbf{x}^k}(\mathbf{x}^{k+1})\rangle\\ &\quad + \frac{L_2}{2}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|_2^2. \qquad (40)\end{aligned}$$

From the definition of $\mathbf{x}^{k+1}$ as a minimizer of $Q_{\mathbf{x}^k}(\mathbf{x})$, we have $\nabla Q_{\mathbf{x}^k}(\mathbf{x}^{k+1}) = \mathbf{0}$. Further, with $Q_{\mathbf{x}^k}(\mathbf{x}^k) = h(\mathbf{x}^k)$, it follows from (39) and (40) that

$$\langle\mathbf{x}^{k+1}-\mathbf{x}^k, \nabla\|\mathbf{x}^k\|_{1,\varepsilon}\rangle + h(\mathbf{x}^{k+1}) \leq h(\mathbf{x}^k) - L_2\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_2^2$$

which together with (38) results in Lemma 1.

Proof of Lemma 2: First, we show that the changes in the dual iterates can be bounded by the changes in the primal iterates

$$\begin{aligned}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|_2^2 &\leq \frac{2L_2^2}{\mu_A}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2\\ &\quad + \frac{2(L_2 + \frac{1}{\varepsilon})^2}{\mu_A}\|\mathbf{x}^k - \mathbf{x}^{k-1}\|_2^2. \qquad (41)\end{aligned}$$

Observe that the approximate $\mathbf{x}$-subproblem actually minimizes $Q_{\mathbf{x}^k}(\mathbf{x})$ in (40), whose minimizer $\mathbf{x}^{k+1}$ satisfies

$$\begin{aligned}\mathbf{0} &= \nabla\|\mathbf{x}^k\|_{1,\varepsilon} + L_2(\mathbf{x}^{k+1} - \mathbf{x}^k)\\ &\quad + \rho\mathbf{A}^T(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{y} - \mathbf{v}^{k+1} - \mathbf{w}^k/\rho). \qquad (42)\end{aligned}$$

Plugging (20) into (42) yields

$$\mathbf{A}^T\mathbf{w}^{k+1} = \nabla\|\mathbf{x}^k\|_{1,\varepsilon} + L_2(\mathbf{x}^{k+1} - \mathbf{x}^k). \qquad (43)$$

Then, it follows that

$$\begin{aligned}&\|\mathbf{A}^T(\mathbf{w}^{k+1} - \mathbf{w}^k)\|_2^2\\ &\leq \left(\left\|\nabla\|\mathbf{x}^k\|_{1,\varepsilon} - \nabla\|\mathbf{x}^{k-1}\|_{1,\varepsilon}\right\|_2 + L_2\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2\right.\\ &\qquad\left. + L_2\|\mathbf{x}^k - \mathbf{x}^{k-1}\|_2\right)^2\\ &\leq \left(\frac{1}{\varepsilon}\|\mathbf{x}^k - \mathbf{x}^{k-1}\|_2 + L_2\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2 + L_2\|\mathbf{x}^k - \mathbf{x}^{k-1}\|_2\right)^2\\ &\leq 2L_2^2\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 + 2\left(L_2 + \frac{1}{\varepsilon}\right)^2\|\mathbf{x}^k - \mathbf{x}^{k-1}\|_2^2 \qquad (44)\end{aligned}$$

where the second inequality follows from (37) that $\nabla\|\mathbf{x}\|_{1,\varepsilon}$ is $\frac{1}{\varepsilon}$-Lipschitz continuous. Further, since $\mathbf{A}\mathbf{A}^T \succeq \mu_A\mathbf{I}_m$ for some $\mu_A > 0$, we have

$$\left\|\mathbf{A}^T(\mathbf{w}^{k+1}-\mathbf{w}^k)\right\|_2^2 \ge \mu_A\left\|\mathbf{w}^{k+1}-\mathbf{w}^k\right\|_2^2$$

which together with (44) results in (41).

From (20) and the definition of $\mathcal{L}_{\rho,\varepsilon}$, we have

$$\begin{aligned}\mathcal{L}_{\rho,\varepsilon}(\mathbf{v}^{k+1},\mathbf{x}^{k+1},\mathbf{w}^{k+1}) - \mathcal{L}_{\rho,\varepsilon}(\mathbf{v}^{k+1},\mathbf{x}^{k+1},\mathbf{w}^{k})\\ = \frac{1}{\rho}\|\mathbf{w}^{k+1}-\mathbf{w}^k\|_2^2.\end{aligned} \tag{45}$$

Moreover, it follows from Lemma 1 that

$$\begin{aligned}\mathcal{L}_{\rho,\varepsilon}(\mathbf{v}^{k+1},\mathbf{x}^{k+1},\mathbf{w}^{k}) - \mathcal{L}_{\rho,\varepsilon}(\mathbf{v}^{k+1},\mathbf{x}^{k},\mathbf{w}^{k})\\ \le -\Big(L_2-\frac{1}{2\varepsilon}\Big)\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_2^2.\end{aligned} \tag{46}$$

Further, from the definition of $\mathbf{v}^{k+1}$ as a minimizer, we have

$$\mathcal{L}_{\rho,\varepsilon}(\mathbf{v}^{k+1},\mathbf{x}^{k},\mathbf{w}^{k}) - \mathcal{L}_{\rho,\varepsilon}(\mathbf{v}^{k},\mathbf{x}^{k},\mathbf{w}^{k}) \le 0. \tag{47}$$

Summing (45), (46) and (47), and using (41) we obtain that

$$\begin{aligned}&\mathcal{L}_{\rho,\varepsilon}(\mathbf{v}^{k+1},\mathbf{x}^{k+1},\mathbf{w}^{k+1}) - \mathcal{L}_{\rho,\varepsilon}(\mathbf{v}^{k},\mathbf{x}^{k},\mathbf{w}^{k})\\ &\le \Big(\frac{2L_2^2}{\rho\mu_A}-L_2+\frac{1}{2\varepsilon}\Big)\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_2^2\\ &\quad+\frac{2\big(L_2+\frac{1}{\varepsilon}\big)^2}{\rho\mu_A}\|\mathbf{x}^{k}-\mathbf{x}^{k-1}\|_2^2\end{aligned}$$

which consequently results in Lemma 2, where c_1 is guaranteed to be positive if

$$\rho > \frac{2L_2^2+2\big(L_2+\frac{1}{\varepsilon}\big)^2}{\mu_A\big(L_2-\frac{1}{2\varepsilon}\big)}.$$

Proof of Lemma 3: First, we show that the sequence $\{\mathbf{z}^k\}$ is bounded if $\mu_A > 0$. From (43), we see that

$$\begin{aligned}\mu_A\|\mathbf{w}^k\|_2^2 &\le \|\mathbf{A}^T\mathbf{w}^k\|_2^2\\ &\le \big(\left\|\nabla\|\mathbf{x}^{k-1}\|_{1,\varepsilon}\right\|_2 + L_2\|\mathbf{x}^k-\mathbf{x}^{k-1}\|_2\big)^2\\ &\le 2\left\|\nabla\|\mathbf{x}^{k-1}\|_{1,\varepsilon}\right\|_2^2 + 2L_2^2\|\mathbf{x}^k-\mathbf{x}^{k-1}\|_2^2\\ &\le 2n + 2L_2^2\|\mathbf{x}^k-\mathbf{x}^{k-1}\|_2^2\end{aligned} \tag{48}$$

where the last inequality follows from $\left\|\nabla\|\mathbf{x}^k\|_{1,\varepsilon}\right\|_2^2 \le n$. Define $\hat{\mathbf{z}}^k := (\mathbf{v}^k,\mathbf{x}^k,\mathbf{w}^k,\mathbf{x}^{k-1})$, since $\hat{\mathcal{L}}(\hat{\mathbf{z}}^k)$ is lower semicontinuous, it is bounded from below. Further, by Lemma 2, $\hat{\mathcal{L}}(\hat{\mathbf{z}}^k)$ is nonincreasing when the condition (29) is satisfied, then, we have

$$\begin{aligned}&\hat{\mathcal{L}}(\hat{\mathbf{z}}^1) \ge \hat{\mathcal{L}}(\hat{\mathbf{z}}^k)\\ &= \frac{1}{\mu}\|\mathbf{v}^k\|_p^p + \|\mathbf{x}^k\|_{1,\varepsilon} + \frac{\rho}{2}\|\mathbf{A}\mathbf{x}^k-\mathbf{y}-\mathbf{v}^{k+1}-\mathbf{w}^k/\rho\|_2^2\\ &\quad-\frac{1}{2\rho}\|\mathbf{w}^k\|_2^2 + c_0\|\mathbf{x}^k-\mathbf{x}^{k-1}\|_2^2\\ &\ge \frac{1}{\mu}\|\mathbf{v}^k\|_p^p + \|\mathbf{x}^k\|_{1,\varepsilon} + \frac{\rho}{2}\|\mathbf{A}\mathbf{x}^k-\mathbf{y}-\mathbf{v}^{k+1}-\mathbf{w}^k/\rho\|_2^2\\ &\quad-\frac{n}{\rho\mu_A} + \frac{2\big(L_2+\frac{1}{\varepsilon}\big)^2-L_2^2}{\rho\mu_A}\|\mathbf{x}^k-\mathbf{x}^{k-1}\|_2^2\end{aligned}$$

where the last inequality follows from (48). Since $\|\mathbf{v}^k\|_p^p$ and $\|\mathbf{x}^k\|_{1,\varepsilon}$ are coercive and by (48), it is easy to see that $\mathbf{v}^k$, $\mathbf{x}^k$ and $\mathbf{w}^k$ are bounded.

Since $\hat{\mathbf{z}}^k$ is bounded, there exists a convergent subsequence $\hat{\mathbf{z}}^{k_j}$ which converges to a cluster point $\hat{\mathbf{z}}^*$. Moreover, $\hat{\mathcal{L}}(\hat{\mathbf{z}}^k)$ is convergent and $\hat{\mathcal{L}}(\hat{\mathbf{z}}^k) \ge \hat{\mathcal{L}}(\hat{\mathbf{z}}^*)$ for any k if $c_1 > 0$. Then, it follows from Lemma 2 that

$$\begin{aligned}c_1\sum_{k=1}^{N}\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|_2^2 &\le \sum_{k=1}^{N}\hat{\mathcal{L}}(\hat{\mathbf{z}}^k)-\hat{\mathcal{L}}(\hat{\mathbf{z}}^{k+1})\\ &= \hat{\mathcal{L}}(\hat{\mathbf{z}}^1)-\hat{\mathcal{L}}(\hat{\mathbf{z}}^{k+1})\\ &\le \hat{\mathcal{L}}(\hat{\mathbf{z}}^1)-\hat{\mathcal{L}}(\hat{\mathbf{z}}^{*}) < \infty.\end{aligned}$$

With $N\to\infty$, we have $\sum_{k=1}^{\infty}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_2^2 < \infty$, which together with (41) implies $\sum_{k=1}^{\infty}\|\mathbf{w}^{k+1}-\mathbf{w}^k\|_2^2 < \infty$. Moreover, it follows from (20) that

$$\begin{aligned}\|\mathbf{v}^{k+1}-\mathbf{v}^k\|_2 \le \|\mathbf{A}(\mathbf{x}^{k+1}-\mathbf{x}^k)\|_2 + \frac{1}{\rho}\|\mathbf{w}^{k+1}-\mathbf{w}^k\|_2\\ +\frac{1}{\rho}\|\mathbf{w}^{k}-\mathbf{w}^{k-1}\|_2.\end{aligned}$$

Thus, we have $\sum_{k=1}^{\infty}\|\mathbf{v}^{k+1}-\mathbf{v}^k\|_2^2 < \infty$. In particular $\sum_{k=1}^{\infty}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_2^2 < \infty$ and $\lim\limits_{k\to\infty}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_2^2 = 0$.

Finally, we show that any cluster point of $\{\mathbf{z}^k\}$ is a stationary point. From the optimality conditions and the definition of $\mathbf{w}^{k+1}$, the iterates satisfy

$$\begin{cases}\mathbf{0} \in \frac{1}{\mu}\partial\|\mathbf{v}^{k+1}\|_p^p + \mathbf{w}^{k+1} + \rho\mathbf{A}(\mathbf{x}^{k+1}-\mathbf{x}^k)\\ \mathbf{0} = \nabla\|\mathbf{x}^k\|_{1,\varepsilon} + L_2(\mathbf{x}^{k+1}-\mathbf{x}^k) - \mathbf{A}^T\mathbf{w}^{k+1}\\ \mathbf{w}^{k+1} = \mathbf{w}^k - \rho(\mathbf{A}\mathbf{x}^{k+1}-\mathbf{y}-\mathbf{v}^{k+1})\end{cases}. \tag{49}$$

For a convergent subsequence $\mathbf{z}^{k_j}$, since $\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_2 \to 0$ as $k\to\infty$, $\mathbf{z}^{k_j}$ and $\mathbf{z}^{k_j+1}$ have the same limit point $\mathbf{z}^* := (\mathbf{v}^*,\mathbf{x}^*,\mathbf{w}^*)$. Moreover, since $\hat{\mathcal{L}}(\hat{\mathbf{z}}^k)$ is convergent, $\|\mathbf{v}^k\|_p^p$ is also convergent. Then, passing to the limit in (49) along the subsequence $\mathbf{z}^{k_j}$ yields

$$-\mathbf{w}^* \in \frac{1}{\mu}\partial\|\mathbf{v}^*\|_p^p,\ \mathbf{A}^T\mathbf{w}^* = \nabla\|\mathbf{x}^*\|_{1,\varepsilon},\ \mathbf{A}\mathbf{x}^*-\mathbf{y} = \mathbf{v}^*.$$

In particular, $\mathbf{z}^*$ is a stationary point of $\mathcal{L}_{\rho,\varepsilon}$.

Proof of Theorem 3: Based on the above lemmas, the proof of Theorem 3 mainly consists of the following two steps:

i) There exists $c_2 > 0$ such that

$$\begin{aligned}&\mathrm{dist}(\mathbf{0},\partial\hat{\mathcal{L}}(\hat{\mathbf{z}}^{k+1}))\\ &\le c_2\big(\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_2 + \|\mathbf{x}^{k}-\mathbf{x}^{k-1}\|_2 + \|\mathbf{x}^{k-1}-\mathbf{x}^{k-2}\|_2\big)\end{aligned}$$

which together with Lemma 3 implies that $\mathrm{dist}(0,\partial\hat{\mathcal{L}}(\hat{\mathbf{z}}^{k+1})) \to 0$ as $k\to\infty$.

ii) Let $\mathbf{z}^k := (\mathbf{v}^k,\mathbf{x}^k,\mathbf{w}^k)$, the generated sequence $\{\mathbf{z}^k\}$ has finite length, i.e.,

$$\sum_{k=0}^{\infty}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_2 < \infty$$

which implies that $\{\mathbf{z}^k\}$ is a *Cauchy sequence* and thus is convergent.

The property (ii) together with Lemma 3 completes the proof. The derivation of property (ii) relies heavily on the Kurdyka-Lojasiewicz (KL) property of $\hat{\mathcal{L}}$, which holds if p is rational since $\|\cdot\|_p$ is semi-algebraic (thus a KL function) in this case. With the above lemmas, the proof of (i) and (ii) follows similarly the proof of Theorem III.3 and Theorem III.4 in [53] with some minor changes, thus is omitted here for succinctness.

References

[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[2] E. J. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[3] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[4] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[5] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.

[6] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[7] D. L. Donoho, "Neighborly polytopes and the sparse solution of underdetermined systems of linear equations," Statist. Dept., Stanford Univ., Stanford, CA, USA, Tech. Rep. 2005–4, 2005.

[8] D. L. Donoho, "High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension," *Discrete Comput. Geom.*, vol. 35, no. 4, pp. 617–652, 2006.

[9] F. Wen, P. Liu. Y. Liu, R. C. Qiu, and W. Yu, "Robust sparse recovery for compressive sensing in impulsive noise using Lp-norm model fitting," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4643–4647.

[10] P. J. Huber, *Robust Statistics*. New York, NY, USA: Wiley, 1981.

[11] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[12] E. J. Candès and P. A. Randall, "Highly robust error correction by convex programming," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 2829–2840, Jul. 2008.

[13] B. Popilka, S. Setzer, and G. Steidl, "Signal recovery from incomplete measurements in the presence of outliers," *Inverse Problems Imag.*, vol. 1, no. 4, pp. 661–672, Nov. 2007.

[14] R. Chan, C.-W. Ho, and M. Nikolova, "Salt-and-pepper noise removal by median-type noise detectors and de tail-preserving regularization," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1479–1485, Oct. 2005.

[15] T. Hashimoto, "Bounds on a probability for the heavy tailed distribution and the probability of deficient decoding in sequential decoding," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 990–1002, Mar. 2005.

[16] L. Bar, A. Brook, N. Sochen, and N. Kiryati, "Deblurring of color images corrupted by impulsive noise," *IEEE Trans. Image Process.*, vol. 16, no. 4, pp. 1101–1111, Apr. 2007.

[17] P. Civicioglu, "Using uncorrupted neighborhoods of the pixels for impulsive noise suppression with ANFIS," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 759–773, Mar. 2007.

[18] P. Windyga, "Fast impulsive noise removal," *IEEE Trans. Image Process.*, vol. 10, no. 1, pp. 173–179, Jan. 2001.

[19] G. R. Arce, *Nonlinear Signal Processing: A Statistical Approach*. New York, NY, USA: Wiley, 2005.

[20] R. E. Carrillo, K. E. Barner, and T. C. Aysal, "Robust sampling and reconstruction methods for sparse signals in the presence of impulsive noise," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 392–408, Apr. 2010.

[21] R. E. Carrillo, T. C. Aysal, and K. E. Barner, "A generalized Cauchy distribution framework for problems requiring robust behavior," *EURASIP J. Adv. Signal Process.*, vol. 2010, no. 1, pp. 1–19, Aug. 2010.

[22] R. E. Carrillo and K. E. Barner, "Lorentzian iterative hard thresholding: Robust compressed sensing with prior information," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4822–4833, Oct. 2013.

[23] J. F. Yang and Y. Zhang, "Alternating direction algorithms for ℓ_1-problems in compressive sensing," *SIAM J. Sci. Comput.*, vol. 33, pp. 250–278, 2011.

[24] Y. Xiao, H. Zhu, and S.-Y. Wu, "Primal and dual alternating direction algorithms for ℓ_1-ℓ_1-norm minimization problems in compressive sensing," *Comput. Optim. Appl.*, vol. 54, no. 2, pp. 441–459, 2013.

[25] S. Cao, Y. Xiao, and H. Zhu, "Linearized alternating directions method for ℓ_1-norm inequality constrained ℓ_1-norm minimization," *Appl. Numer. Math.*, vol. 85, pp. 142–153, 2014.

[26] D. S. Pham and S. Venkatesh, "Improved image recovery from compressed data contaminated with impulsive noise," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 397–405, Jan. 2012.

[27] D. S. Pham and S. Venkatesh, "Efficient algorithms for robust recovery of images from compressed data," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4724–4737, Dec. 2013.

[28] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 372–386, Feb. 2012.

[29] X. Jiang, T. Kirubarajan, and W.-J. Zeng, "Robust sparse channel estimation and equalization in impulsive noise using linear programming," *Signal Process.*, vol. 93, no. 5, pp. 1095–1105, 2013.

[30] J. N. Laska, M. A. Davenport, and R. G. Baraniuk, "Exact signal recovery from sparsely corrupted measurements through the pursuit of justice," in *Proc. 43rd Asilomar Conf. Signals, Syst., Comput.*, 2009, pp. 1556–1560.

[31] C. Studer, P. Kuppinger, G. Pope, and H. Bolcskei, "Recovery of sparsely corrupted signals," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3115–3130, May 2012.

[32] Y. Ji, Z. Yang, and W. Li, "Bayesian sparse reconstruction method of compressed sensing in the presence of impulsive noise," *Circuits, Syst., Signal Process.*, vol. 32, no. 6, pp. 2971–2998, 2013.

[33] J. N. Laska, P. T. Boufounos, M. A. Davenport, and R. G. Baraniuk, "Democracy in action: Quantization, saturation, and compressive sensing," *Appl. Comput. Harmonic Anal.*, 2011, pp. 429–443.

[34] P. Tsakalides and C. L. Nikias, "Robust adaptive beamforming in alpha-stable noise environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1996, pp. 2884–2887.

[35] X. Jiang, W.-J. Zeng, A. Yasotharan, H. C. So, and T. Kirubarajan, "Minimum dispersion beamforming for non-Gaussian signals," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1879–1893, Apr. 2014.

[36] W.-J. Zeng, H. C. So, and L. Huang, "ℓ_p-MUSIC: Robust direction-of-arrival estimator for impulsive noise environments," *IEEE Trans. Signal Process.*, vol. 61, no. 17, pp. 4296–4308, Sep. 2013.

[37] W.-J. Zeng, H.-C. So, and A. M. Zoubir, "An ℓ_p-norm minimization approach to time delay estimation in impulsive noise," *Digit. Signal Process.*, vol. 23, no. 4, pp. 1247–1254, 2013.

[38] F. Moghimi, A. Nasri, and R. Schober, "Lp-norm spectrum sensing for cognitive radio networks impaired by non-Gaussian noise," in *Proc. GLOBECOM*, Honolulu, HI, USA, Nov. 30–Dec. 4, 2009, pp. 1–6.

[39] G. Pope, C. Studer, and M. Baes, "Coherence-based recovery guarantees for generalized basis-pursuit de-quantizing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 3669–3672.

[40] F. Alizadeh and D. Goldfarb, "Second-order cone programming," *Math. Program.*, vol. 95, no. 1, pp. 3–51, 2003.

[41] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9, pp. 589–592, 2008.

[42] C. L. Nikias and M. Shao, *Signal Processing With Alpha-Stable Distributions and Applications*. New York, NY, USA: Wiley, 1995.

[43] P. L. Combettes and J. C. Pesquet, "Proximal splitting methods in signal processing," *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. New York, NY, USA: Springer, 2011, pp. 185–212.

[44] G. Marjanovic and V. Solo, "On ℓ_q optimization and matrix completion," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5714–5724, Nov. 2012.

[45] G. Marjanovic and V. Solo, "ℓ_q matrix completion," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, 2012, pp. 3885–3888.

[46] Z. Xu, X. Chang, F. Xu, and H. Zhang, "L1/2 regularization: A thresholding representation theory and a fast solver," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1013–1027, Jul. 2012.

[47] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[48] 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.*, vol. 3, no. 1, pp. 1–122, 2011.

[49] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, pp. 293–318, 1992.

[50] M. K. Varanasi and B. Aazhang, "Parametric generalized Gaussian density estimation," *J. Acoust. Soc. Amer.*, vol. 86, no. 4, pp. 1404–1415, Oct. 1989.

[51] M. Hong, Z. Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, 2016.

[52] G. Li and T. K. Pong, "Global convergence of splitting methods for nonconvex composite optimization," *SIAM J. Optim.*, vol. 25, no. 4, pp. 2434–2460, Jul. 2015.

[53] F. Wang, Z. Xu, and H.-K. Xu, "Convergence of Bregman alternating direction method with multipliers for nonconvex composite problems," *arXiv preprint*, arXiv:1410.8625, Dec. 2014.

[54] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk, "Fast alternating direction optimization methods," *SIAM J. Imag. Sci.*, vol. 7, no. 3, pp. 1588–1623, 2014.

[55] Y. Tsaig and D. L. Donoho, "Fast solution of ℓ_1-norm minimization problems when the solution maybe sparse," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

[56] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, Jun. 2008.

[57] S. Foucart and M.-J. Lai, "Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 395–407, May 2009.

[58] C. Studer and R. G. Baraniuk, "Stable restoration and separation of approximately sparse signals," *Appl. Comput. Harmon. Anal.*, vol. 37, no. 1, pp. 12–35, 2014.

[59] J. Shang, Z. Wang, and Q. Huang, "A robust algorithm for joint sparse recovery in presence of impulsive noise," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1166–1170, Aug. 2015.

[60] Y. Xiao, T. Zeng, J. Yu, and M. K. Ng, "Restoration of images corrupted by mixed Gaussian-impulse noise via $\ell_1 - \ell_0$ minimization," *Pattern Recognit.*, vol. 44, no. 8, pp. 1708–1720, 2011.

[61] M. Yan, "Restoration of images corrupted by impulse noise and mixed Gaussian impulse noise using blind inpainting," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1227–1245, 2013.

[62] K. Hohm, M. Storath, and A. Weinmann, "An algorithmic framework for Mumford-Shah regularization of inverse problems in imaging," *Inverse Probl.*, vol. 31, no. 11, 2015, Art. no 115011.

[63] G. Yuan and B. Ghanem, "ℓ_0TV: A new method for image restoration in the presence of impulse noise," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5369–5377.

Fei Wen (M'15) received the B.S. degree from the University of Electronic Science and Technology of China (UESTC), Sichuan, China, in 2006, and the Ph.D. degree in communications and information engineering from UESTC in 2013. Since December 2012, he has been a Lecturer at the Air Force Engineering University. He is currently a Research Fellow in the Department of Electronic Engineering, Shanghai Jiao Tong University. His main research interests include statistical signal processing and nonconvex optimization.

Peilin Liu (M'99) received the Ph.D. degree majoring in electronic engineering from the University of Tokyo, Tokyo, Japan, in 1998 and was a Research Fellow in 1999. From 1999 to 2003, she was a Senior Researcher at the Central Research Institute of Fujitsu, Tokyo. She is currently a Professor in the Department of Electronic Engineering, Shanghai Jiao Tong University, an Executive Director in Shanghai Key Laboratory of Navigation and location-based service, and is responsible for a series of important projects, such as BDSSoC platform development, low-power and high-performance communication DSP. Her research interests include signal processing, low-power computing architecture, and application-oriented SoC design and verification. He is the Chair of shanghai chapter of the IEEE CIRCUIT AND SYSTEM.

Yipeng Liu received the B.Sc. degree in biomedical engineering (BME) and the Ph.D. degree in information and communication engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in June 2006 and 2011, respectively. From July 2007 to September 2008, he was a Research Intern at the 10th Institute of China Electronics Technology Group Corporation (CETC-10), China. From November 2010 to May 2011, he was a Visiting Ph.D. student at Tsinghua University, Beijing, China. From June 2011 to November 2011, he was a Research Engineer at Huawei Technologies, China. From November 2011 to November 2014, he was a Postdoctoral Research Fellow at the University of Leuven, Belgium. Since December 2014, he has been an Associate Professor in the School of Electronic Engineering/Center for Robotics/Center for Information in Medicine, University of Electronic Science and Technology of China, Chengdu, China.

Robert C. Qiu (S'93–M'96–SM'01–F'15) received the Ph.D. degree in electrical engineering from New York University (formerly Poly-technic University, Brooklyn, NY), New York, NY, USA. He was with GTE Laboratories, Inc. (currently Verizon), Waltham, MA, USA, and Bell Labs, Alcatel-Lucent, Whippany, NJ, USA. He was the Founder, Chief Executive Officer, and the President of Wiscom Technologies, Inc., which manufactures and markets wideband code-division multiple-access chipsets. He has served as an Adjunct Professor with New York University. He is currently a Full Professor in the Department of Electrical and Computer Engineering, Center for Manufacturing Research, Tennessee Technological University, Cookeville, TN, USA, where he started as an Associate Professor in 2003, before becoming a Full Professor in 2008. He has 15 contributions to the Third-Generation Partnership Project and the IEEE standards bodies. He has co-authored the books entitled *Cognitive Radio Communication and Networking: Principles and Practice* (Hoboken, NJ, USA: Wiley, 2012) and *Cognitive Networked Sensing: A Big Data Way* (New York, NY, USA: Springer, 2013), and authored a book entitled *Introduction to Smart Grid* (Hoboken, NJ, USA: Wiley, 2014). He has authored more than 70 journal papers/book chapters, and holds more than six patents. His current research interests include wireless communication and networking, machine learning, and smart grid technologies. He is a Guest Book Editor of *Ultrawideband Wireless Communications* (New York, NY, USA: Wiley, 2005). He is an Associate Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, other international journals, and three special issues on UWB, including the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE TRANSACTIONS ON SMART GRID. He is as a Member of the Technical Program Committee of the IEEE Global Communications Conference, the IEEE International Conference on Communications, the IEEE Wireless Communications and Networking Conference, the Military Communications Conference, and the International Conference on Ultrawideband.

Wenxian Yu received the B.Sc., M.Sc., and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 1985, 1988, and 1993, respectively. From 1996 to 2008, he was a Professor in the College of Electronic Science and Engineering, National University of Defense Technology, where he was the Deputy Head of the College and an Assistant Director of the National Key Laboratory of Automatic Target Recognition. He was the Executive Dean of the School of Electronic, Information, and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, from 2009 to 2011. He is currently a Yangtze River Scholar Distinguished Professor and the Head of the research part of the School of Electronic, Information, and Electrical Engineering, Shanghai Jiao Tong University, Shanghai.