SIAM J. SCI. COMPUT.
Vol. 40, No. 4, pp. A2062–A2086

ROBUST DECODING FROM 1-BIT COMPRESSIVE SAMPLING WITH ORDINARY AND REGULARIZED LEAST SQUARES*

JIAN HUANG†, YULING JIAO‡, XILIANG LU§, AND LIPING ZHU¶

Abstract. In 1-bit compressive sensing (1-bit CS) where a target signal is coded into a binary measurement, one goal is to recover the signal from noisy and quantized samples. Mathematically, the 1-bit CS model reads $y = \eta \odot \text{sign}(\Psi x^* + \epsilon)$, where $x^* \in \mathcal{R}^n, y \in \mathcal{R}^m$, $\Psi \in \mathcal{R}^{m\times n}$, and ϵ is the random error before quantization and $\eta \in \mathcal{R}^n$ is a random vector modeling the sign flips. Due to the presence of nonlinearity, noise, and sign flips, it is quite challenging to decode from the 1-bit CS. In this paper, we consider a least squares approach under the overdetermined and underdetermined settings. For $m > n$, we show that, up to a constant c, with high probability, the least squares solution x_{ls} approximates x^* with precision δ as long as $m \geq \widetilde{\mathcal{O}}(\frac{n}{\delta^2})$. For $m < n$, we prove that, up to a constant c, with high probability, the ℓ_1-regularized least-squares solution x_{ℓ_1} lies in the ball with center x^* and radius δ provided that $m \geq \mathcal{O}(\frac{s\log n}{\delta^2})$ and $\|x^*\|_0 := s < m$. We introduce a Newton type method, the so-called primal and dual active set (PDAS) algorithm, to solve the nonsmooth optimization problem. The PDAS possesses the property of one-step convergence. It only requires solving a small least squares problem on the active set. Therefore, the PDAS is extremely efficient for recovering sparse signals through continuation. We propose a novel regularization parameter selection rule which does not introduce any extra computational overhead. Extensive numerical experiments are presented to illustrate the robustness of our proposed model and the efficiency of our algorithm.

Key words. 1-bit compressive sensing, ℓ_1-regularized least squares, primal dual active set algorithm, one-step convergence, continuation

AMS subject classifications. 49N45, 65N21

DOI. 10.1137/17M1154102

1. Introduction.

Compressive sensing (CS) is an important approach to acquiring low dimension signals from noisy underdetermined measurements [8, 17, 20, 21]. For storage and transmission, the infinite-precision measurements are often quantized; [6] considered recovering the signals from the 1-bit compressive sensing (1-bit CS) where measurements are coded into a single bit, i.e., their signs. The 1-bit CS is superior to the CS in terms of inexpensive hardware implementation and storage. However, it is much more challenging to decode from nonlinear, noisy, and sign-flipped 1-bit measurements.

*Submitted to the journal's Methods and Algorithms for Scientific Computing section October 26, 2017; accepted for publication (in revised form) May 1, 2018; published electronically July 3, 2018.

http://www.siam.org/journals/sisc/40-4/M115410.html

Funding: The second author's work was supported by National Science Foundation of China (NSFC) 11501579 and National Science Foundation of Hubei Province 2016CFB486. The third author's work was supported by NSFC 11471253 and 91630313, and the fourth author's work was supported by NSFC 11731011 and Chinese Ministry of Education Project of Key Research Institute of Humanities and Social Sciences at Universities 16JJD910002 and National Youth Top-notch Talent Support Program of China.

†Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong 999077, People's Republic of China (j.huang@polyu.edu.hk).

‡Corresponding author. School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430063, People's Republic of China (yulingjiaomath@whu.edu.cn).

§School of Mathematics and Statistics, and Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan 430072, People's Republic of China (xllv.math@whu.edu.cn).

¶Corresponding author. Institute of Statistics and Big Data and Center for Applied Statistics, Renmin University of China, Beijing 100872, People's Republic of China (zhu.liping@ruc.edu.cn).

1.1. Previous work. Since the seminal work of [6], much effort has been devoted to studying the theoretical and computational challenges of the 1-bit CS. Sample complexity was analyzed for support and vector recovery with and without noise [23, 30, 42, 25, 31, 24, 25, 43, 52]. Existing works indicate that $m > \mathcal{O}(s \log n)$ is adequate for both support and vector recovery. The sample size required here has the same order as that required in the standard CS setting. These results have also been refined by adaptive sampling [24, 14, 4]. Extensions include recovering the norm of the target [34, 3] and non-Gaussian measurement settings [1]. Many first order methods [6, 36, 51, 14] and greedy methods [37, 5, 31] are developed to minimize the sparsity promoting nonconvex objected function arising from either the unit sphere constraint or the nonconvex regularizers. To address the nonconvex optimization problem, convex relaxation models are also proposed [52, 43, 42, 53, 44], which often yield accurate solutions efficiently with polynomial-time solvers. See, for example, [40].

1.2. 1-bit CS setting. In this paper we consider the following 1-bit CS model:

$$y = \eta \odot \operatorname{sign}(\Psi x^* + \epsilon), \tag{1.1}$$

where $y \in \mathcal{R}^m$ is the 1-bit measurement, $x^* \in \mathcal{R}^n$ is an unknown signal, $\Psi = [\psi_1^t; \ldots; \psi_m^t] \in \mathcal{R}^{m \times n}$ is a random matrix, $\eta \in \mathcal{R}^m$ is a random vector modeling the sign flips of y, and $\epsilon \in \mathcal{R}^n$ is a random vector with independent and identically distributed (iid) entries modeling errors before quantization. Throughout $\operatorname{sign}(\cdot)$ operates componentwise with $\operatorname{sign}(z) = 1$ if $z \geq 0$ and $\operatorname{sign}(z) = -1$ otherwise, and $\odot$ is the pointwise Hardmard product. Following [42] we assume that the rows of Ψ are iid random vectors sampled from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$ with an unknown covariance matrix Σ, ϵ is distributed as $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_m)$ with an unknown noise level σ, and $\eta \in \mathcal{R}^m$ has independent coordinates $\eta_i s$ satisfying $\mathbb{P}[\eta_i = 1] = 1 - \mathbb{P}[\eta_i = -1] = q \neq \frac{1}{2}$. We assume η_i, ϵ_i, and ψ_i are mutually independent. Because σ is known model (1.1) is invariant in the sense that $\forall \alpha > 0$, $y = \eta \odot \operatorname{sign}(\Psi x^* + \epsilon) = \eta \odot \operatorname{sign}(\alpha \Psi x^* + \alpha \epsilon)$. This indicates that the best one can hope for is to recover x^* up to a scale factor. Without loss of generality we assume $\|x^*\|_\Sigma = 1$, where $\|x\|_\Sigma = (x^t \Sigma x)^{\frac{1}{2}}$ is the elliptic norm of x with respect to Σ.

1.3. Contributions. We study the 1-bit CS problem in both the overdetermined setting with $m > n$ and the underdetermined setting with $m < n$. In the former setting we allow for dense x^*, while in the latter, we assume that x^* is sparse in the sense that $\|x^*\|_0 = s < m$. The basic message is that we can recover x^* with the ordinary least squares or the ℓ_1 regularized least squares.

(1) When $m > n$, we propose to use the least squares solution

$$x_{\text{ls}} \in \arg\min \frac{1}{m} \sum_{i=1}^{m} (y_i - \psi_i^t x)^2$$

to approximate x^*. We show that, with high probability, x_{ls} estimates x^* accurately up to a positive scale factor c defined by (2.2) in the sense that, $\forall \delta \in (0,1)$, $\|x_{\text{ls}}/c - x^*\| \leq \delta$ if $m \geq \widetilde{\mathcal{O}}(\frac{n}{\delta^2})$. We make the following observation:

Up to a constant c, the underlying target x^* can be decoded from 1-bit measurements with the ordinary least squares, as long as the probability of sign flips is not equal to $1/2$.

(2) When $m < n$ and the target signal x^* is sparse, we consider the ℓ_1-regularized least squares solution

$$x_{\ell_1} \in \arg\min \frac{1}{2m}\|y - \Psi x\|_2^2 + \lambda\|x\|_1. \tag{1.2}$$

The sparsity assumption is widely used in modern signal processing [21, 38]. We show that with high probability the error $\|x_{\ell_1}/c - x^*\|$ can be bounded by a prefixed accuracy $\delta \in (0,1)$ if $m \geq \mathcal{O}(\frac{s\log n}{\delta^2})$, which is the same as the order for the standard CS methods to work. Furthermore, the support of x^* can be exactly recovered if the minimum signal magnitude of x^* is larger than $\mathcal{O}(\sqrt{s\log n/m})$. When the target signal is sparse, we obtain the following conclusion:

> *Up to a constant c, the sparse signal x^* can also be decoded from 1-bit measurements with the ℓ_1-regularized least squares, as long as the probability of sign flips is not equal to* 1/2.

(3) We introduce a fast and accurate Newton method, the so-called primal dual active set method (PDAS), to solve the ℓ_1-regularized minimization (1.2). The PDAS possesses the property of one-step convergence. The PDAS solves a small least squares problem on the active set and is thus extremely efficient if combined with continuation. We propose a novel regularization parameter selection rule, which is incorporated with a continuation procedure without additional cost. The code is available at http://faculty.zuel.edu.cn/tjyjxxy/jyl/list.htm.

> *The optimal solution x_{ℓ_1} can be computed efficiently and accurately with the PDAS, a Newton type method which converges after one iteration, even if the objective function* (1.2) *is nonsmooth. Continuation on λ globalizes the PDAS. The regularization parameter can be automatically determined without additional computational cost.*

1.4. Notation and organization. Throughout we denote by $\Psi_i \in \mathcal{R}^{m\times 1}, i = 1,\ldots,m$, and $\psi_j \in \mathcal{R}^{n\times 1}, j = 1,\ldots,n$, the ith column and jth row of Ψ, respectively. We denote a vector of 0 by $\mathbf{0}$, whose length may vary in different places. We use $[n]$ to denote the set $\{1,\ldots,n\}$ and $\mathbf{I}_n$ to denote the identity matrix of size $n \times n$. For $A, B \subseteq [n]$ with length $|A|, |B|$, $x_A = (x_i, i \in A) \in \mathbb{R}^{|A|}$, $\Psi_A = (\Psi_i, i \in A) \in \mathbb{R}^{m\times|A|}$ and $\Psi_{AB} \in \mathbb{R}^{|A|\times|B|}$ denotes a submatrix of Ψ whose rows and columns are listed in A and B, respectively. We use $(\psi_i)_j$ to denote the jth entry of the vector ψ_i and $|x|_{\min}$ to denote the minimum absolute value of x. We use $\mathcal{N}(\mathbf{0}, \Sigma)$ to denote the multivariate normal distribution, with Σ symmetric and positive definite. Let $\gamma_{\max}(\Sigma)$ and $\gamma_{\min}(\Sigma)$ be the largest and the smallest eigenvalues of Σ, respectively, and $\kappa(\Sigma)$ be the condition number $\gamma_{\max}(\Sigma)/\gamma_{\min}(\Sigma)$ of Σ. Let $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}, p \in [1,\infty]$, be the ℓ_p-norm of x. We denote the number of nonzero elements of x by $\|x\|_0$ and let $s = \|x^*\|_0$. The symbols $\|\Psi\|$ and $\|\Psi\|_\infty$ stand for the operator norm of Ψ induced by ℓ_2 norm and the maximum pointwise absolute value of Ψ, respectively. We use $\mathbb{E}[\cdot]$, $\mathbb{E}[\cdot|\cdot]$, $\mathbb{P}[\cdot]$ to denote the expectation, the conditional expectation, and the probability on a given probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. We use C_1 and C_2 to denote generic constants which may vary from place to place. By $\mathcal{O}(\cdot)$ and $\widetilde{\mathcal{O}}(\cdot)$, we ignore some positive numerical constant and $\sqrt{\log n}$, respectively.

The rest of the paper is organized as follows. In section 2 we explain why the least squares works in the 1-bit CS when $m > n$ and obtain a nonasymptotic error bound

for $\|x_{\rm ls}/c - x^*\|$. In section 3 we consider the sparse decoding when $m < n$ and prove a minimax bound on $\|x_{\ell_1}/c - x^*\|$. In section 4 we introduce the PDAS algorithm to solve (1.2). We propose a new regularization parameter selection rule combined with the continuation procedure. In section 5 we conduct simulation studies and compare with existing 1-bit CS methods. We conclude with some remarks in section 6.

2. Least squares when $m > n$. In this section, we first explain why the least squares works in the overdetermined 1-bit CS model (1.1) with $m > n$. We then prove a nonasymptotic error bound on $\|x_{\rm ls}/c - x^*\|$. The following lemma inspired by [7] is our starting point.

LEMMA 2.1. *Let $\tilde{y} = \tilde{\eta} sign(\tilde{\psi}^t x^* + \tilde{\epsilon})$ be the 1-bit model (1.1) at the population level, i.e., the observation data $y_i = \eta_i sign(\psi_i^t x^* + \epsilon_i), i = 1, \ldots, m$, is generated accordingly, where η_i, ψ_i, ϵ_i are random variables with the same distribution as that of $\tilde{\eta}$, $\tilde{\psi}$, $\tilde{\epsilon}$, respectively. Let $\mathbb{P}[\tilde{\eta} = 1] = q \neq \frac{1}{2}$, $\tilde{\psi} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$, $\tilde{\epsilon} \sim \mathcal{N}(0, \sigma^2)$. It follows that*

$$\Sigma^{-1}\mathbb{E}[\tilde{y}\tilde{\psi}]/c = x^*, \tag{2.1}$$

where

$$c = (2q-1)\sqrt{\frac{2}{\pi(\sigma^2+1)}}. \tag{2.2}$$

Proof. The proof is given in Appendix A. ∎

Lemma 2.1 shows that the target x^* is proportional to $\Sigma^{-1}\mathbb{E}[\tilde{y}\tilde{\psi}]$. Note that

$$\mathbb{E}[\Psi^t\Psi/m] = \mathbb{E}\left[\sum_{i=1}^m \psi_i\psi_i^t\right]/m = \Sigma \quad \textit{and} \tag{2.3}$$

$$\mathbb{E}[\Psi^t y/m] = \mathbb{E}\left[\sum_{i=1}^m \psi_i y_i\right]/m = \mathbb{E}[\tilde{y}\tilde{\psi}]. \tag{2.4}$$

As long as $\Psi^t\Psi/m$ is invertible, it is reasonable to expect that

$$x_{\rm ls} = (\Psi^t\Psi/m)^{-1}(\Psi^t y/m) = (\Psi^t\Psi)^{-1}(\Psi^t y)$$

can approximate x^* well up to a constant c even if y consists of sign flips.

THEOREM 2.2. *Consider the ordinary least squares:*

$$x_{ls} \in \arg\min_x \frac{1}{m}\|\Psi x - y\|_2^2. \tag{2.5}$$

If $m \geq 16C_2^2 n$, then with probability at least $1 - 4\exp(-C_1C_2^2 n) - \frac{2}{n^3}$,

$$\|x_{ls}/c - x^*\|_2 \leq \sqrt{\frac{n}{m}}\left(4C_2\sqrt{\kappa(\Sigma)\gamma_{\max}(\Sigma)} + \frac{6(\sigma+1)}{\sqrt{C_1}|2q-1|}\sqrt{\log n}\right), \tag{2.6}$$

where C_1 and C_2 are some generic constants not depending on m or n.

Proof. The proof is given in Appendix C. ∎

Remark 2.1. Theorem (2.2) shows that $\forall \delta \in (0,1)$ if $m \geq \widetilde{\mathcal{O}}(\frac{n}{\delta^2})$, up to a constant, the simple least squares solution can approximate x^* with error of order δ even if y contains very large quantization error and sign flips with probability unequal to $1/2$.

Remark 2.2. To the best of our knowledge, this is the first nonasymptotic error bound for the 1-bit CS in the overdetermined setting. Comparing with the estimation error of the standard least squares in the complete data model $y = \Psi x^* + \epsilon$, the error bound in Theorem 2.2 is optimal up to a logarithm factor $\sqrt{\log n}$, which is due to the loss of information with the 1-bit quantization.

3. Sparse decoding with ℓ_1-regularized least squares.

3.1. Nonasymptotic error bound. Since images and signals are often sparsely represented under certain transforms [38, 16], it suffices for the standard CS to recover the sparse signal $x^* \in \mathcal{R}^n$ with $m = \mathcal{O}(s \log n)$ measurements for $s = \|x^*\|_0$. In this section we show that in the 1-bit CS setting, similar results can be derived through the ℓ_1-regularized least squares (1.2). Model (1.2) has been extensively studied when y is continuous [46, 9, 8, 17]. Here we use model (1.2) to recover x^* from quantized y, which is rarely studied in the literature.

Next we show that, up to the constant c, x_{ℓ_1} is a good estimate of x^* when $m = \mathcal{O}(s \log n)$ even if the signal is highly noisy and corrupted by sign flips in the 1-bit CS setting.

THEOREM 3.1. *Assume* $n > m \geq \max\{\frac{4C_1}{C_2^2} \log n, \frac{64(4\kappa(\Sigma)+1)^2}{C_1} s \log \frac{en}{s}\}$, $s \leq \exp^{(1-\frac{C_1}{2})} n$. *Set* $\lambda = \frac{4(1+C_3|c|)}{\sqrt{C_1}} \sqrt{\frac{\log n}{m}}$. *Then with probability at least* $1 - 2/n^3 - 6/n^2$, *we have*

$$\|x_{\ell_1}/c - x^*\|_2 \leq \frac{816(4\kappa(\Sigma)+1)^2}{\gamma_{min}(\Sigma)} \frac{\sigma + 1 + C_3|q - 1/2|}{\sqrt{C_1}|q - 1/2|} \sqrt{\frac{s \log n}{m}}. \tag{3.1}$$

Proof. The proof is given in Appendix D. ∎

Remark 3.1. Theorem 3.1 shows that $\forall \delta \in (0, 1)$, if $m \geq \mathcal{O}(\frac{s \log n}{\delta^2})$, up to a constant c, the ℓ_1-regularized least squares solution can approximate x^* with precision δ.

Remark 3.2. The error bound in Theorem 3.1 achieves the minimax optimal order $\mathcal{O}(\sqrt{\frac{s \log n}{m}})$ in the sense that it is the optimal order that can be attained even if the signal is measured precisely without 1-bit quantization [39]. From Theorem 3.1 if the minimum nonzero magnitude of x^* is large enough, i.e., $|x^*|_{\min} \geq \mathcal{O}(\sqrt{\frac{s \log n}{m}})$, the support of x_{ℓ_1} coincides with that of x^*.

3.2. Comparison with related works. Assuming $\|x^*\|_2 = 1$ and $\sigma = 0$ and $q = 1$, [6] proposed to decode x^* with

$$\min_{x \in \mathcal{R}^n} \|x\|_1 \quad s.t. \quad y \odot \Psi x \geq 0, \quad \|x\|_2 = 1.$$

A first order algorithm was devised to solve the following Lagrangian version [36], i.e.,

$$\min_{x \in \mathcal{R}^n} \|\max\{\mathbf{0}, -y \odot \Psi x\}\|_2^2 + \lambda \|x\|_1 \quad s.t. \quad \|x\|_2 = 1.$$

In the presence of noise, [31] introduced

$$\min_{x \in \mathcal{R}^n} \mathcal{L}(\max\{\mathbf{0}, -y \odot \Psi x\}) \quad s.t. \quad \|x\|_0 \leq s, \quad \|x\|_2 = 1, \tag{3.2}$$

where $\mathcal{L}(\cdot) = \|\cdot\|_1$ or $\|\cdot\|_2^2$. They used a projected subgradient method, the so-called binary iterative hard thresholding (BITH), to solve (3.2). Assuming that there are sign flips in the noiseless model with $\sigma = 0$, [14] considered

$$\min_{x \in \mathcal{R}^n} \lambda \|\max\{\mathbf{0}, \nu \mathbf{1} - y \odot \Psi x\}\|_0 + \frac{\beta}{2} \|x\|_2^2 \quad s.t. \quad \|x\|_0 \leq s, \tag{3.3}$$

where $\nu > 0, \beta > 0$ are tuning parameters. An adaptive outlier pursuit (AOP) generalization of (3.2) was proposed in [51] to recover x^* and simultaneously detect the entries with sign flips by

$$\min_{x\in\mathcal{R}^n,\Lambda\in\mathcal{R}^m} \mathcal{L}(\max\{\mathbf{0}, -\Lambda\odot y\odot\Psi x\})$$
$$s.t. \quad \Lambda_i\in\{0,1\}, \quad \|\mathbf{1}-\Lambda\|_1\le N, \quad \|x\|_0\le s, \quad \|x\|_2=1,$$

where N is the number of sign flips. Alternating minimization on x and Λ is adopted to solve the optimization problem. Huang et al. [26] considered the pinball loss to deal with both the noise and the sign flips, which reads

$$\min_{x\in\mathcal{R}^n} \mathcal{L}_\tau(\nu\mathbf{1}-y\odot\Psi x\}) \quad s.t. \quad \|x\|_0\le s \quad \|x\|_2=1,$$

where $\mathcal{L}_\tau(t)=t\mathbf{1}_{t\ge 0}-\tau t\mathbf{1}_{t<0}$. Similar to the BITH, the pinball iterative hard thresholding is utilized. With the binary stable embedding, [31] and [14] proved that with high probability, the sample complexity of (3.2) and (3.3) to guarantee estimation error smaller than δ is $m\ge\mathcal{O}(\frac{s}{\delta^2}\log\frac{n}{s})$, which echoes Theorem 3.1. However, there are no theoretical results for other models mentioned above. All the aforementioned models and algorithms are the state-of-the-art works in the 1-bit CS. However, all the methods mentioned above are nonconvex. It is thus hard to justify whether the corresponding algorithms are loyal to their models.

Another line of research concerns convexification. The pioneering work is [42], where they considered the noiseless case without sign flips and proposed the following linear programming method:

$$x_{lp}\in\arg\min_{x\in\mathcal{R}^n}\|x\|_1 \quad s.t. \quad y\odot\Psi x\ge 0, \quad \|\Psi x\|_1=m.$$

As shown in [42], the estimation error is $\|\frac{x_{lp}}{\|x_{lp}\|}-x^*\|\le\mathcal{O}((s\log^2\frac{n}{s})^{\frac{1}{5}})$. The above result is improved to $\|\frac{x_{cv}}{\|x_{cv}\|}-x^*\|\le\mathcal{O}((s\log\frac{n}{s})^{\frac{1}{4}})$ in [43], where both the noise and the sign flips are allowed, through considering the convex problem

$$x_{cv}\in\arg\min_{x\in\mathcal{R}^n} -\langle y,\Psi x\rangle/m \quad s.t. \quad \|x\|_1\le s, \quad \|x\|_2\le 1. \tag{3.4}$$

Comparing with our result in Theorem 3.1, the results derived in [42] and [43] are suboptimal.

In the noiseless case and assuming $\Sigma=\mathbf{I}_n$, [52] considered the Lagrangian version of (3.4)

$$\min_{x\in\mathcal{R}^n} -\langle y,\Psi x\rangle/m+\lambda\|x\|_1 \quad s.t. \quad \|x\|_2\le 1. \tag{3.5}$$

In this special case, the estimation error derived in [52] improved that of [43] and matched our results in Theorem 3.1. However, [52] did not discuss the scenario of $\Sigma\ne\mathbf{I}_n$. Recently [44, 49], proposed a simple projected linear estimator $P_K(\Psi^t y/m)$, where $K=\{x|\|x\|_1\le s,\|x\|_2\le 1\}$, to estimate the low-dimensional structure target belonging to K in high dimensions from noisy and possibly nonlinear observations. They derived the same order of estimation error as that in Theorem 3.1.

Zymnis, Boyd, and Candes [53] proposed an ℓ_1 regularized maximum likelihood estimate, and [26] introduced a convex model through replacing the linear loss in (3.5) with the pinball loss. However, neither studied sample complexity or estimation error.

4. PDAS algorithm. Existing algorithms for (1.2) are mainly first order methods [47, 2, 12]. There are efforts in applying Newton type methods for solving (1.2) in compressed sensing [22, 15]. In this section we use the PDAS method [19, 32], which is a generalized Newton type method [29, 45] to solve (1.2). An important advantage of the PDAS is that it converges after one-step iteration if the initial value is good enough. We globalize it with continuation on regularization parameter. We also propose a novel regularization parameter selection rule which is incorporated along the continuation procedure without any additional computational burden.

4.1. PDAS. In this section we use x to denote x_{ℓ_1} for simplicity. We begin with the following results [13]. Let x be the minimizer of (1.2); then x satisfies

$$\begin{cases} d = \Psi^t(y - \Psi x)/m, \\ x = \mathcal{S}_\lambda(x + d). \end{cases} \tag{4.1}$$

Conversely, if x and d satisfy (4.1), then x is a global minimizer of (1.2), where $\mathcal{S}_\lambda(z)$ is the pointwise soft-thresholding operator defined by $\mathcal{S}_\lambda(z_i) = \arg\min_t \frac{1}{2}(t - z_i)^2 + \lambda|t|_1 = sign(z_i)\max\{|z_i| - \lambda, 0\}$.

Let $Z = \binom{x}{d}$ and $F(Z) = \binom{F_1(Z)}{F_2(Z)} : \mathcal{R}^n \times \mathcal{R}^n \to \mathcal{R}^{2n}$, where $F_1(Z) = x - \mathcal{S}_\lambda(x + d)$ and $F_2(Z) = \Psi^t\Psi x + md - \Psi^t y$. By (4.1), finding the minimizer x of (1.2) is equivalent to finding the root of $F(Z)$. We use the following PDAS method [19, 32] to find the root of $F(Z)$.

Remark 4.1. We can stop when k is greater than a user-predefined *MaxIter*. Since the PDAS converges after one iteration, a desirable property stated in Theorem 4.1, we set $MaxIter = 1$.

The PDAS is actually a generalized Newton method for finding roots of nonsmooth equations [29, 45], since the iteration in Algorithm 1 can be equivalently reformulated as

Algorithm 1. PDAS: $x_\lambda \leftarrow pdas(y, \Psi, \lambda, x^0, MaxIter)$.

1: Input y, Ψ, λ, initial guess x^0, maximum number of iteration *MaxIter*. Let $d^0 = \Psi^t(y - \Psi x^0)/m$.
2: **for** $k = 0, 1, \ldots MaxIter$ **do**
3: Compute the active and inactive sets $\mathcal{A}_k$ and $\mathcal{I}_k$ respectively by

$$\mathcal{A}_k = \{i \in [n] \big| |x_i^k + d_i^k| > \lambda\} \quad \text{and} \quad \mathcal{I}_k = \overline{\mathcal{A}_k}.$$

4: Update x^{k+1} and d^{k+1} by

$$\begin{cases} x_{\mathcal{I}_k}^{k+1} = \mathbf{0}, \\ d_{\mathcal{A}_k}^{k+1} = \lambda sign(x_{\mathcal{A}_k}^k + d_{\mathcal{A}_k}^k), \\ (\Psi_{\mathcal{A}_k}^t \Psi_{\mathcal{A}_k}) x_{\mathcal{A}_k}^{k+1} = \Psi_{\mathcal{A}_k}^t (y - m d_{\mathcal{A}_k}^{k+1}), \\ d_{\mathcal{I}_k}^{k+1} = \Psi_{\mathcal{I}_k}^t (y - \Psi_{\mathcal{A}_k} x_{\mathcal{A}_k}^{k+1})/m. \end{cases}$$

5: If $\mathcal{A}_k = \mathcal{A}_{k+1}$, stop.
6: **end for**
7: Output x^{k+1} and denote it as x_λ.

$$J^k D^k = -F(Z^k), \tag{4.2}$$

$$Z^{k+1} = Z^k + D^k, \tag{4.3}$$

where

$$J^k = \begin{pmatrix} J_1^k & J_2^k \\ \Psi^t\Psi & m\mathbf{I} \end{pmatrix}, \quad J_1^k = \begin{pmatrix} \mathbf{0}_{\mathcal{A}_k,\mathcal{A}_k} & \mathbf{0}_{\mathcal{A}_k,\mathcal{I}_k} \\ \mathbf{0}_{\mathcal{I}_k,\mathcal{A}_k} & \mathbf{I}_{\mathcal{I}_k,\mathcal{I}_k} \end{pmatrix}, \tag{4.4}$$

$$\text{and} \quad J_2^k = \begin{pmatrix} -I_{\mathcal{A}_k,\mathcal{A}_k} & \mathbf{0}_{\mathcal{A}_k,\mathcal{I}_k} \\ \mathbf{0}_{\mathcal{I}_k,\mathcal{A}_k} & \mathbf{0}_{\mathcal{I}_k,\mathcal{I}_k} \end{pmatrix}.$$

We prove this equivalency in Appendix E for completeness.

Local superlinear convergence has been established for generalized Newton methods for nonsmooth equations [29, 45]. The PDAS requires one iteration to convergence. We state the results here for completeness, which is proved in [19].

THEOREM 4.1. *Let x and d satisfy* (4.1). *Denote $\widetilde{\mathcal{A}} = \{i \in [n] \big| |x_i + d_i| \geq \lambda\}$ and $\omega = \min_{i\in[n],|x_i+d_i|\neq\lambda} \{||x_i + d_i| - \lambda|\}$. Let x^0 and $d^0 = \Psi^t(y - \Psi x^0)/m$ be the initial input of Algorithm* 1. *If the columns of $\Psi_{\widetilde{\mathcal{A}}}$ are full-rank and the initial input satisfies $\|x - x^0\|_\infty + \|d - d^0\|_\infty \leq \omega$, then, $x^1 = x$, where x^1 is updated from x^0 after one iteration.*

4.2. Globalization and automatic regularization parameter selection. To apply the PDAS (Algorithm 1) to (1.2), we need to have an initial guess x^0 and specify a proper regularization parameter λ in $pdas(y, \Psi, \lambda, x^0, MaxIter)$. In this section, we address these two issues together with continuation. Since the PDAS is a Newton type algorithm with a fast local convergence rate and x_{ℓ_1} is a piecewise linear function of λ [41], we adopt a continuation to fully exploit the fast local convergence. In particular, this is a simple way to globalize the convergence of PDAS [19]. Observing that $x = \mathbf{0}$ satisfies (4.1) if $\lambda \geq \lambda_0 = \|\Psi^t y/m\|_\infty$, we define $\lambda_t = \lambda_0\rho^t$ with $\rho \in (0, 1)$ for $t = 1, 2, \ldots$. We run Algorithm 1 on the sequence $\{\lambda_t\}_t$ with warmstart, i.e., using the solution x_{λ_t} as an initial guess for the λ_{t+1}-problem. When the whole continuation is done we obtain a solution path of (1.2). For simplicity, we refer to the PDAS with continuation as PDASC, described in Algorithm 2.

The regularization parameter λ in the ℓ_1-regularized 1-bit CS model (1.2), which compromises the trade-off between data fidelity and the sparsity level of the solution, is important for theoretical analysis and practical computation. However, the well-known regularization parameter selection rules such as the discrepancy principle [18, 27], the balancing principle [10, 33, 11, 28], or the Bayesian information criterion [19, 35], are not applicable to the 1-bit CS problem considered here, since the model errors

Algorithm 2. PDASC: $\{x_{\lambda_t}\}_{t\in[MaxGrid]} \leftarrow pdasc(y, \Psi, \lambda_0, x_{\lambda_0}, \rho, MaxGrid, MaxIter)$.

1: Input, y, Ψ, $\lambda_0 = \|\Psi^t y/m\|_\infty$, $x_{\lambda_0} = \mathbf{0}$, $\rho \in (0, 1)$, *MaxGrid*, *MaxIter*.
2: **for** $t = 1, 2, \ldots MaxGrid$ **do**
3: Run algorithm 1 $x_{\lambda_t} \leftarrow pdas(y, \Psi, \lambda, x^0, MaxIter)$ with $\lambda = \lambda_t = \rho^t\lambda_0$, initialized with $x^0 = x_{\lambda_{t-1}}$.
4: Check the stopping criterion.
5: **end for**
6: Output, $\{x_{\lambda_t}\}_{t\in[MaxGrid]}$.

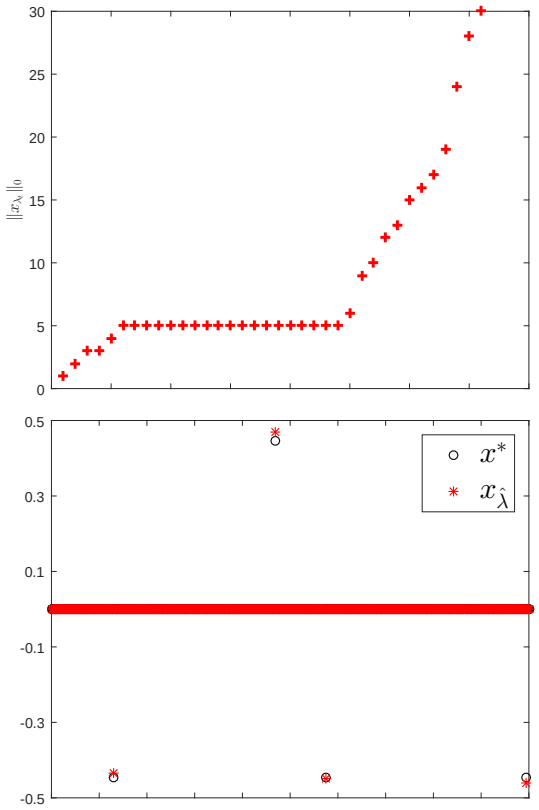


FIG. 4.1. *Active set size on the path (top panel) on data* $\{m = 400, n = 10^3, s = 5, \nu = 0.5, \sigma = 0.01, q = 2.5\%\}$ *and the underlying true signal* x^* *and the solution* $x_{\hat{\lambda}}$ *selected by* (4.5) *(bottom panel).*

are not available directly. Here we propose a novel rule to select the regularization parameter automatically. We run the PDASC to yield a solution path until $\|x_{\lambda_T}\|_0 > \lfloor \frac{m}{\log n} \rfloor$ for the smallest possible T. Let $S_\ell = \{\lambda_t : \|x_{\lambda_t}\|_0 = \ell, t = 1, \ldots, T\}, \ell = 1, \ldots, \lfloor \frac{m}{\log n} \rfloor$, be the set of regularization parameters at which the output of PDAS has ℓ nonzero elements. We determine λ by voting, i.e.,

$$\hat{\lambda} = \max\{S_{\bar{\ell}}\} \quad \text{and} \quad \bar{\ell} = \arg\max_{\ell} \{|S_\ell|\}. \tag{4.5}$$

Therefore, our parameter selection rule is seamlessly integrated with the continuation strategy which serves as a globalization technique without any extra computational overhead.

Here we give an example to show the accuracy of our proposed regularization parameter selection rule (4.5) with data $\{m = 400, n = 10^3, s = 5, \nu = 0.5, \sigma = 0.01, q = 2.5\%\}$. Descriptions of the data can be found in section 5. The top panel of Figure 4.1 shows the size of active set $\|x_{\lambda_t}\|_0$ along the path of PDASC and the lower panel shows the underlying true signal x^* and the solution $x_{\hat{\lambda}}$ selected by (4.5).

5. Numerical simulation. In this section we showcase the performance of our proposed least square decoders (2.5) and (1.2). All the computations were performed on a four-core laptop with 2.90 GHz and 8 GB RAM using MATLAB 2015b. The

MATLAB package `1-bitPDASC` for reproducing all the numerical results can be found at http://faculty.zuel.edu.cn/tjyjxxy/jyl/list.htm.

5.1. Experiment setup. First we describe the data generation process and our parameter choice. In all numerical examples the underlying target signal x^* with $\|x^*\|_0 = s$ is given, and the observation y is generated by $y = \eta \odot \mathrm{sign}(\Psi x^* + \epsilon)$, where the rows of Ψ are iid samples from $\mathcal{N}(\mathbf{0}, \Sigma)$ with $\Sigma_{jk} = \nu^{|j-k|}, 1 \le j, k \le n$. We keep the convention $0^0 = 1$. The elements of ϵ are generated from $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, and $\eta \in \mathcal{R}^m$ has independent coordinate η_i with $\mathbb{P}[\eta_i = 1] = 1 - \mathbb{P}[\eta_i = -1] = q$. Here, we use $\{m, n, s, \nu, \sigma, q\}$ to denote the data generated as above for short. Setting the decay factor ρ close to 1 will make PDAS warm start. Taking *MaxGrid* large enough will ensure $[\lambda_{MaxGrid}, \lambda_0]$ containing a good regularization parameter value. The numerical performance is not sensitive to the choices of these two parameters. We fix $\rho = 0.95, MaxGrid = 200, MaxIter = 1$ in our proposed PDASC algorithm and use (4.5) to determine the value of the regularization parameter λ. All the simulation results are based on 100 independent replications.

5.2. Accuracy and robustness of x_{ls} when $m > n$. Now we present numerical results to illustrate the accuracy of the least square decoder x_{ls} and its robustness to the noise and the sign flips. Figure 5.1 shows the recovery error $\|x_{ls} - x^*\|$ on data set $\{m = 10^3, n = 10, s = 10, \nu = 0.3, \sigma = 0 : 0.05 : 0.5, q = 2.5\%\}$. The top panel of Figure 5.2 shows the recovery error $\|x_{ls} - x^*\|$ on data set $\{m = 1000, n = 10, s = 10, \nu = 0.3, \sigma = 0.01, q = 0 : 1\% : 10\%\}$ and the bottom panel gives recovery error $\|x_{ls} + x^*\|$ on data $\{m = 1000, n = 10, s = 10, \nu = 0.3, \sigma = 0.01, q = 90 : 1\% : 100\%\}$. It is observed that the recovery error $\|x_{ls} - x^*\|$ ($\|x_{ls} + x^*\|$) of the least square decoder is small (around 0.1) and robust to noise level σ and sign flips probability q. This confirms theoretically investigations in Theorem 2.2, which states the error is of order $\widetilde{\mathcal{O}}(\sqrt{\frac{n}{m}}) = 0.1$.

5.3. Support recovery of x_{ℓ_1} when $m < n$. We conduct simulations to illustrate the performance of the model (1.2) PDASC algorithm. We report how the

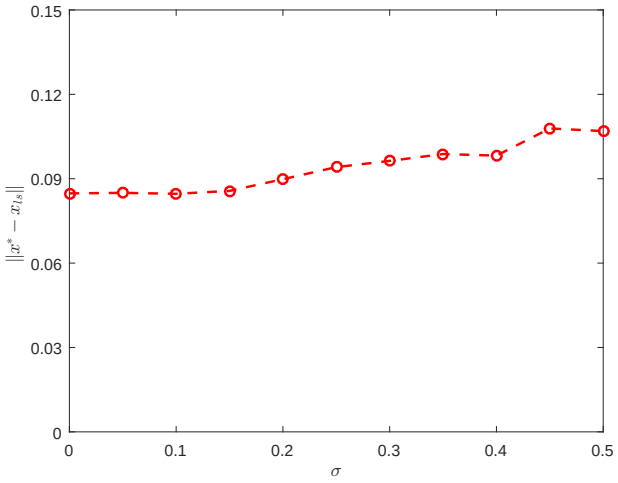


FIG. 5.1. *Recovery error* $\|x_{ls} - x^*\|$ *versus* σ *on* $\{m = 1000, n = 10, s = 10, \nu = 0.3, \sigma = 0 : 0.05 : 0.5, q = 2.5\%\}$.

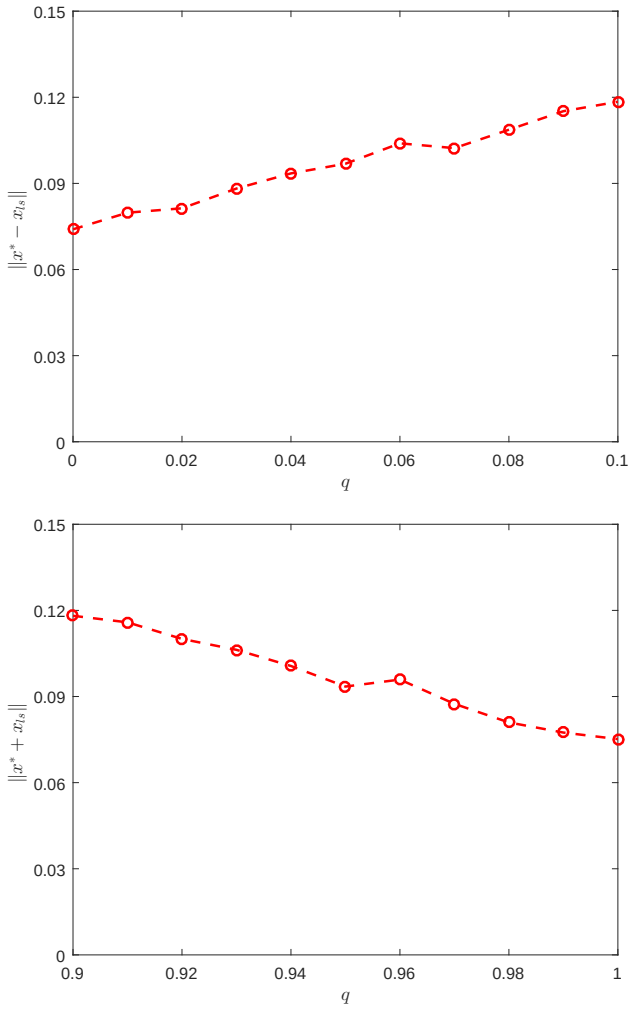
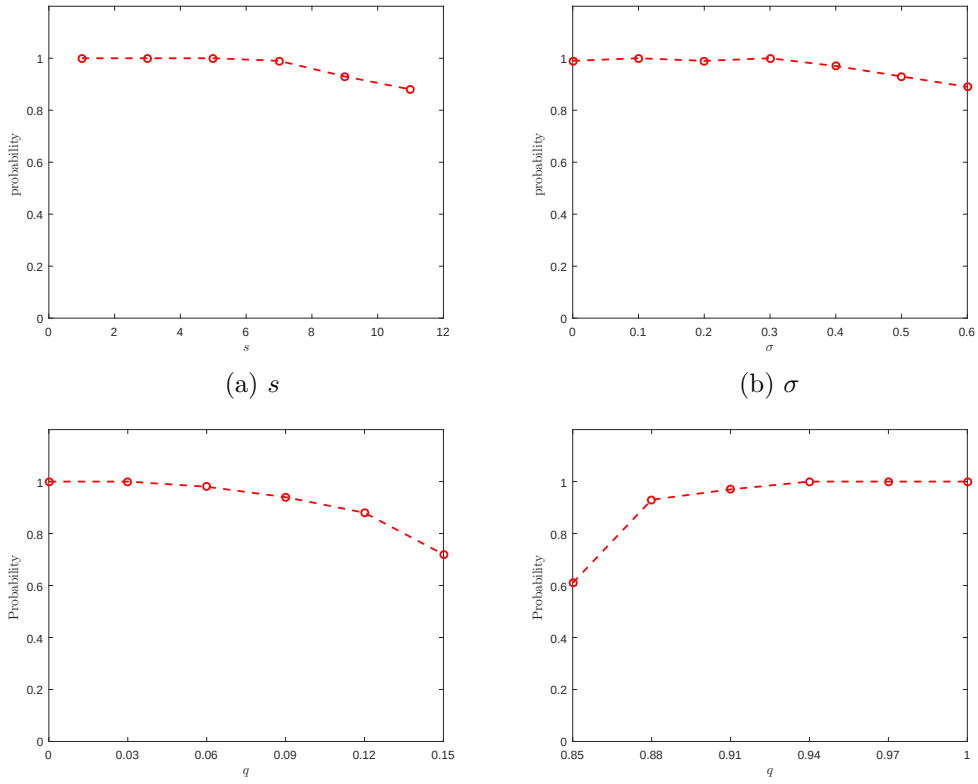


FIG. 5.2. *Recovery error* $\|x_{ls} - x^*\|$ *versus* q *on* $\{m = 10^3, n = 10, s = 10, \nu = 0.3, \sigma = 0.01, q = 0 : 1\% : 10\%\}$ *(top panel) and* $\| - x_{ls} - x^*\|$ *on* $\{m = 1000, n = 10, s = 10, \nu = 0.3, \sigma = 0.01, q = 90\% : 1\% : 100\%\}$ *(bottom panel).*

exact support recovery probability varies with the sparsity level s, the noise level σ and the probability q of sign flips. Figure 5.3 indicates that as long as the sparsity level s is not large, x_{ℓ_1} recovers the underlying true support with high probability even if the measurement contains noise and is corrupted by sign flips. This confirms the theoretical investigations in Theorem 3.1.

5.4. Comparison with other state-of-the-art methods. Now we compare our proposed model (1.2) and PDASC algorithm with several state-of-the-art methods such as BIHT [30] (http://perso.uclouvain.be/laurent.jacques/index.php/Main/BIHTDemo), AOP [51] and PBAOP [26] (both available at http://www.esat.kuleuven.be/stadius/ADB/huang/downloads/1bitCSLab.zip), and linear projection (LP) [49, 44]. BIHT, AOP, LP, and PBAOP are all required to specify the true sparsity level



(a) s (b) σ

(c) q (d) q

FIG. 5.3. *The exact support recovery probability versus* s, σ, *and* q *on* (a) *data set* $\{m = 500, n = 1000, s = 1 : 2 : 12, \nu = 0.1, \sigma = 0.05, q = 1\%\}$, (b) $\{m = 500, n = 1000, s = 5, \nu = 0.3, \sigma = 0 : 0.1 : 0.6, q = 5\%\}$, (c) $\{m = 500, n = 1000, s = 5, \nu = 0.1, \sigma = 0.01, q = 0 : 3\% : 15\%\}$, *and* (d) $\{m = 500, n = 1000, s = 5, \nu = 0.1, \sigma = 0.01, q = 85\% : 3\% : 100\%\}$.

s. Both AOP and PBAOP also require specifying the sign flips probability q. The PDASC does not require specifying the unknown parameter sparsity level s or the probability of sign flips q. We use $\{m = 500, n = 1000, s = 5, \nu = 0.1, \sigma = 0, q = 0\}$, $\{m = 500, n = 1000, s = 5, \nu = 0.3, \sigma = 0.3, q = 5\%\}$, $\{m = 500, n = 1000, s = 5, \nu = 0.5, \sigma = 0.5, q = 10\%\}$, and $\{m = 800, n = 2000, s = 10, \nu = 0.1, \sigma = 0.1, q = 1\%\}$, $\{m = 800, n = 2000, s = 10, \nu = 0.2, \sigma = 0.3, q = 3\%\}$, $\{m = 800, n = 2000, s = 10, \nu = 0.3, \sigma = 0.5, q = 5\%\}$, and $\{m = 5000, n = 20000, s = 50, \nu = 0, \sigma = 0.2, q = 3\}$, $\{m = 5000, n = 20000, s = 50, \nu = 0, \sigma = 0.1, q = 1\}$, $\{m = 5000, n = 20000, s = 5, \nu = 0, \sigma = 0.3, q = 5\%\}$. The average CPU time in seconds (Time (s)), the average of the ℓ_2 error $\|x_\ell - x^*\|$ (ℓ_2-Err), and the probability of exactly recovering true support (PrE (%)) are reported in Table 5.1. The PDASC is comparatively very fast and the most accurate even if the correlation ν, the noise level σ, and the probability of sign flips q are large.

Now we compare the PDASC with the aforementioned competitors to recover a one-dimensional signal. The true signal is sparse under wavelet basis "Db1" [38]. Thus, the matrix Ψ is of size 2500×8000 and consists of random Gaussian and an

TABLE 5.1

Comparison PDASC with state-of-the-art methods on CPU time in seconds (Time (s)), average ℓ_2 error $\|x_\ell - x^*\|$ (ℓ_2-Err), and probability on exactly recovering of true support (PrE (%)).

$\{m = 500, n = 1000, s = 5\}$									
	(a) $\{\nu = 0.1, \sigma = 0.1, q = 1\%\}$			(b) $\{\nu = 0.3, \sigma = 0.3, q = 5\%\}$			(c) $\{\nu = 0.1, \sigma = 0.5, q = 10\%\}$		
Method	Time (s)	ℓ_2-Err	PrE (%)	Time (s)	ℓ_2-Err	PrE (%)	Time	ℓ_2-Err	PrE
BIHT	1.42e-1	1.89e-1	92	1.31e-1	5.73e-1	19	1.32e-1	9.39e-1	0
AOP	2.72e-1	7.29e-2	100	3.55e-1	2.11e-1	92	3.58e-1	4.22e-1	44
LP	8.70e-3	4.19e-1	98	8.50e-3	4.22e-1	93	8.30e-3	4.81e-1	26
PBAOP	1.46e-1	9.08e-2	100	1.36e-1	2.05e-1	90	1.35e-1	4.53e-1	36
PDASC	4.11e-2	6.77e-2	100	4.38e-2	9.40e-2	99	4.56e-2	2.21e-1	71
$\{m = 800, n = 2000, s = 10\}$									
	(a) $\{\nu = 0.1, \sigma = 0.1, q = 1\%\}$			(b) $\{\nu = 0.3, \sigma = 0.2, q = 3\%\}$			(c) $\{\nu = 0.5, \sigma = 0.3, q = 5\%\}$		
Method	Time (s)	ℓ_2-Err	PrE (%)	Time (s)	ℓ_2-Err	PrE (%)	Time	ℓ_2-Err	PrE
BIHT	4.17e-1	2.10-1	84	4.25e-1	4.21e-1	25	4.35e-1	6.46e-1	0
AOP	1.09e-0	7.78-2	100	1.10e-0	1.76e-1	95	1.16e-0	2.86e-1	59
LP	1.95e-2	4.54-1	85	1.99e-2	4.49e-1	71	2.05e-2	5.03e-1	16
PBAOP	4.22e-1	1.00-1	100	4.27e-1	1.58e-1	99	4.31e-1	2.99e-1	51
PDASC	1.23e-1	8.66-2	100	1.27e-1	1.04e-1	98	1.30e-2	1.51e-1	78
$\{m = 5000, n = 20000, s = 50, \nu = 0\}$									
	(a) $\{\sigma = 0.1, q = 1\%\}$			(b) $\{\sigma = 0.2, q = 3\%\}$			(c) $\{\sigma = 0.3, q = 5\%\}$		
Method	Time (s)	ℓ_2-Err	PrE (%)	Time (s)	ℓ_2-Err	PrE (%)	Time	ℓ_2-Err	PrE
BIHT	2.56e+1	2.16e-1	58	2.58e+1	4.54e-1	0	2.58e+1	6.29e-1	0
AOP	6.44e+1	7.56e-2	100	6.46e+1	1.66e-1	96	6.47e+1	2.57e-1	16
LP	2.35e-1	4.47e-1	38	2.30e-1	4.46e-1	34	2.30e-1	4.47e-1	26
PBAOP	2.56e+1	9.89e-2	100	2.58e+1	1.66e-1	95	2.58e+1	2.60e-1	18
PDASC	7.09e-0	7.97e-2	100	7.17e-0	9.17e-2	99	7.23e-0	1.23e-1	86

inverse of one level Harr wavelet transform. The target coefficient has 36 nonzeros. We set $\sigma = 0.5$, $q = 4\%$. The recovered results are shown in Figure 5.4 and Table 5.2. The reconstruction by the PDASC is visually more appealing than others, as shown in Figure 5.4. This is further confirmed by the peak signal-to-noise ratio (PSNR) value reported in Table 5.2, which is defined by $\text{PSNR} = 10 \cdot \log \frac{V^2}{\text{MSE}}$, where V is the maximum absolute value of the true signal, and MSE is the mean squared error of the reconstruction.

6. Conclusions. In this paper we consider decoding from 1-bit measurements with noise and sign flips. For $m > n$, we show that, up to a constant c, with high probability the least squares solution x_{ls} approximates x^* with precision δ as long as $m \geq \widetilde{\mathcal{O}}(\frac{n}{\delta^2})$. For $m < n$, we assume that the underlying target x^* is s-sparse and prove that up to a constant c, with high probability, the ℓ_1-regularized least squares solution x_{ℓ_1} lies in the ball with center x^* and radius δ, provided that $m \geq \mathcal{O}(\frac{s \log n}{\delta^2})$. We introduce the one-step convergent PDAS method to minimize the nonsmooth objection function. We propose a novel tuning parameter selection rule which is seamlessly integrated with the continuation strategy without any extra computational overhead. Numerical experiments are presented to illustrate salient features of the model and the efficiency and accuracy of the algorithm.

There are several avenues for further study. First, many practitioners observed that nonconvex sparse regularization often brings in additional benefit in the standard CS setting. Whether the theoretical and computational results derived in this paper can still be justified when nonconvex regularizers are used deserves further consideration. The 1-bit CS is a kind of nonlinear sampling approach. Analysis of some other nonlinear sampling methods is also of immense interest.

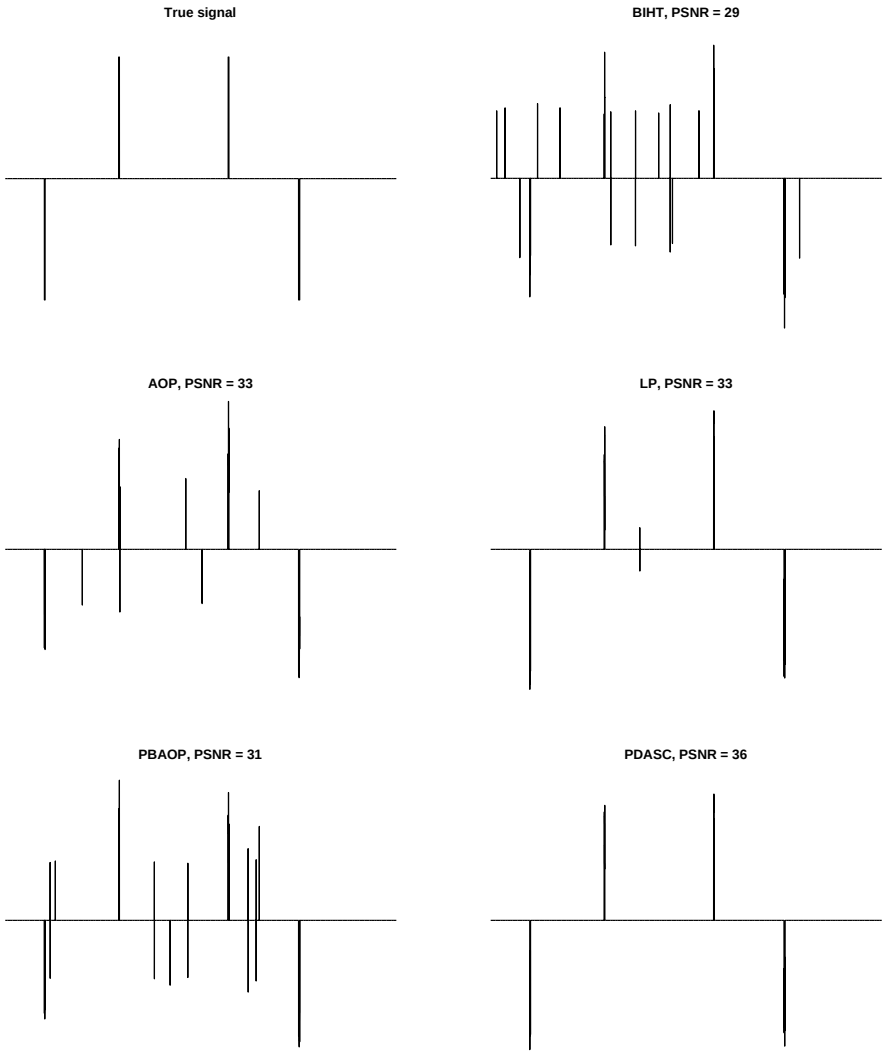


FIG. 5.4. *Reconstruction of the one-dimensional signal with* $\{m = 2500, n = 8000, s = 36, \nu = 0, \sigma = 0.5, q = 4\%\}$.

TABLE 5.2

The CPU time in seconds and the PSNR of one-dimensional signal recovery with $\{m = 2500, n = 8000, s = 36, \nu = 0, \sigma = 0.5, q = 4\%\}$.

Method	CPU time (s)	PSNR
BIHT	4.97	29
AOP	4.98	33
LP	0.11	33
PBAOP	4.93	31
PDASC	3.26	36

Appendix A. Proof of Lemma 2.1.

Proof. Let $u = \tilde{\psi}^t x^*$. Then $u \sim \mathcal{N}(0,1)$ due to $\tilde{\psi} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and the assumption that $\|x^*\|_\Sigma = 1$.

$$\begin{aligned}
\mathbb{E}[\tilde{\psi}\tilde{y}] &= \mathbb{E}[\tilde{\psi}\tilde{\eta}\mathrm{sign}(\tilde{\psi}^t x^* + \tilde{\epsilon})] = \mathbb{E}[\tilde{\eta}]\mathbb{E}[\tilde{\psi}\mathrm{sign}(\tilde{\psi}^t x^* + \tilde{\epsilon})] \\
&= [q - (1-q)]\mathbb{E}[\tilde{\psi}\mathrm{sign}(\tilde{\psi}^t x^* + \tilde{\epsilon})] \\
&= (2q-1)\mathbb{E}[\mathbb{E}[\tilde{\psi}\mathrm{sign}(\tilde{\psi}^t x^* + \tilde{\epsilon})|\tilde{\psi}^t x^*]] \\
&= (2q-1)\mathbb{E}[\mathbb{E}[\tilde{\psi}|\tilde{\psi}^t x^*]\mathrm{sign}(\tilde{\psi}^t x^* + \tilde{\epsilon})] \\
&= (2q-1)\mathbb{E}[\mathbb{E}[\tilde{\psi}|\tilde{\psi}^t x^*]\mathbb{E}[\mathrm{sign}(\tilde{\psi}^t x^* + \tilde{\epsilon})|\tilde{\psi}^t x^*]] \\
&= (2q-1)\mathbb{E}[\Sigma x^* \tilde{\psi}^t x^* \mathbb{E}[\mathrm{sign}(\tilde{\psi}^t x^* + \tilde{\epsilon})|\tilde{\psi}^t x^*]] \\
&= (2q-1)\mathbb{E}[\mathrm{sign}(\tilde{\psi}^t x^* + \tilde{\epsilon})\tilde{\psi}^t x^*]\Sigma x^* \\
&= c\Sigma x^*,
\end{aligned}$$

where $c = (2q-1)\mathbb{E}[\mathrm{sign}(\tilde{\psi}^t x^* + \tilde{\epsilon})\tilde{\psi} x^*]$. The second line follows from the independence assumption and the third from the law of total expectation, the fourth and fifth lines are due to the independence between $\tilde{\epsilon}$ and u, and the sixth line uses the projection interpretation of conditional expectation, i.e., $\mathbb{E}[\tilde{\psi}|\tilde{\psi}^t x^*] = \frac{\mathbb{E}[\tilde{\psi}\tilde{\psi}^t x^*] - \mathbb{E}[\tilde{\psi}^t x^*]\mathbb{E}[\tilde{\psi}]}{\mathbb{E}[(\tilde{\psi}^t x^* - \mathbb{E}[\tilde{\psi}^t x^*])^2]}(\tilde{\psi}^t x^* - \mathbb{E}[\tilde{\psi}^t x^*]) + \mathbb{E}[\tilde{\psi}] = \Sigma x^* \tilde{\psi}^t x^*$, where we use $\mathbb{E}[\tilde{\psi}] = \mathbf{0}$ and $u \sim \mathcal{N}(0,1)$. Let $f(t) = \frac{1}{\sqrt{2\pi}\sigma}\exp\frac{-t^2}{2\sigma^2}$ be the density function of $\tilde{\epsilon} \sim \mathcal{N}(0, \sigma^2)$. Integrating by parts shows that

$$\begin{aligned}
c = (2q-1)\mathbb{E}[\mathrm{sign}(u+\tilde{\epsilon})u] &= (2q-1)\mathbb{E}\left[(1 - 2\mathbb{P}[\tilde{\epsilon} \le -u])u\right] \\
&= (2q-1)\mathbb{E}\left[\frac{\partial(1 - 2\mathbb{P}[\tilde{\epsilon} \le -u)}{\partial u}\right] \\
&= (2q-1)\mathbb{E}[f(-u)] = (2q-1)\sqrt{\frac{2}{\pi(\sigma^2+1)}}.
\end{aligned}$$

The proof is completed by inverting $c\Sigma$. ∎

Appendix B. Preliminaries. We recall some simple properties of subgaussian and subexponential random variables.

LEMMA B.1 (Lemma 2.7.7 of [50] and Remark 5.18 of [48]). *Let ξ_1 and ξ_2 be subgaussian random variables. Then both $\xi_1\xi_2$ and $\xi_1\xi_2 - \mathbb{E}[\xi_1\xi_2]$ are subexponential random variables.*

Lemma B.2 states the nonasymptotic bound on the spectrums of Ψ and the operator norm of $\Psi^t\Psi/m - \Sigma$ when $m \ge \mathcal{O}(n)$.

LEMMA B.2. *Let $\Psi \in \mathcal{R}^{m\times n}$ whose rows ψ_i^t are independent subgaussian vectors in $\mathcal{R}^n$ with mean $\mathbf{0}$ and covariance matrix Σ. Let $m > n$. Then for every $t > 0$ with probability at least $1 - 2\exp(-C_1 t^2)$, one has*

$$(1-\tau)\sqrt{\gamma_{min}(\Sigma)} \le \sqrt{\gamma_{min}\left(\frac{\Psi^t\Psi}{m}\right)} \le \sqrt{\gamma_{max}\left(\frac{\Psi^t\Psi}{m}\right)} \le (1+\tau)\sqrt{\gamma_{max}(\Sigma)} \tag{B.1}$$

and

$$\|\Psi^t\Psi/m - \Sigma\| \le \max\{\tau, \tau^2\}\gamma_{max}(\Sigma), \tag{B.2}$$

where $\tau = C_2\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}$, *and* C_1, C_2 *are generic positive constants depending on the maximum subgaussian norm of rows of* Ψ.

Proof. Let $\Phi = \Psi\Sigma^{-\frac{1}{2}}$. Then the rows of Φ are independent subgaussian isotropic vectors. Equation (B.1) follows from Theorem 5.39 and Lemma 5.36 of [48] and (B.2) is a direct consequence of Remark 5.40 of [48]. ∎

We state the Bernstein-type inequality for the sum of independent and mean 0 subexponential random random variables.

LEMMA B.3 (Corollary 5.17 of [48]). *Let* $\xi_1, \ldots, \xi_m$ *be independent centered subexponential random variables. Then for every* $t > 0$ *one has*

$$\mathbb{P}\left[\left|\sum_{i=1}^{m} \xi_i\right| /m \geq t\right] \leq 2\exp(-\min\{C_1 t^2, C_2 t\}m),$$

where C_1 *and* C_2 *are generic positive constants depending on the maximum subexponential norm of of* ξ_i.

LEMMA B.4. *Let* $\Psi \in \mathcal{R}^{m\times n}$ *whose rows* ψ_i *are independent subgaussian vectors in* $\mathcal{R}^{n\times 1}$ *with mean* $\boldsymbol{0}$ *and covariance matrix* Σ. *Then, with probability at least* $1 - 2\exp(-C_1 C_2^2 n)$,

$$\|\Psi^t\Psi/m - \mathbb{E}[\Psi^t\Psi/m]\| \leq 2C_2\gamma_{max}(\Sigma)\sqrt{\frac{n}{m}}, \tag{B.3}$$

as long as $m \geq 4C_2^2 n$. *Furthermore, if* $m > \frac{4C_1}{C_2^2}\log n$, *then*

$$\left\|\sum_{i=1}^{m}(\mathbb{E}[\psi_i y_i] - \psi_i y_i)/m\right\|_\infty \leq 2\sqrt{\frac{\log n}{C_1 m}} \tag{B.4}$$

holds with probability at least $1 - \frac{2}{n^3}$, *and*

$$\|\Psi^t\Psi/m - \Sigma\|_\infty \leq 2\sqrt{\frac{\log n}{C_1 m}} \tag{B.5}$$

holds with probability at least $1 - \frac{1}{n^2}$.

Proof. By (2.3), $\mathbb{E}[\Psi^t\Psi/m] = \Sigma$; hence (B.3) follows from (B.2) with $t = C_2\sqrt{n}$ and the assumption $m \geq 4C_2^2 n$. Define $G_{i,j} := y_i(\psi_i)_j \in \mathcal{R}^1, i = 1, \ldots, m, j = 1, \ldots, n$, which is subexponential by Lemma B.1. Therefore,

$$\begin{aligned}
\mathbb{P}\left[\left\|\sum_{i=1}^{m}(\mathbb{E}[\psi_i y_i] - \psi_i y_i)/m\right\|_\infty \geq t\right] &= \mathbb{P}\left[\bigcup_{j=1}^{n}\left\{\left|\sum_{i=1}^{m} G_{i,j}/m\right| \geq t\right\}\right] \\
&\leq \sum_{j=1}^{n}\mathbb{P}\left[\left|\sum_{i=1}^{m} G_{k,j}^{(i)}\right| /m \geq t\right] \\
&\leq n\exp(-\min\{C_1 t^2, C_2 t\}m) \\
&\leq 2n\exp(-C_1 t^2 m),
\end{aligned}$$

where the first inequality is due to the union bound, the second follows from Lemma B.3, and the last is because of restrictions $t \leq \frac{C_2}{C_1}$ and $m < n$. Then (B.4) follows from

our assumption that $m > \frac{4C_1}{C_2^2}\log n$ by setting $t{=}2\sqrt{\frac{\log n}{C_1 m}}$. Let $G_{j,k}^i := (\psi_i)_j(\psi_i)_k - \Sigma_{j,k} \in \mathcal{R}^1, i = 1,\dots,n, j = 1,\dots,n, \ell = 1,..,n$, which is mean 0 subexponential by Lemma B.1. Therefore,

$$\begin{aligned}
\mathbb{P}[\|\Psi^t\Psi/m - \Sigma\|_\infty \geq t] &= \mathbb{P}\left[\max_{j,k}\left|\sum_{i=1}^m G_{j,k}^i/m\right| \geq t\right] \\
&= \mathbb{P}\left[\bigcup_{j=1,k=1}^{n,n}\left\{\left|\sum_{i=1}^m G_{j,k}^i/m\right| \geq t\right\}\right] \\
&\leq \sum_{j=1,k=1}^{n,n}\mathbb{P}\left[\left|\sum_{i=1}^m G_{j,k}^i/m\right| \geq t\right] \\
&\leq n^2\exp(-\min\{C_1t^2, C_2t\}m) \\
&\leq n^2\exp(-C_1t^2m),
\end{aligned}$$

where the first inequality is due to the union bound, the second follows from Lemma B.3, and the last inequality is because of restricting $t \leq \frac{C_2}{C_1}$. Then by the assumption that $m > \frac{4C_1}{C_2^2}\log n$, Lemma B.5 follows by setting $t{=}2\sqrt{\frac{\log n}{C_1 m}}$. ∎

Appendix C. Proof of Theorem 2.2.

Proof. First we show that the sample covariance matrix $\Psi^t\Psi/m$ is invertible with probability at least $1 - 2\exp(-C_1C_2^2n)$ as long as $m > 4C_2n$. This follows from (B.1) in Lemma B.2 by setting $t = C_2\sqrt{n}$. Recall

$$\widetilde{x}^* = cx^*. \tag{C.1}$$

Let

$$\Delta = y - \Psi\widetilde{x}^* \tag{C.2}$$

be the error in measuring nonlinearity, sign flips, and noise in the 1-bit CS measurement. Then,

$$\begin{aligned}
\|\Psi^t\Delta/m\|_2 = \|\Psi^t(\Psi\widetilde{x}^* - y)/m\|_2 &= \left\|\frac{\Psi^t\Psi}{m}\widetilde{x}^* - \Psi^t y/m\right\|_2 \\
&= \left\|\left(\frac{\Psi^t\Psi}{m}\widetilde{x}^* - \Sigma\widetilde{x}^*\right) + (\Sigma\widetilde{x}^* - \Psi^t y/m)\right\|_2 \\
&= \left\|\left(\frac{\Psi^t\Psi}{m}\widetilde{x}^* - \mathbb{E}\left[\frac{\Psi^t\Psi}{m}\widetilde{x}^*\right]\right) + (\mathbb{E}[\Psi^t y/m] - \Psi^t y/m)\right\|_2 \\
&\leq |c|\|x^*\|_2\left\|\frac{\Psi^t\Psi}{m} - \mathbb{E}\left[\frac{\Psi^t\Psi}{m}\right]\right\| + \left\|\sum_{i=1}^m(\mathbb{E}[\psi_i y_i] - \psi_i y_i)/m\right\|_2 \\
&\leq |c|\frac{1}{\sqrt{\gamma_{\min}(\Sigma)}}\left\|\frac{\Psi^t\Psi}{m} - \mathbb{E}\left[\frac{\Psi^t\Psi}{m}\right]\right\| + \sqrt{n}\left\|\sum_{i=1}^m(\mathbb{E}[\psi_i y_i] - \psi_i y_i)/m\right\|_\infty, \qquad \text{(C.3)}
\end{aligned}$$

where the fourth equality is due to (2.1), (2.3), and (2.4), the first inequality follows from the triangle inequality and the definition of $\widetilde{x}^*$, and the last inequality uses the assumption $1 = \|x^*\|_\Sigma^2 \geq \gamma_{\min}(\Sigma)\|x^*\|_2^2$ and the fact that

$\|\cdot\|_2 \le \sqrt{n}\|\cdot\|_\infty$. Combining with (B.3) and (B.4), we deduce that, with probability at least $1-2\exp(-C_1C_2^2n)-\frac{2}{n^3}$,

$$\|\Psi^t\Delta/m\|_2 \le \sqrt{\frac{n}{m}}2\left(|c|C_2\sqrt{\kappa(\Sigma)\gamma_{\max}(\Sigma)}+\sqrt{\frac{\log n}{C_1}}\right). \tag{C.4}$$

Now we prove that $\|x_{\rm ls}/c - cx^*\|_2 = \widetilde{\mathcal{O}}(\sqrt{\frac{n}{m}}/c)$ with high probability.

$$\begin{aligned}
|c|\|x_{\rm ls}/c - x^*\|_2 &= \|x_{\rm ls}-\widetilde{x}^*\|_2 = \|(\Psi^t\Psi)^{-1}\Psi^t y - \widetilde{x}^*\|_2\\
&\le \|(\Psi^t\Psi/m)^{-1}\Psi^t(\Psi\widetilde{x}^*+y-\Psi\widetilde{x}^*)/m-\widetilde{x}^*\|_2\\
&= \|(\Psi^t\Psi/m)^{-1}\|\|\Psi^t\Delta/m\|_2\\
&\le \sqrt{\frac{n}{m}}2\left(|c|C_2\sqrt{\kappa(\Sigma)\gamma_{\max}(\Sigma)}+\sqrt{\frac{\log n}{C_1}}\right)/\left(1-2C_2\sqrt{\frac{n}{m}}\right)^2\\
&\le \sqrt{\frac{n}{m}}4\left(|c|C_2\sqrt{\kappa(\Sigma)\gamma_{\max}(\Sigma)}+\sqrt{\frac{\log n}{C_1}}\right),
\end{aligned}$$

where the second inequality follows with probability at least $1-4\exp(-C_1C_2^2n)-\frac{2}{n^3}$ from (C.3) and (B.1) by setting $t=C_2\sqrt{n}$, and the last line is due to the assumption $m \ge 16C_2^2n$. Hence, the proof of Theorem 2.2 is completed by dividing $|c|$ on both sides and some algebra. ∎

Appendix D. Proof of Theorem 3.1.

Proof. Our proof is based on Lemmas D.1–D.3 below. Denote $R = x_{\ell_1} - \widetilde{x}^*$, $\mathcal{A}^* = \mathrm{supp}(x^*)$ and $\mathcal{I}^* = \overline{\mathcal{A}^*}$. The first lemma shows that R is sparse in the sense that its energy is mainly cumulated on $\mathcal{A}^*$ if λ is chosen properly.

LEMMA D.1. *Let*

$$\mathcal{C}_{\mathcal{A}^*} = \{z \in \mathcal{R}^n : \|z_{\mathcal{I}^*}\|_1 \le 3\|z_{\mathcal{A}^*}\|_1\}, \tag{D.1}$$

and define $\mathcal{E} = \{\|\Psi^t\Delta/m\|_\infty \le \lambda/2\}$. *Conditioning on the event* $\mathcal{E}$, *we have* $R \in \mathcal{C}_{\mathcal{A}^*}$.

Proof. The optimality of x_{ℓ_1} implies that $\frac{1}{2m}\|y-\Psi x_{\ell_1}\|_2^2 + \lambda\|x_{\ell_1}\|_1 \le \frac{1}{2m}\|y-\Psi\widetilde{x}^*\|_2^2 + \lambda\|\widetilde{x}^*\|_1$. Recall that $y = \Psi\widetilde{x}^* + \Delta$. Some algebra on the above display shows

$$\begin{aligned}
\frac{1}{2m}\|\Psi R\|_2^2 + \lambda\|R_{\mathcal{I}^*}\|_1 &\le \langle R, \Psi^t\Delta/m\rangle + \lambda\|R_{\mathcal{A}^*}\|_1\\
&\le \|R\|_1\|\Psi^t\Delta/m\|_\infty + \lambda\|R_{\mathcal{A}^*}\|_1 \le \|R\|_1\lambda/2 + \lambda\|R_{\mathcal{A}^*}\|_1,
\end{aligned}$$

where we use the Cauchy–Schwarz inequality and the definition of $\mathcal{E}$. The above inequality shows

$$\frac{1}{m}\|\Psi R\|_2^2 + \lambda\|R_{\mathcal{I}^*}\|_1 \le 3\lambda\|R_{\mathcal{A}^*}\|_1, \tag{D.2}$$

i.e., $R \in \mathcal{C}_{\mathcal{A}^*}$. This finishes the proof of Lemma D.1. ∎

The next lemma gives a lower bound on $\mathbb{P}[\mathcal{E}]$ with a proper regularization parameter λ.

LEMMA D.2. *Let* $C_3 \ge \|x^*\|_1$. *If* $m > \frac{4C_1}{C_2^2}\log n$, *taking* $\lambda = \frac{4(1+|c|C_3)}{\sqrt{C_1}}\sqrt{\frac{\log n}{m}}$, *then with probability at least* $1-2/n^3-2/n^2$, *one has*

$$\|\Psi^t\Delta/m\|_\infty \le \lambda/2. \tag{D.3}$$

Proof.

$$\begin{aligned}
\|\Psi^t\Delta/m\|_\infty &= \|\Psi^t(\Psi\widetilde{x}^* - y)/m\|_\infty = \left\|\left(\frac{\Psi^t\Psi}{m}\widetilde{x}^* - \Sigma\widetilde{x}^*\right) + (\Sigma\widetilde{x}^* - \Psi^t y/m)\right\|_\infty \\
&\le \left\|\left(\frac{\Psi^t\Psi}{m}\widetilde{x}^* - \Sigma\widetilde{x}^*\right)\right\|_\infty + \|(\mathbb{E}[\Psi^t y/m] - \Psi^t y/m)\|_\infty \\
&\le |c|\left\|\left(\frac{\Psi^t\Psi}{m} - \Sigma\right)\|_\infty \|x^*\right\|_1 + \left\|\sum_{i=1}^m (\mathbb{E}[\psi_i y_i] - \psi_i y_i)/m\right\|_\infty \\
&\le |c|C_3 2\sqrt{\frac{\log n}{C_1 m}} + 2\sqrt{\frac{\log n}{C_1 m}} \\
&= \frac{2(1+|c|C_3)}{\sqrt{C_1}}\sqrt{\frac{\log n}{m}},
\end{aligned}$$

where the first inequality is due to the triangle inequality, (2.1), (2.3), and (2.4), the second inequality follows from the definition of $\widetilde{x}^*$ and the Cauchy–Schwarz inequality, and the third one uses (B.4) and (B.5). The proof of Lemma D.2 is completed by setting $\lambda = \frac{4(1+|c|C_3)}{\sqrt{C_1}}\sqrt{\frac{\log n}{m}}$. ∎

The last lemma shows Ψ is strongly convex along the direction contained in the cone $\mathcal{C}_{\mathcal{A}^*}$ defined in (D.1).

LEMMA D.3. *If* $s \le \exp^{(1-\frac{C_1}{2})} n$ *and* $m \ge \frac{64(4\kappa(\Sigma)+1)^2}{C_1} s \log\frac{en}{s}$, *then with probability at least* $1 - 1/n^2$, *we have*

$$\|\Psi z\|_2^2/m \ge \frac{\gamma_{min}(\Sigma)}{68(4\kappa(\Sigma)+1)^2}\|z\|_2^2 \quad \forall z \in \mathcal{C}_{\mathcal{A}^*}.$$

Proof. For all $z \in \mathcal{C}_{\mathcal{A}^*} = \{v \in \mathcal{R}^n : \|v_{\mathcal{I}^*}\|_1 \le 3\|v_{\mathcal{A}^*}\|_1\}$, we sort its entries such that

$$|z_{k_1}| \ge |z_{k_2}| \ge \cdots \ge |z_{k_n}|.$$

Let $\mathcal{A} = \{k_1, \ldots, k_s\}$, and $\mathcal{I} = \bigcup_{t\ge 1}\mathcal{I}_t = \{k_{st+1}, \ldots, k_{(t+1)s}\}$, where $s = \|x^*\|_0$. Then

$$\|z_{\mathcal{I}}\|_1 \le 3\|z_{\mathcal{A}}\|_1. \tag{D.4}$$

By the elementary inequality $\|v\|_2 \le \frac{\|v\|_1}{\sqrt{s}} + \frac{\sqrt{s}\|v\|_\infty}{4}$ $\forall v \in \mathcal{R}^s$, we have

$$\begin{aligned}
\sum_{t\ge 1}\|z_{\mathcal{I}_t}\|_2 &\le \sum_{t\ge 1}\frac{\|z_{\mathcal{I}_t}\|_1}{\sqrt{s}} + \frac{\sqrt{s}\|z_{\mathcal{I}_t}\|_\infty}{4} \\
&= \frac{\|z_{\mathcal{I}}\|_1}{\sqrt{s}} + \sum_{t\ge 1}\frac{\sqrt{s}\|z_{k_{st+1}}\|_\infty}{4} \\
&\le \frac{\|z_{\mathcal{I}}\|_1}{\sqrt{s}} + \frac{\|z_{\mathcal{A}}\|_1 + \|z_{\mathcal{I}}\|_1}{4\sqrt{s}} \\
&\le \frac{3\|z_{\mathcal{A}}\|_1}{\sqrt{s}} + \frac{\|z_{\mathcal{A}}\|_1 + 3\|z_{\mathcal{A}}\|_1}{4\sqrt{s}} \quad (\textit{using (D.4)}) \\
&= \frac{4\|z_{\mathcal{A}}\|_1}{\sqrt{s}} \\
&\le 4\|z_{\mathcal{A}}\|_2, \qquad \text{(D.5)}
\end{aligned}$$

which implies

$$\|z\|_2^2 = \|z_{\mathcal{A}}\|_2^2 + \sum_{t\geq 1}\|z_{\mathcal{I}_t}\|_2^2 \leq \|z_{\mathcal{A}}\|_2^2 + \left(\sum_{t\geq 1}\|z_{\mathcal{I}_t}\|_2\right)^2 = 17\|z_{\mathcal{A}}\|_2^2. \tag{D.6}$$

Define

$$C_{2s}(\Psi) = \inf_{A\subset[n],|A|\leq 2s}\frac{\gamma_{min}(\Psi_A^t\Psi_A)}{m}$$

and

$$O_s(\Psi) = \sup_{A,B\subset[n],A\cap B=\emptyset,|A|\leq s,|B|\leq s}\|\Psi_A^t\Psi_B/m\|.$$

We obtain that

$$\begin{aligned}\langle\Psi_{\mathcal{A}}z_{\mathcal{A}},\Psi_{\mathcal{I}}z_{\mathcal{I}}\rangle/m &= \langle\Psi_{\mathcal{A}}z_{\mathcal{A}},\sum_{t\geq 1}\Psi_{\mathcal{I}_t}z_{\mathcal{I}_t}\rangle/m\\ &\leq \|z_{\mathcal{A}}\|_2\sum_{t\geq 1}\|\Psi_{\mathcal{A}}^t\Psi_{\mathcal{I}_t}\|_2\|z_{\mathcal{I}_t}\|_2\\ &\leq O_s(\Psi)\|z_{\mathcal{A}}\|_2\sum_{t\geq 1}\|z_{\mathcal{I}_t}\|_2\\ &\leq 4O_s(\Psi)\|z_{\mathcal{A}}\|^2, \end{aligned}\tag{D.7}$$

where the first inequality uses the Cauchy–Schwarz inequality, the second inequality follows from the definition of $O_s(\Psi)$, and the third is due to (D.5). Then, $\forall z\in\mathcal{C}_{\mathcal{A}^*}, z\neq 0$, we have

$$\begin{aligned}\|\Psi z\|_2^2/(m\|z\|^2) &\geq \|\Psi z\|_2^2/(17m\|z_{\mathcal{A}}\|^2)\\ &= (\|\Psi_{\mathcal{A}}z_{\mathcal{A}}\|^2 + \|\Psi_{\mathcal{I}}z_{\mathcal{I}}\|^2 + 2\langle\Psi_{\mathcal{A}}z_{\mathcal{A}},\Psi_{\mathcal{I}}z_{\mathcal{I}}\rangle)/(17m\|z_{\mathcal{A}}\|^2)\\ &\geq (\|\Psi_{\mathcal{A}}z_{\mathcal{A}}\|^2 - 8O_s(\Psi)\|z_{\mathcal{A}}\|^2)/(17m\|z_{\mathcal{A}}\|^2)\\ &\geq (C_{2s}(\Psi) - 8O_s(\Psi))/17, \end{aligned}\tag{D.8}$$

where the first inequality uses (D.6), the second inequality follows from (D.7), and the last holds due to the definition of $C_{2s}(\Psi)$. It follows from (D.8) that to complete the proof of this lemma it suffices to derive a lower bound on $C_{2s}(\Psi)$ and an upper bound on $O_s(\Psi)$ with high probability, respectively. Given $A\subset[n], |A|\leq 2s$, we define the event $E_A = \{\sqrt{\frac{\gamma_{min}(\Psi_A^t\Psi_A)}{m}} > \sqrt{\gamma_{min}(\Sigma)}(1-C_2\sqrt{\frac{2s}{m}}-\frac{t}{\sqrt{m}})\}$. Then,

$$\begin{aligned}&\mathbb{P}\left[C_{2s}(\Psi) > \gamma_{min}(\Sigma)\left(1-C_2\sqrt{\frac{2s}{m}}-\frac{t}{\sqrt{m}}\right)^2\right]\\ &= \mathbb{P}\left[\bigcap_{A\in[n],|A|\leq 2s}E_A\right] = \mathbb{P}\left[\bigcap_{A\in[n],|A|=\ell,1\leq\ell\leq 2s}E_A\right]\\ &= 1-\mathbb{P}\left[\bigcup_{A\in[n],|A|=\ell,1\leq\ell\leq 2s}\overline{E_A}\right]\\ &\geq 1-\sum_{\ell=1}^{2s}\sum_{A\subset[n],|A|=\ell}(1-\mathbb{P}[E_A])\end{aligned}$$

$$\begin{aligned}
&\geq 1-\sum_{\ell=1}^{2s}\sum_{A\subset[n],|A|\leq\ell} 2\exp(-C_1t^2)\\
&= 1-\sum_{\ell=1}^{2s}\binom{n}{\ell}2\exp(-C_1t^2)\\
&\geq 1-2\left(\frac{en}{2s}\right)^{2s}\exp(-C_1t^2),
\end{aligned}$$

where the first inequality follows from the union bound, the second inequality follows from (B.1) by replacing Ψ with Ψ_A, and the third inequality holds since $\sum_{\ell=1}^{2s}\binom{n}{\ell}\leq(\frac{n}{2s})^{2s}\sum_{\ell=0}^{2s}\binom{n}{\ell}(\frac{2s}{n})^{\ell}\leq(\frac{n}{2s})^{2s}(1+\frac{2s}{n})^{n}\leq(\frac{en}{2s})^{2s}$. Then, we derive with probability at least $1-2(\frac{en}{2s})^{2s}\exp(-C_1t^2)$,

$$C_{2s}(\Psi)>\gamma_{min}(\Sigma)\left(1-C_2\sqrt{\frac{2s}{m}}-\frac{t}{\sqrt{m}}\right)^2. \tag{D.9}$$

Given $A\subset[n],B\subset[n],|A|\leq s,|B|\leq s,A\cap B=\emptyset$, we define the event $E_{A,B}=\{\|\Psi_A^t\Psi_B/m\|>\gamma_{max}(\Sigma)(C_2\sqrt{\frac{2s}{m}}+\frac{t}{\sqrt{m}})\}$. Denote $C=A\cup B$, $\Phi_C=\Psi_C\Sigma_{CC}^{-\frac{1}{2}}$, $G_C=\Phi_C^t\Phi_C/m-\mathbf{I}_{2s}$. Then each row of Φ_C is a multivariate normal random vector that is sampled from $\mathcal{N}(\mathbf{0},\mathbf{I}_{2s})$. It follows from (B.2) with Ψ and Σ replaced by Φ_C and $\mathbf{I}_{2s}$, respectively, that

$$\mathbb{P}\left[\|G_C\|\geq C_2\sqrt{\frac{2s}{m}}+\frac{t}{\sqrt{m}}\right]\leq 2\exp\left(-C_1t^2\right).$$

Observing $\Sigma_{CC}^{-\frac{1}{2}}\Psi_A^t\Psi_B\Sigma_{CC}^{-\frac{1}{2}}/m$ is a submatrix of G_C, we deduce

$$\mathbb{P}[E_{A,B}]\leq 2\exp\left(-C_1t^2\right).$$

Then, similarly to the proof of (D.9), we have

$$\begin{aligned}
&\mathbb{P}\left[O_s(\Psi)>\gamma_{max}(\Sigma)\left(C_2\sqrt{\frac{2s}{m}}+\frac{t}{\sqrt{m}}\right)\right] \qquad \text{(D.10)}\\
&= \mathbb{P}\left[\bigcup_{A,B\subset[n],A\cap B=\emptyset,|A|\leq s,|B|\leq s}E_{A,B}\right]\\
&= \mathbb{P}\left[\bigcup_{A\subset[n],|A|=\ell,B\subset[n],|B|=\tilde{\ell},A\cap B=\emptyset,1\leq\ell\leq s,1\leq\tilde{\ell}\leq s}E_{A,B}\right]\\
&\leq \sum_{\ell=1,\tilde{\ell}=1}^{s}\sum_{A\in[n],|A|\leq\ell,B\in[n],|B|\leq\tilde{\ell},A\cap B=\emptyset}2\exp(-C_1t^2)\\
&\leq \left(\sum_{\ell=1}^{s}\binom{n}{\ell}\right)^2 2\exp(-C_1t^2)\\
&\leq 2\left(\frac{en}{s}\right)^{2s}\exp(-C_1t^2),
\end{aligned}$$

which implies with probability at least $1-2(\frac{en}{s})^{2s}\exp(-C_1t^2)$,

$$O_s(\Psi) \le \gamma_{max}(\Sigma)\left(C_2\sqrt{\frac{2s}{m}}+\frac{t}{\sqrt{m}}\right). \tag{D.11}$$

Combining (D.9) and (D.11) and setting $t=\sqrt{\frac{4s}{C_1}\log\frac{en}{s}}$, we obtain that with probability at least $1-4/(\frac{en}{s})^{2s}\ge 1-4/n^2$

$$C_{2s}(\Psi)-8O_s(\Psi)\ge\gamma_{min}(\Sigma)f\left(\sqrt{\frac{2s}{m}}+\sqrt{\frac{4s}{mC_1}\log\frac{en}{s}}\right),$$

where the unitary function $f(z)=z^2-(8\kappa(\Sigma)+2)z+1$. It follows from the assumption $s\le\exp^{(1-\frac{C_1}{2})}n$ that $\sqrt{\frac{2s}{m}}\le\sqrt{\frac{4s}{mC_1}\log\frac{en}{s}}$. Then some basic algebra shows that $f(\sqrt{\frac{2s}{m}}+\sqrt{\frac{4s}{mC_1}\log\frac{en}{s}})\ge f(\frac{1}{8\kappa(\Sigma)+2})=\frac{1}{4(4\kappa(\Sigma)+1)^2}$ as long as $m\ge\frac{64(4\kappa(\Sigma)+1)^2}{C_1}s\log\frac{en}{s}$. The proof of Lemma D.3 is completed. ∎

Now we are in the place of combining the above pieces together to finish the proof of Theorem 3.1. Recall $R=x_{\ell_1}-\widetilde{x}^*$. It follows from Lemma D.1 that $R\in\mathcal{C}_{\mathcal{A}^*}$ and (D.2) holds by conditioning on $\mathcal{E}$, i.e.,

$$\frac{1}{m}\|\Psi R\|_2^2+\lambda\|R_{\mathcal{I}^*}\|_1\le 3\lambda\|R_{\mathcal{A}^*}\|_1,$$

which together with Lemma D.3 implies that, with probability at least $1-4/n^2$,

$$\frac{\gamma_{min}(\Sigma)}{68(4\kappa(\Sigma)+1)^2}\|R\|_2^2\le 3\lambda\|R_{\mathcal{A}^*}\|_1\le 3\lambda\sqrt{s}\|R_{\mathcal{A}^*}\|_2,$$

i.e.,

$$|c|\|x_{\ell_1}/c-x^*\|_2=\|x_{\ell_1}-\widetilde{x}^*\|_2\le\frac{204(4\kappa(\Sigma)+1)^2}{\gamma_{min}(\Sigma)}\lambda\sqrt{s}.$$

The proof of Theorem 3.1 is completed by dividing $|c|$ on both sides and using Lemma D.2, which guarantees that (D.2) holds with $\lambda=\frac{4(1+|c|C_3)}{\sqrt{C_1}}\sqrt{\frac{\log n}{m}}$ with probability greater than $1-2/n^3-2/n^2$. ∎

Appendix E. Proof of the equivalency between the PDAS and (4.2)–(4.3).

Proof. Partition Z^k, D^k, and $F(Z^k)$ according to $\mathcal{A}_k$ and $\mathcal{I}_k$ such that

$$Z^k=\begin{pmatrix} x_{\mathcal{A}_k}\\ x_{\mathcal{I}_k}\\ d_{\mathcal{A}_k}\\ d_{\mathcal{I}_k}\end{pmatrix},$$

$$D^k=\begin{pmatrix} D^x_{\mathcal{A}_k}\\ D^x_{\mathcal{I}_k}\\ D^d_{\mathcal{A}_k}\\ D^d_{\mathcal{I}_k}\end{pmatrix}, \tag{E.1}$$

$$F(Z^k) = \begin{bmatrix} -d^k_{\mathcal{A}_k} + \lambda sign(x^k_{\mathcal{A}} + d^k_{\mathcal{A}}) \\ x^k_{\mathcal{I}_k} \\ \Psi^t_{\mathcal{A}_k}\Psi_{\mathcal{A}_k}x^k_{\mathcal{A}_k} + \Psi^t_{\mathcal{A}_k}\Psi_{\mathcal{I}_k}x^k_{\mathcal{I}_k} + md^k_{\mathcal{A}_k} - \Psi^t_{\mathcal{A}_k}y \\ \Psi^t_{\mathcal{I}_k}\Psi_{\mathcal{A}_k}x^k_{\mathcal{A}_k} + \Psi^t_{\mathcal{I}_k}\Psi_{\mathcal{I}_k}x^k_{\mathcal{I}_k} + md^k_{\mathcal{I}_k} - \Psi^t_{\mathcal{I}_k}y \end{bmatrix}. \tag{E.2}$$

Substituting (E.1)–(E.2) and (4.4) into (4.2), we have

$$-(d^k_{\mathcal{A}_k} + D^d_{\mathcal{A}_k}) = -\lambda sign(x^k_{\mathcal{A}} + d^k_{\mathcal{A}}), \tag{E.3}$$

$$x^k_{\mathcal{I}_k} + D^x_{\mathcal{I}_k} = \mathbf{0}, \tag{E.4}$$

(E.5)

$$\Psi^t_{\mathcal{A}_k}\Psi_{\mathcal{A}_k}(x^k_{\mathcal{A}_k} + D^x_{\mathcal{A}_k}) = \Psi^t_{\mathcal{A}_k}y - m(d^k_{\mathcal{A}_k} + D^d_{\mathcal{A}_k}) - \Psi^t_{\mathcal{A}_k}\Psi_{\mathcal{I}_k}(x^k_{\mathcal{I}_k} + D^x_{\mathcal{I}_k}),$$

$$m(d^k_{\mathcal{I}_k} + D^d_{\mathcal{I}_k}) = \Psi^t_{\mathcal{I}_k}y - \Psi^t_{\mathcal{I}_k}\Psi_{\mathcal{A}_k}(x^k_{\mathcal{A}_k} + D^x_{\mathcal{A}_k}) - \Psi^t_{\mathcal{A}_k}\Psi_{\mathcal{I}_k}(x^k_{\mathcal{I}_k} + D^x_{\mathcal{I}_k}). \tag{E.6}$$

It follows from (4.3) that

$$\begin{pmatrix} x^{k+1}_{\mathcal{A}_k} \\ x^{k+1}_{\mathcal{I}_k} \\ d^{k+1}_{\mathcal{A}_k} \\ d^{k+1}_{\mathcal{I}_k} \end{pmatrix} = \begin{pmatrix} x^k_{\mathcal{I}_k} + D^x_{\mathcal{I}_k} \\ d^k_{\mathcal{A}_k} + D^d_{\mathcal{A}_k} \\ x^k_{\mathcal{A}_k} + D^x_{\mathcal{A}_k} \\ d^k_{\mathcal{I}_k} + D^d_{\mathcal{I}_k} \end{pmatrix}. \tag{E.7}$$

Substituting (E.7) into (E.3)–(E.6), we get the iteration procedure of PDAS in Algorithm 1. This completes the proof. ∎

REFERENCES

[1] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin, *One-bit compressed sensing with non-gaussian measurements*, Linear Algebra Appl., 441 (2014), pp. 222–239.

[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, *Optimization with sparsity-inducing penalties*, Found. Trend. Mach. Learn., 4 (2012), pp. 1–106.

[3] R. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wootters, *One-Bit Compressive Sensing of Dictionary-Sparse Signals*, preprint, arXiv:1606.07531, 2016.

[4] R. G. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wootters, *Exponential decay of reconstruction error from binary measurements of sparse signals*, IEEE Trans. Inform. Theory, 63 (2017), pp. 3368–3385.

[5] P. T. Boufounos, *Greedy sparse signal reconstruction from sign measurements*, in Conference Record of the 43rd Asilomar Conference on Signals, Systems and Computers, IEEE, 2009, pp. 1305–1309.

[6] P. T. Boufounos and R. G. Baraniuk, 1-*bit compressive sensing*, in Proceedings of the 42nd Annual Conference on Information Sciences and Systems, IEEE, 2008, pp. 16–21.

[7] D. R. Brillinger, *A generalized linear model with gaussian regressor variables*, in Selected Works of David Brillinger, Springer, New York, 2012, pp. 589–606.

[8] E. J. Candés, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.

[9] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61.

[10] C. Clason, B. Jin, and K. Kunisch, *A duality-based splitting method for l^1-TV image restoration with automatic regularization parameter choice*, SIAM J. Sci. Comput., 32 (2010), pp. 1484–1505.

[11] C. Clason, B. Jin, and K. Kunisch, *A semismooth newton method for L^1 data fitting with automatic choice of regularization parameters and noise calibration*, SIAM J. Imaging Sci., 3 (2010), pp. 199–231.

[12] P. L. Combettes and J. C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, eds., Springer, Berlin, 2011, pp. 185–212.

[13] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.
[14] D.-Q. Dai, L. Shen, Y. Xu, and N. Zhang, *Noisy* 1*-bit compressive sensing: models and algorithms*, Appl. Comput. Harmon. Anal., 40 (2016), pp. 1–32.
[15] I. Dassios, K. Fountoulakis, and J. Gondzio, *A preconditioner for a primal-dual newton conjugate gradient method for compressed sensing problems*, SIAM J. Sci. Comput., 37 (2015), pp. A2783–A2812.
[16] B. Dong and Z. Shen, *MRA based wavelet frames and applications*, IAS Lecture Notes Series, Summer Program on the Mathematics of Image Processing, Park City Mathematics Institute, 2010.
[17] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.
[18] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl., 375, Springer, New York, 1996.
[19] Q. Fan, Y. Jiao, and X. Lu, *A primal dual active set with continuation for compressed sensing*, IEEE Trans. Signal Process, 62 (2014), pp. 6276–6285.
[20] M. Fazel, E. Candes, B. Recht, and P. Parrilo, *Compressed sensing and robust recovery of low rank matrices*, in Proceedings of the 42nd Asilomar Conference on Signals, Systems and Computers, IEEE, 2008, pp. 1043–1047.
[21] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Vol. 1, Birkhäuser, Basel, 2013.
[22] K. Fountoulakis, J. Gondzio, and P. Zhlobich, *Matrix-free interior point method for compressed sensing problems*, Math. Program. Comput., 6 (2014), pp. 1–31.
[23] S. Gopi, P. Netrapalli, P. Jain, and A. Nori, *One-bit compressed sensing: Provable support and vector recovery*, in Proceedings of the International Conference on Machine Learning, 2013, pp. 154–162.
[24] A. Gupta, R. Nowak, and B. Recht, *Sample complexity for* 1*-bit compressed sensing and sparse classification*, in Proceedings of the IEEE International Symposium on Information Theory, IEEE, 2010, pp. 1553–1557.
[25] J. Haupt and R. Baraniuk, *Robust support recovery using sparse compressive sensing matrices*, in Proceedings of the 45th Annual Conference on Information Sciences and Systems, IEEE, 2011, pp. 1–6.
[26] X. Huang, L. Shi, M. Yan, and J. A. Suykens, *Pinball Loss Minimization for One-Bit Compressive Sensing*, preprint, arXiv:1505.03898, 2015.
[27] K. Ito and B. Jin, *Inverse Problems: Tikhonov Theory and Algorithms*, Ser. Appl. Math., 22, World Scientific, River Edge, NJ, 2014.
[28] K. Ito, B. Jin, and T. Takeuchi, *A regularization parameter for nonsmooth tikhonov regularization*, SIAM J. Sci. Comput., 33 (2011), pp. 1415–1438.
[29] K. Ito and K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications*, Adv. Design Control 15, SIAM, 2008.
[30] L. Jacques, K. Degraux, and C. De Vleeschouwer, *Quantized Iterative Hard Thresholding: Bridging* 1*-Bit and High-Resolution Quantized Compressed Sensing*, preprint, arXiv:1305.1786, 2013.
[31] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, *Robust* 1*-bit compressive sensing via binary stable embeddings of sparse vectors*, IEEE Trans. Inform. Theory, 59 (2013), pp. 2082–2102.
[32] Y. Jiao, B. Jin, and X. Lu, *A primal dual active set with continuation algorithm for the* ℓ^0*-regularized optimization problem*, Appl. Comput. Harmon. Anal., 39 (2015), pp. 400–426.
[33] B. Jin, Y. Zhao, and J. Zou, *Iterative parameter choice by discrepancy principle*, IMA J. Numer. Anal., 32 (2012), pp. 1714–1732.
[34] K. Knudson, R. Saab, and R. Ward, *One-bit compressive sensing with norm estimation*, IEEE Trans. Inform. Theory, 62 (2016), pp. 2748–2758.
[35] S. Konishi and G. Kitagawa, *Information Criteria and Statistical Modeling*, Springer, New York, 2008.
[36] J. N. Laska, Z. Wen, W. Yin, and R. G. Baraniuk, *Trust, but verify: Fast and accurate signal recovery from* 1*-bit compressive measurements*, IEEE Trans. Signal Process., 59 (2011), pp. 5289–5301.
[37] W. Liu, D. Gong, and Z. Xu, *One-bit compressed sensing by greedy algorithms*, Numer. Math. Theory Methods Appl., 9 (2016), pp. 169–184.
[38] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, Academic Press, New York, 2008.
[39] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, *A unified framework for high-dimensional analysis of* m*-estimators with decomposable regularizers*, Statist. Sci., 27 (2012), pp. 538–557.

[40] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim., 87, Springer, New York, 2013.
[41] M. R. Osborne, B. Presnell, and B. A. Turlach, *A new approach to variable selection in least squares problems*, IMA J. Numer. Anal., 20 (2000), pp. 389–403.
[42] Y. Plan and R. Vershynin, *One-bit compressed sensing by linear programming*, Comm. Pure Appl. Math., 66 (2013), pp. 1275–1297.
[43] Y. Plan and R. Vershynin, *Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach*, IEEE Trans. Inform. Theory, 59 (2013), pp. 482–494.
[44] Y. Plan, R. Vershynin, and E. Yudovina, *High-dimensional estimation with geometric constraints*, Inform. Inference, 6 (2017), pp. 1–40.
[45] L. Qi and J. Sun, *A nonsmooth version of newton's method*, Math. Program., 58 (1993), pp. 353–367.
[46] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.
[47] J. A. Tropp and S. J. Wright, *Computational methods for sparse solution of linear inverse problems*, Proc. IEEE, 98 (2010), pp. 948–958.
[48] R. Vershynin, *Introduction to the Non-asymptotic Analysis of Random Matrices*, preprint, arXiv:1011.3027, 2010.
[49] R. Vershynin, *Estimation in high dimensions: A geometric perspective*, in Sampling Theory, A Renaissance, Springer, New York, 2015, pp. 3–66.
[50] R. Vershynin, *High Dimensional Probability*, https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf (2017).
[51] M. Yan, Y. Yang, and S. Osher, *Robust 1-bit compressive sensing using adaptive outlier pursuit*, IEEE Trans. Signal Process., 60 (2012), pp. 3868–3875.
[52] L. Zhang, J. Yi, and R. Jin, *Efficient algorithms for robust one-bit compressive sensing*, in Proceedings of the International Conference on Machine Learning, 2014, pp. 820–828.
[53] A. Zymnis, S. Boyd, and E. Candes, *Compressed sensing with quantized measurements*, IEEE Signal Process. Lett., 17 (2010), pp. 149–152.