One-Bit Compressed Sensing via One-Shot Hard Thresholding

Jie Shen
Stevens Institute of Technology
New Jersey, USA
jie.shen@stevens.edu

Abstract

This paper concerns the problem of 1-bit compressed sensing, where the goal is to estimate a sparse signal from a few of its binary measurements. We study a non-convex sparsity-constrained program and present a novel and concise analysis that moves away from the widely used notion of Gaussian width. We show that with high probability a simple algorithm is guaranteed to produce an accurate approximation to the normalized signal of interest under the ℓ_2-metric. On top of that, we establish an ensemble of new results that address norm estimation, support recovery, and model misspecification. On the computational side, it is shown that the non-convex program can be solved via one-step hard thresholding which is dramatically efficient in terms of time complexity and memory footprint. On the statistical side, it is shown that our estimator enjoys a near-optimal error rate under standard conditions. The theoretical results are further substantiated by numerical experiments.

1 INTRODUCTION

The last two decades have witnessed a large demand of learning from high-dimensional data where the number of attributes is of the same order of, or even greater than the number of observations. Consider, for example, the Lou Gehrig's disease: there are millions of possible factors to evaluate but the scientists have a very restricted number of samples for research (each year 2 out of 100,000 individuals are affected by it). It thus turns out to be indispensable to investigate and to resolve two fundamental problems in this scenario: **a)** when is it possible to learn a useful model from the small amount of data; and **b)** how can we construct an accurate estimator with mild computational overhead. Answering these two questions has become a central theme in the field of high-dimensional statistics, and there have been a fruitful literature in a variety of applications such as linear regression [11, 33], classification [16], principal component analysis [22], precision matrix estimation [9], to name a few.

As a special instance of linear regression, compressed sensing (CS) has attracted increasing attention owing to its ease of implementation and the success in practical problems [15]. To be more detailed, CS consists of two phases: data acquisition and signal recovery. Suppose that $\bar{\boldsymbol{x}} \in \mathbb{R}^d$ is the signal of interest. During the first phase, the goal is to efficiently sample it and store the obtained measurements in a device. In light of the fact that such a device may not be computationally powerful, a common paradigm is to adopt linear measurements which mitigate hardware implementation and accelerate data collection. That is, we specify the number of measurements n and the sensing vectors $\{\boldsymbol{a}_i\}_{i=1}^n \subset \mathbb{R}^d$, and record in the device the measurements $\{y_i\}_{i=1}^n$ given by

$$y_i = \langle \boldsymbol{a}_i, \bar{\boldsymbol{x}} \rangle \,, \ \forall\, i = 1, \ldots, n. \tag{1.1}$$

For the purpose of efficient sampling, it is required that $n \ll d$, hence the name compressed sensing.

During the second phase, one has access to $\{(\boldsymbol{a}_i, y_i)\}_{i=1}^n$ and the primary concern is to recover the underlying signal $\bar{\boldsymbol{x}}$ on a possibly powerful machine. While (1.1) is an underdetermined linear system that might have infinite number of feasible solutions, it has been well-understood that with a careful design of the sensing vectors, a broad range of algorithms are capable of producing accurate, or even exact reconstruction of $\bar{\boldsymbol{x}}$ in polynomial time provided that it exhibits certain low-dimensional structure. See, for example, [10, 34, 37, 7, 30].

Though elegant in theory and compelling in practice, [8] pointed out that it is not always realistic to collect the measurements as in (1.1) since they entail infinite-bit precision for the hardware. Alternatively, the measurements

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

are often quantized into finite bits and in the extreme case, only the sign patterns are retained:

$$y_i = \text{sign}\left(\langle \boldsymbol{a}_i, \bar{\boldsymbol{x}} \rangle\right), \ \forall\, i = 1, \ldots, n. \qquad (1.2)$$

The problem of recovering $\bar{\boldsymbol{x}}$ from its binary measurements is referred to as 1-bit compressed sensing, and it bears the potential of savings of physical storage as long as accurate estimation in the 1-bit setting does not require significantly more measurements [20]. Note that as the sign function absorbs the magnitude of $\bar{\boldsymbol{x}}$, we can only hope to reconstruct the direction $\bar{\boldsymbol{x}}/\left\|\bar{\boldsymbol{x}}\right\|$ in general. In this light, the research lines can be roughly divided into three spaces: **(a)** sparse approximation, i.e. identifying the direction; **(b)** norm estimation, i.e. evaluating the norm with extra information; and **(c)** support recovery, i.e. determining the position of the non-zero entries of $\bar{\boldsymbol{x}}$.

Sparse Approximation. Perhaps surprisingly, it was provably shown in [24] that by seeking a global optimum of ℓ_1-norm constrained programs, it is possible to reliably recover the underlying normalized signal with as many observations as in the standard model (1.1). However, it comes up with a computational issue when optimizing the programs over large-scale data sets. From a technical perspective, since the programs are not strongly convex, only a sublinear convergence rate is guaranteed for projected gradient descent [23]. In other words, one has to run the algorithm for a considerable number of iterations in order to attain the optimum. Therefore, a large body of works were dedicated to investigating alternative programs that are endowed with closed-form solutions, which naturally rules out the trouble of slow convergence. For instance, [45] considered optimizing an ℓ_1-norm penalized function which is exactly the Lagrangian function of the program in [24]. Interestingly, with the slight modification, it was shown that a simple element-wise thresholding on a certain vector gives even better error rate. Yet, for the sake of accurate recovery, one has to search for a proper regularization parameter which itself could be expensive. In particular, specifying a large value for it will result in trivial statistical guarantee whereas a small value will lead to a trivial solution (i.e. a zero vector). [19] illustrated that one can derive closed-form solutions for other sparsity-regularized programs such as smoothly clipped absolute deviation penalty [17], minimax concave penalty [44], and ℓ_0-norm (that counts the number of non-zeros of a vector). Unfortunately, no statistical guarantee was established in that work.

Norm Estimation. While it seems implausible to evaluate the norm of $\bar{\boldsymbol{x}}$ under the 1-bit setting, a number of recent works asserted that the premise quickly changes under extra conditions. For example, [21, 5] concurrently showed that by manually adding Gaussian noise before quantization, one may estimate the magnitude by solving an augmented program that incorporates the noise information. This idea was further utilized to deal with heavy-tailed sensing vectors in [14].

Support Recovery. Parallel to the line of estimating the direction of the signal, a plethora of works examined the problem of recovering the support set. In statistics, it is also known as variable selection or feature selection [33, 38]. Note that support recovery sheds light on sparse approximation in that once the support is identified, we can safely eliminate the irrelevant features and apply standard tools from regression theory (that has a rich literature). Unfortunately, it is typically more challenging to establish theoretical guarantee in this space even under the standard CS model [41, 28, 29]. In the regime of 1-bit CS, [18, 1] presented a set of impressive results based on a new family of sensing vectors that move away from the standard Gaussian. The data acquisition is computationally more demanding though.

1.1 OUR CONTRIBUTIONS

In this paper, we propose to study a sparsity-constrained non-convex program for the 1-bit CS problem. We first demonstrate that the global optimum can be computed by a simple hard thresholding operator (to be defined) which is computationally efficient. More importantly, the solution is proved to have a near-optimal approximation error rate to the direction of $\bar{\boldsymbol{x}}$. We develop novel analysis showing that the approximation error is controlled by two factors: one is independent of the signal structure while the other is entirely determined by it. On top of that, we provide theoretical justification that our estimator recovers the support set of $\bar{\boldsymbol{x}}$, can be tailored to offer an accurate estimation of the magnitude of $\bar{\boldsymbol{x}}$, and is resilient to model misspecification. Namely, even when we feed an inappropriate parameter to the program, our estimate still exhibits favorable performance.

1.2 RELATED WORKS

The 1-bit CS problem is closely related to learning halfspaces – a central object of study in learning theory [35]. In compressed sensing, there is an additional structural assumption that the halfspace is embedded in a high-dimensional space, and at the heart of CS is leveraging such prior knowledge for improved sample complexity. In view of the connection to learning theory, a surge of recent works are devoted to understanding their interplay. For example, active learning is a long-established research field in learning theory, which attempts to mitigate the human labor for data annotation by actively querying the labels [4, 40, 42, 43, 31]. Similar ideas are also explored in CS, known as adaptive sensing [3, 5].

It is also worth mentioning that one can expand the observation model (1.2) to 1-bit matrix completion that has found successful applications in social networks. For example, in recommender systems we can construct an incomplete data matrix where each observed entry represents the preference (like or dislike) of a user to an item [12, 39]. The goal is to predict the missing entries and to recommend items that are potentially interesting for the customers. The statistical guarantee for this problem has been established in a series of appealing works [6, 27].

Notation. We use bold lowercase letters such as $\boldsymbol{v}$ to denote a column vector. Its ith element is denoted by v_i. There are three norms that will be involved for a vector $\boldsymbol{v} \in \mathbb{R}^d$: the ℓ_2-norm $\|\boldsymbol{v}\| := \sqrt{v_1^2 + \cdots + v_d^2}$, the ℓ_1-norm $\|\boldsymbol{v}\|_1 := |v_1| + \cdots + |v_d|$, and the ℓ_∞-norm $\|\boldsymbol{v}\|_\infty := \max_{1 \le i \le d} |v_i|$. We write the number of non-zero elements in $\boldsymbol{v}$ as $\|\boldsymbol{v}\|_0$, and with a slight abuse of terminology we call it ℓ_0-norm[1].

For a finite set $\mathcal{S}$, its cardinality is denoted by $|\mathcal{S}|$. The index set of the non-zero elements of a vector $\boldsymbol{v} \in \mathbb{R}^d$ is called the support set, and is denoted by $\operatorname{supp}(\boldsymbol{v})$. We write the index set of the top k elements (in magnitude) as $\operatorname{supp}(\boldsymbol{v}, k)$ with ties being broken lexicographically.

We reserve $\bar{\boldsymbol{x}}$ for the s-sparse target signal that we hope to estimate. We also reserve the upright capital letter C and its subscript variants such as C_0 and C_1 for absolute constants, whose values may change from appearance to appearance. For two scalars a and b, we write $a = O(b)$ if $a \le \mathrm{C} \cdot b$; and we write $a = \Omega(b)$ if $a \ge \mathrm{C} \cdot b$.

Overview. In Section 2, we elaborate the problem setup of 1-bit compressed sensing. The primary theoretical results and the comparison to the prior works are developed in Section 3, and Section 4 presents useful extensions of our main results. In Section 5 we provide empirical evidence to support our analysis. Section 6 concludes the paper and all the proof details are deferred to the Appendix (see the supplementary material).

2 PRELIMINARIES

Let $\bar{\boldsymbol{x}} \in \mathbb{R}^d$ be the s-sparse signal of interest. For now, we presume without loss of generality that it has unit ℓ_2-norm. While our concentration is on the 1-bit CS problem (1.2), we begin our discussion with a more general observation model: for each measurement $y_i \in \{-1, 1\}$, instead of treating it as being generated from a deterministic mapping through the sign function, we consider

$$\mathbb{E}[y_i \mid \boldsymbol{a}_i] = \theta_i\left(\langle \boldsymbol{a}_i, \bar{\boldsymbol{x}} \rangle\right), \ \forall\, i = 1, \ldots, n, \tag{2.1}$$

[1] $\|\boldsymbol{v}\|_0$ is not a norm as it is not absolutely homogeneous.

where the mapping functions $\theta_i(\cdot) \in [-1, +1]$. Note that under the observation model (1.2), $\theta_i(\cdot)$ is exactly the sign function. It is worth mentioning that our model (2.1) is more general than the one considered in prior works [24, 21, 25], since we allow different mapping function for different sample, i.e. θ_i can be mutually distinct.

2.1 THE NON-CONVEX ESTIMATOR

In order to recover $\bar{\boldsymbol{x}}$, we consider the estimate $\hat{\boldsymbol{x}}$ that is a global optimum of the following program:

$$\max_{\boldsymbol{x} \in \mathbb{R}^d} \ \boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{x}, \quad \text{s.t. } \|\boldsymbol{x}\|_0 \le k, \ \|\boldsymbol{x}\| = 1, \tag{2.2}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)^\top$. The two constraints accommodate our prior knowledge on the signal $\bar{\boldsymbol{x}}$, and the objective function seeks for a maximum correlation between the input and output of $\theta_i(\cdot)$. We recall that $\bar{\boldsymbol{x}}$ is s-sparse. The integer $k > 0$ serves as the only parameter in our estimator, and we need to tune it in practice. Ideally, k plays as an upper bound on the true sparsity s, which indicates that our estimator is unbiased, i.e. the true signal is contained in the feasible set. Once the condition is violated, we say the model is misspecified which needs a particular treatment (see Section 4.3).

The optimal solution of the program (2.2) can be obtained by simple algebraic calculation. Define the hard thresholding operator as follows:

$$\mathcal{H}_k(\boldsymbol{z}) := \operatorname*{arg\,min}_{\boldsymbol{u} \in \mathbb{R}^d, \|\boldsymbol{u}\|_0 \le k} \|\boldsymbol{u} - \boldsymbol{z}\|. \tag{2.3}$$

We have the following computational result whose proof is given in Appendix A.1.

Proposition 2.1. *The global optimum $\hat{\boldsymbol{x}}$ of the program (2.2) is given by*

$$\hat{\boldsymbol{x}} = \frac{\mathcal{H}_k(\boldsymbol{v})}{\|\mathcal{H}_k(\boldsymbol{v})\|}, \ \textit{where } \boldsymbol{v} := \boldsymbol{A}^\top \boldsymbol{y}.$$

Computational Cost. The time complexity of calculating $\boldsymbol{A}^\top \boldsymbol{y}$ is $O(dn)$, and applying the hard thresholding costs $O(k \log d)$ since the k-sparse vector $\mathcal{H}_k(\boldsymbol{z})$ can be efficiently computed by first sorting the elements of $\boldsymbol{z}$ in their magnitude, and then retaining the top k of them. Therefore, the total running time is $O(dn)$ given that n is always greater than $k \log d$ in compressed sensing.

3 MAIN RESULTS

We now move on to present the statistical estimation error under standard conditions [24]. We assume that each observation y_i depends on the measurement vector $\boldsymbol{a}_i$ only through $\langle \boldsymbol{a}_i, \bar{\boldsymbol{x}} \rangle$. Namely,

$(A1)$ given the inner product $\langle \boldsymbol{a}_i, \bar{\boldsymbol{x}} \rangle$, y_i and $\boldsymbol{a}_i$ are conditionally independent.

We will focus on the standard Gaussian design of the sensing vectors. That is,

$(A2)$ $\boldsymbol{a}_i \sim N(\boldsymbol{0}, \boldsymbol{I}_{d\times d})$ for all $i = 1, \ldots, n$, and they are mutually independent.

The above sensing scheme offers a clean picture of theoretical understanding. Furthermore, in the context of compressed sensing, we indeed have the control of selecting the sensing vectors. Note that it is possible to relax the assumption to correlated Gaussian design [24], or even non-Gaussian measurements [2, 14].

In order to estimate the s-sparse signal $\bar{\boldsymbol{x}}$ from its non-linear measurements, we have to confine ourselves to a family of mapping functions. As a matter of fact, if all functions θ_i output zero almost surely, then no algorithm is able to recover the underlying signal. Mathematically, the minimum requirement on the functions is that their input and output are correlated. Let

$$\lambda_i := \mathbb{E}_{g\sim N(0,1)}\big[g \cdot \theta_i(g)\big],\ \forall\, i = 1, \ldots, n. \tag{3.1}$$

In light of $(A1)$ and $(A2)$, λ_i essentially characterizes the correlation of interest [24]. While we do not impose individual condition on λ_i, we need to assume

$(A3)$ the average correlation $\lambda := \frac{1}{n}\sum_{i=1}^n \lambda_i > 0$.

Again, since we have the freedom to design $\theta_i(\cdot)$, we can always replace all $\theta_i(\cdot)$ with $-\theta_i(\cdot)$ if we find $\lambda < 0$. We remark that a very recent work [32] studies the interesting case $\lambda = 0$ under extra assumptions.

We are now in the position to present performance guarantee of our estimate $\hat{\boldsymbol{x}}$. We show that as soon as we collect $n = O(k \log d)$ measurements, it is possible to accurately approximate the direction of $\bar{\boldsymbol{x}}$ with high probability from its non-linear measurements.

Theorem 3.1. *Assume $(A1)$, $(A2)$ and $(A3)$. Further assume $\|\bar{\boldsymbol{x}}\| = 1$, $\|\bar{\boldsymbol{x}}\|_0 = s$, and $k \geq s$. With probability at least $1 - d^{-10}$, we have*

$$\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\| \leq \frac{\mathrm{C}}{\lambda}\sqrt{\frac{k \log d}{n}}.$$

Remark 3.2. While our assumption on λ is very mild, it is important to note that in order to obtain near-optimal sample complexity, λ must act as a universal constant, which translates into an implicit requirement that most of the λ_i's are positive constants. Otherwise, Theorem 3.1 may offer trivial guarantee. Consider, for example, $\lambda_1 = \lambda_2 = \ldots \lambda_{n-1} = 0$ and $\lambda_n = 1$. It corresponds to $\lambda = 1/n$ for which we have a trivial upper bound $O(\sqrt{nk \log d})$ on the estimation error. This is not surprising since many zero correlations indicate that the sampling power is wasted on the associated samples (and thus recovery is impossible).

Fortunately, we can show that among many prevalent statistical models, the parameter λ acts as a universal non-zero constant, which indicates that the sample complexity of our estimator is near-optimal in light of the fact that the information-theoretic lower bound for infinite-bit CS is $\Omega(k \log(d/k))$ [26].

3.1 NOISELESS 1-BIT CS

Consider the problem (1.2) where $\theta_i(g) = \operatorname{sign}(g)$. With some calculation,

$$\lambda_i = \mathbb{E}_{g\sim N(0,1)}[g \operatorname{sign}(g)] = \mathbb{E}[|g|] = \sqrt{2/\pi}.$$

Thus, we obtain the following corollary regarding the sample complexity of our estimator.

Corollary 3.3. *Assume the same conditions as in Theorem 3.1. Under the model (1.2),*

$$\Pr\left(\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\| \leq \epsilon\right) \geq 1 - d^{-10}$$

for any $\epsilon \in (0, 1)$ provided that $n = O\left(\epsilon^{-2} k \log d\right)$.

3.2 NOISY 1-BIT CS

In real-world applications, the observations are not only highly quantized, but are also grossly corrupted. In fact, owing to possible systematic errors or human mistakes, the sign may be flipped with some probability. The noisy model can thus be formulated as

$$y_i = \xi_i \operatorname{sign}\left(\langle \boldsymbol{a}_i, \bar{\boldsymbol{x}} \rangle\right), \quad \forall\, i = 1, \ldots, n, \tag{3.2}$$

where ξ_i is independent of $\boldsymbol{a}_i$ and

$$\Pr(\xi_i = 1) = 1 - p_i, \quad \Pr(\xi_i = -1) = p_i$$

for some $p_i \in [0, 0.5)$. This gives a new mapping function and a new correlation parameter as follows:

$$\theta_i(g) = \operatorname{sign}(g) \cdot \mathbb{E}[\xi_i] = (1 - 2p_i)\operatorname{sign}(g),$$
$$\lambda_i = \sqrt{2/\pi}(1 - 2p_i).$$

Corollary 3.4. *Assume the same conditions as in Theorem 3.1. Under the model (3.2), for any $\epsilon \in (0, 1)$*

$$\Pr\left(\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\| \leq \epsilon\right) \geq 1 - d^{-10}$$

provided that $n = O\left(((1 - 2p)\epsilon)^{-2} k \log d\right)$ where $p = \frac{1}{n}\sum_{i=1}^n p_i$.

3.3 PROOF SKETCH OF THEOREM 3.1

Here we sketch the proof and highlight our novelty for the analysis. We defer all technical details to Appendix A. Note that $\bar{\boldsymbol{x}}$ is a feasible solution to (2.2). Owing to the optimality of $\hat{\boldsymbol{x}}$, it holds that

$$\boldsymbol{y}^\top \boldsymbol{A}\hat{\boldsymbol{x}} \geq \boldsymbol{y}^\top \boldsymbol{A}\bar{\boldsymbol{x}}.$$

With some re-arrangement, we have

$$\langle -\lambda\bar{\boldsymbol{x}}, \hat{\boldsymbol{x}} - \bar{\boldsymbol{x}} \rangle \leq \left\langle \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{y} - \lambda\bar{\boldsymbol{x}}, \hat{\boldsymbol{x}} - \bar{\boldsymbol{x}} \right\rangle.$$

Since $\|\bar{\boldsymbol{x}}\| = \|\hat{\boldsymbol{x}}\| = 1$, it follows that $\langle -\lambda\bar{\boldsymbol{x}}, \hat{\boldsymbol{x}} - \bar{\boldsymbol{x}} \rangle = \frac{\lambda}{2}\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|^2$. Thus, we obtain

$$\begin{aligned}\frac{\lambda}{2}\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|^2 &\leq \left\langle \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{y} - \lambda\bar{\boldsymbol{x}}, \hat{\boldsymbol{x}} - \bar{\boldsymbol{x}} \right\rangle \\ &\leq \left\| \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{y} - \lambda\bar{\boldsymbol{x}} \right\|_\infty \cdot \|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|_1 \\ &\leq \left\| \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{y} - \lambda\bar{\boldsymbol{x}} \right\|_\infty \cdot \sqrt{2k}\,\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|, \end{aligned} \tag{3.3}$$

where the second inequality follows from Hölder's inequality. For the third inequality, it follows from the facts that $\|\boldsymbol{v}\|_1 \leq \sqrt{\|\boldsymbol{v}\|_0} \cdot \|\boldsymbol{v}\|$ for all $\boldsymbol{v} \in \mathbb{R}^d$ and that $\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|_0 \leq 2k$. Dividing both sides by $\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|$ gives

$$\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\| \leq \frac{\sqrt{2k}}{\lambda}\left\| \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{y} - \lambda\bar{\boldsymbol{x}} \right\|_\infty. \tag{3.4}$$

It remains to upper bound the infinity norm.

Lemma 3.5 ([45]). *Assume (A1) and (A2). Then for all $i = 1, \ldots, n$, it holds that $\mathbb{E}[\boldsymbol{a}_i y_i] = \lambda_i \bar{\boldsymbol{x}}$.*

In particular, Lemma 3.5 implies $\frac{1}{n}\boldsymbol{A}^\top \boldsymbol{y} - \lambda\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{a}_i y_i - \mathbb{E}[\boldsymbol{a}_i y_i])$. Hence, we can apply Hoeffding's inequality to show that the sum of independent random variables concentrates around its mean with high probability. This is formally stated below.

Lemma 3.6. *Consider the observation model (2.1). Assume (A1), (A2) and (A3). With probability at least $1 - d^{-10}$ (over the random draw of $\boldsymbol{A}$) it holds that*

$$\left\| \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{y} - \lambda\bar{\boldsymbol{x}} \right\|_\infty \leq \mathrm{C}\sqrt{\frac{\log d}{n}}$$

for some absolute constant $\mathrm{C} > 0$.

Applying the inequality of the above lemma, we obtain Theorem 3.1. □

Remark 3.7 (Decomposition of Estimation Error)**.** As can be seen from the proof, if $\bar{\boldsymbol{x}}$ belongs to a generic constraint set $\mathcal{K}$, then we can tailor our analysis as follows. First, we will solve

$$\max_{\boldsymbol{x}} \ \boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{x}, \quad \text{s.t. } \boldsymbol{x} \in \mathcal{K}.$$

We present the closed-form solution for certain $\mathcal{K}$ (a solution for general $\mathcal{K}$ is hard to derive). Let $\boldsymbol{v} = \boldsymbol{A}^\top \boldsymbol{y}$.

- $\mathcal{K} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_0 \leq k, \|\boldsymbol{x}\| = 1, \boldsymbol{x} \geq \boldsymbol{0}\}$. Let $S = \{i : v_i > 0\}$ and $m = |S|$. If $m > k$, redefine S as the index set of the k largest elements of $\boldsymbol{v}$. The global optimum $\hat{\boldsymbol{x}}$ for such $\mathcal{K}$ is given as follows: if $m = 0$, then $\hat{\boldsymbol{x}}$ is the ith standard basis vector where i is the index of the largest entry of $\boldsymbol{v}$; otherwise $\hat{\boldsymbol{x}} = \boldsymbol{v}_S / \|\boldsymbol{v}_S\|$ where $\boldsymbol{v}_S$ is obtained by setting all elements of $\boldsymbol{v}$ outside S to zero.
- $\mathcal{K} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_0 \leq k, \boldsymbol{x} \in \{-1, 0, 1\}^d\}$. Let $S = \mathrm{supp}\,(\boldsymbol{v}, k)$. Then $\hat{\boldsymbol{x}} = \mathrm{sign}\,(\boldsymbol{v}_S)$.

Now suppose that we are able to obtain a global optimum $\bar{\boldsymbol{x}}$. Then following the same reasoning, (3.3) becomes

$$\frac{\lambda}{2}\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|^2 \leq \left\| \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{y} - \lambda\bar{\boldsymbol{x}} \right\|_\infty \cdot \rho_{\mathcal{K}}\,\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|, \tag{3.5}$$

where

$$\rho_{\mathcal{K}} := \max_{\boldsymbol{z} \in \mathcal{K} - \mathcal{K}} \frac{\|\boldsymbol{z}\|_1}{\|\boldsymbol{z}\|} \tag{3.6}$$

is the restricted induced norm which is completely characterized by the signal structure $\mathcal{K}$. While $\rho_{\mathcal{K}} \leq \sqrt{\|z\|_0}$ for general $\mathcal{K}$, it is possible to show an improved bound for specific $\mathcal{K}$. For instance, if $\mathcal{K}$ is the set of s-sparse α-strongly-decaying signals [13] for some $\alpha \geq 2$, then $\rho_{\mathcal{K}}$ acts as an absolute constant, which implies that the sample complexity in Theorem 3.1 can be improved to $O(\log d)$. On the other hand, Lemma 3.6 tells us that the infinity norm in (3.5) is oblivious of $\mathcal{K}$ (but depends on the distribution of $\boldsymbol{A}$ and observation model). Therefore, if one is interested in a sensing matrix $\boldsymbol{A}$ with heavy-tailed distributions, it suffices to derive a new bound as what we did in Lemma 3.6.

3.4 COMPARISON TO PRIOR WORKS

We first compare with two of the most important works on 1-bit CS [24, 25]. Both of them proposed to recover a signal with a generic structure $\mathcal{K}$. Regarding theoretical guarantee, [24, Theorem 1.1] implies that to obtain $\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\| \leq \epsilon$, the sample size $n = O\left(\epsilon^{-4} k \log(d/k)\right)$ which is worse than what we derived in Theorem 3.1 in terms of the dependence on ϵ. Though [25, Theorem 2.1] improved the sample complexity to $n = O\left(\epsilon^{-2} k \log(d/k)\right)$, their proof is technically involved (e.g. they applied high-dimensional geometric arguments)

even specifying $\mathcal{K}$ as the sparsity constraint. In contrast, we depart from their theoretical analysis and reach the same guarantee with fundamental facts in probability theory. Our analysis (Remark 3.7) shows the decomposition of estimation error which is useful to develop new results for different sensing schemes and signal structures.

In the seminal work of [20], a lower bound on the statistical error of sparse approximation was established, and a non-convex program was proposed to achieve the lower bound up to some logarithmic factor. Notably, their guarantee is uniform, meaning that a single draw of the sensing vectors ensures recovery of all sparse signals, whereas our result is specified to a particular signal. Their sample complexity is $n = O(\epsilon^{-1} k \log d)$, which has a better dependence on ϵ than this work. However, due to the non-convex nature, it is not clear how to solve their program in polynomial time. As a matter of fact, an iterative algorithm was devised with compelling performance in practical problems, but it lacks theoretical backend on the convergence behavior.

It is worth noting that [46, Theorem 4.3] claimed a sample complexity bound of $O(s)$ for certain type of $\mathcal{K}$. However, this may not be the true sample complexity since in order to fulfill their conditions, one needs $O(s \log d)$ samples.

4 EXTENSIONS

Theorem 3.1 offers near-optimal guarantee on recovering the direction of the signal $\bar{\boldsymbol{x}}$. In this section, we discuss when we can recover its support set, and even its magnitude under extra conditions. In this section, we focus on the elementary case that $\bar{\boldsymbol{x}}$ is s-sparse.

4.1 SUPPORT RECOVERY

We first describe when our estimator reliably detects the support of $\bar{\boldsymbol{x}}$. We would like point out that in some applications such as medical diagnosis, it is of crucial importance to discover the determinants of a disease (i.e. support recovery). Intuitively, a factor can be identified only when it has "sufficient" impact on the disease. This notion of significance is characterized by the following mathematical quantity:

$$\bar{\boldsymbol{x}}_{\min} := \min_{i \in \operatorname{supp}(\bar{\boldsymbol{x}})} |\bar{x}_i| . \tag{4.1}$$

Throughout the paper, we consider $\bar{\boldsymbol{x}}_{\min} \neq 0$, i.e. the signal of interest is non-zero. We utilize a well-known fact to derive the guarantee of support set of $\bar{\boldsymbol{x}}$.

Lemma 4.1. *For a given signal $\bar{\boldsymbol{x}}$, if*

$$\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| < \bar{\boldsymbol{x}}_{\min},$$

then it holds that $\operatorname{supp}(\bar{\boldsymbol{x}}) \subset \operatorname{supp}(\boldsymbol{x})$.

The lemma can be proved by algebra (see Appendix A.5), and has been widely used in the literature [37]. In allusion to Theorem 3.1 (that is, let the right-hand side therein be less than $\bar{\boldsymbol{x}}_{\min}$), we show that the support set of $\bar{\boldsymbol{x}}$ is contained in that of our estimate.

Theorem 4.2. *Assume the same conditions as in Theorem 3.1. Then* $\operatorname{supp}(\bar{\boldsymbol{x}}) \subset \operatorname{supp}(\hat{\boldsymbol{x}})$ *provided that*

$$\bar{\boldsymbol{x}}_{\min} > \frac{\mathrm{C}}{\lambda} \sqrt{\frac{k \log d}{n}}.$$

In particular, if we know exactly the sparsity of $\bar{\boldsymbol{x}}$, we have $\operatorname{supp}(\bar{\boldsymbol{x}}) = \operatorname{supp}(\hat{\boldsymbol{x}})$.

Remark 4.3. In the special case that $\bar{\boldsymbol{x}}$ is a binary vector, the above theorem indicates exact signal recovery under near-optimal sample complexity.

Remark 4.4. The proof of Lemma 4.1 essentially suggests that the minimum condition for support recovery is $\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|_\infty < \bar{\boldsymbol{x}}_{\min}$. Yet we conjecture that even with such condition Theorem 4.2 may not be significantly improved. Suppose that N samples suffice for support recovery of $\bar{\boldsymbol{x}}$. Consider a two-step scheme of sparse approximation: 1) recover support; 2) linear regression restricted on the obtained support set. Since the second step needs k samples, the total sample size is $N + k$. As lower bound of sparse approximation is $k \log(d/k)$ we must have $N \geq k \log(d/k)$.

4.2 NORM ESTIMATION

In this section we consider $\|\bar{\boldsymbol{x}}\| \leq R$ where R is known, and we hope to estimate the norm of $\bar{\boldsymbol{x}}$. In general, this is impossible in that the sign function will absorb the magnitude information. Thus, we shall make a further assumption for the data acquisition procedure:

$$y_i = \operatorname{sign}\left(\langle \boldsymbol{a}_i, \bar{\boldsymbol{x}} \rangle + b_i\right), \ \forall\, i = 1, \ldots, n, \tag{4.2}$$

where b_i are manually added noise which is known to us. The above observation model is equivalent to

$$y_i = \operatorname{sign}\left(\langle \boldsymbol{a}_i', \bar{\boldsymbol{x}}' \rangle\right), \ \forall\, i = 1, \ldots, n,$$

where

$$\boldsymbol{a}_i' = \begin{pmatrix} \boldsymbol{a}_i \\ b_i / R \end{pmatrix}, \quad \bar{\boldsymbol{x}}' = \frac{1}{\sqrt{\|\bar{\boldsymbol{x}}\|^2 + R^2}} \begin{pmatrix} \bar{\boldsymbol{x}} \\ R \end{pmatrix}.$$

Note that the norm of $\bar{\boldsymbol{x}}$ has been encoded into $\bar{\boldsymbol{x}}'$ and y_i bears the information of b_i, which together paves the way for norm estimation. Also, all the $\boldsymbol{a}_i'$ are i.i.d. standard Gaussian random vectors provided that

(A4) $b_i \sim N\left(0, R^2\right)$ for all $i = 1, \ldots, n$, they are mutually independent, and are independent from all $\boldsymbol{a}_i$.

Since $\bar{\boldsymbol{x}}'$ is an $(s+1)$-sparse signal, and has unit ℓ_2-norm, the estimation of $\bar{\boldsymbol{x}}'$ from (4.2) can be reduced to that of $\bar{\boldsymbol{x}}$ from (1.2) through the following augmented program:

$$\max_{\boldsymbol{x}' \in \mathbb{R}^{d+1}} \ \boldsymbol{y}^\top \boldsymbol{A}' \boldsymbol{x}', \quad \text{s.t. } \|\boldsymbol{x}'\|_0 \le k, \ \|\boldsymbol{x}'\| = 1, \quad (4.3)$$

where $\boldsymbol{A}' = (\boldsymbol{a}_1', \ldots, \boldsymbol{a}_n')^\top$ and $k \ge s+1$. Proposition 2.1 immediately gives the global optimum of (4.3):

$$\hat{\boldsymbol{x}}' = \mathcal{H}_k\left((\boldsymbol{A}')^\top \boldsymbol{y}\right) / \left\|\mathcal{H}_k\left((\boldsymbol{A}')^\top \boldsymbol{y}\right)\right\|. \quad (4.4)$$

With the optimal solution, we are able to evaluate the magnitude of $\bar{\boldsymbol{x}}$, as presented in the following theorem.

Theorem 4.5. *Consider the statistical model (4.2). Assume (A1), (A2), (A3) and (A4). Further assume that $\|\bar{\boldsymbol{x}}\| \le R$, $\|\bar{\boldsymbol{x}}\|_0 = s$, and $k \ge s+1$. Write $\hat{\boldsymbol{x}}'$ as*

$$\hat{\boldsymbol{x}}' = \frac{(\boldsymbol{x}_0;\ t_0)}{\sqrt{\|\boldsymbol{x}_0\|^2 + t_0^2}}.$$

With probability at least $1 - d^{-10}$ over the random draw of $\boldsymbol{A}$, we either have

$$\left\| \frac{R}{\|\boldsymbol{x}_0\|} \boldsymbol{x}_0 - \bar{\boldsymbol{x}} \right\| \le \frac{\mathrm{C} \cdot R}{\lambda} \sqrt{\frac{k \log d}{n}}$$

in the case $t_0 = 0$ (thus $\|\boldsymbol{x}_0\| \ne 0$); or have

$$\left\| \frac{R}{t_0} \boldsymbol{x}_0 - \bar{\boldsymbol{x}} \right\| \le \frac{\mathrm{C} \cdot R}{\lambda} \sqrt{\frac{k \log d}{n}}$$

in the case $t_0 \ne 0$.

Remark 4.6. Interestingly, our theorem implies that when $t_0 = 0$ (i.e. the manually added noise does not play a role in estimation), $\frac{R}{\|\boldsymbol{x}_0\|}\boldsymbol{x}_0$ is a good approximation to $\bar{\boldsymbol{x}}$. In other words, the norm of $\bar{\boldsymbol{x}}$ is extremely close (or even equal) to R. This result can be interpreted from another perspective: once we know the norm of $\bar{\boldsymbol{x}}$ in advance, say $\|\bar{\boldsymbol{x}}\| = R$, it is easy to see that $\bar{\boldsymbol{x}}/R$ is feasible to program (2.2), and Theorem 3.1 already implies

$$\left\| \hat{\boldsymbol{x}} - \frac{\bar{\boldsymbol{x}}}{R} \right\| \le \frac{\mathrm{C}}{\lambda} \sqrt{\frac{k \log d}{n}},$$

which is precisely the first inequality in Theorem 4.5 by noticing $\boldsymbol{x}_0 = \hat{\boldsymbol{x}}$ when $t_0 = 0$. Namely, there is no need to consider model (4.2).

The analysis for $t_0 \ne 0$ follows from Theorem 4 in [21], which showed that as soon as we have a good approximation to the direction of $\bar{\boldsymbol{x}}$, it is possible to estimate the magnitude if the conditions in Theorem 4.5 are fulfilled. However, the scenario $t_0 = 0$ was not addressed therein, and we make efforts to draw a formal analysis. See Appendix A.6 for a full proof.

4.3 MODEL MISSPECIFICATION

In our previous discussion, we always assume that the sparsity k is equal to, or greater than the true sparsity s, i.e. the true signal is contained in the feasible set. However, sometimes we may choose $k < s$ in that we are not aware of s. As a result, recovery of $\bar{\boldsymbol{x}}$ is impossible but we can still hope to approximate its k largest (i.e. most important) components to a high precision.

We now elaborate the new results under the misspecified model. One notable fact is that even when $k < s$, the normalized sparse vector $\mathcal{H}_k(\bar{\boldsymbol{x}})$ is feasible to the non-convex program (2.2). Therefore, for sparse approximation we may apply the same induction and obtain the following.

Theorem 4.7. *Assume the same conditions as in Theorem 3.1 but $k < s$. With probability at least $1 - d^{-10}$,*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{z}\| \le \frac{\mathrm{C}}{\lambda} \sqrt{\frac{k \log d}{n}} + \sqrt{2k} \|\boldsymbol{z} - \bar{\boldsymbol{x}}\|_\infty,$$

where $\boldsymbol{z} := \frac{\mathcal{H}_k(\bar{\boldsymbol{x}})}{\|\mathcal{H}_k(\bar{\boldsymbol{x}})\|}$.

The proof can be found in Appendix A.4. The second term on the right-hand side is the price we pay for model misspecification, and it vanishes as soon as $k \ge s$. It is worth mentioning that it depends exclusively on the nature of the signal $\bar{\boldsymbol{x}}$ rather than on the data acquisition procedure. If the s-sparse signal $\bar{\boldsymbol{x}}$ has a light tail, i.e. the first k components dominate the magnitude, then $\hat{\boldsymbol{x}}$ behaves as a good estimate. To see this, let us write $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_s, 0, \ldots, 0)$ in descending order (according to the magnitude of the elements). In this way $\mathcal{H}_k(\bar{\boldsymbol{x}}) = (\bar{x}_1, \ldots, \bar{x}_k, 0, \ldots, 0)$. Let us denote $\alpha = \|\mathcal{H}_k(\bar{\boldsymbol{x}})\|$ for now. It follows that

$$\|\boldsymbol{z} - \bar{\boldsymbol{x}}\|_\infty = \max\left\{ \left(\frac{1}{\alpha} - 1\right) |\bar{x}_1|, \ |\bar{x}_{k+1}| \right\}. \quad (4.5)$$

If the k leading components dominate the remaining, then $\alpha \approx 1$ and $|\bar{x}_{k+1}| \approx 0$ hold simultaneously. As a consequence, $\|\boldsymbol{z} - \bar{\boldsymbol{x}}\|_\infty$ is close to zero, under which Theorem 4.7 implies that we can accurately identify the principal direction of $\bar{\boldsymbol{x}}$ given sufficient measurements.

Now we move on to discuss the support recovery of the top k elements. In light of Lemma 4.1, the support of $\hat{\boldsymbol{x}}$ is consistent with that of $\mathcal{H}_k(\bar{\boldsymbol{x}})$ as soon as the sample size $n = O\left((\lambda |\bar{x}_k|)^{-2} k \log d\right)$ and

$$\sqrt{2k} \|\boldsymbol{z} - \bar{\boldsymbol{x}}\|_\infty < |\bar{x}_k|.$$

Consider in (4.5) that the infinite norm is given by $(1/\alpha - 1)|\bar{x}_1|$. It follows that a sufficient condition for support recovery is

$$|\bar{x}_{k+1}| \le \left(\frac{1}{\alpha} - 1\right) |\bar{x}_1| < \frac{|\bar{x}_k|}{\sqrt{2k}}.$$

The first inequality, which upper bounds $|\bar{x}_{k+1}|$, indicates that for our purpose, the elements outside of the support of interest cannot be too large. The second inequality, which lower bounds $|\bar{x}_k|$, tells that those inside of the support need to have sufficient magnitude.

Lastly, as we discussed in Section 4.2, a good approximation to the direction of $\mathcal{H}_k(\bar{\boldsymbol{x}})$ implies a good estimation of the norm. We remark that this observation holds for a misspecified model as well.

5 EXPERIMENTS

This section is dedicated to examining the statistical error rate and robustness of the our estimator. We focus on the noiseless model (1.2), and will compare with the Lasso estimator [25] which showed state-of-the-art performance.

Settings. We implement our algorithm and the one of [25] in Matlab 2018, and perform all the experiments on a single server which has two 3.2 GHz Intel Xeon processors, each of which has 8 cores. The sensing vectors are chosen as i.i.d. standard Gaussian. For the s-sparse signal $\bar{\boldsymbol{x}}$, we first randomly choose the support set in a uniform manner, and draw each non-zero element from an i.i.d. uniform distribution over the interval $[-1000, 1000]$. For each experiment to be presented, we generate 100 i.i.d. copies of the true signal $\bar{\boldsymbol{x}}$ and report the averaged performance. If not specified, we always set $k = s$.

Sparse Approximation. We study how the estimation error $\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|$ varies with the sample size n. We fix the sparsity $s = 20$, and consider the dimension $d = 2000$ and $d = 10,000$. For each d, we increase n from 10 to $10,000$, and for each configuration of (d, n) we generate the sensing vectors as aforementioned. The error curves are plotted in Figure 1. It shows that the reconstruction error of our method decays much faster than [25], and it turns out that their estimator incurs large error when the dimension is increased. In contrast, our algorithm consistently produces accurate estimate.

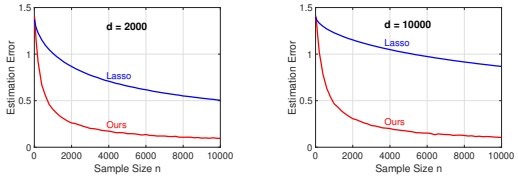

Figure 1: **Approximation error against sample size.**

Support Recovery. We use the same setting as in sparse approximation, but plot the cardinality of the symmetric difference between $\mathrm{supp}(\hat{\boldsymbol{x}})$ and $\mathrm{supp}(\bar{\boldsymbol{x}})$. Since [25] may output arbitrarily dense solution, for fair comparison we apply hard thresholding with the true sparsity to their original estimate. Note that if their original estimate already recovers the support, our post-processing does not hurt it. In this way, the maximum error of both methods is 40 (since $s = 20$), and zero error indicates perfect support recovery. Again, we observe in Figure 2 that our estimator outperforms the state-of-the-art.

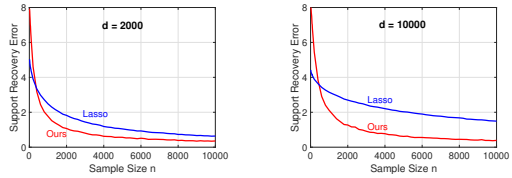

Figure 2: **Support recovery error against sample size.**

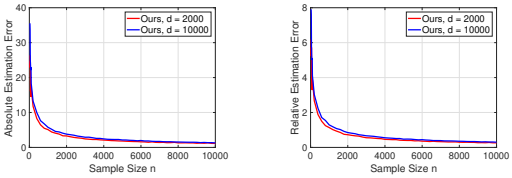

Figure 3: **Norm estimation error against sample size.**

Norm Estimation. We use the estimate presented in Theorem 4.5 to approximate the signal $\bar{\boldsymbol{x}}$ with $\|\bar{\boldsymbol{x}}\| \le R$. We choose $R = 2\|\bar{\boldsymbol{x}}\|$, and illustrate the absolute and relative errors in Figure 3. Note that we did not compare with [25] because norm estimation was not addressed therein. The figure shows that once we have sufficient samples, it is possible to accurately evaluate the norm with the data collection scheme in Section 4.2.

Model Misspecification. In Figure 4 we record the estimation error (in logarithmic scale) of the direction of $\mathcal{H}_k(\bar{\boldsymbol{x}})$ when $k < s$. We observe that even in this challenging scenario our estimate possesses a small error. This matches our theoretical guarantee that our estimator is resilient to model misspecification. It is noticeable that the curve of $k = 1$ bumps more often than others. The reason is that when $k = 1$, the problem boils down to hitting the unique non-zero position, and the estimation error per signal is either 0 or 2 (which results in the bumping phenomenon).

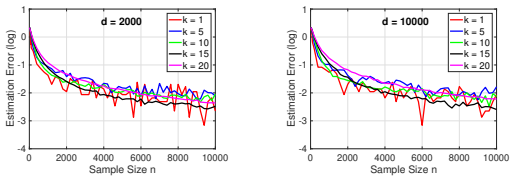

Figure 4: **Approximation error of the leading k components of our method under model misspecification.** The true sparsity is 20.

6 CONCLUSION

In this paper, we have studied an efficient estimator for recovering a sparse signal from its binary measurements. On the computational side, the estimate can be obtained by a one-step hard thresholding operator which enjoys economic computational and memory cost. On the statistical side, we have shown that the estimation error matches the information-theoretic lower bound up to some logarithmic factor. We have also extended our results to support recovery and norm estimation, and have proved near-optimal error rate in these scenarios. For the estimation of all the three facets of a sparse signal, we have offered rigorous theoretical evidence that our estimator is robust to model misspecification. Finally, we have demonstrated through a comprehensive set of experiments that the practical performance of our estimator matches perfectly our analysis.

ACKNOWLEDGEMENTS

We thank Jing Wang and Chicheng Zhang for insightful discussions, and thank the anonymous reviewers for helpful comments. This work is supported by the startup funding from Stevens Institute of Technology.

References

[1] J. Acharya, A. Bhattacharyya, and P. Kamath. Improved bounds for universal one-bit compressive sensing. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 2353–2357, 2017.

[2] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin. One-bit compressed sensing with non-Gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

[3] E. Arias-Castro, E. J. Candès, and M. A. Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2013.

[4] P. Awasthi, M. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM*, 63(6):50:1–50:27, 2017.

[5] R. G. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wootters. Exponential decay of reconstruction error from binary measurements of sparse signals. *IEEE Transactions on Information Theory*, 63(6):3368–3385, 2017.

[6] S. A. Bhaskar. Probabilistic low-rank matrix completion from quantized measurements. *Journal of Machine Learning Research*, 17(60):1–34, 2016.

[7] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[8] P. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *Proceedings of the 42nd Annual Conference on Information Sciences and Systems*, pages 16–21, 2008.

[9] T. T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[10] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[11] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[12] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[13] M. A. Davenport and M. B. Wakin. Analysis of orthogonal matching pursuit using the restricted isometry property. *IEEE Transactions on Information Theory*, 56(9):4395–4401, 2010.

[14] S. Dirksen and S. Mendelson. Non-gaussian hyperplane tessellations and robust one-bit compressed sensing. *CoRR*, abs/1805.09409, 2018.

[15] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[16] J. Fan and Y. Fan. High dimensional classification using features annealed independence rules. *Annals of Statistics*, 36(6):2605–2637, 2008.

[17] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[18] S. Gopi, P. Netrapalli, P. Jain, and A. V. Nori. One-bit compressed sensing: Provable support and vector recovery. In *Proceedings of the 30th International Conference on Machine Learning*, pages 154–162, 2013.

[19] X. Huang and M. Yan. Nonconvex penalties with analytical solutions for one-bit compressive sensing. *Signal Processing*, 144:341–351, 2018.

[20] 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 Transactions on Information Theory*, 59(4):2082–2102, 2013.

[21] K. Knudson, R. Saab, and R. Ward. One-bit compressive sensing with norm estimation. *IEEE Transactions on Information Theory*, 62(5):2748–2758, 2016.

[22] Z. Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

[23] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer US, 2004.

[24] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

[25] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

[26] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

[27] J. Shen, P. Awasthi, and P. Li. Robust matrix completion from quantized observations. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 397–407, 2019.

[28] J. Shen and P. Li. On the iteration complexity of support recovery via hard thresholding pursuit. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3115–3124, 2017.

[29] J. Shen and P. Li. Partial hard thresholding: Towards a principled analysis of support recovery. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pages 3127–3137, 2017.

[30] J. Shen and P. Li. A tight bound of hard thresholding. *Journal of Machine Learning Research*, 18(208):1–42, 2018.

[31] J. Shen and C. Zhang. Attribute-efficient learning of halfspaces with malicious noise: Near-optimal label complexity and noise tolerance. *CoRR*, abs/2006.03781, 2020.

[32] C. Thrampoulidis and A. S. Rawat. Lifting high-dimensional non-linear models with gaussian regressors. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 3206–3215, 2019.

[33] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[34] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

[35] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[36] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *CoRR*, abs/1011.3027, 2010.

[37] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[38] J. Wang, J. Shen, and P. Li. Provable variable selection for streaming features. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5158–5166, 2018.

[39] J. Wang, J. Shen, P. Li, and H. Xu. Online matrix completion for signed link prediction. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, pages 475–484, 2017.

[40] S. Yan and C. Zhang. Revisiting Perceptron: Efficient and label-optimal learning of halfspaces. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pages 1056–1066, 2017.

[41] X.-T. Yuan, P. Li, and T. Zhang. Exact recovery of hard thresholding pursuit. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, pages 3558–3566, 2016.

[42] C. Zhang. Efficient active learning of sparse halfspaces. In *Proceedings of the 31st Annual Conference On Learning Theory*, pages 1856–1880, 2018.

[43] C. Zhang, J. Shen, and P. Awasthi. Efficient active learning of sparse halfspaces with arbitrary bounded noise. *CoRR*, abs/2002.04840, 2020.

[44] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942, 2010.

[45] L. Zhang, J. Yi, and R. Jin. Efficient algorithms for robust one-bit compressive sensing. In *Proceedings of the 31st International Conference on Machine Learning*, pages 820–828, 2014.

[46] R. Zhu and Q. Gu. Towards a lower sample complexity for robust one-bit compressed sensing. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 739–747, 2015.