Exact recoverability from dense corrupted observations via L_1 minimization

Nam H. Nguyen, *Student Member, IEEE,* and Trac D. Tran, *Senior Member, IEEE*

Abstract— **This paper confirms a surprising phenomenon first observed by Wright *et al.* [1] [2] under different setting: given m highly corrupted measurements $y = A_{\Omega\bullet} x^\star + e^\star$, where $A_{\Omega\bullet}$ is a submatrix whose rows are selected uniformly at random from rows of an orthogonal matrix A and $e^\star$ is an unknown sparse error vector whose nonzero entries may be unbounded, we show that with high probability ℓ_1-minimization can recover the sparse signal of interest $x^\star$ exactly from only $m = C\mu^2 k(\log n)^2$ where k is the number of nonzero components of $x^\star$ and $\mu = n \max_{ij} A_{ij}^2$, even if nearly 100% of the measurements are corrupted. We further guarantee that stable recovery is possible when measurements are polluted by both gross sparse and small dense errors: $y = A_{\Omega\bullet} x^\star + e^\star + \nu$ where ν is the small dense noise with bounded energy. Numerous simulation results under various settings are also presented to verify the validity of the theory as well as to illustrate the promising potential of the proposed framework.**

Index Terms— **Compressed sensing, ℓ_1-minimization, sparse signal recovery, discrete Fourier transform, (weak) restricted isometry, random matrix, dense error correction.**

I. INTRODUCTION

Compressed sensing (CS) has been rigorously studied over a past few years as a revolutionary signal sampling paradigm [3], [4], [5]. According to CS, a k-sparse signal $x^\star \in \mathbb{R}^n$ is measured through a set of linear projections $y_i = \langle a_i, x^\star \rangle$, $i = 1, ..., m$, in which vectors $a_i \in \mathbb{R}^n$ form a matrix A of size $m \times n$. The intriguing CS framework advocates the collection of significantly fewer measurements than the ambient dimension of the signal ($m < n$). To reconstruct $x^\star$, a standard ℓ_1-minimization is proposed to solve the inverse problem

$$\min_x \|x\|_1 \qquad \text{subject to} \qquad y = Ax. \qquad (1)$$

It has been well known in the literature that if A obeys Restricted Isometry Property (RIP) [6], [7] - a property essentially implies that every subset of k or fewer columns of A is approximately an orthogonal system, then the linear program in (1) is able to faithfully recover $x^\star$. This RIP condition has been proven to hold for many types of random measurement matrices [8], [9]. For example, random Gaussian or Bernoulli matrices satisfy RIP with high probability as long as the

number of measurements m is on the order of $k \log n$ [8], whereas the sub-orthogonal matrix $A_{\Omega\bullet}$ sampled uniformly from an orthogonal matrix A obeys RIP with high probability when m is on the order of $k \log^4 n$ [9].

In many practical applications, we are often interested in situations in which measurements are contaminated by noise. Mathematically, we often observe

$$y = Ax^\star + e^\star,$$

where $e^\star \in \mathbb{R}^m$ is the vector noise. To reconstruct $x^\star$ from the observation vector y, we minimize the following convex program

$$\min_x \|x\|_1 \qquad \text{subject to} \qquad \|y - Ax\|_2 \le \sigma, \qquad (2)$$

where σ is upper bound of the noise level $\|e^\star\|_2$, which assumes to be known. It has been shown in [6], [10], [11], [12] that if A satisfies RIP and σ is not too large, then by the same amount of measurements as above, solution $\widehat{x}$ of (2) does not depart too far from the optimal solution $x^\star$. In particular, the authors of [6] proved that the reconstruction error proportionally grows with σ as $\|\widehat{x} - x^\star\|_2 \le C\sigma$, where C is a small numerical constant.

This result is elegant when the noise level is low. However, as the noise energy gets larger, $\widehat{x}$ might be unexpectedly very different from $x^\star$. This implies that even a single grossly corrupted measurement may produce $\widehat{x}$ arbitrarily far from the true solution. Unfortunately, gross errors and irrelevant measurements are now ubiquitous in modern applications such as image processing, sensor network, where certain number of measurements may be severely corrupted due to occlusions, sensor failures, transmission error, etc [13], [1], [2]. These examples motivate us to consider a new problem in which we aim to recover a sparse vector $x^\star$ from highly corrupted measurements, $y = Ax^\star + e^\star$. In contrast to previous approaches [6], [10], [11], [12] where only small dense noise term $e^\star$ is considered, in this paper, entries of $e^\star$ can have arbitrarily large magnitude, and their support is assumed to be sparse but unknown. The underlying model has been previously developed by Wright *et al.* [1]. Motivated from the face recognition problem, in which sparse error appears due to a fraction of the query image y being occluded by glasses, hats, etc, the authors proposed to simultaneously minimize the ℓ_1-norm of both x and e,

$$\min_{x,e} \|x\|_1 + \|e\|_1 \qquad \text{subject to} \qquad y = Ax + e. \qquad (3)$$

where columns of matrix A are associated with training images. To analyze the model, they assume A obeys the Gaussian

distribution [2]. That is, entries of A are i.i.d $\mathcal{N}(0, 1/m)$ Gaussian random variables.

As pointed out by Candès and Romberg [14] and Do *et al.* [15], in compressed sensing, completely random measurement matrices might not be relevant in many practical applications. First, we may not be allowed to control measurement matrices. For instance, in MRI or tomography, due to the acquisition system, measurements are inherently frequency based. The second drawback is computationally expensive and memory buffering due to their completely unstructured nature. These weaknesses prevent these fully random sensing matrices from being applied to applications in which both acquisition system (or encoder) and reconstruction system (or decoder) are required to have low complexity and fast implementation.

In this paper, we extend the analysis to a special class of measurement matrices which are constructed from an orthogonal matrix A. Let Ω be a subset of indices of $\{1, ..., n\}$; and measurement matrix $A_{\Omega\bullet}$ is constructed from rows of A associated with indices in Ω. The observation vector y is now obtained by

$$y = A_{\Omega\bullet}x^\star + e^\star, \qquad (4)$$

where we assume that signal $x^\star$ and error $e^\star$ are sparse vectors whose supports are T and S, respectively. These suborthogonal measurement matrices have been carefully studied in the literature such as the partial Fourier ensemble [14] and structurally random matrix (SRM) [15] as a promising replacement for fully random Gaussian/Bernulli sensing matrices. However, so far, none of the previous work guarantees stable reconstruction under highly corrupted sparse error or a combination of both large sparse error and small dense noise. This is our most significant technical contribution.

To recover $x^\star$ and $e^\star$, we propose to solve the following extended ℓ_1-minimization

$$\min_{x,e} \|x\|_1 + \lambda \|e\|_1 \qquad \text{subject to} \qquad y = A_{\Omega\bullet}x + e, \quad (5)$$

where $\lambda > 0$ is a controlled parameter that balance the two ℓ_1-norm terms.

Surprisingly, with an appropriate choice of λ, this simple linear program (5) can assure the exact recovery both $x^\star$ and $e^\star$ exactly, even when the sparsity of $x^\star$ grows almost linearly in the dimension of signal and the errors in $e^\star$ are up to a constant fraction of all the entries. This observation will be confirmed via rigorously mathematical justifications as well as extensive simulations in the next few sections.

A. Motivational applications

There are many important applications in which the observations of interest can be modeled as a linear projection of a sparse signal plus sparse error. Before shifting to the presentation of our main results, we briefly introduce several applications and show how well they fit into our underlying model of interest

- *Image inpainting.* Given an image Y with missing/corrupted pixels, we would like to reconstruct the original image by filling in lost information [16]. If we assume that errors are indicated by a matrix E

whose nonzero-value entries are associated with the missing/corrupted pixels, then Y can be decomposed into two components: the original image B and sparse noise E. In image inpainting, the key hypothesis frequently made to guarantee satisfactory performance is that Y has to be sparsely represented by a few coefficients over an overcomplete dictionary D [17], [16]. This dictionary is typically a concatenation of orthogonal transformations, e.g. wavelet, Fourier, DCT or is learned from a set of training images. By denoting y, b and e as vectorized versions of matrices Y, B and E, we have a mathematical representation, $y = Dx + e$, where x is the sparse coefficient vector. As opposed to previous works in which locations of missing entries are often required to be known in advance, here we do not need to make any of such assumptions in our model. Rather, utilizing the optimization in (5), we let the algorithm guess both the noisy locations and their magnitudes.

- *Compressed sensing for networked data.* In sensor networks, the goal is to design a low-power system but still guarantee reliability in transmission. In this setting [18], each sensor collects information of a signal or object $x^\star$ by simply projecting $x^\star$ onto row vectors a_i of a sensing matrix $A \in \mathbb{R}^{m \times n}$, $b_i = \langle a_i, x^\star \rangle$. As suggested in [18], rather than realizing A in a completely random manner, it is simpler and less computational complex to utilize a matrix A that we can exploit fast implementation and avoid expensive memory buffering such as DCT, Hadarmard or the Fourier transform.

  After having gathered all the data, these sensors send measured information y_i to their neighbors or a central hub for analysis and processing. However, due to the fact that sensors are low cost, it is highly likely that some sensors might fail in collecting data or producing measurements that are not well protected before transmission. This implies that some measurements may be severely corrupted by two types of errors:

  $$y = Ax^\star + e^\star + \nu,$$

  where $e^\star$ is the sparse error, whose entry magnitudes in the support can be arbitrarily large and ν is dense noise with bounded energy σ. To recover both $x^\star$ and $e^\star$, we propose to solve

  $$\min_{x,e} \|x\|_1 + \lambda \|e\|_1 \quad \text{s.t.} \quad \|y - Ax - e\|_2 \leq \sigma.$$

- *Joint source-channel coding.* One potential application of CS is simultaneous joint source channel coding [19], [20]. In contrast to conventional approach where source data $x^\star \in \mathbb{R}^n$ is initially encoded to remove redundancy, then channel-coded for error protection. In CS, $x^\star$ is encoded by a simple linear projection $y = Ax$. In [19], to protect the channel, the authors proposed to use more measurements than the optimal value that CS can recover accurately. In order to retrieve $x^\star$ under channel error, we need to know the probabilistic model of corrupted entries, which is usually unavailable in practice. We believe that ours is a more accessible and more robust approach in recovering such signal.

B. Notations and organization of the paper

We briefly introduce some notations that will be used throughout the paper. We denote x_T as a vector whose entries are selected from the index set $T \in \{1, ..., n\}$ of $x \in \mathbb{R}^n$. Let Ω be a subset of $\{1, ..., n\}$, we denote $A_{\Omega \bullet}$ as a submatrix of A, whose rows are taken from Ω. Similarly, $A_{\Omega T}$ denotes a submatrix of A, containing rows indexed by Ω and columns indexed by T. Further, we reserve the two index sets T and S for signal ($x^\star$) and error ($e^\star$) supports. The sparsity level of $x^\star$ and $e^\star$ are $k = |T|$ and $s = |S|$, respectively. For a vector x, $\text{sgn}(x)$ represents the sign of x componentwise.

We will use several standard vector and matrix norms, which we simply present here for completeness. For $x \in \mathbb{R}^n$, $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ is the ℓ_1-norm, $\|x\|_2$ is the ℓ_2-norm and $\|x\|_\infty = \max_i |x_i|$ is the l_∞-norm. For matrices B, we only use the spectral norm, denoted by $\|B\|$.

We denote by $C, C_1, c, c_1, ...$ positive absolute constants. Finally, when we say that an event occurs with high probability, we mean the occurring probability of the event is at least $1 - cn^{-1}$.

The remainder of this paper is structured as follows. The main results are introduced in Section II. Our proof structure is described Section III. Supporting results are subsequently presented in Sections IV and V. Section VI compares our results with the oracle in which we know in advance the locations of signal and error support. We demonstrate the consistency of our results via extensive simulation in Section VII. Finally, Section VIII summarizes this paper and makes some closing remarks.

II. MAIN RESULTS

A. Sparse model

We begin by studying the easier problem where signal $x^\star$ is perfectly k-sparse and observation vector y is also corrupted by sparse error. A more difficult problem with non-sparse signal $x^\star$ and y being corrupted by both sparse and dense noise will be subsequently investigated in this section. Toward the end, we denote the sparsity indices of $x^\star$ and $e^\star$ as k and s and introduce the (k, s)-*sparse model* defined as follows:

- Signs of $x^\star$ at the support T is independently and equally likely to be 1 or −1.
- Support S of $e^\star$ is uniformly distributed among all sets of cardinality s in Ω.

The random assumption on the sign of $x^\star$ at support T is typical in compressed sensing [14]. This assumption is a sufficient rather than necessary condition and is employed for the convenience of our proof only. Indeed, by sacrificing a factor of $\log n$ to the number of measurements, we can establish similar results when the signs of $x^\star$ arbitrary. We refer the interested readers to a recent paper [21] for more details.

B. Exact recovery as measurements are corrupted by sparse noise

Theorem 1. *Let $x^\star$ be a fixed vector in $\mathbb{R}^n$ and A be an $n \times n$ orthonormal matrix ($A^*A = I$) with $|A_{ij}|^2 \leq \frac{\mu}{n}$,*

where $1 \leq \mu \leq n$, and assume that $(x^\star, e^\star)$ is taken from the (k, s)-sparse model. Suppose we observe m entries from the projection $Ax^\star$ with locations in Ω sampled uniformly at random and these entries are then corrupted by noise $e^\star$. Then there exist numerical constants c and C such that with probability at least $1 - cn^{-1}$, the convex program (5) with $\lambda \sim \sqrt{\frac{n}{\mu m \log n}}$ correctly recovers both the signal and the error (i. e. $\widehat{x} = x^\star$ and $\widehat{e} = e^\star$), provided that

$$m \geq C\mu^2 k (\log n)^2 \qquad and \quad s \leq \gamma m, \qquad (6)$$

for any γ close to 0.9.

In other words, Theorem 1 asserts a surprising message: a sparse signal $x^\star$ can be faithfully recovered with probability converging to one from arbitrary and completely unknown corrupted patterns (as long as they are randomly distributed). We do not place any assumption on the magnitudes or signs of the nonzero entries of $e^\star$. In fact, its magnitude can be arbitrarily large. Theorem 1 is generic in the sense that it only requires signs of nonzero entries of $x^\star$ to be uniformly distributed; everything else is deterministic. We believe that the random assumption on the sign pattern is artificial and can be removed. Indeed, when A is a Fourier matrix, applying advanced techniques in [4], we are able to obtain Theorem 1 for all $x^\star$ supported on T. An interesting open problem is whether this result also holds for other orthogonal sensing matrices.

It is necessary to further clarify Theorem 1. First, higher probabilities of success (i.e. in the form $1 - cn^{-\beta}$ with $\beta \geq 1$) can be obtained at the expense of increasing the number of observations by a factor of β. Next, the theorem addresses that for a particular selection of Ω, exact recovery only holds for an arbitrary *fixed* sparse signal with high probability (as long as signs of such signal at its support are uniformly distributed). In other words, there is no uniform sparse signal recovery guaranteed here. In fact, in order to establish perfect recovery for all sparse signal, we might have to require certain stronger properties for matrix $A_{\Omega \bullet}$ such as RIP [6] or similar to RIP. As shown in [8], [9], $A_{\Omega \bullet}$ obeys RIP with high probability only if the number of measurements exceeds $Ck \log^4 n$, which is a far inferior requirement than our optimal value. By relaxing RIP, we are able to significantly reduce the amount of measurements needed and are still able to guarantee perfect recovery even when the data is highly corrupted.

Theorem 1 also implies that up to a $\log n$ factor from the optimal number of observations as in compressed sensing, we are able to precisely recover the signal in the presence of gross error. In the following theorem, we establish that by the same order of $k \log n$ measurements as in compressed sensing, ℓ_1-minimization is still able to recover precisely both spare signal and high-energy sparse noise. In particular, we draw an interesting relationship between signal sparsity, error sparsity and the parameter λ.

Theorem 2. *Let $x^\star$ be a fixed vector in $\mathbb{R}^n$ and A be an $n \times n$ orthonormal matrix ($A^*A = I$) with $|A_{ij}|^2 \leq \frac{\mu}{n}$, where $1 \leq \mu \leq n$ and assume that $(x^\star, e^\star)$ is taken from the (k, s)-sparse model. Suppose that we observe m entries from the projection*

$Ax^\star$ *with locations in* Ω *sampled uniformly at random and these entries are then further corrupted by noise* $e^\star$*. Then there exist numerical constants c, C_1 and C_2 such that with probability at least* $1 - cn^{-1}$*, the convex program in (5) with* $\lambda = \sqrt{\frac{1}{\gamma \log n} \frac{n}{\mu m}}$*, $\gamma \in (0,1)$ correctly recovers both the signal and the error (i. e. $\widehat{x} = x^\star$ and $\widehat{e} = e^\star$), provided that*

$$m \geq C_1 \mu^2 \max\{\frac{\gamma}{(1-\gamma)^2} k(\log n)^2, k \log n, (\log n)^2\} \quad (7)$$

$$and \quad s \leq C_2 \gamma m. \quad (8)$$

It is easy to check that Theorem 2 implies Theorem 1 by setting $\gamma = 0.9$, or equivalently $\lambda = \sqrt{n/0.9\mu m \log n}$. Later in the paper, we focus on establishing this theorem, then Theorem 1 will automatically follow.

We would like to note the significance of the parameter μ here: μ can be seen as the incoherence of the matrix A, which measure how concentrated or expanded rows of measurement matrix $A_{\Omega\bullet}$ are. Since A is orthonormal, the value of μ ranges between 1 and n. In the worse case scenario when rows of A are maximally concentrated, then $\mu = n$ and A is the identity matrix. It is clear in this case that we cannot retrieve $x^\star$ under a single gross error even if all n measurements (which is now the signal $x^\star$ itself) are observed. On the other hand when $\mu = 1$, entries of A are perfectly spread out and the number of measurements attains its optimally minimum value.

It can be seen that λ in (5) controls the balance between two terms: $\|x\|_1$ and $\|e\|_1$. Specifically, if a large value of λ is selected, we expect to recover the denser-support signal but under sparser error. On the other hand, a smaller choice of λ is better when the error is denser while the signal is sufficiently sparse. Theorem 2 mathematically indicates that it is actually the case. In particular, if γ is chosen to be $1/\log n$, then relying on only $m = Ck \log n$ measurements, linear (convex) programming (5) not only recovers the k-sparse $x^\star$ faithfully, it is also able to correctly identify the noise with arbitrary large magnitude as long as the noise sparsity is proportional to $m/\log n$. On the contrary, if we set γ close to one, then (5) can retrieve $x^\star$ whose support is $m/Ck(\log n)^2$ under error whose support is up to a constant fraction of all the measurements. In fact, the theorem gives a whole range of λ values, whose selection might rely on the prior information we can collect about the sparsity level of the signal as well as of the noise.

C. Stable recovery as measurements are corrupted by both dense and sparse errors

Our result in Theorem 2, although interesting, is limited to the case of noise being exactly sparse only. In practical applications, observations y are also often contaminated by dense noise, which can be either deterministic or stochastic. In this section, we investigate the model where observations are corrupted by both the unknown dense noise ν with small energy bound $\|\nu\|_2 \leq \sigma$ and the sparse noise $e^\star$, whose magnitudes of nonzero entries are arbitrarily large

$$y = A_{\Omega\bullet} x^\star + e^\star + \nu.$$

At first, for the ease of demonstrating our results as well as proving technique, we consider a particular situation where the observation y is only corrupted by dense error whose energy is bounded by σ. The problem is now to recovery $x^\star$ from noisy observation y, where

$$y = A_{\Omega\bullet} x^\star + \nu.$$

To recover $x^\star$, it has been well established that we need to minimize the following convex program

$$\min_x \|x\|_1 \qquad \text{subject to} \qquad \|b - A_{\Omega\bullet} x\|_2 \leq \sigma. \quad (9)$$

When the observation vector y is clean, Candès and Romberg [14] showed that the ℓ_1-minimization is able to recover $x^\star$ precisely. In this section, we extend their result and prove that even with imperfect observations y, the convex program is stable vis a vis perturbations. Particularly, the recovery error is bounded away by a factor of σ. To the best of our knowledge, this is the first robust recovery bound when measurements taken from suborthogonal matrices are corrupted by deterministic noise.

Theorem 3. *Under the same assumptions defined in Theorem 2 and provided that there exists a numerical constant C such that* $m \geq C\mu k \log n$*, for any perturbation ν with $\|\nu\|_2 \leq \sigma$, the solution $\widehat{x}$ to the convex program in (9) yields*

$$\|\widehat{x} - x^\star\|_2 \leq 8\sigma\sqrt{n(1 + 2n/m)} + 2\sigma. \quad (10)$$

Roughly speaking, Theorem 3 states that for a family of matrices A_Ω constructed from any unitary matrix A, minimizing the ℓ_1-norm stably recovers $\widehat{x}$ from just $O(\mu k \log n)$ measurements. A direct consequence of this theorem says that as σ comes closer to zero, the solution of (9) is exact, which coincides with Candès and Romberg's result [14]. Moreover, our result is established for any deterministic noise ν. While preparing this manuscript, we learned of an independent investigation of Candès and Plan [21] into this problem. They place stochastic assumptions on ν, e.g. ν obeys the Gaussian distribution, and thus the resulting error bound is improved.

A more challenging situation occurs when observations are not only contaminated by dense noise with small energy, but they are also corrupted by sparse noise with arbitrarily large magnitude. This model includes the previous settings in Theorems 2 and 3 as the particular cases:

$$y = A_{\Omega\bullet} x^\star + e^\star + \nu. \quad (11)$$

To successfully recover $x^\star$ (as well as $e^\star$), we propose to minimize the following convex program

$$\min_{x,e} \|x\|_1 + \lambda \|e\|_1 \qquad \text{s. t.} \qquad \|b - A_{\Omega\bullet} x - e\|_2 \leq \sigma \quad (12)$$

where σ is the bound of energy noise ν, assumed to be known.

Theorem 4. *Under the same assumptions defined in Theorem 2 and provided that there exist numerical constants C_1 and C_2 such that*

$$m \geq C_1 \mu^2 \max\{\frac{\gamma}{(1-\gamma)^2} k(\log n)^2, k \log n, (\log n)^2\} \quad (13)$$

$$and \quad s \leq C_2 \gamma m, \quad (14)$$

then the pair of solution $(\widehat{x}, \widehat{e})$ *to the convex program (12) obeys*

$$\|\widehat{x} - x^\star\|_2 + \|\widehat{e} - e^\star\|_2 \leq \frac{8(\lambda+1)\sigma}{\min\{1,\lambda\}} \sqrt{n\left(1 + \frac{4n}{m-s}\right)} + 2\sigma. \tag{15}$$

Theorem 4 is significant because it addresses that the convex program in (12) can reliably reconstruct the sparse signal even when the measurements are severely corrupted by both gross sparse and small dense errors. This is the situation that we most likely will encounter in practical applications. When the measurement y is not corrupted by the dense noise ν, the signal can be reconstructed perfectly, regardless of how large the sparse noise is, as previously mentioned in Theorem 2. In addition, we will demonstrate in Section 5 that this reconstruction error is optimal up to a $\sqrt{n}$ factor as compared to the oracle situation in which locations of T nonzero entries of signal $x^\star$ as well as S nonzero entries of the sparse error $e^\star$ are known in prior. In particular, if ignoring this $\sqrt{n}$ factor, (15) is unimprovable.

The preceding results have focused on scenarios where the signal is perfectly sparse. We now consider probably the most general setting, in which $x^\star$ is not exactly sparse, but rather can be approximated well by a sparse vector and the observation vector y is corrupted by both sparse error and dense noise with noise level σ. Denote $x_T^\star \in \mathbb{R}^n$ as a vector containing the k largest magnitude entries of $x^\star$ and zeros elsewhere and assume an uniform distribution on the sign of $x_T^\star$ at the support T, we can now establish the following corollary

Corollary 1. *Under the same assumptions defined in Theorem 2, the pair of solution* $(\widehat{x}, \widehat{e})$ *to the convex program (12) obeys*

$$\begin{aligned} \|\widehat{x} - x^\star\|_2 + \|\widehat{e} - e^\star\|_2 &\leq \frac{8(\lambda+1)}{\min\{1,\lambda\}} \\ &\times \left[\sigma\sqrt{n\left(\frac{4n}{m-s}+1\right)} + 2\|x^\star - x_T^\star\|_1\right] + 2\sigma. \end{aligned} \tag{16}$$

Ignoring the $\sigma\sqrt{8}$ term, one can see how the bound in Corollary 1 shows a natural splitting into two terms. The first can be interpreted as *data error* associated with the noise ν, whereas the second term relates to the *approximation error*, measuring how far the signal $x^\star$ is from the best k-sparse approximation $x_T^\star$.

D. When error is sparsified under an arbitrary basis

Thus far, we have only investigated truly sparse error $e^\star$. That is, $e^\star$ is sparse under the identity transformation. A natural generation is to consider $e^\star$ being sparse under any orthogonal transformation D, including the former as a particular case. Mathematically, we consider the observation model

$$y = A_{\Omega\bullet}x^\star + Dg^\star + \nu, \tag{17}$$

where $e^\star = Dg^\star$ and $g^\star$ is a s-sparse vector. It is clear that via simple algebra, this setting boils down to (11) as

$$D^*y = D^*A_{\Omega\bullet}x^\star + g^\star + D^*\nu.$$

Notice that due to the orthogonality of D, $D^*A_{\Omega\bullet}$ is also an orthogonal matrix. Therefore, all preceding theorems are still relevant in this setting. The parameter μ is now interpreted as the mutual incoherence between the sensing matrix A_Ω and the sparsifying transform D. In particular,

$$\mu = n\max_{i,j}|\langle a_i, d_j\rangle|, \tag{18}$$

where a_i and d_j are columns of matrices $A_{\Omega\bullet}$ and D. As the incoherence of these two matrices is small, fewer measurements are required to still guarantee stable recovery. This results from an intuitive fact that it is easier to decompose y into $x^\star$ and $g^\star$ if two column spaces of $A_{\Omega\bullet}$ and D are sufficiently separated.

E. Contribution and connections to previous works

The problem of recovering the signal from grossly corrupted measurements has initially been formulated by Wright *et al.* in an appealing practical paper [22] and further analyzed in [2]. Taking the sparsity information of $e^\star$ into account, the authors proposed to solve

$$\min_{x,e} \|x\|_1 + \|e\|_1 \qquad \text{subject to} \qquad y = Ax + e. \tag{19}$$

The result of [2] is asymptotic in nature. The authors showed that as n is extremely large and provided x is extremely sparse, then (19) can precisely recover both $x^\star$ and $e^\star$ from almost any error with support fraction bounded away from 100%. Their analysis is based on the Gaussian assumption of the matrix A. Particularly, A is a matrix whose columns a_i's are assumed to be $\mathcal{N}(\mu, \frac{\nu^2}{m}I_m)$, where $\|\mu\|_2 = 1$ and $\|\mu\|_\infty \leq Cm^{-1/2}$. Furthermore, for sufficiently large m, they require the sparsity of x to grow sublinearly with m. This is of course far from the optimal bound, in which k is almost linear with m (i.e. only in the order of $m/\log n$).

One of the appealing consequence of their analysis is an explicit expression between three important terms: the dimension ratio $\delta = \frac{n}{m}$ of the matrix $A \in \mathbb{R}^{m\times n}$, the fraction error $\rho = \frac{s}{m}$ and the signal support density $\alpha = \frac{k}{m}$. However, this relationship is difficult to interpret due to the complicated coupling of these terms.

Employing the idea from [2], Li *et al.* [20] and Laska *et al.* [23] proposed different applications under the same framework. The former considered the problem of joint source-channel coding, and the later proposed a so-called pursuit of justice model to deal with sparse unbounded noise. When the measurement matrix A obeys restricted isometry property (RIP), both of them showed that the combination matrix $[A, \ I]$ also satisfies the RIP with high probability, where I is the identity matrix. A consequent conclusion is that the signal is perfectly recovered as long as signal and error sparsity levels are in the order of $m/\log n$. The main drawback of these papers is that they are not able to show that perfect recovery is guaranteed when the number of corrupted entries is linearly proportional to the total number of observations.

After the initial submission of our paper to Arxiv, we noticed another two independent investigations into this problem: Studer *et al.* [24] and Li [25]. The former studies the more

general observation model, $y = Ax + De$, where A and D are general matrices. The authors established deterministic guarantee, which is weaker than our results in Theorem 1. Using different proof techniques, the latter paper delivered similar results as in Theorem 1 with more general model of the sensing matrix $A_{\Omega \bullet}$. In particular, rows of $A_{\Omega \bullet}$ are sampled independently from a population F obeying $\mathbb{E} a_i a_i^* = I$. However, both papers do not investigate the more realistic model in which both sparse and dense noise present in the observations.

In another direction and much earlier, Candès and Tao investigated the error correction problem [7]. In this problem, the question is how to reconstruct the input vector $x^\star$ from corrupted measurements $y = Bx^\star + e^\star$, where the coding matrix $B \in \mathbb{R}^{m \times n}$ is required to be overcomplete ($m > n$) and $e^\star$ is the channel corruption vector, which is usually assumed to be sparse. They proposed to retrieve $x^\star$ by solving the following ℓ_1-minimization problem

$$\min_x \|b - Bx\|_1 . \qquad (20)$$

Though sharing the same general ℓ_1 model, our approach departs from all previous work in compressed sensing in many aspects:

1) Unlike Wright and Ma's model [2] where Gaussian measurement matrices are analyzed, we study the problem with suborthogonal matrices. These matrices often possess many desirable properties over Gaussian matrices in term of fast and efficient computation [14], [15]. Furthermore, we investigate the more difficult problem in which both sparse and dense error appear in the observations. This model is not studied in [2]. We show a surprising message that the extended ℓ_1 minimization is stable under both perturbations, even if the sparse error is arbitrarily large and its support size is arbitrarily close to the total number of observations. A straight forward consequence of this result is that accurate recovery is achieved when measurements are not perturbed by dense noise.

2) Our model is different from Candès and Tao [7] in two aspects. First, we allow the coding matrix to be underdetermined, that is $m \le n$. Second, the input vector is assumed to be sparse. If we recast the extended ℓ_1-minimization in (5) as

$$\min_x \|x\|_1 + \lambda \|b - A_{\Omega \bullet} x\|_1 , \qquad (21)$$

then one can clearly see the integration of the two ℓ_1-norms in a unified optimization: one is used to impose sparsity of the input vector whereas the other exploits error sparsity as in (20).

3) We propose a minor but subtle modification in the extended ℓ_1 minimization of [2]. By adding a regularization parameter λ into (5), we can balance the ℓ_1-norm of both x and e. Specifically, we can establish an explicit expression for the regularization parameter λ as well as the sparsity levels of both signal and error. This mathematical expression is intuitively interpretable: signal and error sparsity levels should be inversely proportional. If more measurements are corrupted − equivalently, the error is denser − we expect to recover the signal with smaller support size. In contrast, we are able to recover the signal with larger support size when fewer

errors appear in the measurement vector. In practice, when the fraction of error is unknown, we can set a good-for-all parameter $\lambda = \sqrt{\frac{n}{m \log^{1/2} n}}$.

III. STRUCTURE OF OUR PROOF

A. Bernoulli model and derandomization technique

The Bernoulli model. Instead of showing that Theorem 2 holds as Ω and S are sets of size m and s sampled uniformly at random, we find that it is more convenient to prove the theorem for subsets Ω and S sampled according to the Bernoulli model. This way, we can take advantage of the statistical independence of measurements. The same argument as presented in [4], [13] shows that the probability of "failure" under the uniform model is less than two times the probability of failure under the Bernoulli model. Here, "failure" implies the optimization in (5) does not recover exactly the signal. Thus, from now on, we instead consider $\Omega = \{i \in [1, n] : \delta_i = 1\}$ where $\{\delta_i\}_{1 \le i \le n}$ is a sequence of independent identically distributed Bernoulli random variables taking value one with probability η and zero with probability $1 - \eta$, where η is chosen such that the expected cardinality of Ω is $\eta n = m$. Similarly, let $S = \{i \in \Omega : \delta_i' = 1\}$, $1 \le i \le m$ where $\{\delta_i'\}_{i \in \Omega}$ are i.i.d Bernoulli random variables with $\mathbb{P}(\delta_i' = 1) = \rho$ so that the expected cardinality of S is $\rho m = s$. Toward this end, we will write $\Lambda \sim \text{Ber}(\eta)$ as a shorthand for Λ sampled from the Bernoulli model with parameter η.

The following are five important index sets that is frequently used in the sequel.

- Ω are those locations corresponding to observations: $\Omega \sim \text{Ber}(\eta)$ with $\eta = \frac{m}{n}$.
- $S \subset \Omega$ are locations where the measurements are available but absolutely unreliable. It is clear that the distribution of S relies on that of Ω. Conditioning on Ω, we have $S \sim \text{Ber}(\rho)$ with $\rho = \frac{s}{m}$. We can also think S as a subset selected from the set $\{1, ..., n\}$ with parameter $\eta \rho$. That is, $S \sim \text{Ber}(\eta \rho)$.
- $J \subset \Omega$ are locations where the measurements are available and truthworthy. It is clear that $J = \Omega / S$. Conditioned on Ω, we have $J \sim \text{Ber}(1 - \rho)$. In other words, $J \sim \text{Ber}(\rho_0)$ with $\rho_0 := \eta(1 - \rho)$.
- We also consider the index sets $S^c = \{1, ..., n\}/S$ and $J^c = \{1, ..., n\}/J$

Derandomization. In Theorem 2, the sign of $e^\star$ is fixed. During the proof, we need to place an additional assumption on $e^\star$. That is, the sign of $e_S^\star$ is uniformly distributed, receiving value 1 or -1 with probability $1/2$. However, by the same appealing derandomization technique presented in [13], the probability of recovering $e^\star$ whose signs on the support S are arbitrary is at least equal to that of recovering $e^\star$ whose signs are equally likely to be 1 or -1. This is formally stated in the lemma below

Lemma 1 (*Theorem 2.3 of [13]*). *Suppose $x^\star$ obeys conditions of Theorem 2 and the locations of nonzero entries of $e^\star$ follows the Bernoulli model with parameter 2ρ, and signs of $e^\star$ are i.i.d ± 1 with probability $1/2$. Then, if the solution of extended*

ℓ_1-*minimization (5) is exact with high probability, it is also exact with at least the same probability with the model in which the signs of $e^\star$ are fixed and its nonzero entries are selected from the Bernoulli model with parameter ρ.*

B. Dual certificate

The following lemma shows that if there exists a dual pair $(z^{(x)}, z^{(e)})$ satisfying certain conditions, then for any pair (x, e), its ℓ_1-norm sum is no smaller than that of $(x^\star, e^\star)$.

Lemma 2. *Suppose that $\|A_{J^c T}\| < 1$. If there exists a pair of dual vectors $(z^{(x)}, z^{(e)})$ with the following properties,*

1) $z^{(x)} = \lambda A_{\Omega\bullet}^* z^{(e)}$,
2) $z_T^{(x)} = \mathrm{sgn}(x_T^\star)$ *and* $\left\| z_{T^c}^{(x)} \right\|_\infty \leq 3/4$,
3) $z_S^{(e)} = \mathrm{sgn}(e_S^\star)$ *and* $\left\| z_{S^c}^{(e)} \right\|_\infty \leq 3/4$,

then for any perturbation pair (h, f) satisfying $f = -A_{\Omega\bullet}h$, we have

$$\|x^\star + h\|_1 + \lambda \|e^\star + f\|_1 \tag{22}$$
$$\geq \|x^\star\|_1 + \lambda \|e^\star\|_1 + \frac{1}{4}(\|h_{T^c}\|_1 + \lambda \|A_{J\bullet}h\|_1).$$

Before proving this lemma, it is necessary to notice how the Lemma implies the perfect recovery of the linear program in (5). Indeed, denote by $(\widehat{x}, \widehat{e})$ the optimal solution of (5) and let $\widehat{x} := x^\star + h$ and $\widehat{e} := e^\star + f$, then it is obvious that $f = -A_{\Omega\bullet}h$. By the convexity of the objective function, we have $\|x^\star + h\|_1 + \lambda \|e^\star + f\|_1 \leq \|x^\star\|_1 + \lambda \|e^\star\|_1$.

Furthermore, from Lemma 2, assuming the existence of a dual pair $(z^{(x)}, z^{(e)})$ and $\|A_{J^c T}\| < 1$, the inequality (22) obeys. Combining both arguments, we have

$$\frac{1}{4}(\|h_{T^c}\|_1 + \lambda \|A_{J\bullet}h\|_1) \leq 0.$$

It is clear that the left-hand side of the above equation is strictly greater than 0 for every $h \neq 0$. Thus, in order for the equality to occur, it is necessary that $h_{T^c} = 0$ and $A_{JT}h_T = 0$. We can establish that, due to the orthogonality of matrix A, the condition $\|A_{J^c T}\| < 1$ is equivalent to $\|I - A_{JT}^* A_{JT}\| < 1$. This suggests that $A_{JT}^* A_{JT}$ is invertible, and thus, $A_{JT}h_T = 0$ only if $h_T = 0$. We therefore conclude that $h = 0$ and $f = -A_{\Omega\bullet}h = 0$ or in other words, $(\widehat{x}, \widehat{e})$ is the exact solution.

Proof of Lemma 2. Denote as v_0 and w_0 the subgradients of $\|x\|_1$ and $\|e\|_1$ at $x^\star$ and $e^\star$, respectively. It is well-known that $v_{0_T} = \mathrm{sgn}(x_T^\star)$ and $\|v_{0_{T^c}}\|_\infty \leq 1$. Similarly, we have $w_{0_S} = \mathrm{sgn}(e_S^\star)$ and $\|w_{0_{S^c}}\|_\infty \leq 1$. By the definition of subgradients, we derive

$$\|x^\star + h\|_1 + \lambda \|e^\star + f\|_1$$
$$\geq \|x^\star\|_1 + \lambda \|e^\star\|_1 + \langle v_0, h \rangle + \lambda \langle w_0, f \rangle \tag{23}$$
$$= \|x^\star\|_1 + \lambda \|e^\star\|_1 + \langle v_0, h \rangle - \lambda \langle w_0, A_{\Omega\bullet}h \rangle.$$

Let us now consider $\langle v_0, h \rangle - \lambda \langle w_0, A_{\Omega\bullet}h \rangle$. By decomposing v_0 and w_0 into vectors of index sets $\{S, T\}$ and their complements $\{S^c, T^c\}$, we have

$$\langle v_0, h \rangle - \lambda \langle w_0, A_{\Omega\bullet}h \rangle = \langle \mathrm{sgn}(x_T^\star), h_T \rangle - \lambda \langle \mathrm{sgn}(e_S^\star), A_{S\bullet}h \rangle$$
$$+ \langle v_{0_{T^c}}, h_{T^c} \rangle - \lambda \langle w_{0_J}, A_{J\bullet}h \rangle.$$

Now choosing $v_{0_{T^c}}$ such that $\langle v_{0_{T^c}}, h_{T^c} \rangle = \|h_{T^c}\|_1$ and w_{0_J} such that $\langle w_{0_J}, A_{J\bullet}h \rangle = -\|A_{J\bullet}h\|_1$, we can rewrite

$$\langle v_0, h \rangle - \lambda \langle w_0, A_{\Omega\bullet}h \rangle = \langle \mathrm{sgn}(x_T^\star), h_T \rangle$$
$$- \lambda \langle \mathrm{sgn}(e_S^\star), A_{S\bullet}h \rangle + \|h_{T^c}\|_1 + \lambda \|A_{J\bullet}h\|_1. \tag{24}$$

In addition, the identity relation $z^{(x)} = \lambda A_{\Omega\bullet}^* z^{(e)}$ can be reformulated as

$$\begin{pmatrix} \mathrm{sgn}(x_T^\star) \\ 0_{T^c} \end{pmatrix} - \lambda A_{\Omega\bullet}^* \begin{pmatrix} \mathrm{sgn}(e_S^\star) \\ 0_J \end{pmatrix}$$
$$= -\begin{pmatrix} 0_T \\ z_{T^c}^{(x)} \end{pmatrix} + \lambda A_{\Omega\bullet}^* \begin{pmatrix} 0_S \\ z_J^{(e)} \end{pmatrix}.$$

Taking the inner product with h on both sides yields

$$\langle \mathrm{sgn}(x_T^\star), h_T \rangle - \lambda \langle A_{S\bullet}^* \mathrm{sgn}(e_S^\star), h \rangle$$
$$= -\langle z_{T^c}^{(x)}, h_{T^c} \rangle + \lambda \langle A_{J\bullet}^* z_J^{(e)}, h \rangle$$
$$= -\langle z_{T^c}^{(x)}, h_{T^c} \rangle + \lambda \langle z_J^{(e)}, A_{J\bullet}h \rangle.$$

Notice also that $\langle A_{S\bullet}^* \mathrm{sgn}(e_S^\star), h \rangle = \langle \mathrm{sgn}(e_S^\star), A_{S\bullet}h \rangle$. Hence, (24) is equivalent to

$$\langle v_0, h \rangle - \lambda \langle w_0, A_{\Omega\bullet}h \rangle$$
$$= \|h_{T^c}\|_1 + \lambda \|A_{J\bullet}h\|_1 - \langle z_{T^c}^{(x)}, h_{T^c} \rangle + \lambda \langle z_J^{(e)}, A_{J\bullet}h \rangle$$
$$\geq \frac{1}{4}(\|h_{T^c}\|_1 + \lambda \|A_{J\bullet}h\|_1),$$

where the last inequality is due to $\langle z_{T^c}^{(x)}, h_{T^c} \rangle \leq \left\| z_{T^c}^{(x)} \right\|_\infty \|h_{T^c}\|_1 \leq \frac{3}{4} \|h_{T^c}\|_1$ and $\langle z_J^{(e)}, A_{J\bullet}h \rangle \leq \left\| z_J^{(e)} \right\|_\infty \|A_{J\bullet}h\|_1 \leq \frac{3}{4} \|A_{J\bullet}h\|_1$. Substituting this inequality into (23), we complete the proof. $\square$

From the result of Lemma 2, in order to prove exact recovery of the convex program, it suffices to construct a dual certificate $(z^{(x)}, z^{(e)})$ obeying the conditions of Lemma 2. Partitioning $z^{(x)}$, $z^{(e)}$ into two subsets belonging to T and T^c, S and S^c, the identity relation between $z^{(x)}$ and $z^{(e)}$ can be reformulated as follows

$$\begin{pmatrix} \mathrm{sgn}(x_T^\star) \\ 0_{T^c} \end{pmatrix} + \begin{pmatrix} 0_T \\ z_{T^c}^{(x)} \end{pmatrix}$$
$$= \lambda A^* \begin{pmatrix} \mathrm{sgn}(e_S^\star) \\ 0_J \\ 0_{\Omega^c} \end{pmatrix} + \lambda A^* \begin{pmatrix} 0_S \\ z_J^{(e)} \\ 0_{\Omega^c} \end{pmatrix}. \tag{25}$$

If we can construct a pair of vectors $(v^{(x)}, w^{(e)})$ such that $v^{(x)} + w^{(e)}$ is equal to both sides of (25), that is

$$\begin{cases} v_T^{(x)} + w_T^{(e)} = \mathrm{sgn}(x_T^\star) \\ v_{T^c}^{(x)} + w_{T^c}^{(e)} = z_{T^c}^{(x)} \\ A_{S\bullet}(v^{(x)} + w^{(e)}) = \lambda \mathrm{sgn}(e_S^\star) \\ A_{J\bullet}(v^{(x)} + w^{(e)}) = \lambda z_J^{(e)} \\ A_{\Omega^c\bullet}(v^{(x)} + w^{(e)}) = 0, \end{cases}$$

then the existence of the dual certificate $(z^{(x)}, z^{(e)})$ in Lemma 2 is guaranteed. As a consequence, it now suffices to produce

a dual pair $(v^{(x)}, w^{(e)})$ obeying

$$\begin{cases} v_T^{(x)} = \text{sgn}(x_T^\star) \\ \left\| v_{T^c}^{(x)} \right\|_\infty < 3/8 \\ A_{S\bullet} v^{(x)} = 0 \\ \left\| A_{J\bullet} v^{(x)} \right\|_\infty < 3\lambda/8 \\ A_{\Omega^c\bullet} v^{(x)} = 0 \end{cases} \quad \text{and} \quad \begin{cases} w_T^{(e)} = 0 \\ \left\| w_{T^c}^{(e)} \right\|_\infty < 3/8 \\ A_{S\bullet} w^{(e)} = \lambda \, \text{sgn}(e_S^\star) \\ \left\| A_{J\bullet} w^{(e)} \right\|_\infty < 3\lambda/8 \\ A_{\Omega^c\bullet} w^{(e)} = 0. \end{cases}$$
(26)

In the next section, we will establish that the valid dual pair $(v^{(x)}, w^{(e)})$ exists with probability converging to unity.

C. Dual certification constructions

We now propose to construct a dual certificate pair $(v^{(x)}, w^{(e)})$ whose components are described as follows

1) *Construction of $w^{(e)}$ via least-square.* Since $w_T^{(e)} = 0$, the identity conditions $A_{S\bullet} w^{(e)} = \lambda \, \text{sgn}(e_S^\star)$ and $A_{\Omega^c\bullet} w^{(e)} = 0$ can now be represented by a single equation

$$A_{J^c T^c} w_{T^c}^{(e)} = \lambda \begin{pmatrix} \text{sgn}(e_S^\star) \\ 0_{\Omega^c} \end{pmatrix}, \quad (27)$$

where we recall that $J^c = S \cup \Omega^c$. Next, assuming that $\|A_{J^c T}\| < 1$, then we have $\|I - A_{J^c T} A_{J^c T}^*\| = \|A_{J^c T} A_{J^c T}^*\| < 1$. Consequently, matrix $A_{J^c T^c} A_{J^c T^c}^*$ is invertible. We then set

$$w_{T^c}^{(e)} = \lambda A_{J^c T^c}^* (A_{J^c T^c} A_{J^c T^c}^*)^{-1} \begin{pmatrix} \text{sgn}(e_S^\star) \\ 0_{\Omega^c} \end{pmatrix}. \quad (28)$$

Clearly, $w_{T^c}^{(e)}$ is the least-square solution of the linear system in (27). This construction has a natural interpretation: among all solutions of the linear system, $w_{T^c}^{(e)}$ has the minimum ℓ_2-norm. We expect that its ℓ_∞-norm is also sufficiently small to obey the condition in (26).

2) *Construction of $v^{(x)}$.* A simple way to produce $v^{(x)}$ is as follows

$$v^{(x)} = A_{J\bullet}^* A_{JT} (A_{JT}^* A_{JT})^{-1} \text{sgn}(x_T^\star). \quad (29)$$

It is obvious from this construction that $v_T^{(x)} = \text{sgn}(x_T^\star)$. Furthermore, $A_{S\bullet} v^{(x)} = 0$ and $A_{\Omega^c\bullet} v^{(x)} = 0$ due to the orthogonality property of the matrix A. Thus, all three identity relations with respect to $v^{(x)}$ in (26) are guaranteed.

We now state two key lemmas that establish the ℓ_∞-norm bounds for $v^{(x)}$ and $w^{(e)}$.

Lemma 3. *Assume that $\Omega \sim \text{Ber}(\eta)$ and $S \sim \text{Ber}(\eta\rho)$ where parameters $\eta = m/n$ and $\rho = s/m$. Under the same assumptions as in Theorem 2, with high probability, the dual vector $v^{(x)}$ constructed in (29) obeys*

1) $\left\| A_{J\bullet} v^{(x)} \right\|_\infty < 3\lambda/8$,
2) $\left\| v_{T^c}^{(x)} \right\|_\infty < 3/8$.

Lemma 4. *Assume that Ω and S are sampled as in Lemma 3. Under the same assumptions as in Theorem 2, with high probability, the dual vector $w^{(e)}$ constructed in (28) obeys*

1) $\left\| w_{T^c}^{(e)} \right\|_\infty < 3/8$,

2) $\left\| A_{J\bullet} w^{(e)} \right\|_\infty < 3\lambda/8$.

Lemmas 3 and 4 suggest the existence of $(z^{(x)}, z^{(e)})$. In other words, the solution of the convex program in (5) is exact and unique.

IV. PROOFS OF DUAL CERTIFICATES

A. Important auxiliary lemmas

In this section, we first develop several auxiliary results concerning the main proof.

Lemma 5. *Let S_0 be locations sampled randomly from the set $\{1, ..., n\}$, $S_0 \sim \text{Ber}(\rho_0)$. With probability of success at least $1 - n^{-1}$, we have*

$$\left\| I_{k \times k} - \rho_0^{-1} A_{S_0 T}^* A_{S_0 T} \right\| \le \epsilon,$$

provided that $\rho_0 \ge C_0 \frac{\epsilon^{-2} \mu k \log n}{n}$ for $C_0 = 2^{3/4} e^2 \sqrt{\pi e}$.

This result has been known in the literature [26], [27], [28]. However, for completeness, we provide a brief proof which relies on high order moment bound of the spectral norm. We emphasize that the lemma is important since it provides us the bound of $\|A_{J^c T}\|$. In fact, recall that $J \sim \text{Ber}(\rho_0)$ with $\rho_0 = \eta(1 - \rho)$, Lemma 5 suggests that

$$\left\| I - \rho_0^{-1} A_{JT}^* A_{JT} \right\| \le \epsilon, \quad (30)$$

provided that $\rho_0 \ge C\epsilon^{-2} \frac{\mu k \log n}{n}$. Furthermore, from the fact that $A_{JT}^* A_{JT} = I - A_{J^c T}^* A_{J^c T}$, we obtain

$$\epsilon \ge \left\| I - \rho_0^{-1} (I - A_{J^c T}^* A_{J^c T}) \right\|$$
$$\ge \rho_0^{-1} \left\| A_{J^c T}^* A_{J^c T} \right\| - (\rho_0^{-1} - 1).$$

This inequality leads to $\left\| A_{J^c T}^* A_{J^c T} \right\| \le \rho_0 \epsilon + (1 - \rho_0)$. We conclude the argument by the following proposition.

Proposition 1. *Provided that $m - s \ge 4C_0 \mu k \log n$. With probability at least $1 - n^{-1}$, we have*

$$\|A_{J^c T}\| \le \sqrt{1 - \rho_0/2}.$$

Proof of Lemma 5. Define $S_0 = \{i : \delta_i = 1\}$ where δ_i is an independent sequence of Bernoulli variables with parameter ρ_0 and denote u_i ($i \in S_0$) to be row vectors of $A_{S_0 T}$. With these notation, we have

$$A_{S_0 T}^* A_{S_0 T} = \sum_{i \in S_0} u_i \otimes u_i = \sum_{i=1}^n \delta_i u_i \otimes u_i.$$

Applying Theorem 5 of [28] with $q = \log n$, we obtain

$$\left(\mathbb{E} \left\| I - \rho_0^{-1} \sum_{i=1}^n \delta_i u_i \otimes u_i \right\|^{\log n} \right)^{1/\log n}$$
$$\le C \sqrt{\rho_0^{-1} \log n} \max_i \|u_i\|_2$$
$$\le C \sqrt{\rho_0^{-1} (\mu k \log n)/n} := E,$$

where the constant $C = 2^{3/4} \sqrt{\pi e}$, and the last inequality holds from $\|u_i\|_2 \le \sqrt{\mu k/n}$.

From Markov's inequality, we can establish

$$\mathbb{P}\left(\left\|I - \rho_0^{-1}\sum_{i=1}^{n}\delta_i u_i \otimes u_i\right\| \geq \epsilon\right) \leq \frac{E^{\log n}}{(\epsilon)^{\log n}}.$$

By the assumption of the Lemma that $\frac{C\sqrt{\rho_0^{-1}\mu k \log n}}{\epsilon} \leq \frac{1}{e}$, we have with probability of success at least $1 - n^{-1}$,

$$\left\|I - \rho_0^{-1}\sum_{i=1}^{n}\delta_i u_i \otimes u_i\right\| \leq \epsilon,$$

as claimed. $\qquad\square$

The next lemma shows the matrix $A_{S_0 T}$ is almost orthogonal to the matrix $A_{S_0 T^c}$ where S_0 is a random subset selected from columns of the matrix A. This property is important in distinguishing the set T from the set T^c and helping the algorithm identify the true support of $x^\star$. We defer the proof to the Appendix.

Lemma 6. *Let S_0 be locations sampled randomly from the set $\{1,...,n\}$, $S_0 \sim \mathrm{Ber}(\rho_0)$. With probability at least $1 - 3n^{-1}$, the following inequality obeys*

$$\left\|A_{S_0 T}^* u\right\|_2 \leq \sqrt{C'\rho_0 \frac{\mu \max\{k, \log n\}}{n}}$$

for any column vector u of the matrix $A_{S_0 T^c}$, provided that $\rho_0 \geq C\frac{\mu \max\{k, \log n\}}{n}$ where C and C' are numerical constants.

We are now ready to prove Lemmas 3 and 4 regarding the dual certificates.

B. Proof of Lemma 3

Part 1. By the construction of $v^{(x)}$ in (29),

$$A_{J\bullet} v^{(x)} = A_{JT}(A_{JT}^* A_{JT})^{-1}\mathrm{sgn}(x_T^\star).$$

Denote u_i as a row of the matrix A_{JT}, we have

$$\left\|A_{J\bullet} v^{(x)}\right\|_\infty = \max_i |u_i(A_{JT}^* A_{JT})^{-1}\mathrm{sgn}(x_T^\star)|$$
$$:= \max_i |\langle Wu_i^*, \mathrm{sgn}(x_T^\star)\rangle|,$$

where we denote $W := (A_{JT}^* A_{JT})^{-1}$. The right-hand side is a sum of zero mean random variables, which can be bounded by Hoeffding's inequality. Hence,

$$\mathbb{P}(|\langle Wu_i^*, \mathrm{sgn}(x_T^\star)\rangle| \geq \tau) \leq 2\exp\left(-\frac{\tau^2}{2\left\|Wu_i^*\right\|_2^2}\right).$$

Notice from (30) that with probability converging to one, $\|I - \rho_0 A_{JT}^* A_{JT}\| \leq \epsilon$. Thus, $(1-\epsilon)\rho_0 \leq \sigma_{\min}(A_{JT}^* A_{JT}) \leq \sigma_{\max}(A_{JT}^* A_{JT}) \leq (1+\epsilon)\rho_0$ where $\sigma_{\min}$ and $\sigma_{\max}$ are minimum and maximum singular values of the matrix. In addition, we have exploited the fact that spectral norm for any matrix H obeys $\left\|H^{-1}\right\| \leq \frac{1}{\sigma_{\min}(H)}$. Thus, conditioning on the event $\mathcal{E} = \{\|I - \rho_0 A_{JT}^* A_{JT}\| \leq \epsilon\}$, we have

$$\|W\| \leq \frac{1}{(1-\epsilon)\rho_0} \leq \frac{2}{\rho_0},$$

with the choice of $\epsilon \leq 1/2$. Consequently, combining with $\|u_i\|_2^2 \leq \mu\frac{k}{n}$, we conclude that $\|Wu_i^*\|_2^2 \leq \frac{4\mu k}{\rho_0^2 n}$.

Now setting $\tau := \sqrt{\frac{16\mu k \log n}{\rho_0^2 n}}$ and taking the union bound over all row vectors of matrix A_{JT}, we obtain

$$\mathbb{P}\left(\left\|A_{J\bullet} v^{(x)}\right\|_\infty \geq \sqrt{\frac{16\mu k \log n}{\rho_0^2 n}}\right) \leq 2|J|e^{-2\log n} + \mathbb{P}(\mathcal{E}^c)$$
$$\leq 3n^{-1},$$

where the inequality follows from the total probability rule: $\mathbb{P}(F \geq \tau) \leq \mathbb{P}(F \geq \tau|\mathcal{E}) + \mathbb{P}(\mathcal{E}^c)$ with $F := \left\|A_{J\bullet} v^{(x)}\right\|_\infty$. We conclude that $\left\|A_{J\bullet} v^{(x)}\right\|_\infty \leq \frac{\lambda}{4}$ as long as $k \leq C\frac{\lambda^2 \rho_0^2 n}{\mu \log n}$. Replace $\lambda = \sqrt{\frac{1}{\gamma \log n}\frac{n}{m}}$, $\rho_0 = \frac{m-s}{n}$ and $s = \gamma m$, one can see that the upper bound of k automatically follows from the assumption that $k \leq C\frac{(1-\gamma)^2}{\gamma}\frac{m}{\mu^2(\log n)^2}$. $\qquad\square$

Part 2. In this part, we need to show that with high probability,

$$\left\|A_{JT^c}^* A_{JT}(A_{JT}^* A_{JT})^{-1}\mathrm{sgn}(x_T^\star)\right\|_\infty \leq 3/8.$$

Denote u_i as a column vector of the matrix A_{JT^c} and consider $u_i^* A_{JT}(A_{JT}^* A_{JT})^{-1}\mathrm{sgn}(x_T^\star) = \langle (A_{JT}^* A_{JT})^{-1}A_{JT}^* u_i, \mathrm{sgn}(x_T^\star)\rangle$, which is a sum of random variables. Its absolute value can be estimated via Hoeffding's inequality,

$$\mathbb{P}(|u_i^* A_{JT}(A_{JT}^* A_{JT})^{-1}\mathrm{sgn}(x_T^\star)| \geq \tau) \leq 2\exp\left(-\frac{\tau^2}{2\left\|z\right\|_2^2}\right),$$

where $z := (A_{JT}^* A_{JT})^{-1}A_{JT}^* u_i$. As previously showed, conditioning on the event $\mathcal{E}_1 = \{\|I - \rho_0 A_{JT}^* A_{JT}\| \leq \epsilon \leq 1/2\}$, we have $\left\|(A_{JT}^* A_{JT})^{-1}\right\| \leq 2/\rho_0$. In addition, we define the event $\mathcal{E}_2 := \{\|A_{JT}^* u_i\|_2 \leq \sqrt{C'\rho_0\frac{\mu \max\{k, \log n\}}{n}}\}$, which bounds the ℓ_2 norm of $A_{JT}^* u$ with $J \sim \mathrm{Ber}(\rho_0)$. We showed from Lemma 6 that $\mathbb{P}(\mathcal{E}_2) \leq 1 - 3n^{-1}$. Therefore, conditioning on both $\mathcal{E}_1$ and $\mathcal{E}_2$, we get

$$\|z\|_2 \leq \left\|(A_{JT}^* A_{JT})^{-1}\right\|\left\|A_{JT}^* u_i\right\| \leq \sqrt{C'\frac{\mu \max\{k, \log n\}}{\rho_0 n}}.$$

Setting $\tau^2 := 4C'\frac{\mu \max\{k, \log n\}}{\rho_0 n}$ and taking the union bound, we conclude that

$$\mathbb{P}\left(\left\|v_{T^c}^{(x)}\right\|_\infty \geq \tau\right) \leq 2(n-k)e^{-2\log n} + \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c),$$

which is less than $6n^{-1}$. Now replace $\rho_0 = \frac{m-s}{n}$ and assume that $m - s \geq \tilde{C}\mu \max\{k, \log n\}\log n$ where $\tilde{C} = 4(8/3)^2 C'$, we achieve $\left\|v_{T^c}^{(x)}\right\|_\infty \leq 3/8$ as claimed. $\qquad\square$

C. Proof of Lemma 4

1) Preliminary results: In order to set up the bounds of Lemma 4, it is necessary to estimate the spectral norm bound of $\|A_{ST}\|$. The following proposition establishes such a bound.

Proposition 2. *With probability at least $1 - n^{-1}$,*

$$\left\|A_{ST}^* A_{ST}\right\| \leq \left(1 + \sqrt{\frac{\mu k \log n}{s}}\right)\eta\rho.$$

Proof. Recall that $S \sim \text{Ber}(\eta\rho)$. By Lemma 5, with high probability, we have

$$\left\| I - (\eta\rho)^{-1} A_{ST}^* A_{ST} \right\| \leq \epsilon_1, \tag{31}$$

provided $\eta\rho \geq C_0 \frac{\epsilon_1^{-2} \mu k \log n}{n}$. Note that $\eta\rho = \frac{s}{n}$, thus the condition is equivalent to $s \geq \epsilon_1^{-2} \mu k \log n$. This inequality is automatically satisfied by setting $\epsilon_1 = \sqrt{\frac{\mu k \log n}{s}}$. Therefore, (31) gives us

$$\|A_{ST}^* A_{ST}\| \leq \left(1 + \sqrt{\frac{\mu k \log n}{s}} \right) \eta\rho$$

as claimed. $\qquad\qquad\qquad\qquad\qquad\qquad \square$

2) Main proofs:

Part 1. We will start with the construction of $w_{T^c}^{(e)}$ in (28). Our goal is to show that with high probability,

$$V := \lambda \left\| A_{J^c T^c}^* (A_{J^c T^c} A_{J^c T^c}^*)^{-1} \begin{pmatrix} \text{sgn}(e_S^\star) \\ 0_{\Omega^c} \end{pmatrix} \right\|_\infty < 1/4.$$

By series expansion, $(I - H)^{-1} = I + \sum_{j=1}^\infty H^j$. We have

$$\begin{aligned} A_{J^c T^c}^* (A_{J^c T^c} A_{J^c T^c}^*)^{-1} &= A_{J^c T^c}^* (I - A_{J^c T} A_{J^c T}^*)^{-1} \\ &= A_{J^c T^c}^* + \sum_{j \geq 1} A_{J^c T^c}^* (A_{J^c T} A_{J^c T}^*)^j. \end{aligned}$$

Toward that end, denote $r := \begin{pmatrix} \text{sgn}(e_S^\star) \\ 0_{\Omega^c} \end{pmatrix}$. To establish the upper bound of V, we elaborate on the ℓ_∞-norms of two quantities relating to summands of the series expansion. The bound of V is then followed from the triangular inequality. For the first term $V_1 := \lambda \|A_{J^c T^c}^* r\|_\infty$, we have

$$V_1 = \lambda \|A_{ST^c}^* \text{sgn}(e_S^\star)\|_\infty = \lambda \max_i |\langle u_i, \text{sgn}(e_S) \rangle|,$$

where u_i is denoted as a column vector of A_{ST^c}. We notice that $\langle u_i, \text{sgn}(e_S^\star) \rangle$ is a sum of zero mean random variables (by the random assumption on the sign of e_S). Applying Hoeffding's inequality yields

$$\begin{aligned} \mathbb{P}\left(|\langle u_i, \text{sgn}(e_S^\star) \rangle| \geq \tau \right) &\leq 2 \exp \left(-\frac{2\tau^2}{4 \|u_i\|_2^2} \right) \\ &\leq 2 \exp \left(-\frac{\tau^2}{2} \frac{n}{\mu s} \right), \end{aligned}$$

where the last inequality is due to $\|u_i\|_2^2 \leq \frac{\mu s}{n}$. Next, choosing $\tau = \frac{1}{8\lambda}$ and taking the union bound over all $i \in T^c$ yield

$$\mathbb{P}\left(\lambda \|A_{J^c T^c}^* r\|_\infty \geq \frac{1}{8} \right) \leq \exp \left(-\frac{n}{128 \mu \lambda^2 s} + \log(2n) \right),$$

which is bounded away by $e^{-\log n} = n^{-1}$ as long as $s \leq C \frac{n}{\mu \lambda^2 \log n} = C\gamma m$.

For the remainder term, denote the quantity $V_r := \lambda \left\| \sum_{j \geq 1} A_{J^c T^c}^* (A_{J^c T} A_{J^c T}^*)^j r \right\|_\infty$, we have

$$\begin{aligned} V_r &= \lambda \left\| \sum_{j \geq 1} (A_{J^c T^c}^* A_{J^c T})(A_{J^c T}^* A_{J^c T})^{j-1} A_{J^c T}^* r \right\|_\infty \\ &= \lambda \left\| \sum_{j \geq 1} (A_{J T^c}^* A_{J T})(A_{J^c T}^* A_{J^c T})^{j-1} A_{ST}^* \text{sgn}(e_S^\star) \right\|_\infty \\ &= \lambda \max_{i \in T^c} \left| \sum_{j \geq 0} u_i^* A_{JT} (A_{J^c T}^* A_{J^c T})^j A_{ST}^* \text{sgn}(e_S^\star) \right|, \end{aligned}$$

where u_i is denoted as the ith column vector of A_{JT^c}. Notice that vector u_i has length $(m - s)$.

Let $W := \sum_{j \geq 0} A_{JT} (A_{J^c T}^* A_{J^c T})^j A_{ST}^*$. We consider the term inside the max function $V_i = |\langle W^* u_i, \text{sgn}(e_S^\star) \rangle|$. Again, this quantity's bound is an application of Hoeffding's inequality,

$$\mathbb{P}\left(|\langle W^* u_i, \text{sgn}(e_S^\star) \rangle| \geq \tau \right) \leq 2 \exp \left(-\frac{2\tau^2}{4 \|W^* u_i\|_2^2} \right)$$

Next, we have

$$\begin{aligned} \|W^* u_i\| &\leq \|A_{ST}\| \left(\sum_{j \geq 0} \|A_{TJ^c}^* A_{J^c T}\|^j \right) \|A_{TJ}^*\| \|u_i\|_2 \\ &= \frac{\|A_{ST}\| \|A_{JT}^*\|}{1 - \|A_{J^c T}^* A_{J^c T}\|} \|u_i\|_2 . \end{aligned}$$

We now provide the spectral and ℓ_2 norms of these terms. Define the following three events

$$\begin{aligned} \mathcal{E}_1 &:= \{\|A_{J^c T}^* A_{J^c T}\| \leq 1 - \rho_0/2\}, \\ \mathcal{E}_2 &:= \{\|A_{JT}^*\| \leq \sqrt{3\rho_0/2}\}, \quad \text{and} \\ \mathcal{E}_3 &:= \{\|A_{ST}\| \leq (1 + \sqrt{\frac{\mu k \log n}{s}})^{1/2} \sqrt{\eta\rho}\}. \end{aligned}$$

Recall by Proposition 1 that the event $\mathcal{E}_1$ occurs with high probability. Moreover, from Lemma 5, with high probability $\left\| I_{k \times k} - \rho_0^{-1} A_{JT}^* A_{JT} \right\| \leq \epsilon$ provided $\rho_0 \geq C\epsilon^{-2} \frac{\mu k \log n}{n}$. Thus, $\|A_{JT}^*\| \leq \sqrt{\rho_0(1+\epsilon)} \leq \sqrt{3\rho_0/2}$, assuming that $\epsilon \leq 1/2$. Finally, $\mathcal{E}_3$ occurs by Proposition 2 and the fact that $\|u_i\|_2^2 \leq \frac{\mu(m-s)}{n} = \mu\rho_0$. Conditioning on these events, we conclude that

$$\begin{aligned} \|W^* u_i\|_2^2 &\leq \frac{(3\rho_0/2)(\mu\rho_0)(\eta\rho)}{(\rho_0/2)^2} \left(1 + \sqrt{\frac{\mu k \log n}{s}} \right) \\ &\leq 6\mu \frac{s}{n} \left(1 + \sqrt{\frac{\mu k \log n}{s}} \right). \end{aligned}$$

We consider two following cases regarding the size of the set S:

Case 1: if $s \geq \mu k \log n$, then $\|W^* u_i\|_2^2 \leq 12\mu \frac{s}{n}$. Set $\tau = \frac{1}{8\lambda}$ and take the union bound over all $i \in T^c$, we attain

$$\begin{aligned} \mathbb{P}\left(V_r \geq \frac{1}{8\lambda} \right) &\leq 2 \exp \left(-\frac{1}{256 \lambda^2 \mu \frac{s}{n}} + \log n \right) \\ &\quad + \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c) + \mathbb{P}(\mathcal{E}_3^c). \end{aligned}$$

By assuming $s \leq C\gamma m$ with a sufficiently small constant C, $\mu\lambda^2 \frac{s}{n} \leq C\frac{1}{\log n}$. Hence, $V_r \leq \frac{1}{8\lambda}$ with probability $1 - 5n^{-1}$.

Case 2: if $s \leq \mu k \log n$, then $\|W^* u_i\|_2^2 \leq 12\mu \frac{s}{n}\sqrt{\frac{\mu k \log n}{s}}$. Again, setting $\tau = \frac{1}{8\lambda}$ and taking the union bound, we have

$$\mathbb{P}\left(V_r \geq \frac{1}{8\lambda}\right) \leq 2\exp\left(-\frac{1}{256\lambda^2\mu\frac{s}{n}\sqrt{\frac{\mu k \log n}{s}}} + \log n\right)$$
$$\leq 2e^{-\log n} = 2n^{-1},$$

provided that $s \leq C\gamma m$ and $k \leq C'\frac{\gamma m}{\mu \log n}$.

We complete the proof by employing the triangular inequality: $V \leq V_1 + V_r \leq \frac{1}{4\lambda}$. $\qquad\square$

Part 2. In this part, we need to show that with high probability

$$V := \lambda\left\|A_{JT^c}A_{J^cT^c}^*(A_{J^cT^c}A_{J^cT^c}^*)^{-1}\left(\begin{array}{c}\operatorname{sgn}(e_S^\star)\\0_{\Omega^c}\end{array}\right)\right\|_\infty < \frac{\lambda}{4}.$$

Again by series expansion, we first have $(A_{J^cT^c}A_{J^cT^c}^*)^{-1} = \sum_{j\geq 0}(A_{J^cT^c}A_{J^cT^c}^*)^j$. Moreover, since $A_{JT^c}A_{J^cT^c}^* = -A_{JT}A_{J^cT}^*$, we arrive at

$$A_{JT^c}A_{J^cT^c}^*(A_{J^cT^c}A_{J^cT^c}^*)^{-1}\left(\begin{array}{c}\operatorname{sgn}(e_S^\star)\\0_{\Omega^c}\end{array}\right)$$
$$= \sum_{j\geq 0}A_{JT}A_{J^cT}^*(A_{J^cT}A_{J^cT}^*)^j\left(\begin{array}{c}\operatorname{sgn}(e_S^\star)\\0_{\Omega^c}\end{array}\right)$$
$$= \sum_{j\geq 0}A_{JT}(A_{J^cT}^*A_{J^cT})^j A_{ST}^*\operatorname{sgn}(e_S^\star).$$

Let $W := \sum_{j\geq 0}(A_{J^cT}^*A_{J^cT})^j A_{ST}^*$ and let $u_i \in \mathbb{R}^k$ be a row vector of A_{JT}. We consider the following bound $V_i := |\langle W^* u_i^*, \operatorname{sgn}(e_S^\star)\rangle|$. Analogous to the preceding proofs, Hoeffding's inequality is used to estimate V_i,

$$\mathbb{P}(V_i \geq \tau) \leq 2\exp\left(-\frac{2\tau^2}{4\|W^* u_i^*\|_2^2}\right).$$

The spectral norm of W can now be estimated as follows

$$\|W\| \leq \|A_{ST}^*\|\left(\sum_{j\geq 0}\|A_{J^cT}^*A_{J^cT}\|^j\right) = \frac{\|A_{ST}^*\|}{1 - \|A_{J^cT}^*A_{J^cT}\|}.$$

Conditioning on events $\mathcal{E}_1$ and $\mathcal{E}_3$ in Part 1, together with $\|u_i\|_2 \leq \sqrt{\mu\frac{k}{n}}$, we get

$$\|W^* u_i^*\|_2 \leq \|W\|\|u_i\|_2 \leq \sqrt{4\frac{2\eta\rho}{\rho_0^2}\frac{\mu k}{n}}.$$

Set $\tau = 1/4$ and take the union bound over all $i \in J$,

$$\mathbb{P}(V \geq 1/4 \mid \mathcal{E}_1, \mathcal{E}_3) \leq 2\exp\left(-\frac{\rho_0^2 n}{256\mu\eta\rho k} + \log n\right).$$

The right-hand side is less than $2e^{-\log n} = 2n^{-1}$ as long as $\frac{\rho_0^2 n}{256\mu\eta\rho k} = \frac{(m-s)^2}{256\mu s k} \geq 6\log n$. This is automatic from the assumptions that $k \leq C\frac{(1-\gamma)^2}{\gamma}\frac{m}{\mu \log n}$ and $s \leq \gamma m$. $\qquad\square$

V. PROOF OF THEOREMS 3 AND 4: DEALING WITH BOTH SPARSE AND DENSE ERRORS

A. Proof of Theorem 3

Our proof technique is adapted from [29] (see also [30]) but in a different context. In [29], the authors studied the matrix completion problem under noisy observations, while we consider the conventional compressed sensing case. Let $\widehat{x}$ be the optimal solution of (9). Since $x^\star$ is also a feasible solution of (9), $\|A_{\Omega\bullet}x^\star - b\|_2 \leq \sigma$. We have an important observation

$$\|A_{\Omega\bullet}(\widehat{x} - x^\star)\|_2 \leq \|A_{\Omega\bullet}\widehat{x} - b\|_2 + \|A_{\Omega\bullet}x^\star - b\|_2 \leq 2\sigma. \tag{32}$$

Denote $g = \widehat{x} - x^\star$, our goal is to establish a bound for $\|g\|_2$. At first, note that $\|g\|_2^2 = \|Ag\|_2^2$, the triangular inequality gives us

$$\|g\|_2^2 = \|A_{\Omega\bullet}g\|_2^2 + \|A_{\Omega^c\bullet}g\|_2^2 = 4\sigma^2 + \|A_{\Omega^c\bullet}g\|_2^2. \tag{33}$$

It now remains to bound the second term. Our strategy is to bound $\|A_{\Omega^c T}^*A_{\Omega^c\bullet}g\|_2$ and $\|A_{\Omega^c T^c}^*A_{\Omega^c\bullet}g\|_2$ separately, then the bound of $\|A_{\Omega^c\bullet}g\|_2^2$ is obtained via the following expression

$$\|A_{\Omega^c\bullet}g\|_2^2 = \|A_{\Omega^c}^*A_{\Omega^c\bullet}g\|_2^2$$
$$= \|A_{\Omega^c T}^*A_{\Omega^c\bullet}g\|_2^2 + \|A_{\Omega^c T^c}^*A_{\Omega^c\bullet}g\|_2^2, \tag{34}$$

where the first expression follows from $\|A_{\Omega^c}^*A_{\Omega^c\bullet}g\|_2^2 = \langle A_{\Omega^c\bullet}g, A_{\Omega^c}A_{\Omega^c\bullet}^*A_{\Omega^c\bullet}g\rangle = \langle A_{\Omega^c\bullet}g, A_{\Omega^c\bullet}g\rangle = \|A_{\Omega^c\bullet}g\|_2^2$.

To bound $\|A_{\Omega^c T^c}^*A_{\Omega^c\bullet}g\|_2$, we bring Lemma 2 into action: for any perturbation pair $(h, 0)$ satisfying $A_{\Omega\bullet}h = 0$, we have

$$\|x^\star + h\|_1 \geq \|x^\star\|_1 + \frac{1}{4}\|h_{T^c}\|_1. \tag{35}$$

By setting $h = A_{\Omega^c\bullet}^*A_{\Omega^c\bullet}g$, we see that $A_{\Omega\bullet}h = 0$. Hence, applying Lemma 2 yields

$$\|x^\star + A_{\Omega^c\bullet}^*A_{\Omega^c\bullet}g\|_1 \geq \|x^\star\|_1 + \frac{1}{4}\|A_{\Omega^c T^c}^*A_{\Omega^c\bullet}g\|_1. \tag{36}$$

Furthermore, noting that $x^\star + g$ is the optimal solution of the convex program (9). This yields

$$\|x^\star\|_1 \geq \|x^\star + g\|_1 \geq \|x^\star + A_{\Omega^c\bullet}^*A_{\Omega^c\bullet}g\|_1 - \|A_{\Omega\bullet}^*A_{\Omega\bullet}g\|_1.$$

In combination with (36), we have an important inequality: $\|A_{\Omega^c T^c}^*A_{\Omega^c\bullet}g\|_1 \leq 4\|A_{\Omega\bullet}^*A_{\Omega\bullet}g\|_1$. Since the ℓ_1-norm dominates the ℓ_2-norm, $\|A_{\Omega^c T^c}^*A_{\Omega^c\bullet}g\|_2 \leq \|A_{\Omega^c T^c}^*A_{\Omega^c\bullet}g\|_1$ and we have

$$\|A_{\Omega^c T^c}^*A_{\Omega^c\bullet}g\|_2 \leq 4\|A_{\Omega\bullet}^*A_{\Omega\bullet}g\|_1$$
$$\leq 4\sqrt{n}\|A_{\Omega\bullet}^*A_{\Omega\bullet}g\|_2 \tag{37}$$
$$= 4\sqrt{n}\|A_{\Omega\bullet}g\|_2.$$

It is left to develop a bound for $\|A_{\Omega^c T}^*A_{\Omega^c\bullet}g\|_2$. We observe that $A_{\Omega T}A_{\Omega^c T}^* = -A_{\Omega T^c}A_{\Omega^c T^c}^*$ due to the orthogonality property of A. Thus, for any vector u,

$$\|A_{\Omega T}A_{\Omega^c T}^*u\|_2 = \|A_{\Omega T^c}A_{\Omega^c T^c}^*u\|_2$$
$$\leq \|A_{\bullet T^c}A_{\Omega^c T^c}^*u\|_2 = \|A_{\Omega^c T^c}^*u\|.$$

In addition, applying Lemma 5 with $\Omega \sim \mathrm{Ber}(\eta)$ we have $\left\| I - \eta^{-1} A_{\Omega T}^* A_{\Omega T} \right\| \leq \epsilon \leq 1/2$ with high probability. Hence,

$$
\begin{aligned}
&\eta^{-1} \left\| A_{\Omega T} A_{\Omega^c T}^* u \right\|_2^2 \\
&= \eta^{-1} \left\langle A_{\Omega T} A_{\Omega^c T}^* u, A_{\Omega T} A_{\Omega^c T}^* u \right\rangle \\
&= \eta^{-1} \left\langle A_{\Omega^c T}^* u, A_{\Omega T}^* A_{\Omega T} A_{\Omega^c T}^* u \right\rangle \\
&= \left\langle A_{\Omega^c T}^* u, A_{\Omega^c T}^* u \right\rangle - \left\langle A_{\Omega^c T}^* u, (I - \eta^{-1} A_{\Omega T}^* A_{\Omega T}) A_{\Omega^c T}^* u \right\rangle \\
&\geq \left\| A_{\Omega^c T}^* u \right\|_2^2 - \left\| I - \eta^{-1} A_{\Omega T}^* A_{\Omega T} \right\| \left\| A_{\Omega^c T}^* u \right\|_2^2 \\
&\geq \frac{1}{2} \left\| A_{\Omega^c T}^* u \right\|_2^2 .
\end{aligned}
$$

In other words, $\sqrt{\eta/2} \left\| A_{\Omega^c T}^* u \right\|_2 \leq \left\| A_{\Omega T} A_{\Omega^c T}^* u \right\|_2$. Combining these pieces together while setting $u = A_{\Omega^c \bullet} g$ yields

$$
\left\| A_{\Omega^c T}^* A_{\Omega^c \bullet} g \right\|_2 \leq \frac{1}{\sqrt{\eta/2}} \left\| A_{\Omega^c T^c}^* A_{\Omega^c \bullet} g \right\|_2 .
$$

The right-hand side is in turn bounded by $\frac{4\sqrt{n}}{\sqrt{\eta/2}} \left\| A_{\Omega^c T^c}^* \right\|_2$ due to (37). Inserting this bound with the bound in (37) into (34), we obtain

$$
\left\| A_{\Omega^c \bullet} g \right\|_2^2 \leq \left(\frac{2}{\eta} + 1 \right) 16n \left\| A_{\Omega \bullet} g \right\|_2^2 \leq \left(\frac{2}{\eta} + 1 \right) 64 \sigma^2 n,
$$

where the last inequality follows from the known bound in (32). Combining this result with (33) we can conclude that

$$
\left\| g \right\|_2 \leq 2\sigma + 8\sigma \sqrt{ n \left(\frac{2}{\eta} + 1 \right) },
$$

as claimed.

B. Proof of Theorem 4

Proof. The proof of this theorem is considerably more involved since we have to control two residual components $\widehat{x} - x^\star$ and $\widehat{e} - e^\star$, where $(\widehat{x}, \widehat{e})$ is the optimal solution pair of (12). Set $g^{(x)} = \widehat{x} - x^\star$ and $g^{(e)} = \widehat{e} - e^\star$, our goal is to bound $\left\| g^{(x)} \right\|_2 + \left\| g^{(e)} \right\|_2$.

At first, notice that $(x^\star, e^\star)$ and $(\widehat{x}, \widehat{e})$ are pairs of feasible solution, we establish an important bound

$$
\begin{aligned}
\left\| A_{\Omega \bullet} g^{(x)} + g^{(e)} \right\|_2 &\leq \left\| A_{\Omega \bullet} \widehat{x} + \widehat{e} - b \right\|_2 \\
&\quad + \left\| A_{\Omega \bullet} x^\star + e^\star - b \right\|_2 \leq 2\sigma.
\end{aligned}
\tag{38}
$$

To bound $\left\| g^{(x)} \right\|_2 + \left\| g^{(e)} \right\|_2$, we first express $\left\| g^{(x)} \right\|_2$ as $\left\| g^{(x)} \right\|_2^2 = \left\| A g^{(x)} \right\|_2^2 = \left\| A_{\Omega^c \bullet} g^{(x)} \right\|_2^2 + \left\| A_{\Omega \bullet} g^{(x)} \right\|_2^2$. Furthermore, from the fact that $\frac{1}{2} \left\| a + b \right\|_2^2 + \frac{1}{2} \left\| a - b \right\|_2^2 = \left\| a \right\|_2^2 + \left\| b \right\|_2^2$ for any vectors a and b, we get

$$
\begin{aligned}
&\left\| g^{(x)} \right\|_2^2 + \left\| g^{(e)} \right\|_2^2 \\
&= \left\| A_{\Omega^c \bullet} g^{(x)} \right\|_2^2 + \left\| A_{\Omega \bullet} g^{(x)} \right\|_2^2 + \left\| g^{(e)} \right\|_2^2 \\
&\leq \left\| A_{\Omega^c \bullet} g^{(x)} \right\|_2^2 + \frac{1}{2} \left\| A_{\Omega \bullet} g^{(x)} + g^{(e)} \right\|_2^2 + \frac{1}{2} \left\| A_{\Omega \bullet} g^{(x)} - g^{(e)} \right\|_2^2 \\
&\leq 2\sigma^2 + \left\| A_{\Omega^c \bullet} g^{(x)} \right\|_2^2 + \frac{1}{2} \left\| A_{\Omega \bullet} g^{(x)} - g^{(e)} \right\|_2^2 .
\end{aligned}
\tag{39}
$$

It is left to bound the sum of the second and third term on the right-hand side of the equation. We express this sum as

$$
\begin{aligned}
&\left\| A_{\Omega^c \bullet} g^{(x)} \right\|_2^2 + \frac{1}{2} \left\| A_{\Omega \bullet} g^{(x)} - g^{(e)} \right\|_2^2 \\
&= \left\| A_{\Omega^c \bullet} g^{(x)} \right\|_2^2 + \frac{1}{2} \left\| A_{S \bullet} g^{(x)} - g_S^{(e)} \right\|_2^2 + \frac{1}{2} \left\| A_{J \bullet} g^{(x)} - g_J^{(e)} \right\|_2^2 ,
\end{aligned}
$$

where we recall that indices in S are locations where measurements are available but unreliable and indices in J are locations where measurements are available and trustworthy and $\Omega = S \cup J$. To upper bound this sum, we consider the establishment of the upper bounds for each term $M_1 := \left\| A_{\Omega^c \bullet} g^{(x)} \right\|_2^2 + \left\| A_{S \bullet} g^{(x)} - g_S^{(e)} \right\|_2^2$ and $M_2 := \left\| A_{J \bullet} g^{(x)} - g_J^{(e)} \right\|_2^2$ separately.

One of the crucial steps in bounding M_1 and M_2 is the use of Lemma 2, which states that for any perturbation pair (h, f) satisfying $f = -A_{\Omega \bullet} h$,

$$
\begin{aligned}
\left\| x^\star + h \right\|_1 + \lambda \left\| e^\star + f \right\|_1 &\geq \left\| x^\star \right\|_1 + \lambda \left\| e^\star \right\|_1 \\
&\quad + \frac{\lambda}{4} \left\| f_J \right\|_1 + \frac{1}{4} \left\| h_{T^c} \right\|_1 .
\end{aligned}
$$

Now let us denote

$$
f^+ := -\frac{1}{2}(A_{\Omega \bullet} g^{(x)} + g^{(e)}) \quad \text{and} \quad f^- := -\frac{1}{2}(A_{\Omega \bullet} g^{(x)} - g^{(e)}),
$$

as well as

$$
h^+ := -A_{\Omega \bullet}^* f^+ \quad \text{and} \quad h^- := -A_{\Omega \bullet}^* f^- + A_{\Omega^c \bullet}^* A_{\Omega^c \bullet} g^{(x)} .
$$

It is easy to establish the following properties from this construction

$$
\begin{cases}
g^{(x)} = -h^+ + h^- \\
g^{(e)} = -f^+ + f^- \\
\left\| h^+ \right\|_2 = \left\| f^+ \right\|_2 \leq \sigma \\
M_1 = \left\| A_{\Omega^c \bullet} g^{(x)} \right\|_2^2 + 2 \left\| f_S^- \right\|_2^2 \\
M_2 = 2 \left\| f_J^- \right\|_2^2 .
\end{cases}
\tag{40}
$$

1) Bound M_2: At first, since $(x^\star + g^{(x)}, e^\star + g^{(e)})$ is the pair of optimal solution of the convex program, we have $\left\| x^\star \right\|_1 + \left\| e^\star \right\|_1 \geq \left\| x^\star + g^{(x)} \right\|_1 + \left\| e^\star + g^{(e)} \right\|_1$. Furthermore, decomposing $g^{(x)}$ and $g^{(e)}$ and using the triangular inequality, we can derive

$$
\begin{aligned}
&\left\| x^\star + g^{(x)} \right\|_1 + \lambda \left\| e^\star + g^{(e)} \right\|_1 \\
&= \left\| x^\star - h^+ + h^- \right\|_1 + \lambda \left\| e^\star - f^+ + f^- \right\|_1 \\
&\geq \left\| x^\star + h^- \right\|_1 + \lambda \left\| e^\star + f^- \right\|_1 - (\left\| h^+ \right\|_1 + \lambda \left\| f^+ \right\|_1).
\end{aligned}
\tag{41}
$$

Applying Lemma 2 together with the observation that $f^- = -A_{\Omega \bullet} h^-$ yields

$$
\begin{aligned}
&\left\| x^\star + h^- \right\|_1 + \lambda \left\| e^\star + f^- \right\|_1 \\
&\qquad \geq \left\| x^\star \right\|_1 + \lambda \left\| e^\star \right\|_1 + \frac{\lambda}{4} \left\| f_J^- \right\|_1 + \frac{1}{4} \left\| h_{T^c}^- \right\| .
\end{aligned}
$$

Combining these arguments, we get

$$
\frac{\lambda}{4} \left\| f_J^- \right\|_1 + \frac{1}{4} \left\| h_{T^c}^- \right\|_1 \leq \left\| h^+ \right\|_1 + \lambda \left\| f^+ \right\|_1 .
$$

Converting both sides to the ℓ_2-norm using the crude inequality $\|a\|_2 \leq \|a\|_1 \leq \sqrt{n}\|a\|_2$ for all $a \in \mathbb{R}^n$, then applying $\|f^+\|_2 = \|h^+\|_2 \leq \sigma$, we obtain the bound

$$\frac{\min\{\lambda, 1\}}{4}(\|f_J^-\|_2 + \|h_{T^c}^-\|_2) \leq \sqrt{n}(1+\lambda)\|f^+\|_2 \quad (42)$$
$$\leq \sqrt{n}(1+\lambda)\sigma.$$

A specific consequence of this analysis is a bound of M_2

$$M_2 = 2\|f_J^-\|_2^2 \leq 2(\|f_J^-\|_2 + \|h_{T^c}^-\|_2)^2$$
$$\leq 2\left(\frac{4(1+\lambda)}{\min\{\lambda, 1\}}\right)^2 \sigma^2 n. \quad (43)$$

2) Bound M_1: In this section we would like to bound $\|A_{\Omega^c \bullet} g^{(x)}\|_2^2 + 2\|f_S^-\|_2^2$. Denoting $z := \begin{pmatrix} f_S^- \\ -A_{\Omega^c \bullet} g^{(x)} \end{pmatrix}$, then to bound the quantity of interest, it is equivalent to bounding $\|z\|_2$. By the construction of h^- and f^-, we have $A_{\Omega \bullet}^* f^- + h^- - A_{\Omega^c \bullet}^* A_{\Omega^c \bullet} g^{(x)} = 0$ leading to

$$-A_{J \bullet}^* f_J^- - \begin{pmatrix} 0_T \\ h_{T^c}^- \end{pmatrix} = A_{S \bullet}^* f_S^- - A_{\Omega^c \bullet}^* A_{\Omega^c \bullet} g^{(x)} + \begin{pmatrix} h_T^- \\ 0_{T^c} \end{pmatrix}$$
$$= A_{J^c \bullet}^* \begin{pmatrix} f_S^- \\ -A_{\Omega^c \bullet} g^{(x)} \end{pmatrix} + \begin{pmatrix} h_T^- \\ 0_{T^c} \end{pmatrix}$$
$$= A_{J^c \bullet}^* z + \begin{pmatrix} h_T^- \\ 0_{T^c} \end{pmatrix}, \quad (44)$$

where the second identity follows from $J^c = S \cup \Omega^c$.

First we control the upper bound of the ℓ_2-norm of the left-hand side of (44), which can be attained easily from the triangular inequality

$$\left\| A_{J \bullet}^* f_J^- + \begin{pmatrix} 0_T \\ h_{T^c}^- \end{pmatrix} \right\|_2 \leq \|A_{J \bullet}^* f_J^-\|_2 + \|h_{T^c}^-\|_2$$
$$= \|f_J^-\|_2 + \|h_{T^c}^-\|_2.$$

Next, the ℓ_2-norm of the right-hand side of (44) is now lower bounded by

$$\left\| A_{J^c \bullet}^* z + \begin{pmatrix} h_T^- \\ 0_{T^c} \end{pmatrix} \right\|_2^2$$
$$= \|A_{J^c \bullet}^* z\|_2^2 + \|h_T^-\|_2^2 + 2\langle A_{J^c T}^* z, h_T^- \rangle$$
$$\geq \|z\|_2^2 + \|h_T^-\|_2^2 - 2\|A_{J^c T}^*\|\|z\|_2\|h_T^-\|_2$$
$$\geq \|z\|_2^2 + \|h_T^-\|_2^2 - 2\|A_{J^c T}^*\|\|z\|_2\|h_T^-\|_2$$
$$\geq \|z\|_2^2 + \|h_T^-\|_2^2 - 2\sqrt{1-\rho_0/2}\|z\|_2\|h_T^-\|_2$$
$$\geq (1 - \sqrt{1-\rho_0/2})(\|z\|_2^2 + \|h_T^-\|_2^2),$$

where the third inequality follows from Proposition 1: $\|A_{J^c T}^* A_{J^c T}\| \leq 1 - \rho_0/2$ and the last inequality follows from the standard argument $a^2 + b^2 - 2\alpha ab \geq (1 - \alpha)(a^2 + b^2)$.

Combine these pieces together with the fact that $1 - \sqrt{1-\rho_0/2} \geq \frac{\rho_0}{4}$, we attain

$$\|z\|_2^2 + \|h_T^-\|_2^2 \leq \frac{4}{\rho_0}(\|f_J^-\|_2 + \|h_{T^c}^-\|_2)^2. \quad (45)$$

Next, notice that $\|z\|_2^2 = \|f_S^-\|_2^2 + \|A_{\Omega^c \bullet} g^{(x)}\|_2^2$ and together with (43), we get the following bound of M_1

$$M_1 \leq 2(\|z\|_2^2 + \|h_T^-\|_2^2) \leq \frac{8}{\rho_0}(\|f_J^-\|_2 + \|h_{T^c}^-\|_2)^2$$
$$\leq \frac{8}{\rho_0}\left(\frac{4(1+\lambda)}{\min\{1, \lambda\}}\right)^2 \sigma^2 n. \quad (46)$$

Obviously, from combining these two previous inequalities on M_1 and M_2, we can establish the bound of the sum $M_1 + M_2$. However, we can tighten this bound by a constant factor from the following simple steps:

$$M_1 + M_2 \leq 2(\|A_{\Omega^c \bullet} g^{(x)}\|_2^2 + \|f_S^-\|_2^2 + \|f_J^-\|_2^2)$$
$$\leq 2\left[\frac{4}{\rho_0}(\|f_J^-\|_2 + \|h_{T^c}^-\|_2)^2 + \|f_J^-\|_2^2\right]$$
$$\leq 2\left(\frac{4}{\rho_0} + 1\right)(\|f_J^-\|_2 + \|h_{T^c}^-\|_2)^2$$
$$\leq 2\left(\frac{4}{\rho_0} + 1\right)\left(\frac{4(1+\lambda)}{\min\{\lambda, 1\}}\right)^2 \sigma^2 n,$$

where the second inequality follows from (45) and the last inequality follows from (42).

Inserting the above bound into (39) leads to

$$\|g^{(x)}\|_2^2 + \|g^{(e)}\|_2^2 \leq 2\sigma^2 + 2\left(\frac{4}{\rho_0} + 1\right)\left(\frac{4(\lambda+1)}{\min\{1, \lambda\}}\right)^2 \sigma^2 n.$$

Finally, applying $(\|g^{(x)}\|_2 + \|g^{(e)}\|_2)^2 \leq 2(\|g^{(x)}\|_2^2 + \|g^{(e)}\|_2^2)$ will complete our proof. □

VI. ORACLE INEQUALITIES

In this section we would like to discuss the optimality of the reconstruction error bound in Theorem 4. In particular, we compare this result with the best possible accuracy one can achieve. Suppose we had available an oracle informing us in advance the locations of T nonzero coefficients of the signal as well as S nonzero coefficients of the sparse noise. Then one can use this valuable information to construct the ideal estimator pair $(x^{\text{Oracle}}, e^{\text{Oracle}})$ by least-square projection. To see this, we decompose y into two components: y_S and y_J, where y_J is not affected by sparse error. Thus,

$$y_J = A_{JT} x_T^\star + \nu_J.$$

Recall from (30), $A_{JT}^* A_{JT}$ is invertible. In particular, $\rho_0/2 \leq \sigma_{\min}(A_{JT}^* A_{JT}) \leq \sigma_{\max}(A_{JT}^* A_{JT}) \leq 3\rho_0/2$ where $\sigma_{\min}$ and $\sigma_{\max}$ are the minimum and the maximum singular value of the matrix, respectively. Therefore, the least-square solution of this linear system is

$$x_T^{\text{Oracle}} = (A_{JT}^* A_{JT})^{-1} A_{JT}^* y_J.$$

The oracle error bound on the signal is now estimated by

$$\|x_T^{\text{Oracle}} - x_T^\star\|_2 = \|(A_{JT}^* A_{JT})^{-1} A_{JT}^* \nu_J\|_2.$$

It is obvious that $\|H^{-1}\| \leq \frac{1}{\sigma_{\min}(H)}$ for any matrix H. Therefore,

$$\|x_T^{\text{Oracle}} - x_T^\star\|_2 \leq \|(A_{JT}^* A_{JT})^{-1}\|\|A_{JT}\|\|\nu_J\|_2 \leq \sigma\sqrt{6/\rho_0}. \quad (47)$$

Now the oracle solution of the error can be found from the identity equation on the set S: $y_S = A_{ST} x_T^{\text{Oracle}} + e_S^{\text{Oracle}} + \nu_S$. This leads to

$$e_S^{\text{Oracle}} = e_S^\star + A_{ST}(x_T^\star - x_T^{\text{Oracle}}).$$

Recall in Proposition 2 that

$$\|A_{ST}\| \leq (\eta\rho)^{1/2} \left(1 + \sqrt{\frac{\mu k \log n}{s}}\right)^{1/2}$$

$$= \eta^{1/2} \left(\frac{s}{m} + \sqrt{\frac{s}{m}} \sqrt{\frac{\mu k \log n}{m}}\right)^{1/2} \leq \sqrt{2},$$

provided that $m \geq \mu k \log n$. We conclude that the oracle error bound on $e^\star$ has to satisfy

$$\left\|e_S^{\text{Oracle}} - e_S^\star\right\|_2 \leq \sqrt{2} \left\|x_T^{\text{Oracle}} - x_T^\star\right\|_2 \leq \sqrt{12/\rho_0}.$$

In conclusion, with the help of the oracle, we have

$$\left\|x^{\text{Oracle}} - x^\star\right\|_2 + \left\|e^{\text{Oracle}} - e^\star\right\|_2 \leq 3\sigma\sqrt{6/\rho_0}. \quad (48)$$

with adversarial noise. Consequently, our error bound in Theorem 4 loses a $\sqrt{n}$ vis-a-vis over the ideal bound achieved via the oracle help.

VII. NUMERICAL EXPERIMENTS

In this section, we provide extensive simulations to confirm the validity of our theoretical results. Since the observation model in (4) can be expressed as $y = [A_{\Omega\bullet} \quad \frac{1}{\lambda}I]z^\star = Bz^\star$ where $z^\star = [x^{\star^T}, \lambda e^{\star^T}]^T$ and I is the $m \times m$ identity matrix, the extended ℓ_1-minimization in (5) and the noisy version in (12) can be recast as conventional ℓ_1 programs

$$\min_z \|z\|_1 \quad \text{s.t.} \quad y = Bz,$$

and

$$\min_z \|z\|_1 \quad \text{s.t.} \quad \|b - Bz\| \leq \sigma.$$

In this section, we use the Homotopy solver introduced in [31] for our experiments. Another important implementation detail is the choice of the parameter λ. For moderate signal dimensions (e.g $n \leq 10^8$), we suggest to set $\lambda = \sqrt{\frac{n}{m(\log n)^{1/2}}}$. With this choice, measurements are allowed to be corrupted up to 25% as presented in our theorems. Of course, if we know in prior that the signal is very sparse, reducing the value of λ will help retrieve the signal under more corrupted measurements. In practical applications, we recommend $\lambda = \sqrt{\frac{n}{m(\log n)^{1/2}}}$ as a "good-for-all" parameter.

A. Exact recovery from grossly corrupted measurements

We first illustrate the correct recoverability of the signal under gross error as provided in Theorem 2. We consider random signals $x^\star$ of varying lengths $n = \{1024, 2048, 4096, 8192\}$. For each n, we generate signals of sparsity k where k varies from 1 to 60 with step size 2. Here, magnitudes of nonzero entries are Gaussian distributed and their locations are chosen uniformly at random. For each sparsity level, the measurement matrix $A_{\Omega\bullet}$ is produced by uniformly selecting $m = 500$
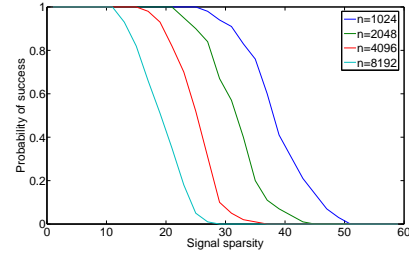

Fig. 1. The probability of success as a function of signal sparsity for various signal dimensions. Here, a total of $m = 500$ measurements are observed and $1/4$ of them are grossly corrupted.
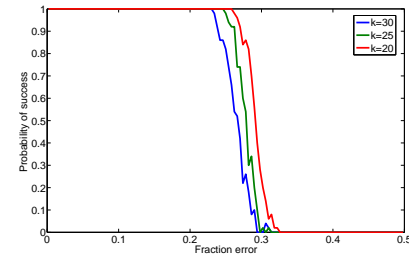

Fig. 2. The probability of success as a function of fraction error s/m. Here, we fix signal dimension to $n = 1024$, a total of $m = 500$ measurements are used and the signal sparsity is $k = [25, 30, 35]$.

rows at random from the Fourier matrix A. Error vector $e^\star$ is generated to have uniformly distributed support with cardinality $s = m/4$ and the polarity of nonzero entries being equally likely positive or negative. We set magnitudes of $e^\star$ such that $\|e^\star\|_2 \geq 100 \|x^\star\|_2$. The reader should note that this setting yields an observed signal that is significantly dominated by the noise.

For each value of the signal sparsity k, we repeat the experiment 100 times and keep track of the probability of exact recovery. In all experiments, we set $\lambda = \sqrt{\frac{n}{m(\log n)^{1/2}}}$. The algorithm is declared to be successful if the relative error with respect to $x^\star$ satisfies $\|\hat{x} - x^\star\|_2 / \|x^\star\|_2 \leq 10^{-3}$. The performance curve is plotted in Fig. 1. Numerical values on the x-axis denote signal sparsity whereas those on the y-axis denote the probability of exact recovery. Interestingly, this experiment demonstrates that the theory provides an accurate prediction of the simulation results even for relatively small problem sizes. In particular, perfect recovery is still attained with signals of moderate sparsity level even if 25% measurements are grossly perturbed. Furthermore, the sparsity level is proportional with $\frac{m}{(\log n)^{3/2}}$ as expected.

Next, we fix the signal dimension to $n = 1024$ and performs the same experiments with varying signal sparsity $k = [20, 25, 30]$. Fig 2 demonstrates the probability of success with varying fraction error s/m. Note that as the signal's sparsity level increases, we expect to recover the signal under fewer corrupted measurements.
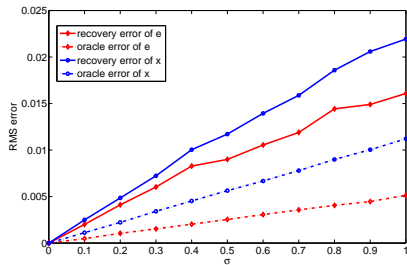
Fig. 3. RMS error as a function of σ with $n = 1024$, $m = 500$, $k = 20$ and $s = m/4$.
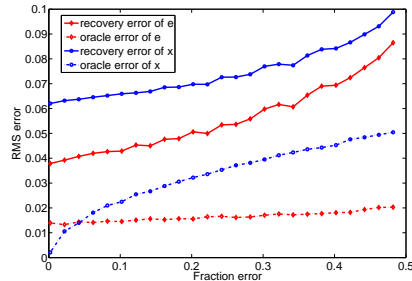

Fig. 4. RMS error as a function of s with $n = 1024$, $m = 500$, $k = 20$ and $\sigma = 1$.

B. Stable recovery from both dense and sparse corrupted measurements

We now demonstrate stable recoverability when measurements are both contaminated by gross sparse and small dense error. We generate small noise ν from i.i.d. $\mathcal{N}(0, \delta^2)$. The signal $x^\star$, the sparse error $e^\star$ and the measurement matrix $A_{\Omega\bullet}$ are constructed similarly as in previous experiments. For each setting, we perform the simulations 100 times and report the average error.

We first evaluate the performance of (12) with the signal $x^\star$ whose dimension and sparsity level are fixed to be $n = 1024$ and $k = 20$. We also set the number of measurements and the error sparsity to be $m = 500$ and $s = m/4$. Non-zero entries of the signal and the sparse errors are i.i.d. $\mathcal{N}(0, 10)$. Estimation errors are quantified by the root-mean square (RMS), which is defined as $\|\widehat{x} - x^\star\|_2 / n$ and $\|\widehat{e} - e^\star\|_2 / n$, respectively. Fig. 3 shows the RMS error with varying noise level. We also demonstrate in this figure the RMS errors of an oracle obtained from Section VI. Fig. 3 clearly illustrates that the RMS errors grow almost linearly with the noise level. Furthermore, RMS errors attained by solving (12) is just twice the RMS error achieved by the oracle.

Now we fix $\sigma = 1$ and run the optimization in (12) for varying values of error sparsity. Fig. 4 establishes fact that as s decreases, we expect to achieve more accurate recovery.

C. Experiments with images

In our last experiment, we consider the problem of recovering an image from highly corrupted undersampled Fourier coefficients. As usual, the data is given by $y = A_{\Omega\bullet}x^\star + e^\star + \nu$ where $A_{\Omega\bullet}$ is a partial Fourier matrix obtained from

subsampling rows of the full 2D Fourier matrix A, $e^\star$ is a sparse error vector whose nonzero entries can have arbitrarily large magnitudes, and ν is a small dense noise vector. In this experiment, $x^\star$ is the Shepp-Logan phantom image (see Fig. 5), which is not sparse in the spatial domain but in the gradient domain. Therefore, to reconstruct $x^\star$, we use the total variation (TV) criterion and minimize

$$\min_{x,e} \|x\|_{\mathrm{TV}} + \lambda \|e\|_1 \quad \text{s.t.} \quad \|y - A_{\Omega\bullet}x - e\|_2 \le \sigma, \quad (49)$$

where $\|\nu\|_2 \le \sigma$ is assumed to be known and $\|x\|_{\mathrm{TV}}$ is the ℓ_1-norm of the gradient, also known as the total-variation of x. This norm is formally defined as

$$\|x\|_{\mathrm{TV}} = \sum_{ij} \sqrt{(\nabla_h x)_{ij}^2 + (\nabla_v x)_{ij}^2}, \quad (50)$$

where ∇_h and ∇_v denote the discrete finite difference operators along the horizonal and vertical coordinates. To optimize (49), we employ the classic alternating direction method (ADM) as presented in [32]. In this particular experiment, we perform a two-step algorithm

1) We solve (49) via the ADM method. The optimal solution is denoted as $(\widehat{x}, \widehat{e})$.
2) Next, we select $J \in \{1, ..., m\}$ as locations where coefficients of $\widehat{e}$ are zeros or approximately zeros. These locations correspond to reliable observations. Then, we solve the following optimization

$$\min_{x} \|x\|_{\mathrm{TV}} \quad \text{s.t.} \quad \|y_J - A_{J\bullet}x\|_2 \le \sigma, \quad (51)$$

where only clean observations are considered. The output of (51) is what we expect to get.

In this experiment, we sample 12267 Fourier coefficients of the 256×256 phantom image $x^\star$ along a number of radical lines (as seen in the top right of Fig. 5, 45 radical lines are sampled). We then select 50% of these coefficients uniformly at random and purposely add them to a deterministic large noise vector whose magnitudes are twice larger than the magnitudes of Fourier coefficients. This process assumes that half of the observed Fourier coefficients are significantly corrupted during the data acquisition. We note that the locations of these missing entries are unknown. All the Fourier coefficients is afterward contaminated by a Gaussian noise vector with zero mean and standard deviation 0.01. Fig. 5 on the bottom left and right shows the reconstruction from minimizing the TV only and from the aforementioned two-step algorithm, respectively. In the optimization (49), λ is set to be $\sqrt{\frac{n}{m \log n}}$. It is clear that while the conventional TV minimization fails to recover the original image, our proposed method recovers the image almost exactly. Notably, the relative error $\frac{\|x^\star - x_{\mathrm{recovered}}\|_2}{\|x^\star\|_2}$ of our method is 0.0887.

VIII. DISCUSSION AND CONCLUSION

In this paper, we present a complete analysis of a surprising phenomenon: one can recover perfectly a sparse signal from grossly corrupted measurements by linear programming (5), even if the corruption is up to a significant fraction of all the entries. More specifically, we deliver an explicit connection
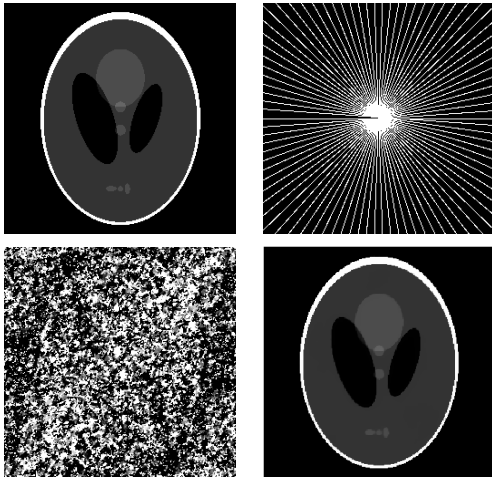
Fig. 5. Top left: original 256×256 phantom image. Top right: Fourier domain sampling positions with 45 radical lines. Bottom left: recovered image from the TV only. Bottom right: recovered image from our proposed optimization in (49).

between sparsity levels of the signal and the error. Our result can be interpreted as a generalization of compressed sensing, where measurements are both incomplete and corrupted by sparse errors. Furthermore, our results indicate that robustness is still retained even in a more challenging situation: the convex program (12) can stably recover a sparse signal under measurements perturbed by both gross sparse and small dense errors. Particularly, recovery error lies within a constant fraction of the dense noise level. We also establish stable recovery for a much more general class of signals − approximately sparse signals.

As exhibited in Theorem 1, when the fraction of error is close to 1 − or in other words, most of the measurements are corrupted, signal sparsity k is still allowed to be proportional to $\frac{m}{\mu^2 log^2 n}$ in order to retain accurate recovery. We conjecture that this bound is optimal. That is, we cannot achieve perfect reconstruction when $k \sim O(\frac{m}{\mu^2 \log n})$ and the error support size s is close to m. In fact, we claim this conjecture in our upcoming paper for a class of Gaussian measurement matrices [33]. How to establish a similar result for suborthogonal measurement matrices is an interesting open problem.

We would like to mention a related work that describes a similar phenomenon. Recently, Candès *et al.* [13], [34], Chandrasekaran *et al.* [35], Xu *et al.* [36] and Agarwal *et al.* [37] have shown that one can exactly recover a low-rank matrix $L \in \mathbb{R}^{n_1 \times n_2}$ from its grossly corrupted entries $M = L + S$ by solving the following convex program:

$$\min_{L,S} \|L\|_* + \lambda \|S\|_1 \qquad \text{subject to} \qquad M = L + S. \quad (52)$$

More specifically, the authors of [13], [34] proved that as long as the rank of L is an order of $\frac{n}{\log^2 n}$ with $n = \max\{n_1, n_2\}$, then the solution of (52) with an appropriate choice of parameter λ is exact even if almost all entries of L are arbitrarily perturbed. Interestingly, the results in these papers shares similar behavior as what presented here in our paper. We believe that similar phenomena also holds for other high-dimensional signal and error models as well.

IX. APPENDIX

Proof of Corollary 1. At first, we observe a variant of Lemma 2. Assuming the existence of a dual vector $(z^{(x)}, z^{(e)})$ satisfying properties of Lemma 2, then for any perturbation pair (f, h) such that $f = -A_{\Omega \bullet} h$, we have

$$\begin{aligned} \|x^\star + h\|_1 + \lambda \|e^\star + f\|_1 &\geq \|x_T^\star\|_1 + \lambda \|e^\star\|_1 \\ &- \|x_{T^c}^\star\|_1 + \frac{1}{4}(\|h_{T^c}\|_1 + \lambda \|A_{J\bullet} h\|_1). \end{aligned} \quad (53)$$

The proof is essentially analogous to that of Lemma 2. The only difference is the non-sparse nature of $x^\star$. Now decompose $x^\star$ into $x_T^\star$ and $x_{T^c}^\star$ and use the triangular inequality to provide a lower bound for $\|x^\star + h\|_1$, we have

$$\|x^\star + h\|_1 + \lambda \|e^\star + f\|_1 \geq \|x_T^\star + h\|_1 + \lambda \|e^\star + f\|_1 - \|x_{T^c}^\star\|_1.$$

Applying Lemma 2 to the bound $\|x_T^\star + h\|_1 + \lambda \|e^\star + f\|_1$ will lead to the inequality (53).

Following closely the proof of Theorem 4, except in bounding the quantity M_2, we employ the inequality in (53). With the same notations, we have $\|x^\star + g^{(x)}\|_1 + \|e^\star + g^{(e)}\|_1 \leq \|x^\star\|_1 + \|e^\star\|_1 = \|x_T^\star\|_1 + \|x_{T^c}^\star\|_1 + \|e^\star\|_1$. Using the lower bound of $\|x^\star + g^{(x)}\|_1 + \|e^\star + g^{(e)}\|_1$ in (41) together with (53), we get a similar result as in (42)

$$\frac{\min\{\lambda, 1\}}{4}(\|f_J^-\|_2 + \|h_{T^c}^-\|_2) \leq \sqrt{n}(1 + \lambda)\sigma + 2 \|x_{T^c}^\star\|_1.$$

The rest of our proof follows exactly from the analysis of Theorem 4. □

Proof of Lemma 6. The proof is essentially analogous to the one presented in [14]. We first establish a bound for $\mathbb{E} \|A_{S_0 T}^* u\|_2$, and then show that $\|A_{S_0 T}^* u\|_2$ concentrates around its expectation.

Define $S_0 = \{i : \delta_i = 1\}$ where δ_i is an independent sequence of Bernoulli variables with parameter ρ_0 and denote by $v_i \in R^k$ the ith column of matrix $A_{S_0 T}^*$. With these notations, we have

$$A_{S_0 T}^* u = \sum_{i \in S_0} u_i v_i = \sum_{i=1}^n \delta_i u_i v_i.$$

Notice that from the orthogonality property of A, $\sum_{i=1}^n u_i v_i = A_{\bullet T}^* a = 0$ where a is a column of matrix $A_{\bullet T^c}$. Thus, by subtracting this zero term from $A_{S_0 T}^* u$, one can see that $A_{S_0 T}^* u$ is a sum of zero-mean random variable

$$A_{S_0 T}^* u = \sum_{i=1}^n (\delta_i - \rho_0) u_i v_i.$$

We can now estimate $\mathbb{E} \|A_{S_0 T}^* u\|_2$ as follows

$$\begin{aligned} \mathbb{E} \|A_{S_0 T}^* u\|_2^2 &= \mathbb{E} \sum_{i=1}^n (\delta_i - \rho_0)^2 u_i^2 \langle v_i, v_i \rangle \\ &+ \mathbb{E} \sum_{i,j; i \neq j} (\delta_i - \rho_0)(\delta_j - \rho_0) u_i u_j \langle v_i, v_j \rangle. \end{aligned}$$

The second term vanishes due to the independence of δ_i, $i = 1, ..., n$. Furthermore, $\mathbb{E}(\delta_i - \rho_0)^2 = \rho_0(1 - \rho_0) \leq \rho_0$. Hence,

$$\mathbb{E} \left\| A^*_{S_0 T} u \right\|_2^2 \leq \rho_0 \max_i \|v_i\|_2^2 \left(\sum_{i=1}^n u_i^2 \right)$$

$$= \rho_0 \max_i \|v_i\|_2^2 \leq \rho_0 \frac{\mu k}{n}.$$

Therefore, by Jensen's inequality, we conclude that $\mathbb{E} \left\| A^*_{S_0 T} u \right\|_2 \leq \sqrt{\mathbb{E} \left\| A^*_{S_0 T} u \right\|_2^2} \leq \sqrt{\rho_0 \frac{\mu k}{n}}$.

We now apply a remarkable result from Talagrand that bounds the supremum of a sum of independent random variables. Let $Z_1, ..., Z_n$ be a sequence of independent random variables and let M be the supremum defined by

$$M = \sup_{g \in \mathcal{G}} \sum_{i=1}^n g(Z_i),$$

where g is a family of real-valued functions.

Theorem 5. *If $|g| \leq B$ for every $g \in \mathcal{G}$ and $\{g(Z_i)\}_{i=1,...,n}$ have zero mean for every $g \in \mathcal{G}$, then for all $\tau \geq 0$,*

$$\mathbb{P}(|M - \mathbb{E}M| \geq \tau) \leq 3 \exp \left(-\frac{t}{C_T B} \log \left(1 + \frac{B\tau}{\sigma^2 + B\mathbb{E}\overline{M}} \right) \right),$$

where $\sigma^2 = \sup_{g \in \mathcal{G}} \sum_{i=1}^n \mathbb{E}g^2(Z_i)$, $\overline{M} = \sup_{g \in \mathcal{G}} |\sum_{i=1}^n g(Z_i)|$ and $C_T > 0$ is a small numerical constant.

By the definition of norm, we have

$$M := \left\| A^*_{S_0 T} u \right\|_2 = \max_{\|g\|_2 \leq 1} \left\langle A^*_{S_0 T} u, g \right\rangle$$

$$= \max_{\|g\|_2 \leq 1} \sum_{i=1}^n (\delta_i - \rho_0) u_i \left\langle v_i, g \right\rangle.$$

Denote $Z_i = (\delta_i - \rho_0) u_i v_i$, we have M is the supremum sum of independent random variable $g(Z_i)$ where $g(Z_i) := (\delta_i - \rho_0) u_i \left\langle v_i, g \right\rangle$. Since $M \geq 0$, $\mathbb{E}M = \mathbb{E}\overline{M}$. The absolute value of $g(Z_i)$ is bounded by

$$|g(Z_i)| \leq \|(\delta_i - \rho_0) u_i v_i\|_2 \|g\|_2 \leq |u_i| \|v_i\|_2 \leq \frac{\mu}{n} \sqrt{k} := B.$$

In addition, from $\mathbb{E}(\delta_i - \rho_0)^2 = \rho_0(1 - \rho_0)$, σ^2 is computed from the argument

$$\sum_{i=1}^n \mathbb{E}g^2(Z_i) = \sum_{i=1}^n \rho_0(1 - \rho_0) u_i^2 \left\langle v_i, g \right\rangle^2$$

$$\leq \rho_0 \max_i u_i^2 g(\sum_{i=1}^n v_i v_i^*) g \leq \rho_0 \frac{\mu}{n} \|g\|_2^2 \left\| \sum_{i=1}^n v_i v_i^* \right\|.$$

Notice that $\sum_{i=1}^n v_i v_i^* = A^*_{T\bullet} A_{\bullet T} = I$ by the orthogonality property of A. Then, $\sigma^2 \leq \max_{\|g\|_2 \leq 1} \rho_0 \frac{\mu}{n} \|g\|_2^2 \leq \rho_0 \frac{\mu}{n}$. Applying Talagrand's inequality yields

$$\mathbb{P}(M \geq \mathbb{E}M + \tau)$$

$$\leq 3 \exp \left(-\frac{\tau}{C_T \sqrt{k} \mu/n} \log \left(1 + \frac{\tau \sqrt{k}}{\rho_0 + k\sqrt{\rho_0}(\mu/n)^{1/2}} \right) \right). \tag{54}$$

We need to consider two cases

1) If $\rho_0 \geq k\sqrt{\rho_0 \frac{\mu}{n}}$, or equivalently, $\rho_0 \geq \frac{\mu k^2}{n}$, we select τ such that $\tau \leq \rho_0/\sqrt{k}$. Thus, the right-hand side of (54) is bounded by

$$3 \exp \left(-\frac{\tau}{C_T \sqrt{k} \mu/n} \log \left(1 + \frac{\tau \sqrt{k}}{2\rho_0} \right) \right),$$

which is in turn smaller than $3 \exp \left(-\frac{\tau^2}{3C_T \rho_0 \mu/n} \right)$ due to the simple observation that $\log(1 + x) \geq 2x/3$ for $0 \leq x \leq 1$. Set $\tau^2 := C\rho_0 \frac{\mu \log n}{n}$ where $C = 15C_T$, the right-hand side of (54) will be less than $3e^{-\log n^5} = 3n^{-1}$. Note that this choice of τ is consistent with the condition $\tau \leq \rho_0/\sqrt{k}$ as long as $\rho_0 \geq C\frac{\mu k \log n}{n}$. We conclude that in this case

$$\mathbb{P}\left(M \geq \sqrt{\rho_0 \frac{\mu k}{n}} + \sqrt{C\rho_0 \frac{\mu \log n}{n}} \right) \leq 3n^{-1}.$$

In other words, with high probability, $M \leq \sqrt{C'\rho_0 \frac{\max\{k, \log n\}}{n}}$.

2) On the other hand, if $\rho_0 \leq \frac{\mu k^2}{n}$, we select τ such that $\tau \leq \sqrt{\rho_0 \frac{\mu k}{n}}$. The right-hand side of (54) is now less than

$$3 \exp \left(-\frac{\tau}{C\sqrt{k}\mu/n} \log \left(1 + \frac{\tau}{2(\rho_0 k\mu/n)^{1/2}} \right) \right)$$

$$\leq 3 \exp \left(-\frac{\tau^2}{3C_T k\rho_0^{1/2}(\mu/n)^{3/2}} \right).$$

Similarly, the right-hand side of (54) will be less than $3n^{-1}$ by setting $\tau^2 := Ck\rho_0^{1/2}(\frac{\mu}{n})^{3/2} \log n$. This choice of τ is consistent with its bound as long as $\rho_0 \geq C\frac{\mu(\log n)^2}{n}$. Therefore,

$$\mathbb{P}\left(M \geq \sqrt{\rho_0 \frac{\mu k}{n}} + \sqrt{Ck \log n} \rho_0^{1/4} (\frac{\mu}{n})^{3/4} \right) \leq 3n^{-1}.$$

In other words, with high probability, $M \leq \sqrt{C'\rho_0 \frac{\mu k}{n}}$ and the proof is completed.

$\square$

REFERENCES

[1] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[2] J. Wright and Y. Ma, "Dense error correction via l1 minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3540–3560, July 2010.

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

[4] E. J. Candès, J.Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 5406–5425, Feb. 2006.

[5] H. Rauhut, "Random sampling of sparse trigonometric polynomials," *Appl. Comput. Harmon. Anal.*, vol. 22, no. 1, pp. 201–224, Jan. 2007.

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

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

[8] ——, "Near-optimal signal recovery from random projections: Universal encoding strategies," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Nov. 2005.

[9] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Comm. Pure Applied Math.*, vol. 61, no. 8, pp. 1025–1045, April 2008.

[10] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the present of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[11] H. Rauhut, "Stability results for random sampling of sparse trigonometric polynomials," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5661–5670, Dec. 2008.

[12] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 1030–1051, Mar. 2006.

[13] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, pp. 1–37, May 2011.

[14] E. J. Candès and J.Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, pp. 969–985, April 2007.

[15] T. T. Do, L. Gan, N. H. Nguyen, and T. D. Tran, "Fast compressive sampling with structurally random matrices," *IEEE Trans. Signal Process.*, Jan. 2012, to appear.

[16] M. Elad, J.-L. Starck, P. Querre, and D. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis," *Applied Comput. Harmon. Anal.*, vol. 19, pp. 340–358, Nov. 2005.

[17] J.-L. Starck, M. Elad, and D. Donoho, "Image decomposition via the combination of sparse representation and a variational approach," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1570–1582, Oct. 2005.

[18] J. Haupt, W. Bajwa, M. Rabbat, and R. Nowak, "Compressed sensing for networked data," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 92–101, Mar. 2008.

[19] Z. M. Charbiwala, S. Chakraborty, S. Zahedi, Y. Kim, T. He, C. Bisdikian, and M. B. Srivastav, "Compressive oversampling for robust data transmission in sensor networks," in *Proc. INFOCOM conf.*, San Diego, CA, USA, Mar. 2010, pp. 1–9.

[20] Z. Li, F. Wu, and J. Wright, "On the systematic measurement matrix for compressed sensing in the presence of gross error," in *Data compress. conf. (DCC)*, Snowbird, UT, USA, Mar. 2010, pp. 356–365.

[21] E. J. Candès and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Trans. Inf. Theory*, 2012, to appear.

[22] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *the Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031 – 1044, June 2010.

[23] J. N. Laska, M. A. Davenport, and R. G. Baraniuk, "Exact signal recovery from sparsely corrupted measurements through the pursuit of justice," in *Asilomar conf. Sig. Sys. Comput.*, Pacific Grove, CA, USA, Nov. 2009, pp. 1556–1560.

[24] C. Studer, P. Kuppinger, G. Pope, and H. Bolcskei, "Sparse signal recovery from sparsely corrupted measurements," in *Proc. Inter. Symp. Inf. Theory (ISIT)*, St. Pertersburg, Russia, Aug. 2011, pp. 1422–1426.

[25] X. Li, "Compressed sensing and matrix completion with constant proportion of corruptions," April 2011, preprint at http://arxiv.org/abs/1104.1041.

[26] M. Rudelson and R. Vershynin, "Sampling from large matrices: An approach through geometric functional analysis," *Journal of the ACM*, vol. 54, no. 4, pp. 1–19, July 2007.

[27] J. A. Tropp, "On the conditioning of random subdictionaries," *Appl. Comput. Harmon. Anal.*, vol. 25, pp. 1–24, 2008.

[28] N. H. Nguyen, T. T. Do, and T. D. Tran, "A fast and efficient algorithm for low-rank approximation of a matrix," in *Proc. 41st ACM Symp. Theory Comput.*, Bethesda, MD, USA, June 2009, pp. 215–224.

[29] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, June 2010.

[30] Z. Zhou, X. Li, J. Wright, E. J. Candès, and Y. Ma, "Stable principal component pursuit," in *Proc. Inter. Symp. Inf. Theory (ISIT)*, Austin, TX, USA, June 2010, pp. 1518–1522.

[31] M. S. Asif and J. Romberg, "Dynamic updating for l1 minimization," *IEEE Sel. Topics Signal Proc.*, vol. 4, no. 2, pp. 421–434, April 2010.

[32] J. Yang, Y. Zhang, and W. Yin, "A fast alternating direction method for tvl1-l2 signal reconstruction from partial fourier data," *IEEE Sel. Topics Signal Proc.*, vol. 4, no. 2, pp. 288–297, April 2010.

[33] N. H. Nguyen, N. M. Nasrabadi, and T. D. Tran, "Robust lasso with missing and grossly corrupted observations," in *Ad. Neural Infor. Proc. Sys. (NIPS)*, Granada, Spain, Dec. 2011.

[34] A. Ganesh, X. Li, J. Wright, E. J. Candès, and Y. Ma, "Dense error correction for low-rank matrices via principal component pursuit," in *Proc. Inter. Symp. Inf. Theory (ISIT)*, Austin, TX, USA, June 2010, pp. 1513–1517.

[35] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal Opt.*, vol. 21, no. 2, pp. 572–596, June 2011.

[36] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," in *Ad. Neural Infor. Proc. Sys. (NIPS)*, Vancouver, BC, Canada, Dec. 2010, pp. 2496–2504.

[37] A. Agarwal, S. Negahban, and M. Wainwright, "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions," in *Proc. 28th Inter. Conf. Mach. Learn. (ICML)*, Bellevue, Washington, USA, June 2011, pp. 1129–1136.