Dense Error Correction via ℓ^1-Minimization

John Wright, *Member,* and Yi Ma, *Senior Member.*

Abstract—This paper studies the problem of recovering a sparse signal $x \in \mathbb{R}^n$ from highly corrupted linear measurements $y = Ax + e \in \mathbb{R}^m$, where e is an unknown error vector whose nonzero entries may be unbounded. Motivated by an observation from face recognition in computer vision, this paper proves that for highly correlated (and possibly overcomplete) dictionaries A, any sufficiently sparse signal x can be recovered by solving an ℓ^1-minimization problem:

$$\min \|x\|_1 + \|e\|_1 \quad \text{subject to} \quad y = Ax + e.$$

More precisely, if the fraction of the support of the error e is bounded away from one and the support of x is a very small fraction of the dimension m, then as m becomes large the above ℓ^1-minimization succeeds for all signals x and almost all sign-and-support patterns of e. This result suggests that accurate recovery of sparse signals is possible and computationally feasible even with nearly 100% of the observations corrupted. The proof relies on a careful characterization of the faces of a convex polytope spanned together by the standard crosspolytope and a set of iid Gaussian vectors with nonzero mean and small variance, dubbed the "cross-and-bouquet" model. Simulations and experiments corroborate the findings, and suggest extensions to the result.

Index Terms—Dense error correction, sparse signal recovery, measure concentration, Gaussian matrices, ℓ^1-minimization, polytope neighborliness.

I. INTRODUCTION

Recovery of high-dimensional sparse signals or errors has been one of the fastest growing research areas in signal processing in the past few years. At least two factors have contributed to this explosive growth. On the theoretical side, the progress has been propelled by powerful tools and results from multiple mathematical areas such as measure concentration [1]–[3], statistics [4]–[6], combinatorics [7], and coding theory [8]. On the practical side, a lot of excitement has been generated by remarkable successes in real-world applications in areas such as signal (image or speech) processing [9], communications [10], computer vision and pattern recognition [11]–[13] and more.

A. A motivating example

One notable, and somewhat surprising, successful application of sparse representation is automatic face recognition. The basic setting of face recognition is as follows. For each person, a set of training images are taken under different illuminations. We can view each image as a vector by stacking its columns and put all the training images as column vectors of a matrix, say $A \in \mathbb{R}^{m \times n}$. Then, m is the number of pixels in an image and n is the total number of images for all the subjects of interest. Given a new query image, again we can stack it as a vector $y \in \mathbb{R}^m$. To identify which subject the image belongs to, we can try to represent y as a linear combination of all the images, i.e., $y = Ax$ for some $x \in \mathbb{R}^n$. Since in practice n can potentially be larger than m, the equations can be underdetermined and the solution x may not be unique. In this context, it is natural to seek the sparsest solution for x whose large non-zero coefficients then provide information about the subject's true identity. This can be done by solving the typical ℓ^1-minimization problem:

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

The problem becomes more interesting if the query image y is severely occluded or corrupted, as shown in Figure 1 left, column (a). In this case, one needs to solve a corrupted set of linear equations $y = Ax + e$, where $e \in \mathbb{R}^m$ is an unknown vector whose nonzero entries correspond to the corrupted pixels. For sparse errors e and tall matrices A ($m > n$), Candes and Tao [14] proposed to multiply the equation $y = Ax + e$ with a matrix B such that $BA = 0$, and then use ℓ^1-minimization to recover the error vector e from the new linear equation $By = Be$.

As we mentioned earlier, in face recognition (and many other applications), n can be larger than m and the matrix A can be full rank. One cannot directly apply the above technique even if the error e is known to be very sparse. To resolve this difficulty, in [11], the authors proposed to instead seek $[x, e]$ together as the sparsest solution to the extended equation $y = [A \; \mathtt{I}]w$ with $w = \begin{bmatrix} x \\ e \end{bmatrix} \in \mathbb{R}^{m+n}$, by solving the extended ℓ^1-minimization problem:

$$\min_{x, e} \|x\|_1 + \|e\|_1 \quad \text{subject to} \quad y = Ax + e. \qquad (2)$$

This seemingly minor modification to the previous error correction approach has drastic consequences for the performance of robust face recognition. Solving the modified ℓ^1-minimization enables almost perfect recognition even with more than 60% of the pixels of the query image arbitrarily corrupted (see Figure 1 for an example), far beyond the amount of error that can theoretically be corrected by the previous error correction method [14].

Although ℓ^1-minimization is expected to recover sufficiently sparse solutions with overwhelming probability for general systems of linear equations (see [16]), it is rather surprising that it works for the equation $y = Ax + e$ at all. In the
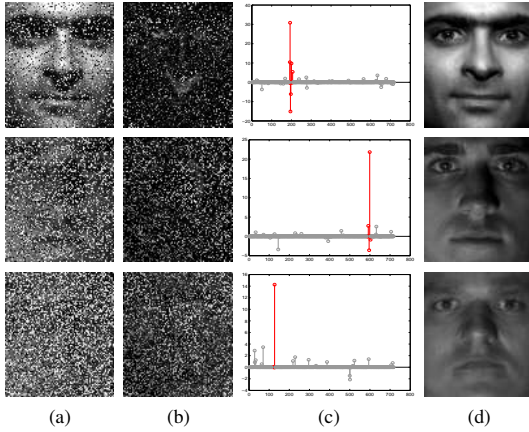
Fig. 1. **Face recognition under random corruption.** Left: (a) Test images $\boldsymbol{y}$ with random corruption from the database presented in [15]. Top row: 30% of pixels are corrupted, Middle row: 50% corrupted, Bottom row: 70% corrupted. (b) Estimated errors $\hat{\boldsymbol{e}}$. (c) Estimated sparse coefficients $\hat{\boldsymbol{x}}$. (d) Reconstructed images $\boldsymbol{y}_r = A\hat{\boldsymbol{x}}$. The extended ℓ^1-minimization (2) correctly recovers and identifies all three corrupted face images. Right: The recognition rate across the entire range of corruption for all the 38 subjects in the database. It performs almost perfectly upto 60% random corruption.

application described above, the columns of A are highly correlated. As m becomes large (i.e., the resolution of the image becomes high), the convex hull spanned by all face images of all subjects is only an extremely tiny portion of the unit sphere $\mathbb{S}^{m-1}$. For example, the images in Figure 1 lie on $\mathbb{S}^{8,063}$. The smallest inner product with their normalized mean is 0.723; they are contained within a spherical cap of volume $\leq 1.47 \times 10^{-229}$. These vectors are tightly bundled together as a "bouquet," whereas the vectors associated with the identity matrix and its negative $\pm \mathtt{I}$ together[1] form a standard "cross" in $\mathbb{R}^m$, as illustrated in Figure 2. Notice that such a "cross-and-bouquet" matrix $[A \ \mathtt{I}]$ is neither incoherent nor (restrictedly) isometric, at least not uniformly. Also, the density of the desired solution $\boldsymbol{w}$ is not uniform either. The $\boldsymbol{x}$ part of $\boldsymbol{w}$ is usually a very sparse[2] vector, but the $\boldsymbol{e}$ part can be very dense. Existing results for recovering sparse signals suggest that ℓ^1-minimization may have difficulty in dealing with such signals, contrary to its empirical success in face recognition.

We have experimented with similar cross-and-bouquet type models where the matrix A is a random matrix with highly correlated column vectors. The simulation results in Section IV indicate that what we have seen in face recognition is not an isolated phenomenon. In fact, the simulations reveal something even more striking and puzzling: As the dimension m increases (and the sample size n grows in proportion), the percentage of errors that the ℓ^1-minimization (2) can correct seems to approach 100%. At first sight, this may seem rather surprising, but this paper explains why this should be expected when the space dimension is high.

B. Assumptions and the main result

Motivated by the face recognition example above, this paper aims to resolve this apparent discrepancy between theory and

---

[1] Here, we allow the entries of the error $\boldsymbol{e}$ to assume either positive or negative signs.

[2] Ideally, $\boldsymbol{x}$ is concentrated only on images of the same subject. In the standard face recognition scenario where the number of subjects in the database grows, but the number of images per subject is fixed, the desired sparsity of $\boldsymbol{x}$ is actually a constant.
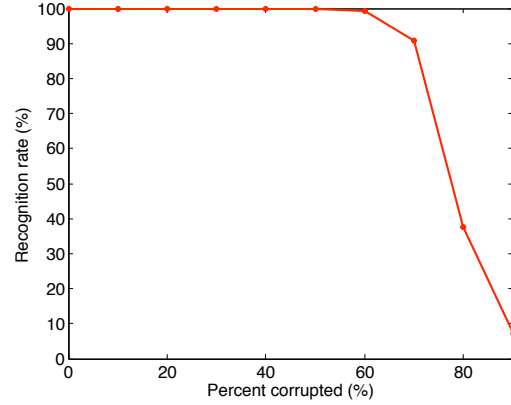

Fig. 2. **The "cross-and-bouquet" model.** Left: the bouquet A and the crosspolytope spanned by the matrix $\pm \mathtt{I}$. Right: the tip of the bouquet magnified; it is a collection of i.i.d. Gaussian vectors with small variance σ^2 and common mean vector $\boldsymbol{\mu}$. The cross-and-bouquet polytope is spanned by vertices from both the bouquet A and the cross $\pm \mathtt{I}$.

practice of ℓ^1-minimization. To this end, we need to give a more careful characterization of its behavior in recovering a sparse signal $\boldsymbol{x}_0 \in \mathbb{R}^n$ from highly corrupted linear measurements $\boldsymbol{y} \in \mathbb{R}^m$:

$$\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{e}_0,$$

where $\boldsymbol{e}_0 \in \mathbb{R}^m$ is a vector of errors of arbitrarily large magnitude and support. We are interested in the situation when $[A, \mathtt{I}]$ forms a cross-and-bouquet (CAB) model, for which we give a precise definition below.

The model for $A \in \mathbb{R}^{m \times n}$ should capture the idea that it consists of small deviations about a mean, hence a "bouquet:"

Assumption 1 (Bouquet Model): We consider the case where the columns of A are i.i.d. samples from a Gaussian distribution: $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] \in \mathbb{R}^{m \times n}$, with

$$\boldsymbol{a}_i \sim_{iid} \mathcal{N}\left(\boldsymbol{\mu}, \frac{\nu^2}{m}\mathtt{I}_m\right), \ \|\boldsymbol{\mu}\|_2 = 1, \ \|\boldsymbol{\mu}\|_\infty \leq C_\mu m^{-1/2}, \ (3)$$

for some constant $C_\mu \geq 1$.

Together, the two conditions on the variance and mean force the bouquet to remain both tight and incoherent with the standard basis (or "cross") as the dimension m increases.

We assume little prior knowledge in the signal $\boldsymbol{x}_0$ and error $\boldsymbol{e}_0$ to be recovered. For ease of analysis, we assume:

Assumption 2 (Signs of $\boldsymbol{x}_0$ and $\boldsymbol{e}_0$): The signs $\boldsymbol{\sigma}$ of $\boldsymbol{e}_0$ are chosen at random while the signs $\boldsymbol{\sigma}_x$ of $\boldsymbol{x}_0$ can be arbitrary.

Geometrically, for $\boldsymbol{x}_0$ and $\boldsymbol{e}_0$ to be recoverable from the ℓ^1 minimization (2), it would require vertices from the "bouquet" A to "see" through the crosspolytope $\pm I$ to vertices of the "negative bouquet" $-A$ that are nearly antipodal to them. In a low-dimensional space, it is difficult to imagine this is even possible at all. However, our analysis will reveal that in high-dimensional spaces, as long as $\boldsymbol{x}_0$ is sparse enough, this seemingly-impossible event occurs with probability approaching one.

We study the behavior of the solution to the ℓ^1-minimization (2) in the following growth framework:

Assumption 3 (Proportional Growth): We say a sequence of signal-error problems exhibits proportional growth with parameters $\delta > 0$, $\rho \in (0,1)$, $\alpha > 0$ if

$$n = \lfloor \delta m \rfloor, \quad \|\boldsymbol{e}_0\|_0 = \lfloor \rho m \rfloor, \quad \|\boldsymbol{x}_0\|_0 = \lfloor \alpha m \rfloor. \quad (4)$$

In the above assumption, while the support size of the error e can be any fraction $\rho < 1$ of the dimension m, the support size of the sparse signal x is only a very small fraction of m: $\|\boldsymbol{x}\|_0 = \alpha m$ for some small constant $\alpha > 0$ (which may depend on δ and ρ). This condition differs from the typical assumption in the sparse representation literature in the sense that the signals of interest x and e have different support density. Through the proof of the main result, one will see why the support density α of the signal x can only be very small while we allow the support density ρ of the error e to be arbitrarily close to 1. In fact, α has to be a small fraction of $1 - \rho$. However small α is, such a linear growth of sparsity for x is more than adequate for signals arising in many practical problems where the support size x can often be bounded or grows sublinearly in the dimension. For example in face recognition, the support size of x is bounded by a constant – the number of images per subject.

Under Assumptions 1, 2, and 3, we have the following result:

Theorem 1 (Dense Error Correction with the CAB Model): Fix any $\delta > 0$, $\rho < 1$. Suppose that A is distributed according to (3) with ν sufficiently small, that $J \subset [m]$ is a uniform random subset of size ρm, and that $\boldsymbol{\sigma} \in \mathbb{R}^m$ with $\boldsymbol{\sigma}_J$ iid uniform ± 1 (independent of J) and $\boldsymbol{\sigma}_{J^c} = 0$, and that m is sufficiently large. Then with probability at least $1 - C\exp(-\varepsilon^\star m)$ in $A, J, \boldsymbol{\sigma}$, for all $\boldsymbol{x}_0$ with $\|\boldsymbol{x}_0\|_0 \leq \alpha^\star m$ and any $\boldsymbol{e}_0$ with signs and support $(J, \boldsymbol{\sigma})$,

$$(\boldsymbol{x}_0, \boldsymbol{e}_0) = \arg\min \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{s.t.} \quad A\boldsymbol{x} + \boldsymbol{e} = A\boldsymbol{x}_0 + \boldsymbol{e}_0,$$

and the minimizer is uniquely defined.

Here, C is numerical, $\alpha^\star$ and $\varepsilon^\star$ are postive constants (wrt m), which depend on δ, ρ, ν. By "ν sufficiently small" and "m sufficiently large" we mean that $0 < \nu < \nu^\star$ and $m > m^\star$, where $\nu^\star(\delta, \rho) > 0$ and $m^\star(\delta, \rho, \nu) > 0$ are constant wrt m.

Below, we comment further on the nature of the constants in Theorem 1. We find it convenient to introduce one special function, which is defined as follows. Let $f : (0, 1/e) \rightarrow (0, 1/e)$ be defined by

$$f(x) = x \log(1/x).$$

It is easy to verify that f is strictly increasing on $(0, 1/e)$, and hence has a well-defined inverse. We let

$$\Upsilon : (0, 1/e) \rightarrow (0, 1/e), \quad \Upsilon[t] = f^{-1}(t)$$

denote this inverse.[3] We note that because f is strictly increasing, Υ is also strictly increasing. Moreover, for all $x \in (0, 1/e)$, $f(x) \geq x$, and so

$$\Upsilon[t] \leq t \quad \forall\, t \in (0, 1/e). \quad (5)$$

Finally, notice that for any fixed a and $t < a/e$,

$$x \log(a/x) \leq t \iff x \leq a\Upsilon(t/a), \quad (6)$$

a property that will be useful in taking union bounds.

Remark 2: One can give explicit expressions for the constants in Theorem 1. We first introduce

$$\tau = \exp\left(-48^{-2}\delta^{-2}\nu^{-2}\right), \quad (7)$$
$$C_s = \Upsilon\left[2^{-17}C_\mu^{-4}\delta^{-2}(1-\rho)\right]. \quad (8)$$

One can then choose $C = 31$, and set

$$\begin{cases} \nu_{\delta,\rho}^\star &= 2^{-12}\sqrt{C_s/\delta}, \\ \varepsilon_{\delta,\rho,\nu}^\star &= \nu^2(1-\rho)\tau^2/2, \\ \alpha_{\delta,\rho,\nu}^\star &= \nu^2\,\Upsilon[\varepsilon^\star/\delta], \\ m_{\delta,\rho,\nu}^\star &= 2^{17}/(1-\rho)\,C_s. \end{cases} \quad (9)$$

Above, we have added subscripts to emphasize the dependence of these quantites on the properties of the bouquet (δ, ν) and the density of the error, ρ. In calculations below, we occasionally drop these constants for the sake of compactness; e.g., $\alpha^\star$ will refer to $\alpha_{\delta,\rho,\nu}^\star$.

One immediately notices that the given constants are extremely small (or, in the case of $m^\star$, extremely large). We did not attempt to obtain the best possible constants. Indeed, it is reasonable to believe that the best constants that can be obtained via our analysis will be quite pessimistic. Our analysis uses linear algebraic manipulations to reduce the quantity of interest to a function of sparse singular values and then applies standard concentration results to argue that they are well-behaved for all sparse vectors. This is reminiscent of (and inspired by) the RIP analysis of the standard compressed sensing setup. That analysis obtains, in a very clean and simple manner, guarantees for sparse recovery that are order-optimal. However, in a finite-m setting RIP results are pessimistic compared to those obtained with tools from geometric functional analysis [18] or combinatorics [19].

It should also be noted that if the support of $\boldsymbol{x}_0$ is sublinear (i.e., $\|\boldsymbol{x}_0\|_0 = o(m)$), exact recovery holds for sufficiently large m. In other words, as long as the bouquet is sufficiently

---

[3]One can futher write $\Upsilon[t] = \exp(\mathfrak{W}_{-1}(-t))$, where $\mathfrak{W}_{-1}$ is the negative branch of the Lambert W function [17].

tight, asymptotically ℓ^1-minimization recovers any sufficiently sparse signal from almost any error with support fraction bounded away from 100%. Notice that we makes no assumption on the signal $\boldsymbol{x}$, except that its support being extremely small. Existing results in the sparse representation literature typically either pertain to all sparse signals (i.e., the strong threshold of [20]) or almost all sparse signals, in terms of signs and support (i.e., the weak threshold of [20]). Our result has a hybrid flavor: it holds for all signals $\boldsymbol{x}$, but only for almost all errors $\boldsymbol{e}$, in terms of signs and support.

C. Relations to previous results

a) Restricted isometry and incoherence of the cross-and-bouquet model: As mentioned earlier, typical results in the literature for sparse signal recovery do not apply to equations of the type $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$. The cross-and-bouquet matrix $[A \ \mathtt{I}]$ is neither restrictedly isometric nor incoherent. As a result, greedy algorithms such as Orthogonal Matching Pursuit [21], [22] succeed only when the error $\boldsymbol{e}$ is very sparse (see Section IV) for the simulation results and comparison with our method). However, this does not mean that the restricted isometry property is irrelevant to the new problem. On the contrary, similar ideas play an important role in our proof. Moreover, unlike the typical compressed sensing setting, the solution $[\boldsymbol{x}, \boldsymbol{e}]$ sought has very uneven density (or sparsity). This is reminiscent of the block sparsity studied in [23]. However, as we will see, the special block structure of the cross-and-bouquet model enables sparse recovery far beyond the breakdown point for general sparse (or block sparse) signals.

b) Error correction: From an error correction viewpoint, the above result seems surprising: One can correctly solve a set of linear equations with *almost all* the equations randomly and arbitrarily corrupted. This is perhaps surprising considering that the best error-correcting codes (in the binary domain $\mathbb{Z}_2$), constructed based on expander graphs, normally correct a fixed fraction of errors [24]–[26]. The exact counterpart of our result in the binary domain is not clear.[4] While there are superficial similarities between our result and [25], [27] in the use of linear programming for decoding and analysis via polytope geometry, those works do not consider real valued signals. In particular, the negative result of [27] for specific families of binary codes admitting linear programming decoders does not apply here.

We can, however, draw the following comparisons with existing error correction methods in the domain of real numbers:

- When $n < m$, the range of A is a subspace in $\mathbb{R}^m$. In such an overdetermined case, one could directly apply the method of Candes and Tao [14] mentioned earlier. However, the error vector $\boldsymbol{e}$ needs to be sparse for that approach whereas our result suggests even dense errors (with support far beyond 50%) can be corrected by instead solving the extended ℓ^1-minimization (2). Thus, even in the overdetermined case, solving (2) has advantages for coherent matrices A, *if* the signal $\boldsymbol{x}$ to be

recovered is known to be sparse. This will be verified by simulations in Section IV.
- The small linear growth of the support of $\boldsymbol{x}$ in m is the best one can hope for in the regime of dense errors. In general, we need at least $\|\boldsymbol{x}\|_0$ uncorrupted linear measurements to recover $\boldsymbol{x}$ uniquely. If a fraction $\rho < 1$ of the m equations is totally corrupted by $\boldsymbol{e}$, the support of $\boldsymbol{x}$ has to be smaller than the number of remaining $(1 - \rho)m$ uncorrupted measurements. On the other hand, the sparser the error $\boldsymbol{e}$ is, the larger support $\boldsymbol{x}$ can be. Simulation results in Section IV also confirm this phenomenon. However, in this paper, we are mainly interested in how ℓ^1-minimization behaves with dense errors, for ρ very close to 1.
- When $n \geq m$, the matrix A is full rank and the method of Candes and Tao [14] no longer applies.[5] Our result suggests that as long as A is highly correlated, the extended ℓ^1-minimization (2) can still recover the sparse signal $\boldsymbol{x}$ exactly even if almost all the equations might be corrupted. This is verified by the simulation results in Section IV.

c) Polytope geometry: The success of ℓ^1-minimization in recovering sparse solutions $\boldsymbol{x}$ from underdetermined systems of linear equations $\boldsymbol{y} = A\boldsymbol{x}$ can be viewed as a consequence of a surprising property of high-dimensional polytopes. If the column vectors of A are random samples from a zero-mean Gaussian $\mathcal{N}(0, \mathtt{I})$, and m and n are allowed to grow proportionally, then with overwhelming probability the convex polytope spanned by the columns of A is highly centrally symmetric neighborly [20], [28]. Neighborliness provides the necessary and sufficient condition for uniform sparse recovery: the ℓ^1-minimization (1) correctly recovers $\boldsymbol{x}$ if and only if the columns associated with the nonzero entries of $\boldsymbol{x}$ span a face of the polytope $\mathrm{conv}(A)$.

In our case, the columns of the matrix A are i.i.d. Gaussian vectors with nonzero mean $\boldsymbol{\mu}$ and small variance σ^2, whereas the vectors of the identity matrix $\mathtt{I}$ are completely fixed. To characterize when the extended ℓ^1-minimization (2) is able to recover the solution $[\boldsymbol{x}, \boldsymbol{e}]$ correctly, we need to examine the geometry of the peculiar centrally symmetric convex polytope $\mathrm{conv}(\pm A, \pm \mathtt{I})$ spanned together by columns of A and $\mathtt{I}$. In a sense, our result implies that the "cross-and-bouquet" polytope is neighborly[6]. Or in other words, the vertices associated with the nonzero entries of $\boldsymbol{x}$ and $\boldsymbol{e}$ form a face of the polytope with probability approaching one as the dimension m becomes large. Precisely due to this kind of neighborliness of the cross-and-bouquet polytope, the extended ℓ^1-minimization (2) is able to correctly recover the desired solution, even though the

---

[4]It is possible that under an analogous growth model (see Section I-B), the LP decoder of [25] could also correct large fractions of binary errors.

[5]One could choose to premultiply the equation $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$ with an "approximate orthogonal complement" of A, say the orthogonal complement of the mean vector $\boldsymbol{\mu}$, which is an $(m - 1) \times m$ matrix B. This gives $B\boldsymbol{y} = B\boldsymbol{e} + \boldsymbol{z}$ where $\boldsymbol{z} = BA\boldsymbol{x}$. If the norm of $\boldsymbol{x}$ is bounded, then $\boldsymbol{z}$ is a signal with small magnitude due to the near-orthogonality of B and A. In this case, one can view $\boldsymbol{z}$ as a noise term and try to recover $\boldsymbol{e}$ as a sparse signal via ℓ^1-minimization. However, the recovery will not be exact, and it is then unclear how to subsequently recover the desired sparse signal $\boldsymbol{x}$.

[6]Although, the precise notion of neighborliness is not the same - our result can be viewed as a hybrid of the "weak" neighborliness of [28] (since the signs and support of $\boldsymbol{e}$ are random) and "strong" (or central) neighborliness [28] (since the signs and support of $\boldsymbol{x}$ are arbitrary).

part of the solution corresponding to e might be dense.

D. Implications for applications

a) Robust reconstruction, classification, and source separation: The new result about the cross-and-bouquet model has strong implications for robust reconstruction, classification, and separation of highly correlated classes of signals such as faces or voices, despite severe corruption. It helps explain the surprising performance of robust face recognition that we discussed earlier. It further suggests that if the resolution of the image increases in proportion with the size of the database (say, due to the increasing number of subjects), the ℓ^1-minimization would tolerate an even higher level of corruption, far beyond the 60% at the 80×60 resolution experimented with in [11]. Other applications where this kind of model could be useful and effective include speech recognition/imputation, audio source separation, video segmentation, or activity recognition from motion sensors.

b) Communication through an almost random channel: The result suggests that we can use the cross-and-bouquet model to accurately send information through a highly corrupting channel. Hypothetically, we can imagine a channel through which we can send one real number at a time, say as one packet of binary bits, and each packet has a high probability $\rho < 1$ of being totally corrupted. One can use the sparse vector x (or its support) to represent useful information, and use a set of highly correlated high-dimensional vectors as the encoding transformation A. The high correlation in A ensures that there is sufficient redundancy built in the encoded message Ax so that the information about x will not be lost even if most entries of Ax can be corrupted while being sent through such a channel (although the coding "rate" for x is pessimistically small with the current estimates, roughly characterized by α which is a small fraction of $1 - \rho$). Our result suggests that the decoding can be done correctly and efficiently using linear programming.

c) Encryption and information hiding: One can potentially use the cross-and-bouquet model for encryption. For instance, if both the sender and receiver share the same encoding matrix A (say a randomly chosen Gaussian matrix), the sender can deliberately corrupt the message Ax with arbitrary random errors e before sending it to the receiver. The receiver can use linear programming to decode the information x, whereas any eavesdropper will not be able to make much sense out of the highly corrupted message $y = Ax + e$. The higher ρ is, the more "secured" the encrypted message would be (in exchange for a lower coding rate for x). Of course, the long-term security of such an encryption scheme relies on the difficulty of learning the encoding matrix A after gathering many instances of corrupted message. It is not even clear whether it is easy to learn A from instances of uncorrupted message $y = Ax$ – existing algorithms do not guarantee global optimality [9]. Presumably, in the presence of corruption the problem becomes even more daunting.

II. ROADMAP OF PROOF

In this section, we will lay out the roadmap of the proof for our main result. After introducing some basic notation,

we will prove two key lemmas. In subsection II-B, we first introduce a certificate q whose existence guarantees the uniqueness and optimality of (x_0, e_0). In subsection II-C, the second lemma gives a set of sufficient conditions under which the certificate of optimality can be constructed via a simple iterative procedure. We postpone to Section III the detailed probabilistic analysis that proves that these conditions hold with overwhelming probability under the assumptions of the theorem. The proof will borrow many standard measure concentration results that are summarized in the appendix for completeness. We put such technical details in a separate section and the appendix so that the reader might want to skip them in a first reading, and directly jump to Section IV for empirical verification and experiments.

A. Notation

Before stating our main result, we fix some additional notation. For any $n \in \mathbb{Z}_+$, $[n]$ denotes the set $\{1, \ldots, n\}$. Let $I = \operatorname{supp}(x_0) \subset [n]$, $J = \operatorname{supp}(e_0) \subset [m]$. Let $k_1 = |I|$ be the support size of the signal x_0 and $k_2 = |J|$ the support size of the error e_0, and let $\sigma = \operatorname{sign}(e_0(J)) \in \{\pm 1\}^{k_2}$. For $M \in \mathbb{R}^{m \times n}$, M_i will denote the i-th row of M; for $I \subseteq [m]$, M_I will denote the submatrix of M containing the rows indexed by I. To reduce confusion between the index set I and the identity matrix, we use $\mathbb{I}$ to denote the latter.

We will use several norms on vectors and matrices. For $x \in \mathbb{R}^m$, $\|x\|_1 = \sum_i |x_i|$ is the ℓ^1-norm, $\|x\|$ is the ℓ^2 norm. For matrices M, $\|M\| = \sigma_{max}(M)$ denotes the ℓ^2 operator norm, $\|M\|_F = \sqrt{\operatorname{tr}[M^*M]}$ denotes the Frobenius norm, $\|M\|_{2,\infty} = \max_i \|M_i\|$ denotes the $\ell^2 \to \ell^\infty$ operator norm. For an $m \times n$ matrix M with ordered singular values $\sigma_1 \geq \sigma_2 \geq \ldots$, we will use $\sigma_{min}(M)$ to denote $\sigma_{\min\{m,n\}}(M)$, i.e., the smallest singular value. For a linear subspace $V \subseteq \mathbb{R}^n$, $P_V \in \mathbb{R}^{n \times n}$ will denote the projection operator onto V.

We will make extensive use of the shrinkage operator $\mathcal{S}$ with threshold $1/2$, defined for scalars t as

$$\mathcal{S}[t] = \operatorname{sign}(t) \max\{t - 1/2, 0\}, \qquad (10)$$

and extended to vectors by applying it elementwise.

Below, wherever the symbol C occurs with no subscript, it should be read as "some constant" (with respect to dimension m). When used in different sections, it need not refer to the same constant.

B. Optimality conditions

Lemma 1: Fix $x_0 \in \mathbb{R}^n$ and $e_0 \in \mathbb{R}^m$. Let $I = \operatorname{supp}(x_0)$, $\sigma_x = \operatorname{sign}(x_0(I))$, $J = \operatorname{supp}(e_0)$, $\sigma = \operatorname{sign}(e_0(J))$. Suppose that $\mu_{J^c} \neq 0$, that the following system of equations has a solution $q \in \mathbb{R}^{m-k_2+n-k_1}$ with $\|q\|_\infty < 1$:

$$\begin{cases} \begin{bmatrix} \nu(Z_{J^c, I^c})^* & \mathbb{I} \\ \nu(Z_{J^c, I})^* & 0 \end{bmatrix} q = \begin{bmatrix} -\nu(Z_{J, I^c})^* \sigma \\ \sigma_x - \nu(Z_{J, I})^* \sigma \end{bmatrix}, \\ \langle \mu_{J^c}, q_{1 \ldots m-k_2} \rangle = -\langle \mu_J, \sigma \rangle, \end{cases} \qquad (11)$$

and that

$$A_{J^c, I} \text{ has full rank } k_1. \qquad (12)$$

Then $(\boldsymbol{x}_0, \boldsymbol{e}_0)$ is the unique optimal solution to

$$\min \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subject to} \quad A\boldsymbol{x} + \boldsymbol{e} = A\boldsymbol{x}_0 + \boldsymbol{e}_0. \quad (13)$$

Proof: Duality (e.g., Theorem 4 of [29]) shows that if there exists $\boldsymbol{h} \in \mathbb{R}^m$ satisfying

$$\begin{cases} (A^*)_I\,\boldsymbol{h} = \boldsymbol{\sigma}_x, \ \ \boldsymbol{h}_J = \boldsymbol{\sigma}, \\ \|(A^*)_{I^c}\,\boldsymbol{h}\|_\infty < 1, \ \ \|\boldsymbol{h}_{J^c}\|_\infty < 1, \end{cases} \quad (14)$$

and

$$B \doteq \begin{bmatrix} A_{[m],I} & (\mathtt{I})_{[m],J} \end{bmatrix} \in \mathbb{R}^{m \times (k_1+k_2)} \text{ has full rank } k_1 + k_2, \quad (15)$$

then $(\boldsymbol{x}_0, \boldsymbol{e}_0)$ is the unique optimal solution to (13).

Notice that $\operatorname{rank}(B) = k_2 + \operatorname{rank}(A_{J^c,I})$, and so (12) implies (15). Consider the system

$$\begin{bmatrix} (A_{J^c,I^c})^* & \mathtt{I} \\ (A_{J^c,I})^* & 0 \end{bmatrix} \boldsymbol{q} = \begin{bmatrix} -(A_{J,I^c})^*\boldsymbol{\sigma} \\ \boldsymbol{\sigma}_x - (A_{J,I})^*\boldsymbol{\sigma} \end{bmatrix}, \quad \|\boldsymbol{q}\|_\infty < 1, \quad (16)$$

in unknown $\boldsymbol{q} \in \mathbb{R}^{m+n-k_1-k_2}$. If there exists $\boldsymbol{q}$ satisfying (16), then setting $\boldsymbol{h}_J = \boldsymbol{\sigma}$ and $\boldsymbol{h}_{J^c} = \boldsymbol{q}_{1\ldots m-k_2}$ gives a dual vector $\boldsymbol{h}$ satisfying (14). Since any $\boldsymbol{q}$ satisfying (11) also satisfies (16), this completes the proof. ∎

Thus, we merely need to construct a vector $\boldsymbol{q}$ of small ℓ^∞ norm satisfying (11). It will be convenient to rescale (11) and introduce notation for its various components. To this end, let $U \doteq (Z_{J^c,I^c})^* \in \mathbb{R}^{(n-k_1)\times(m-k_2)}$,

$$W \doteq \begin{bmatrix} (1-\rho)^{-1/2}(Z_{J^c,I})^* \\ \|\boldsymbol{\mu}_{J^c}\|^{-1}\boldsymbol{\mu}_{J^c}^* \end{bmatrix}, \quad R \doteq \begin{bmatrix} -(Z_{J,I^c})^* \\ 0 \\ 0 \end{bmatrix},$$

$$T \doteq \begin{bmatrix} 0 \\ -(1-\rho)^{-1/2}(Z_{J,I})^* \\ -\|\boldsymbol{\mu}_{J^c}\|^{-1}\boldsymbol{\mu}_J^* \end{bmatrix}, \quad \boldsymbol{z} \doteq \begin{bmatrix} 0 \\ (1-\rho)^{-1/2}\nu^{-1}\boldsymbol{\sigma}_x \\ 0 \end{bmatrix},$$

and further set

$$\Phi \doteq \begin{bmatrix} \nu U & \mathtt{I} \\ W & 0 \end{bmatrix}. \quad (17)$$

Then $\boldsymbol{q}$ is a solution to (11) if and only if $\boldsymbol{q}$ satisfies

$$\Phi\boldsymbol{q} = \nu R\boldsymbol{\sigma} + T\boldsymbol{\sigma} + \boldsymbol{z}. \quad (18)$$

C. Iterative construction of separator

The following lemma allows us to relax our requirements on $\boldsymbol{q}$: rather than directly producing a solution with $\|\boldsymbol{q}\|_\infty < 1$, we can instead produce a solution of small ℓ^2 norm that does not too severely violate the ℓ^∞ constraint, in the sense that $\|\mathcal{S}[\boldsymbol{q}]\|_2$ is small. For simplicity we often write $\mathcal{S}[\boldsymbol{q}]$ as $\mathcal{S}\boldsymbol{q}$.

Lemma 2: Consider an underdetermined system of equations $\Phi\boldsymbol{q} = \boldsymbol{w}$. Let P_Φ denote the projection operator onto $\operatorname{range}(\Phi^*)$, and let $\xi_s(\Phi)$ denote the norm P_Φ, restricted to sparse vectors:

$$\xi_s(\Phi) \doteq \sup_{\substack{\|\boldsymbol{x}\|_0 \le s, \\ \|\boldsymbol{x}\|_2 \le 1}} \|P_\Phi\boldsymbol{x}\|_2. \quad (19)$$

Suppose $\xi_s(\Phi) < 1$ and there exists a solution $\boldsymbol{q}_0$ with $\Phi\boldsymbol{q}_0 = \boldsymbol{w}$ satisfying

$$\|\boldsymbol{q}_0\| + \frac{\xi_s(\Phi)}{1 - \xi_s(\Phi)}\|\mathcal{S}\boldsymbol{q}_0\| \le \sqrt{s}/2. \quad (20)$$

Then there exists a solution $\boldsymbol{q}^\star$ satisfying $\|\boldsymbol{q}^\star\|_\infty \le 1/2$.

Proof: We construct a convergent sequence of vectors $\boldsymbol{q}_0, \boldsymbol{q}_1, \ldots$ whose limit $\boldsymbol{q}^\star$ satisfies $\Phi\boldsymbol{q}^\star = \boldsymbol{w}$, $\|\boldsymbol{q}^\star\|_\infty \le 1/2$. Set

$$\boldsymbol{q}_i = \boldsymbol{q}_{i-1} - (\mathtt{I} - P_\Phi)\mathcal{S}[\boldsymbol{q}_{i-1}], \quad (21)$$

and notice that for all i, $\Phi\boldsymbol{q}_i = \Phi\boldsymbol{q}_0 = \boldsymbol{w}$.

We will show by induction on i that the sequence $(\boldsymbol{q}_i)$ obeys the following estimates

(i)	$\|\boldsymbol{q}_i\|$	$\le$	$\|\boldsymbol{q}_0\| + \|\mathcal{S}\boldsymbol{q}_0\| \sum_{j=1}^{i} \xi_s(\Phi)^j,$
(ii)	$\|\mathcal{S}\boldsymbol{q}_i\|$	$\le$	$\xi_s(\Phi)^i\|\mathcal{S}\boldsymbol{q}_0\|,$
(iii)	$\|\mathcal{S}\boldsymbol{q}_i\|_0$	$\le$	$s.$

For the base case $i = 0$, (i),(ii) hold trivially. For (iii), notice that since each nonzero element of $\mathcal{S}[\boldsymbol{q}_0]$ corresponds to an element of $\boldsymbol{q}_0$ of size at least $1/2$, $\|\mathcal{S}[\boldsymbol{q}_0]\|_0 \times 1/4 \le \|\boldsymbol{q}_0\|^2$. By (20), $\|\boldsymbol{q}_0\|^2 \le s/4$, establishing (iii).

Now, suppose (i)-(iii) hold for $0, \ldots, i-1$. For any scalar t, $|t - \mathcal{S}(t)| \le |t|$, $\|\boldsymbol{q}_{i-1} - \mathcal{S}[\boldsymbol{q}_{i-1}]\| \le \|\boldsymbol{q}_{i-1}\|$. Hence, using that $\mathcal{S}[\boldsymbol{q}_{i-1}]$ is s-sparse,

$$\begin{aligned} \|\boldsymbol{q}_i\| &= \|\boldsymbol{q}_{i-1} - \mathcal{S}[\boldsymbol{q}_{i-1}] + P_\Phi\mathcal{S}[\boldsymbol{q}_{i-1}]\|, \\ &\le \|\boldsymbol{q}_{i-1} - \mathcal{S}[\boldsymbol{q}_{i-1}]\| + \|P_\Phi\mathcal{S}[\boldsymbol{q}_{i-1}]\|, \\ &\le \|\boldsymbol{q}_{i-1}\| + \xi_s(\Phi)\|\mathcal{S}[\boldsymbol{q}_{i-1}]\|, \\ &\le \|\boldsymbol{q}_0\| + \|\mathcal{S}[\boldsymbol{q}_0]\| \sum_{j=1}^{i} \xi_s(\Phi)^j, \end{aligned}$$

and so (i) holds for i. Bounding the summation by $\sum_{j=1}^{\infty} \xi_s(\Phi)^j = \frac{\xi_s}{1-\xi_s}$ and applying (20), $\|\boldsymbol{q}_i\|^2 \le s/4$, and hence $\|\mathcal{S}[\boldsymbol{q}_i]\|_0 \le s$, establishing (iii) for i.

Finally, notice that for any vectors $\boldsymbol{p}, \boldsymbol{q}$,

$$\|\mathcal{S}[\boldsymbol{q} + \boldsymbol{p}]\| \le \|\mathcal{S}[\boldsymbol{q}]\| + \|\boldsymbol{p}\|,$$

and further, that $\mathcal{S}[\boldsymbol{q} - \mathcal{S}[\boldsymbol{q}]] = 0$. Hence,

$$\begin{aligned} \|\mathcal{S}[\boldsymbol{q}_i]\| &= \|\mathcal{S}[\boldsymbol{q}_{i-1} - \mathcal{S}[\boldsymbol{q}_{i-1}] + P_\Phi\mathcal{S}[\boldsymbol{q}_{i-1}]]\| \\ &\le \|\mathcal{S}[\boldsymbol{q}_{i-1} - \mathcal{S}[\boldsymbol{q}_{i-1}]]\| + \|P_\Phi\mathcal{S}[\boldsymbol{q}_{i-1}]\| \\ &\le \xi_s(\Phi)\|\mathcal{S}[\boldsymbol{q}_{i-1}]\|. \end{aligned}$$

Applying the inductive hypothesis establishes (ii), and so the three conditions hold for all i.

Since $\|\boldsymbol{q}_i - \boldsymbol{q}_{i-1}\| \le \|\mathcal{S}[\boldsymbol{q}_{i-1}]\| \le \xi_s^i\|\mathcal{S}[\boldsymbol{q}_0]\|$, the sequence $(\boldsymbol{q}_i)$ converges to limit $\boldsymbol{q}^\star$. Moreover, $\|\mathcal{S}[\boldsymbol{q}^\star]\| = \lim_{i\to\infty} \|\mathcal{S}[\boldsymbol{q}_i]\| = 0$, and hence $\mathcal{S}[\boldsymbol{q}^\star] = 0$ and so all of the elements of $\boldsymbol{q}^\star$ have magnitude $\le 1/2$. ∎

D. Ideas for the remainder of the proof.

Lemmas 1 and 2 have shown that to verify that a given $\boldsymbol{x}_0, \boldsymbol{e}_0$ with sign-and-support triplet $(I, J, \boldsymbol{\sigma})$ is the optimal solution to the ℓ^1-minimization (2), it is sufficient to show that (20) holds.[7] This reduces the problem of verifying ℓ^1-recoverability to analyzing three quantities: the restricted operator norm ξ_s, the norm of the initial guess $\|\boldsymbol{q}_0\|_2$, and the norm of the initial violations $\|\mathcal{S}\boldsymbol{q}_0\|_2$. In the next section,

---

[7]Notice that the condition (20) depends on the random matrix A and the random sign-and-support $(I, J, \boldsymbol{\sigma})$ of $\boldsymbol{x}_0$ and $\boldsymbol{e}_0$, through the construction of the matrix Φ.

taking $\boldsymbol{q}_0$ as the least ℓ^2-norm solution to the equation (18), we will bound these quantities, and show that as $m \to \infty$, for any sequence of signal supports I, (20) holds with probability approaching one in the random matrix A and error $(J, \boldsymbol{\sigma})$. As we will formally establish in Section III-D, the probability of failure for a given I will be small enough to allow a union bound over all I, establishing our main Theorem 1. Readers who are less interested in the details can skip the next section in the first reading, and directly jump to the empirical verification in Section IV.

III. PROBABILISTIC ANALYSIS AND FINAL PROOF

In this section, we complete the proof of the main result, by showing that the condition in (20) is satisfied with overwhelming probability. To do so, we must bound three quantities: the norm P_Φ restricted to s-sparse vectors (i.e., $\xi_s(\Phi)$), the norm of the least squares solution $\|\boldsymbol{q}_0\|$, and the norm of the soft-thresholded least squares solution $\|\mathcal{S}\boldsymbol{q}_0\|$.

A. Tools and preliminaries

Our probabilistic analysis depends on measure concentration results for Lipschitz functions of random matrices, in particular their norms and singular values. In the following, let M be an $am \times bm$, $a > b$, random matrix with entries iid $\mathcal{N}(0, 1/m)$. Then by [30]

$$\mathbb{P}[\|M\| > \sqrt{a} + \sqrt{b} + \sqrt{2\varepsilon}] \leq \exp(-\varepsilon m), \qquad (22)$$

$$\mathbb{P}[\sigma_{min}(A) < \sqrt{a} - \sqrt{b} - \sqrt{2\varepsilon}] \leq \exp(-\varepsilon m). \qquad (23)$$

Similarly (e.g., via [3] Equation (2.35))

$$\mathbb{P}[\|M\|_F > \sqrt{abm} + \sqrt{2\varepsilon}] \leq \exp(-\varepsilon m). \qquad (24)$$

We will be interested in the action of random matrices on sparse vectors, which can be studied using the following extremal submatrix singular values[8]:

$$\beta_s(M) \doteq \sup_{\substack{\|\boldsymbol{x}\|_2 \leq 1, \|\boldsymbol{x}\|_0 \leq s}} \|M\boldsymbol{x}\|_2,$$

$$\gamma_s(M) \doteq \inf_{\substack{\|\boldsymbol{x}\|_2 = 1, \|\boldsymbol{x}\|_0 \leq s}} \|M\boldsymbol{x}\|_2,$$

$$\phi_s(M) \doteq \sup_{\substack{\|\boldsymbol{x}\|_2 \leq 1, \|\boldsymbol{x}\|_0 \leq s \\ \|\boldsymbol{y}\|_2 \leq 1, \|\boldsymbol{y}\|_0 \leq s}} \boldsymbol{x}^* M \boldsymbol{y}.$$

Now let $a > 0$ be arbitrary and suppose that $c < \min(a, b)$. Since $\binom{m}{k} < (me/k)^k$, the $am \times bm$ matrix M has $\binom{bm}{cm} \leq \exp\left(mc \log \frac{b}{c}\right)$ submatrices of M_I of size $am \times cm$. Union

---

[8]The quantities β, γ, ϕ have a strong relationship to the restricted isometry property of the matrices in question. In particular, our argument suggests the possibility of a deterministic analogue of Theorem 1 under RIP-like conditions on the matrix Z

bounds then show that

$$\mathbb{P}\left[\beta_{cm}(M) \geq \sqrt{a} + \sqrt{c} + \sqrt{2c \log\left(\frac{b}{c}\right) + 2\varepsilon}\right] < \exp(-\varepsilon m). \qquad (25)$$

$$\mathbb{P}\left[\gamma_{cm}(M) \leq \sqrt{a} - \sqrt{c} - \sqrt{2c \log\left(\frac{b}{c}\right) + 2\varepsilon}\right] < \exp(-\varepsilon m). \qquad (26)$$

$$\mathbb{P}\left[\phi_{cm}(M) \geq 2\sqrt{c} + \sqrt{2c \log\left(\frac{a}{c}\right) + 2c \log\left(\frac{b}{c}\right) + 2\varepsilon}\right]$$
$$< \exp(-\varepsilon m). \qquad (27)$$

We also note that if S is (possibly random) subspace of dimension $a'm < am$ that is probabilistically independent of M, then by the rotational invariance of the Gaussian distribution, the above estimates hold for $P_S M$, with a replaced by a'. We will apply the above lemmas to bound the norms of the components of (11).

B. Projection of sparse vectors

We next show that sparse vectors cannot be *too* aligned with the range of Φ^*:

Lemma 3: Fix any $\delta \geq 1$, $\rho \geq 15/16$, and suppose that $\nu \leq \nu^\star_{\rho,\delta}$, and that $m > m^\star_{\rho,\delta,\nu}$. Fix any support $I \subset [n]$ of size no larger than

$$|I| \leq \alpha^\star_{\delta,\rho,\nu} m, \qquad (28)$$

and suppose that J is chosen uniformly at random. Construct the matrix Φ according to (17). Then for

$$s = C_s(\delta, \rho) m, \qquad (29)$$

the restricted norm $\xi_s(\Phi)$ satisfies

$$1 - \xi_s(\Phi) \geq \nu^4 \delta (1 - \rho)/512, \qquad (30)$$

with probability at least $1 - 10 \exp(-2\varepsilon^\star_{\rho,\delta,\nu} m)$ in Z, J. Here, $C_s(\delta, \rho) > 0$ does not depend on m or ν (and is given explicitly in (8)).

To avoid the possibility for confusion, please notice that the quantity $s = C_s m$ in Lemma 3 represents our bound on the number of elements of $\mathcal{S}[\boldsymbol{q}]$ (i.e., the number of violations of the box constraint), rather than the sparsity of the signal $\boldsymbol{x}$ itself.

Proof: (of Lemma 3) The proof has two parts: we first reduce the quantity of interest to a function of the singular values of several submatrices of U and W, and then apply Gaussian measure concentration to bound these quantities.

a) Lower bounding $1 - \xi$: Notice that

$$1 - \xi_s(\Phi) \geq \tfrac{1}{2}(1 - \xi_s(\Phi)^2) = \tfrac{1}{2}\left(1 - \sup_{\substack{\|\boldsymbol{w}\| \leq 1, \\ \|\boldsymbol{w}\|_0 \leq s}} \|P_\Phi \boldsymbol{w}\|^2\right)$$

$$= \tfrac{1}{2}\left(1 - \sup_{\substack{\|\boldsymbol{w}\| \leq 1, \\ \|\boldsymbol{w}\|_0 \leq s}} \|\boldsymbol{w}\|^2 - \|\boldsymbol{w} - P_\Phi \boldsymbol{w}\|^2\right)$$

$$= \tfrac{1}{2}\left(1 - \sup_{\substack{\|\boldsymbol{w}\| = 1, \\ \|\boldsymbol{w}\|_0 \leq s}} 1 - \|\boldsymbol{w} - P_\Phi \boldsymbol{w}\|^2\right)$$

$$= \tfrac{1}{2} \inf_{\substack{\|\boldsymbol{w}\| = 1, \\ \|\boldsymbol{w}\|_0 \leq s}} \|\boldsymbol{w} - P_\Phi \boldsymbol{w}\|^2 = \tfrac{1}{2} \inf_{\substack{\|\boldsymbol{w}\| = 1, \\ \|\boldsymbol{w}\|_0 \leq s}} \min_{\boldsymbol{r}} \|\boldsymbol{w} - \Phi^* \boldsymbol{r}\|^2.$$

We can write this more explicitly as

$$\frac{1}{2} \inf_{\substack{\|\boldsymbol{w}_1\|^2+\|\boldsymbol{w}_2\|^2=1 \\ \|\boldsymbol{w}_1\|_0+\|\boldsymbol{w}_2\|_0\leq s}} \min_{\boldsymbol{r}} \left\| \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{bmatrix} - \begin{bmatrix} \nu U^* & W^* \\ \mathtt{I} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \end{bmatrix} \right\|^2$$

$$= \frac{1}{2} \inf_{\boldsymbol{w}_1,\boldsymbol{w}_2} \min_{\boldsymbol{u}_1,\boldsymbol{u}_2} \left\| \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{bmatrix} - \begin{bmatrix} \nu U^* & W^* \\ \mathtt{I} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 + \boldsymbol{w}_2 \\ \boldsymbol{u}_2 \end{bmatrix} \right\|^2,$$

$$= \frac{1}{2} \inf_{\boldsymbol{w}} \min_{\boldsymbol{u}} \|\boldsymbol{w}_1 - \nu U^*\boldsymbol{u}_1 - \nu U^*\boldsymbol{w}_2 - W^*\boldsymbol{u}_2\|^2 + \|\boldsymbol{u}_1\|^2. \tag{31}$$

Notice that $W \in \mathbb{R}^{(k_1+1)\times(1-\rho)m}$ has a nontrivial nullspace; let $\Gamma \subset \mathbb{R}^{(1-\rho)m} = \text{null}(W)$, let P_Γ denote the projection onto Γ, and P_W denote the projection onto range(W^*), so that

$$P_W + P_\Gamma = \mathtt{I}, \quad P_W P_\Gamma = 0, \quad \text{and} \quad P_\Gamma W^* = 0.$$

Then in terms of these operators, the first term in (III-B0a) can be written as

$$\|\boldsymbol{w}_1 - \nu U^*\boldsymbol{u}_1 - \nu U^*\boldsymbol{w}_2 - W^*\boldsymbol{u}_2\|^2$$
$$= \|P_\Gamma(\boldsymbol{w}_1 - \nu U^*\boldsymbol{u}_1 - \nu U^*\boldsymbol{w}_2 - W^*\boldsymbol{u}_2)\|^2$$
$$\quad + \|P_W(\boldsymbol{w}_1 - \nu U^*\boldsymbol{u}_1 - \nu U^*\boldsymbol{w}_2 - W^*\boldsymbol{u}_2)\|^2,$$
$$\geq \|P_\Gamma(\boldsymbol{w}_1 - \nu U^*\boldsymbol{u}_1 - \nu U^*\boldsymbol{w}_2 - W^*\boldsymbol{u}_2)\|^2,$$
$$= \|\nu P_\Gamma U^*\boldsymbol{u}_1 - (P_\Gamma\boldsymbol{w}_1 - \nu P_\Gamma U^*\boldsymbol{w}_2)\|^2.$$

Plugging into (III-B0a), we find that $1 - \xi(\Phi)$ is bounded below by

$$\frac{1}{2} \inf_{\boldsymbol{w}_1,\boldsymbol{w}_2} \min_{\boldsymbol{u}_1} \|\nu P_\Gamma U^*\boldsymbol{u}_1 - (P_\Gamma\boldsymbol{w}_1 - \nu P_\Gamma U^*\boldsymbol{w}_2)\|^2 + \|\boldsymbol{u}_1\|^2. \tag{32}$$

For any matrix M and vector and $\boldsymbol{y}$, the unique solution to

$$\min_{\boldsymbol{x}} \|M\boldsymbol{x} - \boldsymbol{y}\|^2 + \|\boldsymbol{x}\|^2$$

is given by $\boldsymbol{x}^\star(\boldsymbol{y}) = (\mathtt{I} + M^*M)^{-1}M^*\boldsymbol{y}$. Hence, in (32), for any $\boldsymbol{w} = [\boldsymbol{w}_1^*, \boldsymbol{w}_2^*]^*$, the optimizer $\boldsymbol{u}_1^\star(\boldsymbol{w})$ has the explicit form

$$\hat{\boldsymbol{u}}_1^\star(w) = \nu(\mathtt{I} + \nu^2 U P_\Gamma U^*)^{-1}U P_\Gamma(\boldsymbol{w}_1 - \nu U^*\boldsymbol{w}_2)$$
$$= \nu\left(\mathtt{I} + \nu^2 U P_\Gamma U^*\right)^{-1} U\Psi\boldsymbol{w}, \tag{33}$$

where we have used Ψ for the matrix $[P_\Gamma \mid -\nu P_\Gamma U^*]$. Notice that for any $\boldsymbol{w}$,

$$\|\boldsymbol{u}^\star(\boldsymbol{w})\| \geq \frac{\nu\,\sigma_{min}(U)}{\|\mathtt{I} + \nu^2 U P_\Gamma U^*\|}\|\Psi\boldsymbol{w}\| \geq \frac{\nu\,\sigma_{min}(U)}{1 + \nu^2\|U\|^2}\|\Psi\boldsymbol{w}\|.$$

Hence, for any $\boldsymbol{w}$ satisfying $\|\boldsymbol{w}\| = 1$ and $\|\boldsymbol{w}\|_0 \leq s$,

$$\|\boldsymbol{u}^\star(\boldsymbol{w})\| \geq \frac{\nu\,\sigma_{min}(U)}{1 + \nu^2\|U\|^2}\gamma_s(\Psi). \tag{34}$$

Applying (32), we have that

$$1 - \xi_s(\Phi) \geq \frac{1}{2} \inf_{\substack{\|\boldsymbol{w}\|=1 \\ \|\boldsymbol{w}\|_0\leq s}} \|\boldsymbol{u}^\star(\boldsymbol{w})\|^2 \geq \frac{\nu^2\,\sigma_{min}^2(U)}{2\left(1 + \nu^2\|U\|^2\right)^2} \gamma_s^2(\Psi). \tag{35}$$

We apply a sequence of straightfoward algebraic manipulations to lower bound $\gamma_s(\Psi)$ in terms of the singular values of

submatrices of U and W. Each step below should be intuitive; their correctness is proved in Lemmas 6-9 of the appendix:

$$\gamma_s(\Psi) \geq \nu\gamma_s\left([P_\Gamma \mid P_\Gamma U^*]\right), \qquad \text{by (59)},$$
$$= \nu\gamma_s\left([\mathtt{I} - P_W \mid (\mathtt{I} - P_W)U^*]\right),$$
$$\geq \nu\left(\gamma_s\left([\mathtt{I}\ U^*]\right) - \beta_s(P_W) - \beta_s(P_W U^*)\right) \qquad \text{by (60)},$$
$$\geq \nu\left(\gamma_s\left([\mathtt{I}\ U^*]\right) - \frac{\beta_s(W)}{\sigma_{min}(W)} - \beta_s(P_W U^*)\right), \quad \text{by (62)}.$$

Finally, by (63),

$$\gamma_s\left([\mathtt{I}\ U^*]\right) \geq \sqrt{\min\{1, \gamma_s^2(U^*)\} - \phi_s(U^*)}.$$

b) Concentration of measure: For convenience, below we write $\varepsilon = 2\varepsilon_{\delta,\rho,\nu}^\star$. We will prove that when C_s is chosen according to (8), the following bounds hold:

(i) $\quad\quad \|U\| \leq 3\sqrt{\delta}$ w.p. $\geq 1 - \exp(-\varepsilon m)$,

(ii) $\quad\quad \sigma_{min}(U) \geq \sqrt{\delta}/2$ w.p. $\geq 1 - \exp(-\varepsilon m)$,

(iii) $\quad\quad \gamma_s(U^*) \geq \sqrt{(1-\rho)/2}$ w.p. $\geq 1 - \exp(-\varepsilon m)$,

(iv) $\quad\quad \phi_s(U^*) \leq \dfrac{1-\rho}{4}$ w.p. $\geq 1 - \exp(-\varepsilon m)$,

(v) $\quad \beta_s(P_W U^*) \leq \dfrac{\sqrt{1-\rho}}{8}$ w.p. $\geq 1 - \exp(-\varepsilon m)$,

(vi) $\quad\quad \sigma_{min}(W) \geq 1/4$ w.p. $\geq 1 - 3\exp(-\varepsilon m)$,

(vii) $\quad\quad \beta_s(W) \leq \dfrac{\sqrt{1-\rho}}{32}$ w.p. $\geq 1 - 2\exp(-\varepsilon m)$.

When each of the above bounds holds, then

$$\gamma_s(\Psi) \geq \frac{\nu\sqrt{1-\rho}}{4},$$

and by (35)

$$1 - \xi_s(\Phi) \geq \frac{\nu^2\delta}{8(1 + 9\nu^2\delta)^2}\gamma_s^2(\Psi) \geq \frac{\nu^2\delta}{32}\gamma_s^2(\Psi),$$

where we have used that $\nu < 1/3\sqrt{\delta}$. Combining the two bounds establishes that $1 - \xi \geq \nu^4\delta(1-\rho)/512$ on this intersection of events; summing the failure probabilities shows that the bound holds with probability at least $1 - 10\exp(-\varepsilon m)$.

We conclude the proof by establishing the probabilistic bounds (i)-(vii). It is convenient to note that since $\Upsilon(t) \leq t$, the choice of C_s in (8) and the fact that $C_\mu \geq 1$ imply that

$$\sqrt{C_s} \leq C_\mu\sqrt{C_s} \leq 2^{-8}(1-\rho). \tag{36}$$

Furthermore, using (6) and the fact that $C_\mu, \delta \geq 1$,

$$C_s\log\left(\frac{\delta}{C_s}\right) \leq 2^{-17}C_\mu^{-2}\delta^{-1}(1-\rho)^2 \leq 2^{-17}(1-\rho)^2$$

and so

$$\sqrt{2C_s\log\left(\frac{\delta}{C_s}\right)} \leq 2^{-8}(1-\rho). \tag{37}$$

We note three remaining bounds of interest. First, notice from (9) that $m^\star = 2^{17}/(1-\rho)C_s$, and so using (36),

$$m^\star \geq 2^{33}(1-\rho)^{-2}. \tag{38}$$

Furthermore, from (9), and (36)

$$\nu^\star \leq \sqrt{2^{-24}\delta^{-1}C_s} \leq 2^{-20}(1-\rho),$$

where we have used $\delta \geq 1$ to drop the δ^{-1} term. Since in (9), $\varepsilon^\star \leq \nu^2$, we have that $\varepsilon^\star \leq 2^{-40}(1-\rho)^2$. In (9), $\alpha^\star = \Upsilon[\varepsilon^\star/\delta] \leq \varepsilon^\star/\delta \leq \varepsilon^\star$. Hence, $\alpha^\star \leq 2^{-40}(1-\rho)^2$, and for any $m \geq m^\star$,

$$
\begin{aligned}
\sqrt{\alpha^\star + \tfrac{1}{m}} &\leq \sqrt{\alpha^\star} + \sqrt{\tfrac{1}{m^\star}} \\
&\leq 2^{-33/2}(1-\rho) + 2^{-20}(1-\rho) \leq 2^{-8}(1-\rho).
\end{aligned} \tag{39}
$$

Using that $\varepsilon = 2\varepsilon^\star$, it is convenient to record the weak bound

$$
\sqrt{2\varepsilon} \leq 2^{-8}(1-\rho). \tag{40}
$$

(i). Notice that U is an $(n-k_1) \times (m-k_2)$ iid $\mathcal{N}(0,1/m)$ matrix. Applying (22), with probability at least $1-\exp(-\varepsilon m)$,

$$
\|U\| \leq \sqrt{\delta - \alpha^\star} + \sqrt{1-\rho} + \sqrt{2\varepsilon}.
$$

Since $\sqrt{1-\rho} < 1 \leq \sqrt{\delta}$ and (applying (40)) $\sqrt{2\varepsilon} \leq 2^{-8}(1-\rho) \ll 1 \leq \sqrt{\delta}$, on the above good event $\|U\| \leq 3\sqrt{\delta}$.

(ii). Similarly, with probability at least $1-\exp(-\varepsilon m)$,

$$
\begin{aligned}
\sigma_{min}(U) &\geq \sqrt{\delta - \alpha^\star} - \sqrt{1-\rho} - \sqrt{2\varepsilon} \\
&\geq \sqrt{\delta} - \sqrt{\alpha^\star} - \sqrt{1-\rho} - \sqrt{2\varepsilon}
\end{aligned}
$$

Since $\rho \geq 15/16$, $\sqrt{1-\rho} \leq 1/4 \leq \sqrt{\delta}/4$. Similarly, $\sqrt{2\varepsilon} \leq 2^{-8}(1-\rho) \leq 2^{-8}\sqrt{\delta}$. Finally, by (39), $\sqrt{\alpha^\star} \leq 2^{-8} \leq 2^{-8}\sqrt{\delta}$ and so on the above good event,

$$
\sigma_{min}(U) \geq \sqrt{\delta} \times (1 - 2^{-8} - 1/4 - 2^{-8}) > \sqrt{\delta}/2.
$$

(iii). Similarly, applying (26) to U^* gives that with probability at least $1-\exp(-\varepsilon m)$,

$$
\begin{aligned}
\gamma_s(U^*) &\geq \sqrt{1-\rho} - \sqrt{C_s} - \sqrt{2C_s \log\left(\tfrac{\delta}{C_s}\right)} - \sqrt{2\varepsilon}, \\
&\geq \sqrt{1-\rho} - 3 \times 2^{-8} \times (1-\rho), \\
&\geq (1 - 3 \times 2^{-8})\sqrt{1-\rho} > \sqrt{(1-\rho)/2},
\end{aligned}
$$

where above we have used (36), (37) and (40).

(iv). Using (27), with probability at least $1-\exp(-\varepsilon m)$,

$$
\begin{aligned}
\phi_s(U^*) &\leq 2\sqrt{C_s} + \sqrt{2C_s \log\left(\tfrac{\delta}{C_s}\right) + 2C_s \log\left(\tfrac{1-\rho}{C_s}\right) + 2\varepsilon}, \\
&< 2\sqrt{C_s} + 2\sqrt{2C_s \log\left(\tfrac{\delta}{C_s}\right)} + \sqrt{2\varepsilon}, \\
&\leq 5 \times 2^{-8}(1-\rho) < (1-\rho)/4.
\end{aligned}
$$

Above we have used $\log\left(\tfrac{1-\rho}{C_s}\right) \leq \log\left(\tfrac{\delta}{C_s}\right)$ (which follows since $\delta \geq 1 > 1-\rho$), and have applied (36), (37) and (40).

(v). Next consider the term $P_W U^*$. The matrix P_W projects the columns of U^* onto the independent $k_1 + 1$-dimensional subspace $\text{range}(W^*)$. Since W and U are probabilistically independent, $\beta_s(P_W U^*)$ is distributed as $\beta_s(M)$, where M is a $(k_1+1) \times (n-k_1)$ iid $\mathcal{N}(0,1/m)$ matrix. In particular, using (25), with probability at least $1-\exp(-\varepsilon m)$,

$$
\begin{aligned}
\beta_s(P_W U^*) &\leq \sqrt{\tfrac{k_1+1}{m}} + \sqrt{\tfrac{s}{m}} + \sqrt{2\tfrac{s}{m}\log\left(\tfrac{n-k_1}{s}\right) + 2\varepsilon}, \\
&\leq \sqrt{\alpha^\star + \tfrac{1}{m}} + \sqrt{C_s} + \sqrt{2C_s \log\left(\tfrac{\delta}{C_s}\right)} + \sqrt{2\varepsilon}, \\
&\leq 4 \times 2^{-8}(1-\rho) < 2^{-6}(1-\rho) < 2^{-3}\sqrt{1-\rho},
\end{aligned}
$$

where we have used (39), (36), (37) and (40) to bound the four terms.

(vi). Lemma 11, proved in the appendix, shows that with probability at least $1 - 3\exp(-\varepsilon m)$, $\sigma_{min}(W) > 1/4$.

(vii). Using the definition of W and (61),

$$
\beta_s(W) \leq \frac{\beta_s((Z_{J^c,I})^*)}{\sqrt{1-\rho}} + \frac{\beta_s(\boldsymbol{\mu}_{J^c}^*)}{\|\boldsymbol{\mu}_{J^c}\|}. \tag{41}
$$

With probability at least $1 - \exp(-\varepsilon m)$,

$$
\begin{aligned}
\beta_s((Z_{J^c,I})^*) &\leq \sqrt{\tfrac{k_1}{m}} + \sqrt{\tfrac{s}{m}} + \sqrt{\tfrac{2s}{m}\log\left(\tfrac{(1-\rho)m}{s}\right) + 2\varepsilon} \\
&\leq \sqrt{\alpha^\star} + \sqrt{C_s} + \sqrt{2C_s \log\left(\tfrac{1-\rho}{m}\right)} + \sqrt{2\varepsilon}, \\
&\leq 4 \times 2^{-8}(1-\rho) = 2^{-6}(1-\rho).
\end{aligned}
$$

For the second term in (41), notice that $\boldsymbol{\mu}_{J^c}^*$ is simply a row vector of length $m - k_2$, and so $\beta_s(\boldsymbol{\mu}_{J^c}^*)$ is simply the largest Euclidean norm of any vector formed from s elements of $\boldsymbol{\mu}_{J^c}$:

$$
\begin{aligned}
\beta_s(\boldsymbol{\mu}_{J^c}^*) &\leq \beta_s(\boldsymbol{\mu}^*) = \sup_{|L| \leq s} \|\boldsymbol{\mu}_L\| = C_\mu \sqrt{s/m} \\
&= C_\mu \sqrt{C_s} \leq 2^{-8}(1-\rho),
\end{aligned}
$$

where we have used that $\|\boldsymbol{\mu}\|_\infty \leq C_\mu m^{-1/2}$, and invoked (36) to bound $C_\mu \sqrt{C_s}$. Lemma 10 shows that with probability at least $1 - \exp(-\varepsilon m)$,

$$
\|\boldsymbol{\mu}_{J^c}\|^{-1} \leq 2(1-\rho)^{-1/2}.
$$

Hence, with overall probability at least $1 - 2\exp(-\varepsilon m)$,

$$
\beta_s(W) \leq 2^{-6}\sqrt{1-\rho} + 2^{-7}\sqrt{1-\rho} < 2^{-5}\sqrt{1-\rho}.
$$

This completes the proof. ∎

C. Initial certificate

In this section, we analyze the initial separator $\boldsymbol{q}_0$, obtained as the minimum ℓ^2-norm solution to the equation (18). We upper bound both $\|\boldsymbol{q}_0\|_2$ and $\|\mathcal{S}\boldsymbol{q}_0\|_2$. These bounds provide the second half of the conditions needed in Lemma 2 to show that $\boldsymbol{q}_0$ can be refined by alternating projections to give a certificate of optimality.

Lemma 4: Fix any $\delta \geq 1$, $\rho < 1$. Suppose that $\nu < \nu_{\delta,\rho}^\star$ and $m > m_{\delta,\rho,\nu}^\star$. Fix any subset $I \subset [n]$ of size

$$
|I| \leq \alpha_{\delta,\rho,\nu}^\star m. \tag{42}
$$

Then the minimum ℓ^2-norm solution $\boldsymbol{q}_0$ to the system of equations (18) satisfies

$$
\begin{aligned}
\|\boldsymbol{q}_0\|_2 &\leq C\nu\sqrt{\delta m} + O(1). \\
\|\mathcal{S}\boldsymbol{q}_0\|_2 &\leq C\nu\sqrt{\delta m}\exp(-C'\delta^{-1}\nu^{-2}) + O(1).
\end{aligned} \tag{43}
$$

simultaneously for all $\boldsymbol{x}_0$ with $\text{supp}(\boldsymbol{x}_0) \subseteq I$, with probability at least $1 - 21\exp(-2\varepsilon_{\delta,\rho,\nu}^\star m)$ in Z, J, σ.

Here, C and C' are numerical constants. In particular, the result holds with $C = 312$ and $C' = 1/48^2$. With these choices, the $O(1)$ term is upper bounded by $48(1-\rho)^{-1/2}$.

The proof of Lemma 4 is largely based on the following general concentration lemma for functions of iid Rademacher sequences. Its proof is straightforward; we delay it to the appendix.

Lemma 5: Fix any matrix $M \in \mathbb{R}^{m \times n}$ and let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ be a sequence of iid Rademacher random variables. Let $\mathcal{S}$ denote the shrinkage operator. Then for any $\varepsilon > 0$,

$$\mathbb{P}\left[\|M\boldsymbol{\sigma}\| > 2\|M\|_F + 4\|M\|\sqrt{\varepsilon n}\right] < 4\exp(-\varepsilon n). \quad (44)$$

$$\mathbb{P}\left[\|\mathcal{S}[M\boldsymbol{\sigma}]\| > 2\|M\|_F \exp\left(-1/4\|M\|_{2,\infty}^2\right) + 4\|M\|\sqrt{\varepsilon n}\right]$$
$$< 4\exp(-\varepsilon n). \quad (45)$$

With Lemma 5 in hand, to establish Lemma 4 we need only estimate the relevant norms for our system of equations:

Proof: (of Lemma 4) Below, we will let

$$\varepsilon \doteq 2\varepsilon_{\delta,\rho,\nu}^{\star} = \nu^2\tau^2(1-\rho), \quad (46)$$

where in the second step we have used (9). It is further convenient to note that by (9), $\alpha^{\star} = \nu^2\Upsilon[\varepsilon^{\star}/\delta] \le \nu^2\varepsilon^{\star}/\delta \le \frac{1}{2}\nu^4\tau^2(1-\rho)/\delta$, where we have used that $\Upsilon[t] \le t$ for all t. Using that $\delta \ge 1$, we record the slightly weaker bound

$$\alpha_{\delta,\rho,\nu}^{\star} \le \nu^4\tau^2(1-\rho). \quad (47)$$

Of course, it is easy to see that $\alpha^{\star}$ and $\sqrt{2\varepsilon}$ are both much smaller than one.

Conditioning of Φ. Whenever $\Phi\Phi^*$ is invertible, the pseudoinverse operator

$$\Phi^{\dagger} \doteq \Phi^*(\Phi\Phi^*)^{-1}$$

is well-defined, and $\|\Phi^{\dagger}\| = 1/\sigma_{min}(\Phi)$. We will let $\mathcal{E}_{\Phi}$ denote the event

$$\mathcal{E}_{\Phi} = \left\{\Phi\Phi^* \text{ is invertible, and } \|\Phi^{\dagger}\| \le 8\right\}.$$

Notice that

$$\sigma_{min}(\Phi) \ge \min(1, \sigma_{min}(W)) - \nu\|U\|.$$

Lemma 11, proved in the appendix, shows that with probability at least $1 - 3\exp(-\varepsilon m)$, $\sigma_{min}(W) \ge 1/4$. Applying (22) to U shows that with probability $1 - \exp(-\varepsilon m)$,

$$\nu\|U\| \le \nu(\sqrt{\delta} + \sqrt{1-\rho} + \sqrt{2\varepsilon}) \le 3\nu\sqrt{\delta} \le 1/8,$$

where we have used that

$$\nu^{\star} < 1/24\sqrt{\delta}.$$

So, with probability at least $1 - 4\exp(-\varepsilon m)$, $\sigma_{min}(\Phi) \ge 1/8$, and

$$\mathbb{P}[\mathcal{E}_{\Phi}] \ge 1 - 4\exp(-\varepsilon m).$$

Bounding the solution. Hence, on $\mathcal{E}_{\Phi}$, the minimum ℓ^2 norm solution $\boldsymbol{q}_0$ to (18) is given by

$$\boldsymbol{q}_0 = \nu\Phi^{\dagger}R\boldsymbol{\sigma} + \Phi^{\dagger}T\boldsymbol{\sigma} + \Phi^{\dagger}\boldsymbol{z}.$$

We define four bad events, three of which involve the norms of the elements of $\boldsymbol{q}_0$ deviating above their expectations, and the one of which involves the norm of a soft-thresholded version of the first component deviating above its expectation:

$$\mathcal{E}_1 = \left\{\sigma_{min}(\Phi) = 0, \text{ or } \|\nu\Phi^{\dagger}R\boldsymbol{\sigma}\| > 128\nu\sqrt{\delta m}\right\},$$

$$\mathcal{E}_2 = \left\{\sigma_{min}(\Phi) = 0, \text{ or } \|\mathcal{S}[\nu\Phi^{\dagger}R\boldsymbol{\sigma}]\| > 128\nu\tau\sqrt{\delta m}\right\},$$

$$\mathcal{E}_3 = \left\{\sigma_{min}(\Phi) = 0, \text{ or } \|\Phi^{\dagger}T\boldsymbol{\sigma}\| > 176\nu\tau\sqrt{\delta m} + \frac{48}{\sqrt{1-\rho}}\right\},$$

$$\mathcal{E}_4 = \left\{\sigma_{min}(\Phi) = 0, \text{ or } \|\Phi^{\dagger}\boldsymbol{z}\| > 8\nu\tau\sqrt{\delta m}\right\}.$$

We will see below that the forms of the bounds in these events arise naturally from the structure of (18); the numerical constants however are pessimistic. On $\cap_i \mathcal{E}_i^c$, using that $\tau \le 1$

$$\|\boldsymbol{q}_0\| \le \|\nu\Phi^{\dagger}R\boldsymbol{\sigma}\| + \|\Phi^{\dagger}T\boldsymbol{\sigma}\| + \|\Phi^{\dagger}\boldsymbol{z}\|$$
$$\le 312\nu\sqrt{\delta m} + O(1).$$

Similarly, noting that for any vectors $\boldsymbol{x}, \boldsymbol{y}$ $\|\mathcal{S}[\boldsymbol{x} - \boldsymbol{y}]\| \le \|\mathcal{S}[\boldsymbol{x}]\| + \|\boldsymbol{y}\|$, on $\cap_i \mathcal{E}_i^c$,

$$\|\mathcal{S}[\boldsymbol{q}_0]\| \le \|\mathcal{S}[\nu\Phi^{\dagger}R\boldsymbol{\sigma}]\| + \|\Phi^{\dagger}T\boldsymbol{\sigma}\| + \|\Phi^{\dagger}\boldsymbol{z}\|$$
$$\le 312\nu\tau\sqrt{\delta m} + O(1).$$

We next bound the probability of these four bad events.

$\mathcal{E}_1$: **bounding** $\|\nu\Phi^{\dagger}R\boldsymbol{\sigma}\|$. With probability at least $1 - 2\exp(-\varepsilon m)$, the following two bounds hold:

$$\|R\| \le \sqrt{\delta} + \sqrt{\rho} + \sqrt{2\varepsilon} \le 3\sqrt{\delta}, \quad (48)$$
$$\|R\|_F \le \sqrt{\delta\rho m} + \sqrt{2\varepsilon} \le 2\sqrt{\delta m}, \quad (49)$$

where we have used the weak bound $\sqrt{2\varepsilon} < 1 \le \sqrt{\delta}$ (which is a consequence, e.g., of (46)) and $\sqrt{\rho} \le 1 \le \sqrt{\delta}$. Let $\mathcal{E}_R$ denote the event that (48)-(49) hold. On $\mathcal{E}_{\Phi} \cap \mathcal{E}_R$,

$$2\|\nu\Phi^{\dagger}R\|_F + 4\|\nu\Phi^{\dagger}R\|\sqrt{\varepsilon m}$$
$$\le 2\nu\|\Phi^{\dagger}\|\|R\|_F + 4\nu\|\Phi^{\dagger}\|\|R\|\sqrt{\varepsilon m}$$
$$\le 32\nu\sqrt{\delta m} + 96\nu\sqrt{\delta\varepsilon m} \le 128\nu\sqrt{\delta m},$$

where again we have used that $\varepsilon < 1$. Then,

$$\mathbb{P}[\mathcal{E}_1] = \mathbb{P}[\mathcal{E}_1 \mid \mathcal{E}_{\Phi} \cap \mathcal{E}_R]\mathbb{P}[\mathcal{E}_{\Phi} \cap \mathcal{E}_R]$$
$$+ \mathbb{P}[\mathcal{E}_1 \mid (\mathcal{E}_{\Phi} \cap \mathcal{E}_R)^c]\mathbb{P}[(\mathcal{E}_{\Phi} \cap \mathcal{E}_R)^c]$$
$$\le \mathbb{P}[\mathcal{E}_1 \mid \mathcal{E}_{\Phi} \cap \mathcal{E}_R] + \mathbb{P}[\mathcal{E}_{\Phi}^c \cup \mathcal{E}_R^c]$$
$$\le \sup_{Z_0, J_0 \in \mathcal{E}_{\Phi} \cap \mathcal{E}_R} \mathbb{P}[\mathcal{E}_1 \mid Z_0, J_0] + \mathbb{P}[\mathcal{E}_{\Phi}^c] + \mathbb{P}[\mathcal{E}_R^c]$$
$$\le \sup_{Z_0, J_0 \in \mathcal{E}_{\Phi} \cap \mathcal{E}_R} \mathbb{P}[\mathcal{E}_1 \mid Z_0, J_0] + 5\exp(-\varepsilon m). \quad (50)$$

For any $(Z_0, J_0) \in \mathcal{E}_{\Phi} \cap \mathcal{E}_R$, applying Lemma 5 with $\Phi^{\dagger}R$ fixed gives

$$\mathbb{P}[\mathcal{E}_1 \mid Z_0, J_0] = \mathbb{P}\left[\|\nu\Phi^{\dagger}R\boldsymbol{\sigma}\| > 128\nu\sqrt{\delta m} \mid Z_0, J_0\right]$$
$$\le \mathbb{P}\left[\|\nu\Phi^{\dagger}R\boldsymbol{\sigma}\| > 2\|\nu\Phi^{\dagger}R\|_F + 4\|\nu\Phi^{\dagger}R\|\sqrt{\varepsilon m} \mid Z_0, J_0\right]$$
$$\le \exp(-\varepsilon m),$$

and $\mathbb{P}[\mathcal{E}_1] \le 6\exp(-\varepsilon m)$.

$\mathcal{E}_2$: **bounding** $\|\mathcal{S}[\nu\Phi^{\dagger}R\boldsymbol{\sigma}]\|$. By the same calculation as (50)

$$\mathbb{P}[\mathcal{E}_2] \le \sup_{Z_0, J_0 \in \mathcal{E}_{\Phi} \cap \mathcal{E}_R} \mathbb{P}[\mathcal{E}_2 \mid Z_0, J_0] + 5\exp(-\varepsilon m).$$

3ION THEORY, 2009.

11

Since
$$\|\nu\Phi^\dagger R\|_{2,\infty} \le \|\nu\Phi^\dagger R\| \le \nu\|\Phi^\dagger\|\|R\|,$$

for any fixed $Z_0, J_0 \in \mathcal{E}_\Phi \cap \mathcal{E}_R$,

$$\exp\left(-\frac{1}{4\|\nu\Phi^\dagger R\|_{2,\infty}^2}\right) \le \exp\left(-\frac{1}{4\nu^2\|\Phi\|^2\|R\|^2}\right)$$
$$\le \exp\left(-\frac{1}{48^2\delta\nu^2}\right) = \tau,$$

and further

$$2\|\nu\Phi^\dagger R\|_F \exp\left(-\frac{1}{4\|\nu\Phi^\dagger R\|_{2,\infty}^2}\right) + 4\|\nu\Phi^\dagger R\|\sqrt{\varepsilon m}$$
$$\le 32\nu\tau\sqrt{\delta m} + 96\nu\sqrt{\delta\varepsilon m} \le 128\nu\tau\sqrt{\delta m},$$

where in the last step we have used that $\varepsilon < \tau^2$ (see (46)). Applying Lemma 5,

$$\mathbb{P}[\mathcal{E}_2 \mid Z_0, J_0] = \mathbb{P}[\|\mathcal{S}[\nu\Phi^\dagger R\boldsymbol{\sigma}]\| > 128\nu\tau\sqrt{\delta m} \mid Z_0, J_0]$$
$$\le \mathbb{P}\left[\|\mathcal{S}[\nu\Phi^\dagger R\boldsymbol{\sigma}]\| > \binom{2\|\nu\Phi^\dagger R\|_F e^{-\frac{1}{4\|\nu\Phi^\dagger R\|_{2,\infty}^2}}}{+ 4\|\nu\Phi^\dagger R\|\sqrt{\varepsilon\rho m}} \mid Z_0, J_0\right]$$
$$\le \exp(-\varepsilon m).$$

We conclude that $\mathbb{P}[\mathcal{E}_2] < 6\exp(-\varepsilon m)$.

$\mathcal{E}_3$: **bounding** $\|\Phi^\dagger T\boldsymbol{\sigma}\|$. Notice that if $\|\cdot\|_\diamond$ is either the Frobenius or operator norm,

$$\|T\|_\diamond \le \frac{\|Z_{J,I}\|_\diamond}{\sqrt{1-\rho}} + \frac{\|\boldsymbol{\mu}_J\|}{\|\boldsymbol{\mu}_{J^c}\|}.$$

Lemma 10 shows that with probability at least $1 - \exp(-\varepsilon m)$,

$$\|\boldsymbol{\mu}_{J^c}\|^2 > (1-\rho)/4,$$

and so with the same probability

$$\|\boldsymbol{\mu}_J\|/\|\boldsymbol{\mu}_{J^c}\| \le 1/\|\boldsymbol{\mu}_{J^c}\| \le 2(1-\rho)^{-1/2}.$$

Notice further that with probability at least $1 - 2\exp(-\varepsilon m)$,

$$\|Z_{J,I}\| \le \sqrt{\rho} + \sqrt{\alpha^\star} + \sqrt{2\varepsilon} \le 3,$$
$$\|Z_{J,I}\|_F \le \sqrt{\alpha^\star\rho m} + \sqrt{2\varepsilon} \le \sqrt{\alpha^\star m} + 1.$$

The final bounds above are loose, but only impact the numerical constants below. Since (47) implies that $\alpha^\star \le \nu^2\tau^2$, with probability at least $1 - 3\exp(-\varepsilon m)$,

$$\|T\| \le \frac{\|Z_{J,I}\|}{\sqrt{1-\rho}} + \frac{\|\boldsymbol{\mu}_J\|}{\|\boldsymbol{\mu}_{J^c}\|} \le \frac{5}{\sqrt{1-\rho}}, \quad (51)$$
$$\|T\|_F \le \frac{\|Z_{J,I}\|_F}{\sqrt{1-\rho}} + \frac{\|\boldsymbol{\mu}_J\|}{\|\boldsymbol{\mu}_{J^c}\|} \le \nu\tau\sqrt{m} + \frac{3}{\sqrt{1-\rho}}. \quad (52)$$

Let $\mathcal{E}_T$ denote the event that (51)-(52) hold. On $\mathcal{E}_T \cap \mathcal{E}_\Phi$,

$$2\|\Phi^\dagger T\|_F + 4\|\Phi^\dagger T\|\sqrt{\varepsilon m}$$
$$\le 2\|\Phi^\dagger\|\|T\|_F + 4\|\Phi^\dagger\|\|T\|\sqrt{\varepsilon m}$$
$$\le 16\nu\tau\sqrt{m} + 48(1-\rho)^{-1/2} + 160(1-\rho)^{-1/2}\sqrt{\varepsilon m},$$
$$\le 176\nu\tau\sqrt{\delta m} + 48(1-\rho)^{-1/2},$$

where we have used (46) to bound $\varepsilon \le \nu^2\tau^2(1-\rho)$. Then,

$$\mathbb{P}[\mathcal{E}_3] \le \sup_{Z_0, J_0 \in \mathcal{E}_\Phi \cap \mathcal{E}_T} \mathbb{P}[\mathcal{E}_3 \mid Z_0, J_0] + \mathbb{P}[\mathcal{E}_\Phi^c] + \mathbb{P}[\mathcal{E}_T^c]$$
$$\le \sup_{Z_0, J_0 \in \mathcal{E}_\Phi \cap \mathcal{E}_T} \mathbb{P}[\mathcal{E}_3 \mid Z_0, J_0] + 6\exp(-\varepsilon m).$$

Applying Lemma 5 with $\Phi^\dagger T$ fixed,

$$\mathbb{P}[\mathcal{E}_3 \mid Z_0, J_0]$$
$$\le \mathbb{P}[\|\Phi^\dagger T\boldsymbol{\sigma}\| \ge 2\|\Phi^\dagger T\|_F + 4\|\Phi^\dagger T\|\sqrt{\varepsilon m} \mid Z_0, J_0]$$
$$\le \exp(-\varepsilon m).$$

and $\mathbb{P}[\mathcal{E}_3] \le 7\exp(-\varepsilon m)$.

$\mathcal{E}_4$: **bounding** $\|\Phi^\dagger z\|$. Finally, on $\mathcal{E}_\Phi$,

$$\|\Phi^\dagger z\| \le \|\Phi^\dagger\|\|z\| \le \frac{8}{\nu}\sqrt{\frac{\alpha^\star m}{1-\rho}} \le 8\nu\tau\sqrt{m}.$$

Here we have used that by (47), $\alpha^\star \le \nu^4\tau^2(1-\rho)$. So, $P[\mathcal{E}_4] \le P[\mathcal{E}_\Phi^c] \le 4\exp(-\varepsilon m)$. Summing the failure probabilities establishes the lemma. ∎

D. Proof of Theorem 1

Proof: We first prove our result under the assumption that the matrix A is wide ($\delta \ge 1$) and the error e is not too sparse ($\rho \ge 15/16$). We then show that this more difficult case implies the result for arbitrary δ, ρ.

c) Underdetermined A, dense errors.: Fix any $\delta \ge 1$ and $\rho \ge 15/16$. Fix any $\nu \le \nu_{\delta,\rho}^\star$, and set $\alpha^\star, \varepsilon^\star, m^\star$ as specified in Remark 2. Set $k_1 = \lfloor \alpha^\star m \rfloor$ and choose $C_s(\delta, \rho) > 0$ according to (8) and set $s = \lfloor C_s m \rfloor$. This choice of constants is consistent with assumptions and conclusions of Lemmas 3 and 4.

For each subset $I \in \binom{[n]}{k_1}$, let $\mathcal{E}_I$ be the event that the two bounds in Lemma 3 and 4 hold simultaneously. The two lemmas imply that

$$\forall I \in \binom{[n]}{\alpha^\star m}, \quad \mathbb{P}[\mathcal{E}_I^c] \le 31\exp(-2\varepsilon^\star m). \quad (53)$$

Notice that on $\mathcal{E}_I$,

$$\|\boldsymbol{q}_0\| + \frac{\|\mathcal{S}\boldsymbol{q}_0\|}{1-\xi_s(\Phi)} < 312\nu\sqrt{\delta m}\left(1 + \frac{512\tau}{\nu^4\delta(1-\rho)}\right) + \frac{48}{\sqrt{1-\rho}}.$$

Below, we show that for any $\nu < \nu^\star$,

$$\text{(i).} \qquad \frac{512\tau}{\nu^4\delta(1-\rho)} \le 1,$$
$$\text{(ii).} \qquad 2 \cdot 312 \cdot \nu\sqrt{\delta m} \le \tfrac{1}{4}\sqrt{C_s m},$$
$$\text{(iii).} \qquad \frac{48}{\sqrt{1-\rho}} < \tfrac{1}{4}\sqrt{C_s m}.$$

Hence, on $\mathcal{E}_I$,

$$\|\boldsymbol{q}_0\| + \frac{\|\mathcal{S}\boldsymbol{q}_0\|}{1-\xi_s(\Phi)} < \sqrt{C_s m}/2 = \sqrt{s}/2, \quad (54)$$

and so, by Lemmas 1 and 2, on $\mathcal{E}_I$, for any $\boldsymbol{x}_0$ with $\text{supp}(\boldsymbol{x}_0) \subseteq I$ and any $\boldsymbol{e}_0$ with $\text{supp}(\boldsymbol{e}_0) = J$ and $\text{sign}(\boldsymbol{e}_0) = \boldsymbol{\sigma}$, $(\boldsymbol{x}_0, \boldsymbol{e}_0)$ is the unique solution to

$$\min \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subject to} \quad A\boldsymbol{x} + \boldsymbol{e} = A\boldsymbol{x}_0 + \boldsymbol{e}_0.$$

Now, further notice that by our choice of $\alpha^\star$ in (9),

$$\alpha^\star \leq \Upsilon[\varepsilon^\star/\delta] \leq \delta\Upsilon[\varepsilon^\star/\delta].$$

Hence, by (6)

$$\alpha^\star \log(\delta/\alpha^\star) \leq \varepsilon^\star,$$

and

$$\begin{aligned}
\mathbb{P}[\cap_I \mathcal{E}_I] &\geq 1 - \sum_I \mathbb{P}[\mathcal{E}_I^c], \\
&\geq 1 - \#\binom{n}{k_1} \times 31\exp(-2\varepsilon^\star m), \\
&\geq 1 - 31\exp\left(\alpha^\star \log(\delta/\alpha^\star)\, m - 2\varepsilon^\star m\right), \\
&\geq 1 - 31\exp(-\varepsilon^\star m),
\end{aligned}$$

Hence, to establish the result for $\delta \geq 1$ and $\rho \in (15/16, 1)$, it just remains to prove (i)-(iii).

(i). We establish (i) by proving two intermediate bounds (55) and (56) below hold whenever $\nu \leq \nu^\star$.

- $$\sqrt{\tau} \leq \nu^4. \tag{55}$$

To show (55), it is enough to argue that $\nu^{-1}\tau^{1/8} \leq 1$. Now,

$$\tau^{1/8} = \exp(-8^{-1}48^{-2}\delta^{-1}\nu^{-2}).$$

By (9), for $\nu < \nu^\star$,

$$\nu \leq 2^{-12}\sqrt{C_s/\delta} < 2^{-12}\delta^{-1} < 8^{-1}48^{-2}\delta^{-1},$$

where in the second inequality, we have used that $\Upsilon[t] \leq t$, and so by (9), $C_s \leq 2^{-17}C_\mu^{-2}\delta^{-2}(1-\rho) \ll \delta^{-1}$. So $\tau^{1/8} \leq \exp(-\nu^{-1})$ and

$$\nu^{-1}\tau^{1/8} \leq \nu^{-1}\exp\left(-\nu^{-1}\right) \doteq g(\nu).$$

Since

$$dg/d\nu = \nu^{-2}\exp(-\nu^{-1})(\nu^{-1} - 1),$$

g is strictly increasing for $\nu < 1$. Hence, it is enough to show that $g(\tilde{\nu}) \leq 1$ for some $\tilde{\nu}$ satisfying $\nu^\star \leq \tilde{\nu} < 1$. Notice that $\nu^\star \leq 2^{-12}$. It is easy to verify that $g(2^{-12}) = 2^{12}\exp(-2^{12}) \ll 1$, which establishes (55).

- $$\sqrt{\tau} \leq \frac{\delta(1-\rho)}{512}. \tag{56}$$

To show this, notice that

$$\begin{aligned}
\sqrt{\tau} &\leq \exp(-2^{-13}\delta^{-1}\nu^{-2}) \\
&\leq \exp(-2^{15}/C_s) \leq \exp(-2^{32}/(1-\rho)),
\end{aligned}$$

where we have used (9) and the fact that $C_s \leq 2^{-17}(1-\rho)$. Consider

$$h(x) \doteq 512x\exp(-2^{32}x).$$

Since $dh/dx = (512 - 2^{32}x)\exp(-2^{32}x)$, h is decreasing for $x > 512/2^{32}$, and for all $x > 1$, $h(x) < h(1) \ll 1$. Notice that for any $\rho > 0$,

$$\frac{512}{1-\rho}\sqrt{\tau} \leq h\left(\frac{1}{1-\rho}\right) < h(1) \ll 1 < \sqrt{\delta},$$

establishing (56).

(ii). For $\nu \leq \nu^\star$, by (9),

$$2 \cdot 312 \cdot \nu\sqrt{\delta m} \leq 624 \cdot 2^{-12}\sqrt{C_s m} \leq \tfrac{1}{4}\sqrt{C_s m}. \tag{57}$$

(iii). For $m \geq m^\star$, by (9),

$$\tfrac{1}{4}\sqrt{C_s m} \geq \tfrac{1}{4}\sqrt{\frac{2^{17}}{1-\rho}} \gg \frac{48}{\sqrt{1-\rho}}.$$

This establishes the theorem for $\delta \geq 1$, $\rho \in (15/16, 1)$.

d) Extension to arbitrary $\delta > 0$, $\rho \in (0,1)$: Now, fix any $\delta > 0$. Replace δ in the above choice of constants with $\max(\delta, 1)$. Generate

$$\tilde{A} = \boldsymbol{\mu}\mathbf{1}^* + \nu\tilde{Z} \in \mathbb{R}^{m\times m}$$

according to the bouquet model. Then with overwhelming probability in $\tilde{Z}, J, \boldsymbol{\sigma}$ for any $\boldsymbol{e}_0$ with signs and support $J, \boldsymbol{\sigma}$ and any sufficiently sparse $\tilde{\boldsymbol{x}}_0$,

$$(\tilde{\boldsymbol{x}}_0, \boldsymbol{e}_0) = \arg\min \|\tilde{\boldsymbol{x}}\|_1 + \|\boldsymbol{e}\|_1 \text{ s.t. } \tilde{A}\tilde{\boldsymbol{x}} + \boldsymbol{e} = \tilde{A}\tilde{\boldsymbol{x}}_0 + \boldsymbol{e}_0. \tag{58}$$

and the minimizer is uniquely defined. Form $A \in \mathbb{R}^{m\times\delta m}$ by selecting the first δm columns of $\tilde{A}$. Clearly, A is distributed according to the bouquet model. Lemma 12 shows that for any $\boldsymbol{x}_0 \in \mathbb{R}^{\delta m}$, if $[\boldsymbol{x}_0^*\ 0\ \dots\ 0]^*$, $\boldsymbol{e}_0$ exactly recovered by (58), $(\boldsymbol{x}_0, \boldsymbol{e}_0)$ is the unique optimal solution to

$$\min \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \text{ s.t. } \tilde{A}\boldsymbol{x} + \boldsymbol{e} = \tilde{A}\tilde{\boldsymbol{x}}_0 + \boldsymbol{e}_0,$$

and so with at least the same large probability, ℓ^1-minimization in the reduced system $A\boldsymbol{x} + \boldsymbol{e} = \boldsymbol{y}$ recovers all sparse $\boldsymbol{x}_0$ from a random error $\boldsymbol{e}_0$.

Finally, a straightforward argument given in Lemma 13 shows if ℓ^1-minimization uniquely recovers $(\boldsymbol{x}_0, \tilde{\boldsymbol{e}}_0)$, then for any $\boldsymbol{e}_0$ with $J \doteq \text{supp}(\boldsymbol{e}_0) \subseteq \text{supp}(\tilde{\boldsymbol{e}}_0)$ and $\text{sign}(\boldsymbol{e}_0(J)) = \text{sign}(\tilde{\boldsymbol{e}}_0(J))$, then ℓ^1-minimization uniquely recovers $(\boldsymbol{x}_0, \boldsymbol{e}_0)$. Fix any $\rho \in (0,1)$. Choose constants as above, but with ρ replaced by $\max\{\rho, 15/16\}$. For $\rho < 15/16$, generate the signs and support of $\boldsymbol{e}_0$ as follows: generate a random subset $\tilde{J} \subset [m]$ of size $15m/16$, and a random sign pattern $\tilde{\boldsymbol{\sigma}}$ with $\tilde{\boldsymbol{\sigma}}(\tilde{J}^c) = 0$, and $\tilde{\boldsymbol{\sigma}}(\tilde{J})$ iid Rademacher. Then with overwhelming probability, ℓ^1 minimization uniquely recovers $(\boldsymbol{x}_0, \tilde{\boldsymbol{e}}_0)$ for all sufficiently sparse $\boldsymbol{x}_0$. Generate J as a random subset of J' of size ρm, and let $\boldsymbol{\sigma} = \tilde{\boldsymbol{\sigma}}(J)$. By Lemma 13, ℓ^1-minimization, with the same overwhelmingly large probability ℓ^1-minimization also uniquely recovers $(\boldsymbol{x}_0, \boldsymbol{e}_0)$ for all sufficiently sparse $\boldsymbol{x}_0$. ∎

IV. SIMULATIONS AND EXPERIMENTS

In this section, we perform simulations verifying the conclusions of Theorem 1, and investigating the effect of various model parameters on the error correction capability of the ℓ^1-minimization (2). In the simulations below we use the publicly available ℓ^1-magic package [31], except for one (higher-dimensional) face recognition example, which requires a customized interior point method. Since ℓ^1-recoverability depends only on the signs and support of $(\boldsymbol{x}_0, \boldsymbol{e}_0)$, in the

simulations below we choose $\boldsymbol{x}_0(i) \in \{0,1\}$ and $\boldsymbol{e}_0(i) \in \{-1,0,1\}$. We will judge an output $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{e}})$ to be correct if $\max(\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_\infty, \|\hat{\boldsymbol{e}} - \boldsymbol{e}_0\|_\infty) < 0.01$.

a) Comparison with alternative approaches: We first compare the performance of the extended ℓ^1-minimization

$$\min \ \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subject to} \quad \boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$$

to two alternative approaches. The first is the error correction approach of [14], which multiplies by a full rank matrix B such that $BA = 0$,[9] solves

$$\min \ \|\boldsymbol{e}\|_1 \quad \text{subject to} \quad B\boldsymbol{e} = B\boldsymbol{y},$$

and then subsequently recovers $\boldsymbol{x}$ from the clean system of equations $A\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{e}$. The second is the Regularized Orthogonal Matching Pursuit (ROMP) algorithm [32], a state-of-the-art greedy method for recovering sparse signals.[10] For this algorithm, we use the implementation from http://math.ucdavis.edu/~dneedell/.
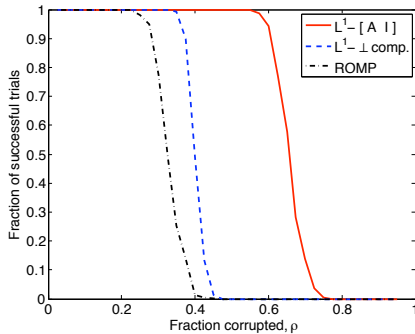

Fig. 3. **Comparison with alternative approaches.** Here, we fix $m = 500$, $\delta = 0.25$, $\nu = 0.05$, and $k_1 = 15$, and compare three approaches to recovering the sparse signal $\boldsymbol{x}_0$ from error $\boldsymbol{e}_0$. The first, denoted "$L^1 - [A\ \text{I}]$" solves the extended ℓ^1 minimization advocated in this work. The second, denoted "$L^1 - \perp$ comp" premultiplies by the orthogonal complement of A, and then solves an underdetermined system of linear equations for the sparse error $\boldsymbol{e}$ [14]. The final approach is the greedy Regularized Orthogonal Matching Pursuit (ROMP) [32].

For this experiment, the ambient dimension is $m = 500$; the parameters of the CAB model are $\nu = 0.05$ and $\delta = 0.25$. We fix the signal support to be $k_1 = 15$, and vary the fraction of errors from 0 to 0.95. For each error fraction, we generate 500 independent problems. Figure 3 plots the fraction of successes for each of the three algorithms, as a function of error density ρ. There the extended ℓ^1-minimization is denoted "$L^1 - [A\ \text{I}]$" (red curve), while the alternative approach of [14] is denoted "$L^1 - \perp$ comp" (blue curve). Whereas both ROMP and the ℓ^1 approach of [14] break down around 40% corruption, the extended ℓ^1-minimization continues to succeed with high probability even beyond 60% corruption.

b) Error correction capacity: While the previous experiment demonstrates the advantages of the extended ℓ^1-minimization (2) for the CAB model, Theorem 1 suggests that more is true: If the support of $\boldsymbol{x}$ is a diminishing fraction of the

dimension, as the dimension increases, the fraction of errors that the extended ℓ^1-minimization can correct should approach one. We generate problem instances with $\delta = 0.25$, $\nu = 0.05$, for varying $m = 100, 200, 400, 800, 1600$. For each problem size, and for each error fraction $\rho = 0.05, 0.1, \ldots, 0.95$, we generate 500 random problems, and plot the fraction of correct recoveries in Figure 4. At left, we fix $k_1 = 1$, while at right, k_1 grows as $k_1 = m^{1/2}$. In both cases, as m increases, the fraction of errors that can be corrected also increases, with no sign of stopping anywhere before $\rho = 1$.
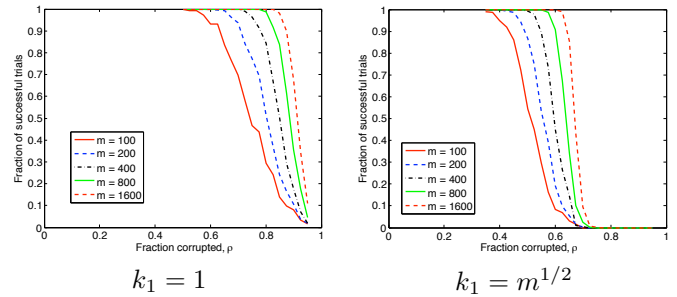

$$k_1 = 1 \qquad\qquad\qquad k_1 = m^{1/2}$$

Fig. 4. **Error correction for $\boldsymbol{x}$ with sublinear support.** We fix $\delta = 0.25$, $\nu = 0.05$, and plot the fraction of successful recoveries as a function of the error density ρ, for each $m = 100, 200, 400, 800, 1600$. At left, k_1 is fixed at 1; at right, $k_1 = m^{1/2}$. In both cases, as m increases, the fraction of errors that can be corrected approaches 1.

c) Varying model parameters: We next investigate the effect of varying δ (Figure 5 left) and ν (Figure 5 right). We first fix $m = 400$, $\nu = .3$, and consider different bouquet sizes $n = 100, 200, 300, 400, 500$. Figure 5 left plots the fraction of correct trials for varying error densities ρ, for each of these bouquet sizes. For this fixed m, the error correction capability decreases only slightly as n increases.
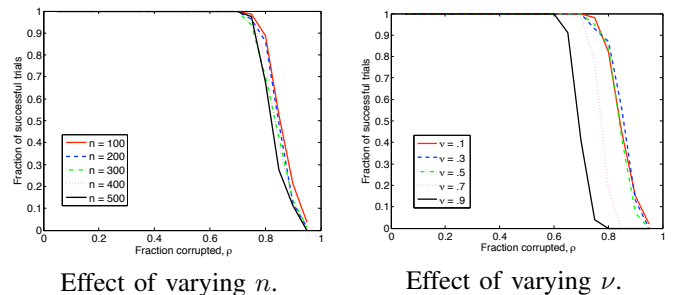

Effect of varying n.   Effect of varying ν.

Fig. 5. **Effect of varying n and ν.** At left, we fix $m = 400$, $\nu = .3$, and consider varying $n = 100, 200, \ldots, 500$. For each of these model settings, we plot the fraction of correct recoveries as a function of the fraction of errors. Notice that the error correction capacity decreases only slightly as n increases. At right, we fix $m = 400$, $n = 200$, and vary ν from .1 to .9. Again, we plot the fraction of correct recoveries for each error fraction. As expected from Theorem 1, as ν decreases, the error correction capacity of ℓ^1 increases.

We next fix $m = 400$, $n = 200$, and consider the effect of varying ν. Figure 5 plots the result for $\nu = .1, .3, .5, .7, .9$. Notice that as ν decreases (i.e., the bouquet becomes tighter), the error correction capacity increases: for any fixed fraction of successful trials, the fraction of error that can be corrected increases by approximately 15% as ν decreases from .9 to .5.

d) Phase transition in proportional growth: Theorem 1 claims that for any $\rho < 1$, the ℓ^1-minimization works

---

[9]This comparison requires $n \ll m$ although our method is not limited to this case.

[10]For the models considered here, less sophisticated greedy methods such as the standard orthogonal matching pursuit fail even for small error fractions.

even when the signal support k_1 grows proportionally to m: $k_1 \leq \alpha m$. However, it gives rather pessimistic estimate on how large the fraction can be. Based on intuition from more homogeneous polytopes (especially the work of Donoho and Tanner on Gaussian matrices [20]), we might expect that when k_1 exhibits proportional growth, an asymptotically sharp phase transition between guaranteed recovery and guaranteed failure will occur at some critical error fraction $\rho^* \in (0,1)$. We investigate this empirically here by again setting $\delta = 0.25$, $\nu = 0.05$, but this time allowing $k_1 = 0.05m$. Figure 6 plots the fraction of correct recovery for varying error fractions ρ, as m grows: $m = 100, 200, 400, 800, 1600$. In this proportional growth setting, we see an increasingly sharp phase transition, near $\rho = 0.6$.
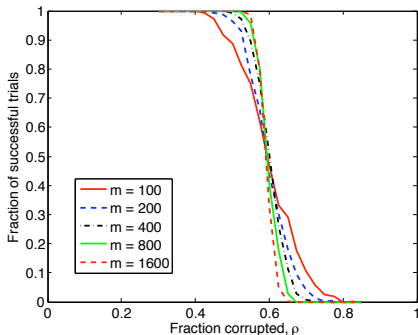

Fig. 6. **Phase transition in proportional growth.** When the signal support grows in proportion to the dimension ($k_1/m \to \alpha \in (0,1)$), we observe an asymptotically sharp phase transition in the probability of correct recovery, similar to that investigated in [20]. Here, for $\delta = 0.25$, $\nu = 0.05$, $k_1 = 0.05\,m$, we indeed see a sharp phase transition at $\rho = 0.6$.

e) Error correction with real face images: Finally, we return to the motivating example of face recognition under varying illumination and random corruption. For this experiment, we use the Extended Yale B face database [15], which tests illumination sensitivity of face recognition algorithms. As in [11], we form the matrix A from images in Subsets 1 and 2, which contain mild-to-moderate illumination variations. Each column of the matrix A is a $w \times h$ face image, stacked as a vector in $\mathbb{R}^m$ ($m = w \times h$). Here, the weak proportional growth setting corresponds to the case when the total number of image pixels grows proportionally to the number n of face images. Since the number of images per subject is fixed, this is the same as the total image resolution growing proportionally to the number of subjects. We vary the image resolutions through the range 34×30, 48×42, 68×60, 96×84.[11] The matrix A is formed from images of $4, 9, 19, 38$ subjects, respectively, corresponding to $\delta \approx 0.09$. Here, $\nu \approx 0.3$. In face recognition, the sublinear growth of $\|x_0\|_0$ comes from the fact that the observation should ideally be a linear combination of only images of the same subject. Various estimates of the required number of images, k_1, appear in the literature, ranging from 5 to 9. Here, we fix $k_1 = 7$, and generate the (clean) test image synthetically as a linear combination of k_1 training images from a single subject. The reason for using synthetic linear

[11]Thus, the total dimension $m = 1020, 2016, 4080, 8064$ grows roughly by a factor of 2 from one curve to the next, similar to the simulations above.

combinations as opposed to real test images is simply that it allows us to verify whether x_0 was correctly recovered; in the real data experiments of the introduction of this work and of [11], success could only be judged in terms of the recognition rate of the entire classification pipeline.

For each resolution considered, and for each error fraction, we generate 75 trials. Figure 7 (left) plots the fraction of successes as a function of the fraction of corruption. Notice that as predicted by Theorem 1, the fraction of errors that can be corrected again approaches 1 as the data size increases. Figure 7 (right) gives a visual demonstration of the algorithm's capability. In the test images in Figure 7 (right, top), the amount of corruption is chosen to correspond to a 50% probability of success according to the plots in Figure 7 (left). Below each corrupted test image, the "clean" image recovered by our method is shown.
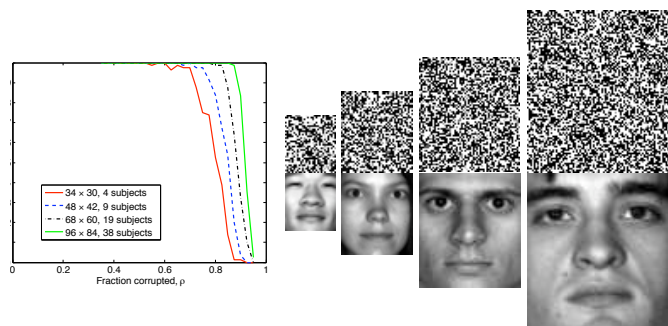

Fig. 7. **Error correction with real face images.** We simulate weak proportional growth in the Extended Yale B face database, with the resolution of the images growing in proportion to the number of subjects. Left: fraction of correct recoveries for varying levels of occlusion. Right: examples of correct recovery for each resolution considered. Top: corrupted test image. The fraction of corruption is chosen so that the probability of correct recovery is 50%. Bottom: clean image, from correctly recovered x_0.

V. DISCUSSIONS AND FUTURE WORK

a) Compressed sensing for signals with varying sparsity: In the conventional setting for recovering a sparse signal, one often implicitly assumes that each entry of the signal has an equal probability of being nonzero. As a result, one typically requires that the incoherence (or coherence) of the dictionary is somewhat uniform. In this paper, we saw quite a different example. If we view both x and e as the signal that we want to recover, then the sparsity or density of the combined signal is quite uneven – x is very sparse but e can be very dense. Nevertheless, our result suggests that if the incoherence of the dictionary is adaptive to the distribution of the density – more coherent for the sparse part and less for the dense part, then ℓ^1-minimization will be able to recover such uneven signals even if bounds based on the even sparsity assumption suggest otherwise. Thus, if one has some prior knowledge about which part of the signal is likely to be more sparse or more dense, one can achieve much better performance with ℓ^1-minimization by using a dictionary with matching incoherence. More generally, for any given distribution of sparsity, one may ask the question whether there exists an optimal dictionary

with matching incoherence such that ℓ^1-minimization has the highest chance of success.

b) Stability with respect to noise: Although in our model we do not explicitly consider any noise, say $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e} + \boldsymbol{z}$, where $\boldsymbol{z}$ adds small Gaussian noise to each entry, we have observed empirically in our simulations and also in experiments with face images that the ℓ^1-minimization (2) for the cross-and-bouquet model is surprisingly stable with respect to measurement or numerical noise. In fact, as the method is able to deal with dense errors regardless of their magnitude, large noisy entries in $\boldsymbol{z}$ might be treated like errors and be absorbed into $\boldsymbol{e}$. This is especially the case if the signal $\boldsymbol{x}$ is extremely sparse – in the case of face recognition, the support of $\boldsymbol{x}$ is nearly constant. Nevertheless, for less sparse signals $\boldsymbol{x}$, the noise tolerance can likely be further improved by relaxing the equality constraint in (2), in a similar spirit to the Lasso [4]–[6], [33] or Dantzig selector [34]:

$$\min_{\boldsymbol{x},\boldsymbol{e}} \|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subject to} \quad \|\boldsymbol{y} - A\boldsymbol{x} + \boldsymbol{e}\| \leq \varepsilon^2.$$

To our best knowledge, a more precise characterization of the effect of entry-wise small noise or perturbation $\boldsymbol{z}$ on the estimate of $\boldsymbol{x}$ and $\boldsymbol{e}$ through the above convex program remains an open problem.

c) Neighborliness of polytopes: As we have seen in this paper, a precise characterization of the performance of ℓ^1-minimization requires us to analyze the geometry of polytopes associated with the specific dictionaries in question. In practice, we often use ℓ^1-minimization for purposes other than signal reconstruction or error correction. For instance, using machine learning techniques, we can learn from exemplars a dictionary that is optimal for certain tasks such as data classification [13]. The polytope associated with such a dictionary may be very different from those that are normally studied in signal processing or coding theory or error correction, leading to qualitatively different behavior of the ℓ^1-minimization. Thus, we should expect that in the coming years, many new classes of high-dimensional polytopes with even more interesting properties may arise from other applications and practical problems.

d) Dense error correction for low-rank matrices: Recent studies of low-rank matrix completion have shown that it is possible to exactly complete a low-rank matrix A from knowing only a diminishing fraction of its entries [35]. Our recent work has shown that one can exactly recover a low-rank matrix A with grossly corrupted entries $M = A + E$ by solving a similar convex program:

$$\min \|A\|_* + \lambda\|E\|_1 \quad \text{subject to} \quad M = A + E,$$

as long as the matrix E is sufficiently sparse [36]. Moreover, empirically one can observe a similar *dense error correction* phenomenon: when the rank of the matrix A is sublinear in its dimension, as the dimension increases, the fraction of entries corrupted by random errors with random signs that the convex program can correct also approaches to 100%. As an extension to the analysis of [35], one can prove this is indeed the case if the signs of the errors in E are chosen at random [37]. It seems that the dense error correction property

discovered in this work may be an instance of a more general property of error correction problems involving structured high-dimensional signals and errors with random support and signs.

e) Modifications of the error model: In our model, we have assumed the support and signs of the error $\boldsymbol{e}_0$ are both random, which plays a crucial role in the proof of the main theorem. In many practical applications, the errors may not be completely random and have additional structures. For instance, the signs of the error can even be adversarial. To reduce the randomness in our error model, one can invoke similar derandomization arguments of [35] (Theorem 2.3) and show that Theorem 1 holds for any error $\boldsymbol{e}_0$ with arbitrary signs but with support strictly less than 50%. Then the only randomness left would be the support of the errors. In many applications, we may also have certain prior knowledge about the support of the errors. For example, in image processing, corruption incurred by occlusion normally has a spatially contiguous support. Recent empirical studies in face recognition [38] have demonstrated that if such prior knowledge can be properly harnessed, one can significantly boost the error correction capability of the ℓ^1-minimization. However, a rigorous mathematical justification for such success remains an open problem.

APPENDIX
TECHNICAL LEMMAS

This appendix summarizes a few results from measure concentration and proves several lemmas used in the main text.

OBSERVATIONS ON SUBMATRIX SINGULAR VALUES

Lemma 6: For any block matrix $[M\ N]$, and scalar t with $|t| \leq 1$,

$$\gamma_s([M \mid tN]) \geq |t|\gamma_s([M\ N]). \tag{59}$$

Proof:

$$\gamma_s\left([M \ tN]\right) = \inf_{\substack{\|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2\geq 1 \\ \|\boldsymbol{x}\|_0+\|\boldsymbol{y}\|_0\leq s}} \left\|[M \ tN]\begin{bmatrix}\boldsymbol{x}\\\boldsymbol{y}\end{bmatrix}\right\|$$

$$= \inf_{\substack{\|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2\geq 1, \\ \|\boldsymbol{x}\|_0+\|\boldsymbol{y}\|_0\leq s.}} \left\|[M \ N]\begin{bmatrix}\boldsymbol{x}\\t\boldsymbol{y}\end{bmatrix}\right\|$$

$$= \inf_{\substack{\|t\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2\geq 1, \\ \|\boldsymbol{x}\|_0+\|\boldsymbol{y}\|_0\leq s.}} \left\|[M \ N]\begin{bmatrix}t\boldsymbol{x}\\t\boldsymbol{y}\end{bmatrix}\right\|$$

$$\geq |t| \inf_{\substack{\|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2\geq 1, \\ \|\boldsymbol{x}\|_0+\|\boldsymbol{y}\|_0\leq s.}} \left\|[M \ N]\begin{bmatrix}\boldsymbol{x}\\\boldsymbol{y}\end{bmatrix}\right\| = |t|\gamma_s([M \ N]).$$

∎

Lemma 7: For any matrices M, N,

$$\gamma_s(M+N) \geq \gamma_s(M) - \beta_s(N), \qquad (60)$$

$$\text{and} \quad \beta_s(M+N) \leq \beta_s(M) + \beta_s(N). \qquad (61)$$

Proof: Notice that

$$\gamma_s(M+N) = \inf_{\|\boldsymbol{x}\|=1,\|\boldsymbol{x}\|_0\leq s} \|M\boldsymbol{x}+N\boldsymbol{x}\|$$

$$\geq \inf_{\|\boldsymbol{x}\|=1,\|\boldsymbol{x}\|_0\leq s} (\|M\boldsymbol{x}\| - \|N\boldsymbol{x}\|)$$

$$\geq \inf_{\|\boldsymbol{x}\|=1,\|\boldsymbol{x}\|_0\leq s} \|M\boldsymbol{x}\| - \sup_{\|\boldsymbol{y}\|=1,\|\boldsymbol{y}\|_0\leq s} \|N\boldsymbol{y}\|$$

$$= \gamma_s(M) - \beta_s(N).$$

The calculation for (61) is identical. ∎

Lemma 8: Let $W \in \mathbb{R}^{m\times n}$, $(m < n)$ be any matrix with full row rank m, and let $P_W \in \mathbb{R}^{n\times n}$ denote projection operator onto the range of W^*. Then

$$\beta_s(P_W) \leq \beta_s(W)/\sigma_{min}(W). \qquad (62)$$

Proof:

$$\beta_s(P_W) = \sup_{\|\boldsymbol{x}\|\leq 1,\|\boldsymbol{x}\|_0\leq s} \|W^*(WW^*)^{-1}W\boldsymbol{x}\|$$

$$\leq \|W^*(WW^*)^{-1}\| \sup_{\substack{\|\boldsymbol{x}\|\leq 1, \\ \|\boldsymbol{x}\|_0\leq s}} \|W\boldsymbol{x}\| = \sigma_{min}^{-1}(W)\,\beta_s(W).$$

∎

Lemma 9: For any matrix M

$$\gamma_s^2\left([\mathtt{I} \ M]\right) \geq \min\{1,\gamma_s^2(M)\} - \phi_s(M). \qquad (63)$$

Proof:

$$\gamma_s^2\left([\mathtt{I} \ M]\right) = \inf_{\substack{\|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2=1 \\ \|\boldsymbol{x}\|_0+\|\boldsymbol{y}\|_0\leq s}} \|\boldsymbol{x}+M\boldsymbol{y}\|^2$$

$$= \inf_{\boldsymbol{x},\boldsymbol{y}} \|\boldsymbol{x}\|^2 + \|M\boldsymbol{y}\|^2 + 2\boldsymbol{x}^*M\boldsymbol{y}$$

$$\geq \inf_{\boldsymbol{x},\boldsymbol{y}} \|\boldsymbol{x}\|^2 + \gamma_s^2(M)\|\boldsymbol{y}\|^2 - 2\|\boldsymbol{x}\|\|\boldsymbol{y}\|\phi_s(M)$$

$$\geq \inf_{\boldsymbol{x},\boldsymbol{y}} \min\{1,\gamma_s^2(M)\}(\|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2) - \phi_s(M)$$

$$= \min\{1,\gamma_s^2(M)\} - \phi_s(M),$$

where we have used that for $t \in [0,1]$, $\sqrt{t(1-t)} \leq 1/2$. ∎

TWO CONCENTRATION RESULTS

Lemma 10: Suppose $\boldsymbol{\mu} \in \mathbb{R}^m$ $\|\boldsymbol{\mu}\|_\infty \leq C_\mu m^{-1/2}$, $\|\boldsymbol{\mu}\| = 1$, and let $J \subset [m]$ be uniformly distributed amongst all subsets of size ρm. Then for any $\varepsilon > 0$, with probability at least $1 - \exp(-\varepsilon m)$ in J,

$$\|\boldsymbol{\mu}_{J^c}\|^2 \geq 1 - \rho - 4C_\mu^2\sqrt{2\varepsilon}. \qquad (64)$$

In particular, if $\delta \geq 1$ then with probability at least $1 - \exp(-2\varepsilon^\star_{\delta,\rho,\nu}m)$,

$$\|\boldsymbol{\mu}_{J^c}\|^2 \geq (1-\rho)/4, \qquad (65)$$

where $\varepsilon^\star_{\delta,\rho,\nu}$ is specified in (9).

Proof: Consider the group of permutations of $[m]$ with distance $d(\pi_1,\pi_2) = m^{-1}\#\{i \mid \pi_1(i) \neq \pi_2(i)\}$. Let π be a random permutation sampled according to the uniform measure. Then random variable $\|\boldsymbol{\mu}_{J^c}\|^2$ is equal in distribution to $f(\pi) = \sum_{i=1}^{(1-\rho)m} \boldsymbol{\mu}(\pi(i))^2$. Since $\|\boldsymbol{\mu}\|_\infty < C_\mu m^{-1/2}$, $|f(\pi_1) - f(\pi_2)| \leq 2C_\mu^2 m^{-1}\#\{\pi_1(i) \neq \pi_2(i)\} = 2C_\mu^2 d(\pi_1,\pi_2)$. Corollary 4.3 of [3] shows that

$$\mathbb{P}\left[f(\pi) \leq \mathbb{E}f(\pi) - t\right] < \exp(-mt^2/32C_\mu^4).$$

Noticing that $\mathbb{E}f = 1 - \rho$ establishes (64).

For (65), notice that by (9),

$$\sqrt{\varepsilon^\star} \leq \nu\sqrt{1-\rho} \leq 2^{-12}\sqrt{C_s(1-\rho)/\delta}$$

$$\leq 2^{-12}\sqrt{2^{-17}C_\mu^{-4}\delta^{-3}(1-\rho)^2}$$

$$\leq 2^{-20}C_\mu^{-2}(1-\rho),$$

where in the last step we have used that $\delta \geq 1$. Hence, $4C_\mu^2\sqrt{2\times 2\varepsilon^\star} \leq 2^{-17}(1-\rho) \ll 3(1-\rho)/4$, and by (64) with probability at least $1 - \exp(-2\varepsilon^\star m)$, $\|\boldsymbol{\mu}_{J^c}\|^2 \geq (1-\rho)/4$. ∎

Lemma 11: Let $\varepsilon^\star_{\delta,\rho,\nu}$ be defined as in (9). With probability at least $1 - 3\exp(-2\varepsilon^\star_{\delta,\rho,\nu}m)$, $\boldsymbol{\mu}_{J^c} \neq \boldsymbol{0}$ and the matrix W defined by

$$W^* = \left[\begin{array}{cc} \frac{\boldsymbol{\mu}_{J^c}}{\|\boldsymbol{\mu}_{J^c}\|} & \frac{Z_{J^c,I}}{\sqrt{1-\rho}} \end{array}\right] \in \mathbb{R}^{(m-k_2)\times(k_1+1)}$$

satisfies

$$\sigma_{min}(W) \geq 1/4. \qquad (66)$$

Proof: For compactness, let

$$\boldsymbol{\vartheta} = \begin{cases} \boldsymbol{\mu}_{J^c}/\|\boldsymbol{\mu}_{J^c}\|, & \boldsymbol{\mu}_{J^c} \neq \boldsymbol{0}, \\ \boldsymbol{0}, & \text{else}, \end{cases}$$

and let $P_\vartheta = \boldsymbol{\vartheta}\boldsymbol{\vartheta}^*$ be the matrix that projects onto the span of $\boldsymbol{\vartheta}$. Let $Q \in \mathbb{R}^{(m-k_2)\times(k_1+1)}$ be defined by

$$Q \doteq \left[\begin{array}{cc} \boldsymbol{\vartheta} & \frac{Z_{J^c,I}}{\sqrt{1-\rho}} \end{array}\right].$$

Let $\mathcal{E}_1$ denote the event

$$\mathcal{E}_1 = \{\boldsymbol{\mu}_{J^c} = \boldsymbol{0}\},$$

and notice that on $\mathcal{E}_1^c$, W is well defined and $\sigma_{min}(W) = \sigma_{k+1}(Q)$. Furthermore,

$$Q = \left[\begin{array}{cc} \boldsymbol{\vartheta} & (\mathtt{I}-P_\vartheta)\frac{Z_{J^c,I}}{\sqrt{1-\rho}} \end{array}\right] + \left[\begin{array}{cc} \boldsymbol{0} & P_\vartheta\frac{Z_{J^c,I}}{\sqrt{1-\rho}} \end{array}\right],$$

and so

$$\sigma_{k_1+1}(Q) \geq \sigma_{k_1+1}\left[\boldsymbol{\vartheta} \quad (\mathtt{I} - P_{\boldsymbol{\vartheta}})\frac{Z_{J^c,I}}{\sqrt{1-\rho}}\right] - \left\|P_{\boldsymbol{\vartheta}}\frac{Z_{J^c,I}}{\sqrt{1-\rho}}\right\|.$$

For any matrix $M = [A \; B] \in \mathbb{R}^{p \times (q_1+q_2)}$ with $A \in \mathbb{R}^{p \times q_1}$ and $B \in \mathbb{R}^{p \times q_2}$ and $A^*B = 0$, $\sigma_{q_1+q_2}(M) = \min\{\sigma_{q_1}(A), \sigma_{q_2}(B)\}$. Since $\|P_{\boldsymbol{\vartheta}} Z_{J^c,I}\| = \|\boldsymbol{\vartheta}^* Z_{J^c,I}\|$,

$$\sigma_{min}(Q) \geq \min\left\{\|\boldsymbol{\vartheta}\|, \frac{\sigma_{k_1}[(\mathtt{I}-P_{\boldsymbol{\vartheta}})Z_{J^c,I}]}{\sqrt{1-\rho}}\right\} - \frac{\|\boldsymbol{\vartheta}^* Z_{J^c,I}\|}{\sqrt{1-\rho}}$$
$$\geq \min\left\{\|\boldsymbol{\vartheta}\|, \frac{\sigma_{k_1}[Z_{J^c,I}]}{\sqrt{1-\rho}}\right\} - 2\frac{\|\boldsymbol{\vartheta}^* Z_{J^c,I}\|}{\sqrt{1-\rho}}.$$

Further define the events

$$\mathcal{E}_2 = \left\{\|\boldsymbol{\vartheta}^* Z_{J^c,I}\| > \sqrt{\alpha^\star} + \sqrt{4\varepsilon^\star}\right\},$$
$$\mathcal{E}_3 = \left\{\sigma_{k_1}[(\mathtt{I} - P_{\boldsymbol{\vartheta}})Z_{J^c,I}] < \sqrt{1-\rho-\tfrac{1}{m}} - \sqrt{\alpha^\star} - \sqrt{4\varepsilon^\star}\right\}.$$

On $\cap_{i=1}^3 \mathcal{E}_i^c$, $\boldsymbol{\mu}_{J^c} \neq \mathbf{0}$, $\|\boldsymbol{\vartheta}\| = 1$, W is well-defined, $\sigma_{min}(W) = \sigma_{k_1+1}(Q)$, and

$$\sigma_{k_1+1}(Q) \geq \sqrt{1 - \tfrac{1}{(1-\rho)m^\star}} - 3\sqrt{\alpha^\star} - 6\sqrt{\varepsilon^\star}.$$

From (9), $m^\star = 2^{17}(1-\rho)^{-1}C_s^{-1} > 2^{17}(1-\rho)^{-1}$ (where we have used that $C_s < 1$. Moreover, from (9), $\alpha^\star \leq \nu^2 < 2^{-24}$, and similarly, $\varepsilon^\star \leq \nu^2 < 2^{-24}$. Plugging in to the above, on $\cap_{i=1}^3 \mathcal{E}_i$,

$$\sigma_{k_1+1}(Q) \geq \sqrt{1 - 2^{-17}} - 9 \cdot 2^{-12} > 1/4.$$

It remains to be shown that $\mathbb{P}[\cap_{i=1}^3 \mathcal{E}_i^c]$ is large. By Lemma 10, with probability at least $1 - \exp(-2\varepsilon^\star m)$,

$$\|\boldsymbol{\mu}_{J^c}\|^2 \geq (1-\rho)/4 > 0,$$

and so $\mathbb{P}[\mathcal{E}_1] \leq \exp(-2\varepsilon^\star m)$. Now, on $\mathcal{E}_1$, $\|\boldsymbol{\vartheta}^* Z_{J^c,I}\| = 0$. On $\mathcal{E}_1^c$, for any fixed J (and hence any fixed $\boldsymbol{\vartheta}$, the conditional joint distribution of the k_1-dimensional vector $\boldsymbol{\vartheta}^* Z_{J^c,I}$ is simply iid $\mathcal{N}(0, 1/m)$. Hence, by (24), for any $J_0 \in \mathcal{E}_1^c$,

$$\mathbb{P}[\|\boldsymbol{\vartheta}^* Z_{J^c,I}\| > \sqrt{\alpha^\star} + \sqrt{4\varepsilon^\star}] \leq \exp(-2\varepsilon^\star m), \qquad (67)$$

and

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_2] &= \mathbb{P}[\mathcal{E}_2 \mid \mathcal{E}_1]\mathbb{P}[\mathcal{E}_1] + \mathbb{P}[\mathcal{E}_2 \mid \mathcal{E}_1^c]\mathbb{P}[\mathcal{E}_1^c] \\
&\leq 0 + \mathbb{P}[\mathcal{E}_2 \mid \mathcal{E}_1^c] \\
&\leq \sup_{J_0 \in \mathcal{E}_1^c} \mathbb{P}\left[\|\boldsymbol{\vartheta}^* Z_{J^c,I}\| > \sqrt{\alpha^\star} + \sqrt{4\varepsilon^\star} \mid J_0\right] \\
&\leq \exp(-2\varepsilon^\star m).
\end{aligned}$$

Similarly, on $\mathcal{E}_1$, $(\mathtt{I} - P_{\boldsymbol{\vartheta}})Z_{J^c,I} = Z_{J^c,I}$ is simply a $(1-\rho)m \times k_1$ matrix, and so (23) gives that $\forall J_0 \in \mathcal{E}_1$,

$$\mathbb{P}\left[\sigma_{min}(Z_{J^c,I}) < \sqrt{1-\rho} - \sqrt{\alpha^\star} - \sqrt{4\varepsilon^\star} \mid J_0\right]$$
$$\leq \exp(-2\varepsilon^\star m).$$

Similarly, on $\mathcal{E}_1^c$, $(\mathtt{I} - P_{\boldsymbol{\vartheta}})$ projects the columns of $Z_{J^c,I}$ onto a subspace of dimension $(1-\rho)m - 1$. From the rotational invariance of the Gaussian distribution, for any fixed $J_0 \in \mathcal{E}_1^c$, the first $(1-\rho)m - 1$ singular values of $(\mathtt{I} - P_{\boldsymbol{\vartheta}})Z_{J^c,I}$ are equal in distribution to the first $(1-\rho)m-1$ singular values of

an iid Gaussian matrix of size $(1-\rho)m - 1 \times k_1$. In particular, for any fixed $J_0 \in \mathcal{E}_1^c$, by (23)

$$\mathbb{P}\left[\sigma_{min}[(I - P_{\boldsymbol{\vartheta}})Z_{J^c,I}] < \sqrt{1-\rho-\tfrac{1}{m}} - \sqrt{\alpha^\star} - \sqrt{4\varepsilon^\star} \mid J_0\right]$$
$$\leq \exp(-2\varepsilon^\star m).$$

Hence,

$$\mathbb{P}[\mathcal{E}_3] \leq \sup_{J_0} \mathbb{P}[\mathcal{E}_3 \mid J = J_0] \leq \exp(-2\varepsilon^\star m).$$

Summing, we find that

$$\mathbb{P}[\cap_{i=1}^3 \mathcal{E}_i^c] \geq 1 - \sum_{i=1}^3 \mathbb{P}[\mathcal{E}_i] \geq 1 - 3\exp(-2\varepsilon^\star m),$$

establishing the result. ∎

PROOF OF LEMMA 5

Proof: The lemma is an application of a concentration result on the cube due to Talagrand [39]. We use the version stated in Corollary 4.10 and Equation (4.10) of [3]. The result states that if $\boldsymbol{\sigma} \in \mathbb{R}^n$ is distributed according to any product measure on $[-1,1]^n$, $f : \mathbb{R}^n \to \mathbb{R}$ is a convex, $\|f\|_{lip}$-Lipschitz function, and m_f is any median of f, then

$$\mathbb{P}[|f(\boldsymbol{\sigma}) - m_f| \geq t] \leq 4\exp\left(-t^2/16\|f\|_{lip}^2\right). \qquad (68)$$

To show (44), notice that

$$\begin{aligned}
\mathbb{E}[\|M\boldsymbol{\sigma}\|^2] &= \mathbb{E}\left[\sum_i \left(\sum_j M_{ij}\sigma_j\right)^2\right] \\
&= \sum_i \mathbb{E}\left[\sum_{j,k} M_{ij}M_{ik}\sigma_j\sigma_k\right] = \sum_i \sum_j M_{ij}^2 = \|M\|_F^2.
\end{aligned}$$

So, $\mathbb{E}\|M\boldsymbol{\sigma}\| \leq \sqrt{\mathbb{E}[\|M\boldsymbol{\sigma}\|^2]} \doteq \|M\|_F$. Since $f(\cdot) \doteq \|M \cdot \|$ is nonnegative, by the Markov inequality, f has a median no larger than $2\mathbb{E}f \leq 2\|M\|_F$. Since f is $\|M\|$-Lipschitz, plugging into (68) completes the proof of (44).

For (45), let $g : \mathbb{R} \to \mathbb{R}$ with $g(t) = |\mathcal{S}(t)| = \max(t - 1/2, -t - 1/2, 0)$. As the pointwise maximum of three linear functions, g is convex. Let $h : \mathbb{R}^n \to \mathbb{R}$ as $h(x_1, \ldots, x_n) = \|(g(x_1), \ldots, g(x_n))\|$. Notice that g is nonnegative, and futher that for nonnegative $\boldsymbol{z}$, $\|\boldsymbol{z}\|$ is nondecreasing in each z_i. Hence, $h(\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}) = \|(g(\alpha x_1+(1-\alpha)y_1), \ldots, g(\alpha x_2 + (1 - \alpha)y_n))\| \leq \|\alpha(g(x_1), \ldots, g(x_n)) + (1-\alpha)(g(y_1), \ldots, g(y_n))\| \leq \alpha h(\boldsymbol{x}) + (1-\alpha)h(\boldsymbol{y})$. Thus h is a convex function. $f(\boldsymbol{\sigma}) \doteq \|\mathcal{S}[M\boldsymbol{\sigma}]\| = h(M\boldsymbol{\sigma})$ is the composition of a convex function with a linear function and hence is also convex.

Moreover, since for any scalars a, b, $|\mathcal{S}(a) - \mathcal{S}(b)| \leq |a-b|$, for vectors $\boldsymbol{x}, \boldsymbol{y}$, $\|\mathcal{S}(\boldsymbol{x}) - \mathcal{S}(\boldsymbol{y})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$, and so g is 1-Lipschitz. As the composition of a 1-Lipschitz function with an $\|M\|$-Lipschitz function, f is $\|M\|$-Lipschitz. Hence, (68) applies, (again noticing that there is a median no larger that $2\mathbb{E}f$), giving

$$\mathbb{P}\left[\|\mathcal{S}[M\boldsymbol{\sigma}]\| > 2\mathbb{E}\|\mathcal{S}[M\boldsymbol{\sigma}]\| + 4\|M\|\sqrt{\varepsilon n}\right] < 4\exp(-\varepsilon n). \qquad (69)$$

It remains to estimate the expectation. Notice that $(M\boldsymbol{\sigma})_i = M_i\boldsymbol{\sigma} = \sum_j M_{ij}\sigma_j$ is a sum of independent, bounded random variables. Applying Hoeffding's inequality,

$$\mathbb{P}[|(M\boldsymbol{\sigma})_i| > t] < 2\exp\left(-2t^2/\|M_i\|^2\right). \qquad (70)$$

To simplify notation, let X be an arbitrary random variable satisfying a normal tail bound $\mathbb{P}[|X| > t] < C_1\exp(-C_2 t^2)$. Then

$$\mathbb{E}[\mathcal{S}[X]^2] = 2\int_{t=0}^{\infty} t\mathbb{P}[|\mathcal{S}[X]| > t]\, dt$$
$$= 2\int_{t=1/2}^{\infty} (t - 1/2)\mathbb{P}[|X| > t]\, dt$$
$$\leq C_1\int_{t=1/2}^{\infty} 2t\exp(-C_2 t^2)\, dt = \frac{C_1}{C_2}\exp(-C_2/4).$$

Hence,

$$\mathbb{E}[\|\mathcal{S}[M\boldsymbol{\sigma}]\|^2] = \sum_i \mathbb{E}[\mathcal{S}[M_i\boldsymbol{\sigma}]^2]$$
$$\leq \sum_i \|M_i\|^2\exp(-1/2\|M_i\|^2)$$
$$\leq \exp(-1/2\max_i\|M_i\|^2)\sum_i\|M_i\|^2$$
$$= \exp(-1/2\|M\|_{2,\infty}^2)\|M\|_F^2.$$

Bounding $\mathbb{E}[f]$ by $\sqrt{\mathbb{E}[f^2]}$ and plugging in to (69) completes the proof. ∎

REDUCTION TO LARGE ERROR FRACTIONS

Lemma 12: Let $A \in \mathbb{R}^{m\times n}$, and suppose $(\boldsymbol{x}_0, \boldsymbol{e}_0)$ is the unique optimal solution to

$$\min\|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subject to} \quad A\boldsymbol{x} + \boldsymbol{e} = A\boldsymbol{x}_0 + \boldsymbol{e}_0. \quad (71)$$

Then if $\text{supp}(\boldsymbol{x}_0) \subseteq I$, then $(\boldsymbol{x}_0(I), \boldsymbol{e}_0)$ is the unique optimal solution to

$$\min\|\boldsymbol{x}'\|_1 + \|\boldsymbol{e}\|_1 \text{ subject to } A_I\boldsymbol{x}' + \boldsymbol{e} = A_I\boldsymbol{x}_0(I) + \boldsymbol{e}_0. \quad (72)$$

Proof: Any solution $(\boldsymbol{x}', \boldsymbol{e})$ to $A_I\boldsymbol{x}' + \boldsymbol{e} = A_I\boldsymbol{x}_0(I) + \boldsymbol{e}_0$ corresponds to a solution $(\boldsymbol{x}, \boldsymbol{e})$ to $A\boldsymbol{x} + \boldsymbol{e} = A_I\boldsymbol{x}_0 + \boldsymbol{e}_0$, by setting $\boldsymbol{x}(I) = \boldsymbol{x}'$, $\boldsymbol{x}(I^c) = 0$. If $(\boldsymbol{x}_0(I), \boldsymbol{e}_0)$ is not unique or optimal for (72), then $(\boldsymbol{x}_0, \boldsymbol{e}_0)$ is not unique or optimal for (71). ∎

This lemma mimics [36]:

Lemma 13: Suppose that $(\boldsymbol{x}_0, \boldsymbol{e}_0)$ is the unique optimal solution to

$$\min\|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subject to} \quad A\boldsymbol{x} + \boldsymbol{e} = A\boldsymbol{x}_0 + \boldsymbol{e}_0. \quad (73)$$

Let $J = \text{supp}(\boldsymbol{e}_0)$, and suppose $J' \subset J$, and that $\boldsymbol{e}_0'$ is constructed by setting $\boldsymbol{e}_0'(J') = \boldsymbol{e}_0(J')$, $\boldsymbol{e}_0'(J'^c) = 0$. Then $(\boldsymbol{x}_0, \boldsymbol{e}_0')$ is the unique optimal solution to

$$\min\|\boldsymbol{x}\|_1 + \|\boldsymbol{e}\|_1 \quad \text{subject to} \quad A\boldsymbol{x} + \boldsymbol{e} = A\boldsymbol{x}_0 + \boldsymbol{e}_0'. \quad (74)$$

Proof: Suppose on the contrary, that there exist $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{e}}$ satisfying $A\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{e}} = A\boldsymbol{x}_0 + \boldsymbol{e}_0'$ and

$$\|\tilde{\boldsymbol{x}}\|_1 + \|\tilde{\boldsymbol{e}}\|_1 \leq \|\boldsymbol{x}_0\|_1 + \|\boldsymbol{e}_0'\|_1.$$

Set $\boldsymbol{h} = \boldsymbol{e}_0 - \boldsymbol{e}_0'$. Notice that $A\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{e}} + \boldsymbol{h} = A\boldsymbol{x}_0 + \boldsymbol{e}_0$, and so $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{e}} + \boldsymbol{h})$ is feasible for (73). But

$$\|\boldsymbol{x}_0\|_1 + \|\boldsymbol{e}_0\|_1 = \|\boldsymbol{x}_0\|_1 + \|\boldsymbol{e}_0'\|_1 + \|\boldsymbol{h}\|_1$$
$$\geq \|\tilde{\boldsymbol{x}}\|_1 + \|\tilde{\boldsymbol{e}}\|_1 + \|\boldsymbol{h}\|_1 \geq \|\tilde{\boldsymbol{x}}\|_1 + \|\tilde{\boldsymbol{e}} + \boldsymbol{h}\|_1.$$

Since the optimizer to (73) is unique, we conclude that $\tilde{\boldsymbol{x}} = \boldsymbol{x}_0$, $\tilde{\boldsymbol{e}} = \boldsymbol{e}_0'$. ∎

REFERENCES

[1] T. Figiel, J. Lindenstrauss, and V. D. Milman, "The dimension of almost spherical sections of convex bodies," *Acta Mathematica*, vol. 139, no. 1-2, pp. 53–94, 2008.

[2] B. S. Kashin, "The widths of certain finite-dimensional sets and classes of smooth functions," *Izvestia*, vol. 41, no. 2, pp. 334–351, 2008.

[3] M. Ledoux, *The Concentration of Measure Phenomenon, Mathematical Surveys and Monographs 89.* Providence, RI: American Mathematical Society, 2001.

[4] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[5] W. Fu and K. Knight, "Asymptotics for Lasso-type estimators," *Annals of Statistics*, vol. 28, no. 5, pp. 1356–1378, 2000.

[6] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *Annals of Statistics*, vol. 37, no. 1, pp. 246–270, 2009.

[7] R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss, "Combining geometry and combinatorics: A unified approach to sparse signal recovery," in *Allerton Conference on Communication, Control and Computing*, 2008, pp. 798–805.

[8] V. Guruswami, J. R. Lee, and A. Razborov, "Almost Euclidean subspaces of ℓ_1^n via expander codes," in *Symposium on Discrete Algorithms*, 2008, pp. 353–362.

[9] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[10] W. Bajwa, J. Haupt, G. Raz, and R. Nowak, "Compressed channel sensing," in *Proceedings of Conference on Information Sciences and Systems*, 2008, pp. 5–10.

[11] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[12] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[13] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[14] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[15] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[16] D. Donoho, "For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[17] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth, "On the Lambert W function," *Advances in Computational Mathematics*, vol. 5, no. 329, 1996.

[18] M. Rudelson and R. Vershynin, "On sparse reconstruction from fourier and gaussian measurements," *Communications on Pure and Applied Mathematics*, vol. 61, pp. 1025–1045, 2008.

[19] D. Donoho and J. Tanner, "Exponential bounds implying the construction of compressed sensing matrices, error-correcting codes and neighborly polytopes by random sampling," *IEEE Transactions on Information Theory*, 2010.

[20] ——, "Counting faces of randomly projected polytopes when the projection radically lowers dimension," *Journal of the American Mathematical Society*, vol. 22, no. 1, pp. 1–53, 2009.

[21] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[22] D. Needell and J. Tropp, "CoSAMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[23] Y. Eldar and M. Mishali, "Robust recovery of signals from a union of subspaces," *IEEE Transactions on Information Theory,* to be published.

[24] M. Sipser and D. A. Spielman, "Expander codes," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 1710–1722, 1996.

[25] J. Feldman, T. Malkin, R. Servedio, C. Stein, and M. Wainwright, "LP decoding corrects a constant fraction of errors," *IEEE Transactions on Information Theory*, vol. 53, no. 1, pp. 82–89, 2007.

[26] M. R. Capalbo, O. Reingold, S. P. Vadhan, and A. Wigderson, "Randomness conductors and constant-degree lossless expanders," in *Proceedings of the 34th ACM Symposium on Theory of Computing*, 2002, pp. 659–668.

[27] N. Kashyap, "A decomposition theorem for binary linear codes," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 3035–3058, 2008.

[28] D. Donoho, "Neighborly polytopes and sparse solution of underdetermined linear equations," preprint, 2005. [Online]. Available: www-stat.stanford.edu/Reports/abstracts/05-04.pdf.

[29] J. Fuchs, "On sparse representation in arbitrary redundant bases," *IEEE Transactions on Information Theory*, vol. 50, no. 6, pp. 1341–1344, 2004.

[30] S. Szarek, "Condition numbers of random matrices," *Journal of Complexity*, vol. 7, pp. 131–149, 1991.

[31] E. Candes and J. Romberg, "ℓ^1-magic: Recovery of sparse signals via convex programming," 2005. [Online]. Available: http://www.acm.caltech.edu/l1magic/.

[32] D. Needell and R. Vershynin, "Signal recovery from inaccurate and incomplete measurements via regularized orthogonalized matching pursuit," *Journal of Selected Topics in Signal Processing,* to be published.

[33] D. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ^1-norm near solution approximates the sparest solution," preprint, 2004. [Online]. Available: www-stat.stanford.edu/~donoho/Reports/2004/l1l0approx.pdf.

[34] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2005.

[35] ——, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory,* submitted for publication.

[36] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *preprint*, 2009.

[37] A. Ganesh, J. Wright, X. Li, E. Candes, and Y. Ma, "Dense error correction for low-rank matrices via principal component pursuit," in *International Symposium on Information Theory (ISIT), submitted for publication*, 2010.

[38] Z. Zhou, A. Wagner, H. Mobahi, and Y. Ma, "Face recognition with contiguous occlusion using markov random fields," in *Proceedings of International Conference on Computer Vision*, 2009.

[39] M. Talagrand, "Concentration of measure and isoperimetric inequalities in product spaces," *Publications Mathematiques de l'I.H.E.S.*, vol. 81, pp. 73–205, 1995.

Yi Ma is an associate professor at the Electrical and Computer Engineering Department of the University of Illinois at Urbana-Champaign. He is also the research manager of the Visual Computing group at Microsoft Research Asia in Beijing since January 2009. His main research interest is in computer vision, high-dimensional data analysis, and systems theory. He is the first author of the popular vision textbook "An Invitation to 3-D Vision," published by Springer in 2003. Yi Ma received two Bachelors' degree in Automation and Applied Mathematics from Tsinghua University (Beijing, China) in 1995, a Master of Science degree in EECS in 1997, a Master of Arts degree in Mathematics in 2000, and a PhD degree in EECS in 2000, all from the University of California at Berkeley. Yi Ma received the David Marr Best Paper Prize at the International Conference on Computer Vision 1999, the Longuet-Higgins Best Paper Prize at the European Conference on Computer Vision 2004, and the Sang Uk Lee Best Student Paper Award with his students at the Asian Conference on Computer Vision in 2009. He also received the CAREER Award from the National Science Foundation in 2004 and the Young Investigator Award from the Office of Naval Research in 2005. He is an associate editor of IEEE Transactions on Pattern Analysis and Machine Intelligence and has served as the chief guest editor for special issues for the Proceedings of IEEE and the IEEE Signal Processing Magazine. He will also serve as Program Chair for ICCV 2013 in Sydney, Australia. He is a senior member of IEEE and a member of ACM, SIAM, and ASEE.

John Wright is a researcher with the Visual Computing Group at Microsoft Research Asia. He received his PhD in Electrical Engineering from the University of Illinois at Urbana-Champaign. His graduate work focused on developing efficient and provably correct algorithms for error correction with high-dimensional data, and on their application in automatic face recognition. His research interests encompass a number of topics in vision and signal processing, including minimum description length methods for clustering and classification, error correction and inference with non-ideal data, video analysis and tracking, as well as face and object recognition. His work has received a number of awards and honors, including a UIUC Distinguished Fellowship, Carver Fellowship, Microsoft Research Fellowship, the UIUC Martin Award for Outstanding Graduate Research, and the Lemelson-Illinois Prize for Innovation.