July 2008
UILU-ENG-08-2210
DC-237

DENSE ERROR CORRECTION VIA ℓ^1 -MINIMIZATION

John Wright and Yi Ma

Coordinated Science Laboratory
1308 West Main Street, Urbana, IL 61801
University of Illinois at Urbana-Champaign

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 2008	

4. TITLE AND SUBTITLE
Dense Error Correction via l^1-Minimization

5. FUNDING NUMBERS
NSF CRS-EHS-0509151
NSF CCF-TF-0514955,
ONR YIP N00014-05-1-0633
NSF IIS 07-03756

6. AUTHOR(S)
John Wright and Yi Ma

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
1308 West Main Street
Urbana, Illinois 61801-2307

8. PERFORMING RGANIZATION REPORT NUMBER
UILU-ENG-08-2210
DC-237

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
NSF, 4201 Wilson Blvd, Arlington, VA 22203

ONR, Ballston Centre, Tower 1, 800 N. Quincy, Arlington, VA 22217-5660

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official position, policy, or decision, unless so designated by other documentation

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In this paper, we study 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 could be unbounded. Motivated by the problem of face recognition in computer vision, we will prove that if a signal has a sufficiently sparse representation with respect to a highly correlated dictionary A (either overcomplete or not), then with overwhelming probability, it can be recovered by solving the following ℓ_1-minimization problem: $\min \|x\|_1 + \|e\|_1$ subject to $y = Ax + e$; even for very dense e. More precisely, in this paper we prove that under the above conditions, for any $\rho < 1$, as m goes to infinity, solving the above ℓ_1-minimization problem correctly recovers any sparse enough non-negative signal x from almost any error e with support size $\rho \cdot m$. This result suggests that accurate recovery of sparse signals is possible and computationally feasible even with errors asymptotically approaching 100%! The proof relies on a careful characterization of the neighborliness of a convex polytope spanned together by the standard cross polytope and a nonzero mean Gaussian ensemble with a small variance, which we call the "cross-and-bouquet" model. The high neighborliness of this polytope enables the striking error correction ability of the above ℓ_1-minimization. We will also show simulations and experimental results that corroborate our findings.

14. SUBJECT TERMS
Sparse Signal Recovery, Dense Error Correction, ℓ_1-minimization, Gaussian Random Ensemble, Polytope Neighborliness.

15. NUMBER OF PAGES
36

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Dense Error Correction via ℓ^1-Minimization

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

Abstract

In this paper, we study 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 could be unbounded. Motivated by the problem of face recognition in computer vision, we will prove that if a signal has a sufficiently sparse representation with respect to a highly correlated dictionary A (either overcomplete or not), then with overwhelming probability, it can be recovered by solving the following ℓ^1-minimization problem:

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

even for very dense e. More precisely, in this paper we prove that under the above conditions,

for any $\rho < 1$, as m goes to infinity, solving the above ℓ^1-minimization problem correctly recovers any sparse enough non-negative signal x from almost any error e with support size $\leq \rho m$.

This result suggests that accurate recovery of sparse signals is possible and computationally feasible even with errors asymptotically approaching 100%! The proof relies on a careful characterization of the neighborliness of a convex polytope spanned together by the standard cross polytope and a nonzero mean Gaussian ensemble with a small variance, which we call the "cross-and-bouquet" model. The high neighborliness of this polytope enables the striking error correction ability of the above ℓ^1-minimization. We will also show simulations and experimental results that corroborate our findings.

Index Terms

Sparse Signal Recovery, Dense Error Correction, ℓ^1-minimization, Gaussian Random Ensemble, Polytope Neighborliness.

J. Wright and Y. Ma are with the Electrical and Computer Engineering Department, University of Illinois at Urbana-Champaign. Corresponding author: John Wright, 146 Coordinated Science Lab, 1308 W. Main St., Urbana, Illinois 61801. Email: jnwright@uiuc.edu.

I. INTRODUCTION

Recovery of high-dimensional sparse signals or errors has been one of the fastest growing research areas in signal processing for the past few years. At least two factors have contributed to this explosive progress. On the theoretical side, the progress has been propelled by powerful tools and results from multiple mathematical areas such as random matrices, discrete geometry, combinatorics, and coding theory. On the practical side, a lot of excitement has been generated by its remarkable success in many real-world applications in areas like signal (image or speech) processing, computer vision, and pattern recognition.

A. A Motivating Example

One notable successful application of sparse representation is automatic face recognition. As described in [1], face recognition can be cast as a sparse representation problem 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 y is, 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 under-determined 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:

$$\min_x \|x\|_1 \quad \text{subject to} \quad y = Ax. \tag{1}$$

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 non-zero entries correspond to the corrupted pixels. To correct the error e, if A is a tall matrix, i.e. $m > n$, Candes and Tao [2] has proposed to multiply the equation $y = Ax + e$ with the orthogonal complement of A, say B, and then use ℓ^1-minimization to recover the error vector e from the new linear equation $By = Be$ if e is sparse.

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 trick even if the error e is known to be sparse. To resolve this difficulty, in [1], the authors have proposed instead to solve $[x, e]$ altogether

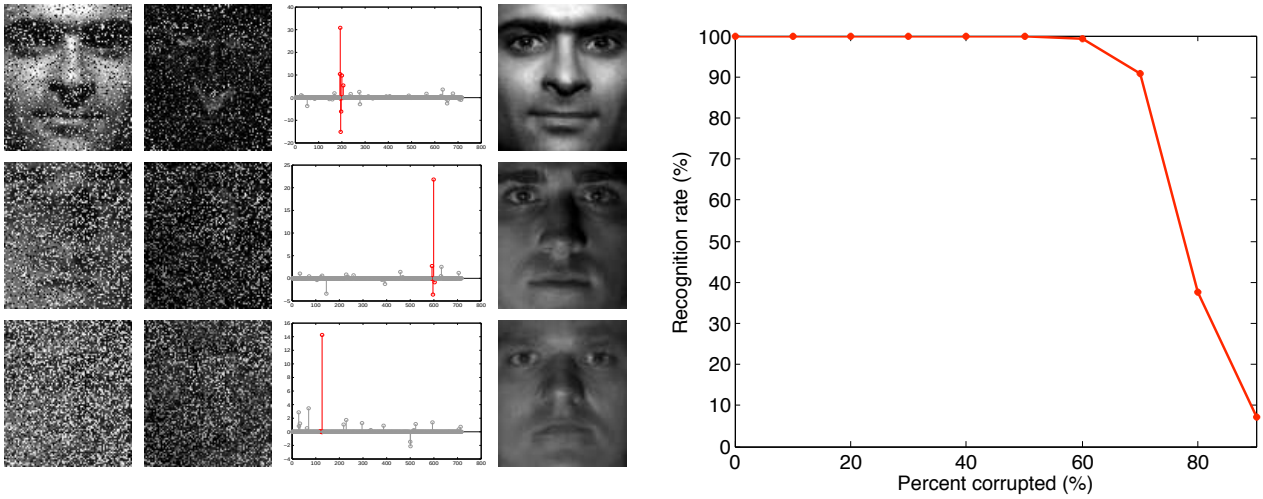


Fig. 1. **Face recognition under random corruption.** Left: (a) Test images $\boldsymbol{y}$ from Extended Yale B, with random corruption. Top row: 30% of pixels are corrupted, Middle row: 50% corrupted, Bottom row: 70% corrupted. (b) Estimated errors $\hat{e}$. (c) Estimated sparse coefficients $\hat{x}$. (d) Reconstructed images $y_r = A\hat{x}$. 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.

as the sparsest solution to the extended equation $y = [A\ I]w$ with $w = \left[\begin{smallmatrix} x \\ e \end{smallmatrix}\right] \in \mathbb{R}^{m+n}$, and w can be found by solving the extended ℓ^1-minimization problem:

$$\min_{w} \|w\|_1 \quad \text{subject to} \quad y = [A\ \ I]w. \tag{2}$$

This seemingly minor modification to the previous error correction approach has drastic consequences on the performance of robust face recognition. Solving the modified ℓ^1-minimization enables almost perfect recognition even with more than 60% pixels of the query image are 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 [2].

Although ℓ^1-minimization is expected to recover sparse solutions with high probability for general systems of linear equations, it is rather surprising that it works for the equation $y = [A\ I]w$ at all. The columns of A are highly correlated in the case of face recognition. 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

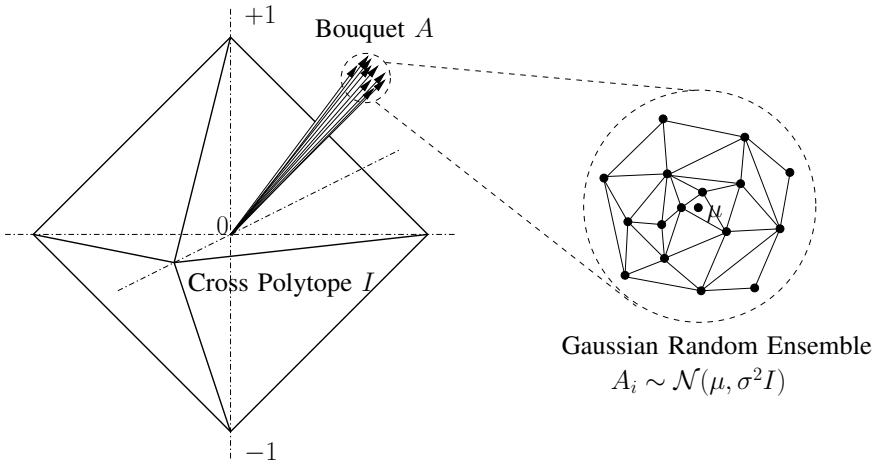


Fig. 2. The "cross-and-bouquet" model. Left: the bouquet A and the crosspolytope spanned by the matrix $\pm I$. Right: the tip of the bouquet magnified, which is a Gaussian random ensemble with a small variance σ^2 around the mean vector μ. The cross-and-bouquet polytope is spanned by vertices from both the bouquet A and the cross $\pm I$.

only an extremely tiny portion of the unit sphere $\mathbb{S}^{m-1}$. [1] Geometrically, the vectors in A are all tightly bundled together as a "bouquet," whereas the vectors associated with the identity matrix and its negative $\pm I$ together[2] form a standard "cross" in $\mathbb{R}^m$, as illustrated in Figure 2. Notice that such a "cross-and-bouquet" type matrix $[A\ I]$ is neither incoherent nor (restrictedly) isometric, at least not uniformly. Also, the density of the desired solution w is not uniform either. The x part of w is usually a very sparse non-negative vector, but the e part can be very dense and have arbitrary signs. 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 could be any random matrix with highly correlated column vectors. The simulation results in Section III 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 to 100% asymptotically!

[1] At first sight, this seems somewhat surprising as faces of different people look so different to human eyes. That is probably because human brain has adapted to distinguish highly correlated visual signals such as faces or voices. The result of this paper may help understand why such tasks can be done accurately and robustly as long as the dimension of the signal is high enough.

[2] Here we allow the entries of the error e to assume either positive or negative signs.

B. The Main Model and Result

Motivated by the above empirical observations, this paper aims to resolve the apparent discrepancy between theory and practice of ℓ^1-minimization and gives a more careful characterization of its behavior in recovering $[x, e]$ from the cross-and-bouquet (CAB) type models:

$$y = Ax + e = [A \;\; I]w. \tag{3}$$

We model the bouquet, the columns of A, as iid samples from a multivariate Gaussian distribution $\mathcal{N}(\mu, \sigma^2 I)$, where $\sigma = \nu m^{-1/2}$ with ν sufficiently small and $\|\mu\|_2 = 1$ and $\|\mu\|_\infty \leq C_\mu m^{-1/2}$ for some $C_\mu \in \mathbb{R}_+$. The first condition ensures that the bouquet remains tight as the dimension m grows; the second condition requires that the mean of the bouquet is mostly incoherent with the columns of the cross $\pm I$.

We will consider proportional growth for m and n, that is, $m/n \to \delta \in \mathbb{R}_+$ as $m \to \infty$. However, the support size of the sparse signal x is only allowed to grow *sublinearly* in m, or more precisely $\|x\|_0 = o(m^{1-\eta})$ for some $\eta > 0$. Be aware that this is different from the typical assumption in the sparse representation literature where the support normally grows proportionally with the dimension – no matter how small that proportion might be. There are technical reasons why the support of the signal x can only be sublinear if we allow the support of the error e to be arbitrarily dense, which we will explain soon. In practice, this sublinear bound of sparsity is in fact more than adequate for signals in many practical problems, including the face recognition problem where ideally the support of x is bounded by a constant – the number of images per subject.

This paper proves that under the above conditions

> *for any $\rho < 1$, as m goes to infinity, solving the above ℓ^1-minimization problem correctly recovers any non-negative sparse signal x from almost any error e with support size $\leq \rho m$.*

While we will leave a more precise statement and the proof of the fact to Section II, for the rest of this section, we will discuss some of the main implications of this result, in the broader context of sparse signal recovery and many of its potential applications.

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 simply do not apply to solving this new type of equations $y = Ax + e$. The cross-and-bouquet type matrix $[A \; I]$ is neither highly isometric nor incoherent and the solution $[x, e]$ sought has very uneven density (or sparsity). As a result, greedy algorithms such as

Matching Pursuit [3], [4] consistently fail for this kind of problems, unless both x and e are unrealistically sparse. However, this does not mean that the concepts of restricted isometry or others become irrelevant to the new problem. On the contrary, the proof of our results precisely rely on characterizing a special type of restricted isometry associated with this new problem, see Lemma 4.

b) Error correction: From an error correction viewpoint, the above result seems almost impossible: One is able to correctly solve a set of linear equations with *almost all* the equations are randomly and arbitrarily corrupted! This is especially surprising given the fact that in the binary domain $\mathbb{Z}_2$, the best error-correction codes that are constructed based on the best expander graphs can normally correct errors that are only a fraction of the code [5]. The relationships between our result and binary error-correction codes remain to be uncovered. Here we have the following observations about its relations to exiting error correction methods in the domain of real numbers:

- When $n < m$, the range of A is a subspace in $\mathbb{R}^m$. In this case, one can directly apply the results of Candes and Tao. However, the error vector e needs to be sparse for that approach whereas our result suggests even dense errors (with support far beyond 50%) can be corrected by solving instead the extended ℓ^1-minimization (2).
- The sublinear growth of the support of x in m is the best one can hope for in the regime of dense errors. In general, we need at least $\|x\|_0$ independent linear equations to be able to recover x correctly. If $n \approx m$, as an arbitrary portion of the m equations can be totally corrupted by e, one cannot ensure any fixed portion of the equations remain good for recovering x. Of course, if the error e is sparse, then the ℓ^1-minimization (2) will be able to recover x with linear growth in support, as ensured by the existing theory [2], [6], [7]. In this paper, we are only interested in how the ℓ^1-minimization behaves with dense errors.
- When $n \geq m$, in general the Gaussian matrix A is full rank and the method of Candes and Tao simply does not apply any more in this situation. Our result suggests that as long as A is highly correlated, the ℓ^1-minimization (2) can still recover the sparse signal x correctly with high probability even if almost all the equations might be corrupted. One may also choose to pre-multiply the equation $y = Ax + e$ with an "approximate" orthogonal complement of A, say the orthogonal complement of the mean vector μ, which is an $(m-1) \times m$ matrix B. Then the equation becomes $By = Be + z$ where $z = BAx$ is a signal with small magnitude due to the almost orthogonality between B and A. One can view z as noise and try to recover e as a sparse signal via ℓ^1-minimization. However, in theory the breakdown point for such ℓ^1-minimization is far below 50% unless e is non-negative in which case the breakdown point could approach 100% [8].

c) Polytope geometry: The reason why one can recover sparse solutions from a system of linear equations $y = Ax$ with high probability relies on a fundamental (and surprising) property of high-dimensional random polytopes: As m and n grow proportionally, if the column vectors of A are random samples from a zero-mean Gaussian $\mathcal{N}(0, I)$, the convex polytope spanned by the vectors, denoted as $\text{conv}(A)$, is highly neighborly [7], [9]. Neighborliness of Gaussian (or other) random polytopes has been well characterized in the literature. These properties provide the necessary and sufficient conditions when ℓ^1-minimization (1) is able to recover the sparse solution x for the equation $y = Ax$. More precisely, the ℓ^1-minimization (1) can correctly recover the sparse solution x if and only if the columns associated with the non-zero entries of x span a face of the polytope $\text{conv}(A)$.

In our case, the column vectors of the matrix A are a Gaussian random ensemble with non-zero mean and small variance whereas vectors of the cross $\pm I$ are completely fixed. To characterize when the extended ℓ^1-minimization (2) is able to recover the solution $[x, e]$ correctly, we need to examine the geometry of the peculiar convex polytope spanned together by the random bouquet A and the fixed cross $\pm I$, denoted as $\text{conv}(A, \pm I)$. Thus, it comes at no surprise that the proof of our main result relies on a careful study of the geometry of this cross-and-bouquet polytope. As we will show that indeed, the vertices associated with the non-zero entries of x and e form a face of the polytope with overwhelming probability as the dimension m becomes large. Precisely due to this special neighborliness of the cross-and-bouquet polytope, the extended ℓ^1-minimization (2) is able to correctly recover the desired solution, regardless it is sparse, dense, or a mixture of both.[3]

D. Implications on Applications

d) Robust classification and source separation: The new result about the cross-and-bouquet model obviously has strong implications on robust classification or separation of highly correlated classes of signals such as faces or voices, despite severe corruption. It helps explain the surprising performance of 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, the ℓ^1-minimization would tolerate even higher level of corruption, far beyond the 60% at the resolution of [1]. Other applications where this kind of model can be extremely useful and effective include speech recognition/imputation, audio source separation, video segmentation, or activity recognition from motion sensors.

[3]Notice that the equation $y = [A\ I]w$ is under-determined. The kernel of the matrix $[A\ I]$ is non-trivial and so the solution to this equation is not unique. The main result essentially guarantees that due to the geometry of the cross-and-bouquet polytope, the solution found by the ℓ^1-minimization is the correct one even if it might not be sparse.

e) 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, as one packet of binary streams, and each packet has a high probability 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. Obviously, the high correlation in A is to ensure that there is sufficient redundancy built in the encoded message Ax so that the information about x will not be lost even if many entries of Ax can be corrupted while being sent through this channel.

f) 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 randomly chosen from a Gaussian ensemble), the sender can deliberately corrupt the message Ax with arbitrary random errors e before sending it to the receiver. The receiver only has to run a linear programming to recover the information x, whereas any eavesdropper will not be able to make much sense out of the highly corrupted message $y = Ax + e$. Of course, the long-term security of such an encryption scheme relies on how hard it is for one to learn 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 $y = Ax$. Even if the dimensions of the matrix A are given, to effectively learn A from a set of observed messages $Y = [y_1, y_2, \ldots, y_k]$ is still a largely open problem, known in the literature as the "dictionary-learning" problem. The best known algorithm is iterative in nature and with no theoretical guarantee of convergence [10]. The problem with highly corrupted observations is expected to be a much more daunting problem for any code breaker. But its hardness is still open.

II. Roadmap of the Proof

In this section, we fist give a precise statement for the main result in Section II-A. We will then lay out the roadmap for the proof of the main result, starting with the key geometric picture behind the proof described in Section II-B. In Section II-C, we will prove the main result by assuming that two technical conditions in Lemma 2 hold. In the interest of space, we leave the lengthy proofs for the two technical conditions to the Appendix, and will only discuss the main ideas behind their proofs in Section II-D.

A. Problem Statement

For the cross-and-bouquet model (3), let $y = Ax_0 + e_0$ for some signal-error pair (x_0, e_0) where x_0 is a non-negative sparse signal and e_0 a dense vector with arbitrary signs. We are interested in the conditions

under which the ℓ^1-minimization (2) can recover the correct solution.

For any $n \in \mathbb{Z}$, $[n]$ will denote $\{1, \ldots, n\}$. Denote $\text{supp}(x_0) = I \subset [n]$, $\text{supp}(e_0) = J \subset [m]$, $\text{sgn}(e_0(J)) = \sigma$, and 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.

Assumption 1 (Weak Proportional Growth): We say that a sequence of signal-error problems exhibits weak proportional growth if $m \to \infty$, $n/m \to \delta \in \mathbb{R}_+$, $k_2/m \to \rho \in (0,1)$ and $k_1 = o(m)$. More precisely, we assume $\exists\, C_0,\ \eta_0 > 0$ such that $k_1 \leq C_0\, m^{1-\eta_0}$.

We will consider matrices A drawn from certain distributions. The model for $A \in \mathbb{R}^{m \times n}$ needs to capture the idea that it consists of small deviations about a mean, hence a "bouquet". In this paper, we consider the columns of A are iid samples from a Gaussian distribution:

$$A_i \sim_{iid} \mathcal{N}\left(\mu, \frac{\nu^2}{m} I_m\right). \tag{4}$$

Assumption 2 (Centrality of the Bouquet): There exists $C_\mu \in \mathbb{R}_+$, m_0, such that for all $m > m_0$, $\|\mu\|_2 = 1$ and $\|\mu^{(m)}\|_\infty \leq C_\mu m^{-1/2}$.

In the following, we will say the cross-and-bouquet model is ℓ^1*-recoverable at* (I, J, σ) if for all $(x_0 \geq 0, e_0)$ with support (I, J) and e_0 with the signs σ, we have

$$(x_0, e_0) = \arg\min \|x\|_1 + \|e\|_1 \quad \text{subject to} \quad Ax + e = Ax_0 + e_0, \tag{5}$$

and the minimizer is uniquely defined. From the geometry of ℓ^1-minimization, if (5) does not hold for some pair (x_0, e_0), then it does not hold for any (x, e) with the same signs and support as (x_0, e_0) [9]. Thus, understanding ℓ^1-recoverability at each (I, J, σ) completely characterizes which solutions to $y = Ax + e$ can be correctly recovered. In this language, our main result can be stated more precisely as:

Theorem 1 (Error Correction with the Cross-and-Bouquet Model): For any $\delta > 0$, and $\rho < 1$, there exists a $\nu_0(\delta, \rho, C_\mu) > 0$ such that for a sequence of CAB models $A^{(m)}$ with $n(m) = \lfloor \delta m \rfloor,\ k_1(m) = o(m),\ k_2(m) = \lfloor \rho m \rfloor$, $\nu < \nu_0$ and errors $e^{(m)}$ with support $J^{(m)}$ of the error is chosen uniformly at random from $\binom{[m]}{k_2}$ and signs $\sigma^{(m)}$ chosen uniformly at random from $\{\pm 1\}^{k_2}$,

$$\lim_{m \to \infty} P_{A,J,\sigma}\left[\, \forall\, I \in \binom{[n]}{k_1},\ \ell^1\text{-recoverability at } (I, J, \sigma)\right] \;=\; 1. \tag{6}$$

B. Problem Geometry

We prove the above theorem by first restating geometrically the necessary and sufficient conditions for ℓ^1-recoverability, as separation of a higher-dimensional ℓ^1-ball and an affine subspace (see Figure 3). To witness this separation, we must show the existence of a separating hyperplane, whose normal we will denote by q.

Let $A \in \mathbb{R}^{m\times n}$, $I \subset [n]$, $J \subset [m]$ and $\sigma \in \{\pm 1\}^{|J|}$. $A_{J,I}$ will denote the $|J| \times |I|$ submatrix of A indexed by these quantities. Sometimes we will use $A_{J,\bullet}$ as a shorthand for $A_{J,[n]}$. Also, we will use 1_I (or 1_J) to represent a vector in $\mathbb{R}^n$ (or $\mathbb{R}^m$) that has ones on the support I (or J) and zeros elsewhere. To lessen confusion between the index set I and the identity matrix, we will use $\mathtt{I}$ to denote the latter.

Let $w \doteq A^*_{J,\bullet}\sigma - 1_I \in \mathbb{R}^n$ and define

$$G = \begin{bmatrix} A_{J^c,I} & A_{J^c,I^c} \\ 0 & \mathtt{I}_{n-k_1} \end{bmatrix} \in \mathbb{R}^{p\times n}, \quad p = m + n - k_1 - k_2. \tag{7}$$

Below as necessary, we will use $R_1 = \{1, \ldots, m-k_2\} \subset [p]$ to index the upper rows of G (corresponding to A), and $R_2 = [p] \setminus R_1$ to index the lower rows.

Lemma 1: Suppose G has full column rank n.[4] Then the model is ℓ^1-recoverable at (I, J, σ) iff

$$\exists\, q \in \mathbb{R}^p \;:\; \|q\|_\infty < 1 \quad \text{and} \quad G^*q = w. \tag{8}$$

Proof: The pair (x_0, e_0) to $y = Ax + e$ is the unique minimum ℓ^1-norm solution to this equation iff

$$\nexists\, (\Delta x, \Delta e) \neq 0 \;:\; A\Delta x = -\Delta e,\; \|x + \Delta x\|_1 + \|e + \Delta e\|_1 \leq \|x\|_1 + \|e\|_1. \tag{9}$$

That, is (x_0, e_0) is optimal iff there is no perturbation $(\Delta x, \Delta e)$ that respects the constraint while not increasing the 1-norm. Assume wlog that $x = 1_I$ and $e \in \{-1, 0, 1\}^m$, and that $\|\Delta x\|_\infty < 1,\ \|\Delta e\|_\infty < 1$; we lose no generality in making the second assumption because the problem is convex – if there exists any nonzero perturbation that does not increase $\|\cdot\|_1$, we may scale it to produce an arbitrarily small perturbation that also does not increase $\|\cdot\|_1$. Then,

$$\|x + \Delta x\|_1 = \|x\|_1 + 1_I^*\Delta x + \|\Delta x_{I^c}\|_1, \quad \text{and} \quad \|e + \Delta e\|_1 = \|e\|_1 + e^*\Delta e + \|\Delta e_{J^c}\|_1.$$

Substituting into (9) and using $\Delta e = -A\Delta x$ yields that (x, e) is optimal iff

$$\nexists\, \Delta x \neq 0 \;:\; \|A_{J^c,\bullet}\Delta x\|_1 + \|\Delta x_{I^c}\|_1 \leq \langle A^*e - 1_I, \Delta x\rangle \tag{10}$$

Condition (10) is satisfied iff

$$\forall\, \Delta x \neq 0,\ \|G\Delta x\|_1 > \langle w, \Delta x\rangle \tag{11}$$

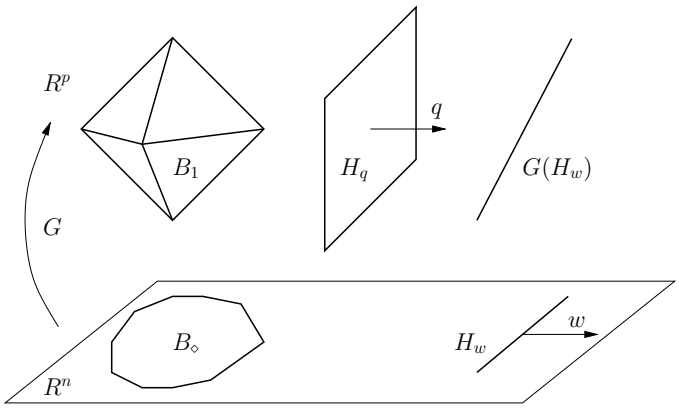


Fig. 3. Geometry for the proof of Lemma 1: The unit ball $B_\diamond$ is separable from H_w in $\mathbb{R}^n$ if and only if in the lifted space $\mathbb{R}^p$, the ℓ^1-ball B_1 is separable from the image of H_w under the injective map G. H_q is the separating hyperplane with a normal vector q. Such an H_q might not be unique in $\mathbb{R}^p$, and q_0 would be the normal to the special separating hyperplane that contains $G(H_w)$.

Let $H_w \subset \mathbb{R}^n$ be the affine subspace $\{x : \langle w, x\rangle = 1\}$. The function $\|G \cdot\|_1$ defines a norm $\|\cdot\|_\diamond$ on $\mathbb{R}^n$. Geometrically, (11) is satisfied iff the unit ball $B_\diamond$ of $\|\cdot\|_\diamond$ is contained in the halfspace $H_w^- = \{\langle w, \cdot\rangle < 1\}$, as illustrated in Figure 3. This unit ball is a convex polytope, given by the inverse image (under the injective map G) of the intersection of $\mathcal{R}(G)$ and the unit ℓ^1 ball B_1 in $\mathbb{R}^p$:

$$B_\diamond = G^{-1}[\,\mathcal{R}(G) \cap B_1(\mathbb{R}^p)\,]\,. \tag{12}$$

Now, $B_\diamond \subset H_w^-$ iff $[\mathcal{R}(G) \cap B_1(\mathbb{R}^p)] \subset G[\,H_w^-\,]$ iff $B_1(\mathbb{R}^p) \cap G(\mathrm{cl} H_w^+) = \emptyset$. These two closed convex sets are nonintersecting iff there is a hyperplane[5] $H_q = \{v \in \mathbb{R}^p \ : \ \langle q, v\rangle = 1\} \subset \mathbb{R}^p$ separating them (see Figure 3 again). We lose no generality in assuming that $B_1 \subset H_q^-$, that $G[\mathrm{cl}\, H_w^+] \subset \mathrm{cl}\, H_q^+$, and that H_q meets the relative boundary $\mathrm{rbd}\, G[\mathrm{cl}\, H_w^+] = G[H_w]$. The first condition occurs iff $\|q\|_\infty < 1$, while the second occurs iff $G^* q = w$. ■

The most natural candidate for a normal vector q is the minimum ℓ^2-norm solution to this equation,

$$q_0 = (G^\dagger)^* w = (GG^*)^{-1} G w. \tag{13}$$

When we use this particular normal q_0, we are demanding that the *projection* of B_1 onto $\mathcal{R}(G)$ lies in $G[H_w^-]$. Since the projection contains the intersection, $B_1 \subset \{\langle q_0, \cdot\rangle < 1\}$ is a sufficient, but not necessary condition. It is not surprising, then, that this condition often does not hold – empirically, $\|q_0\|_\infty \geq 1$

[4]In the model outlined above, this occurs with probability one for m sufficiently large.

[5]Notice H_q cannot contain $0 \in \mathrm{interior}(B_1)$, so the normalization $\langle q, v\rangle = 1$ is appropriate.

with fairly large probability. However, the key observation is that the set of violations $\{|q_0(i)| \geq 1\}$ is often quite small, and we can improve q_0, through an iterative scheme, to a valid q with $\|q\|_\infty < 1$.

C. Iterative Construction of Separator

We next give a lemma that argues that if we are given an initial guess at a normal vector $q_0 \in \mathbb{R}^p$ whose hyperplane $H_0 = \{\langle q_0, \cdot \rangle = 1\}$ separates $G[H_w]$ from *most* of the vertices of B_1, then we can refine q_0 to a q_∞ that separates $G[H_w]$ and *all* of the vertices of B_1. In general, finding such a q_∞ requires solving a linear programming problem. We will analyze the feasibility of this linear program by considering an iteration that is essentially equivalent to the alternating projection method for finding a pair of closest points between two convex sets. In this case, the two convex sets of interest are the hypercube of radius $1-\varepsilon$ and the affine subspace $q_0 + \mathcal{R}(G)^\perp$.

In the following lemma, $q_0 \in \mathbb{R}^p$ is arbitrary (though $q_0 = G^{\dagger *} w$ is natural). We will construct a sequence of vectors $q_0, q_1, \ldots, q_k \ldots$ Let T_k be the "bad set" of indices at iteration k:

$$T_k = \{j : |q_k(j)| > 1\}. \tag{14}$$

Fix a small constant $\varepsilon > 0$, and define the operator θ which takes the part of a vector that sticks out above $1-\varepsilon$:

$$[\theta x]_i \doteq \begin{cases} 0, & \text{for } |x_i| \leq 1-\varepsilon, \\ \operatorname{sgn}(x_i)(|x_i| - 1 + \varepsilon), & \text{for } |x_i| > 1-\varepsilon. \end{cases} \tag{15}$$

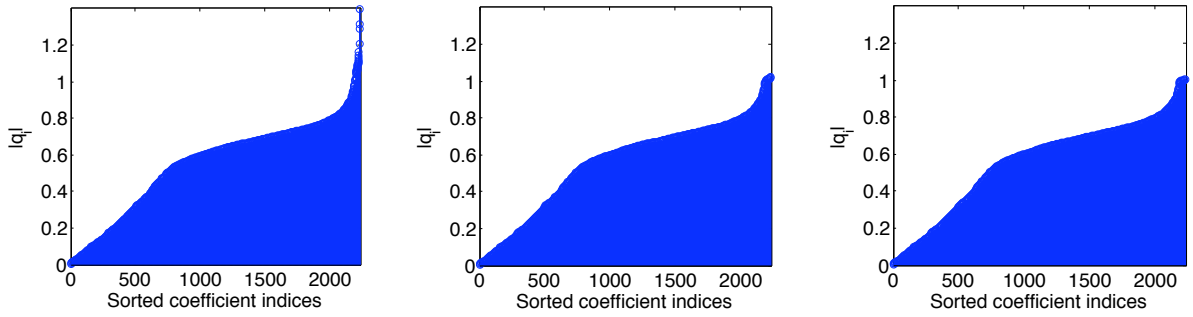
We iteratively construct q_∞ by setting

$$q_{i+1} \;=\; q_i - \pi_{\mathcal{R}(G)^\perp} \theta q_i \;=\; q_i - \theta q_i + \pi_{\mathcal{R}(G)} \theta q_i. \tag{16}$$

Notice that by construction, $G^* q_k = G^* q_0 = w$ for all k. So if $\theta q_i \to 0$, then $\|q_i\|_\infty < 1$ eventually, and q_∞ is a valid separator.

Before proving that this iteration produces a valid separator with high probability, we first demonstrate its behavior on a simulated example with $m = 3,000$, $\delta = .4$, $\nu = .1$, $\rho = .65$, and $k_1 = 10$. Figure 4 plots the sorted absolute values of entries of q_i. Notice that the sorted coefficients clearly divide into two parts; these correspond to the upper (R_1) and lower (R_2) indices. The initial separator, q_0 cleanly separates $G(H_w)$ from most of the vertices of B_1: only 39 entries protrude above $1-\varepsilon$. These entries are quickly iterated away: $\|\theta q\|$ decreases geometrically until after 5 iterations a valid separator is obtained.

Let Γ_k denote the arrangement of all k-dimensional coordinate subspaces (i.e., the set of all $\leq k$-sparse vectors). A simple inductive argument gives the following lemma:



Iteration 0 Iteration 1 Iteration 2

$\#T_0 = 39,\ \|\theta q_0\|_2 = 1.176$ $\#T_1 = 39,\ \|\theta q_1\|_2 = 0.072$ $\#T_2 = 38,\ \|\theta q_2\|_2 = 0.006$

Fig. 4. Iterative refinement producing a separating hyperplane. Here, $m = 3000$, $\delta = .4$, $\nu = .1$, $\rho = .65$, $k_1 = 10$. We plot the sorted magnitudes of the entries of q_i. At left, q_0 separates $G(H_w)$ from most of the vertices of B_1: only 39 violations occur. The distinct bimodal characteristic of q_0 is due to the differences between the statistics of the top (R_1) and bottom (R_2) indices. Applying the iteration decreases $\|\theta q_i\|$ geometrically; after 5 iterations a valid separator is obtained.

Lemma 2: Suppose $\exists\, c \in (0,1)$ such that

$$\xi \doteq \sup_{x \in \Gamma_{cp} \setminus \{0\}} \frac{\|\pi_{\mathcal{R}(G)} x\|_2}{\|x\|_2} < 1 \tag{17}$$

and

$$\|q_0\|_2 + \frac{1}{1-\xi}\|\theta q_0\|_2 \leq (1-\varepsilon)\sqrt{cp}. \tag{18}$$

Then $\lim_{k\to\infty} \theta q_k = 0$.

Proof: Consider the following three statements:

$$\|q_k\|_2 \leq \|q_0\|_2 + \|\theta q_0\|_2 \sum_{i=0}^{k} \xi^i, \tag{19}$$

$$\#T_k \leq cp, \tag{20}$$

$$\|\theta q_k\|_2 \leq \|\theta q_0\|_2\, \xi^k. \tag{21}$$

We will show, by induction on k that these statements hold for all k, giving the desired result. Notice that (19) and (21) are trivially true for $k = 0$. For (20), notice that by (18),

$$\#T_0 \leq \frac{\|q_0\|_2^2}{(1-\varepsilon)^2} \leq cp.$$

Now, suppose the three statements hold for $0, \ldots, k$. Since θq_k has the same signs and smaller magnitude

than q_k, $\|q_k - \theta q_k\|_2 \leq \|q_k\|_2$; combining this with the inductive hypothesis we have

$$\begin{aligned}\|q_{k+1}\|_2 &= \|q_k - \theta q_k + \pi_{\mathcal{R}(G)}\theta q_k\| \leq \|q_k - \theta q_k\| + \|\pi_{\mathcal{R}(G)}\theta q_k\| \leq \|q_k\| + \xi^{k+1}\|\theta q_0\| \\ &\leq \|q_0\|_2 + \|\theta q_0\|_2 \sum_{i=0}^{k+1} \xi^i,\end{aligned}$$

establishing (19) for $k+1$. Similarly, notice that since $\pi_{\mathcal{R}(G)}\theta q_k$ dominates $\theta(q_k - \theta q_k + \pi_{\mathcal{R}(G)}\theta q_k)$ elementwise,

$$\|\theta q_{k+1}\| \leq \|\pi_{\mathcal{R}(G)}\theta q_k\| \leq \xi\|\theta q_k\| \leq \xi^{k+1}\|\theta q_0\|.$$

and (21) holds at $k+1$. Finally, to get the sparsity result (20), note that

$$\|q_{k+1}\|_2 \leq \|q_0\|_2 + \|\theta q_0\| \sum_{i=0}^{k+1} \xi^i \leq \|q_0\|_2 + \frac{1}{1-\xi}\|\theta q_0\|_2 \leq (1-\varepsilon)\sqrt{cp},$$

and so θq_{k+1} must be (cp)-sparse. ■

D. Putting All Together

According to Lemmas 1 and 2, if we can show the two conditions (17) and (18) in Lemma 2 hold asymptotically with overwhelming probability in A as a Gaussian ensemble, that essentially proves the main Theorem 1.

In this subsection, we lay out the main ideas for the rest of the proof, which essentially consists of two parts, one for each of the conditions in Lemma 2: 1. We provide a more accurate bound on the projection ratio ξ for sparse vectors in (17); 2. We show that the initial normal of the separating hyperplane $q_0 = (G^\dagger)^* w$ satisfies the second condition (18), hence it can be fixed by the iterative scheme given in the proof of Lemma 2. More precisely, we will establish the following facts:

1) For a small enough constant c, the projection ratio ξ for cm-sparse signals onto $\mathcal{R}(G)$ is bounded below 1 by a polynomial function in ν. More precisely, $\xi < 1 - C\nu^8$ for some constant $C > 0$. As a result, the coefficient $\frac{1}{1-\xi}$ in the second condition (18) is bounded by $C^{-1}\nu^{-8}$.
2) As m goes to infinity, the ℓ^2-norm of the initial separating normal vector $\|q_0\|_2$ is bounded above by $\nu O(m^{1/2})$, and $\|\theta q_0\|_2$ is bounded above by $e^{-\alpha/\nu^2}O(m^{1/2})$ for some constant α.

Putting these results together, the initial separating normal vector q_0 satisfies:

$$\|q_0\|_2 + \frac{1}{1-\xi}\|\theta q_0\|_2 \;\leq\; \nu O(m^{1/2}) + C^{-1}\nu^{-8}e^{-\alpha/\nu^2}O(m^{1/2}). \tag{22}$$

As long as the deviation of the Gaussian ensemble ν is small enough, the second condition (18) of Lemma 2 will be satisfied since the right hand side is $O(m^{1/2})$. Hence, by Lemma 2, the initial normal

q_0 will converge to a valid normal vector that separates the ℓ^1-ball B_1 from the subspace $G(H_w)$, which essentially proves the main Theorem 1. This intuition is made rigorous in Section C of the appendix.

Whereas Lemmas 1 and 2 have simple geometric and algebraic proofs, the above results require more detailed analysis of large Gaussian matrices. We therefore leave many of the technical details to the appendix, and in this section outline only the main ideas and steps for their proofs. Their derivation is based on recent (and now widely-used) results on concentration of Lipschitz functions, which state that if x is a d-dimensional iid $N(0,1)$ random vector and $f : \mathbb{R}^n \to \mathbb{R}$ is 1-Lipschitz, then [11]

$$P\left[|f(x) - \mathbb{E}f(x)| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\pi^2}\right). \tag{23}$$

Two special cases are particularly of interest here. First, the norm concentrates according to (e.g., for this form see [12]):

$$P\left[\|x\| \geq \beta\sqrt{d}\right] \leq \exp\left(-\frac{2(\beta-1)^2 d}{\pi^2}\right). \tag{24}$$

We will also return to (23) in the proof of Lemma 8 of the appendix. Second, as has been widely exploited in the compressed sensing literature (e.g., [2], [6]), the singular values of rectangular Gaussian matrices with aspect ratio α concentrate about the values $1 \pm \sqrt{\alpha}$ predicted by the Marchenko-Pasteur law:

Fact 1 (Concentration of singular values [11]): Let $A \in \mathbb{R}^{m\times n}$, $(m > n)$ be a random matrix with entries iid $N(0, \frac{1}{m})$. Then if $m \to \infty$ and $n/m \to \delta \in (0,1)$, then for any $t > 0$,

$$P\left[\sigma_{max}(A) > 1 + \sqrt{n/m} + o(1) + t\right] \leq e^{-mt^2/2}, \tag{25}$$

$$P\left[\sigma_{min}(A) < 1 - \sqrt{n/m} + o(1) - t\right] \leq e^{-mt^2/2}. \tag{26}$$

For technical convenience, below we always assume we are in the large error regieme, with $\bar{\rho} \doteq 1-\rho < \delta$. The conclusion still follows for smaller error fractions, since whenever (I, J, σ) is ℓ^1-recoverable, so is $(I, J', \sigma_{J'})$ for any $J' \subset J$. Below, wherever the symbol C occurs with no subscript, it should be read as "some constant." When used in different sections, it need not refer to the same constant.

1) Projection of Sparse Vectors: In this subsection, we upper-bound norm of the projection of any sparse vector onto $\mathcal{R}(G)$. Notice that since the lower coordinates of

$$G \doteq \begin{bmatrix} M_1 & M_2 \\ 0 & \mathtt{I} \end{bmatrix} = \begin{bmatrix} Z_1 + \mu_{J^c} 1^*_{k_1} & Z_2 + \mu_{J^c} 1^*_{\delta m - k_1} \\ 0 & \mathtt{I} \end{bmatrix}$$

contain an identity matrix, when the variance ν^2/m of the perturbations Z_1, Z_2 is small, we expect that sparse vectors with support on R_2 will be very close to $\mathcal{R}(G)$. The following lemma verifies that this is the case, but argues that the projection residual is at least $\Omega(\nu^8)$. The technical conditions appear

complicated, but simply assert that: 1. c is sufficiently small. 2. ν is sufficiently small. 3. we are in the large-error regieme: $\bar{\rho}$ is sufficiently small.

Lemma 3 (Projection of Sparse Vectors): Suppose that

$$c \;<\; \min\left(\frac{\bar{\rho}}{64(1+2C_\mu)^2}, \frac{\bar{\rho}}{1024}\right), \quad \bar{\rho}H(c/\bar{\rho}) + \delta H(c/\delta) < \bar{\rho}/512, \tag{27}$$

$$\bar{\rho} \;<\; \min\left(\delta, \tfrac{1}{2} - \tfrac{1}{C_\mu}\right), \quad \nu < \min\left(\frac{1}{24\sqrt{\bar{\rho}}}, (2\delta)^{-1/4}\right) \tag{28}$$

Then the projection of a sparse vector onto the range of G is bounded as

$$\sup_{y\in\Gamma_{cm}\setminus\{0\}} \frac{\|\pi_{\mathcal{R}(G)}y\|_2}{\|y\|_2} \;<\; 1 - \frac{\nu^8}{16}\left(\frac{(\sqrt{\delta}-\sqrt{\bar{\rho}})(\frac{\sqrt{\bar{\rho}}}{4}-6\sqrt{c})}{1+4\nu^2\left(\sqrt{\delta}+\sqrt{\bar{\rho}}\right)^2}\right)^4. \tag{29}$$

on the complement of a bad event with probability $\asymp e^{-Cm}$.

Proof: The projection of an observation $y = \left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right] \in \Gamma_{cm}$ onto $\mathcal{R}(G)$ solves

$$\min_x \left\|\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right] - Gx\right\|_2^2 = \min_{w_1,w_2} \left\|\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right] - G\left[\begin{smallmatrix} w_1 \\ y_2+w_2 \end{smallmatrix}\right]\right\|_2^2 = \min_{w_1,w_2} \|y_1 - M_1w_1 - M_2(y_2+w_2)\|_2^2 + w_2^*w_2.$$

Since in general M_1 has full rank k_1, we can find the unique optimal w_1 by minimizing the first term:

$$w_1 \;=\; (M_1^*M_1)^{-1}M_1^*y_1 - (M_1^*M_1)^{-1}M_1^*M_2(y_2+w_2)$$

and subsequently, the optimal w_2 satisfies:

$$-M_2^*y_1 + M_2^*M_1w_1 + M_2^*M_2(y_2+w_2) + w_2 = 0 \Rightarrow \left(\mathtt{I} + M_2^*\pi_{M_1^\perp}M_2\right)w_2 = M_2^*\pi_{M_1^\perp}y_1 - M_2^*\pi_{M_1^\perp}M_2y_2,$$

where $\pi_{M_1^\perp}$ denotes the projection matrix onto the orthogonal complement of $\mathcal{R}(M_1)$.

Write $M_2^*\pi_{M_1^\perp} = USV^*$ with $U \in \mathbb{R}^{\delta m - k_1 \times \bar{\rho}m - k_1}$ and $V \in \mathbb{R}^{\bar{\rho}m \times \bar{\rho}m - k_1}$ orthogonal matrices, and the diagonal of $S \in \mathbb{R}^{\bar{\rho}m-k_1 \times \bar{\rho}m-k_1}$ containing the nonzero singular values of $M_2^*\pi_{M_1^\perp}$. Then if w_2 is the solution to the above equation

$$\begin{aligned} \|\, y - \pi_{\mathcal{R}(G)}y \,\|_2 \;\geq\; \|w_2\|_2 \;&=\; \left\|(S^2+\mathtt{I})^{-1}SV^*\left[\begin{smallmatrix} \mathtt{I} & -M_2 \end{smallmatrix}\right]\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right]\right\|_2 \\ &=\; \left\|(S^2+\mathtt{I})^{-1}S\left[\begin{smallmatrix} V^* & -SU^* \end{smallmatrix}\right]\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right]\right\|_2. \end{aligned} \tag{30}$$

Above is the norm of the product of a diagonal matrix $(S^2+\mathtt{I})^{-1}S$, a wide matrix $\left[\begin{smallmatrix} V^* & -SU^* \end{smallmatrix}\right]$, and a sparse vector y. We will bound it by lower bounding the elements of the diagonal matrix, and then lower bounding the "restricted minimum singular value"

$$\gamma_{cm}\left(\left[\begin{smallmatrix} V^* & -SU^* \end{smallmatrix}\right]\right) \;\doteq\; \inf_{y\in\Gamma_{cm}\setminus\{0\}} \|[V^* \quad -SU^*]y\,\|/\|y\|.$$

First, however, we drop the top row of $(S^2+\mathtt{I})^{-1}S[V^* \quad -SU^*]$. This allows us to uniformly lower bound the diagonal of $(S^2+\mathtt{I})^{-1}S$. While σ_1 can be quite large due to the inhomogeneous term ($\mu_{J^c}1^*$), and hence $\frac{\sigma_1}{\sigma_1^2+1}$ can be quite small, for the remaining singular values $\frac{\sigma_i}{\sigma_i^2+1}$ is at least on the order of ν.

To this end, let $\tilde{S} \in \mathbb{R}^{\bar{\rho}m-k_1-1 \times \bar{\rho}m-k_1-1}$ be the diagonal matrix obtained by dropping the row and column of S corresponding to the largest singular value; $\tilde{V}$ and $\tilde{U}$ are obtained by dropping the corresponding columns. From (30),

$$\|w_2\|_2 \;\geq\; \left\|(\tilde{S}^2+\mathtt{I})^{-1}\tilde{S}[\tilde{V}^* \quad -\tilde{S}\tilde{U}^*]\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right]\right\|_2 \;\geq\; \frac{\sigma_{min}(M_2^*\pi_{M_1^\perp})}{1+\sigma_2^2(M_2^*\pi_{M_1^\perp})}\,\gamma_{cm}([\tilde{V}^* \quad -\tilde{S}\tilde{U}^*])\,\|y\|_2, \quad (31)$$

where $\sigma_{min}(M_2^*\pi_{M_1^\perp})$ is the smallest nonzero singular value and $\sigma_2(M_2^*\pi_{M_1^\perp})$ is the second largest singular value.

a) Bounding the second largest singular value $\sigma_2(M_2^*\pi_{M_1^\perp})$: Write $\hat{\mu} \doteq \pi_{M_1^\perp}\mu_{J^c}$, and notice that

$$\begin{aligned}\sigma_2(M_2^*\pi_{M_1^\perp}) &= \inf_{u\neq 0}\sup_{v\neq 0}\frac{\|M_2^*\pi_{M_1^\perp}\pi_{u^\perp}v\|_2}{\|v\|_2} = \inf_{u\neq 0}\sigma_1(M_2^*\pi_{M_1^\perp}\pi_{u^\perp}) \\ &\leq \sigma_1(M_2^*\pi_{M_1^\perp}\pi_{\hat{\mu}^\perp}) = \sigma_1(Z_2^*\pi_{(\mu_{J^c},Z_1)^\perp}).\end{aligned}$$

Choose any orthonormal basis for the subspace $\Sigma = (\mathcal{R}(Z_1)+\mathcal{R}(\mu_{J^c}))^\perp$. Since Σ is probabilistically independent of Z_2, the representation of the projection $Z_2^*\pi_{(\mu_{J^c},Z_1)^\perp}$ with respect to the chosen basis is simply distributed as a $\delta m - k_1 \times \bar{\rho}m - k_1 - 1$ random matrix $\hat{Z}_2$ with entries $N(0,\nu^2/m)$. Since $\frac{\sqrt{m}}{\nu\sqrt{\delta m-k_1}}\hat{Z}_2$ is distributed as $N(0,\frac{1}{\delta m-k_1})$, by Fact 1,

$$P\left[\sigma_1\left(\tfrac{\sqrt{m}}{\nu\sqrt{\delta m-k_1}}\hat{Z}_2\right) \geq 1+\sqrt{\tfrac{\bar{\rho}m-k_1-1}{\delta m-k_1}}+t\right] \leq \exp\left(-(t+o(1))^2(\delta m-k_1)/2\right), \quad (32)$$

and so $P\left[\sigma_1(\hat{Z}_2) \geq 2\nu(\sqrt{\delta}+\sqrt{\bar{\rho}})\right] \asymp e^{-Cm}$. On the complement of this bad event, $\sigma_2^2(M_2^*\pi_{M_1^\perp}) \leq 4\nu^2(\sqrt{\delta}+\sqrt{\bar{\rho}})^2$.

b) Bounding the smallest nonzero singular value $\sigma_{min}(M_2^*\pi_{M_1^\perp}) = \inf_{x\in M_1^\perp}\frac{\|M_2^*x\|_2}{\|x\|_2}$: Choose any orthonormal basis for $M_1^\perp$. Since M_1 is independent of M_2, with respect to this basis, we can write $M_2^*\pi_{M_1^\perp} = \hat{Z}_2 + 1\hat{\mu}^*$, where $\hat{Z}_2 \in \mathbb{R}^{\delta m-k_1\times\bar{\rho}m-k_1}$ is an iid $N(0,\nu^2/m)$ random matrix and $\hat{\mu}$ is the expression of $\pi_{M_1^\perp}\mu_{J^c}$ in this basis.

$$\sigma_{min}(M_2^*\pi_{M_1^\perp}) = \sigma_{min}\left(\pi_{1^\perp}\hat{Z}_2 + 1(m^{-1}1^*\hat{Z}_2+\hat{\mu}^*)\right) \geq \sigma_{min}(\pi_{1^\perp}\hat{Z}_2),$$

where in the final step we have used the orthogonality of $\mathcal{R}(\pi_{1^\perp}\hat{Z}_2)$ and the rank-one perturbation $1(m^{-1}1^*\hat{Z}_2+\hat{\mu}^*)$ to drop the perturbation. Finally, with respect to any orthonormal basis for $1^\perp$, $\pi_{1^\perp}\hat{Z}_2$ is distributed as a $\delta m-k_1-1\times\bar{\rho}m-k_1$ random matrix Z_2' with entries iid $N(0,\nu^2/m)$. Again by Fact 1, $\sigma_{min}(\frac{1}{\nu\sqrt{\delta}}Z_2') \to 1-\sqrt{\frac{\bar{\rho}}{\delta}}$, and similarly by measure concentration,

$$P\left[\sigma_{min}(Z_2') < \frac{\nu}{2}(\sqrt{\delta}-\sqrt{\bar{\rho}})\right] \asymp e^{-Cm}. \quad (33)$$

On the complement of this bad event, $\sigma_{min}(M_2^* \pi_{M_1^\perp}) \geq \frac{\nu}{2}(\sqrt{\delta} - \sqrt{\bar{\rho}})$.

The reader may notice that the above bounds essentially agree with the Marchenko-Pasteur law for the submatrix Z_2. This should be expected, since dropping the largest singular value eliminates most of the influence of μ_{J^c}, while projecting onto a subspace $\mathcal{R}(M_1^\perp)$ of (very small) codimension $k_1 = o(m)$ does not essentially change the conditioning of Z_2^*.

Finally, as will be stated more precisely in Lemma 4 and eventually proved in Appendix A, we will show that with probability at least $1 - e^{-Cm(1+o(1))}$, the restricted singular value γ_{cm} in (31) is lower bounded as

$$\gamma_{cm}([\tilde{V}^* \quad -\tilde{S}\tilde{U}^*]) \quad \geq \quad \frac{\nu\sqrt{\bar{\rho}}}{4} - 6\nu\sqrt{c}.$$

Combining the three results, we have that for all $y \in \Gamma_{cm}$,

$$\frac{\|y - \pi_G y\|_2}{\|y\|_2} \;\geq\; \frac{\nu(\sqrt{\delta} - \sqrt{\bar{\rho}})/2}{1 + 4\nu^2(\sqrt{\bar{\rho}} + \sqrt{\delta})^2} \left(\frac{\nu\sqrt{\bar{\rho}}}{4} - 6\nu\sqrt{c} \right). \tag{34}$$

Notice that $\frac{\|\pi_G y\|}{\|y\|} = \sqrt{1 - \left(\frac{\|y - \pi_G y\|}{\|y\|}\right)^2} \leq 1 - \left(\frac{\|y - \pi_G y\|}{\|y\|}\right)^4$, (where we have used that $1 - x^4 > \sqrt{1 - x^2}$ for $x < 1/\sqrt{2}$). Combined with (34), this implies the desired result (29). ■

We here give a more precise statement about the restricted isometry property about $[\tilde{V}^* \quad -\tilde{S}\tilde{U}^*]$ used in the proof of the above lemma. For an arbitrary matrix M, define $\gamma_k(M)$ as

$$\gamma_k(M) = \inf_{\|y\|_0 \leq k} \frac{\|My\|_2}{\|y\|_2}. \tag{35}$$

We are interested in knowing $\gamma_{cm}([\tilde{V}^* \quad -\tilde{S}\tilde{U}^*])$, where $\tilde{U}$, $\tilde{S}$, and $\tilde{V}$ come from a (compact) singular value decomposition[6] of $P \doteq M_2^* \pi_{M_1^\perp}$, after dropping the largest singular value. That is, if v_1 and u_1 are the (right- and left-) first singular vectors of P, $\tilde{U}\tilde{S}\tilde{V}^*$ is a compact singular value decomposition of $\pi_{u_1^\perp} P \pi_{v_1^\perp}$.

Lemma 4 (Restricted Isometry): Consider the matrix $P \doteq M_2^* \pi_{M_1^\perp} \in \mathbb{R}^{\delta m - k_1 \times \bar{\rho} m}$, and let u_1, v_1 be its first singular vectors. Let $\tilde{U}\tilde{S}\tilde{V}^*$ denote a compact singular value decomposition of $\pi_{u_1^\perp} P \pi_{v_1^\perp}$. If the conditions (27) and (28) hold, then for all m sufficiently large,

$$\gamma_{cm}([\tilde{V}^* \quad -\tilde{S}\tilde{U}^*]) \geq \frac{\nu\sqrt{\bar{\rho}}}{4} - 6\nu\sqrt{c} \tag{36}$$

on the complement of a bad event with probability $e^{-Cm(1+o(1))}$.

[6]With probability one, the matrices U and V are unique upto multiplication of their columns by a common set of signs. The quantity of interest, γ, does not depend on the choice of representative signs.

We postpone the rather technical proof of this lemma to Appendix A, but notice again that the lower bound given in the lemma agrees with (and in fact is looser than) the Marchenko-Pasteur law for a $\bar{\rho}m \times cm$ Gaussian $N(0, \nu^2/m)$ matrix (i.e., the concentration result of Fact 1). In fact, the proof essentially follows by arguing that the parts of this matrix are probabilistically independent, transforming to an equivalent pair of Gaussian matrices, and applying Fact 1. The somewhat technical conditions (27) introduced here are necessary because γ_{cm} involves a minimization over a very large set: all subsets of cm columns. More delicate balancing is needed to ensure that a union bound over this set remains small.

2) Initial Separating Hyperplane: In this section, we analyze the initial separator q_0, obtained as the minimum 2-norm solution to the equation $G^*q = w$. We upper bound both $\|q_0\|_2$ and $\|\theta q_0\|_2$, where θ is the operator that retains the portion of a vector that protrudes above $1 - \varepsilon$ in absolute value. These bounds provide the second half of the conditions needed in Lemma 2 to show that q_0 can be refined by alternating projections to give a true separator. In the following lemma, the exact numerical constants involved in the bounds are less important than the fact that, with respect to decreasing ν, $\|q_0\| = O(\nu)$ and $\|\theta q_0\| = O(e^{-C/\nu^2})$.

Lemma 5: Suppose $\nu < 1$. There exist constants α_1, α_2 such that the initial separator q_0 satisfies

$$\|q_0\|_2 \;\leq\; \alpha_1 \, \nu \, m^{1/2} \;+\; o(m^{1/2}) \tag{37}$$

$$\|\theta q_0\|_2 \;\leq\; \alpha_2 \, e^{-\frac{\bar{\rho}}{128\nu^2}} \; m^{1/2} \;+\; o(m^{1/2}). \tag{38}$$

on the complement of a bad event of probability $\leq e^{-Cm^{1-\eta_0/2}(1+o(1))}$.

Proof: Recall that $w = Z^*_{J,\bullet}\sigma - 1_I + \langle \mu_J, \sigma \rangle 1 \in \mathbb{R}^{\delta m}$, and $q_0 = G^{\dagger *} w$. Notice that

$$G^{\dagger *} = G(G^*G)^{-1} = \begin{bmatrix} Z_1 & Z_2 \\ 0 & \mathtt{I} \end{bmatrix} (G^*G)^{-1} + \begin{bmatrix} \mu_{J^c} 1^* \\ 0 \end{bmatrix} (G^*G)^{-1} \tag{39}$$

where $Z_1 = Z_{J^c,I}$ and $Z_2 = Z_{J^c,I^c}$. Expanding its product with w gives

$$\begin{aligned} q_0 \;=\; & \begin{bmatrix} Z_1 & Z_2 \\ 0 & \mathtt{I} \end{bmatrix} (G^*G)^{-1} Z^*_{J,\bullet}\sigma \;+\; \begin{bmatrix} Z_1 & Z_2 \\ 0 & \mathtt{I} \end{bmatrix} \Big(-(G^*G)^{-1} 1_I + \langle \mu_J, \sigma \rangle \, (G^*G)^{-1} 1 \Big) \\ + \; & \begin{bmatrix} \mu_{J^c} \\ 0 \end{bmatrix} \Big(1^*(G^*G)^{-1} Z^*_{J,\bullet}\sigma \;-\; 1^*(G^*G)^{-1} 1_I \;+\; \langle \mu_J, \sigma \rangle \, 1^*(G^*G)^{-1} 1 \Big). \end{aligned} \tag{40}$$

In this section, we concentrate our efforts on the first term above. A more detailed analysis of $(G^*G)^{-1}$, which we postpone to Lemma 7 of the appendix, shows that the remaining terms are all negligible, contributing $o(m^{1/2})$ to $\|q_0\|$. This is essentially due to the presence of a large common term μ_{J^c} in the

columns of G: the most significant term in G^*G is $\mu^*_{J^c}\mu_{J^c}11^*$, and $(G^*G)^{-1}$ shrinks the ones vector. More precisely, Lemma 7 shows that with probability at least $1-e^{-Cm^{1-\eta_0/2}(1+o(1))}$,

$$\left\| q_0 - \left[\begin{smallmatrix} Z_1 & Z_2 \\ 0 & \mathtt{I} \end{smallmatrix}\right] (G^*G)^{-1} Z^*_{J,\bullet}\,\sigma \right\| \;\le\; Cm^{1/2-\eta_0/4}.$$

This remaining term can be further simplified by splitting out several of the inhomogeneous parts of $(G^*G)^{-1}$. Define $Q = Z^*_{J^c,\bullet}Z_{J^c,\bullet} + \left[\begin{smallmatrix} 0 & 0 \\ 0 & \mathtt{I} \end{smallmatrix}\right] = \left[\begin{smallmatrix} Z_1^*Z_1 & Z_1^*Z_2 \\ Z_2^*Z_1 & Z_2^*Z_2+\mathtt{I} \end{smallmatrix}\right] \in \mathbb{R}^{n\times n}$. Similarly, let $y = Z^*_{J^c,\bullet}\mu_{J^c} \in \mathbb{R}^n$. In terms of these variables, $G^*G = Q + y1^* + 1y^* + \alpha 11^*$. Applying the matrix inversion lemma,

$$(G^*G)^{-1} = Q^{-1} - Q^{-1/2}M\Xi M^*Q^{-1/2}, \tag{41}$$

where $M = \left[\begin{smallmatrix} \frac{Q^{-1/2}1}{\|Q^{-1/2}1\|_2} & \frac{Q^{-1/2}y}{\|Q^{-1/2}y\|_2} \end{smallmatrix}\right] \in \mathbb{R}^{n\times 2}$, and Ξ is an appropriate 2×2 matrix. Since $\vartheta \doteq Z^*_{J,\bullet}\,\sigma \in \mathbb{R}^n$ is iid $N(0,\nu^2\rho)$ independent of G, with overwhelming probability it is almost orthogonal to the rank-2 perturbation $\Gamma \doteq Q^{-1/2}M\Xi M^*Q^{-1/2}$: $P\left[\|\pi_\Gamma\vartheta\| \ge m^{1/2-\eta_0/4}\right] \asymp e^{-Cm^{1-\eta_0/2}}$. Since furthermore $\|\Gamma\| \le \|(G^*G)^{-1}\| + \|Q^{-1}\| \le C_G + \frac{4}{\nu^2\bar\rho}$ is bounded by a constant,

$$\left\|\left[\begin{smallmatrix} Z_1 & Z_2 \\ 0 & \mathtt{I} \end{smallmatrix}\right]\Gamma\vartheta\right\| \quad\le\quad \left(1+2\nu^2(\sqrt{\bar\rho}+\sqrt{\delta})^2\right)\left(C_G + \frac{4}{\nu^2\bar\rho}\right) m^{1/2-\eta_0/4}$$

and the remaining part of q_0 is

$$\left[\begin{smallmatrix} Z_1 & Z_2 \\ 0 & \mathtt{I} \end{smallmatrix}\right] Q^{-1}\vartheta \;=\; \left[\begin{smallmatrix} Z_1 \\ 0 \end{smallmatrix}\right]\left[Q^{-1}\right]_{I,\bullet}\vartheta \;+\; \left[\begin{smallmatrix} Z_2 \\ \mathtt{I} \end{smallmatrix}\right]\left[Q^{-1}\right]_{I^c,I}\vartheta_I \;+\; \left[\begin{smallmatrix} Z_2 \\ \mathtt{I} \end{smallmatrix}\right]\left[Q^{-1}\right]_{I^c,I^c}\vartheta_{I^c}.$$

The first two terms involve projections of ϑ onto k_1-dimensional subspaces, and hence are of lower order. That is, for $\Sigma \doteq \mathrm{null}([Q^{-1}]_{I,\bullet})^\perp$, we have $P\left[\|\pi_\Sigma\vartheta\|_2 \ge m^{1/2-\eta_0/4}\right] \asymp e^{-Cm^{1-\eta_0/2}}$. Since $\|Z_1\|$ and $\|Q^{-1}\|$ are bounded by constants with overwhelming probability, with probability at least $1-e^{-Cm^{1-\eta_0/2}(1+o(1))}$, $\left\|\left[\begin{smallmatrix} Z_1 \\ 0 \end{smallmatrix}\right]\left[Q^{-1}\right]_{I,\bullet}\vartheta\right\| \le C'm^{1/2-\eta_0/4}$. Identical reasoning shows that on the complement of a bad event of probability $\asymp e^{-Cm^{1-\eta_0/2}}$, $\left\|\left[\begin{smallmatrix} Z_2 \\ \mathtt{I} \end{smallmatrix}\right]\left[Q^{-1}\right]_{I^c,I}\vartheta_I\right\| \le C''m^{1/2-\eta_0/4}$.

This leaves $\left[\begin{smallmatrix} Z_2 \\ \mathtt{I} \end{smallmatrix}\right]\left[Q^{-1}\right]_{I^c,I^c}\vartheta_{I^c}$. Expressing Q as $\left[\begin{smallmatrix} U & V^* \\ V & W \end{smallmatrix}\right]$ and applying the Schur complement formula gives $[Q^{-1}]_{I^c,I^c} \;=\; W^{-1}+W^{-1}V(U^{-1}-V^*W^{-1}V)^{-1}V^*W^{-1}$, where $W = Z_2^*Z_2+\mathtt{I}$, $V = Z_2^*Z_1$, and $U = Z_1^*Z_1$. Because $W \succeq \mathtt{I}$, $\|W^{-1}\| \le 1$. With probability at least $1-e^{-Cm(1+o(1))}$, $\|U\| = \|Z_1\|^2 \le 2\nu^2\bar\rho$, $\sigma_{min}(U) \ge \frac{\nu^2\bar\rho}{2}$, and $\|V\| \le \|Z_1\|\|Z_2\| \le 2\,\nu^2\,(\sqrt{\bar\rho\delta}+\bar\rho)$ and so

$$\left\|W^{-1}V(U^{-1}-V^*W^{-1}V)^{-1}V^*W^{-1}\right\| \;\le\; \frac{\|W^{-1}\|^2\|V\|^2}{\sigma_{min}(U^{-1}) - \|V\|^2\|W^{-1}\|} \;\le\; \frac{8\nu^6(1+\sqrt{\bar\rho\delta})^2}{1-8\nu^6(1+\sqrt{\bar\rho\delta})^2}$$

is bounded by a constant. Let Σ' denote the k_1-dimensional range of this matrix. With probability $\ge 1-e^{-Cm^{1-\eta_0/2}(1+o(1))}$, $\|\pi_{\Sigma'}\vartheta\| \le m^{1/2-\eta_0/4}$, and so

$$\left\|\left[\begin{smallmatrix} Z_2 \\ \mathtt{I} \end{smallmatrix}\right]W^{-1}V(U^{-1}-V^*W^{-1}V)^{-1}V^*W^{-1}\vartheta\right\| \;\le\; C'''\,m^{1/2+\eta_0/4},$$

leaving only $\hat{q}_0 \doteq \left[\begin{smallmatrix} Z_2 \\ \mathtt{I} \end{smallmatrix}\right] (Z_2^* Z_2 + \mathtt{I})^{-1} \vartheta_{I^c}$. With probability at least $1 - e^{-Cm(1+o(1))}$, $\|\vartheta_{I^c}\| \leq \sqrt{2}\,\nu\sqrt{\rho\delta}\,m^{1/2}$, and so

$$\|\hat{q}_0\|_2 \;\leq\; \left\|\left[\begin{smallmatrix} Z_2 \\ \mathtt{I} \end{smallmatrix}\right]\right\| \|\vartheta_{I^c}\| \;\leq\; \sqrt{1+\|Z_2\|_2^2}\,\|\vartheta_{I^c}\| \;\leq\; \nu\sqrt{2\,\delta\,\rho\left(1+2\nu^2\left(\sqrt{\delta}+\sqrt{\bar{\rho}}\right)^2\right)}\,m^{1/2} \tag{42}$$

establishing the first part of the lemma.

For the second part, will bound the upper (R_1) and lower (R_2) parts of $\hat{q}_0$ elementwise with a pair of iid Gaussian vectors, and then argue that the Lipschitz function $\|\theta\cdot\|$ is concentrated about its (very small) expectation. For the upper block, write $Z_2 = QR$, where $Q \in \mathbb{R}^{\bar{\rho}m \times \bar{\rho}m}$ is an orthogonal matrix, and $R \in \mathbb{R}^{\bar{\rho}m \times \delta m - k_1}$ is an upper-triangular matrix with non-negative elements on the diagonal. With probability one (as long as $\mathrm{rank}(Z_2) = \bar{\rho}m$), Q and R are uniquely determined by Z_2. Moreover, Q is a uniform random orthogonal matrix, probabilistically independent of R.[7] Since $\hat{q}_0(R_1) = QR\,(R^*R + \mathtt{I})^{-1}\vartheta_{I^c}$ is the product of a uniform random orthogonal matrix and an independent vector $R(R^*R+\mathtt{I})^{-1}\vartheta_{I^c}$, $\frac{\hat{q}_0(R_1)}{\|\hat{q}_0(R_1)\|}$ is uniformly distributed on $\mathbb{S}^{\bar{\rho}m-1}$. With probability $\geq 1 - e^{-Cm(1+o(1))}$, $\|q_0(R_1)\| \leq \|Z_2\|\|\vartheta_{I^c}\| \leq 2\,\nu^2(\delta+\sqrt{\delta})\,m^{1/2}$. Introduce an independent random variable λ_1 distributed as the norm of a $(\bar{\rho}m)$-dimensional iid $N(0,\sigma^2)$ vector with $\sigma = \frac{4\,\nu^2\left(\delta+\sqrt{\delta}\right)}{\sqrt{\bar{\rho}}}$ (i.e., an appropriately scaled $\chi_{\bar{\rho}m}$ rv), and define

$$\phi_1 \doteq \lambda_1 \frac{\hat{q}_0(R_1)}{\|\hat{q}_0(R_1)\|}. \tag{43}$$

Since ϕ_1 is the product of a uniform random unit vector and an appropriate χ random variable, its distribution is iid $N(0,\sigma^2)$. With probability $1 - e^{-Cm(1+o(1))}$, $\|\phi_1\| \geq \sigma\sqrt{\bar{\rho}m} \geq \|\hat{q}_0(R_1)\|$, so ϕ_1 dominates $\hat{q}_0(R_1)$ elementwise and $\|\theta\phi_1\| \geq \|\theta\hat{q}_0(R_1)\|$. Applying Lemma 8, with probability $1 - e^{-Cm(1+o(1))}$,

$$\|\theta\phi_1\|_2 \;\leq\; 4\exp\left(-\frac{1}{16\sigma^2}\right)\sqrt{\bar{\rho}m} \;=\; 4\sqrt{\bar{\rho}}\,\exp\left(-\frac{\bar{\rho}}{256\,\nu^4(\delta+\sqrt{\delta})^2}\right)m^{1/2} \tag{44}$$

For the lower (R_2) coordinates, write $Z_2^* = \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right]\left[\begin{array}{c} R_1 \\ 0 \end{array}\right] \doteq QR$ where $R_1 \in \mathbb{R}^{\bar{\rho}m \times \bar{\rho}m}$ is an upper triangular matrix with nonnegative diagonal elements, Q_1 is an orthogonal matrix, and Q_2 is a random orthobasis for $\mathcal{R}(Q_1)^\perp$ (so that $Q \in \mathbb{R}^{n-k_1 \times n-k_1}$ is an orthogonal matrix). Again from the rotational invariance of the Gaussian distribution, Q is a uniform random orthogonal matrix, independent of R, and

$$\hat{q}_0(R_2) = (Z_2^* Z_2 + \mathtt{I})^{-1}\vartheta_{I^c} = Q(RR^* + \mathtt{I})^{-1}Q^*\vartheta_{I^c} \doteq Q(RR^*+\mathtt{I})^{-1}\gamma, \tag{45}$$

[7]This follows from the rotational invariance of the Gaussian distribution: left multiplication by an independent orthogonal matrix sampled according to the invariant measure yields an independent pair $Q'R = Z_2' \equiv_d Z_2$.

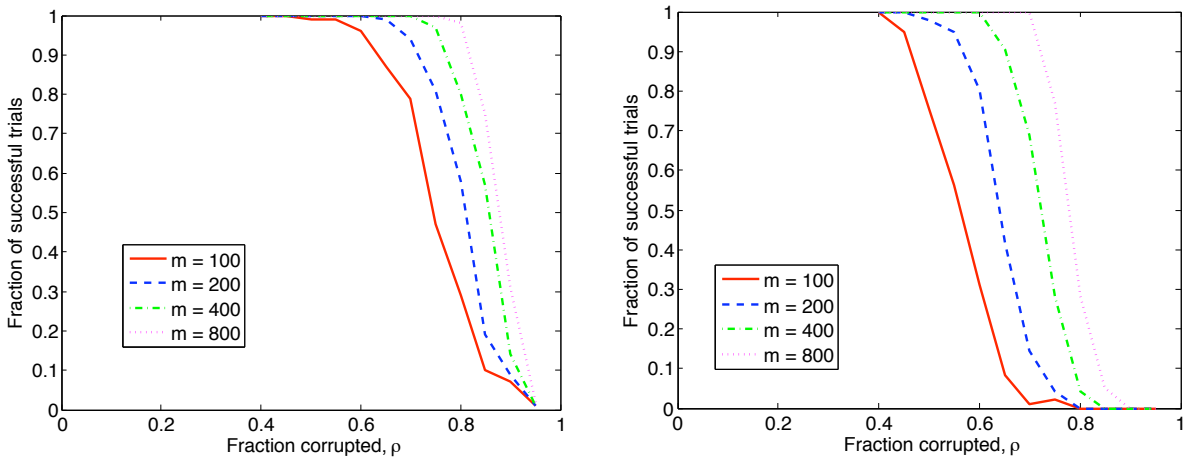


Fig. 5. Error correction in weak proportional growth. We fix $\delta = .5$, $\nu^2 = .4$, and plot the fraction of successful recoveries as a function of the error density ρ, for each $m = 100, 200, 400, 800$. At left, k_1 is fixed at 1; at right, $k_1 = 5$. In both cases, as m increases, the fraction of errors that can be corrected approaches 1.

where $\gamma \doteq Q^* \vartheta_{I^c}$ is an iid $N(0, \nu^2\rho)$ random vector, *independent of* Q. Hence, $\hat{q}_0(R_2)$ is the product of a uniform random orthogonal matrix Q, and a probabilistically independent vector $(RR^* + \mathtt{I})^{-1}\gamma$, and its orientation $\frac{\hat{q}_0(R_2)}{\|\hat{q}_0(R_2)\|}$ is a uniform random vector on $\mathbb{S}^{n-k_1-1}$. As above, introduce an independent random variable λ_2 distributed as the norm of an $(n-k_1)$-dimensional iid $N(0, 4\nu^2\rho)$ random vector, and define

$$\phi_2 = \lambda_2 \frac{\hat{q}_0(R_2)}{\|\hat{q}_0(R_2)\|}. \tag{46}$$

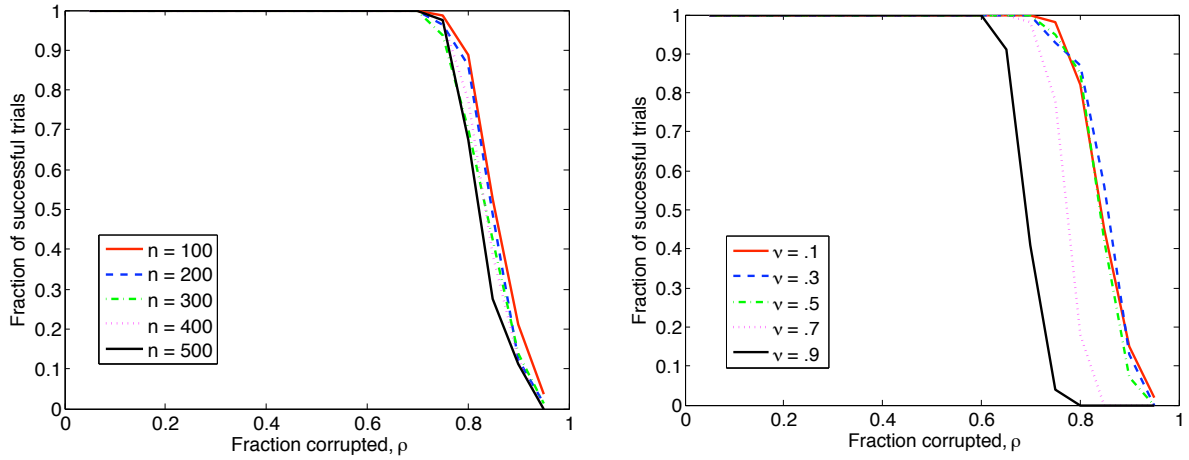
The product of an independent unit vector and (appropriately scaled) χ_{n-k_1} scalar, ϕ_2 is distributed as an iid $N(0, 4\nu^2\rho)$ vector. With probability at least $1 - e^{-Cm(1+o(1))}$, $\|\phi_2\| \geq \sqrt{2}\nu\sqrt{\rho}\sqrt{n-k_1}$, and $\|\hat{q}_0(R_2)\| \leq \|\vartheta_{I^c}\| \leq \sqrt{2}\nu\sqrt{\rho}\sqrt{n-k_1}$. Therefore, ϕ_2 dominates $\hat{q}_0(R_2)$ elementwise, and $\|\theta\phi_2\| \geq \|\theta\hat{q}_0(R_2)\|$. By Lemma 8,

$$\|\theta\phi_2\|_2 \;\leq\; 4\sqrt{\delta}\exp\left(-\frac{1}{64\nu^2\rho}\right) m^{1/2}. \tag{47}$$

Combining the bounds on $\|\theta\phi_1\|$ and $\|\theta\phi_2\|$ gives the second part of the lemma. ■

III. Simulations and Experiments

In this section, we perform simulations verifying the basic phenomenon of Theorem 1, and investigating the effect of various model parameters on the error correction capability of the ℓ^1-minimization.



Effect of varying n. Effect of varying ν.

Fig. 6. 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.

a) Error correction capacity: We first verify the ability of ℓ^1-minimization to correct increasingly large fractions of random errors in the weak proportional growth setting. We generate problem instances with $\delta = 1/2$, $\nu^2 = .4$, for varying $m = 100, 200, 400, 800$. For each problem size, and for each error fraction $\rho = 0.05, 0.1, \ldots, 0.95$, we generate 100 random problems, and plot the fraction of correct recoveries in Figure 5. At left, we set $k_1 = 1$, while at right, $k_1 = 5$. In both cases, as m increases, the fraction of errors that can be corrected also increases.

b) Varying model parameters.: We next investigate the effect of varying δ (Figure 6 left) and ν (Figure 6 right). We first fix $m = 400$, $\nu = .3$, and consider different bouquet sizes $n = 100, 200, 300, 400, 500$. Figure 6 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.

We next fix $m = 400$, $n = 200$, and consider the effect of varying ν. Figure 6 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.

c) Phase transition in total proportional growth: Theorem 1 does not provide any explicit information about the behavior of ℓ^1-minimization when the signal support k_1 in proportion to m: $k_1/m \to \rho_1 \in$

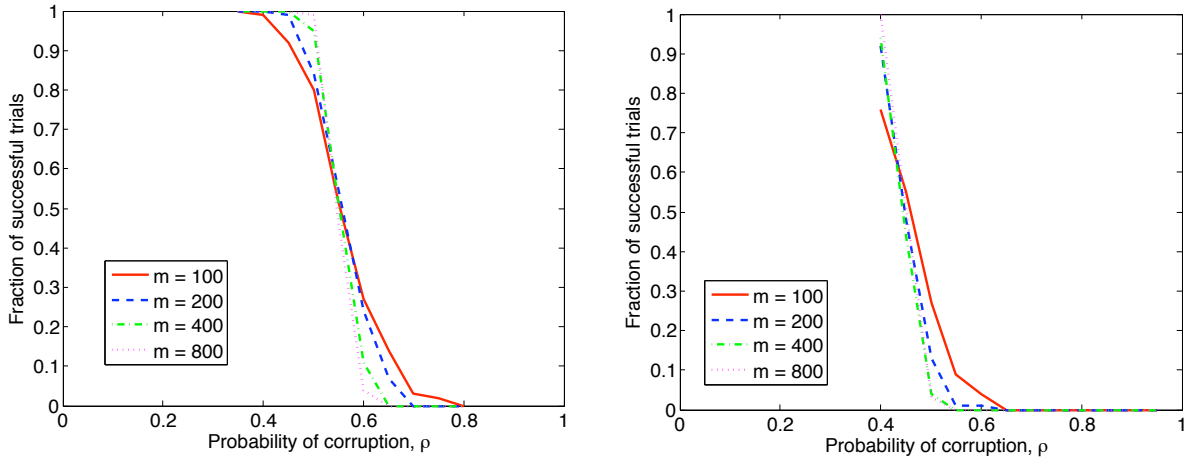


Fig. 7. Phase transition in total proportional growth. When the signal support grows in proportion to the dimension ($k_1/m \to \rho_1 \in (0,1)$), we observe an asymptotically sharp phase transition in the probability of correct recovery, similar to that investigated in [7]. Left: $\rho = 0.05$. Right: $\rho = 0.1$.

$(0,1)$. Based on intuition from the study of more homogeneous polytopes (especially the work of Donoho and Tanner on the Gaussian ensemble [7]), we might expect that when k_1 also exhibits proportional growth, an asymptotically sharp phase transition between guaranteed recovery and guaranteed failure will occur at some critical error fraction $\rho^* \in (0,1)$. Simulations suggest that this is the case, as shown in Figure 7.

IV. Discussions and Future Work

d) 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 shows 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 corresponding incoherence. More generally, for any

given distribution of sparsity, one may ask the question whether there exists an optimal dictionary with corresponding incoherence such that ℓ^1-minimization has the highest chance of success.

e) Stability with respect to noise: Although in our model, we do not explicitly consider any noise (say $y = Ax + e + z$ where z is Gaussian noise), it is known that ℓ^1-minimization is stable to small noise [13]. This is also what we have observed empirically in our simulations or in the experiments with the face images: ℓ^1-minimization for the cross-and-bouquet model is surprisingly stable 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 z will be treated like errors and be absorbed into e. So the estimate of x is likely to be affected only by noises with small magnitude. However, a more precise characterization of the effect of noise (say Gaussian) on the estimate of the sparse signal x and the error e remains an open problem.

f) Neighborliness of polytopes: From our study, we see that in order to precisely characterize the performance of ℓ^1-minimization, one needs to analyze the geometry of the type of polytopes associated with the special 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. The polytope associated with such a dictionary can be very different from what we have normally studied in signal processing or coding theory, so will be the performance of ℓ^1-minimization for finding the desired representation, either sparse or not. Thus, we should expect that in coming years, many new classes of polytopes with interesting properties may arise from other applications and practical problems.

ACKNOWLEDGMENTS

This work was partially supported by the grants NSF CRS-EHS-0509151, NSF CCF-TF-0514955, ONR YIP N00014-05-1-0633, and NSF IIS 07-03756. John Wright also gratefully acknowledges a Microsoft Live Labs Fellowship.

REFERENCES

[1] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *To appear in IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[2] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Information Theory*, vol. 51, no. 12, 2005.

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

[4] D. Needell and J. Tropp, "CoSAMP: Iterative signal recovery from incomplete and inaccurate samples," *To appear in Applied and Computational Harmonic Analysis*, 2008.

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

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

[7] D. Donoho and J. Tanner, "Counting faces of randomly projected polytopes when the projection radically lowers dimension," preprint*, http://www.math.utah.edu/ tanner/*, 2007.

[8] ——, "Sparse nonnegative solution of underdetermined linear equations by linear programming," preprint, 2005.

[9] D. Donoho, "Neighborly polytopes and sparse solution of underdetermined linear equations," *preprint*, 2005.

[10] A. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *To appear in SIAM Review*, 2008.

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

[12] S. Dasgupta, D. Hsu, and N. Verma, "A concentration theorem for projections," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

[13] D. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ^1-norm near solution approximates the sparest solution," *preprint*, 2004.

[14] P. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, pp. 99–111, 1972.

[15] N. Alon and J. Spencer, *The Probabilistic Method*. Wiley-Interscience, 2001.

APPENDIX
TECHNICAL LEMMAS AND RESULTS

A. Restricted Isometry for Sparse Vectors

In this appendix, we prove Lemma 4 of Section II-D, which states that as long as c is sufficiently small, i.e., $c < \min\left(\frac{\bar\rho}{64\,(1+2C_\mu)^2}, \frac{\bar\rho}{1024}\right)$ and $\bar\rho H(c/\bar\rho) + \delta H(c/\delta) < \bar\rho/512$, with overwhelming probability

$$\gamma_{cm}\left([\tilde V^* \ -\tilde S\tilde U^*]\right) \ \geq \ \frac{\nu\sqrt{\bar\rho}}{4} - 6\nu\sqrt{c}.$$

Proof: Recall that $\tilde U\tilde S\tilde V^*$ is a compact singular value decomposition of $\pi_{u_1^\perp} P \pi_{v_1^\perp}$, where $P = M_2^*\pi_{M_1^\perp}$ and u_1, v_1 are its leading singular vectors. Notice that the conditional distribution of P given M_1 is Gaussian: $P = Z_2^*\pi_{M_1^\perp} + 1\mu_{J^c}^*\pi_{M_1^\perp} \doteq Z_2^*\pi_{M_1^\perp} + 1\hat\mu^*$. We will argue that the second term dominates.

a) $1\hat\mu^*$ determines the leading singular vectors: Since the columns of M_1 are k_1 small perturbations of μ_{J^c}, the distance $\|\hat\mu\| = \|\pi_{M_1^\perp}\mu_{J^c}\|$ of μ_{J^c} to $\mathcal{R}(M_1)$ should be small. However, we will argue that $\|\pi_{M_1^\perp}\mu_{J^c}\|$ is at least $\Omega(k_1^{-1/2})$. Choose an orthonormal basis whose first element is $\frac{\mu_{J^c}}{\|\mu_{J^c}\|}$. With respect to this basis, M_1 can be expressed as $\left[\begin{smallmatrix}0\\ \tilde Z_2\end{smallmatrix}\right] + e_1(\tilde z_1^* + \|\mu_{J^c}\|1^*) \doteq \left[\begin{smallmatrix}0\\ \tilde Z_2\end{smallmatrix}\right] + e_1 v^*$, where $\tilde z_1$ and $\tilde Z_2$ are iid $N(0,\nu^2/m)$. Then $\left\|\pi_{M_1}\frac{\mu_{J^c}}{\|\mu_{J^c}\|}\right\|_2^2$ is

$$e_1^*\left(\left[\begin{smallmatrix}0\\ \tilde Z_2\end{smallmatrix}\right] + e_1 v^*\right)\left(vv^* + \tilde Z_2^*\tilde Z_2\right)^{-1}\left([\,0\ \tilde z_2^*\,] + ve_1^*\right)e_1 \quad = \quad \frac{v^*(\tilde Z_2^*\tilde Z_2)^{-1}v}{1 + v^*(\tilde Z_2^*\tilde Z_2)^{-1}v}.$$

Applying Fact 1 to the $\bar\rho m - 1 \times k_1$ $N(0,\nu^2/m)$ matrix $\tilde Z_2$, one can easily show that $P\left[\left\|(\tilde Z_2^*\tilde Z_2)^{-1}\right\| > \frac{2}{\nu^2\bar\rho}\right] \asymp e^{-Cm}$. The norm of the k_1-dimensional $N(0,\nu^2/m)$ vector $\tilde z_1$ also concentrates[8]: $P\left[\|\tilde z_1\| > \sqrt{k_1}\right] \asymp e^{-C'mk_1}$. On the complement of these bad events, $\|v\| \leq \|\tilde z_1\| + \|\mu_{J^c}1\| = (1 + \|\mu_{J^c}\|)\sqrt{k_1} \leq 2\sqrt{k_1}$, and $v^*(\tilde Z_2^*\tilde Z_2)^{-1}v \leq \frac{8}{\nu^2\bar\rho}k_1$. Moreover,

$$\left\|\frac{\mu_{J^c}}{\|\mu_{J^c}\|} - \pi_{M_1}\frac{\mu_{J^c}}{\|\mu_{J^c}\|}\right\|_2^2 \quad = \quad \frac{1}{1 + v^*(\tilde Z_2^*\tilde Z_2)^{-1}v} \quad \geq \quad \frac{1}{1 + \frac{8}{\nu^2\bar\rho}k_1}. \tag{48}$$

Lemma 6 below shows if $\bar\rho < \frac{1}{2} - \frac{1}{C_\mu}$, with probability $\geq 1 - e^{-Cm(1+o(1))}$ in the random support of the error e, $\|\mu_{J^c}\| \geq \frac{\bar\rho}{2C_\mu}$.[9] Together with (48), this implies that $\|\hat\mu\| = \|\mu_{J^c} - \pi_{M_1}\mu_{J^c}\|_2 \geq \frac{\bar\rho}{4C_\mu^2}\sqrt{\frac{1}{1+\frac{8}{\nu^2\bar\rho}k_1}}$. Since $\|1\|_2 = \sqrt{\delta m - k_1}$, on this good event $\|1\hat\mu^*\|_2 \geq C_1\, m^{\eta_0/2}$ for some constant C_1 and m sufficiently large. From Fact 1, $\|Z_2\|$ is bounded by some constant C_2 with probability at least $1 - e^{-Cm(1+o(1))}$.

[8]For a d-dimensional iid $N(0,\sigma^2)$ vector x, $P[\|x\| \geq \beta\sigma\sqrt{d}] \leq e^{-(\beta-1)^2 d/2}$ [11]. To obtain the result above, set $\beta = \sqrt{m}/\nu$.

[9]This follows because demanding that $\|\mu\|_2 = 1$ and $\|\mu\|_\infty \leq C_\mu m^{-1/2}$ forces μ to spread over its coordinates; the probability that a randomly chosen support misses almost all of the energy of μ is overwhelmingly small.

Treating $Z_2^* \pi_{M_1^\perp}$ as a nuisance perturbation of $1\hat{\mu}^*$ and applying Wedin's perturbation bound for principal subspaces [14] then gives

$$\|\pi_{u_1^\perp} - \pi_{1^\perp}\| \;\leq\; 4 \sin \angle(u_1, 1) \;\leq\; 4 \frac{\|Z_2^* \pi_{M_1^\perp}\|}{\|1\hat{\mu}^*\|} \;\leq\; \frac{4\, C_2}{C_1\, m^{\eta_0/2}}.$$

and similarly $\|\pi_{v_1^\perp} - \pi_{\hat{\mu}^\perp}\| \leq \frac{4\, C_2}{C_1\, m^{\eta_0/2}}$.

$$\|\pi_{u_1}^\perp P \pi_{v_1}^\perp - \pi_{1^\perp} P \pi_{\hat{\mu}^\perp}\| \;\;\leq\;\; \|\pi_{u_1^\perp} - \pi_{1^\perp}\| \|P \pi_{v_1^\perp}\| + \|\pi_{1^\perp} P\| \|\pi_{v_1^\perp} - \pi_{\hat{\mu}^\perp}\|$$

Now, $\|\pi_{1^\perp} P\| \leq \|Z_2\| \leq C_2$, and $\|P \pi_{v_1^\perp}\| = \sigma_2(P) \leq \sqrt{2}\nu(\sqrt{\bar{\rho}} + \sqrt{\delta})$ simultaneously with probability $\geq 1 - e^{-Cm(1+o(1))}$ (the second bound was established in part (a) of the proof of Lemma 3). Hence, $\exists\, C_3$ such that $P\left[\, \|\pi_{u_1^\perp} P \pi_{v_1^\perp} - \pi_{1^\perp} P \pi_{\hat{\mu}^\perp}\|_2 \;>\; C_3\, m^{-\eta_0/2} \,\right] \asymp e^{-Cm}$. For an arbitrary matrix W, write $f(W) = \gamma_{cm}([\pi_{\mathcal{R}(W^*)} \; - W^*])$. We are interested in $f(\pi_{u_1^\perp} P \pi_{v_1^\perp})$.[10] Using the fact that singular values of submatrices are Lipschitz and applying Wedin's $\sin\Theta$ theorem [14] to $\pi_{R(W^*)}$, it is not difficult to show that if $\mathrm{rank}(W + \Delta) = \mathrm{rank}(W)$,

$$|\, f(W + \Delta) - f(W)\,| \;\leq\; \left(1 + \frac{2}{\sigma_{min}(W) - \|\Delta\|}\right) \|\Delta\|, \tag{49}$$

where $\sigma_{min}(W)$ is the smallest nonzero singular value. Applying this bound with $W = \pi_{u_1^\perp} P \pi_{v_1^\perp}$, $\Delta \doteq \pi_{u_1^\perp} P \pi_{v_1^\perp} - \pi_{1^\perp} P \pi_{\hat{\mu}^\perp}$, and noticing that $\sigma_{min}(\pi_{u_1^\perp} P \pi_{v_1^\perp})$ is bounded below by a positive constant with overwhelming probability, we have that $\left|\, f\left(\pi_{u_1^\perp} P \pi_{v_1^\perp}\right) - f\left(\pi_{1^\perp} P \pi_{\hat{\mu}^\perp}\right) \,\right| < \frac{\nu\sqrt{\bar{\rho}}}{16}$ with probability at least $1 - e^{-Cm(1+o(1))}$. We henceforth restrict our attention to $f(\pi_{1^\perp} P \pi_{\hat{\mu}^\perp})$.

b) Analysis via Gaussian measure concentration: Let Σ denote the subspace $(\mathcal{R}(Z_1) + \mathcal{R}(\mu_{J^c}))^\perp$, and let V_0 be some orthonormal basis for this subspace, chosen independently of Z_2. From the above reasoning, we can restrict our attention to $\pi_{1^\perp} P \pi_{\hat{\mu}^\perp} = \pi_{1^\perp} Z_2 \pi_\Sigma$. Let $\pi_{1^\perp} Z_2 \pi_\Sigma = U' S' V'^*$ be a compact singular value decomposition of this matrix. Then,

$$\gamma_{cm}\left(\left[\begin{array}{cc} V'^* & -S'U'^* \end{array}\right]\right) \;\;=\;\; \gamma_{cm}\left(V'^* \left[\begin{array}{cc} \mathtt{I} & \pi_\Sigma Z_2^* \pi_{1^\perp} \end{array}\right]\right) \;\;=\;\; \gamma_{cm}\left(V_0^* \left[\begin{array}{cc} \mathtt{I} & \pi_\Sigma Z_2^* \pi_{1^\perp} \end{array}\right]\right).$$

Where the final step follows because γ_{cm} is invariant under left multiplication of its argument by an orthogonal matrix. Now, $V_0^* \pi_\Sigma Z_2^* = V_0^* Z_2^*$ is simply distributed as a $\bar{\rho}m - k_1 - 1 \times \delta m - k_1$ iid $N(0, \nu^2/m)$ random matrix. Finally, introduce an additional uniformly distributed random orthogonal matrix $Q \in \mathbb{R}^{\bar{\rho}m - k_1 - 1 \times \bar{\rho}m - k_1 - 1}$, chosen independently of Z_2, and define $\Psi \doteq Q V_0^* \pi_\Sigma Z_2^*$. This is again an iid $N(0, \nu^2/m)$ matrix. From the rotational invariance of the Gaussian distribution it is easy to show that Ψ and Q are independent random variables. Hence, $\gamma_{cm}\left(\left[\begin{array}{cc} V'^* & -S'U'^* \end{array}\right]\right) = \gamma_{cm}\left(\left[\begin{array}{cc} QV_0^* & \Psi\pi_{1^\perp} \end{array}\right]\right)$.

[10] Since left multiplication by an orthogonal matrix does not change γ_{cm}, $f(\pi_{u_1^\perp} P \pi_{v_1^\perp}) = \gamma_{cm}([\tilde{V}^* \; - \tilde{S}\tilde{U}^*])$.

Here, QV_0^* is the transpose of random orthobasis for the subspace Σ; it can be realized by orthogonalizing the projection of a Gaussian matrix onto Σ. To this end, introduce an iid $N(0,\nu^2/m)$ matrix $\Phi \in \mathbb{R}^{\bar{\rho}m-k_1-1\times\bar{\rho}m}$ independent of Σ and Ψ. Then, $\gamma_{cm}\left(\left[\begin{array}{cc} QV_0^* & \Psi\pi_{1^\perp}\end{array}\right]\right)$ is equal in distribution to $\gamma_{cm}\left(\left[\begin{array}{cc} (\Phi\pi_\Sigma\Phi^*)^{-1/2}\,\Phi\pi_\Sigma & \Psi\pi_{1^\perp}\end{array}\right]\right)$. This "transfer to Gaussianity" makes it easier to provide bounds on γ_{cm}. Let $\Lambda \doteq (\Phi\pi_\Sigma\Phi^*)^{-1/2}$. Now,

$$\begin{aligned}\gamma_{cm} \;&=\; \min_{\#L_1\cup L_2=cm}\sigma_{min}\left(\left[\begin{array}{cc}[\Lambda\Phi\pi_\Sigma]_{\bullet,L_1} & [\Psi\pi_{1^\perp}]_{\bullet,L_2}\end{array}\right]\right) \;\geq\\ &\min_{\#L_1=\#L_2=cm}\min\Big\{\sigma_{min}([\Lambda\Phi\pi_\Sigma]_{\bullet,L_1}),\,\sigma_{min}(\pi_{\Sigma'^\perp}[\Psi\pi_{1^\perp}]_{\bullet,L_2})\Big\} \;-\; \max_{\#L_1=\#L_2=cm}\left\|\pi_{\Sigma'}\,[\Psi\pi_{1^\perp}]_{\bullet,L_2}\right\|\end{aligned}$$

where Σ' denotes the subspace $\mathcal{R}([\Lambda\Phi\pi_\Sigma]_{\bullet,L_1})$.

c) Bounding $\sigma_{min}[\Lambda\Phi\pi_\Sigma]_{\bullet,L}$: Applying Fact 1 to $\Phi\pi_\Sigma$ gives that $P\left[\,\|\Phi\pi_\Sigma\|_2 \geq 3\nu\sqrt{\bar{\rho}}\,\right] \asymp e^{-\bar{\rho}m/2}$. On the complement of this bad event, $\sigma_{min}(\Lambda) \geq \frac{1}{3\nu\sqrt{\bar{\rho}}}$. Write

$$\begin{aligned}[\Phi\pi_\Sigma]_{\bullet,L} \;&=\; \Phi_{\bullet,L} - [\Phi\pi_{\Sigma^\perp}]_{\bullet,L} \;=\; \Phi_{\bullet,L}(\mathtt{I}-[\pi_{\Sigma^\perp}]_{L,L}) - \Phi_{\bullet,L^c}[\pi_\Sigma]_{L^c,L}\\ \Longrightarrow\quad \sigma_{min}([\Phi\pi_\Sigma]_{\bullet,L}) \;&\geq\; \sigma_{min}(\Phi_{\bullet,L})(1-\|[\pi_{\Sigma^\perp}]_{L,L}\|) - \|\pi_{\Phi_{\bullet,L}}\Phi_{\bullet,L^c}\pi_{[\pi_\Sigma]_{L^c,L}}\|.\end{aligned}$$

Straightforward application of Fact 1 shows that $P\left[\sigma_{min}(\Phi_{\bullet,L}) \leq \frac{\nu\sqrt{\bar{\rho}}}{2} - \nu\sqrt{c}\right] \asymp e^{-\bar{\rho}m/8}$, while for any[11] $\varepsilon_1 > 0$, $P\left[\|\pi_{\Phi_{\bullet,L}}\Phi_{\bullet,L^c}\pi_{[\pi_\Sigma]_{L^c,L}}\| \geq 2\nu\sqrt{c}+\nu\sqrt{\bar{\rho}\varepsilon_1}\right] \asymp e^{-\bar{\rho}\varepsilon_1 m/2}$. Finally, consider the matrix $\Upsilon \doteq \left[\begin{array}{cc} Z_1 & \nu\sqrt{\bar{\rho}}\frac{\mu_{J^c}}{\|\mu_{J^c}\|}\end{array}\right] \in \mathbb{R}^{\bar{\rho}m\times k_1+1}$. We are interested in $\|[\pi_{\Sigma^\perp}]_{L,L}\| \;=\; \left\|\Upsilon_{L,\bullet}(\Upsilon^*\Upsilon)^{-1}\Upsilon^*_{\bullet,L}\right\| \leq \frac{\|\Upsilon_{L,\bullet}\|^2}{\sigma^2_{min}(\Upsilon)}$. It is not difficult to show[12] that with probability at least $1-e^{-\frac{\bar{\rho}m}{8}(1+o(1))}$, $\sigma_{min}(\Upsilon) \geq \frac{\nu\sqrt{\bar{\rho}}}{2}$. Meanwhile for any $\varepsilon_2 > 0$, $P\left[\,\|Z_{1L,\bullet}\| \geq \nu\sqrt{c}+\nu\sqrt{\bar{\rho}\varepsilon_2}\right] \asymp e^{-\bar{\rho}\varepsilon_2 m/2}$. On the complement of this bad event (and invoking Lemma 6)

$$\|\Upsilon_{L,\bullet}\| \;\leq\; \|Z_{1L,\bullet}\| + \left\|\nu\sqrt{\bar{\rho}}\frac{\mu_{J^c}(L)}{\|\mu_{J^c}\|}\right\| \;\leq\; \nu\sqrt{c}+\nu\sqrt{\bar{\rho}\varepsilon_2}+\nu\sqrt{\bar{\rho}}\frac{C_\mu\sqrt{c}}{\sqrt{\bar{\rho}}/2C_\mu} \;=\; \nu\left(\sqrt{\bar{\rho}\varepsilon_2}+\sqrt{c}\,\left(1+2C_\mu^2\right)\right).$$

By the assumptions of the lemma, $\sqrt{c}\,(1+2C_\mu^2) \leq \sqrt{\bar{\rho}}/8$, and $\|[\pi_{\Sigma^\perp}]_{L,L}\| \leq \frac{\|\Upsilon_{L,\bullet}\|^2}{\sigma^2_{min}(\Upsilon)} \leq 4(\sqrt{\varepsilon_2}+1/8)^2$. Setting $\varepsilon_1 = \varepsilon_2 = \frac{1}{64}$

$$\sigma_{min}\left([\Phi\pi_\Sigma]_{\bullet,L}\right) \;\geq\; \left(\frac{\nu\sqrt{\bar{\rho}}}{2}-\nu\sqrt{c}\right)\left(1-\frac{1}{4}\right)-\left(2\,\nu\sqrt{c}+\frac{\nu\sqrt{\bar{\rho}}}{8}\right) \;=\; \frac{\nu\sqrt{\bar{\rho}}}{4}-\frac{11\nu\sqrt{c}}{4}, \qquad (50)$$

and $\sigma_{min}\left([\Lambda\Phi\pi_\Sigma]_{\bullet,L}\right) \geq \frac{1}{12}-\frac{11}{12}\sqrt{\frac{c}{\bar{\rho}}} > \frac{1}{24}$ on the complement of a bad event of probability $e^{-\frac{\bar{\rho}m}{128}(1+o(1))}$. The number of subsets L of size cm is $e^{\bar{\rho}mH(c/\bar{\rho})(1+o(1))}$. The probability any L is bad is bounded by

[11] Since $\Phi_{\bullet,L^c}$ is independent of $\Phi_{\bullet,L}$ and Σ, the norm of $\pi_{\Phi_{\bullet,L}}\Phi_{\bullet,L^c}\pi_{[\pi_\Sigma]_{L^c,L}}$ is simply distributed as the norm of a $cm\times cm$ iid $N(0,\nu^2/m)$ matrix. By Fact 1, $P\left[\|\pi_{\Phi_{\bullet,L}}\Phi_{\bullet,L^c}\pi_{[\pi_\Sigma]_{L^c,L}}\| \geq 2\nu\sqrt{c}+t\nu\sqrt{c}\right] \leq e^{-(t-o(1))^2cm/2}$. Set $t=\sqrt{\frac{\bar{\rho}\varepsilon_1}{c}}$.

[12] Write $\sigma_{min}(\Upsilon) \;\geq\; \sigma_{min}\left(\left[\begin{array}{cc}\pi_{\mu_{J^c}^\perp}Z_1 & \nu\sqrt{\bar{\rho}}\frac{\mu_{J^c}}{\|\mu_{J^c}\|}\end{array}\right]\right) - \|\pi_{\mu_{J^c}}Z_1\| \;\geq\; \min(\sigma_{min}(\pi_{\mu_{J^c}^\perp}Z_1),\nu\sqrt{\bar{\rho}}) - \|\pi_{\mu_{J^c}}Z_1\|$, apply Fact 1 to the singular value and standard tail bounds to the k_1 dimensional $N(0,\nu^2/m)$ vector $\mu^*_{J^c}Z_1$.

$e^{\bar{\rho}m\left(H(c/\bar{\rho})-\frac{1}{128}\right)(1+o(1))}$, which falls off exponentially when $H(c/\bar{\rho}) < 1/128$. This is guaranteed for $c/\bar{\rho} < 1/1024$.

d) Bounding $\sigma_{min}\left(\pi_{\Sigma'^{\perp}}[\Psi\pi_{1^{\perp}}]_{\bullet,L}\right)$: Choose any orthonormal basis for the $[(\bar{\rho}-c)m-k_1-1]$-dimensional subspace $\Sigma'^{\perp}$, where Σ' denotes the cm-dimensional range of $[\Lambda\Phi\pi_{\Sigma}]_{\bullet,L_1}$. The expression of the columns of $\pi_{\Sigma'^{\perp}}\Psi$ with respect to this basis is a $(\bar{\rho}-c)m-k_1-1 \times \delta m - k_1$ matrix $\tilde{\Psi}$ with entries $N(0,\nu^2/m)$. Split $\tilde{\Psi}\pi_{1^{\perp}}$ as

$$[\tilde{\Psi}\pi_{1^{\perp}}]_{\bullet,L} \;=\; \tilde{\Psi}_{\bullet,L} - \frac{1}{m}\tilde{\Psi}_{\bullet,L^c}11^* - \frac{1}{m}\tilde{\Psi}_{\bullet,L}11^*.$$

Using the independence of $\frac{1}{m}\tilde{\Psi}_{\bullet,L^c}1$ and $\tilde{\Psi}_{\bullet,L}$ and applying Fact 1, it is not difficult to show[13] that

$$P\left[\sigma_{min}\left(\tilde{\Psi}_{\bullet,L} - \frac{1}{m}\tilde{\Psi}_{\bullet,L^c}11^*\right) \le \frac{\nu\sqrt{\bar{\rho}-c}}{2} - \nu\sqrt{c}\right] \le e^{-\frac{(\bar{\rho}-c)m}{8}(1+o(1))}. \tag{51}$$

The final term involves $\frac{1}{\sqrt{m}}\tilde{\Psi}_{\bullet,L}1$, an iid $N(0,c\nu^2/m)$ vector of dimension $(\bar{\rho}-c)m-k_1-1$,

$$P\left[\left\|\frac{1}{\sqrt{m}}\tilde{\Psi}_{\bullet,L}1\right\| \;\ge\; 2\nu\sqrt{c(\bar{\rho}-c)}\right] \;\asymp\; e^{-\frac{(\bar{\rho}-c)m}{2}(1+o(1))}. \tag{52}$$

Combining these results, we have that on the complement of a bad event of probability $\asymp e^{-(\bar{\rho}-c)m/8}$,

$$\begin{aligned}\sigma_{min}([\tilde{\Psi}\pi_{1^{\perp}}]_{\bullet,L}) \;&\ge\; \sigma_{min}\left(\tilde{\Psi}_{\bullet,L} - \frac{1}{m}\tilde{\Psi}_{\bullet,L^c}11^*\right) - \left\|\frac{1}{m}\tilde{\Psi}_{\bullet,L}11^*\right\| \;\ge\; \frac{\nu\sqrt{\bar{\rho}-c}}{2} - \nu\sqrt{c} - 2\nu c\sqrt{\bar{\rho}-c} \\ &\ge\; \sqrt{\frac{1023}{1024}}\frac{\nu\sqrt{\bar{\rho}}}{2} - 3\nu\sqrt{c} \quad \text{by the assumptions of the lemma.}\end{aligned} \tag{53}$$

There are $\asymp e^{\bar{\rho}mH(c/\bar{\rho})}$ subsets L_1 of size cm and $\asymp e^{\delta mH(c/\delta)}$ subsets L_2 of size cm, where H denotes the (base-e) binary entropy function. The total number of choices of L_1, L_2 is asymptotic to $e^{\left(\bar{\rho}H\left(\frac{c}{\bar{\rho}}\right)+\delta H\left(\frac{c}{\delta}\right)\right)m}$, and the probability that any pair is bad is bounded by

$$\exp\Big(\big(\,\bar{\rho}H(c/\bar{\rho}) + \delta H(c/\delta) - \tfrac{\bar{\rho}-c}{8}\big)\,m\,(1+o(1))\Big).$$

Under the assumptions of the lemma, the exponent is negative. Since for $\nu < \frac{1}{24\sqrt{\bar{\rho}}}$, this bound is smaller than the bound for $\Lambda\Phi\pi_{\Sigma}$, this bound controls the overall behavior.

[13]Independent translations do not substantially affect Fact 1: for an $m \times n$ iid $N(0,1/m)$ matrix M and an independent translation x, $\sigma_{min}(M + x1^*) \ge \sigma_{min}(\pi_{x^{\perp}}M)$, which obeys the same concentration result, now applied to an $(m-1)\times n$ matrix. Appropriate rescaling of the $(\bar{\rho}-c)m-k_1-1 \times cm$ $N(0,\nu^2/m)$ matrix $\tilde{\Psi}_{\bullet,L}$ yields the desired expression.

e) Bounding the cross-coherence $\left\|\pi_{\Sigma''}\left[\Psi\pi_{1^\perp}\right]_{\bullet,L_2}\right\|$*:* Begin by fixing L_1 and L_2. Let Σ'' denote the subspace $\mathcal{R}([\Lambda\Phi\pi_\Sigma]_{\bullet,L_1})$. Notice that Σ'' and Ψ are probabilistically independent. Now,

$$\left\|\pi_{\Sigma''}\left[\Psi\pi_{1^\perp}\right]_{\bullet,L_2}\right\| \;\le\; \|\pi_{\Sigma''}\Psi_{\bullet,L_2}\| + \left\|\frac{1}{\sqrt{m}}\pi_{\Sigma''}\Psi 1\right\|\sqrt{c}.$$

Now, $\|\pi_{\Sigma''}\Psi_{\bullet,L_2}\|$ is distributed as the norm of a $cm\times cm$ iid $N(0,\nu^2/m)$ matrix, and so for any $\varepsilon_1>0$,

$$P\left[\|\pi_{\Sigma''}\Psi_{\bullet,L_2}\| \ge 2\nu\sqrt{c} \,+\, \varepsilon_1\nu\sqrt{\bar\rho}\right] \;\asymp\; e^{-\varepsilon_1^2\bar\rho m/2}. \tag{54}$$

Similarly, the vector $\frac{1}{\sqrt{m}}\pi_{\Sigma''}\Psi 1$ is has the same norm as a cm-dimensional iid $N(0,\nu^2/m)$ vector, so

$$P\left[\left\|\frac{1}{\sqrt{m}}\pi_{\Sigma''}\Psi 1\right\| \ge \nu\sqrt{c}+\varepsilon_2\nu\sqrt{\bar\rho}\right] \asymp e^{-\varepsilon_2^2\bar\rho m/2}. \tag{55}$$

On the complements of these two bad events, $\left\|\pi_{\Sigma''}\left[\Psi\pi_{1^\perp}\right]_{\bullet,L_2}\right\| \;\le\; (\varepsilon_1+\varepsilon_2)\,\nu\sqrt{\bar\rho}+3\,\nu\sqrt{c}$. Set $\varepsilon_1=\varepsilon_2=1/16$. The probability the union of the two bad events is then asymptotic to $e^{-\bar\rho m/512}$. As in the previous lemma, we close with a union bound over choices of L_1, L_2. The number of such choices is asymptotic to $e^{\left(\bar\rho H\left(\frac{c}{\bar\rho}\right)+\delta H\left(\frac{c}{\delta}\right)\right)m}$, and the probability that there exists some bad pair is bounded by a function asymptotic to

$$\exp\left(\left(\bar\rho H\left(\frac{c}{\bar\rho}\right)+\delta H\left(\frac{c}{\delta}\right)-\bar\rho/512\right)m\right) \tag{56}$$

Under the hypotheses of the lemma, the coefficient of this exponent is negative.

f) Pulling the bounds together: Pulling these three bounds together, with probability at least $1-e^{-Cm(1+o(1))}$,

$$\gamma_{cm}\left(\left[\,(\Phi\pi_\Sigma\Phi^*)^{-1/2}\Phi\pi_\Sigma\;\;\Psi\pi_{1^\perp}\,\right]\right) \;\ge\; \sqrt{\frac{1023}{1024}}\frac{\nu\sqrt{\bar\rho}}{2}-3\nu\sqrt{c}-\frac{\nu\sqrt{\bar\rho}}{8}-3\nu\sqrt{c} \;\ge\; \frac{5}{16}\nu\sqrt{\bar\rho}-6\nu\sqrt{c}. \tag{57}$$

Since $\left|\gamma_{cm}\left(\left[\,\tilde V^*\;-\tilde S\tilde U^*\,\right]\right)-\gamma_{cm}\left(\left[\,(\Phi\pi_\Sigma\Phi^*)^{-1/2}\Phi\pi_\Sigma\;\;\Psi\pi_{1^\perp}\,\right]\right)\right| \le \frac{\nu\sqrt{\bar\rho}}{16}$, the desired bound follows. ■

Lemma 6: Suppose $\bar\rho < 1/2 - 1/C_\mu^2$. With probability at least $1-e^{-Cm(1+o(1))}$ over random error supports $J\in\binom{[m]}{\rho m}$, the "clean part" of the mean, μ_{J^c} satisfies $\|\mu_{J^c}\|\ge\frac{\sqrt{\bar\rho}}{2\,C_\mu}$.

Proof: For $\beta>1$, define $L_\beta \doteq \{\; i \;:\; \mu_i^2 > m^{-1}/\beta\,\}\subset[m]$. Then since $\mu_i^2\le C_\mu^2/m$ for all i, and $\sum_i\mu_i^2=1$,

$$\frac{C_\mu^2}{m}\#L_\beta+\frac{1}{\beta m}(m-\#L_\beta)\ge 1,$$

and so $\gamma_\beta\doteq\frac{\#L_\beta}{m}\ge\frac{\beta-1}{\beta C_\mu^2-1}$. We will show that for any constant $\tau>1$, $P\left[\#J^c\cap L_\beta<\frac{\bar\rho\,\gamma_\beta m}{\tau}\right]\asymp e^{-Cm}$, so that with overwhelming probability

$$\|\mu_{J^c}\|_2^2 \;\ge\; \|\mu_{J^c\cap L_\beta}\|_2^2 \;\ge\; \frac{\bar\rho\,\gamma_\beta m}{\tau}\frac{1}{m\beta} \;\ge\; \frac{\bar\rho\,\gamma_\beta}{\beta\,\tau}. \tag{58}$$

Choosing $\beta = \tau = 2$ gives the bound claimed above. To verify that the intersection of J^c and L_β not too small, notice that

$$P\left[\#J^c \cap L_\beta \le \frac{\bar\rho\,\gamma_\beta m}{\tau}\right] \;\le\; \frac{\sum_{k=0}^{\frac{\bar\rho\,\gamma_\beta m}{\tau}} \binom{\gamma_\beta m}{k}\binom{(1-\gamma_\beta)m}{m-k}}{\binom{m}{\bar\rho m}} \;\le\; \frac{m\,\binom{\gamma_\beta m}{\frac{\bar\rho\gamma_\beta}{\tau}m}\binom{(1-\gamma_\beta)m}{\bar\rho m}}{\binom{m}{\bar\rho m}}$$

Where we have used that $\bar\rho\gamma_\beta/\tau < 1/2$ for the upper bound $\binom{\gamma_\beta m}{k} \le \binom{\gamma_\beta m}{\frac{\bar\rho\gamma_\beta}{\tau}m}$, and $m-k \le \bar\rho m \le \frac{1-\gamma_\beta}{2}$ for the bound $\binom{(1-\gamma_\beta)m}{m-k} \le \binom{(1-\gamma_\beta)\,m}{\bar\rho m}$, and finally, crudely upper bounded the number of terms in the summation by m. The logarithm of the numerator is

$$\gamma_\beta\, mH(\bar\rho/\tau) + (1-\gamma_\beta)\, mH(\bar\rho) + o(m)$$

while the logarithm of the denominator is simply $mH(\bar\rho)+o(m)$. For $\tau > 1$, the denominator dominates, and the logarithm of the complete bound is $-\gamma_\beta m(H(\bar\rho) - H(\bar\rho/\tau) + o(m)$. ■

B. Technical Lemmas for Initial Separating Hyperplane

This section contains several results used above in controlling the initial separator q_0. We will first justify the assertion that $\left[\begin{smallmatrix} Z_1 & Z_2 \\ 0 & \mathtt{I}\end{smallmatrix}\right](G^*G)^{-1}Z^*_{J,\bullet}\,\sigma$ contributes $O(m^{1/2})$ to $\|q_0\|$. We close with a measure concentration result for $\|\theta\cdot\|$, also used in the proof of Lemma 5.

Lemma 7 (Lower order terms for q_0): Suppose that $\bar\rho < \min\left(\delta,\, \frac{1}{2} - \frac{1}{C_\mu}\right)$. Then $\exists\,\nu_0 > 0$ such that if $\nu < \nu_0$, there exist constants (wrt m) C_G and C_q such that

$$\|(G^*G)^{-1}\| \;\le\; C_G \qquad \text{and} \qquad \left\|q_0 - \left[\begin{smallmatrix} Z_1 & Z_2 \\ 0 & \mathtt{I}\end{smallmatrix}\right](G^*G)^{-1}Z^*_{J,\bullet}\sigma\right\| \;\le\; C_q\, m^{1/2-\eta_0/4} \tag{59}$$

simultaneously on the complement of a bad event of probability $\le e^{-Cm^{1-\eta_0/2}(1+o(1))}$.

Proof: We first show that $1^*(G^*G)^{-1}1$, $\|(G^*G)^{-1}1\|$, and $\|(G^*G)^{-1}\|$ are simultaneously bounded by constants w.p. $\ge 1 - e^{-Cm(1+o(1))}$. Write $Q = \left[\begin{smallmatrix} Z_1^*Z_1 & Z_1^*Z_2 \\ Z_2^*Z_1 & Z_2^*Z_2+\mathtt{I}\end{smallmatrix}\right] \in \mathbb{R}^{n\times n}$, and $y = Z^*_{J^c,\bullet}\mu_{J^c} \in \mathbb{R}^n$. The Grammian can be expressed as $G^*G = Q + y1^* + 1y^* + \alpha 11^*$, where $\alpha = \mu^*_{J^c}\mu_{J^c}$, and

$$(G^*G)^{-1} = Q^{-1} - Q^{-1}\left[\begin{array}{cc} 1 & y\end{array}\right]\left[\begin{array}{cc} 1^*Q^{-1}1 & 1^*Q^{-1}y+1 \\ 1^*Q^{-1}y+1 & y^*Q^{-1}y-\alpha\end{array}\right]^{-1}\left[\begin{array}{c} 1^* \\ y^*\end{array}\right]Q^{-1}. \tag{60}$$

Let $b \doteq 1^*Q^{-1}1$, $c \doteq 1^*Q^{-1}y$, $d \doteq y^*Q^{-1}y$, and explicitly invert the above 2×2 matrix:

$$\begin{aligned}(G^*G)^{-1}1 &= Q^{-1}1 - Q^{-1}\left[\begin{array}{cc} 1 & y\end{array}\right]\frac{\left[\begin{array}{cc} d-\alpha & -c-1 \\ -c-1 & b\end{array}\right]}{b(d-\alpha)-(c+1)^2}\left[\begin{array}{c} b \\ c\end{array}\right] \\ &= \frac{c+1}{b\,(\alpha-d)+(c+1)^2}\,Q^{-1}1 \;-\; \frac{1}{\alpha-d+(c+1)^2/b}\,Q^{-1}y \;\doteq\; \lambda_1\,Q^{-1}1 + \lambda_2\,Q^{-1}y.\end{aligned}$$

Similarly, $1^*(G^*G)^{-1}1 \;=\; \frac{-b}{b(d-\alpha)-(c+1)^2} \;\leq\; \frac{1}{\alpha-d}$.

We bound the quadratic terms b, c, and d. Applying Fact 1 to the $\delta m \times \bar{\rho} m$ iid $N(0,\nu^2/m)$ matrix $Z_{J^c,\bullet} = [Z_1\; Z_2]$ gives that $\|Z_{J^c,\bullet}\|_2 \leq \sqrt{2}\nu\left(\sqrt{\delta}+\sqrt{\bar{\rho}}\right)$ on the complement of an event of probability $\asymp e^{-Cm}$. On the complement of that bad event,

$$b \;=\; 1^*Q^{-1}1 \;\geq\; \frac{\|1\|_2^2}{\|Q\|} \;\geq\; \frac{\delta m}{1+\|Z_{J^c,\bullet}\|^2} \;\geq\; \frac{\delta m}{1+2\,\nu^2\,(\sqrt{\delta}+\sqrt{\bar{\rho}})^2} \;\doteq\; C_b\, m \tag{61}$$

For $c = 1^*Q^{-1}y$, notice that $y = Z^*_{J^c,\bullet}\,\mu_{J^c}$ is iid $N(0,\nu^2\alpha/m)$ random vector. We would like to assert that this vector is almost orthogonal to $Q^{-1}1$. To do so, first split out the part of $Z_{J^c,\bullet}$ that is not probabilistically independent of y: write

$$Q \;=\; Z^*_{J^c,\bullet}\pi_{\mu^\perp_{J^c}} Z_{J^c,\bullet} + \left[\begin{smallmatrix} 0 & 0 \\ 0 & \mathtt{I}\end{smallmatrix}\right] + \frac{1}{\alpha} y y^* \;\doteq\; L + \frac{1}{\alpha}\, y\, y^*,$$

and

$$Q^{-1} = L^{-1} - L^{-1} y \frac{1}{\alpha + y^* L^{-1} y} y^* L^{-1}.$$

Then, $|1^*Q^{-1}y| \;=\; \left|1^*L^{-1}y\left(\frac{\alpha}{\alpha+y^*L^{-1}y}\right)\right| \;\leq\; |1^*L^{-1}y|$. Now, $\|L^{-1}1\|_2 \leq \frac{\sqrt{\delta m}}{\sigma_{min}(L)}$. It is not difficult to show[14] that for any block matrix $M = \left[\begin{smallmatrix} A & B \\ 0 & \mathtt{I}\end{smallmatrix}\right]$ with $\|A\|\|B\| < 1-\sigma^2_{min}(A)$,

$$\sigma^2_{min}(M) \;\geq\; \sigma^2_{min}(A) - \frac{\|A\|^2\,\|B\|^2}{1-\sigma^2_{min}(A)}.$$

The relevant singular value is

$$\sigma_{min}(L) \;=\; \sigma^2_{min}\left(\left[\begin{array}{cc} \pi_{\mu^\perp_{J^c}} Z_1 & \pi_{\mu^\perp_{J^c}} Z_2 \\ 0 & \mathtt{I}\end{array}\right]\right) \;\geq\; \sigma^2_{min}(\pi_{\mu^\perp_{J^c}} Z_1) \;-\; \frac{\|\pi_{\mu^\perp_{J^c}} Z_1\|\|\pi_{\mu^\perp_{J^c}} Z_2\|}{1-\|\pi_{\mu^\perp_{J^c}} Z_1\|^2}$$

On the complement of an event of probability $\asymp e^{-Cm}$,

$$\sigma^2_{min}(\pi_{\mu^\perp_{J^c}} Z_1) \geq \frac{\nu^2\bar{\rho}}{2} \qquad \|\pi_{\mu^\perp_{J^c}} Z_1\|^2 \leq \|Z_1\|^2 \leq 2\nu^2\bar{\rho} \qquad \|\pi_{\mu^\perp_{J^c}} Z_2\|^2 \leq \|Z_2\|^2 \leq \sqrt{2}\nu^2\left(\sqrt{\delta}+\sqrt{\bar{\rho}}\right).$$

The first comes by identifying $\pi_{\mu^\perp_{J^c}} Z_1$ with a $\bar{\rho}m - 1 \times k_1$ Gaussian matrix and applying Fact 1, while the second and third follow directly from Fact 1. Plugging in, $\sigma_{min}(L) \geq \frac{\nu^2\bar{\rho}}{2} - \frac{8\nu^4(\sqrt{\delta}+\sqrt{\bar{\rho}})}{1-8\nu^4(\sqrt{\delta}+\sqrt{\bar{\rho}})} \geq \frac{\nu^2\bar{\rho}}{4}$ for ν sufficiently small.[15] Returning to the quantity of interest, $\|L^{-1}1\|_2 \leq \frac{4\sqrt{\delta}}{\nu^2\bar{\rho}}\, m^{1/2}$. Meanwhile, since y is independent of L, $\left\langle \frac{L^{-1}1}{\|L^{-1}1\|_2}, y\right\rangle$ is simply an $N(0,\nu^2\alpha/m)$ random variable, and so for any $\varepsilon > 0$

$$P\left[\left|1^*L^{-1}y\right| > \varepsilon m^{1/2}\right] \;\leq\; P\left[\|L^{-1}1\| > \frac{4\sqrt{\delta}}{\nu^2\bar{\rho}} m^{1/2}\right] + P\left[\left|\left\langle \frac{L^{-1}1}{\|L^{-1}1\|_2}, y\right\rangle\right| > \varepsilon\frac{\nu^2\bar{\rho}}{4\sqrt{\delta}}\right] \;\asymp\; e^{-C_\varepsilon m}$$

[14] Write $\sigma^2_{min}(M) \geq \min_{\|x_1\|_2^2+\|x_2\|_2^2=1} \left(\|Ax_1\|_2 - \|Bx_2\|_2\right)^2 + \|x_2\|_2^2$

$$\geq \min_{\lambda\in[0,1]} \sigma^2_{min}(A) + (1-\sigma^2_{min}(A))(1-\lambda) - 2\|A\|\|B\|\sqrt{1-\lambda}.$$

[15] For example, $\nu < \min\left(\frac{1}{8}\sqrt{\frac{\bar{\rho}}{\sqrt{\delta}+\sqrt{\bar{\rho}}}}, \frac{1}{2}\left(\sqrt{\delta}+\sqrt{\bar{\rho}}\right)^{1/4}\right)$ suffices.

where $C_\varepsilon > 0$ is a constant (wrt m) depending on ε. So, with overwhelming probability, $c = 1^* Q^{-1} y$ has magnitude bounded by $\varepsilon\, m^{1/2}$.

The final quadratic term is $d = y^* Q^{-1} y = y^* L^{-1} y \frac{y^* L^{-1} y}{\alpha + y^* L^{-1} y} \leq y^* L^{-1} y$. Rather than simply independently bounding $\|y\|$ and $\|L^{-1}\|$, we obtain finer control by exploiting the fact that for most vectors L is well-conditioned, due to the presence of the identity matrix in $\left[\begin{smallmatrix} Z_1 & Z_2 \\ 0 & \mathtt{I} \end{smallmatrix}\right]$. Consider the subspace $\Sigma = \{x : x_I = 0\} \subset \mathbb{R}^n$. Since for all $x \in \Sigma$, $\|Lx\|_2 \geq \|x\|_2$, $\left\|L^{-1}\big|_{L\Sigma}\right\| \leq 1$, and

$$\begin{aligned} y^* L^{-1} y &= y^* \left(L^{-1}\big|_{L\Sigma} \pi_{L\Sigma}\, y + L^{-1} \pi_{(L\Sigma)^\perp} y \right) \\ &\leq \|y\|_2^2 \left\|L^{-1}\big|_{L\Sigma}\right\|_2 + \|L^{-1}\|_2 \|y\|_2 \left\|\pi_{(L\Sigma)^\perp} y\right\|_2 \quad \leq \quad 2\nu^2 \alpha\delta \;+\; \frac{4\sqrt{2\alpha\delta}}{\nu\bar{\rho}} \left\|\pi_{(L\Sigma)^\perp} y\right\|_2 . \end{aligned}$$

The final term, $\|\pi_{(L\Sigma)^\perp} y\|$ is the projection of the $N(0, \nu^2\alpha/m)$ vector y onto an independent k_1-dimensional subspace; for any $\varepsilon' > 0$, $P\left[\left\|\pi_{(L\Sigma)^\perp} y\right\| \geq \varepsilon' \nu \sqrt{\alpha}\right] \asymp e^{-\varepsilon'^2 m/2}$. For appropriate choice of ε, with overwhelming probability, $d \leq y^* L^{-1} y \leq 4\nu^2\alpha\delta$.

We now have everything we need to demonstrate the first two claims of the lemma. Due to the centrality assumption, the energy of μ is well-spread: $\alpha \geq \frac{\bar{\rho}}{4C_\mu^2}$ w.p. $\geq 1 - e^{-Cm(1+o(1))}$ (Lemma 6). So, with probability at least $1 - e^{-Cm(1+o(1))}$,

$$1^*(G^*G)^{-1}1 \;\leq\; \frac{1}{\alpha - d} \;\leq\; \frac{1}{\alpha\,(1 - 4\nu^2\delta)} \;\leq\; \frac{4\,C_\mu^2}{\bar{\rho}\,(1 - 4\nu^2\delta)} \;\doteq\; C_1. \tag{62}$$

For the coefficient λ_1 in $(G^*G)^{-1}1$, for any $\varepsilon > 0$

$$|\lambda_1| \;\leq\; \frac{|\,c + 1\,|}{b\,(\alpha - d)} \;\leq\; \frac{\varepsilon\, m^{1/2} + 1}{\frac{\delta m}{1+8\nu^2}\alpha(1 - 4\nu^2\delta)} \tag{63}$$

with overwhelming probability. Hence for any $\varepsilon'' > 0$, $|\lambda_1| \leq \varepsilon'' m^{-1/2}$ for m sufficiently large, on the complement of a bad event of probability $\asymp e^{-Cm}$. Similarly, $|\lambda_2| \leq \frac{1}{\alpha - d} \leq \frac{4\,C_\mu^2}{\bar{\rho}\,(1 - 4\nu^2\delta)}$. An identical argument to the one given for $\|L^{-1}\|$ above shows that $\|Q^{-1}\| \leq \frac{4}{\nu^2\bar{\rho}}$ w.p. $\geq 1 - e^{-Cm(1+o(1))}$, and so

$$\|(G^*G)^{-1}1\|_2 \;\leq\; |\lambda_1|\|Q^{-1}\|\|1\| + |\lambda_2|\|Q^{-1}\|\|y\| \quad \leq \quad \frac{4\,\varepsilon''\sqrt{\delta}}{\nu^2\bar{\rho}} \;+\; \frac{16\sqrt{2\delta} C_\mu}{\bar{\rho}\,(1 - 4\nu^2\delta)} \;\doteq\; C_2.$$

We next bound $\|(G^*G)^{-1}\|$, establishing the second part of the lemma. Introduce matrices $M = \left[\begin{array}{cc} \frac{Q^{-1/2}1}{\|Q^{-1/2}1\|_2} & \frac{Q^{-1/2}y}{\|Q^{-1/2}y\|_2} \end{array}\right]$ and

$$\Xi = \left[\begin{array}{cc} \sqrt{1^*Q^{-1}1} & 0 \\ 0 & \sqrt{y^*Q^{-1}y} \end{array}\right] \left[\begin{array}{cc} 1^*Q^{-1}1 & 1^*Q^{-1}y{+}1 \\ 1^*Q^{-1}y{+}1 & y^*Q^{-1}y{-}\alpha \end{array}\right]^{-1} \left[\begin{array}{cc} \sqrt{1^*Q^{-1}1} & 0 \\ 0 & \sqrt{y^*Q^{-1}y} \end{array}\right].$$

In terms of these matrices, $(G^*G)^{-1} \;=\; Q^{-1} \,-\, Q^{-1/2} M \Xi M^* Q^{-1/2}$. Now,

$$\Xi = \frac{\left[\begin{array}{cc} 1^*Q^{-1}1\,(\alpha - y^*Q^{-1}y) & -\sqrt{1^*Q^{-1}1}\sqrt{y^*Q^{-1}y}\,\left(1^*Q^{-1}y + 1\right) \\ -\sqrt{1^*Q^{-1}1}\sqrt{y^*Q^{-1}y}\,\left(1^*Q^{-1}y + 1\right) & (y^*Q^{-1}y)(1^*Q^{-1}1) \end{array}\right]}{1^*Q^{-1}1\;(\alpha - y^*Q^{-1}y) - (1^*Q^{-1}y + 1)^2}.$$

Applying the quadratic product bounds derived above, w.p. $\geq 1 - e^{-C_\varepsilon m(1+o(1))}$, the denominator is at least

$$\frac{C_b(1-4\nu^2\delta)\bar{\rho}}{4C_\mu}\, m - (\varepsilon m^{1/2}+1)^2 \;\geq\; C_{denom}\, m \qquad \text{eventually}$$

for any constant $C_{denom} < \frac{C_b(1-4\nu^2\delta)\bar{\rho}}{4C_\mu}$ and corresponding choice of ε. Since each of the terms in the numerator is bounded by Cm for some constant C, each of the terms in the 2×2 matrix Ξ is bounded by some constant, and $\|\Xi\| \leq C_\Xi$ w.p. $\geq 1 - e^{-Cm(1+o(1))}$ for appropriate constant C_Ξ. Hence,

$$\|(G^*G)^{-1}\| \;\leq\; \|Q^{-1}\| + \|Q^{-1}\|\,\|M\|^2\,\|\Xi\| \;\leq\; \frac{4}{\nu^2\bar{\rho}} + \frac{4}{\nu^2\bar{\rho}}\,2\,C_\Xi \;\doteq\; C_G,$$

a constant, establishing the first assertion of the lemma.

For the second assertion, recall that

$$\begin{aligned} q_0 \;&=\; \left[\begin{smallmatrix} Z_1 & Z_2 \\ 0 & \mathtt{I} \end{smallmatrix}\right](G^*G)^{-1}Z^*_{J,\bullet}\,\sigma \;+\; \left[\begin{smallmatrix} Z_1 & Z_2 \\ 0 & \mathtt{I} \end{smallmatrix}\right]\Big(-(G^*G)^{-1}1_I + \langle \mu_J,\sigma\rangle\,(G^*G)^{-1}1\Big) \\ &+\; \left[\begin{smallmatrix} \mu_{J^c} \\ 0 \end{smallmatrix}\right]\Big(-1^*(G^*G)^{-1}1_I \;+\; \langle \mu_J,\sigma\rangle\,1^*(G^*G)^{-1}1\Big) + \left[\begin{smallmatrix} \mu_{J^c} \\ 0 \end{smallmatrix}\right]1^*(G^*G)^{-1}Z^*_{J,\bullet}\,\sigma \end{aligned} \tag{64}$$

Before proceeding, we bound $|\langle \mu_J,\sigma\rangle|$. Consider the Martingale $(X_i)_{i=0}^{\rho m}$ given by $X_0 = 0$, $X_i = \sum_{j=1}^{i}\mu_J(j)\sigma(j)$. We are interested in $X_{\rho m} = \langle \mu_J,\sigma\rangle$. Since $|X_i - X_{i-1}| \leq \mu_J(j)$, by Hoeffding's inequality [15],

$$P\left[\,|X_{\rho m}| \geq t\,\right] \;\leq\; 2\exp\left(-\frac{t^2}{2\sum_{j=1}^{\rho m}\mu^2_{J(j)}}\right) \;\leq\; 2\,e^{-\frac{t^2}{2}}, \tag{65}$$

and so with probability $\geq 1 - e^{-Cm^{1-\eta_0/2}}$, $|\langle \mu_J,\sigma\rangle| \leq m^{1/2-\eta_0/4}$.

The second term of (64), $\left\|\left[\begin{smallmatrix} Z_1 & Z_2 \\ 0 & \mathtt{I} \end{smallmatrix}\right]\Big(-(G^*G)^{-1}1_I + \langle \mu_J,\sigma\rangle\,(G^*G)^{-1}1\Big)\right\|$ is bounded above by

$$\Big(1 + 2\nu^2(\sqrt{\delta}+\sqrt{\bar{\rho}})^2\Big)\Big(C_G\sqrt{C_0}\,m^{1/2-\eta_0/2} + m^{1/2-\eta_0/4}C_2\Big) \;\leq\; C_4 m^{1/2-\eta_0/4}$$

w.p. $\geq 1 - e^{-Cm^{1-\eta_0/2}(1+o(1))}$, for appropriate C_4. Similarly, for the third term of (64)

$$\left\|\left[\begin{smallmatrix} \mu_{J^c} \\ 0 \end{smallmatrix}\right]\Big(-1^*(G^*G)^{-1}1_I \;+\; \langle \mu_J,\sigma\rangle\,1^*(G^*G)^{-1}1\Big)\right\| \;\leq\; C_1 + C_1 m^{1/2-\eta_0/4}$$

For the final term of (64), notice that $\vartheta \doteq Z^*_{J,\bullet}\sigma$ is distributed as an iid $N(0,\nu^2\rho)$ vector, independent of G, and so

$$P\left[\left|\left\langle \frac{(G^*G)^{-1}1}{\|(G^*G)^{-1}1\|}, \vartheta\right\rangle\right| \geq m^{1/2-\eta_0/4}\right] \asymp e^{-Cm^{1-\eta_0/2}}. \tag{66}$$

On the complement of this bad event,

$$\left\|\mu_{J^c}1^*(G^*G)^{-1}\vartheta\right\| \;\leq\; \|(G^*G)^{-1}1\|\cdot\left|\left\langle \frac{(G^*G)^{-1}1}{\|(G^*G)^{-1}1\|}, \vartheta\right\rangle\right| \;\leq\; C_2\, m^{1/2-\eta_0/4}. \tag{67}$$

■

Lemma 8 (Concentration for Gaussian tops): Fix $\sigma \le 1$, $\varepsilon \le 1/2$. Let x be a d-dimensional random vector with entries iid $N(0,\sigma^2)$, and let θ be the operator that takes the part of x that is greater than $1-\varepsilon$:

$$\theta : \mathbb{R}^d \to \mathbb{R}^d \;\; s.t. \;\; [\theta x]_i = \begin{cases} sgn(x_i)(|x_i| - 1 + \varepsilon) & |x_i| > 1-\varepsilon \\ 0 & else \end{cases} \tag{68}$$

Then

$$P\left[\, \|\theta x\|_2 \ge 4e^{-\frac{1}{16\sigma^2}}\, d^{1/2} \,\right] \asymp e^{-C_\sigma d}, \tag{69}$$

where C_σ is a constant depending only on σ.

Proof: Let $y \in \mathbb{R}^d$ be iid $N(0,1)$, then $\|\theta x\|_2$ is equal in distribution to $\|\theta\sigma y\|_2$. Now, $\mathbb{E}\|\theta\sigma y\|_2^2 = d\cdot\mathbb{E}(\theta x_i)^2 = \frac{d}{\sigma}\sqrt{\frac{2}{\pi}}\int_{1-\varepsilon}^{\infty} t^2 e^{-t^2/2\sigma^2} dt$. Integrating by parts[16] yields

$$d^{-1}\mathbb{E}\|\theta\sigma y\|_2^2 \;\; = \;\; \frac{(1-\varepsilon)\,\sigma}{\sqrt{\pi/2}} e^{-\frac{(1-\varepsilon)^2}{2\sigma^2}} + 2\,\sigma^2 Q\left(\tfrac{1-\varepsilon}{\sigma}\right) \;\; \le \;\; \sigma\sqrt{\frac{2}{\pi}}\,\frac{1+\sigma^2}{1-\varepsilon}\, e^{-\frac{(1-\varepsilon)^2}{2\sigma^2}} \;\; \le \;\; 4\sigma e^{-\frac{1}{8\sigma^2}},$$

and $\mathbb{E}[\,\|\theta\sigma y\|_2\,] \le 2\, e^{-\frac{1}{16\sigma^2}} d^{1/2}$. Meanwhile, $\mathbb{E}\sqrt{\sum_{i=1}^d |\theta\sigma y_i|^2} = \sqrt{d}\;\mathbb{E}\sqrt{\frac{\sum_{i=1}^d |\theta\sigma y_i|^2}{d}}$. It is not difficult to show[17] that $\mathbb{E}\sqrt{\frac{\sum_{i=1}^d |\theta\sigma y_i|^2}{d}} \to C'_\sigma$ for some constant $C'_\sigma > 0$ depending only on σ and so $\mathbb{E}\|\theta\sigma y\|_2 \ge C'_\sigma\, d^{1/2}$. Since $f(\cdot) = \|\theta\sigma\cdot\|_2$ is 1-Lipschitz for $\sigma \le 1$, $P\,[\,\|\theta\sigma y\|_2 \ge 2\,\mathbb{E}\|\theta\sigma y\|_2\,] \le \exp\left(-8(\mathbb{E}\|\theta\sigma y\|_2)^2/\pi^2\right)$ [11]. Plugging in the upper and lower bounds on $\mathbb{E}|\theta\sigma y\|_2$ yields the result. ■

C. Details of the proof of Theorem 1

Proof: For a given $I \in \binom{[n]}{k_1}$. By Lemma 2, (I,J,σ) is ℓ^1-recoverable if $\exists\, \beta \in (0,1)$ such that

$$\|q_0\|_2 + \frac{1}{1-\xi_{\beta p}}\|\theta q_0\|_2 \;\; \le \;\; (1-\varepsilon)\sqrt{\beta p} \;\; = \;\; (1-\varepsilon)\sqrt{\beta(\bar\rho+\delta)}\, m^{1/2} - o(m^{1/2}).$$

Choose $c \doteq \beta(\bar\rho+\delta)$ small enough that $c < \min\left(\frac{\bar\rho}{64(1+2C_\mu)^2}, \frac{\bar\rho}{1024}\right)$ and $\bar\rho H(c/\bar\rho) + \delta H(c/\delta) < \bar\rho/512$. Then by Lemma 3, for m sufficiently large, and with probability at least $1 - e^{-Cm^{1-\eta_0/2}(1+o(1))}$, as long as $\nu < \min\left(\frac{1}{24\sqrt{\bar\rho}}, (2\delta)^{-1/4}, \frac{1}{2(\sqrt{\delta}+\sqrt{\bar\rho})}\right)$,

$$\frac{1}{1-\xi_{\beta p}} \;=\; \frac{1}{1-\xi_{cm}} \;\le\; \frac{16}{\nu^8}\left(\frac{32}{\sqrt{\delta\bar\rho}-\bar\rho}\right)^4.$$

while

$$\|q_0\|_2 \;\le\; \alpha_1 \nu m^{1/2} + o(m^{1/2}) \qquad \|\theta q_0\|_2 \;\le\; \alpha_2\, e^{-\bar\rho/128\nu^2} m^{1/2} + o(m^{1/2}),$$

[16] And noting that $Q(z) \le \frac{1}{z\sqrt{2\pi}} e^{-z^2/2}$.

[17] Apply the strong law of large numbers to $d^{-1}\sum|\theta\sigma y_i|^2$ and Slutsky's theorem to argue that $\mathbb{E}\sqrt{d^{-1}\sum|\theta\sigma y_i|^2} \to \sqrt{\mathbb{E}|\theta\sigma y_i|^2}$.

and so

$$\|q_0\|_2 + \frac{1}{1-\xi}\|\theta q_0\|_2 \le \alpha_1 \nu m^{1/2} + \frac{C}{\nu^8} e^{-\bar{\rho}/128\nu^2} m^{1/2} + o(m^{1/2}) \;\le\; (1-\varepsilon)\sqrt{c}m^{1/2} - o(m^{1/2})$$

for ν sufficiently small and m sufficiently large. Hence, under these conditions probability that (I, J, σ) is not ℓ^1-recoverable is bounded by $e^{-Cm^{1-\eta_0/2}(1+o(1))}$. The number of subsets I is $\binom{\delta m}{k_1} \le (\delta m)^{k_1} \le e^{C_0 m^{1-\eta_0}\log(\delta m)}$, and so the probability that $\exists\, I$ such that (I, J, σ) is not ℓ^1-recoverable is bounded by

$$e^{C_0 m^{1-\eta_0}\log(\delta m)}\; e^{-Cm^{1-\eta_0/2}(1+o(1))} \;\to\; 0 \quad \text{as} \quad m \to \infty.$$

■