For Most Large Underdetermined Systems of Linear Equations the Minimal ℓ_1-norm Solution Is Also the Sparsest Solution

DAVID L. DONOHO
Stanford University

Abstract

We consider linear equations $y = \Phi x$ where y is a given vector in $\mathbb{R}^n$ and Φ is a given $n \times m$ matrix with $n < m \leq \tau n$, and we wish to solve for $x \in \mathbb{R}^m$. We suppose that the columns of Φ are normalized to the unit ℓ_2-norm, and we place uniform measure on such Φ. We prove the existence of $\rho = \rho(\tau) > 0$ so that for large n and for all Φ's except a negligible fraction, the following property holds: *For every y having a representation $y = \Phi x_0$ by a coefficient vector $x_0 \in \mathbb{R}^m$ with fewer than $\rho \cdot n$ nonzeros, the solution x_1 of the ℓ_1-minimization problem*

$$\min \|x\|_1 \quad \textit{subject to} \quad \Phi x = y$$

is unique and equal to x_0. In contrast, heuristic attempts to sparsely solve such systems—greedy algorithms and thresholding—perform poorly in this challenging setting. The techniques include the use of random proportional embeddings and almost-spherical sections in Banach space theory, and deviation bounds for the eigenvalues of random Wishart matrices. © 2006 Wiley Periodicals, Inc.

1 Introduction

Many situations in science and technology call for solutions to underdetermined systems of equations, i.e., systems of linear equations with fewer equations than unknowns. Examples in array signal processing, inverse problems, and genomic data analysis all come to mind. However, any educated person working in such fields would agree with the statement: "You have a system of linear equations with fewer equations than unknowns. There are infinitely many solutions." And indeed, they would have been taught well. However, the intuition imparted through such teaching would be misleading.

On closer inspection, many of the applications ask for *sparse* solutions of such systems, i.e., solutions with few nonzero elements, the interpretation being that we are sure that "relatively few" of the candidate sources, pixels, or genes are turned on, we just don't know a priori which ones those are. Finding sparse solutions to such systems would better match the real underlying situation. It would also in many cases have important practical benefits, i.e., allowing us to install fewer

Communications on Pure and Applied Mathematics, Vol. LIX, 0797–0829 (2006)

antenna elements, make fewer measurements, store less data, or investigate fewer genes.

The search for sparse solutions can transform the problem completely, in many cases making unique solutions possible (Lemma 2.1 below; see also [8, 10, 16, 17, 28, 29]). Unfortunately, this only seems to change the problem from an impossible one to an intractable one! Finding the sparsest solution to a general underdetermined system of equations is NP-hard [23]; many classic combinatorial optimization problems can be cast in that form.

In this paper we will see that for most underdetermined systems of equations, when a sufficiently sparse solution exists, it can be found by convex optimization. More precisely, for a given ratio m/n of unknowns to equations, there is a threshold $\rho > 0$ so that most large $n \times m$ matrices generate systems of equations with two properties:

- If we run convex optimization to find the ℓ_1-minimal solution and happen to find a solution with fewer than ρn nonzeros, then this is the unique sparsest solution to the equations.
- If the result does not happen to have ρn nonzeros, there is no solution with $< \rho n$ nonzeros.

In such cases, if a sparse solution would be *very desirable*—needing far fewer than n coefficients—it may be found by convex optimization. If it is *of relatively small value*—needing close to n coefficients—finding the optimal solution requires combinatorial optimization.

1.1 Background: Sparse Signal Representation

To place the results of this paper in context, we describe their genesis.

In recent years, a large body of research has focused on the use of overcomplete signal representations, in which a given signal $y \in \mathbb{R}^n$ is decomposed as $y = \sum x(i)\phi_i$ using a dictionary of $m > n$ atoms. Equivalently, we try to solve $y = \Phi x$ for Φ an $n \times m$ matrix. Overcompleteness implies that $m > n$, so the problem is underdetermined. The goal is to use the freedom this allows to provide a sparse representation.

Motivations for this viewpoint were first obtained empirically, where representations of signals were obtained in the early 1990s with the use of combinations of several orthonormal bases by Coifman and collaborators [4, 5] and combinations of several frames in Mallat and Zhang's work on matching pursuit [21] and in the mid 1990s [3] by Chen, Donoho, and Saunders.

A theoretical perspective showing that there is a sound mathematical basis for overcomplete representation has come together rapidly in recent years; see [8, 10, 14, 16, 17, 28, 29]. An early result was the following: Suppose that Φ is the concatenation of two orthobases, so that $m = 2n$. Suppose that the *coherence*—the maximal inner product between any pair of columns of Φ—is at most M. Suppose that $y = \Phi x_0$ where x_0 has at most N nonzeros. If $N < M^{-1}$, x_0 provides the unique optimally sparse representation of y.

Consider the solution x_1 to the problem

$$\min \|x\|_1 \quad \text{subject to} \quad y = \Phi x.$$

If $N \leq (1 + M^{-1})/2$ we have $x_1 = x_0$. In short, we can recover the sparsest representation by solving a convex optimization problem.

As an example, a signal of length n that is a superposition of no more than $\sqrt{n}/2$ total spikes and sinusoids is uniquely representable in that form and can be uniquely recovered by ℓ_1-optimization (in this case $M = 1/\sqrt{n}$). The sparsity bound required in this result, comparable to $1/\sqrt{n}$, is disappointingly small; however, it was surprising at the time that any such result was possible. Many substantial improvements on these results have since been made [8, 14, 16, 17, 28, 29].

It was mentioned in [10] that the phenomena proved there represented only the tip of the iceberg. Computational results published there showed that for randomly generated systems Φ one could get unique recovery even with as many as about $n/5$ nonzeros in a twofold overcomplete representation. Hence, empirically, even a mildly sparse representation could be exactly recovered by ℓ_1-optimization.

Very recently, Candès, Romberg, and Tao [2] showed that for partial Fourier systems, formed by taking n rows *at random* from an $m \times m$ standard Fourier matrix, the resulting $n \times m$ matrix with overwhelming probability allowed exact equivalence between the sparsest solution and the ℓ_1-solution in all cases where the number N of nonzeros was smaller than $cn/\log(n)$. This very inspiring result shows that equivalence is possible with a number of nonzeros almost proportional to n. Furthermore, [2] showed empirical examples where equivalence held with as many as $n/4$ nonzeros.

1.2 This Paper's Contributions

In previous work, equivalence between the minimal ℓ_1-solution and the optimally sparse solution required that the sparse solution have an asymptotically negligible fraction of nonzeros. The fraction $O(n^{-1/2})$ could be accommodated in results of [8, 10, 14, 16, 28], and $O(1/\log(n))$ in [2].

In this paper we construct a large class of examples where equivalence holds even when the number of nonzeros is proportional to n. More precisely, we show that there is a constant $\rho(\tau) > 0$ so that all but a negligible proportion of large $n \times m$ matrices Φ with $n < m \leq \tau n$ have the following property: *for every system $y = \Phi x$ allowing a solution with fewer than ρn nonzeros, ℓ_1-minimization uniquely finds that solution.* Here "proportion of matrices" is taken by using the natural uniform measure on the space of matrices with columns of unit ℓ_2-norm.

We show that, in contrast, greedy algorithms and thresholding algorithms seem to fail in this setting.

An interesting feature of our analysis is its use of techniques from Banach space theory, in particular, quantitative extensions of Dvoretsky's almost-spherical sections theorem (by Milman, Kashin, Schechtman, and others) and other related tools exploiting randomness in high-dimensional spaces, including properties of the minimum eigenvalue of Wishart matrices.

Section 2 gives a formal statement of the result and the overall proof architecture; Sections 3 through 5 prove key lemmas; Section 6 describes a geometric interpretation of these results; Section 7 discusses the failure of greedy and thresholding procedures. Section 8 mentions a heuristic based on the proof given here that correctly predicts the observable empirical behavior of $\rho(\tau)$. Section 9 discusses stability and well-posedness.

2 Overview

Let $\phi_1, \dots, \phi_m$ be random points on the unit sphere $\mathbb{S}^{n-1}$ in $\mathbb{R}^n$, independently drawn from the uniform distribution. Let $\Phi = [\phi_1 \cdots \phi_m]$ be the matrix obtained by concatenating the resulting vectors. We denote this as $\Phi_{n,m}$ when we wish to emphasize the size of the matrix.

For a vector $y \in \mathbb{R}^n$ we are interested in the sparsest possible representation of y using columns of Φ; this is given by

$$(\mathrm{P}_0) \qquad \min \|x\|_0 \quad \text{subject to} \quad \Phi x = y,$$

It turns out that if (P_0) has *any* sufficiently sparse solutions, then it will typically have a unique sparsest one.

LEMMA 2.1 *On an event E having probability* 1*, the matrix Φ has the following* unique sparsest solution *property*:

> *For every vector x_0 having $\|x_0\|_0 < n/2$ the vector $y = \Phi x_0$ generates an instance of problem* (P_0) *whose solution is uniquely x_0.*

PROOF: With probability 1, the ϕ_i are in general position in $\mathbb{R}^n$. If there were two solutions, both with fewer than $n/2$ nonzeros, we would have $\Phi x_0 = \Phi x_1$ implying $\Phi(x_1 - x_0) = 0$, a linear relation involving n conditions satisfied using fewer than n points, contradicting the general position. □

In general, solving (P_0) requires combinatorial optimization and is impractical. The ℓ_1-norm is in some sense the convex relaxation of the ℓ_0-norm. So consider instead the minimal ℓ_1-norm representation:

$$(\mathrm{P}_1) \qquad \min \|x\|_1 \quad \text{subject to} \quad \Phi x = y.$$

This poses a convex optimization problem, and so in principle is more tractable than (P_0). Surprisingly, when the answer to (P_0) is sparse, it can be the same as the answer to (P_1).

DEFINITION 2.2 The *equivalence breakdown point* of a matrix Φ, $\mathrm{EBP}(\Phi)$, is the maximal number N such that, for every x_0 with fewer than N nonzeros, the corresponding vector $y = \Phi x_0$ generates a linear system $y = \Phi x$ for which problems (P_1) and (P_0) have identical unique solutions, both equal to x_0.

Using known results, we have immediately that the EBP typically exceeds $c\sqrt{n/\log(m)}$.

LEMMA 2.3 *For each $\eta > 0$,*

$$\mathrm{Prob}\left\{\mathrm{EBP}(\Phi_{n,m}) > \sqrt{\frac{n}{(8+\eta)\log(m)}}\right\} \to 1, \quad n \to \infty.$$

PROOF: The mutual coherence $M = \max_{i \neq j} |\langle \phi_i, \phi_j \rangle|$ obeys

$$M < \sqrt{\frac{2\log(m)}{n}}(1 + o_p(1));$$

compare calculations in [8, 10]. Applying [8], (P_0) and (P_1) have the same solution whenever $\|x_0\|_0 < (1 + M^{-1})/2$.

While it may seem that $O(\sqrt{n/\log(m)})$ is already surprisingly large, more than we "really deserve," more soberly, this is asymptotically only a vanishing *fraction* of nonzeros. In fact, the two problems have the same solution over an even much broader range of sparsity $\|x_0\|_0$, extending up to a *nonvanishing* fraction of nonzeros. Here is our main result.

THEOREM 2.4 *For each $\tau > 1$, there is a constant $\rho^*(\tau) > 0$ so that for every sequence (m_n) with $m_n \leq \tau n$,*

$$\mathrm{Prob}\{n^{-1}\,\mathrm{EBP}(\Phi_{n,m_n}) \geq \rho^*(\tau)\} \to 1, \quad n \to \infty.$$

In words, the overwhelming majority of $n \times m$ matrices Φ have the property that, for every underdetermined system of equations $y = \Phi x$ possessing a solution with at most $\rho^* n$ nonzeros, that solution is both the sparsest possible solution and the minimal ℓ_1-solution.

Remark 2.5. The space of $n \times m$ matrices having columns with unit norm is, of course,

$$\overbrace{\mathbb{S}^{n-1} \times \cdots \times \mathbb{S}^{n-1}}^{m \text{ terms}}.$$

Now the probability measure we are assuming on our random matrix Φ is precisely the canonical uniform measure on this space. Hence, the above result shows that having $\mathrm{EBP}(\Phi) \geq \rho^* n$ is a *generic* property of matrices, experienced on a set of nearly full measure.

Remark 2.6. We emphasize that no probability measure is placed on x_0. Given a matrix Φ obeying $\mathrm{EBP}(\Phi) \geq \rho^* n$, ℓ_1-minimization is successful for every x_0 having at most $\rho^* n$ nonzeros.

Remark 2.7. The proof shows that the probability in question tends to 1 exponentially fast.

Remark 2.8. The proof shows that the function $\rho^*(\tau) \geq c/\log(\tau)$ on $\tau \geq 1$.

Remark 2.9. An explicit lower bound for ρ^* can be given based on our proof, but it is exaggeratedly small. Empirical studies observed in computer simulations set $\frac{3}{10}n$ as the empirical breakdown point when $\tau = 2$, and a heuristic based on our proof quite precisely predicts the same breakdown point—see Section 8 below.

2.1 Geometric Intuition

Our proof is based on a geometric viewpoint about the uniqueness question borrowed from [9, 11]. With x_1 the solution of (P$_1$) and x_0 the solution of (P$_0$), we have

$$\|x_1\|_1 \leq \|x_0\|_1, \tag{2.1}$$

since x_0 is merely feasible for (P$_1$), while x_1 is optimal. At the same time

$$\Phi x_1 = \Phi x_0. \tag{2.2}$$

Let B_1^m denote the ℓ_1^m-ball consisting of all vectors x obeying $\|x\|_1 \leq \|x_0\|_1$. Let $\mathcal{N}_{x_0}$ denote the affine manifold consisting of all vectors x obeying $\Phi(x - x_0) = 0$. It is the translate of the null space $\mathcal{N}_0 = \{x : \Phi x = 0\}$ of Φ. Both x_0 and x_1 belong to the intersection $B_1^m \cap \mathcal{N}_{x_0}$.

Our proof of Theorem 2.4 is an argument showing that, when x_0 is sufficiently sparse, the intersection $B_1^m \cap \mathcal{N}_{x_0}$ reduces to a point.

Figure 2.1 suggests studying the intersection between the translated null space $\mathcal{N}_{x_0}$ and a certain convex cone. The cone contains all the directions that initially reduce the ℓ_1-norm:

$$\mathcal{K}_{x_0} = \{\delta : \|x_0 + t\delta\|_1 \leq \|x_0\|_1 \text{ for small } t > 0\}.$$

If the angle between these two objects is significantly nonzero, it is visually apparent that the intersection $B_1^m \cap \mathcal{N}_{x_0}$ reduces to a point.

What does it mean to control the angle between various facets of the convex cone $\mathcal{K}_{x_0}$ and the affine subspace $\mathcal{N}_{x_0}$? Roughly, it requires the enumeration of every facet of $\mathcal{K}_{x_0}$ and the verification of a clean intersection between the corresponding hyperplane and the subspace $\mathcal{N}_{x_0}$. What we are effectively going to show is that, when x_0 has at most ρn nonzero elements, we can control this intersection, keeping it a singleton.

2.2 Analytic Intuition

Behind all our analysis is the observation that the null space $\mathcal{N}_0$ is a random subspace of $\mathbb{R}^m$. We are thus trying to show that a random subspace has a clean intersection with each member of a fixed, discrete family of hyperplanes.

We will be appealing to the (rough) principle that that all vectors in a random k-dimensional subspace of an n-dimensional space look "nearly Gaussian" in the

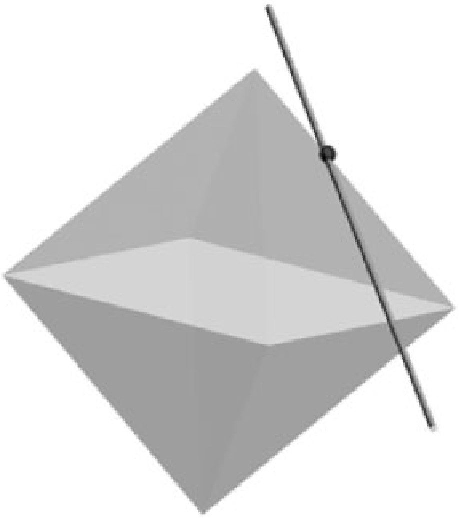

FIGURE 2.1. Geometry of the unique ℓ_1-solution. Intersection of linear subspace $\{x : y = \Phi x\}$ with B_1^m reduces to a single point. Both x_0 and x_1 must lie in the intersection. Hence the ℓ_1-solution is the sparsest solution.

sense that the histograms of the entries in all such vectors have approximately a Gaussian shape. Vectors with this nearly Gaussian behavior have all norms roughly equivalent, with constants of equivalence that can be approximated by assuming the entries in the vector are samples from a Gaussian distribution. Hence $\|x\|_1 \approx \sqrt{2n/\pi} \cdot \|x\|_2$ for all x in such a low-dimensional subspace. Note the factor $\sqrt{n}$; the ℓ_1-norm is much larger than the ℓ_2-norm. Also, if we consider the norm $J(y) \equiv \mathrm{val}(P_1)$, then $J(y) \approx c \cdot \sqrt{n}\|y\|_2$ for all y in such a subspace.

A back-of-the-envelope calculation based on such observations points in the direction of our theorem's conclusion. Roughly, when x_0 is sufficiently sparse, any perturbation of x_0 that obeys $\Phi x = y$ has entries that look nearly Gaussian off the support of x_0, and so the component of the perturbation off the support must have a very large ℓ_1-norm; this will more than offset any gain that the perturbation could cause on the support of x_0. Since no perturbation to x_0 can reduce the ℓ_1-norm, x_0 is the unique solution.

Turning such a rough idea into a proof, of course, takes substantial work! The nearly Gaussian property seems much too vague to yield a proof, and instead we argue based on norm equivalence properties coming out of the theory of random matrices and the geometry of Banach spaces.

2.3 Proof Outline

Let $y = \Phi x_0$ and let $I = \text{supp}(x_0)$. Suppose there is an alternate decomposition

$$y = \Phi(x_0 + \delta)$$

where the perturbation δ obeys $\Phi\delta = 0$. Partitioning $\delta = (\delta_I, \delta_{I^c})$, we have

$$\Phi_I \delta_I = -\Phi_{I^c}\delta_{I^c}.$$

We will simply show that, on a certain event $\Omega_n(\rho, \tau)$,

$$\|\delta_I\|_1 < \|\delta_{I^c}\|_1 \tag{2.3}$$

uniformly over every I with $|I| < \rho n$ and over every $\delta_I \neq 0$. Now

$$\|x_0 + \delta\|_1 - \|x_0\|_1 \geq \|\delta_{I^c}\|_1 - \|\delta_I\|_1.$$

It is then always the case that any perturbation $\delta \neq 0$ increases the ℓ_1-norm relative to the unperturbed case $\delta = 0$. In words, every perturbation hurts the ℓ_1-norm more off the support of x_0 than it helps the norm on the support of x_0, so it hurts the ℓ_1-norm overall, so every perturbation leads away from what, by convexity, must therefore be the global optimum.

It follows that the ℓ_1-minimizer is unique whenever $|I| < \rho n$, and the event $\Omega_n(\rho, \tau)$ occurs.

Formally, the event $\Omega_n(\rho, \tau)$ is the intersection of three subevents Ω_n^i, $i = 1, 2, 3$. These depend on positive constants η_i and ρ_i to be chosen later. The subevents are:

- Ω_n^1: The minimum singular value of Φ_I exceeds η_1, uniformly in I with $|I| < \rho_1 n$.
- Ω_n^2: Denote $v = \Phi_I\delta_I$. The ℓ_1-norm $\|v\|_1$ exceeds $\eta_2\sqrt{n}\|v\|_2$, uniformly in I with $|I| < \rho_2 n$.
- Ω_n^3: Let δ_{I^c} obey $v = -\Phi_{I^c}\delta_{I^c}$. The ℓ_1-norm $\|\delta_{I^c}\|_1$ exceeds $\eta_3\|v\|_1$ uniformly in I with $|I| < \rho_3 n$.

Lemmas 3.1, 4.4, and 5.1 show that one can choose the ρ_i and η_i so that the complement of each of the Ω_n^i, $i = 1, 2, 3$, has probability tending to zero exponentially fast in n. We so choose. It follows, with $\rho_4 \equiv \min_i \rho_i$, that the intersection event $E_{\rho_4,n} \equiv \bigcap_i \Omega_n^i$ is overwhelmingly likely for large n.

When we are on the event $E_{\rho_4,n}$, Ω_n^1 gives us

$$\begin{aligned}\|\delta_I\|_1 &\leq \sqrt{|I|} \cdot \|\delta_I\|_2 \\ &\leq \sqrt{|I|}\frac{\|v\|_2}{\lambda_{\min}^{1/2}(\Phi_I^{\mathsf{T}}\Phi_I)} \\ &\leq \eta_1^{-1}|I|^{1/2}\|v\|_2.\end{aligned}$$

At the same time, Ω_n^2 gives us

$$\|v\|_1 \geq \eta_2\sqrt{n}\|v\|_2.$$

Finally, Ω_n^3 gives us

$$\|\delta_{I^c}\|_1 \geq \eta_3 \|v\|_1,$$

and hence, provided

$$|I|^{1/2} < \eta_1 \cdot \eta_2 \cdot \eta_3 \cdot \sqrt{n},$$

we have (2.3), and hence ℓ_1 succeeds. In short, we just need to bound the fraction $|I|/n$.

Now pick $\rho^* = \min(\rho_4, (\eta_1 \cdot \eta_2 \cdot \eta_3)^2)$ and set $\Omega_n(\rho^*, \tau) = E_{\rho_4,n}$; we get $\mathrm{EBP}(\Phi) \geq \rho^* n$ on $\Omega_n(\rho^*, \tau)$. □

It remains to prove Lemmas 3.1, 4.4, and 5.1 supporting the above analysis.

3 Controlling the Minimal Eigenvalues

We first show there is, with overwhelming probability, a uniform bound $\eta_1(\rho, \tau)$ on the minimal singular value of every matrix Φ_I that can be constructed from the matrix Φ with $|I| < \rho n$. This is of independent interest; see Section 9.

LEMMA 3.1 *Let $\lambda < 1$. Define the event*

$$\Omega_{n,m,\rho,\lambda} = \{\lambda_{\min}(\Phi_I^{\mathsf{T}} \Phi_I) \geq \lambda \ \forall |I| < \rho \cdot n\}.$$

There is $\rho_1 = \rho_1(\lambda, \tau) > 0$ so that, along sequences (m_n) with $m_n \leq \tau n$,

$$P(\Omega_{n,m_n,\rho_1,\lambda}) \to 1, \quad n \to \infty.$$

The bound $\eta_1(\rho, \tau) > 0$ is implied by this result; simply invert the relation $\lambda \mapsto \rho_1(\lambda, \tau)$ and put $\eta_1 = \lambda^{1/2}$.

3.1 Individual Result

We first study $\lambda_{\min}(\Phi_I^{\mathsf{T}} \Phi_I)$ for a single fixed I.

LEMMA 3.2 *Let $\rho > 0$ be sufficiently small. There exist $\eta = \eta(\rho) > 0$, $\beta(\rho) > 0$, and $n_1(\rho)$ so that for $k = |I| \leq \rho n$ we have*

$$P\{\lambda_{\min}(\Phi_I^{\mathsf{T}} \Phi_I) \leq \eta^2\} \leq \exp(-n\beta), \quad n > n_1.$$

PROOF: Effectively, our idea is to show that Φ_I is related to matrices of independent and identically distributed (i.i.d.) Gaussians, for which such phenomena are already known.

Without loss of generality suppose that $I = \{1, \dots, k\}$. Let R_i, $i = 1, \dots, k$, be i.i.d. random variables distributed $\chi_n/\sqrt{n}$, where χ_n denotes the χ_n-distribution. These can be generated by taking i.i.d. standard normal random variables Z_{ij} that are independent of (ϕ_i) and setting

$$R_i = \left(n^{-1} \sum_{j=1}^{n} Z_{ij}^2\right)^{1/2}. \tag{3.1}$$

Let $\psi_i = R_i \cdot \phi_i$; then the ψ_i are i.i.d. $N(0, \frac{1}{n} I_n)$, and we view them as the columns of Ψ. With $R = \mathrm{diag}((R_i)_i)$, we have $\Phi_I = \Psi R^{-1}$, and so

$$\lambda_{\min}(\Phi_I^{\mathsf{T}} \Phi_I) = \lambda_{\min}((R^{-1})^{\mathsf{T}} \Psi^{\mathsf{T}} \Psi R^{-1}) \ge \lambda_{\min}(\Psi^{\mathsf{T}} \Psi) \cdot (\max_i R_i)^{-2}. \tag{3.2}$$

Hence, for a given $\eta > 0$ and $\epsilon > 0$, the two events

$$E = \{\lambda_{\min}(\Psi^{\mathsf{T}} \Psi) \ge (\eta + \epsilon)^2\}, \qquad F = \left\{ \max_i R_i < 1 + \frac{\epsilon}{\eta} \right\}$$

together imply the event

$$\{\lambda_{\min}(\Phi_I^{\mathsf{T}} \Phi_I) \ge \eta^2\}.$$

The following lemma will be proved in the next subsection:

LEMMA 3.3 *For $u > 0$,*

$$P\{\max_i R_i > 1 + u\} \le \exp\left\{-n \frac{u^2}{2}\right\}. \tag{3.3}$$

There we will also prove the following:

LEMMA 3.4 *Let Ψ be an $n \times k$ matrix of i.i.d. $N(0, 1/n)$ Gaussians, $k < n$. Let $\lambda_{\min}(\Psi^{\mathsf{T}} \Psi)$ denote the minimum eigenvalue of $\Psi^{\mathsf{T}} \Psi$. For $\epsilon > 0$ and $k/n \le \rho$,*

$$P\left\{\lambda_{\min}(\Psi^{\mathsf{T}} \Psi) < (1 - \sqrt{\rho} - \epsilon - t)^2\right\} \le \exp\left(-n \frac{t^2}{2}\right), \quad n > n_0(\epsilon, \rho). \tag{3.4}$$

Pick now $\eta > 0$ with $\eta < 1 - \sqrt{\rho}$, and choose ϵ so $2\epsilon < 1 - \sqrt{\rho} - \eta$; finally, put $t = 1 - \sqrt{\rho} - 2\epsilon - \eta$. Define $u = \epsilon/\eta$. Then by Lemma 3.4

$$P(E^{\mathrm{c}}) \le \exp\left(-n \frac{t^2}{2}\right), \quad n > n_0(\epsilon, \rho),$$

while by Lemma 3.3

$$P(F^{\mathrm{c}}) \le \exp\left(-n \frac{u^2}{2}\right).$$

Setting $\beta < \min(t^2/2, u^2/2)$, we conclude that, for $n_1 = n_1(\epsilon, \rho, \beta)$,

$$P\left\{\lambda_{\min}(\Phi_I^{\mathsf{T}} \Phi_I) < \eta^2\right\} \le \exp(-n\beta), \quad n > n_1(\epsilon, \rho, \beta).$$

□

3.2 Invoking Concentration of Measure

We now prove Lemma 3.3. Now (3.1) exhibits each R_i as a function of n i.i.d. standard normal random variables, Lipschitz with respect to the standard Euclidean metric, with Lipschitz constant $1/\sqrt{n}$. Moreover, $\max_i R_i$ itself is such a Lipschitz function. By concentration of measure for Gaussian variables [20], (3.3) follows.

The proof of Lemma 3.4 depends on the observation—see Szarek [27], Davidson and Szarek [6], or El Karoui [13]—that the singular values of Gaussian matrices obey concentration of measure:

LEMMA 3.5 *Let Ψ be an $n \times k$ matrix of i.i.d. $N(0, \frac{1}{n})$ Gaussians, $k < n$. Let $s_\ell(\Psi)$ denote the ℓ^{th} largest singular value of Ψ, $s_1 \geq s_2 \geq \cdots$. Let $\sigma_{\ell;k,n} = \mathrm{Median}(s_\ell(\Psi))$. Then*

$$P\{s_\ell(\Psi) < \sigma_{\ell;k,n} - t\} \leq \exp\left(-n\frac{t^2}{2}\right).$$

The idea is that a given singular value, viewed as a function of the entries of a matrix, is Lipschitz with respect to the Euclidean metric on $\mathbb{R}^{nk}$. Then one applies concentration of measure for scaled Gaussian variables.

As for the median $\sigma_{k;k,n}$, we remark that the well-known Marčenko-Pastur law implies that, if $k_n/n \to \rho$,

$$\sigma_{k_n;k_n,n} \to 1 - \sqrt{\rho}, \quad n \to \infty.$$

Hence, for a given $\epsilon > 0$ and all sufficiently large $n > n_0(\epsilon, \rho)$, $\sigma_{k_n;k_n,n} > 1 - \sqrt{\rho} - \epsilon$. Observing that $s_k(\Psi)^2 = \lambda_{\min}(\Psi^\mathsf{T}\Psi)$ gives the conclusion (3.4).

3.3 Proof of Lemma 3.1

We now combine estimates for individual I's obeying $|I| \leq \rho n$ to obtain the simultaneous result.

We need a standard combinatorial fact, used here and below:

LEMMA 3.6 *For $p \in (0, 1)$, let $H(p) = p\log(1/p) + (1-p)\log(1/(1-p))$ be Shannon entropy. Then*

$$\log\binom{N}{\lfloor pN \rfloor} = NH(p)(1 + o(1)), \quad N \to \infty.$$

Now for a given $\lambda \in (0, 1)$ and each index set I, define the event

$$\Omega_{n,I;\lambda} = \{\lambda_{\min}(\Phi_I^\mathsf{T}\Phi_I) \geq \lambda\}.$$

Then

$$\Omega_{n,m,\rho,\lambda} = \bigcap_{|I| \leq \rho n} \Omega_{n,I;\lambda}.$$

By Boole's inequality,

$$P(\Omega^\mathsf{c}_{n,m,\rho,\lambda}) \leq \sum_{|I| \leq \rho n} P(\Omega^\mathsf{c}_{n,I;\lambda}),$$

so

$$\log P(\Omega^\mathsf{c}_{n,m,\rho,\lambda}) \leq \log \#\{I : |I| \leq \rho n\} + \log P(\Omega^\mathsf{c}_{n,I;\lambda}), \tag{3.5}$$

and we want the right-hand side to tend to $-\infty$. By Lemma 3.6,

$$\log \#\{I : |I| \leq \rho n\} = \log\binom{m_n}{\lfloor \rho n \rfloor} = \tau n H\left(\frac{\rho}{\tau}\right)(1 + o(1)).$$

Invoking now Lemma 3.2, we get a $\beta > 0$ so that for $n > n_0(\rho, \lambda)$, we have

$$\log P(\Omega^\mathsf{c}_{n,I;\lambda}) \leq -\beta n.$$

We wish to show that the $-\beta n$ in this relation can outweigh $\tau n H(\rho/\tau)$ in the preceding one, giving a combined result in (3.5) tending to $-\infty$. Now note that the Shannon entropy $H(p) \to 0$ as $p \to 0$. Hence for small enough ρ, $\tau H(\rho/\tau) < \beta$. Picking such a ρ—call it ρ_1—and setting $\beta_1 = \beta - \tau H(\rho_1/\tau) > 0$, we have for $n > n_0$ that

$$\log(P(\Omega^{\mathsf{c}}_{n,m,\rho_1,\lambda})) \le \tau n H\left(\frac{\rho_1}{\tau}\right)(1 + o(1)) - \beta n,$$

which implies an n_1 so that

$$P(\Omega^{\mathsf{c}}_{n,m,\rho,\lambda}) \le \exp(-\beta_1 n), \quad n > n_1(\rho, \lambda).$$

4 Almost-Spherical Sections

Dvoretsky's theorem [12, 24] says that every infinite-dimensional Banach space contains very-high-dimensional subspaces on which the Banach norm is nearly proportional to the Euclidean norm. This is called the spherical sections property, since it says that slicing the unit ball in the Banach space by intersection with an appropriate finite-dimensional linear subspace will result in a slice that is effectively spherical. We need a quantitative refinement of this principle for the ℓ_1-norm in $\mathbb{R}^n$, showing that, with overwhelming probability, every operator Φ_I for $|I| < \rho n$ affords a spherical section of the ℓ_1^n-ball. The basic argument we use derives from refinements of Dvoretsky's theorem in Banach space theory, going back to the work of Milman and others [15, 22, 26].

DEFINITION 4.1 Let $|I| = k$. We say that Φ_I offers an *ϵ-isometry between $\ell_2(I)$ and ℓ_1^n* if

$$(4.1) \qquad (1 - \epsilon) \cdot \|\alpha\|_2 \le \sqrt{\frac{\pi}{2n}} \cdot \|\Phi_I \alpha\|_1 \le (1 + \epsilon) \cdot \|\alpha\|_2 \quad \forall \alpha \in \mathbb{R}^k.$$

Remark 4.2. The scale factor $\sqrt{\pi/2n}$ embedded in the definition is reciprocal to the expected ℓ_1^n-norm of a standard i.i.d. Gaussian sequence.

Remark 4.3. In Banach space theory, the same notion would be called an $(1 + \epsilon)$–isometry [15, 24].

LEMMA 4.4 (Simultaneous ϵ-isometry) *Consider the event Ω_n^2 ($\equiv \Omega_n^2(\epsilon, \rho)$) that* every *$\Phi_I$ with $|I| \le \rho \cdot n$ offers an ϵ-isometry between $\ell_2(I)$ and ℓ_1^n. For each $\epsilon > 0$, there is $\rho_2(\epsilon) > 0$ so that*

$$P(\Omega_n^2(\epsilon, \rho_2)) \to 1, \quad n \to \infty.$$

4.1 Proof of Simultaneous Isometry

Our approach is based on a result for individual I, which will later be extended to get a result for every I. This individual result is well-known in Banach space theory, going back to [15, 18, 26]. For our proof, we repackage key elements from the proof of theorem 4.4 in Pisier's book [24]. Pisier's argument shows that for *one specific* I, there is a *positive probability* that Φ_I offers an ϵ-isometry. We add extra "bookkeeping" to find that the probability is actually overwhelming and later conclude that there is an *overwhelming* probability that *every* I with $|I| < \rho n$ offers such isometry.

LEMMA 4.5 (Individual ϵ-isometry) *Fix* $\epsilon > 0$. *Choose* δ *so that*

$$(4.2)\qquad (1-3\delta)(1-\delta)^{-1} \geq (1-\epsilon)^{1/2} \quad \textit{and} \quad (1+\delta)(1-\delta)^{-1} \leq (1+\epsilon)^{1/2}.$$

Choose $\rho_0 = \rho_0(\epsilon) > 0$ *so that*

$$\rho_0 \cdot \left(1 + \frac{2}{\delta}\right) < \delta^2 \frac{2}{\pi},$$

and let $\beta(\epsilon)$ *denote the difference between the two sides. For a subset* I *in* $\{1, \dots, m\}$, *let* $\Omega_{n,I}$ *denote the event* $\{\Phi_I$ offers an ϵ-isometry to $\ell_1^n\}$. *Then as* $n \to \infty$,

$$\max_{|I| \leq \rho_0 n} P(\Omega_{n,I}^{\mathsf{c}}) \leq 2\exp(-\beta(\epsilon) n(1+o(1))).$$

This lemma will be proved in Section 4.2. We first show how it implies Lemma 4.4.

With $\beta(\epsilon)$ as given in Lemma 4.5, we choose $\rho_2(\epsilon) < \rho_0(\epsilon)$ that satisfies

$$\tau H\left(\frac{\rho_2}{\tau}\right) < \beta(\epsilon),$$

where $H(p)$ is the Shannon entropy, and let $\gamma > 0$ be the difference between the two sides. Now

$$\Omega_n^2 = \bigcap_{|I| < \rho_2 n} \Omega_{n,I}.$$

It follows that

$$P((\Omega_n^2)^{\mathsf{c}}) \leq \#\{I : |I| \leq \rho_2 n\} \cdot \max_{|I| \leq \rho n} P(\Omega_{n,I}^{\mathsf{c}}).$$

Hence

$$\log(P((\Omega_n^2)^{\mathsf{c}})) \leq n\left[\tau H\left(\frac{\rho_2}{\tau}\right) - \beta(\epsilon)\right](1+o(1)) = -\gamma n \cdot (1+o(1)) \to -\infty.$$

4.2 Proof of Individual Isometry

We temporarily Gaussianize our dictionary elements ϕ_i. Let R_i be i.i.d. random variables distributed $\chi_n/\sqrt{n}$, where χ_n denotes the χ_n-distribution. This can be generated by taking i.i.d. standard normal random variables Z_{ij} that are independent of (ϕ_i) and setting

$$R_i = \left(n^{-1}\sum_{j=1}^{n} Z_{ij}^2\right)^{1/2}. \tag{4.3}$$

Let $\psi_i = R_i \cdot \phi_i \cdot \sqrt{\pi/2n}$. Then ψ_i are i.i.d. n-vectors with entries i.i.d. $N(0, \pi/2n^2)$. It follows that

$$E\|\psi_i\|_1 = 1.$$

Define, for each $\alpha \in \mathbb{R}^k$, $f_\alpha(\psi_1, \dots, \psi_k) = \|\sum_i \alpha_i \psi_i\|_1$. If $\alpha \in \mathbb{S}^{k-1}$, the distribution of $\sum_i \alpha_i \psi_i$ is $N(0, \frac{\pi}{2n} I_n)$; hence $E f_\alpha = 1$ for all $\alpha \in \mathbb{S}^{k-1}$. More transparently,

$$E\left\|\sum \alpha_i \psi_i\right\|_1 = \|\alpha\|_2 \quad \forall \alpha \in \mathbb{R}^k.$$

In words, there is exact isometry between the ℓ_2-norm and the expectation of the ℓ_1-norm. We now show that over individual realizations there is approximate isometry; i.e., individual realizations are close to their expectations.

We need two standard lemmas in Banach space theory [15, 18, 22, 26]; we simplify versions in Pisier [24, chap. 4]:

LEMMA 4.6 *Let $u_i \in \mathbb{R}^n$, $i = 1, \dots, k$. For each $\epsilon > 0$, choose δ obeying* (4.2). *Let $\mathcal{N}_\delta$ be a δ-net for $\mathbb{S}^{k-1}$ under the ℓ_2^k-metric. The validity on this net of norm equivalence,*

$$1 - \delta \le \left\|\sum_i \alpha_i u_i\right\|_1 \le 1 + \delta \quad \forall \alpha \in \mathcal{N}_\delta,$$

implies norm equivalence on the whole space:

$$(1-\epsilon)^{1/2}\|\alpha\|_2 \le \left\|\sum_i \alpha_i u_i\right\|_1 \le (1+\epsilon)^{1/2}\|\alpha\|_2 \quad \forall \alpha \in \mathbb{R}^k.$$

LEMMA 4.7 *There is a δ-net $\mathcal{N}_\delta$ for $\mathbb{S}^{k-1}$ under the ℓ_2^k-metric obeying*

$$\log(\#\mathcal{N}_\delta) \le k\left(1 + \frac{2}{\delta}\right).$$

So, given $\epsilon > 0$ in the statement of our lemma, invoke Lemma 4.6 to get a workable δ, and invoke Lemma 4.7 to get a net $\mathcal{N}_\delta$ obeying the required bound. Corresponding to each element α in the net $\mathcal{N}_\delta$, define now the event

$$E_\alpha = \left\{1 - \delta \le \left\|\sum_i \alpha_i \psi_i\right\|_1 \le 1 + \delta\right\}.$$

On the event $E = \bigcap_{\alpha \in \mathcal{N}_\delta} E_\alpha$, we may apply Lemma 4.6 to conclude that the system $(\psi_i : 1 \le i \le k)$ gives ϵ-equivalence between the ℓ_2-norm on $\mathbb{R}^k$ and the ℓ_1-norm on $\mathrm{Span}(\psi_i : 1 \le i \le k)$.

Now $E_\alpha^{\mathrm{c}} \equiv \{|f_\alpha - Ef_\alpha| > \delta\}$. We note that f_α may be viewed as a function g_α on kn i.i.d. standard normal random variables, where g_α is a Lipschitz function on $\mathbb{R}^{kn}$ with respect to the ℓ_2-metric, having Lipschitz constant $\sigma = \sqrt{\pi/2n}$. By concentration of measure for Gaussian variables [20, secs. 1.2–1.3],

$$P\{|f_\alpha - Ef_\alpha| > t\} \le 2\exp\left\{-\frac{t^2}{2}\sigma^2\right\}.$$

Hence

$$P(E_\alpha^{\mathrm{c}}) \le 2\exp\left\{-\delta^2 \cdot n \cdot \frac{2}{\pi}\right\}.$$

From Lemma 4.7 we have

$$\log \#\mathcal{N}_\delta \le k\left(1 + \frac{2}{\delta}\right)$$

and so

$$\log(P(E^{\mathrm{c}})) \le k \cdot \left(1 + \frac{2}{\delta}\right) + \log 2 - \delta^2 \cdot n \cdot \frac{2}{\pi} < \log(2) - n\beta(\epsilon).$$

We conclude that the ψ_i give a near isometry with overwhelming probability.

We now de-Gaussianize. We argue that, with overwhelming probability, we also get an ϵ-isometry of the desired type for Φ_I. Setting $\gamma_i = \alpha_i \cdot \sqrt{\pi/2n} \cdot R_i$, observe that

$$\sum_i \alpha_i \psi_i = \sum_i \gamma_i \phi_i. \tag{4.4}$$

Pick η so that

$$(1+\eta) < (1-\epsilon)^{-1/2}, \qquad (1-\eta) > (1+\epsilon)^{-1/2}. \tag{4.5}$$

Consider the event

$$G = \{(1-\eta) < R_i < (1+\eta) : i = 1, \dots, n\}.$$

On this event we have the isometry

$$(1-\eta) \cdot \|\alpha\|_2 \le \sqrt{\frac{2n}{\pi}} \cdot \|\gamma\|_2 \le (1+\eta) \cdot \|\alpha\|_2.$$

It follows that on the event $G \cap E$, we have

$$\begin{aligned} \frac{(1-\epsilon)^{1/2}}{(1+\eta)} \cdot \sqrt{\frac{2n}{\pi}} \|\gamma\|_2 &\le (1-\epsilon)^{1/2} \|\alpha\|_2 \\ &\le \Big\| \sum_i \alpha_i \psi_i \Big\|_1 \quad \Big(= \Big\| \sum_i \gamma_i \phi_i \Big\|_1 \text{ by (4.4)}\Big) \end{aligned}$$

$$\leq (1+\epsilon)^{1/2}\|\alpha\|_2 \leq \frac{(1+\epsilon)^{1/2}}{(1-\eta)} \cdot \sqrt{\frac{2n}{\pi}}\|\gamma\|_2;$$

taking into account (4.5), we indeed get an ϵ-isometry. Hence, $\Omega_{n,I} \supset G \cap E$.

Now

$$P(G^{\mathrm{c}}) = P\{\max_i |R_i - 1| > \eta\}.$$

By (4.3), we may also view $|R_i - 1|$ as a function of n i.i.d. standard normal random variables, Lipschitz with respect to the standard Euclidean metric, with Lipschitz constant $1/\sqrt{n}$. This gives

$$P\{\max_i |R_i - 1| > \eta\} \leq 2m \exp\left\{-n\frac{\eta^2}{2}\right\} = 2m \exp\{-n\beta_G(\epsilon)\}. \tag{4.6}$$

Combining these, we get that on $|I| < n\rho$,

$$P(\Omega^{\mathrm{c}}_{n,I}) \leq P(E^{\mathrm{c}}) + P(G^{\mathrm{c}}) \leq 2\exp(-\beta(\epsilon)n) + 2m\exp(-\beta_G(\epsilon)n).$$

We note that $\beta_G(\epsilon)$ will certainly be larger than $\beta(\epsilon)$.

5 Sign Pattern Embeddings

Our claim for Ω^3_n in the proof of Theorem 2.4 asserts that if $|I| < \rho_3 n$, then

$$\|\delta_{I^{\mathrm{c}}}\|_1 \geq \eta_3\|v\|_1 \tag{5.1}$$

whenever v belongs to $V_I = \mathrm{range}(\Phi_I)$ and $\delta_{I^{\mathrm{c}}}$ obeys $v = -\Phi_{I^{\mathrm{c}}}\delta_{I^{\mathrm{c}}}$.

Here is a brief set of slogans that may help render this inequality plausible. Recall that the ℓ_1-norm is a measure of the sparsity of a vector. The subspace V_I is a random variable independent of $(\phi_i : i \in I^{\mathrm{c}})$. A typical $v \in V_I$ may be viewed as a random vector in $\mathbb{R}^n$, with direction $v/\|v\|_2$ uniformly distributed in $\mathbb{R}^n$. Attempting to represent such a random vector in $\mathbb{R}^n$ by a linear combination of independent random vectors $(\phi_i : i \in I^{\mathrm{c}})$ is not a very good idea; one will not get coefficients with a very small ℓ_1-norm. Hence (5.1) says that such a representation is not materially better than simply representing v in terms of the standard unit vector basis in $\mathbb{R}^n$; the ℓ_1-norm of the representation in the standard unit vector basis will be roughly as good as the ℓ_1-norm in the random dictionary. In the terminology of signal processing, equation (5.1) says that the dictionary of unit basis vectors is just as good at sparsely representing "noise" v as any other dictionary.

Such slogans capture only part of the meaning of (5.1). The relation (5.1) is uniform across *all* v in the subspace V_I, simultaneously for every I with $|I| < \rho_3 n$. We hence turn to a more formal approach.

The left side of (5.1) is lower-bounded by the value of the linear program

$$\min \|\delta_{I^{\mathrm{c}}}\|_1 \quad \text{subject to} \quad \Phi_{I^{\mathrm{c}}}\delta_{I^{\mathrm{c}}} = -v. \tag{5.2}$$

The value function of this linear program, $\mathrm{val}_{I^{\mathrm{c}}}(v) : \mathbb{R}^n \mapsto \mathbb{R}$, is a norm on $\mathbb{R}^n$. We plan to establish (5.1) by showing that $\mathrm{val}_{I^{\mathrm{c}}}(v) \geq \eta_3\|v\|_1$ for all v in the linear subspace V_I. Thus we approach (5.1) as norm comparison on a subspace.

The norm comparison problem is related to an established topic in geometric functional analysis: Milman's *quotient of a subspace* theorem [24]. That theorem states (in an equivalent form, the "subspace of a quotient" theorem) that any Banach space has a certain quotient and a certain subspace that offers an ϵ-Euclidean section of the quotient norm. Here *quotient norm* means precisely a norm of the form (5.2) above. As a result, the norm comparison (5.1) can be viewed as saying that, typically, on each subspace V_I, the quotient norm val_{I^c} and the ℓ_1-norm are each equivalent to the Euclidean norm. This is obviously related to the development of Section 4 above, and it seems clear that one could proceed along the lines of Section 4 to prove (5.1). Instead, we develop an argument that is useful in understanding why other algorithms don't work in this setting (Section 7 below).

We reexpress the norm comparison problem using the duality theorem for linear programming. The left side of (5.1) is lower-bounded by the value of the primal program (5.2), which is at least the value of the dual

$$\max\langle v, \xi\rangle \quad \text{subject to} \quad |\langle \phi_i, \xi\rangle| \le 1, \ \ i \in I^{\mathrm{c}}.$$

Lemma 5.1, stated below, gives us a supply of dual feasible vectors and hence a lower bound on the dual program. Its conclusion, (5.3), tells us that for all sufficiently small $\epsilon > 0$, for every sign vector $\sigma = \mathrm{sgn}(v)$ arising from $v \in \mathrm{range}(\Phi_I)$ there exists a ξ that is dual feasible and a relatively good approximation to $\epsilon\sigma$: $\|\xi - \epsilon\sigma\| < \epsilon\delta(\epsilon)\|\sigma\|_2$, with $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. Using this, we get

$$\langle v, \xi\rangle \ge \langle v, \epsilon\sigma\rangle - \|\xi - \epsilon\sigma\|_2\,\|v\|_2 \ge \epsilon\|v\|_1 - \epsilon\delta(\epsilon)\|\sigma\|_2\,\|v\|_2;$$

picking ϵ sufficiently small and taking into account the spherical-sections theorem, we arrange that $\delta(\epsilon)\|\sigma\|_2\|v\|_2 \le \frac{1}{4}\|v\|_1$ uniformly over $v \in V_I$ where $|I| < \rho_3 n$; (5.1) follows with $\eta_3 = \frac{3}{4}\epsilon$.

It remains to state and prove Lemma 5.1. With I any collection of indices in $\{1, \dots, m\}$, $\mathrm{range}(\Phi_I)$ is a linear subspace of $\mathbb{R}^n$, and on this subspace a subset Σ_I of possible *sign patterns* can be realized, i.e., sequences in $\{\pm 1\}^n$ generated by

$$\sigma(k) = \mathrm{sgn}\left\{\sum_I \alpha_i \phi_i(k)\right\}, \qquad 1 \le k \le n.$$

We need to show that for every $v \in \mathrm{range}(\Phi_I)$, some approximation ξ to $\mathrm{sgn}(v)$ satisfies $|\langle \xi, \phi_i\rangle| \le 1$ for $i \in I^{\mathrm{c}}$.

LEMMA 5.1 (Simultaneous sign pattern embedding) *Positive functions $\delta(\epsilon)$ and $\rho_3(\epsilon; \tau)$ can be defined on $(0, \epsilon_0)$ so that $\delta(\epsilon) \to 0$ as $\epsilon \to 0$, yielding the following properties: For each $\epsilon < \epsilon_0$, there is an event $\Omega_n^3 (\equiv \Omega_{n,\epsilon}^3)$ with*

$$P(\Omega_n^3) \to 1, \quad n \to \infty.$$

On this event, for every subset I with $|I| < \rho_3 n$ and every sign pattern $\sigma \in \Sigma_I$, there is a vector ξ $(\equiv \xi_\sigma)$ with

$$\|\xi - \epsilon\sigma\|_2 \le \epsilon \cdot \delta(\epsilon) \cdot \|\sigma\|_2 \tag{5.3}$$

and

$$|\langle \phi_i, \xi \rangle| \leq 1, \quad i \in I^c. \tag{5.4}$$

In words, a small multiple $\epsilon\sigma$ of any sign pattern σ almost lives in the dual ball $\{\xi : |\langle \phi_i, \xi \rangle| \leq 1\}$. The key aspects are the *proportional dimension* of the constraint ρn and the *proportional distortion* required to fit in the dual ball.

5.1 Proof of Simultaneous Sign Pattern Embedding

The proof introduces a function $\beta(\epsilon)$, positive on $(0, \epsilon_0)$, which places a constraint on the size of ϵ allowed. The bulk of the effort concerns the following lemma, which demonstrates approximate embedding of a *single* sign pattern in the dual ball. The β-function allows us to cover many individual such sequences, producing our result.

LEMMA 5.2 (Individual sign pattern embedding) *Let* $\sigma \in \{-1, 1\}^n$, $\epsilon > 0$, *and* $\xi_0 = \epsilon\sigma$. *There is an iterative algorithm, described below, producing a vector* ξ *as output, that obeys*

$$|\langle \phi_i, \xi \rangle| \leq 1, \quad i = 1, \dots, m. \tag{5.5}$$

Let $(\phi_i)_{i=1}^m$ *be i.i.d. uniform on* $\mathbb{S}^{n-1}$; *there is an event* $\Omega_{\sigma,\epsilon,n}$, *described below, having probability controlled by*

$$\operatorname{Prob}(\Omega^c_{\sigma,\epsilon,n}) \leq 2n \exp\{-n\beta(\epsilon)\} \tag{5.6}$$

for a function $\beta(\epsilon)$ *that can be explicitly given and is positive for* $0 < \epsilon < \epsilon_0$. *On this event,*

$$\|\xi - \xi_0\|_2 \leq \delta(\epsilon) \cdot \|\xi_0\|_2, \tag{5.7}$$

where $\delta(\epsilon) = \epsilon/\sqrt{1-\epsilon^2}$.

In short, a single sign pattern, "shrunken" appropriately, obeys (5.5) with overwhelming probability (see (5.6)) after a slight modification (indicated by (5.7)). Lemma 5.2 will be proven in a section of its own. We now show that it implies Lemma 5.1.

LEMMA 5.3 *Let* $V = \operatorname{range}(\Phi_I) \subset \mathbb{R}^n$. *The number of different sign patterns* σ *generated by vectors* $v \in V$ *obeys*

$$\#\Sigma_I \leq \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{|I|}.$$

PROOF: This is known to statisticians as a consequence of the Vapnik-Chervonenkis (VC) class theory. See Pollard [25, chap. 4]. □

Let again $H(p) = p\log(1/p) + (1-p)\log(1/(1-p))$ be the Shannon entropy. Notice that if $|I| < \rho n$, then

$$\log(\#\Sigma_I) \leq nH(\rho)(1 + o(1)),$$

while also

$$\log \#\{I : |I| < \rho n,\ I \subset \{1, \dots, m\}\} \le n \cdot \tau \cdot H\left(\frac{\rho}{\tau}\right) \cdot (1 + o(1)).$$

Hence, the total number of all sign patterns generated by all operators Φ_I obeys

$$\log \#\{\sigma : \sigma \in \Sigma_I,\ |I| < \rho n\} \le n\left(H(\rho) + \tau H\left(\frac{\rho}{\tau}\right)\right)(1 + o(1)).$$

Now, the function $\beta(\epsilon)$ introduced in Lemma 5.2 is positive, and $H(p) \to 0$ as $p \to 0$. Hence, for each $\epsilon \in (0, \epsilon_0)$, there is $\rho_3(\epsilon) > 0$ obeying

$$H(\rho_3) + \tau H\left(\frac{\rho_3}{\tau}\right) < \beta(\epsilon).$$

Define

$$\Omega_n^3 = \bigcap_{|I| < \rho_3 n} \bigcap_{\sigma \in \Sigma_I} \Omega_{\sigma,I},$$

where $\Omega_{\sigma,I}$ denotes the instance of the event (called $\Omega_{\sigma,\epsilon,n}$ in Lemma 5.2) generated by a specific combination of σ and I. On the event Ω_n^3, *every* sign pattern associated with *any* Φ_I obeying $|I| < \rho_3 n$ is almost dual feasible. Now

$$\begin{aligned} &P((\Omega_n^3)^{\mathsf{c}}) \\ &\quad \le \sum_{|I| < \rho_3 n} \sum_{\sigma \in \Sigma_I} P(\Omega_{\sigma,I}^{\mathsf{c}}) \\ &\quad \le \exp\left\{ n\left(H(\rho_3) + \tau H\left(\frac{\rho_3}{\tau}\right)\right)(1 + o(1))\} \cdot \exp\{-n\beta(\epsilon)(1 + o(1))\right\} \\ &\quad = \exp\left\{ -n\left(\beta(\epsilon) - \left(H(\rho_3) + \tau H\left(\frac{\rho_3}{\tau}\right)\right)\right)(1 + o(1))\right\} \to 0, \quad n \to \infty. \end{aligned}$$

5.2 Proof of Individual Sign Pattern Embedding

An Embedding Algorithm

We now develop an algorithm to create a dual feasible point y starting from a nearby almost-feasible point ξ_0. It is an instance of the successive projection method for finding feasible points for systems of linear inequalities [1].

Let I_0 be the collection of indices $1 \le i \le m$ with

$$|\langle \phi_i, \xi_0 \rangle| > \frac{1}{2},$$

and then set

$$\xi_1 = \xi_0 - P_{I_0}\xi_0,$$

where P_{I_0} denotes the least-squares projector $\Phi_{I_0}(\Phi_{I_0}^{\mathsf{T}}\Phi_{I_0})^{-1}\Phi_{I_0}^{\mathsf{T}}$. In effect, we identify the components where ξ_0 exceeds half the forbidden level $|\langle \phi_i, \xi_0 \rangle| > 1$, and we "kill" those components. Repeat the process, this time on ξ_1, and with a new threshold $t_1 = \frac{3}{4}$. Let I_1 be the collection of indices $1 \le i \le m$ where

$$|\langle \phi_i, \xi_1 \rangle| > \frac{3}{4},$$

and set

$$\xi_2 = \xi_0 - P_{I_0 \cup I_1} \xi_0,$$

again killing the "offending" subspace. Continue in the obvious way, producing ξ_3, ξ_4, etc., with stage-dependent thresholds $t_\ell \equiv 1 - 2^{-\ell-1}$ successively closer to 1. Set

$$I_\ell = \{i : |\langle \phi_i, \xi_\ell \rangle| > t_\ell\},$$

and, putting $J_\ell \equiv I_0 \cup \cdots \cup I_\ell$,

$$\xi_{\ell+1} = \xi_0 - P_{J_\ell} \xi_0.$$

This process can terminate in two ways. First, if J_ℓ has more than n elements, then the process terminates in failure. If I_ℓ is empty, the process terminates successfully, and we set $\xi = \xi_\ell$. Termination must occur at stage $\ell^* \le n$. (In simulations, termination often occurs at $\ell = 1, 2$, or 3). At successful termination,

$$|\langle \phi_i, \xi \rangle| \le 1 - 2^{-\ell^*-1}, \quad i = 1, \ldots, m.$$

Hence ξ is definitely dual feasible. The only question is how close to y_0 it is.

Analysis Framework

In our analysis of the algorithm, we will study

$$\alpha_\ell = \|\xi_\ell - \xi_{\ell-1}\|_2$$

and

$$|I_\ell| = \#\{i : |\langle \phi_i, \xi_\ell \rangle| > 1 - 2^{-\ell-1}\}.$$

We will propose upper bounds $\mu_{\ell;\epsilon,n}$ and $\nu_{\ell;\epsilon,n}$ for these quantities, of the form

$$\mu_{\ell;\epsilon,n} = \|\xi_0\|_2 \cdot \epsilon^\ell \quad (= \sqrt{n}\epsilon^{\ell+1}),$$

$$\nu_{\ell;\epsilon,n} = n \cdot \lambda_0 \cdot \frac{\epsilon^{2\ell+4}}{4};$$

here λ_0 can in principle be taken as any number in $(0, 1)$, but here and below we always take it as $\frac{1}{2}$. This choice determines the range $(0, \epsilon_0)$ for ϵ and restricts the upper limit on ρ. We define subevents

$$E_\ell = \{\alpha_j \le \mu_j,\ j = 1, \ldots, \ell;\ |I_j| \le \nu_j,\ j = 0, \ldots, \ell - 1\}.$$

Now define

$$\Omega_{\sigma,\epsilon,n} = \bigcap_{\ell=1}^{n} E_\ell;$$

this event implies, first of all, that the embedding algorithm terminates successfully:

$$|J_{\ell^*}| \le \sum \nu_\ell \le n \cdot \frac{\lambda_0}{4} \cdot \sum_{\ell=0}^{\infty} \epsilon^{2\ell+4} \le n \cdot \frac{\lambda_0}{4} \cdot \frac{\epsilon^4}{1-\epsilon^2} < n.$$

It also implies that

$$\|\xi - \xi_0\|_2 \le \Bigl(\sum \alpha_\ell^2\Bigr)^{1/2} \le \Bigl(\sum \mu_\ell^2\Bigr)^{1/2} \le \|\xi_0\|_2 \cdot \frac{\epsilon}{(1-\epsilon^2)^{1/2}}.$$

To define the quantities ϵ_0 and β in the statement of Lemma 5.2, we need four further definitions. Recall the quantity $\eta_1(\rho, \tau)$ from Lemma 3.1. Define $\epsilon_1 > 0$ as the solution to $\eta_1(\lambda_0 \epsilon_1^2, \tau)^2 = \lambda_0$. Then $\eta_1(\lambda_0 \epsilon^2, \tau)^2 \ge \lambda_0$ for all $\epsilon \in (0, \epsilon_1]$. Let $\epsilon_2 = \epsilon_2(\tau)$ be a solution to

$$\epsilon^4 \log\left(\frac{\epsilon^2}{64\tau}\right) = -\frac{1}{64},$$

and define the function β by

$$\beta(\epsilon; \tau) = \frac{\epsilon^2}{32} - 2\tau e^{-1/(64\epsilon^4)};$$

then $\beta(\epsilon) > 0$ for $0 < \epsilon < \epsilon_2$.

We now restrict ourselves to choosing ϵ in the range $\epsilon < \epsilon_0 = \min(\epsilon_1, \epsilon_2)$. We will show that for $\ell = 1, 2, \dots,$

$$P(E_{\ell+1}^{\mathsf{c}} \mid E_\ell) \le 2\exp\{-\beta(\epsilon)n\}. \tag{5.8}$$

This implies

$$P(\Omega_{\sigma,\epsilon,n}^{\mathsf{c}}) \le 2n\exp\{-\beta(\epsilon)n\},$$

and the lemma follows.

Transfer to Gaussianity

As in Sections 3 and 4, we again Gaussianize. Let ψ_i denote random vectors in $\mathbb{R}^n$ that are i.i.d. $N(0, \frac{1}{n}I_n)$. We will analyze the algorithm below as if the ψ's rather than the ϕ's made up the columns of Φ.

As already described in Sections 3.1 and 4.2, there is a natural coupling between spherical ϕ's and Gaussian ψ's that justifies this transfer. As earlier, let R_i, $i = 1, \dots, m$, be i.i.d. random variables that are independent of (ϕ_i) and individually $\chi_n/\sqrt{n}$. Then define

$$\psi_i = R_i \phi_i, \quad i = 1, \dots, m.$$

If the ϕ_i are uniform on $\mathbb{S}^{n-1}$, then the ψ_i are indeed $N(0, \frac{1}{n}I_n)$. The R_i are all quite close to 1 for large n. According to (4.6), for fixed $\eta > 0$,

$$P\{1 - \eta < R_i < 1 + \eta,\ i = 1, \dots, m\} \ge 1 - 2m\exp\left\{-n\frac{\eta^2}{2}\right\}.$$

Hence it should be plausible that the difference between the ϕ_i and the ψ_i is immaterial. Arguing more formally, we notice the equivalence

$$|\langle \phi_i, \xi \rangle| < 1 \Leftrightarrow |\langle \psi_i, \xi \rangle| < R_i.$$

Running the algorithm using the ψ's instead of the ϕ's, with thresholds calibrated to $1-\eta$ via $t_0 = (1-\eta)/2$, $t_1 = (1-\eta)\cdot 3/4$, etc., will produce a result y obeying $|\langle \psi_i, \xi \rangle| < 1-\eta\ \forall i$. Therefore, with overwhelming probability, the result will also obey $|\langle \phi_i, \xi \rangle| < 1\ \forall i$.

However, such rescaling of thresholds is completely equivalent to rescaling of the input ξ_0 from $\epsilon\sigma$ to $\epsilon'\sigma$, where $\epsilon' = \epsilon(1-\eta)$. Hence, if we can prove results with functions $\delta(\epsilon)$ and $\beta(\epsilon)$ for the Gaussian ψ's, the same results are proven for the spherical ϕ's with functions $\delta'(\epsilon) = \delta(\epsilon') = \delta(\epsilon(1-\eta))$ and $\beta'(\epsilon) = \min(\beta(\epsilon'), \eta^2/2)$.

Adapted Coordinates

It will be useful to have coordinates specially adapted to the analysis of the algorithm. Given $\xi_0, \xi_1, \dots$, define $\zeta_0, \zeta_1, \dots$ by Gram-Schmidt orthonormalization. In terms of these coordinates we have the following equivalent construction: Let $\alpha_0 = \|\xi_0\|_2$, and let ξ_i, $1 \le i \le m$, be i.i.d. vectors $N(0, \frac{1}{n} I_n)$. We will sequentially construct vectors ψ_i, $i = 1, \dots, m$, in such a way that their joint distribution is i.i.d. $N(0, \frac{1}{n} I_n)$, but so that the algorithm has an explicit trajectory.

At stage 0, we realize m scalar Gaussians $Z_i^0 \sim^{\text{i.i.d.}} N(0, \frac{1}{n})$, threshold at level t_0, say, and define I_0 to be the set of indices so that $|\alpha_0 Z_i^0| > t_0$. For such indices i only, we define

$$\psi_i = Z_i^0 \zeta_0 + P_{\zeta_0}^{\perp} \xi_i, \quad i \in I_0.$$

For all other i, we retain Z_i^0 for later use. We then define $\xi_1 = \xi_0 - P_{I_0}\xi_0$, $\alpha_1 = \|\xi_1 - \xi_0\|_2$, and ζ_1 by orthonormalizing $\xi_1 - \xi_0$ with respect to ζ_0.

At stage 1, we realize m scalar Gaussians $Z_i^1 \sim^{\text{i.i.d.}} N(0, \frac{1}{n})$, and define I_1 to be the set of indices not in I_0 so that $|\alpha_0 Z_i^0 + \alpha_1 Z_i^1| > t_1$. For such indices i only, we define

$$\psi_i = Z_i^0 \zeta_0 + Z_i^1 \zeta_1 + P_{\zeta_0,\zeta_1}^{\perp} \xi_i, \quad i \in I_1.$$

For i neither in I_0 nor I_1, we retain Z_i^1 for later use. We then define $\xi_2 = \xi_0 - P_{I_0 \cup I_1}\xi_0$, $\alpha_2 = \|\xi_2 - \xi_1\|_2$, and ζ_2 by orthonormalizing $\xi_2 - \xi_1$ with respect to ζ_0 and ζ_1.

Continuing in this way, at some stage ℓ^* we stop, (i.e., I_{ℓ^*} is empty), and we define ψ_i for all i not in $I_0 \cup \dots \cup I_{\ell^*-1}$ (if there are any such) by

$$\psi_i = \sum_{j=0}^{\ell^*-1} Z_i^j \zeta_j + P_{\zeta_0,\dots,\zeta_{\ell^*-1}}^{\perp} \xi_i, \quad i \notin I_0 \cup \dots \cup I_{\ell^*-1}.$$

We claim that we have produced a set m of i.i.d. $N(0, \frac{1}{n} I_n)$'s for which the algorithm has the indicated trajectory we have just traced. A proof of this fact repeatedly uses independence properties of orthogonal projections of standard normal random vectors.

It is immediate that, for each ℓ up to termination, we have expressions for the key variables in the algorithm in terms of the coordinates. For example:

$$\xi_\ell - \xi_0 = \sum_{j=1}^{\ell} \alpha_j \zeta_j; \qquad \|\xi_\ell - \xi_0\|_2 = \left(\sum_{j=1}^{\ell} \alpha_j^2 \right)^{1/2}.$$

Control on $\boldsymbol{\alpha_\ell}$

We now develop a bound for

$$\alpha_{\ell+1} = \|\xi_{\ell+1} - \xi_\ell\|_2 = \|P_{I_\ell}(\xi_{\ell+1} - \xi_\ell)\|_2.$$

Recalling that

$$P_{I_\ell} v = \Psi_{I_\ell} (\Psi_{I_\ell}^{\mathsf{T}} \Psi_{I_\ell})^{-1} \Psi_{I_\ell}^{\mathsf{T}} v$$

and putting $\lambda(I_\ell) = \lambda_{\min}(\Psi_{I_\ell}^{\mathsf{T}} \Psi_{I_\ell})$, we have

$$\|P_{I_\ell}(\xi_{\ell+1} - \xi_\ell)\|_2 \le \lambda(I_\ell)^{-1/2} \|\Psi_{I_\ell}^{\mathsf{T}}(\xi_{\ell+1} - \xi_\ell)\|_2.$$

But $\Psi_{I_\ell}^{\mathsf{T}} y_{\ell+1} = 0$ because $\xi_{\ell+1}$ is orthogonal to every ψ_i, $i \in I_\ell$, by construction. Now for $i \in I_\ell$,

$$|\langle \psi_i, \xi_\ell \rangle| \le |\langle \psi_i, \xi_\ell - \xi_{\ell-1} \rangle| + |\langle \psi_i, \xi_{\ell-1} \rangle| \le \alpha_\ell |Z_i^\ell| + t_\ell,$$

and so

$$\left\| \Psi_{I_\ell}^{\mathsf{T}} \xi_\ell \right\|_2 \le t_\ell |I_\ell|^{1/2} + \alpha_\ell \left(\sum_{i \in I_\ell} (Z_i^\ell)^2 \right)^{1/2}. \tag{5.9}$$

We remark that

$$\{i \in I_\ell\} \Rightarrow \{|\langle \psi_i, \xi_\ell \rangle| > t_\ell,\ |\langle \psi_i, \xi_{\ell-1} \rangle| < t_{\ell-1}\} \Rightarrow \{|\langle \psi_i, \xi_\ell - \xi_{\ell-1} \rangle| \ge t_\ell - t_{\ell-1}\};$$

putting $u_\ell = (t_\ell - t_{\ell-1})/\alpha_\ell = 2^{-\ell-1}/\alpha_\ell$ and noting $\langle \psi_i, \xi_\ell - \xi_{\ell-1} \rangle = \alpha_\ell Z_i^\ell$, this gives

$$\sum_{i \in I_\ell} (Z_i^\ell)^2 \le \sum_{i \in J_{\ell-1}^{\mathsf{c}}} (Z_i^\ell)^2 1_{\{|Z_i^\ell| > u_\ell\}}.$$

We conclude from $(a+b)^2 \le 2(a^2 + b^2)$ and $t_\ell < 1$ that

$$\alpha_{\ell+1}^2 \le 2 \cdot \lambda(I_\ell)^{-1} \left[|I_\ell| + \alpha_\ell^2 \left(\sum_{i \in J_{\ell-1}^{\mathsf{c}}} (Z_i^\ell)^2 1_{\{|Z_i^\ell| > u_\ell\}} \right) \right]. \tag{5.10}$$

Large Deviations

Define the events

$$F_\ell = \{\alpha_\ell \le \mu_{\ell;\epsilon,n}\}, \qquad G_\ell = \{|I_\ell| \le \nu_{\ell;\epsilon,n}\},$$

so that

$$E_{\ell+1} = F_{\ell+1} \cap G_\ell \cap E_\ell, \quad \ell = 1, 2, \dots,$$

and $E_1 = F_1 \cap G_0$. We wish to show that (5.8) holds; our approach will be to establish pairs of bounds:

$$P\{F^{\mathsf{c}}_{\ell+1} \mid G_\ell, E_\ell\} \le \exp(-n\beta(\epsilon)), \quad P\{G^{\mathsf{c}}_\ell \mid E_\ell\} \le \exp(-n\beta(\epsilon)), \tag{5.11}$$

which of course combine to give (5.8). Put

$$\rho_0(\epsilon) = \lambda_0 \epsilon^2.$$

In the event E_ℓ, $|J_\ell| \le \rho_0(\epsilon)n$. Recall the definition of ϵ_1. Since $\epsilon_1 \ge \epsilon_0 > \epsilon$, $\eta_1(\rho_0(\epsilon), \tau)^2 \ge \lambda_0$. On E_ℓ, $\lambda(I_\ell) \ge \lambda_0$. Also on E_ℓ, $u_j = 2^{-j-1}/\alpha_j > 2^{-j-1}/\mu_j = v_j$ (say) for $j \le \ell$. Now

$$\begin{aligned} P\{F^{\mathsf{c}}_{\ell+1} \mid G_\ell, E_\ell\} &\le P\Big\{2 \cdot \lambda_0^{-1}\Big[\nu_\ell + \mu_\ell^2\Big(\sum_i (Z_i^\ell)^2 1_{\{|Z_i^\ell| > v_\ell\}}\Big)\Big] > \mu_{\ell+1}^2\Big\} \\ &= P\Big\{\sum_i (Z_i^\ell)^2 1_{\{|Z_i^\ell| > v_\ell\}} > \Delta_\ell\Big\}, \end{aligned} \tag{5.12}$$

where

$$\Delta_\ell = \frac{\lambda_0 \mu_{\ell+1}^2/2 - \nu_\ell}{\mu_\ell^2} = \frac{\epsilon^2}{8}.$$

Also,

$$P\{G^{\mathsf{c}}_\ell \mid E_\ell\} \le P\Big\{\sum_i 1_{\{|Z_i^\ell| > v_\ell\}} > \nu_\ell\Big\}. \tag{5.13}$$

Both (5.12) and (5.13) require us to bound the probability that sums of i.i.d. random variables exceed certain thresholds. We need two simple large-deviations bounds.

LEMMA 5.4 *Let Z_i be i.i.d. $N(0,1)$, $k \ge 0$, $t > 2$, $\Delta > 0$. Then*

$$\frac{1}{m-k} \log P\Big\{\sum_{i=1}^{m-k} Z_i^2 1_{\{|Z_i| > t\}} > (m-k)\Delta\Big\} \le e^{-t^2/4} - \frac{\Delta}{4}$$

and

$$\frac{1}{m-k} \log P\Big\{\sum_{i=1}^{m-k} 1_{\{|Z_i| > t\}} > (m-k)\Delta\Big\} \le e^{-t^2/2} - \frac{\Delta}{4}.$$

Note that the random variables Z_i^ℓ are $N(0, 1/n)$, whereas the random variables Z_i in the lemma are standardized. Taking this into account,

$$\frac{1}{n} \log P\{F_{\ell+1}^{\mathsf{c}} \mid G_\ell, E_\ell\} \le 2\tau e^{-t_\ell^2/4} - \frac{\Delta_\ell}{4},$$

where

$$t_\ell^2 = n \cdot v_\ell^2 = \frac{2^{-2\ell-2}}{\epsilon^{2+2\ell}}.$$

Note that $\Delta_{\ell'} = \Delta_\ell$ for $\ell \neq \ell'$, and since $\epsilon < \frac{1}{2}$, $t_{\ell+1} > t_\ell$,

$$\exp\left\{\frac{-1}{4(2\epsilon)^4}\right\} = \exp\left\{-\frac{t_1^2}{4}\right\} \ge \exp\left\{-\frac{t_\ell^2}{4}\right\}, \quad \ell > 1.$$

Hence

$$2\tau e^{-t_\ell^2/4} - \frac{\Delta_\ell}{4} \le \beta(\epsilon; \tau), \quad \ell = 1, 2, \ldots.$$

We conclude that

$$P\{F_{\ell+1}^{\mathsf{c}} \mid G_\ell, E_\ell\} \le \exp(-n\beta(\epsilon)).$$

Hence the first half of (5.11) holds. A similar analysis holds for the G_ℓ's and for E_1.

Remark 5.5. The large-deviations bounds stated in Lemma 5.4 are far from the best possible; we merely found them convenient in producing an explicit expression for β. Better bounds would be helpful in deriving reasonable estimates on the constant $\rho^*(\tau)$ in Theorem 2.4.

6 Polytope Interpretation

Our result has an appealing interpretation from the viewpoint of convex polytopes. Let Q^n denote the absolute convex hull of $\phi_1, \ldots, \phi_m$,

$$Q^n = \left\{ y \in \mathbb{R}^n : y = \sum_i x(i)\phi_i, \ \sum_i |x(i)| \le 1 \right\}.$$

Equivalently, Q^n is exactly the set of vectors where $\mathrm{val}(P_1) \le 1$. Similarly, let the cross-polytope (a.k.a. octahedron) $C^m \in \mathbb{R}^m$ be the absolute convex hull of the standard Kronecker basis $(e_i)_{i=1}^m$:

$$C^m = \left\{ x \in \mathbb{R}^m : x = \sum_i x(i)e_i, \ \sum_{i=1}^m |x(i)| \le 1 \right\}.$$

Note that each set is a convex polytope, and it is almost true that the vertices $\{\pm e_i\}$ of C^m map under Φ into vertices $\{\pm\phi_i\}$ of Q^n. More precisely, the vertices of Q^n are among the image vertices $\{\pm\phi_i\}$; because Q^n is a convex hull, there is the possibility that for some i, ϕ_i lies strictly in the interior of Q^n.

Now if ϕ_i were strictly in the interior of Q^n, then we could write

$$\phi_i = \Phi\alpha_1, \quad \|\alpha_1\|_1 < 1,$$

where $i \notin \mathrm{supp}(\alpha_1)$. It would follow that a singleton α_0 generates ϕ_i through $\phi_i = \Phi\alpha_0$, so α_0 necessarily solves (P_0), but, because

$$\|\alpha_0\|_1 = 1 > \|\alpha_1\|_1,$$

α_0 is not the solution of (P_1). So, when any ϕ_i is strictly in the interior of Q^n, (P_1) and (P_0) are inequivalent problems.

Now on the event $\Omega_n(\rho^*, \tau)$, the matrix Φ has the property that (P_1) and (P_0) share the same solution whenever (P_0) has a solution with $k = 1 < \rho^* n$ nonzeros. We conclude that in the event $\Omega_n(\rho^*, \tau)$, *the vertices of Q^n are in one-to-one correspondence with the vertices of C^m*. Letting $\mathrm{skel}_0(Q)$ denote the set of vertices of a polytope Q, this correspondence says

$$\mathrm{skel}_0(\Phi C^m) = \Phi[\mathrm{skel}_0(C^m)].$$

Something much more general is true. By a $(k-1)$–*simplex* we mean a set

$$\Sigma(v_1, \dots, v_k) = \left\{ x = \sum_j \alpha_j v_{i_j},\ \alpha_j \ge 0,\ \sum \alpha_j = 1 \right\}$$

where the vectors $v_i \in \mathbb{R}^n$ are affinely independent. By the $(k-1)$–*face* of a polytope Q with vertex set $v = \{v_1, \dots\}$, we mean a $(k-1)$–simplex $\Sigma(v_{i_1}, \dots, v_{i_k})$, which is equal to the intersection of a Q with a supporting hyperplane of Q. By $(k-1)$–*skeleton* $\mathrm{skel}_{k-1}(Q)$ of a polytope Q, we mean the collection of all $(k-1)$–faces.

The 0-skeleton is the set of vertices, the 1-skeleton is the set of edges, etc. In general, one can say that the $(k-1)$–faces of Q^n form a subset of the images under Φ of the $(k-1)$–faces of O_n:

$$\mathrm{skel}_{k-1}(\Phi C^m) \subset \Phi[\mathrm{skel}_{k-1}(C^m)], \quad 1 \le k < n.$$

Indeed, some of the image faces $\Phi\Sigma(\pm e_{i_1}, \dots, \pm e_{i_k})$ could be at least partially interior to Q^n, and hence they could be not part of the $(k-1)$–skeleton of Q^n.

Our main result says that much more is true; Theorem 2.4 is equivalent to this geometric statement:

THEOREM 6.1 *There is a constant $\rho^* = \rho^*(\tau) > 0$ so that for $n < m < \tau n$, on an event $\Omega_n(\rho^*, \tau)$ whose complement has negligible probability for large n,*

$$\mathrm{skel}_{k-1}(\Phi C^m) = \Phi[\mathrm{skel}_{k-1}(C^m)], \quad 1 \le k < \rho^* \cdot n.$$

In particular, with overwhelming probability, *the topology of the $(k-1)$–skeleton of Q^n is the same as for the corresponding $(k-1)$–skeleton of C^m*, even for k proportional to n. The topology of the skeleton of C^m is of course obvious.

We do not prove here that Theorem 2.4 is equivalent to Theorem 6.1. This equivalence is discussed at length in the article [7].

7 Other Algorithms Fail

Several algorithms besides ℓ_1-minimization have been proposed for the problem of finding sparse solutions [9, 23, 28]. In this section we show that two standard approaches fail in the current setting where ℓ_1 of course succeeds.

7.1 Subset Selection Algorithms

Consider two algorithms that attempt to find sparse solutions to $y = \Phi x$ by selecting subsets I and then attempting to solve $y = \Phi_I x_I$.

The first is *simple thresholding*. One sets a threshold t and selects a subset $\hat{I}$ of terms "highly correlated with y,"

$$\hat{I} = \{i : |\langle y, \phi_i \rangle| > t\},$$

and then attempts to solve $y \approx \Phi_{\hat{I}} x_{\hat{I}}$. Statisticians have been using methods like this on noisy data for many decades; the approach is sometimes called "subset selection by preliminary significance testing from univariate regressions."

The second is *greedy subset selection*. One selects a subset iteratively, starting from $r_0 = y$ and $\ell = 0$ and proceeding stagewise, through stages $\ell = 0, 1, \dots$. At the 0th stage, one identifies the best-fitting single term,

$$i_0 = \mathrm{argmax}_i\, |\langle r_0, \phi_i \rangle|,$$

and then, putting $x_{i_0} = \langle r_0, \phi_{i_0} \rangle$, subtracts that term away:

$$r_1 = r_0 - x_{i_0} \phi_{i_0};$$

at stage 1 one behaves similarly, getting $i_1$and r_2, etc. In general,

$$i_\ell = \mathrm{argmax}_i\, |\langle r_{\ell-1}, \phi_i \rangle|$$

and

$$r_\ell = y - P_{i_1,\dots,i_\ell} y.$$

One stops as soon as $r_\ell = 0$. Procedures of this form have been used routinely by statisticians since the 1960s under the name stepwise regression; the same procedure is called "orthogonal matching pursuit" in signal analysis and "greedy approximation" in the approximation theory literature. For further discussion, see [9, 28, 29].

Under sufficiently strong conditions, both methods can work.

THEOREM 7.1 (Tropp [28]) *Suppose that the dictionary Φ has coherence $M = \max_{i \neq j} |\langle \phi_i, \phi_j \rangle|$. Suppose that x_0 has $k \leq M^{-1}/2$ nonzeros, and apply the greedy algorithm to $y = \Phi x_0$. The algorithm will stop after k stages, having selected at each stage one of the terms corresponding to the k nonzero entries in x_0, at the end having precisely found the unique sparsest solution x_0.*

A parallel result can be given for thresholding.

THEOREM 7.2 *Let $\eta \in (0, 1)$. Suppose that x_0 has $k \leq \eta M^{-1}/2$ nonzeros, and that the nonzero coefficients obey $|x_0(i)| \geq (\eta/\sqrt{k})\|x\|_2$ (thus, they are all about the same size). Choose a threshold so that exactly k terms are selected. These k terms will be exactly the nonzeros in x_0, and solving $y = \Phi_{\hat{I}} x_{\hat{I}}$ will recover the underlying optimal sparse solution x_0.*

PROOF: We need to show that a certain threshold which selects exactly k terms selects only terms in I. Consider the preliminary threshold $t_0 = \eta/(2\sqrt{k})\|x_0\|_2$. We have, for $i \in I$,

$$\begin{aligned}
|\langle \phi_i, y\rangle| &= \left| x_i + \sum_{j \neq i} x_0(j)\langle \phi_i, \phi_j \rangle \right| \\
&\geq |x_i| - M \sum_{j \neq i} |x_0(j)| \\
&> |x_i| - M\sqrt{k}\|x_0\|_2 \\
&\geq \|x_0\|_2 \cdot \left(\frac{\eta}{\sqrt{k}} - M\sqrt{k} \right) \\
&\geq \|x_0\|_2 \cdot \frac{\eta}{2\sqrt{k}} = t_0.
\end{aligned}$$

Hence for $i \in I$, $|\langle \phi_i, y\rangle| > t_0$. On the other hand, for $j \notin I$,

$$\begin{aligned}
|\langle \phi_j, y\rangle| &= \left| \sum_i x_0(i)\langle \phi_i, \phi_j \rangle \right| \\
&\leq M\sqrt{k}\|x_0\|_2 \leq t_0.
\end{aligned}$$

Hence, for small enough $\delta > 0$, the threshold $t_\delta = t_0 + \delta$ selects exactly the terms in I. □

7.2 Analysis of Subset Selection

The present article considers situations where the number of nonzeros is proportional to n. As it turns out, this is far beyond the range where previous general results about greedy algorithms and thresholding would work. Indeed, in this article's setting of a random dictionary Φ, we have (see Lemma 2.3) coherence $M \sim \sqrt{2\log(m)}/\sqrt{n}$. Theorems 7.1 and 7.2 therefore apply only for $|I| = o(\sqrt{n}) \ll \rho n$. In fact, it is not merely that the theorems don't apply; the nice behavior asserted in Theorems 7.1 and 7.2 is absent in this more challenging setting.

THEOREM 7.3 *Let n, m, τ, and ρ^* be as above. For a sequence of events Ω_n having overwhelming probability for large n, there is a vector y with unique sparsest representation using at most $k < \rho^* n$ nonzero elements, for which the following are true*:

- *The minimal ℓ_1-norm solution is also the optimally sparse solution.*
- *The thresholding algorithm can only find a solution using n nonzeros.*
- *The greedy algorithm makes a mistake in its first stage, selecting a term not appearing in the optimally sparse solution.*

PROOF: The statement about ℓ_1-minimization is of course just a reprise of Theorem 2.4. The other two claims depend on the following:

LEMMA 7.4 *Let n, m, τ, and ρ^* be as in Theorem* 2.4. *Let* $I = \{1, \dots, k\}$, *where* $\rho^*/2n < k < \rho^* n$. *There exists* $C > 0$ *so that, for each* $\eta_2 > 0$, *for all sufficiently large n, with overwhelming probability* ***some*** $y \in \text{range}(\Phi_I)$ *has* $\|y\|_2 = \sqrt{n}$ *but*

$$|\langle y, \phi_i \rangle| < C, \quad i \in I,$$

and

$$\min_{i \in I} |\langle y, \phi_i \rangle| < \eta.$$

This lemma will be proved in the next subsection. Let's see what it says about thresholding. The construction of y guarantees that it is a random variable independent of ϕ_i, $i \notin I$. With R_i as introduced in (4.3), the coefficients $\langle y, \phi_i \rangle R_i$, $i \in I^{\mathrm{c}}$, are i.i.d. with standard normal distribution, and by (4.6) these differ trivially from $\langle y, \phi_i \rangle$. This implies that for $i \in I^{\mathrm{c}}$, the coefficients $\langle y, \phi_i \rangle$ are i.i.d. with a distribution that is nearly standard normal. In particular, for some $a = a(C) > 0$, with overwhelming probability for large n, we will have

$$\#\{i \in I^{\mathrm{c}} : |\langle y, \phi_i \rangle| > C\} > a \cdot m,$$

and, if η is the parameter used in the invocation of the lemma above, with overwhelming probability for large n we will also have

$$\#\{i \in I^{\mathrm{c}} : |\langle y, \phi_i \rangle| > \eta\} > n.$$

Hence, thresholding will actually select $a \cdot m$ terms not belonging to I *before* any term belonging to I. Also, if the threshold is set so that thresholding selects $< n$ terms, then some terms from I will *not* be among those terms (in particular, the terms where $|\langle \phi_i, y \rangle| < \eta$ for η small).

With probability 1, the points ϕ_i are in general position. Because of Lemma 2.1, we can only obtain a solution to the original equations if one of two things is true:

- We select *all* terms of I.
- We select n terms (and then it doesn't matter which ones).

If *any* terms from I are omitted by the selection $\hat{I}$, we cannot get a sparse representation. Since with overwhelming probability *some* of the k terms appearing in I are not among the n best terms for the inner product with the signal, thresholding does not give a solution until n terms are included, and there must be n nonzero coefficients in the solution obtained.

Now let's see what the lemma says about greedy subset selection. Recall that the $\langle y, \phi_i \rangle R_i$, $i \in I^{\mathrm{c}}$, are i.i.d. with standard normal distribution, and these differ trivially from $\langle y, \phi_i \rangle$. Combining this fact with standard extreme-value theory for i.i.d. Gaussians [19], we conclude that for each $\delta > 0$, with overwhelming probability for large n,

$$\max_{i \in I^{\mathrm{c}}} |\langle y, \phi_i \rangle| > (1-\delta)\sqrt{2\log(m - |I|)}.$$

On the other hand, with overwhelming probability for large n,

$$\max_{i \in I} |\langle y, \phi_i \rangle| < C.$$

Since $\rho^* < \frac{1}{2}$, $m_n - \rho^* n \to \infty$ as $n \to \infty$. Thus $\sqrt{\log(m - |I|)} \gg C$ for large n. Hence with overwhelming probability for large n, the first step of the greedy algorithm will select a term not belonging to I. □

Not proved here but strongly suspected is that there exist y so that greedy subset selection cannot find any exact solution until it has been run for at least n stages.

7.3 Proof of Lemma 7.4

Let $V = \mathrm{range}(\Phi_I)$; pick any orthobasis (ζ_i) for V, and let $Z_1, \dots, Z_k$ be i.i.d. standard Gaussian $N(0, 1)$. Set $v = \sum_i Z_j \zeta_j$. Then for $i \in I$, $\langle \phi_i, v \rangle \sim N(0, 1)$.

Now let (φ_{ij}) be the array defined by $\varphi_{ij} = \langle \phi_i, \zeta_j \rangle$. Note that the φ_{ij} are independent of v and are approximately $N(0, \frac{1}{k})$. (More precisely, with R_i the random variable defined earlier at (4.3), $R_i G_{ij}$ is exactly $N(0, \frac{1}{k})$). Let φ_i denote the vector (φ_{ij}).

The proof of Lemma 5.2 shows that result to have nothing to do with signs; it can be applied to *any* vector rather than some sign pattern vector σ. Make the temporary substitutions $n \leftrightarrow k$, $\sigma \leftrightarrow (Z_j)$, and $\phi_i \leftrightarrow \varphi_i$, and choose $\epsilon > 0$. Apply Lemma 5.2 to σ. Get a vector ξ obeying

$$|\langle \varphi_i, \xi \rangle| \le 1, \quad i = 1, \dots, k. \tag{7.1}$$

Now define

$$y \equiv \frac{\sqrt{n}}{\|\xi\|_2} \cdot \sum_{j=1}^{k} \xi(j) \zeta_j.$$

Lemma 5.2 stipulated an event, E_n, on which the algorithm delivers

$$\|\xi - \epsilon v\|_2 \le \epsilon \delta(\epsilon) \|v\|_2.$$

This event has probability exceeding $1 - \exp\{-\beta n\}$. On this event

$$\|\xi\|_2 \ge \epsilon(1 - \delta(\epsilon)) \|v\|_2.$$

Arguing as in (4.6), the event $F_n = \{\|v\|_2 \ge (1-\eta)\sqrt{k}\}$ has

$$P(F_n^{\mathrm{c}}) \le 2 \exp\left\{-\frac{k\eta^2}{2}\right\}.$$

Hence on an event $E_n \cap F_n$,

$$\|\xi\|_2 \geq \epsilon(1-\delta(\epsilon))(1-\eta)\sqrt{k}.$$

We conclude, using (7.1), that for $i = 1, \ldots, k$,

$$|\langle \phi_i, y\rangle| = \frac{\sqrt{n}}{\|\xi\|_2}|\langle \varphi_i, \xi\rangle| \leq \frac{1}{\epsilon(1-\delta(\epsilon))(1-\eta)\sqrt{\rho^*/2}} \cdot 1 \equiv C, \text{ say.}$$

This is the first claim of the lemma.

For the second claim, notice that this would be trivial if $\langle y, \phi_i\rangle$ were i.i.d. $N(0,1)$. This is not quite true, because of conditioning involved in the algorithm underlying Lemma 5.2. However, an application of Sidák's lemma shows that the conditioning only makes the indicated event even more likely than for an i.i.d. sequence.

8 Breakdown Point Heuristics

In our proof we identify three operative factors in determining the equivalence breakdown point. The point of the proof is, of course, to have a rigorously valid argument identifying the correct qualitative behavior. It seems, however, that the three factors identified in the proof provide a good quantitative understanding of the value ρ at which breakdown of equivalence sets in. In an article [30] by Yaacov Tsaig and the author (an outgrowth of the first version of the paper you are reading), it is shown that a quantitative prediction of the breakdown point can be based on the three factors operating in the proof of Theorem 2.4, and that this prediction is fairly accurate.

9 Stability

Skeptics may object that our discussion of sparse solutions to underdetermined systems is irrelevant because the whole concept cannot be stable. Actually, the concept is stable, as a byproduct of Lemma 3.1. There we showed that, with overwhelming probability for large n, all singular values of every submatrix Φ_I with $|I| < \rho n$ exceed $\eta_1(\rho, \tau)$. Now invoke the following:

THEOREM 9.1 (Donoho, Elad, and Temlyakov [9]) *Let Φ be given and set*

$$\eta(\rho, \Phi) = \min\left\{\lambda_{\min}^{1/2}(\Phi_I^{\mathsf{T}}\Phi_I) : |I| < \rho n\right\}.$$

Suppose we are given the vector $y = \Phi x_0 + z$, $\|z\|_2 \leq \epsilon$, $\|x_0\|_0 \leq \rho n/2$. Consider the optimization problem

$$(\mathrm{P}_{0,\epsilon}) \qquad \min \|x\|_0 \quad \textit{subject to} \quad \|y - \Phi x\|_2 \leq \epsilon,$$

and let $\hat{x}_{0,\epsilon}$ denote any solution. Then

- $\|\hat{x}_{0,\epsilon}\|_0 \leq \|x_0\|_0 \leq \rho n/2$ *and*
- $\|\hat{x}_{0,\epsilon} - x_0\|_2 \leq 2\epsilon/\eta$ *where* $\eta = \eta(\rho, \Phi)$.

Applying Lemma 3.1, we see that the problem of obtaining a sparse approximate solution to noisy data is a stable problem: if the noiseless data have a solution with at most $\rho n/2$ nonzeros, then an error of size $\leq \epsilon$ in measurements can lead to a reconstruction error of size $\leq 2\epsilon/\eta_1(\rho, \tau)$. We stress that we make no claim here about stability of the ℓ_1-reconstruction; only that stability *by some method* is in principle possible. Detailed investigation of stability is being pursued separately.

Acknowledgment. The author received partial support from National Science Foundation grants DMS 00-77261, DMS 01-40698 (FRG), and DMS 05-05303, and from an ONR-MURI project. Thanks to Emmanuel Candès for calling my attention to a blunder in a previous version, to Noureddine El Karoui for discussions about singular values of random matrices, and to Yaacov Tsaig for the experimental results mentioned here.

Bibliography

[1] Bauschke, H. H.; Borwein, J. M. On projection algorithms for solving convex feasibility problems. *SIAM Review* **38** (1996), no. 3, 367–426.

[2] Candès, E. J.; Romberg, J.; Tao, T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, to appear.

[3] Chen, S. S.; Donoho, D. L.; Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** (1998), no. 1, 33–61.

[4] Coifman, R. R.; Meyer, Y.; Quake, S.; Wickerhauser, M. V. Signal processing and compression with wavelet packets. *Progress in wavelet analysis and applications (Toulouse, 1992)*, 77–93. Frontières, Gif-sur-Yvette, 1993.

[5] Coifman, R. R.; Wickerhauser, M.V. Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory* **38** (1992), 713–718.

[6] Davidson, K. R.; Szarek, S. J. Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces*, Vol. 1, 317–366. North-Holland, Amsterdam, 2001.

[7] Donoho, D. L. Neighborly polytopes and sparse solutions of underdetermined linear equations. *IEEE Trans. Inform. Theory*, to appear.

[8] Donoho, D. L.; Elad, M. Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proc. Natl. Acad. Sci. USA* **100** (2003), no. 5, 2197–2202 (electronic).

[9] Donoho, D. L.; Elad, M.; Temlyakov, V. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** (2006), no. 1, 6–18.

[10] Donoho, D. L.; Huo, X. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47** (2001), no. 7, 2845–2862.

[11] Donoho, D. L.; Johnstone, I. M.; Hoch, J. C.; Stern, A. S. Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** (1992), no. 1, 41–81.

[12] Dvoretsky, A. Some results on convex bodies and Banach spaces. *Proc. Internat. Sympos. Linear Spaces (Jerusalem, 1960)*, 123–160. Jerusalem Academic Press, Jerusalem; Pergamon, Oxford, 1961.

[13] El Karoui, N. New results about random covariance matrices and statistical applications. Doctoral dissertation, Stanford University, 2004.

[14] Elad, M.; Bruckstein, A. M. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Inform. Theory* **48** (2002), no. 9, 2558–2567.

[15] Figiel, T.; Lindenstrauss, J.; Milman, V. D. The dimension of almost spherical sections of convex bodies. *Acta Math.* **139** (1977), 53–94.

[16] Fuchs, J. J. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inform. Theory* **50** (2004), no. 6, 1341–1344.

[17] Gribonval, R.; Nielsen, M. Sparse representations in unions of bases. *IEEE Trans. Inform. Theory* **49** (2003), no. 12, 3320–3325.

[18] Johnson, W. B.; Schechtman, G. Embedding ℓ_p^m into ℓ_1^n. *Acta Math.* **149** (1982), 71–85.

[19] Leadbetter, R.; Lindgren, G.; Rootzén, H. *Extremes and related properties of random sequences and processes*. Springer, New York–Berlin, 1983.

[20] Ledoux, M. *The concentration of measure phenomenon.* Mathematical Surveys and Monographs, 89. American Mathematical Society, Providence, R.I., 2001.

[21] Mallat, S.; Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **41** (1993), no. 12, 3397–3415.

[22] Milman, V. D.; Schechtman, G. *Asymptotic theory of finite-dimensional normed spaces.* Lecture Notes in Mathematics, 1200. Springer, Berlin, 1986.

[23] Natarajan, B. K. Sparse approximate solutions to linear systems. *SIAM J. Comput.* **24** (1995), 227–234.

[24] Pisier, G. *The volume of convex bodies and Banach space geometry*. Cambridge Tracts in Mathematics, 94. Cambridge University Press, Cambridge, 1989.

[25] Pollard, D. *Empirical processes: theory and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, Calif.; American Statistical Association, Alexandria, Va., 1990.

[26] Schechtman, G. Random embeddings of Euclidean spaces in sequence spaces. *Israel J. Math.* **40** (1981), no. 2, 187–192.

[27] Szarek, S. J. Spaces with large distance to ℓ_∞^n and random matrices. *Amer. J. Math.* **112** (1990), no. 6, 899–942.

[28] Tropp, J. A. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory* **50** (2004), no. 10, 2231–2242.

[29] Tropp, J. A. Just relax: convex programming methods for subset selection and sparse approximation. *IEEE Trans Inform. Theory*, to appear.

[30] Tsaig, Y.; Donoho, D. L. Breakdown of equivalence between the minimal ℓ^1-norm solution and the sparsest solution. *European J. Appl. Signal Process.*, to appear.

DAVID L. DONOHO
Statistics Department
Stanford University
Stanford, CA 94305
E-mail: donoho@stanford.edu

Received August 2004.
Revised September 2005.