For Most Large Underdetermined Systems of Equations, the Minimal ℓ^1-norm Near-Solution Approximates the Sparsest Near-Solution

David L. Donoho
Department of Statistics
Stanford University

August, 2004

Abstract

We consider inexact linear equations $y \approx \Phi\alpha$ where y is a given vector in $\mathbf{R}^n$, Φ is a given n by m matrix, and we wish to find an $\alpha_{0,\epsilon}$ which is sparse and gives an approximate solution, obeying $\|y - \Phi\alpha_{0,\epsilon}\|_2 \leq \epsilon$. In general this requires combinatorial optimization and so is considered intractable. On the other hand, the ℓ^1 minimization problem

$$\min \|\alpha\|_1 \text{ subject to } \|y - \Phi\alpha\|_2 \leq \epsilon,$$

is convex, and is considered tractable. We show that for most Φ the solution $\hat{\alpha}_{1,\epsilon} = \hat{\alpha}_{1,\epsilon}(y, \Phi)$ of this problem is quite generally a good approximation for $\hat{\alpha}_{0,\epsilon}$.

We suppose that the columns of Φ are normalized to unit ℓ^2 norm 1 and we place uniform measure on such Φ. We study the *underdetermined* case where $m \sim An$, $A > 1$ and prove the existence of $\rho = \rho(A)$ and $C > 0$ so that for large n, and for all Φ's except a negligible fraction, the following *approximate sparse solution* property of Φ holds: *For every y having an approximation $\|y - \Phi\alpha_0\|_2 \leq \epsilon$ by a coefficient vector $\alpha_0 \in \mathbf{R}^m$ with fewer than $\rho \cdot n$ nonzeros, we have*

$$\|\hat{\alpha}_{1,\epsilon} - \alpha_0\|_2 \leq C \cdot \epsilon.$$

This has two implications. First: for most Φ, whenever the combinatorial optimization result $\alpha_{0,\epsilon}$ would be very sparse, $\hat{\alpha}_{1,\epsilon}$ is a good approximation to $\alpha_{0,\epsilon}$. Second: suppose we are given noisy data obeying $y = \Phi\alpha_0 + z$ where the unknown α_0 is known to be sparse and the noise $\|z\|_2 \leq \epsilon$. For most Φ, noise-tolerant ℓ^1-minimization will stably recover α_0 from y in the presence of noise z.

We study also the barely-determined case $m = n$ and reach parallel conclusions by slightly different arguments.

The techniques include the use of almost-spherical sections in Banach space theory and concentration of measure for eigenvalues of random matrices.

Key Words and Phrases. Solution of Underdetermined Linear Systems. Approximate Sparse Solution of Linear equations. Almost-Spherical Sections of Banach Spaces. Eigenvalues of Random Matrices.

Acknowledgements. Partial support from NSF DMS 00-77261, and 01-40698 (FRG) and ONR. Thanks to Noureddine El Karoui and Stanislaw Szarek for helpful correspondence, preprints, and reprints.

1 Introduction

Underdetermined systems of linear equations appear naturally in many important problems in science and technology, ranging from array signal processing to image processing to genomic data analysis. Such systems, with fewer equations than unknowns, may have many solutions, but often the solution of interest is the *sparsest* solution – the one having the fewest possible nonzeros. In a companion paper [8], it was shown that for "most" underdetermined systems, the sparsest solution – if it is sufficiently sparse – can be recovered uniquely by solving a *convex* optimization problem, namely, by finding the solution with smallest ℓ^1 norm. Here by *sufficient sparsity*, we mean that the number of nonzeros in the solution was only a certain fraction of the number of equations.

In "most" applications in science and technology, of course, the underlying model will not be perfectly correct and measurements will not be perfectly accurate. It is essential to use procedures which are robust against the effects of measurement noise and modelling error. In this paper we consider a noise-tolerant approach: searching among the many near-solutions which satisfy the system of equations to within a specified accuracy, and selecting the near-solution with the smallest ℓ^1 norm. We show that "most" matrices underlying underdetermined systems have the following property. *when there exists any sufficiently sparse near-solution, the near-solution with minimal ℓ^1 norm is a good approximation to it.*

1.1 Background

Now for some context. In recent years, there has been rapid development in the theory of sparse overcomplete signal representations [5, 13, 14, 29, 30, 16]. In this literature, one attempts to represent a signal $S \in \mathbf{R}^n$ sparsely, using an overcomplete set, for example, the union of several orthonormal bases or frames. Formally, one has an n by m matrix Φ with columns ϕ_i, $i = 1, \ldots, m$. Following [20, 2], Φ is also called the dictionary, and the ϕ_i are also called atoms; in the cases they envisioned, Φ for example could be the concatenation of several bases (sinusoids, wavelets, spikes, etc.). Now the problem of solving for α in the system $S = \Phi\alpha$ is in general ill-posed, since $m > n$. This literature identified a class of matrices Φ where this ill-posedness could be resolved by sparsity. For such matrices, the coherence $M(\Phi) = \max_{i \neq j} |\langle \phi_i, \phi_j \rangle|$ is small. Thus, for example, the concatenation of the sinusoid basis and the natural basis for $\mathbf{R}^n$ has coherence $1/\sqrt{n}$. It was found that, for noiseless observations, if the matrix Φ has appropriately small coherence then, whenever there truly is a sparse solution $S = \Phi\alpha_0$, where α_0 has at most $N \leq (1 + M^{-1})/2$ nonzeros, then this solution will be found, either by ℓ^1 norm minimization [6] or by stepwise greedy approximation [29] run for N steps. Since in some interesting cases $M(\Phi) = 1/\sqrt{n}$, so for certain matrices Φ these results permit unique recovery of the sparsest representation, whenever that sparsest representation has fewer than $\sqrt{n}/2$ nonzeros.

In a recent paper, Donoho, Elad, and Temlyakov [7], considered the stability of such representations in the presence of noise; related results were also obtained by Tropp [30]. They suppose we have a vector y of interest and consider the convex optimization problem

$$(P_{1,\epsilon}) \qquad \min \|\alpha\|_1 \text{ subject to } \|y - \Phi\alpha\|_2 \leq \epsilon$$

They suppose the vector y of interest happens to have an approximate sparse representation $\|y - \Phi\alpha_0\|_2 \leq \epsilon$ where $\alpha_0 \in \mathbf{R}^m$ is unknown, with at most $N \leq (1 + M^{-1})/4$ nonzeros, and show that the solution $\hat{\alpha}_{1,\epsilon}$ of this problem obeys

$$\|\hat{\alpha}_{1,\epsilon} - \alpha_0\|_2 \leq 3 \cdot \epsilon.$$

In short, although the problem is underdetermined, if there is a nearby sparse solution, the minimal ℓ^1 near-solution will approximate it.

1.2 This Paper

In the result just cited, approximation was proven under the condition that the number of nonzeros was at most $(1 + M^{-1})/4$. Since for an n by m matrix with $m > n$, the incoherence is bounded by $M \geq 1/\sqrt{n}$ [5], this result accomodates at most $O(\sqrt{n})$ nonzeros.

In the present paper, we will establish approximation bounds under much weaker sparsity conditions, allowing $\sim \rho n$ nonzeros. Moreover, our results apply to "most" underdetermined systems.

In our first result, we let Φ be n by n, so the problem is not necessarily underdetermined, and suppose the columns are normalized by $\|\phi_i\|_2 = 1$, $i = 1, \ldots, n$. We place uniform measure on the space $\mathbf{S}^{n-1} \times \cdots \times \mathbf{S}^{n-1}$ of such matrices. The smallest singular value of Φ is then typically of size $O(1/\sqrt{n})$ [10, 27], so the problem, though not underdetermined, is poorly conditioned. Letting $\hat{\alpha} = \Phi^{-1} y$ be the usual solution of the linear equations, the best approximation bound it obeys is of the form

$$\|\hat{\alpha} - \alpha_0\|_2 \leq c\sqrt{n} \cdot \epsilon.$$

Hence the problem gets increasingly poorly posed for large n. Standard results on approximate sparse representation by a greedy process of incremental model building [22] also fall apart in this case, because of the poor conditioning.

In contrast, consider the problem $(P_{1,\epsilon})$ discussed above (and in [7]). We show there are constants $\rho > 0$ and $C > 0$ so that, when n is large, all but a vanishing fraction of such matrices Φ have the following sparse approximation property: *whenever the given data y permit a sparse approximate solution $\|y - \Phi\alpha_0\|_2 \leq \epsilon$ having $\|\alpha_0\|_0 \leq \rho n$, then the solution $\hat{\alpha}_{1,\epsilon}$ to $(P_{1,\epsilon})$ obeys*

$$\|\hat{\alpha}_{1,\epsilon} - \alpha_0\|_2 \leq C \cdot \epsilon. \tag{1.1}$$

In short, although the traditional solution of the system of linear equations would at best obey an approximation bound scaling poorly as $n \to \infty$, the ℓ^1 penalization gives an approximation bound with behavior $C\epsilon$, in a natural sense best possible in both n and ϵ.

In our second result, we let Φ be n by m, with $n < m < An$ where $A > 1$. Now the problem is certainly underdetermined. With again ℓ^2-normalized columns and uniform measure on the space of such n by m matrices, we again show the existence of $\rho = \rho(A)$ so that, an overwhelming fraction of n by m matrices have the sparse approximation property: whenever there is a sparse approximate solution α_0 obeying $\|\alpha_0\|_0 \leq \rho n$ and $\|y - \Phi\alpha_0\|_2 \leq \epsilon$, the ℓ^1 solution approximates α_0 with the approximation bound (1.1).

An interesting aspect of our proofs is the role played by key results in the geometry of Banach spaces; namely the spherical sections theorem (Dvoretsky, Milman, ...) and more particularly, the refinement for octahedra, due to Kashin. We also rely on concentration of measure estimates for singular values of random matrices, quoting heavily from work of Szarek.

Section 2 of this paper develops our results in the $m = n$ case; Section 3 discusses the $m \leq An$ case.

We indulge in a small sin of usage. We allow ourselves to say things like "$n/2$-dimensional" even if n is odd. Whenever an expression such as ρn refers to a dimension or other naturally integral quantity, we implicitly assume that, if the expression is not integral, it is rounded down.

1.3 Potential Applications

There are two ways to view our results.

On the one hand, they say that sparse modeling in underdetermined linear systems – a vast enterprise throughout science and technology – has a respectable intellectual justification.

In such endeavors we have a model $y \approx \Phi\alpha$ and we believe that there is a sparse near-solution, i.e. there is some vector α_0 with relatively few nonzeros satisfying $\|y - \Phi\alpha_0\|_2 \leq \epsilon$. The inaccuracy in our model might be due to measurement error or to modeling error. Call α_0 the ideal sparse representation and $S_0 = \Phi\alpha_0$ the ideal noiseless data.

If Φ is like "most" n-by-m matrices, then the solution to $(P_{1,\epsilon})$ is a good approximation to the ideal noiseless representation, i.e. the representation we could recover if there were *no* modeling error, *no* measurement error, and the equations were *not* underdetermined. Hence, heuristic models based on sparse representations are not silly starting points; small violations of sparsity and small measurement errors can be tolerated.

On the other hand, our results have an algorithmic interpretation. They say that although sparse solution of underdetermined linear systems is in general computationally intractable, a valuable and effective substitute is available. In many cases, we simply solve $(P_{1,\epsilon})$, and check if the result is a sufficiently good approximation to a sparse vector. If it is not, we can be sure there is no highly sparse near-solution to the equations. And conversely, in those same cases, if there is a highly-sparse near-solution to the equations, it must be near the solution to $(P_{1,\epsilon})$.

Of course, in specific applications, what matters is not "most" matrices but the specific matrix in actual use. Nevertheless, researchers using methods related to $(P_{1,\epsilon})$ in the setting of large underdetermined systems are currently reporting good results [2, 28, 11]. Our results provide theoretical support for their empirical success.

2 The Case $m = n$

Let $\phi_1, \phi_2, \ldots, \phi_m$ be random points uniformly distributed on the unit sphere $\mathbf{S}^{n-1}$ in $\mathbf{R}^n$. Let $\Phi = [\phi_1 \ldots \phi_m]$ be the matrix obtained by concatenating the corresponding column vectors. The space of n by m matrices having columns with unit norm is, of course,

$$\mathbf{\Phi}_{n,m} = \overbrace{\mathbf{S}^{n-1} \times \cdots \times \mathbf{S}^{n-1}}^{\leftarrow m \text{ terms} \rightarrow} .$$

Now the probability measure we are assuming on the random matrix Φ is just the natural uniform measure on $\mathbf{\Phi}_{n,m}$. Hence, probabilistic statements about properties of Φ are interpretable as statements about the fraction of matrices $\Phi \in \mathbf{\Phi}_{n,m}$ with a certain property. When we say that a property of Φ holds *with overwhelming probability for large* n, for example, we mean that, for each $\delta > 0$, for an understood sequence (n, m_n) with $n \to \infty$, the fraction of such matrices in $\mathbf{\Phi}_{n,m_n}$ eventually exceeds $1 - \delta$ as $n \to \infty$.

For a vector $S \in \mathbf{R}^n$ we are interested in the sparsest possible representation; this is given by:

$$(P_0) \qquad \min \|\alpha\|_0 \text{ subject to } \Phi\alpha = S,$$

It was pointed out in [8] that if (P_0) has *any* sparse solution, then it will have a unique sparsest one.

Lemma 2.1 *With probability* 1, Φ *has the following* **unique sparsest representation** *property:*

For every vector α_0 having $\|\alpha_0\|_0 < n/2$ the instance of (P_0) generated by the data $S = \Phi\alpha_0$ is uniquely solved by α_0.

In short, it makes sense to speak of *the* sparsest solution to $S = \Phi\alpha$.

Consider the *sparse approximation problem*

$$(P_{0,\epsilon}) \qquad \min \|\alpha\|_0 \text{ subject to } \|y - \Phi\alpha\|_2 \le \epsilon.$$

and let $\hat{\alpha}_{0,\epsilon}$ denote any minimizer. This asks for the sparsest near-solution. In the setting considered here, [8, Section 9] was able to combine results from [7] and inequalities from [8] to get a result which implies the following.

Corollary 2.1 *Consider the setting where $m = n$. There exists $\rho > 0$ so that with overwhelming probability for large n, the matrix Φ has the following* **sparse approximation property***:*

Whenever a vector y obeys $\|y - \Phi\alpha_0\|_2 \le \epsilon$ for some α_0 obeying $\|\alpha_0\|_0 < \rho n$, then any solution to the instance of $(P_{0,\epsilon})$ generated by y obeys

$$\|\hat{\alpha}_{0,\epsilon} - \alpha_0\|_2 \le 4\epsilon.$$

This shows, for example, that the set of all sparse approximants with prescribed sparsity is confined to a small neighborhood. Problem $(P_{0,\epsilon})$ is not computationally feasible in practice; in general it requires combinatorial optimization, enumerating subsets of the m variables and checking to see which, if any, permit an ϵ-approximation. We view this result merely as indicating that the *question* of finding approximate sparse solutions from noisy data is well-posed.

Consider instead the *convex* optimization problem

$$(P_{1,\epsilon}) \qquad \min \|\alpha\|_1 \text{ subject to } \|y - \Phi\alpha\|_2 \le \epsilon.$$

and let $\hat{\alpha}_{1,\epsilon}$ denote any solution. Convexity makes $P_{1,\epsilon})$ a far more computationally appealing problem than $(P_{0,\epsilon})$. In this section we prove the following.

Theorem 2.1 *Consider the setting where $m = n$. There exist $\rho > 0$ and $C > 0$ so that with overwhelming probability for large n, Φ has the following sparse approximation property:*

Whenever a vector y has an approximate representation $\|y - \Phi\alpha_0\|_2 \le \epsilon$, with an α_0 obeying $\|\alpha_0\|_0 < \rho n$, then any solution to $(P_{1,\epsilon})$ obeys

$$\|\hat{\alpha}_{1,\epsilon}(y) - \alpha_0\|_2 \le C\epsilon.$$

Again, the notion of probability here refers to uniform measure on the space of $n \times n$-matrices with unit-norm columns. Hence, the above result shows that the minimal-ℓ^1 near-solution *generically* approximates the sparsest near-solution, whenever that solution is sufficiently sparse. Here by *generic* we mean "experienced on a set of matrices of nearly full measure".

2.1 Proof Outline

We now describe the overall architecture of the proof, which requires several lemmas proved in later subsections. As in [7], we first note that

$$\|\hat{\alpha}_{1,\epsilon}\|_1 \le \|\alpha_0\|_1,$$

since α_0 is merely feasible for $(P_{1,\epsilon})$, while $\hat{\alpha}_{1,\epsilon}$ is optimal. At the same time

$$\|\Phi\hat{\alpha}_{1,\epsilon} - \Phi\alpha_0\|_2 \leq 2\epsilon,$$

by the triangle inequality.

We now view $\hat{\alpha}_{1,\epsilon}$ as a perturbed version $\alpha_0+\beta$ of α_0, where the perturbation obeys additional properties. Letting B_{ϵ,α_0} denote the collection of perturbations β to α_0 obeying

$$\|\Phi\beta\|_2 \leq 2\epsilon, \qquad \|\alpha_0 + \beta\|_1 \leq \|\alpha_0\|_1,$$

then, indeed,

$$\hat{\alpha}_{1,\epsilon} - \alpha_0 \in B_{\epsilon,\alpha_0}.$$

Defining then

$$(Q_{\epsilon,\alpha_0}) \qquad \sup \|\beta\|_2 \text{ subject to } \beta \in B_{\epsilon,\alpha_0}$$

we must have

$$\|\hat{\alpha}_{1,\epsilon} - \alpha_0\|_2 \leq val(Q_{\epsilon,\alpha_0}).$$

Now let $I = supp(\alpha_0)$ and suppose without loss of generality that $I = \{1, \ldots, |I|\}$. Partitioning $\beta = (\beta_I, \beta_{I^c})$, note that

$$\|\alpha_0 + \beta\|_1 - \|\alpha_0\|_1 \leq \|\beta_I\|_1 - \|\beta_{I^c}\|_1.$$

Consider then the new optimization problem

$$(R_{\epsilon,I}) \qquad \sup \|\beta\|_2 \text{ subject to } \|\Phi\beta\|_2 \leq 2\epsilon, \quad \|\beta_I\|_1 \geq \|\beta_{I^c}\|_1.$$

We have $val(R_{\epsilon,I}) \geq val(Q_{\epsilon,\alpha_0})$.

We now define an event $\Omega_n(\rho)$ - this may be viewed as a set of matrices $\Phi \in \mathbf{\Phi}_{n,n}$ - and show that on this event we have

$$val(R_{\epsilon,I}) \leq C_\rho\epsilon, \qquad \forall |I| < \rho n. \tag{2.1}$$

The event $\Omega_n(\rho)$ is the intersection of 5 subevents Ω_n^i, $i = 1, \ldots, 5$, implicitly parametrized by certain constants $\rho_i > 0$, $\eta_i > 0$.

The first three events refer to the eigenvalues of Gram matrices $\Phi^T\Phi$ or their submatrices $\Phi_I^T\Phi_I$, where Φ_I denotes the $n \times |I|$ matrix with columns taken from $i \in I$. We let λ_{min} and λ_{max} denote the smallest and largest eigenvalues, with λ_k for intermediate ones, $\lambda_{min} = \lambda_1 \leq \cdots \leq \lambda_k \leq \cdots \leq \lambda_{max}$. The singular values of Φ, Φ_I etc, are just the square roots of the corresponding eigenvalues. The events are:

Ω_n^1 $\lambda_{min}(\Phi_I^T\Phi_I) \geq \eta_1^2 > 0$, uniformly in I with $|I| < \rho_1 n$.

Ω_n^2 $\lambda_{max}(\Phi^T\Phi) \leq \eta_2^2$.

Ω_n^3 With $\ell = \lfloor \rho_3 n \rfloor$, $\lambda_\ell(\Phi_{I^c}^T\Phi_{I^c}) \geq \eta_3^2 > 0$, uniformly in I with $|I| < \rho_3 n$.

In the next subsection, it is shown that, for appropriate ρ_i, $\eta_i > 0$, these events all have $P((\Omega_n^i)^c) \leq \exp(-n\beta_i)$, where $\beta_i > 0$, $i = 1, 2, 3$.

The remaining subevents require additional definitions. Our argument will turn around the vectors

$$v = -\Phi_I\beta_I, \qquad w = \Phi_{I^c}\beta_{I^c};$$

they obey $\|v - w\|_2 = \|\Phi\beta\|_2 \leq \epsilon$. To exploit this closeness, we must cope with the fact that Φ has numerous small singular values, and deal with an associated subspace separately. We

associate to each I a random subspace $W_{I,0}$ spanned by the singular vectors associated to the $\lfloor \rho_4 n \rfloor$ lowest singular values of Φ_{I^c}. There is also the random subspace $V_I = Range(\Phi_I)$. We also let $W_{I,1}$ denote the orthocomplement of $W_{I,0}$ in $Range(\Phi_{I^c})$, and $V_{I,1} = V_I \cap W_{I,1}$. Finally, $W_{I,2}$ is the remaining orthocomplement in $\mathbf{R}^n$. The vectors v and w can be resolved into components $w = w_0 + w_1$, $v = v_0 + v_1 + v_2$.

Corresponding to these, we have subspaces $B_{I,0}$ and $B_{I,1}$ in R^m, consisting of vectors β_0, β_1 solving

$$\Phi_{I^c}\beta_0 = w_0, \qquad \Phi_{I^c}\beta_1 = w_1.$$

(Recall that the n by n matrix Φ is nonsingular with probability one). Naturally, $\beta_{I^c} = \beta_0 + \beta_1$. Similarly, define $\gamma \in B_{I,0} + B_{I,1}$ by $\gamma = \gamma_0 + \gamma_1$, where

$$\Phi_{I^c}\gamma_0 = v_0, \qquad \Phi_{I^c}\gamma_1 = v_1.$$

Let $C_{I,1}$ denote the subspace of all such γ_1's in $B_{I,1}$.

The remaining subevents of Ω_n are now definable:

Ω_n^4 The subspaces $W_{I,0}$ and V_I have positive angle, so that

$$\angle(W_{I,0}, V_I) \geq \eta_4 > 0,$$

uniformly in $|I| < \rho_4 n$; for the definition of angle between subspaces, see (2.8) below.

Ω_n^5 On every subspace $B_{I,0} + C_{I,1}$ the ℓ_n^1 norm is *almost-Euclidean*:

$$\frac{1}{2}\|\beta_0 + \gamma_1\|_2 \leq \sqrt{\frac{\pi}{2n}} \cdot \|\beta_0 + \gamma_1\|_1 \leq \frac{3}{2}\|\beta_0 + \gamma_1\|_2$$

uniformly over $\beta_0 \in B_{I,0}$, $\gamma_1 \in C_{I,1}$ and $|I| < \rho_5 n$.

It will be shown in later subsections that for ρ_4 and ρ_5 chosen appropriately, $P((\Omega_n^i)^c) \leq \exp(-n\beta_i)$, where $\beta_i > 0$, $i = 4, 5$. Combining all this, let $\rho_6 = \min_{i=1}^5 \rho_i$ and $\beta = \min_i \beta_i$. Set

$$E_n = \cap_{i=1}^5 \Omega_n^i(\rho_6),$$

and notice that $P((E_n)^c) \leq 5 \cdot \exp(-n\beta)$. Since E_n is overwhelmingly likely for all large n, assume for the rest of the proof that the event E_n holds.

Our goal is, once again, to estimate the value of the optimization problem $(R_{\epsilon,I})$, so we will be interested in bounding $\|\beta_I\|_2^2 + \|\beta_{I^c}\|_2^2$ using the control available from $\|v - w\|_2 \leq \epsilon$ and $\|\beta_{I^c}\|_1 \leq \|\beta_I\|_1$. In the argument below c_k, $k = 1, \ldots, 8$ denote positive constants whose precise values are not relevant for the proof itself, but may be of interest later on.

We plan to invoke the constraint $\|\beta_I\|_1 \geq \|\beta_{I^c}\|_1$. Using Ω_n^1,

$$\|v\|_2 \geq \eta_1 \|\beta_I\|_2,$$

and so

$$\sqrt{|I|}\|v\|_2/\eta_1 \geq \|\beta\|_1. \tag{2.2}$$

Now

$$\begin{aligned} \|\beta_{I^c}\|_1 &\geq \|\beta_0 + \beta_1\|_1 \\ &\geq \|\beta_0 + \gamma_1\|_1 - \|\beta_1 - \gamma_1\|_1. \end{aligned}$$

By $\Omega_n^5 \cap \Omega_n^4$ we have

$$\begin{aligned} \|\beta_0 + \gamma_1\|_1 &\geq c_1\sqrt{n} \cdot \|\beta_0 + \gamma_1\|_2 \\ &\geq c_2\sqrt{n} \cdot (\|\beta_0\|_2^2 + \|\gamma_1\|_2^2)^{1/2}. \end{aligned}$$

Applying Ω_n^3,

$$\begin{aligned} \|\beta_1 - \gamma_1\|_2 &\leq \|w_1 - v_1\|_2/\eta_3 \\ &\leq \|w - v\|_2/\eta_3 \leq \epsilon/\eta_3. \end{aligned}$$

Hence,

$$\|\beta_{I^c}\|_1 \geq c_2\sqrt{n} \cdot (\|\beta_0\|_2^2 + \|\gamma_1\|_2^2)^{1/2} - \sqrt{n}\epsilon/\eta_3.$$

Now from $\|\beta_I\|_1 \geq \|\beta_{I^c}\|_1$ and (2.2) we have

$$\sqrt{|I|}\|v\|_2/\eta_1 \geq \sqrt{n} \cdot (c_2\|\gamma_1\|_2 - \epsilon/\eta_3),$$

yielding

$$\epsilon/\eta_3 \geq c_2\|\gamma_1\|_2 - \sqrt{\frac{|I|}{n}}\|v\|_2/\eta_1.$$

From Ω_n^4, $\|v\|_2 \leq c_3\|v_1\|_2$, and by Ω_n^2, $\|v_1\|_2 \leq \eta_2\|\gamma_1\|_2$. Combining these, we get

$$\epsilon/\eta_3 \geq \|v\|_2(c_2/(c_3\eta_2) - \sqrt{\frac{|I|}{n}}/\eta_1).$$

Picking $\rho_7 > 0$ small enough, for $\frac{|I|}{n} < \rho_7$, we have, for some $c_4 > 0$

$$c_4\epsilon > \|v\|_2.$$

We now use this bound on the size of v to control the size of both β_I and β_{I^c}. We immediately get, due to Ω_n^1

$$\|\beta_I\|_2 \leq \|v\|_2/\eta_1 = c_5\epsilon.$$

We also easily get

$$\|\beta_1\|_2 \leq \eta_3^{-1}\|w_1\|_2 \leq \eta_3^{-1}(\|v_1 - w_1\|_2 + \|v_1\|_2) \leq \eta_3^{-1}(\epsilon + c_4\epsilon) \equiv c_6\epsilon.$$

Meanwhile,

$$\begin{aligned} \|\beta_0\|_2 &\leq c_7\|\beta_0\|_1/\sqrt{n} \qquad \text{by } (\Omega_n^5) \\ &\leq c_7/\sqrt{n} \cdot (\|\beta_0 + \beta_1\|_1 + \|\beta_1\|_1) \\ &= c_7/\sqrt{n} \cdot (\|\beta_I\|_1 + \|\beta_1\|_1) \\ &\leq c_7/\sqrt{n} \cdot (\sqrt{|I|}\|\beta_I\|_2 + \sqrt{n}\|\beta_1\|_2) \\ &\leq c_7(\sqrt{\rho_7}c_5 + c_6)\epsilon. \end{aligned}$$

We conclude that

$$\|\beta\|_2 \leq \|\beta_I\|_2 + \|\beta_0\|_2 + \|\beta_1\|_2 \leq c_8\epsilon,$$

with c_8 independent of n, and $|I|$ assumed $\leq \rho_7 n$. Hence defining $\rho = \min(\rho_6, \rho_7)$ and setting $\Omega_n(\rho) \equiv E_n$ we get (2.1). QED.

Remark 1. It may be of interest to estimate the size of the coefficients c_i used in the proof. Note that, as $\rho \to 0$, $\eta_1 = 1 + o(1)$, $\eta_2 = 1 + o(1)$, $\eta_3 = O(\rho_3)$, $\eta_4 = \pi/2 - o(1)$. Hence, we can arrange so that, as $\rho \to 0$,

$$\begin{aligned} c_1 &= \frac{1}{\sqrt{2\pi}}(1 + o(1)) \\ c_2 &= c_1 + o(1) \\ c_3 &= 1 + o(1) \\ c_4 &= (1 - \rho)/\eta_3 \cdot (1 + o(1)) \\ c_5 &= 1 + o(1) \\ c_6 &= 1 + c_4 \\ c_7 &= \sqrt{\pi/2} + o(1) \\ c_8 &= (c_5^2 + c_6^2 + c_7^2)^{1/2} \end{aligned}$$

Here the final "output" we are interested in, $C_\rho \le c_8$. Notice that c_8 is well-bounded, except for the presence of the η_3^{-1} factor in c_4. The estimate from the proof of Lemma 2.5 gives $\eta_3 > \rho_3/\sqrt{2e}$. Hence, c_4 'blows up' as $\rho_3 \to 0$. The best control results under an alternate asymptotic in which $\max_{i \ne 3} \rho_i \to 0$ while $\rho_3 = const$.

Remark 2. In the case $\epsilon = 0$, this gives a different approach to the main result in [8] (in the case $m = n$), overlapping in the use of eigenvalue bounds and spherical sections, but using a subspace angle principle in place of the sign-embeddings in [8].

2.2 Control of Eigenvalues

We now show that one can set parameters yielding the claimed properties for Ω_n^i $i = 1, 2, 3$.

The first Lemma, from [8], is more general than we need in this section. For later use, it allows a range of $m \ge n$, namely $n \le m \le An$, where $A > 1$; for now we need only $n = m$, which is the case $A = 1$.

Lemma 2.2 *[8] Define the event*

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

For each $\rho \in (0, 1/2]$ and $A \ge 1$, there is $\lambda = \lambda(\rho, A) > 0$ so that along sequences of (n, m) with $m \le An$

$$P(\Omega_{n,m,\rho,\lambda}) \to 1, n \to \infty.$$

The second lemma supports our claims for Ω_n^2.

Lemma 2.3 *Let ϕ_i be iid uniform on S^{n-1}. For some $\beta > 0$ and $n > n_0$,*

$$P\{\lambda_{max}(\Phi^T \Phi) > 3\} \le \exp(-n\beta)$$

Proof. We use existing bounds for Gaussian iid $N(0, \frac{1}{n} I_n)$ vectors and the standard relationship [24, Chapter 4] between Gaussians and uniform spherical vectors.

Szarek [27] proved that for the $n \times n$ matrix X defined with iid Gaussian entries $X_{ij} \sim N(0, \frac{1}{n})$, we have

$$P\{\lambda_{max}(X^T X) > 7/3\} \le \exp(-n\beta_G), \qquad n \ge n_0. \tag{2.3}$$

As in the companion paper [8], this immediately implies results for the uniform spherical case; we spell this out as it will be helpful again below. Let R_i be iid random variables distributed $\chi_n/\sqrt{n}$, where χ_n denotes the χ_n distribution. These can be generated by taking iid standard normal RV's Z_{ij} which are independent of (ϕ_i) and setting

$$R_i = (n^{-1}\sum_{j=1}^{n} Z_{ij}^2)^{1/2}. \tag{2.4}$$

Let $x_i = R_i \cdot \phi_i$; then the x_i are iid $N(0, \frac{1}{n}I_n)$, and we view them as the columns of X. With $R = diag((R_i)_i)$,

$$\lambda_{max}(X^T X) = \lambda_{max}(R^T \Phi^T \Phi R) \geq \|R^{-1}\|^2 \lambda_{max}(\Phi^T \Phi), \tag{2.5}$$

where $\|R^{-1}\| = \max_i R_i^{-1}$. Define η by $(1-\eta)^2 = 9/21$, and note that on the event $\{\max_i |R_i - 1| < \eta\}$, $\lambda_{max}(\Phi^T\Phi) \leq 3$.

Now (2.4) exhibits each R_i as a function of n iid standard normal random variables, Lipschitz with respect to the standard Euclidean metric, with Lipschitz constant $1/\sqrt{n}$. Therefore, by concentration of measure [19],

$$P\{\max_i |R_i - 1| > \eta\} \leq 2n \exp\{-n\eta^2/2\} = 2n \exp\{-n\beta_\chi\}, \tag{2.6}$$

for $\beta_\chi > 0$. Hence $\lambda_{max}(\Phi^T\Phi) \leq 3$, except on an event of probability bounded by $\exp\{-n\beta_G\} + 2n\exp\{-n\beta_\chi\}$. QED

We need the Cauchy Interlace Theorem [23, 186-187].

Lemma 2.4 *Let G be an $n \times n$ real symmetric matrix and let G_I be the $n-k$ by $n-k$ principal submatrix obtained by deleting k columns and k rows corresponding to indices $i \in I$, $|I| = k$. Then for $1 \leq \ell \leq n-k$,*

$$\lambda_\ell(G) \leq \lambda_\ell(G_I) \leq \lambda_{\ell+k}(G).$$

This is well-known in the case $k = 1$ as a consequence of the Courant-Fischer min-max characterization of eigenvalues [17], and can be proved inductively starting from the case $k = 1$.

Lemma 2.5 *Fix $\rho < 1/2$. There is $\beta_3 > 0$ so that, on an event $\Omega_n^3(\rho)$ with probability exceeding $1 - \exp(-n\beta_3)$ for $n > n_3$,*

$$\lambda_{\lfloor \rho n \rfloor}(\Phi_{I^c}^T \Phi_{I^c}) \geq \rho^2/2e \qquad \forall |I| \leq \rho n.$$

Proof. Now consider matrices Φ_{I^c} formed by deleting columns from $1, \ldots, n$ which belong to I. Applying Lemma 2.4 to $G = \Phi^T\Phi$, we have for $k \geq 1$,

$$\lambda_k(\Phi_{I^c}^T \Phi_{I^c}) \geq \lambda_k(\Phi^T\Phi).$$

We now again use the connection between uniforms and Gaussians. Letting R_i as in (2.4) then $X = \Phi R$ has columns which are iid $N(0, \frac{1}{n}I_n)$. Analogously to (2.5),

$$\lambda_k(X^T X) \leq \|R\|^2 \lambda_k(\Phi^T\Phi).$$

Picking $\eta = \sqrt{2} - 1$ we have from (2.6) that the event $E_n^c \equiv \{\|R\|^2 > 2\}$ has probability bounded above by $\exp\{-n\beta)$ where $\beta = 3/4 - 1/\sqrt{2} > 0$.

Szarek [27, Theorem 1.2] shows that if X is an n by n matrix with entries iid $N(0, \frac{1}{n} I_n)$, then

$$P\{\lambda_k^{1/2}(X^T X) \le \alpha \frac{k}{n}\} \le (\sqrt{2e}\alpha)^{k^2}. \tag{2.7}$$

Picking $\alpha = 1/2e$ in (2.7) gives for the event $F_n = \{\lambda_k(X^T X) \le \frac{1}{2e}(\frac{k}{m})^2\}$

$$\log P(F_n^c) \le -k^2 \log(2e)/2.$$

Picking $k = \rho n$, we get that on an event $E_n \cap F_n$ having overwhelming probability for large n,

$$\lambda_{\rho n}(\Phi^T \Phi) \ge \rho^2/(2e).$$

2.3 Angle Between Subspaces

As above $W_{I,0}$ is the span of the $\rho_3 n$ first singular vectors of Φ. By the iid character of the ϕ_i's, this is a random uniform subspace of dimension $\rho_3 n$ inside $Range(\Phi_{I^c})$. On the other hand, $V_I = Range(\Phi_I)$, is a random uniform subspace of dimension $|I|$ inside $\mathbf{R}^n$, and independent of $W_{I,0}$. Its orthogonal projection on $Range(\Phi_I^c)$ splits as $V_{I,0} + V_{I,1}$, $V_{I,0} \subset W_{I,0}$, $V_{I,1} \subset W_{I,1}$.

Given subspaces A and B, the angle between them is defined so

$$\cos(\angle(A, B)) = \sup\{\frac{\langle a, b\rangle}{\|a\|_2 \|b\|_2} : a \in A, b \in B\}. \tag{2.8}$$

Lemma 2.6 *For each $\eta > 0$, there is $\rho_4 > 0$ so that, with overwhelming probability,*

$$\cos(\angle(W_{I,0}, V_I)) \le \eta, \qquad \forall |I| < \rho_4 n.$$

To prove this, we first need a lemma about an individual pair of random subspaces.

Lemma 2.7 *For sufficiently small $\rho > 0$, let A and B be independent random uniform subspaces in $\mathbf{R}^n$ of dimension $\le \rho n$. For some $\beta > 0$ and all $n > n_0$, the angle between these subspaces obeys*

$$P\{\cos(\angle(A, B)) > 3\sqrt{\rho}\} \le \exp(-n\beta).$$

Proof. We give the argument merely for $k = \lfloor \rho n \rfloor$. Without loss of generality assume that coordinates have been chosen so that A is just the span of the first k standard unit basis vectors. Let $x_i, i = 1, \ldots, k$ be iid Gaussian vectors $N(0, \frac{1}{n} I_n)$; without loss of generality, we may take $B = span\{x_1, \ldots, x_k\}$. Form the random matrix X by concatenating the columns x_i. Let Y be the matrix obtained from X by Gram-Schmidt orthogonalization, and let Z be the upper k-by-k submatrix, Then, it is well known (to statisticians, at least) that $\cos(\angle(A, B))$ is just the top singular value of Z. To see this, note that if a and b are unit norm vectors in A and B, with u the vector of first k-entries in a and $b = Yv$, then

$$\cos(\angle(A, B)) \ge |\langle a, b\rangle| = |u' Z v|,$$

and the right side is maximized at the top singular value. Hence

$$\cos(\angle(A, B)) = \lambda_{max}(Z^T Z)^{1/2}.$$

Now $Y = XT$ where T is a triangular matrix implementing Gram-Schmidt orthogonalization. Hence, if $\tilde{Z}$ denotes the upper k rows of X, $Z = \tilde{Z}T$. The columns of $\tilde{Z}$ are distributed as $N(0, \frac{1}{n} I_k)$. Put $\overline{Z} = \sqrt{n/k} \cdot \tilde{Z}$. Then

$$\begin{aligned} \lambda_{max}(Z^T Z) &= \lambda_{max}(T^{-1} \tilde{Z}^T \tilde{Z} T^{-1}) \le \|T^{-1}\|^2 \lambda_{max}(\tilde{Z}^T \tilde{Z}) \\ &= \|T^{-1}\|^2 \cdot \frac{k}{n} \cdot \lambda_{max}(\overline{Z}^T \overline{Z}). \end{aligned}$$

Now $\overline{Z}$ is a 'standard' matrix with columns $N(0, \frac{1}{k} I_k)$. Applying (2.3) (replacing n by k), we get exponential bounds for $\{\lambda_{max}(\overline{Z}^T \overline{Z}) \ge 7/3\}$. Note that $\|T^{-1}\|^2 = 1/\lambda_{\min}(X^T X)$. Applying (the idea behind) Lemma 2.2, we can get (for small enough ρ) exponential bounds on $\{\|T^{-1}\|^2 > 9/7\}$. QED.

To adapt this individual result, for one pair of random subspaces, to obtain Lemma 2.6, which is simultaneous across many such pairs, we need the following Lemma, adapted from [8], where it is used several times for the same purpose.

Lemma 2.8 *Consider a family of events $\Omega_{n,I}$, indexed by subsets $I \subset \{1, \dots, n\}$ with $|I| \le \rho n$. Suppose these obey, for a common $\beta > 0$ and n_0,*

$$P(\Omega_{n,I}^c) \le \exp(-n\beta), \qquad \forall |I| \le \rho n, \qquad n \ge n_0.$$

Then for some $\rho' > 0$, $\beta' > 0$, and n_0',

$$P(\cup_{|I| < \rho' n} \Omega_{n,I}^c) \le \exp(-n\beta'), \quad n \ge n_0'.$$

Proof. Let $H(p)$ be Shannon entropy. Then

$$\log \binom{N}{\rho N} = N H(\rho)(1 + o(1)), \qquad N \to \infty;$$

in fact, if for $k < n/2$, $S_{n,k}$ denotes the cumulative sum

$$S_{n,k} = \binom{n}{1} + \cdots + \binom{n}{k}$$

then also

$$\log S_{n,\rho n} = n H(\rho)(1 + o(1)), \qquad n \to \infty.$$

Now

$$P(\cup_{|I| \le \rho n} \Omega_{n,I}^c) \le \sum_{|I| \le \rho n} P(\Omega_{n,I}) \le S_{n,n\rho} \exp(-n\beta).$$

Then

$$\log(P(\cup_{|I| < \rho n} \Omega_{n,I}^c)) \le n H(\rho)(1 + o(1)) - n\beta.$$

Now $H(\rho) \to 0$ as $\rho \to 0$, so, for sufficiently small ρ',

$$n H(\rho') < n\beta/2, \qquad n > n_0',$$

so

$$\log(P(\cup_{|I| \le \rho' n} \Omega_{n,I}^c) \le -n\frac{\beta}{2} = -n\beta', \qquad n > n_0'.$$

QED.

Proof of Lemma 2.6. For a given $\eta > 0$, define ρ_0 so that $\eta = 3\sqrt{\rho_0}$. Also define events

$$\Omega_{n,I} = \{\cos(\angle(W_{I,0}, V_{I,1})) \le \eta\}, \quad |I| \le \rho_0 n.$$

Applying the individual result Lemma 2.7 to each such event, we are immediately in a position to apply Lemma 2.8, turning "input" ρ_0 into "output" ρ'. Defining $\rho_4 = \min(\rho_0, \rho')$, we are done.

2.4 Almost-Euclidean Sections of ℓ_n^1

The subspaces V_I and $W_{I,0}$ are random subspaces of $\mathbf{R}^n$, and independent random variables. Moreover, they are each uniformly distributed random subspaces. They induce subspaces $B_{I,0}$ and C_I, via

$$W_{I,0} = \Phi_{I^c} B_{I,0}, \qquad V_{I,1} = \Phi_{I^c} C_I.$$

$B_{I,0}$ is a uniform random subspace of $\mathbf{R}^{n-k}$, and C_I is a uniform random subspace of the orthocomplement of $B_{I,0}$ inside $\mathbf{R}^{n-k}$.

Such subspaces will typically give *almost-Euclidean sections* of the ℓ_n^1 ball. As background, Dvoretsky's theorem [9, 24] says that every infinite-dimensional Banach space contains very high-dimensional subspaces on which the Banach norm is essentially the Euclidean norm. With more precision:

Definition 2.1 *We say that the k-dimensional subspace $A \subset \mathbf{R}^n$ offers an ϵ-Euclidean section of ℓ_n^1 if*

$$(1-\epsilon) \cdot \|a\|_2 \le \sqrt{\frac{\pi}{2n}} \cdot \|a\|_1 \le (1+\epsilon) \cdot \|a\|_2, \quad \forall a \in A. \tag{2.9}$$

Since we are taught in school that the ℓ_n^1 norm and the ℓ_n^2 norm are quite different, this seems counterintuitive; but in fact "most" subspaces give almost-Euclidean sections [15, 18]. We now push this to extremes:

Lemma 2.9 *There is $\rho_5 > 0$ so that, on an event Ω_n^5, every $B_{I,0} + C_{I,1}$ where $|I| < \rho_5 n$ gives an ϵ-Euclidean section of ℓ_n^1 with $\epsilon = 1/2$. The exception probability $P((\Omega_n^5)^c) \le \exp(-n\beta_5)$, where $\beta_5 > 0$.*

This depends on a standard result about generating random subspaces.

Lemma 2.10 *Let $\ell > 2k$, let B be a random uniform k-dimensional subspace of R^ℓ, and let C be a random uniform k-dimensional subspace of the orthocomplement of B in R^ℓ. Suppose that, conditionally on the orthocomplement of B, C is independent of B. Let $A = B + C$. Then A is a random uniform $2k$-dimensional subspace of R^ℓ.*

We omit the proof, which merely says that convolutions between uniform measures on different Grassmanians are again uniform. We also need a known result on obtaining almost-Euclidean sections by random subspaces.

Lemma 2.11 *Fix $\epsilon > 0$. There is $\rho(\epsilon) > 0$ so that $\rho < 1/4$ and the following holds for any $k < \rho n$. On an event $\Omega_{n,k,\epsilon}$, a random uniform $2k$-dimensional subspace A of $\mathbf{R}^{n-k}$ offers an ϵ-Euclidean section of ℓ_n^1. The exception $\Omega_{n,k,\epsilon}^c$ has probability at most $\exp(-n\beta(\epsilon))$, $n > n_0$.*

Proof of Lemma 2.11. The proof is obtained by a straightforward adaptation of known results [15, 24], this version, together with specifics on $\rho(\epsilon)$ and $\beta(\epsilon)$ has been worked out carefully in [8, Lemma 3.2]). We omit the details. QED

Proof of Lemma 2.9. This follows by the same approach as in the proof of Lemma 2.6, where an individual result for an individual pair of random subspaces was generalized to many pairs of random subspaces. The individual result, Lemma 2.11, gives us $\rho(1/2) > 0$ for "input" to Lemma 2.8, and we get ρ' as "output". Then we set $\rho_5 = \min(\rho(1/2), \rho')$. QED

3 The Case $n < m < An$

We now turn to the underdetermined case in which the number of equations is still proportional to the number of unknowns.

Theorem 3.1 *Let $A > 1$. Consider a sequence of problems (n, m_n) where $n < m < An$. There exist $\rho(A) > 0$ and $C > 0$ so that for all large n, the overwhelming majority of all $n \times m_n$ matrices Φ have the following property:*

> *For each vector y admitting an approximation $\|y - \Phi\alpha_0\|_2 \leq \epsilon$, by some vector α_0 obeying $\|\alpha_0\|_0 < \rho n$, the solution of $(P_{1,\epsilon})$ obeys*
>
> $$\|\hat{\alpha}_{1,\epsilon}(y) - \alpha_0\|_2 \leq C\epsilon.$$

3.1 Proof Outline

The proof of Theorem 3.1 has parallels to the $m = n$ case, with a few specific differences concerning the definition of the events Ω_n^i.

The events Ω_n^i, $i = 1, 2$ are exactly the same and the claims are the same. The support for our claims about Ω_n^1 in the $n < m < An$ case was already provided in Lemma 2.2, as noted earlier. The support for our claims about Ω_n^2 is given in the following subsections.

For the third event, we must take into account the fact that Φ has $m - n$ singular values which are exactly zero.

Ω_n^3 With $k = \lfloor \rho_3 n \rfloor$, and $\ell = m - n + k$, $\lambda_\ell(\Phi_{I^c}^T \Phi_{I^c}) \geq \eta_3^2 > 0$, uniformly in I with $|I| < \rho_3 n$.

The next subsection shows that for appropriate ρ_3, $\eta_3 > 0$ and $\beta_3 > 0$, $P((\Omega_n^3)^c) \leq \exp(-n\beta_3)$, $n > n_3$.

The remaining subevents again concern the vectors

$$v = -\Phi_I \beta_I, \qquad w = \Phi_{I^c} \beta_{I^c}.$$

We again associate to each I a random subspace $W_{I,0}$; this time, because of the null space of Φ, $W_{I,0}$ is associated to the $m - n + \lfloor \rho_3 n \rfloor$ lowest singular values of Φ_{I^c}. There is also the random subspace $V_I = Range(\Phi_I)$. We also let $W_{I,1}$ denote the orthocomplement of $W_{I,0}$ in $Range(\Phi_{I^c})$, and $V_{I,1} = V_I \cap W_{I,1}$. With probability one, $W_{I,2}$ the remaining orthocomplement in $\mathbf{R}^n$, is $\{0\}$. The vectors v and w can be resolved into components $w = w_0 + w_1$, $v = v_0 + v_1$.

Corresponding to this, we have subspaces $B_{I,0}$ and $B_{I,1}$ in R^m, consisting of vectors β_0, β_1 solving

$$\Phi_{I^c} \beta_0 = w_0, \qquad \Phi_{I^c} \beta_1 = w_1.$$

(Note that there is now a nullspace of Φ_{I^c}, so we additionally take $B_{I,1}$ in the orthocomplement of $B_{I,0}$.) Naturally, $\beta_{I^c} = \beta_0 + \beta_1$. Similarly, define $\gamma \in B_{I,0} + B_{I,1}$ by $\gamma = \gamma_0 + \gamma_1$, where

$$\Phi_{I^c} \gamma_0 = v_0, \qquad \Phi_{I^c} \gamma_1 = v_1.$$

Let $C_{I,1}$ denote the subspace of all such γ_1's in $B_{I,1}$. Note that, for a given w_0 and w_1, β_1 will be uniquely defined, but β_0 will not be; similarly for γ_0, γ_1.

Ω_n^4

$$\|v\|_2 \leq \eta_4 \cdot \|v_1\|_2, \qquad \forall v \in V_I,$$

uniformly in $|I| < \rho_4 n$.

Ω_n^5

$$\eta_5 \cdot \sqrt{A} \cdot \|\beta_0 + \gamma_1\|_2 \le \|\beta_0 + \gamma_1\|_1/\sqrt{n} \le \sqrt{A} \cdot \|\beta_0 + \gamma_1\|_2$$

uniformly over $\beta_0 \in B_{I,0}$, $\gamma_1 \in C_{I,1}$ and $|I| < \rho_5 n$.

Later subsections support our claims that for appropriate positive η_i and ρ_i, these events have probabilities exceeding $1 - \exp(-n\beta_i)$, $\beta_i > 0$, $i = 4, 5$. Once these claims are established, the proof outline can go through just as in the $m = n$ case, although the implicitly defined constants c_k, $k = 1, \ldots, 8$ will be different.

3.2 Control of Eigenvalues

We first justify our claims for Ω_n^2.

Lemma 3.1 *For $\eta_2 > 0$ and $\beta_2 > 0$, we have*

$$P\{\lambda_{max}(\Phi^T\Phi) \ge \eta_2\} \le \exp(-n\beta_2), \qquad n > n_0.$$

This is implied by the following Lemma about extreme singular values of nonsquare matrices, taken from Theorem 2.13 in Davidson-Szarek [4]; the Lemma will be used elsewhere below.

Lemma 3.2 *Let Z be a q by p matrix of iid $N(0, \frac{1}{q})$ Gaussians, $p < q$. Let $s_{max}(Z)$ denote the largest singular value of this matrix, and $s_{min}(Z)$ denote the smallest singular value. Then, with $\kappa = p/q$,*

$$P\{s_{max}(Z) > 1 + \sqrt{\kappa} + t\} \le \exp(-qt^2)$$
$$P\{s_{min}(Z) < 1 - \sqrt{\kappa} - t\} \le \exp(-qt^2)$$

Proof of Lemma 3.1. With (R_i) a collection of m independent $\chi_n/\sqrt{n}$ random variables, construct $X = \sqrt{\frac{n}{m}} diag(R)\Phi^T$. Then

$$\lambda_{max}(\Phi^T\Phi)^{1/2} \le \sqrt{A} \cdot \|R^{-1}\| \cdot s_{max}(X).$$

But X is standard Gaussian $N(0, \frac{1}{m})$. Applying Lemma 3.2 to $Z = X^T$, with $q = m$ and $p = n$, we get

$$P\{s_{max}(Z) > 1 + 1/\sqrt{A} + t\} \le \exp(-mt^2).$$

The result follows from this and (2.6). QED

We next supply the needed estimates for Ω_n^3.

Lemma 3.3 *For each $\rho_3 \in (0, 1)$ there are $\eta_3 > 0$ and $\beta_3 > 0$ so that*

$$P\{\lambda_{m-n+\rho_3 n}(\Phi_{I^c}^T\Phi_{I^c}) \ge \eta_3^2 \quad \forall |I| \le \rho_3 n\} \ge 1 - \exp(-n\beta_3).$$

This follows from a Lemma about intermediate singular values, similar to Lemma 3.2; this is derived in El Karoui [12]:

Lemma 3.4 *Let Z be a q by p matrix of iid $N(0, \frac{1}{q})$ Gaussians, $p < q$. Let $s_\ell(Z)$ denote the ℓ-th singular value where $s_1 = s_{min}$ etc. Let $\sigma_{\ell;p,q} = Median(s_\ell(Z))$ Then*

$$P\{s_\ell(Z) < \sigma_{\ell;p,q} - t\} \le \exp(-qt^2/2).$$

The proof idea is the same one behind Davidson and Szarek's proof for Lemma 3.2 - show that the singular values are Lipschitz functions on Euclidean space, and then use concentration of measure.

Proof of Lemma 3.3. We begin by observing that the Cauchy Interlace theorem reduces the problem to considering $\lambda_{m-n+\rho n}(\Phi^T\Phi)$. As in the proof of Lemma 3.1, set $X = \sqrt{\frac{n}{m}}\cdot diag(R)\cdot\Phi^T$, where R contains as usual iid $\chi_n/\sqrt{n}$ RV's. Then again

$$\lambda_{m-n+\ell}(\Phi^T\Phi) \geq \frac{m}{n}\cdot\|R^{-1}\|^2\cdot\lambda_{m-n+\ell}(XX^T),$$

and again $\|R^{-1}\|^2 \approx 1$ with good exponential bounds.

Note that X is standard normal with columns $N(0, \frac{1}{m}I_m)$. We now employ standard Random Matrix Theory terminology explained more fully in e.g. El Karoui [12]. The median singular value $\sigma_{\ell;m,n}(X)$ has a limit given by

$$\sigma^2_{\ell;n,m} \to F_A^{-1}(p), \quad m/n \to A > 1, \quad \ell/m \to p, \quad n \to \infty,$$

where F_A denotes the Marčenko-Pastur distribution function, having a jump of size $1-1/A$ at 0 and a density on $a_A = (1-1/\sqrt{A})^2 < x < b_A = (1+1/\sqrt{A})^2$ given by

$$f_A(x) = \frac{A}{2\pi}\sqrt{(b_A-x)(x-a_A)}.$$

Now

$$F_A(a_A+\delta) - F_A(a_A) \leq \frac{A}{3\pi}\delta^{3/2}, \qquad \delta > 0.$$

Putting $\delta_{A\alpha} = (3\pi\alpha/A)^{2/3}$, we get

$$\liminf_{n\to\infty}\sigma^2_{An-n+A\alpha n} \geq a_A + \delta_{A\alpha}.$$

Thus

$$P\{s^2_{An-n+A\alpha n} \leq a_A + \delta_{A\alpha}/2\} \leq \exp(-nt^2_{A\alpha}/2),$$

where

$$\begin{aligned} t_{A\alpha} &= \sqrt{a_A+\delta_{A\alpha}} - \sqrt{a_A+\delta_{A\alpha}/2} \\ &\geq \frac{\delta_{A\alpha}}{4(a_A+\delta_{A\alpha})^{1/2}} \geq \delta^{1/2}_{A\alpha}/4. \end{aligned}$$

Putting now

$$\xi_{A\alpha} = a_A + \delta_{A\alpha}/2, \qquad \beta_{A\alpha} = (3\pi\alpha/A)^{2/3}/32,$$

we get

$$P\{\lambda_{An-n+A\alpha n} \leq \xi_{A\alpha}\} \leq \exp(-n\beta_{A\alpha}).$$

Defining $\zeta_{A\alpha} = (a_A + \delta_{A\alpha}/2)/a_A$ we also have

$$P\{\|R^{-1}\|^2 > \zeta_{A\alpha}\} \leq m\cdot\exp(-n(\zeta_{A\alpha}-1)^2/2).$$

combining these gives the required estimates for Ω_n^3, with $\rho_3/A = \alpha$, $\eta_3 = \xi_{A\alpha}/\zeta_{A\alpha} = (1-\sqrt{A})^2$ and $\beta_3 = \min(\beta_{A\alpha}, (\zeta_{A\alpha}-1)^2/2)$. QED

3.3 Angle Between Subspaces

For each I and each $v \in Range(\Phi_I)$, let $v_{I,1}$ denote the component of v in the subspace $W_{I,1}$

Lemma 3.5 *There are $\rho_4 \in (0, 1/2)$ and $\eta_4 > 0$ so that on an event Ω_n^4*

$$\|v\|_2 \le \eta_4 \|v_{I,1}\|_2 \qquad \forall v \in V_I, \quad |I| \le \rho_4 n,$$

and $P((\Omega_n^4)^c) \le \exp(-n\beta_4)$, $n > n_0$.

Geometrically, this says that every $W_{I,0}$ makes an angle with its corresponding V_I which is well-bounded away from 0.

We begin by considering an individual result, for one specific I. $W_{0,I}$ and V_I are independent uniform random subspaces of R^m of appropriate dimensions. Applying the same reasoning as in Lemma 2.7, we get

Lemma 3.6 *Let A be a random $m-n+k$-dimensional subspace of R^m and let B be an independent random k-dimensional subspace of R^m. For every $\eta > 0$ sufficiently small, there exist $\beta > 0$, n_0 so that*

$$P\{\cos(\angle(A, B)) > 1 - \eta\} \le \exp(-n\beta), \qquad n > n_0.$$

Proof. As in Lemma 2.7, we let X be an $m \times k$ matrix of iid Gaussians $N(0, \frac{1}{m})$, let Y be the result of Gram-Schmidt orthonormalization, and let Z be the first $m-n+k$ rows of Y. We need an upper bound on the top singular value of the matrix Z. We consider instead the matrix $\tilde{Z}$ based on the upper $m-n+k$ rows of X. We note that $\tilde{Z} = \sqrt{\frac{m-n+k}{m}} \cdot \overline{Z}$, where $\overline{Z}$ again has iid Gaussian entries, now standardized so that columns have expected length 1. For such matrices, we invoke Lemma 3.2 and get that for the top singular value, the event

$$\{s_{max}(\overline{Z}) > 1 + \sqrt{k/(m-n+k)} + t\}$$

has probability bounded by $\exp(-(m-n+k)t^2)$. We conclude that the top singular value for $\tilde{Z}$ obeys

$$s_{max}(\tilde{Z}) > (1+t)(\frac{m-n+k}{m})^{1/2} + \sqrt{\frac{k}{m}}$$

with probability bounded by $\exp(-(m-n+k)t^2)$. Choose $\eta \in (0, 1/A)$ so that

$$(1-\eta)^{1/2} > (1 - 1/A)^{1/2},$$

then for small enough $\rho < 1/2$,

$$(1-\eta)^{1/2} - (1 - 1/A + \rho/A)^{1/2} > \sqrt{\rho/A};$$

and we can define $t > 0$ by the solution to

$$(1-\eta)^{1/2} = (1+t)(1 - 1/A + \rho/A)^{1/2} + \sqrt{\rho/A}.$$

Then the event

$$\{s_{max}(\tilde{Z}) > (1-\eta)^{1/2}\}$$

has probability bounded by $\exp(-n(A-1)t^2)$. Now $Z = \tilde{Z}T$, where T is again the triangular matrix that implements Gram-Schmidt on the columns of X and so

$$s_{max}(Z) \le s_{max}(\tilde{Z})\|T\|$$

Again $\|T\| \leq 1/s_{min}(X)$. Invoking again Lemma 3.2 we have that

$$s_{min}(X) \geq 1 - \sqrt{\rho/A} - t$$

except on an event of probability $\leq \exp(-mt^2/2)$. Picking ρ also small enough that

$$1 - \sqrt{\rho/A} > (1+\eta)^{-1/2}$$

we get $P\{\|T\| > (1+\eta)^{1/2}\} \leq \exp(-n\beta)$, for $\beta > 0$. Combining these we get

$$\cos(\angle(A,B)) \leq s_{max}(\tilde{Z})\|T\| \leq (1-\eta)^{1/2} \cdot (1-\eta)^{1/2},$$

except on an event of probability $\leq \exp(-n(A-1)t^2) + \exp(-n\beta)$. QED

Proof of Lemma 3.5. Since

$$\|v\|_2^2 = \|v_0\|_2^2 + \|v_1\|_2^2$$

while

$$\|v_0\|_2 \leq \cos(\angle(W_{I,0}, V_I))\|v\|_2,$$

we get

$$\|v\|_2 \leq (1 - \cos^2(\angle))^{-1/2}\|v_1\|.$$

Applying Lemma 2.8 together with Lemma 3.6 gives that for some $\eta > 0$, on an event Ω_n^4

$$\cos(\angle(W_{I,0}, V_I)) \leq 1 - \eta, \quad |I| < \rho_4 n.$$

Hence Lemma 3.5 holds with $\eta_4 = 1/(2\eta - \eta^2)$. QED

3.4 Equivalence to Euclidean Norm

We now discuss equivalence between the Euclidean norm and the ℓ^1 norm on the subspaces $B_{I,0}$ and $C_{I,1}$.

Lemma 3.7 *For small enough $\rho_5 > 0$ there is $\eta_5 > 0$ so that, on an event Ω_n^5,*

$$\eta_5\sqrt{A} \cdot \|\beta_0 + \gamma_1\|_2 \leq \|\beta_0 + \gamma_1\|_1/\sqrt{n} \leq \sqrt{A} \cdot \|\beta_0 + \gamma_1\|_2$$

uniformly over $\beta_0 \in B_{I,0}$, $\gamma_1 \in C_{I,1}$ and $|I| < \rho_5 n$. The probability of the exceptional event $(\Omega_n^5)^c$ is bounded by $\exp(-n\beta)$ for $n > n_0$, where $\beta > 0$.

Recall that in the previous case $m = n$, we showed that sections were almost spherical, i.e. that the constants in such norm equivalence statements could be made close to 1. In that case, the subspaces involved were of small dimension relative to the ambient space R^m. Now we are considering cases where the subspaces are of substantial dimension $> ((A-1)/A)m$ relative to m. We no longer get equivalence between ℓ^1 norms and ℓ^2 norms with constants close to 1, but we still get equivalence. The insight that this can happen goes back to Kashin [18].

A convenient expression for this phenomenon has been developed in Pisier's book [24, Chapter 6]; it is based on the volume ratio notion introduced by Szarek [25], and shows that random subspaces will work (which is what we need).

Lemma 3.8 *Let $\ell < m$ and let U be a random uniform $(m-\ell)$-dimensional subspace of R^m. On an event E_m we have the norm equivalence*

$$c_{l,m}\|u\|_2 \le \|u\|_1/\sqrt{m} \le \|u\|_2 \qquad \forall u \in U,$$

with

$$1/c_{l,m} = (8e)^{\frac{m}{m-\ell}}.$$

The exception probability $P(E_m^c) \le 2^{-m}$.

This immediately implies the following result for one single pair $B_{I,0}$, $C_{I,1}$.

Lemma 3.9 *For $\rho \in (0,1/2)$ set $\eta_\rho = c_{\ell,m}$, where $\ell = \lfloor m - n + 2\rho n \rfloor$. On an event $\Omega_{n,I}$,*

$$\eta_\rho\sqrt{A}\|\beta_0+\gamma_1\|_2 \le \|\beta_0+\gamma_1\|_1/\sqrt{n} \le \sqrt{A}\|\beta_0+\gamma_1\|_2$$

uniformly over $\beta_0 \in B_{I,0}$, $\gamma_1 \in C_{I,1}$. The probability of the exceptional event $(\Omega_{n,I})^c$ is bounded by 2^{-n}.

Proof of Lemma 3.7. Fix a $\rho \in (0,1/2)$ as provided by Lemma 3.9. Set $\eta_5 = \eta_\rho\sqrt{A}$. Apply Lemma 2.8 with ρ as input, getting ρ' as output. Set $\rho_5 = \rho'$ and $\eta_5 = \sqrt{A}\eta_\rho$. QED.

References

[1] E.J. Candès, J. Romberg and T. Tao. (2004) Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. Manuscript.

[2] Chen, S., Donoho, D.L., and Saunders, M.A. (1999) Atomic Decomposition by Basis Pursuit. *SIAM J. Sci Comp.*, **20**, 1, 33-61.

[3] R.R. Coifman, Y. Meyer, S. Quake, and M.V. Wickerhauser (1990) Signal Processing and Compression with Wavelet Packets. in *Wavelets and Their Applications*, J.S. Byrnes, J. L. Byrnes, K. A. Hargreaves and K. Berry, eds. 1994,

[4] K.R. Davidson and S.J. Szarek (2001) Local Operator Theory, Random Matrices and Banach Spaces. *Handbook of the Geometry of Banach Spaces, Vol. 1* W.B. Johnson and J. Lindenstrauss, eds. Elsevier.

[5] Donoho, D.L. and Huo, Xiaoming (2001) Uncertainty Principles and Ideal Atomic Decomposition. *IEEE Trans. Info. Thry.* **47** (no. 7), Nov. 2001, pp. 2845-62.

[6] Donoho, D.L. and Elad, Michael (2002) Optimally Sparse Representation from Overcomplete Dictionaries via ℓ^1 norm minimization. *Proc. Natl. Acad. Sci. USA* March 4, 2003 **100** 5, 2197-2002.

[7] Donoho, D., Elad, M., and Temlyakov, V. (2004) Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise. Submitted. URL: http://www-stat.stanford.edu/~donoho/Reports/2004.

[8] Donoho, D.L. (2004) For most underdetermined systems of linear equations, the minimal ℓ^1 solution is also the sparsest solution. Manuscript. Submitted. URL: http://www-stat.stanford.edu/~donoho/Reports/2004

[9] A. Dvoretsky (1961) Some results on convex bodies and Banach Spaces. *Proc. Symp. on Linear Spaces.* Jerusalem, 123-160.

[10] A. Edelman, Eigenvalues and condition numbers of random matrices, SIAM J. Matrix Anal. Appl. 9 (1988), 543-560.

[11] B. Efron, I. Johnstone, T. Hastie, and R. Tibshirani (2004) Least Angle Regression. *Ann. Statist.* **32**, 407-451.

[12] Noureddine El Karoui (2004) *New Results About Random Covariance Matrices and Statistical Applications.* Ph.D. Thesis, Stanford University.

[13] M. Elad and A.M. Bruckstein (2002) A generalized uncertainty principle and sparse representations in pairs of bases. *IEEE Trans. Info. Thry.* **49** 2558-2567.

[14] J.J. Fuchs (2002) On sparse representation in arbitrary redundant bases. Manuscript.

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

[16] R. Gribonval and M. Nielsen. Sparse Representations in Unions of Bases. To appear *IEEE Trans Info Thry.*

[17] G. Golub and C. van Loan.(1989) *Matrix Computations.* Johns Hopkins: Baltimore.

[18] Boris S. Kashin (1977) Diameters of certain finite-dimensional sets in classes of smooth functions. *Izv. Akad. Nauk SSSR, Ser. Mat.* **41** (2) 334-351.

[19] Michel Ledoux. *The Concentration of Measure Phenomenon.* Mathematical Surveys and Monographs **89**. American Mathematical Society 2001.

[20] S. Mallat, Z. Zhang, (1993). "Matching Pursuits with Time-Frequency Dictionaries," IEEE Transactions on Signal Processing, 41(12):3397–3415.

[21] V.D. Milman and G. Schechtman (1986) *Asymptotic Theory of Finite-Dimensional Normed Spaces.* Lect. Notes Math. **1200**, Springer.

[22] B.K. Natarajan (1995) Sparse Approximate Solutions to Linear Systems. *SIAM J. Comput.* **24**: 227-234.

[23] B. N. Parlett (1980) *The Symmetric Eigenvalue Problem.* Prentice Hall.

[24] G. Pisier (1989) *The Volume of Convex Bodies and Banach Space Geometry.* Cambridge University Press.

[25] Szarek, S.J. (1978) On Kashin's almost -Euclidean orthogonal decomposition of ℓ_1^n. *Bull. Acad. Polon. Sci Sér. Sci. Math. Astronom. Phys.* **26**, 691-694.

[26] Szarek, S.J. (1990) Spaces with large distances to ℓ_∞^n and random matrices. *Amer. Jour. Math.* **112**, 819-842.

[27] Szarek, S.J.(1991) Condition Numbers of Random Matrices. *J. Complexity* **7**, 131-149.

[28] R. Tibshirani (1995) Regression Shrinkage and Subset Selection with the Lasso. *Journ. Roy. Stat. Soc. B*, **58**, 267-288.

[29] J.A. Tropp (2003) Greed is Good: Algorithmic Results for Sparse Approximation To appear, *IEEE Trans Info. Thry.*

[30] J.A. Tropp (2004) Just Relax: Convex programming methods for Subset Sleection and Sparse Approximation. Manuscript.