On the Role of Sparsity in Compressed Sensing and Random Matrix Theory

Roman Vershynin
Department of Mathematics
University of Michigan
Ann Arbor, MI 48109
Email: romanv@umich.edu

***Abstract*—We discuss applications of some concepts of Compressed Sensing in the recent work on invertibility of random matrices due to Rudelson and the author. We sketch an argument leading to the optimal bound $\Omega(N^{-1/2})$ on the median of the smallest singular value of an $N \times N$ matrix with random independent entries. We highlight the parts of the argument where sparsity ideas played a key role.**

I. Introduction

A concept that underlies the many recent developments in the area of Compressed Sensing is *sparsity*. Much earlier, sparsity has been used in a similar way (but often implicitly) in theoretical mathematics, and most notably in Geometric Functional Analysis. Recently the understanding of the role of sparsity led to formalizing some of the connections between the statements in those areas, leading to a new interplay between "pure" and "applied" mathematics.

While applications of "pure" mathematics to Compressed Sensing are expected and indeed quite common, the reverse direction – from Compressed Sensing to mathematics – is still rarely seen. This paper discusses one such application to the problem of *invertibility of random matrices*, which has been addressed in particular in the papers of Rudelson and the author [6], [7], [11]. Specifically, we will outline the role of sparsity in the proof of a conjecture that goes back to von Neumann and his collaborators, and which states that the smallest singular value $s_N(A)$ of an $N \times N$ matrix A with random independent centered entries is order $N^{-1/2}$ with high probability. This was proved for general random matrices in [6], [8], and extended to rectangular random matrices in [7], [11].

Since we would like to focus here on techniques rather than results, we will often make oversimplifying assumptions and state weaker forms of the results. For the same reason, we discuss very little of history of these results and related work. The interested reader is encouraged to look at the original papers cited above for the statements of complete results, and for bibliography.

We will denote positive absolute constants by $C, c, C_1, \ldots$; their values may change from line to line.

II. Sparsity as entropy control

One normally thinks of sparsity as a way to represent objects (vectors or functions) in a certain basis in an economical way – such that only a small number of basis elements can be used to accurately represent each object. In this discussion, we shall identify the basis with the canonical basis of $\mathbb{R}^N$, and our objects will be vectors in $\mathbb{R}^N$. We then say that a vector in $\mathbb{R}^N$ is *s-sparse* if it has few non-zero coordinates:

$$|\mathrm{supp}(x)| \le s \ll N.$$

We shall denote the set of all such vectors in $\mathbb{R}^N$ by $Sparse(N, s)$. This set clearly consists of the union of all s-dimensional subspaces of $\mathbb{R}^N$.

An efficient way to use sparsity is through control of the metric entropy of the space $Sparse(N, s)$. Recall that, given a subset S of a metric space and a number $\varepsilon > 0$, the *covering number* $\mathcal{N}(S, \varepsilon)$ is the smallest cardinality of an ε-net of S, i.e. the smallest number of ε-balls centered at points in S needed to cover S. The logarithm of the covering number is often called *metric entropy*.

A simple argument based on comparison of volumes leads to an exponential bound of the metric entropy of many natural subsets of $\mathbb{R}^N$, and in particular of the Euclidean sphere S^{N-1}, see e.g. Lemma 9.5 in [5]. This bound for the interesting range $\varepsilon \in (0, 1)$ reads as

$$\mathcal{N}(S^{N-1}, \varepsilon) \le (3/\varepsilon)^N. \tag{1}$$

This bound improves significantly for the set of sparse vectors. Since there are $\binom{N}{s}$ ways to choose the support of a sparse vector, we have

$$\mathcal{N}\big(Sparse(N, s) \cap S^{N-1}, \varepsilon\big) \le \binom{N}{s} \mathcal{N}(S^{N-1}, \varepsilon).$$

Using (1) along with the bound $\binom{N}{s} \le (eN/s)^s$ valid for $1 \le s \le N/2$, which follows from Stirling's formula, we conclude with

$$\mathcal{N}\big(Sparse(N, s) \cap S^{N-1}, \varepsilon\big) \le \Big(\frac{CN}{s\varepsilon}\Big)^s. \tag{2}$$

Comparing this with (1), we see that sparse vectors enjoy significantly smaller entropy than the whole sphere – the covering number is essentially exponential in the sparsity s rather than the dimension N. This advantage is crucially used in many arguments, such as in the following one.

In Compressed Sensing, a basic quality of matrices that guarantees their good performance as measurement operators is the *Restricted Isometry Condition*. An $n \times N$ matrix A with $n \leq N$ is said to satisfy the Restricted Isometry Condition (RIC) if A acts as an approximate isometry when restricted to the set of sparse vectors. Formally, for every integer $s \leq N$ we define the RIC constant δ_s of the matrix A as the minimal number that satisfies the two-sided inequality

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2 \qquad (3)$$

for all $x \in \mathit{Sparse}(N,s)$. Candes and Tao [2] have shown (with constant improved in [1]) that, given a matrix A with $\delta_{2s} \leq \sqrt{2}-1$, one can exactly recover every s-sparse vector x from its "measurement vector" $y = Ax$ by solving the convex optimization problem

$$\min_{x \in \mathbb{R}^N} \|x\|_1 \quad \text{subject to} \quad Ax = y.$$

While it is difficult to explicitly construct matrices with good dimensions and RIC constants, random constructions are abundant in the literature (see Section V in [3]). Here we sketch the known argument for Gaussian matrices, which will highlight sparsity as entropy control, and will lead to our discussion of more difficult questions in Random Matrix Theory.

Proposition 1 (Gaussian matrices): Let $\bar{A}$ be an $n \times N$ matrix whose entries are independent standard normal random variables. Let $1 \leq s \leq N$ and $\delta > 0$. If

$$n \geq C(\delta) s \log(N/s)$$

then, with high probability, the matrix $A = \frac{1}{\sqrt{n}}\bar{A}$ satisfies RIC with constant $\delta_s \leq \delta$. Here $C(\delta) > 0$ only depends on δ.

Proof: (Sketch) An approximation argument shows that it is enough to check (3) for all x in any fixed δ-net of $\mathit{Sparse}(N,s) \cap S^{N-1}$. So we choose such a net $\mathcal{N}$ of cardinality controlled as in (2), and we fix a vector $x \in \mathcal{N}$. Due to independence of the rows of A and the rotation invariance of the normal distribution, the random variable $\|Ax\|_2^2$ is distributed identically with $\chi^2 := \sum_{i=1}^n g_i^2$, where g_i are independent standard normal random variables. By the known concentration properties of the χ^2 distribution, or alternatively by the standard exponential concentration inequalities, one has

$$(1-\delta)n \leq \sum_{i=1}^n g_i^2 \leq (1+\delta)n$$

with probability at least $1 - e^{-c(\delta)n}$. In other words, with this probability, the Restricted Isometry Condition (3) holds for a fixed vector $x \in \mathcal{N}$. Taking the union bound and using the bound (2) on the cardinality of the net, we see that (3) holds for *all* vectors $x \in \mathcal{N}$ with probability at least

$$1 - |\mathcal{N}|e^{-c(\delta)n} \geq 1 - (CN/s)^s e^{-c(\delta)n}.$$

By the condition we made on the dimensions, the proof is complete. ∎

The argument above can be easily generalized to distributions other than normal by using standard exponential concentration inequalities; suitable moment bounds (subgaussian) are sufficient for this purpose.

III. Invertibility of random matrices

One can view the Restricted Isometry Condition (3) as the condition that all submatrices of A with a given number of columns are well conditioned. The question of how well conditioned random matrices are goes back to at least Von Neumann and his collaborators, in connection with their work on large matrix inversion. Some history of the work on this problem is described in [6] and [7], and some new results appeared since then, see [10]. Here we shall focus on the original prediction going back to Von Neumann and his group – that the smallest singular value $s_N(A)$ of an $N \times N$ matrix with random independent centered entries is typically of order $N^{-1/2}$. Coupled with the known estimate on the largest singular value $\mathbb{E}s_1(A) \leq N^{1/2}$ (valid under suitable moment assumptions), the prediction implies that the condition number $\kappa(A) = s_1(A)/s_N(A) = O(N)$, i.e. is typically linear in the dimension.

This prediction was verified for Gaussian matrices in [4], [9] using the explicit formula for the joint density of their eigenvalues, and was first proved for general random matrices in [6] under some mild moment assumptions. Ideas based on sparsity play an important role in [6]. We will discuss this role the in the rest of the paper, and sketch the proof of the prediction above:

Theorem 2 ([6]): Let A be an $N \times N$ matrix whose entries are independent identically distributed random variables with mean zero, unit variance, and fourth moment bounded by a constant. Then the median of $s_N(A)$ is bounded below by $cN^{-1/2}$.

Note that the result is sharp – it was proved in [8] that the median of $s_N(A)$ is bounded above by $CN^{-1/2}$.

IV. Invertibility on sparse vectors

Our plan is to first prove Theorem 2 for Gaussian matrices A (whose all entries are standard normal random variables), and then to indicate how to modify the proof for general distributions.

The smallest singular value has the following convenient expression:

$$s_N(A) = \min_{x \in S^{N-1}} \|Ax\|_2.$$

Our goal is then to bound $\|Ax\|_2$ below uniformly for all unit vectors x.

We already know how to achieve this goal for all *sparse* vectors x. Indeed, by Proposition 1 the Gaussian matrices satisfy the Restricted Isometry Property. If we choose $n = N$ and $s = cN$ with sufficiently small absolute constant $c > 0$, Proposition 1 shows that, with high probability,

$$\min_{x \in \mathit{Sparse}(N,cN) \cap S^{N-1}} \|Ax\|_2 \geq cN^{1/2}.$$

(Although the choice $n = N$ is not a typical one in Compressed Sensing where one would like to take significantly smaller number of measurements n then the signal dimension N, for our problem on square matrices one has to start with $n = N$).

Note that this bound is much better than we need in Theorem 2 – we would be happy with $cN^{-1/2}$ in the right hand side.

Now we need to handle the non-sparse vectors.

V. Invertibility on spread vectors

Our success with sparse vectors is due the fact that there are "not too many" of them. As we have seen by comparing (1) to (2), the metric entropy of the set of sparse vectors is much smaller than that of all vectors. Such a nice entropy control allowed us to handle all sparse vectors by taking a union bound (in the proof of Proposition 1) without paying too much price in the probability estimates.

Repeating a similar argument for non-sparse vectors is hopeless, as they lack a nice entropy control. Instead, we could first try to identify the class of vectors which is entirely opposite to the sparse vectors, and try to handle this class. These are *spread vectors* – those vectors in S^{N-1} whose all coordinates have the same order $N^{-1/2}$, i.e. $x \in S^{N-1}$ is spread if $|x_i| = \Omega(N^{-1/2})$ for all $i \le N$. An advantage of spread vectors over sparse ones is that we know the magnitude of all their coefficients. So we develop the following *geometric* argument to prove the invertibility on the set of spread vectors.

Let us begin with a qualitative argument. Suppose the matrix A performs extremely poor, and we have $s_N(A) = 0$; in other words, A is a singular matrix. Therefore one of its columns X_k of A lies in the span $H_k = \mathrm{span}(X_i)_{i \ne k}$ of the others.

This simple observation can be made into a quantitative argument, which will work very well with the spread vectors. Suppose $x = (x_1, \dots, x_N) \in \mathbb{R}^N$ is a spread vector. Then, for every $k = 1, \dots, N$, we have

$$\begin{aligned} \|Ax\|_2 &\ge \mathrm{dist}(Ax, H_k) = \mathrm{dist}\Big(\sum_{i=1}^N x_i X_i, H_k\Big) \qquad (4) \\ &= \mathrm{dist}(x_k X_k, H_k) = |x_k| \cdot \mathrm{dist}(X_k, H_k) \\ &\ge cN^{-1/2}\, \mathrm{dist}(X_k, H_k). \end{aligned}$$

Since the right hand side does not depend on x, we have proved in particular (for $k = 1$) that

$$\min_{\text{Spread } x} \|Ax\|_2 \ge cN^{-1/2}\, \mathrm{dist}(X_1, H_1).$$

It remains to estimate the distance between the random vector X_1 and the independent random hyperplane H_1. Since X_1 is a Gaussian vector, it is easy to check (using the rotation invariance of the Gaussian disstibution) that $\mathrm{dist}(X_1, H_1)$ is distributed identically with the absolute value of a standard normal random variable g. But the probability density function of g is bounded by the absolute constant $(2\pi)^{-1/2}$, which yields

$$\mathrm{dist}(X_1, H_1) = |g| = \Omega(1)$$

with arbitrarily high constant probability (say, 0.999). (Here and thereafter the notation $f = \Omega(g)$ means that $f \ge \mathrm{const} \cdot g$ for some constant that does not depend on the dimensions (such as N, s) in question.

We have thus shown that, with arbitrarily high constant probability,

$$\min_{\text{Spread } x} \|Ax\|_2 \ge cN^{-1/2}.$$

This is a desired uniform bound for the spread vectors.

VI. Bridging sparse and spread vectors

There are of course many vectors that are neither sparse nor spread, but it will now be relatively easy to bridge these two classes.

Consider all vectors in S^{N-1} that are within a small absolute constant distance $c' > 0$ from the set of sparse vectors $\mathit{Sparse}(N, cN) \cap S^{N-1}$. We shall call such vectors *compressible*, and the rest of the vectors on the sphere are *incompressible*. The intuition, which again is coming from sparse recovery, suggests that compressible vectors should behave similarly to sparse vectors, while incompressible vectors should be similar to spread vectors.

Indeed, a trivial approximation argument extends our invertibility bound from sparse to compressible vectors (one just need to approximate a compressible vector by a sparse one and use that the error of this approximation c' can only blow up by a factor $\|A\| = O(N^{1/2})$). So we have the desired bound

$$\min_{\text{Compressible } x} \|Ax\|_2 \ge cN^{1/2}.$$

For incompressible vectors, instead of an approximation argument (which won't work) one makes the following simple observation: every incompressible vector has $\Omega(N)$ coordinates of magnitude $\Omega(N^{-1/2})$. This is a way how incompressible vectors are similar to spread ones.

To complete the proof for incompressible vectors, we again use the geometric argument, but stop just before the last estimate in (4):

$$\|Ax\|_2 \ge \max_k |x_k| \cdot \mathrm{dist}(X_k, H_k).$$

As we already know, for each $k = 1, \dots, N$, the distance satisfies $\mathrm{dist}(X_k, H_k) = \Omega(1)$ with arbitrarily large probability. Therefore, still with high probability, most of these distances (arbitrarily high constant proportion of them) are $\Omega(1)$. On the other hand, we also know that some fixed proportion of the coordinates x_k are of magnitude $\Omega(N^{-1/2})$. Therefore, intersecting these two events, we see that for any incompressible vector x there exists a coordinate k that satisfies both bounds. This implies that, with high probability,

$$\min_{\text{Incompressible } x} \|Ax\|_2 \ge cN^{-1/2}.$$

This is a desired bound which, along with the already proved estimate for compressible vectors, implies the final result:

$$s_N(A) = \min_{x \in S^{N-1}} \|Ax\|_2 \ge cN^{-1/2}.$$

VII. Extensions and further remarks

The above argument generalizes from Gaussian to general distributions. There are two places where we used rotation invariance of the Gaussian distribution. One such place was the use of Proposition 1 in the treatment of sparse vectors. As we already mentioned, Proposition 1 can be easily extended to more general distributions using the standard large deviation inequalities.

The other place where Gaussian distribution was used was in the argument for spread vectors. We argued there that the distance $\mathrm{dist}(X_1, H_1)$ between a random vector and a random independent hyperplane is $\Omega(1)$ with arbitrarily high probability. For Gaussian distribution of the entries, this followed by a direct and easy computation. For more general distributions, this estimate is still true, but it requires more work.

Let us condition on a realization of the hyperplane H_1, and let $a \in \mathbb{R}^N$ be a unit normal vector of H_1. Then clearly

$$\mathrm{dist}(X_1, H_1) = \langle a, X_1 \rangle.$$

Writing this in coordinates for $a = (a_1, \dots, a_N)$ and $X_1 = (\xi_1, \dots, \xi_N)$ we see that

$$\mathrm{dist}(X_1, H_1) = \Big| \sum_{i=1}^{N} a_i \xi_i \Big| =: |S|$$

where S is clearly a sum of independent random variables.

Our goal is to show that $|S| = \Omega(1)$ with high probability. One way to do this is to use a Central Limit Theorem (in the form of Berry-Esseen) to approximate S by a standard normal random variable g, for which we already have the desired result. For the The Central Limit Theorem to work, one obviously needs that many coordinates of a are not too small (for example, it will clearly not work if a is 1-sparse, as the sum S will consist of just one term). However, sparsity ideas can be again of help here. Running an argument similar to the one above for compressible vectors, one can show that, with high probability, the normal a to the random hyperplane H_1 is incompressible. We then condition on such H_1, and the Central Limit Theorem works well:

$$\big| \mathbb{P}(|S| < \varepsilon) - \mathbb{P}(|g| < \varepsilon) \big| = O(N^{-1/2}). \tag{5}$$

This proves the desired bound $\mathrm{dist}(X_1, H_1) = \Omega(1)$ with high probability $1 - O(N^{-1/2})$.

In [7], [11], Theorem 2 was extended to *rectangular matrices* $N \times n$, where $N \geq n$. Under the same assumptions, the median of the smallest singular value $s_n(A)$ of such random matrices is bounded below by $c(\sqrt{N} - \sqrt{n-1})$, which is asymptotically optimal. Note that for square matrices, where $N = n$, this bound equals $cN^{-1/2}$, which agrees with Theorem 2.

Under stronger moment assumptions on the entries (subgaussian), not only the median of the smallest singular value can be estimated, but also strong probability inequalities can be proved. For example, square matrices satisfy

$$\mathbb{P}(s_N(A) < \varepsilon N^{-1/2}) \leq C\varepsilon + e^{-cN}$$

and rectangular matrices satisfy

$$\mathbb{P}\big(s_n(A) \leq \varepsilon(\sqrt{N} - \sqrt{n-1})\big) \leq (C\varepsilon)^{N-n+1} + e^{-cN}$$

for all $\varepsilon \geq 0$. Proving such exponential inequalities is more difficult, because one can not afford a polynomial error in probability $O(N^{-1/2})$ which one necessarily obtains when applying Central Limit Theorem in (5). Instead of using Central Limit Theorems, one develops a Littlewood-Offord Theory, whose probability estimates are fine-tuned to the additive structure of the coefficients of a. Since the sparsity does not play a key role in these arguments, we will not discuss this direction here. The interested reader is encouraged to consult the papers [7], [11].

Acknowledgment

This research was partially supported by NSF FRG grant DMS 0918623.

References

[1] E. J. Candes, *The restricted isometry property and its implications for compressed sensing*, Compte Rendus de l'Academie des Sciences - Mathematique 346 (2008), 589-592.

[2] E. J. Candes, T. Tao, *Decoding by linear programming*, IEEE Trans. Information Theory 51 (2005), 4203–4215.

[3] E. J. Candes and M. Wakin, *An introduction to compressive sampling*, IEEE Signal Processing Magazine, March 2008, 21–30.

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

[5] M. Ledoux and M. Talagrand, *Probability in Banach spaces. Isoperimetry and processes*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 23. Springer-Verlag, Berlin, 1991.

[6] M. Rudelson, R. Vershynin, *The Littlewood-Offord problem and invertibility of random matrices*, Advances in Mathematics 218 (2008), 600–633.

[7] M. Rudelson, R. Vershynin, *The smallest singular value of a random rectangular matrix*, Communications on Pure and Applied Mathematics, to appear.

[8] M. Rudelson, R. Vershynin, *The least singular value of a random square matrix is $O(n^{-1/2})$*, Comptes rendus de l'Academie des sciences - Mathematique 346 (2008), 893–896.

[9] S. Szarek, *Condition numbers of random matrices*, J. Complexity 7 (1991), 131–149.

[10] T. Tao, V. Vu, *Random matrices: The distribution of the smallest singular values*, Geometric and Functional Analysis, to appear.

[11] R. Vershynin, *Spectral norm of products of random and deterministic matrices*, to appear.