

Sparse Representations and Invertibility of Random Matrices

Roman Vershynin Mark Rudelson

University of California, Davis

University of Missouri-Columbia

Von Neumann Symposium
Sparse Representation and High Dimensional Geometry
Snowbird, Utah, July 2007

Sparse Representations and Random Matrices

- Sparse representation theory relies on advancements in several fields of Mathematics, in particular

Random Matrix Theory \longrightarrow Sparse Representations.

(Example: random measurements in Compressed Sensing).

- This talk: the reverse application

Sparse Representations \longrightarrow Random Matrix Theory.

Sparse Representations and Random Matrices

- Question. Are random square matrices well or ill conditioned?

Conjecture (Von Neumann, Goldstine, Smale).

The condition number of an $n \times n$ random matrix is $O(n)$.

- This talk: *sparse representations and high dimensional geometry* in application to this problem.

Singular values and condition number

- A : an $n \times n$ matrix with real or complex entries.

The *singular values* $s_k(A)$ are the eigenvalues of $|A| = \sqrt{A^*A}$ in the non-increasing order.

- Thus all singular values are between

$$s_1(A) = \sup_{x: \|x\|=1} \|Ax\| = \|A\| \quad \text{and} \quad s_n(A) = \inf_{x: \|x\|=1} \|Ax\| = \frac{1}{\|A^{-1}\|}.$$

- The *condition number*

$$\kappa(A) = \frac{s_1(A)}{s_n(A)}$$

measures the worst *distortion* of vectors in \mathbb{R}^n under A .

Singular values and condition number

- It is well known that for random matrices with i.i.d. entries

$$s_1(A) \sim n^{1/2} \quad \text{with high probability.}$$

Von Neumann and Goldstine proved this for Gaussian matrices. They speculated (1947, 1963) based on numerical evidence that

$$s_n(A) \sim n^{-1/2} \quad \text{with high probability?}$$

Smale (1985) added a conjectural bound on the “high probability”.

- In other words, both $\|A\|$ and $\|A^{-1}\|$ should be of order $n^{1/2}$. Thus the condition number should grow **linearly** in the dimension:

$$\kappa(A) \sim n \quad \text{with high probability?}$$

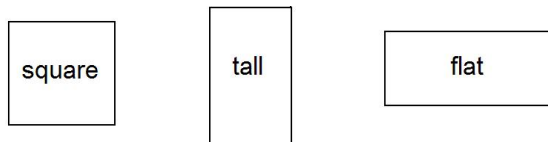
[Von Neumann and Goldstine, 1947]

Smallest singular value

Conjecture (Von Neumann-Goldstine 1947, Smale 1985).

For an $n \times n$ matrix A with i.i.d. centered entries, $s_n(A) \gtrsim n^{-1/2}$ w.h.p.

- **Difficulty.** Square matrices are *at the border* between *tall* matrices (well conditioned) and *flat* matrices (ill conditioned).



- **Previous positive results.**
True for *Gaussian matrices* (Edelman, 1988).
For more general matrices, *polynomial bounds* were proved by Rudelson (2005) and Tao-Vu (2006).
- **Simplest previously unsolved case.**
Bernoulli matrices (± 1 uniform entries).
Spielman-Teng (ICM 2002) stated a very precise conjecture:

Smallest singular value

Conjecture (Von Neumann-Goldstine 1947, Smale 1985).

For an $n \times n$ matrix A with i.i.d. centered entries, $s_n(A) \gtrsim n^{-1/2}$ w.h.p.

Conjecture (Spielman-Teng, ICM 2002).

For an $n \times n$ random Bernoulli matrix A , and for every $\varepsilon > 0$,

$$\mathbb{P}(s_n(A) \leq \varepsilon n^{-1/2}) \leq \varepsilon + c^n, \quad c = \text{const} \in (0, 1).$$

- **Singularity probability.** A remarkable partial case is for $\varepsilon = 0$:

$$\mathbb{P}(A \text{ is singular}) \leq c^n.$$

It is even nontrivial that the singularity probability $\rightarrow 0$ as $n \rightarrow \infty$; this was proved by Komlós (1967). Exponential bounds:

Kahn-Komlós-Szemerédi (1995), Tao-Vu (2006).

Erdős conjecture: $c = \frac{1}{2} + o(1)$.

Smallest singular value

Conjecture (Von Neumann-Goldstine 1947, Smale 1985).

For an $n \times n$ matrix A with i.i.d. centered entries, $s_n(A) \gtrsim n^{-1/2}$ w.h.p.

Conjecture (Spielman-Teng, ICM 2002).

For an $n \times n$ random Bernoulli matrix A , and for every $\varepsilon > 0$,

$$\mathbb{P}(s_n(A) \leq \varepsilon n^{-1/2}) \leq \varepsilon + c^n, \quad c = \text{const} \in (0, 1).$$

- **Main result.** Both these conjectures are true (in full generality).
- They hold for arbitrary *subgaussian* matrices:
i.i.d. centered entries, all moments bounded by the corresponding moments of the standard Gaussian random variable.

Examples: Gaussian, Bernoulli, uniform.

Smallest singular value

Theorem (Main)

For an $n \times n$ matrix A with i.i.d. subgaussian entries, and for $\varepsilon > 0$,

$$\mathbb{P}(s_n(A) \leq \varepsilon n^{-1/2}) \leq C\varepsilon + c^n.$$

This bound is *optimal*. It implies:

- Von Neumann-Goldstine-Smale's conjecture:

$$s_n(A) \gtrsim n^{-1/2} \quad \text{with high probability.}$$

- In particular, **condition number** is linear in the dimension:

$$\kappa(A) = O(n) \quad \text{with high probability.}$$

- Spielman-Teng's conjecture (for general matrices).
- Kahn-Komlós-Szemerédi bound on the singularity probability

$$\mathbb{P}(A \text{ is singular}) \leq c^n.$$

(take $\varepsilon = 0$). Holds for general matrices, not only Bernoulli.

Argument: sparse representations

$$s_n(A) = \inf_{x \in S^{n-1}} \|Ax\| \gtrsim n^{-1/2} \quad \text{with high probability?}$$

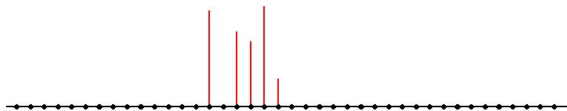
- The **proof** uses *sparse representations*. The idea to use sparsity is not new; e.g. [Kashin, 1977] on Euclidean sections of L_1 ball.
- **Case study**. We illustrate the method on the important example of **Gaussian matrices**, where the result is known [Edelman, 1988].
- **Difficulty**. Edelman used the explicit formula for the *joint density* of the singular values of A . Our method has to be completely different (Bernoulli matrix has *no density!*)
- **Sparsity**. Rather than bound $\|Ax\|$ below for all $x \in S^{n-1}$ *at once*, we shall first prove it separately for the two extreme types of vectors x : **sparse** and **spread**.

We start with *sparse* vectors...

Argument 1. Sparse vectors

- A vector $x \in \mathcal{S}^{n-1}$ is *sparse* if it has few nonzero coordinates:

$$|\text{supp}(x)| \leq \delta n \quad (\delta \sim 0.01)$$



- Want to show: $\|Ax\|$ is bounded below for all sparse vectors x .
- This follows from the **Restricted Isometry Property** (R.I.P.) of A , introduced by [Candes-Tao, 2004] in Compressed Sensing:

$$0.8\|x\| \leq \left\| \frac{1}{\sqrt{n}} Ax \right\| \leq 1.2\|x\| \quad \text{for all sparse } x.$$

- Let us look only at the left hand side:

$$\inf_{x \in \text{Sparse}} \|Ax\| \geq \sqrt{n} \gg n^{-1/2}.$$

So we are done for the sparse vectors.

- We move on to the opposite class of *spread* vectors... Challenge.

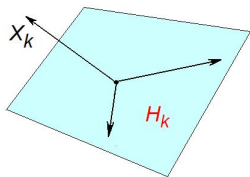
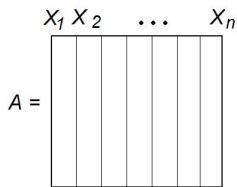
Argument 2. Spread vectors

- A vector $x \in S^{n-1}$ is *spread* if all its coordinates are about the same: $|x_k| \sim \frac{1}{\sqrt{n}}$



- **Difficulty.** Spread vectors contain *more information* than sparse. Mathematically, the set of sparse vectors has a small ϵ -net. The set of spread vectors does *not*.
- **Question.** How can **non-sparsity** be an advantage?
- **Answer.** We know *the magnitudes* of all coefficients (we did not know these for sparse vectors!)
- So, a completely different *geometric argument* for spread vectors:

Argument 2. Spread vectors



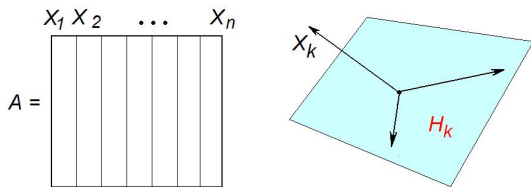
- The trivial rank argument:

A nonsingular \Leftrightarrow each column X_k is not in the span H_k of the others

- A *quantitative* version should look something like:

$$\inf_x \|Ax\| \geq \dots \text{dist}(X_k, H_k)$$

Argument 2. Spread vectors



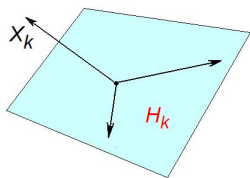
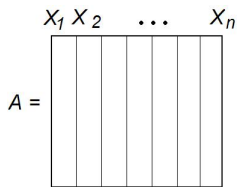
Here is such a geometric argument for spread x :

$$\begin{aligned}\|Ax\| &\geq \text{dist}(Ax, H_k) = \text{dist}\left(\sum x_k X_k, H_k\right) = \text{dist}(x_k X_k, H_k) \\ &= |x_k| \cdot \text{dist}(X_k, H_k) \sim \frac{1}{\sqrt{n}} \text{dist}(X_k, H_k).\end{aligned}$$

The right hand side does not depend on x . Thus

$$\inf_{x \in \text{Spread}} \|Ax\| \gtrsim \frac{1}{\sqrt{n}} \text{dist}(X_n, H_n).$$

Argument 2. Spread vectors



- We have shown: $\inf_{x \in \text{Spread}} \|Ax\| \gtrsim \frac{1}{\sqrt{n}} \text{dist}(X_n, H_n)$ w.h.p.
It remains to estimate the distance.
- X_n is a Gaussian vector, H_n is an independent hyperplane. Then

$$\text{dist}(X_n, H_n) = |\text{gaussian}| \sim \text{const} \quad \text{with high probability.}$$

- This proves the invertibility of A on spread vectors:

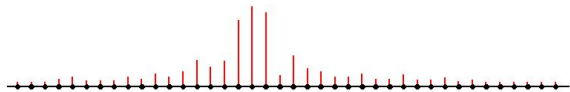
$$\inf_{x \in \text{Spread}} \|Ax\| \geq n^{-1/2} \quad \text{with high probability.}$$

Argument 3. Compressible, incompressible vectors

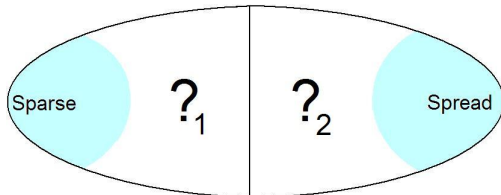
- **Conclusion.** The invertibility of A for sparse and spread vectors:

$$\inf_{x \in \text{Sparse} \cup \text{Spread}} \|Ax\| \geq n^{-1/2} \quad \text{with high probability.}$$

- Unfortunately, there \exists vectors that are *neither sparse nor spread*:



- So, how to fill the gap – how to truly decompose S^{n-1} ?



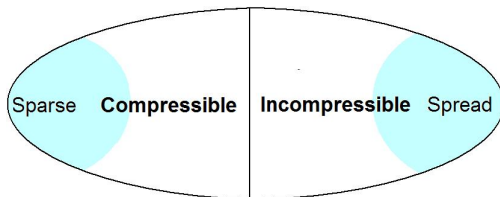
- **Example of $?_1$:** wavelet coefficients. They are not sparse, but are well approximated by sparse vectors. These are *compressible* vectors:

Argument 3. Compressible, incompressible vectors

- A vector in S^{n-1} is called *compressible* if it can be approximated by a sparse vector to within $\varepsilon \sim 0.001$ in the Euclidean norm.



- The other vectors are called *incompressible*. They have cn coefficients $|x_k| \sim \frac{1}{\sqrt{n}}$ (easy). Thus *incompressible* \approx *spread*.
- Now we have a genuine decomposition of S^{n-1} :



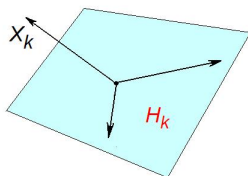
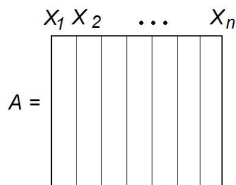
- The previous invertibility argument easily generalizes: sparse \rightarrow compressible; spread \rightarrow incompressible.
- **Conclusion.** For Gaussian matrices, $s_n(A) \gtrsim n^{-1/2}$ w.h.p. □

Argument 4. General random matrices

- **Question.** Where did we use that the matrix is *Gaussian*?
- **Answer.** In the *distance bound*, which we used for spread vectors:

Distance bound

The distance from a column X_k to the span H_k of the other columns is \sim constant w.h.p.



- This bound is easy for Gaussian distribution: the distance = |gaussian| \sim const w.h.p.
- This *argument fails* for non-Gaussian distributions (e.g. Bernoulli). But the *distance bound holds*. This is the main challenge.

Argument 5. Distance bound

Theorem (Distance bound)

Let X_1, \dots, X_n be random vectors in \mathbb{R}^n with i.i.d. subgaussian coordinates. Then the distance from X_n to the span of the other vectors H_n is \sim constant w.h.p.

- More precisely, for $\varepsilon > 0$,

$$\mathbb{P}(\text{dist}(X_n, H_n) \leq \varepsilon) \lesssim \varepsilon + c^n.$$

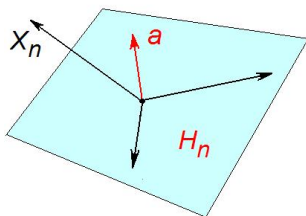
This bound is *optimal*.

- A weaker bound by Tao-Vu (2006): $\varepsilon \sim \frac{1}{n}$; $\mathbb{P}(\dots) \lesssim \frac{1}{\sqrt{\log n}}$.
- **Argument.** Usual tools *fail* (ε -nets, concentration of measure).

Argument 6. Littlewood-Offord Theory

- Develop a probabilistic tool related to **additive combinatorics**: “Inverse Littlewood-Offord theory” (proposed by Tao-Vu).
- Write the distance as a *sum of independent random variables*

$$S = \text{dist}(X_n, H_n) = \langle a, X_n \rangle = \sum a_k \xi_k$$



where $X_n = (\xi_1, \dots, \xi_n)$ and $a = (a_1, \dots, a_n)$ is the normal of H_n . (Condition on H_n , thus on a).

Argument 6. Littlewood-Offord Theory

- Deviation inequalities for sums of independent random variables

$$S = \sum a_k \xi_k$$

via embedding of the coefficient vector $a = (a_1, \dots, a_n)$ into an *arithmetic progression* (approximately).

- Let $D(a)$ = the shortest length of such arithmetic progression. The bigger $D(a)$, the less **additive structure** is in a , the better (smaller) should behave the random sum S . Main result:

$$\mathbb{P}(|S| < \varepsilon) \lesssim \varepsilon + \frac{1}{D(a)}.$$

- **Argument:** Ergodic approach.

Argument 7. Partition according to additive structure

- Previously, we decomposed according to the *sparsity*:

$$S^{n-1} = \text{Sparse} \cup \text{Spread}.$$

This is too coarse (on the spread side).

- Finer decomposition – according to the *additive structure*:

$$\text{Spread} = \bigcup_{D=1}^{\infty} S_D \quad S_D = \{a : D(a) \sim D\}.$$

- Use an ε -net *argument* for each part S_D .
- As D increases, the ε -nets become *bigger*, but the probability bounds become *better* (less additive structure). Tradeoff. □

Summarize:

Smallest singular value

Theorem (Main)

For an $n \times n$ matrix A with i.i.d. subgaussian entries, and for $\varepsilon > 0$,

$$\mathbb{P}(s_n(A) \leq \varepsilon n^{-1/2}) \leq C\varepsilon + c^n.$$

This bound is *optimal*. It implies:

- Von Neumann-Goldstine-Smale's conjecture:

$$s_n(A) \gtrsim n^{-1/2} \quad \text{with high probability.}$$

- In particular, **condition number** is linear in the dimension:

$$\kappa(A) = O(n) \quad \text{with high probability.}$$

- Spielman-Teng's conjecture (for general matrices).
- Kahn-Komlós-Szemerédi bound on the singularity probability

$$\mathbb{P}(A \text{ is singular}) \leq c^n.$$

(take $\varepsilon = 0$). Holds for general matrices, not only Bernoulli.