

# Small Ball Probability, Arithmetic Structure and Random Matrices

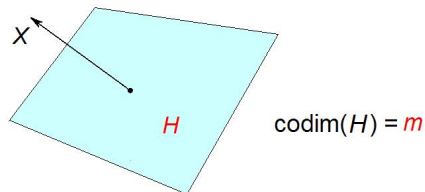
Roman Vershynin

University of California, Davis

April 23, 2008

# Distance Problems

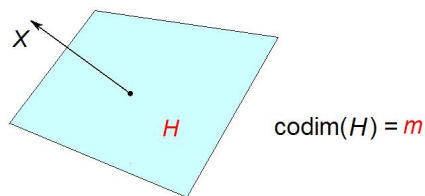
How far is a random vector  $X$  from a given subspace  $H$  in  $\mathbb{R}^N$ ?



- **Model:** coordinates of  $X$  are i.i.d. random variables.
- **Simplest distributions:** continuous, e.g. **Gaussian**,  $N(0, 1)$  entries.  
**Hardest distributions:** discrete, e.g. **Bernoulli**,  $\pm 1$  entries.

## Distance Problems

How far is a random vector  $X$  from a given subspace  $H$  in  $\mathbb{R}^N$ ?



- For a standard **Gaussian** random vector  $X$ ,

$$\text{dist}(X, H) = \|P_{H^\perp} X\|_2 = \|g\|_2,$$

where  $g$  is a standard Gaussian random vector in  $\mathbb{R}^m$ .

- The Euclidean norm of  $g$  is typically  $\sqrt{m}$ . Thus

$$\text{dist}(X, H) \sim \sqrt{m} \quad \text{with high probability.}$$

- Is this true for other distributions, e.g. Bernoulli?

## Distance Problems

- **Expectation.** It is an exercise to show that for a random vector  $X$  and a subspace  $H$  of codimension  $m$ , one has

$$\mathbb{E} \operatorname{dist}(X, H)^2 = m.$$

- **Large deviations.** Standard concentration technique gives:

$$\mathbb{P}(|\operatorname{dist}(X, H) - \sqrt{m}| > t) \lesssim e^{-ct^2}, \quad t > 0$$

see e.g. [TAO-VU'05] for Bernoulli; also true for general distributions.

- Large deviation inequality gives both *upper* and *lower* bounds on  $\operatorname{dist}(X, H)$ . However, *the lower bound is often too weak*.
- For a nontrivial lower bound, one needs  $t \leq \sqrt{m}$ . So, if  $m = O(1)$  large deviations give a useless lower bound.

**Problem.** For  $\varepsilon > 0$ , estimate the *small ball probability*

$$\mathbb{P}(\operatorname{dist}(X, H) \leq \varepsilon\sqrt{m}) \leq \dots$$

- Small ball probability is critical for singularity of random matrices.

## Distance Problems

**Problem.** For a random vector  $X$  and a subspace  $H$  of codimension  $m$ , estimate the *small ball probability*

$$\mathbb{P}(\text{dist}(X, H) \leq \varepsilon\sqrt{m}) \leq \dots$$

- Recall that for the simplest distribution, Gaussian,

$$\text{dist}(X, H) = \|g\|_2$$

where  $g$  is the standard Gaussian vector in  $\mathbb{R}^m$ .

- Since the Gaussian *density*  $\approx$  *constant* near the origin,

$$\mathbb{P}(\text{dist}(X, H) \leq \varepsilon\sqrt{m}) \sim \varepsilon^m.$$

- This is an ideal estimate. It is *independent of the subspace  $H$* .
- But this completely *fails for general distributions*, e.g. Bernoulli:

## Distance Problems

**Problem.** For a random vector  $X$  and a subspace  $H$  of codimension  $m$ , estimate the *small ball probability*

$$\mathbb{P}(\text{dist}(X, H) \leq \varepsilon\sqrt{m}) \leq \dots$$

- Consider *Bernoulli* distribution.
- For  $H = \{x : x_1 = x_2\}$ , we have  $\mathbb{P}(X \in H) = \frac{1}{2}$ .
- For  $H = \{x : x_1 + \dots + x_N = 0\}$ , we have  $\mathbb{P}(X \in H) = \frac{1}{\sqrt{N}}$ .
- The unfortunate fact is that these probabilities are *not exponentially small*, unlike for gaussian distribution.
- But may be these are bad examples of subspaces  $H$ . If so, can we *describe* all bad subspaces  $H$ ? And how can we *deal* with them?

# Distance Problems

**Problem.** For a random vector  $X$  and a subspace  $H$  of codimension  $m$ , estimate the *small ball probability*

$$\mathbb{P}(\text{dist}(X, H) \leq \varepsilon\sqrt{m}) \leq \dots$$

**Our goals:** (1) describe all obstacles to the small ball probabilities;  
(2) eliminate them.

- 1 **Description:** The only obstacle is the strong *additive structure* of  $H$ . This is a development of *Littlewood-Offord Theory*.
- 2 **Elimination:** *Random  $H$*  have no additive structure.

## Distance Problems

**Problem.** For a random vector  $X$  and a subspace  $H$  of codimension  $m$ , estimate the *small ball probability*

$$\mathbb{P}(\text{dist}(X, H) \leq \varepsilon\sqrt{m}) \leq \dots$$

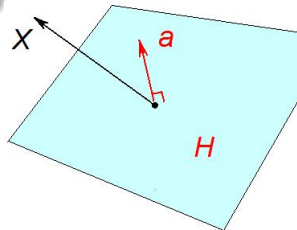
- Will illustrate our strategy for  $m = 1$  (so  $H$  is a hyperplane).

**Observation.** The distance  $\text{dist}(X, H)$  is a *sum of independent random variables*.

Indeed, consider the normal  $\mathbf{a}$  of  $H$ . Then

$$\text{dist}(X, H) = |\langle \mathbf{a}, X \rangle| = \left| \sum_{k=1}^N a_k \xi_k \right|$$

where  $\mathbf{a} = (a_1, \dots, a_N)$  and  $X = (\xi_1, \dots, \xi_N)$ .



## Small ball probability

- Study of the small ball probability for sums of independent random variables goes back to Kolmogorov, Lévy and Esseen (*local limit theorems*).
- For i.i.d. random variables  $\xi_k$ , we consider the sum

$$S = \sum_{k=1}^N a_k \xi_k, \quad a_k \in \mathbb{R} \text{ or } \mathbb{R}^m.$$

**Problem.** Estimate the small ball probability  $\mathbb{P}(|S| \leq \varepsilon)$ .

- More generally, estimate the *Lévy concentration function*

$$p_\varepsilon(S) := \sup_v \mathbb{P}(|S - v| \leq \varepsilon).$$

- *Large deviation inequalities* demonstrate that  $S$  concentrates nicely about its mean. *Small ball probability* goes in the opposite direction:  $S$  can not concentrate too much.

## Small ball probability

$$S = \sum_{k=1}^N a_k \xi_k, \quad a_k \in \mathbb{R}.$$

- **Examples** (for Bernoulli random variables  $\xi_k = \pm 1$ ):

For  $a = (1, 1, 0, \dots, 0)$ , we have  $\mathbb{P}(S = 0) = \frac{1}{2}$ .

For  $a = (1, 1, 1, \dots, 1)$ , we have  $\mathbb{P}(S = 0) \sim N^{-1/2}$ .

Generally, if  $|a_k| \geq 1$  then  $\mathbb{P}(|S| \leq 1) \lesssim N^{-1/2}$   
[LITTLEWOOD-OFFORD '43, ERDÖS '45].

For  $a = (1, 2, 3, \dots, N)$ , we have  $\mathbb{P}(S = 0) \sim N^{-3/2}$ .

Generally, if  $|a_j - a_k| \geq 1$  then  $\mathbb{P}(S = 0) \lesssim N^{-3/2}$   
[ERDÖS-MOSER '65, SÁRKÖZI-SZEMERÉDI '65, HÁLASZ '77].

- **Can we see a pattern?** For what  $a$  do we have  $\mathbb{P}(S = 0) \sim N^{-5/2}$ ?

# Littlewood-Offord Theory

$$S = \sum_{k=1}^N a_k \xi_k, \quad \mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}^N.$$

- Program put forth in [TAO-VU '06]:
- If the small ball probability is big, then the coefficient vector  $\mathbf{a}$  has strong *additive structure*. This allowed [TAO-VU '06] to go down to probabilities of an *arbitrary polynomial order*:  $\mathbb{P}(S = 0) \lesssim n^{-A}$ .
- New result [RUDELSON-V. '07]: can go down to probabilities of *arbitrary order* up to exponentially small,  $c^N$ .

# Littlewood-Offord Theory

$$S = \sum_{k=1}^N a_k \xi_k, \quad \mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}^N.$$

- How to quantify the *additive structure* in the coefficient vector  $\mathbf{a}$ ?
- By how commensurate the coefficients  $a_k$  are. If  $a_k$  are *rational* then we can measure this with the *least common denominator*

$$\text{LCD}(\mathbf{a}) = \inf\{t > 0 : t\mathbf{a} \in \mathbb{Z}^N\}.$$

- For general *real* coefficients, we allow some error  $\alpha > 0$ :

$$\text{LCD}_\alpha(\mathbf{a}) = \inf\{t > 0 : \text{dist}(t\mathbf{a}, \mathbb{Z}^N) \leq \alpha\}.$$

- LCD is “inversely proportional” to the amount of structure in  $\mathbf{a}$ .
- A related measure of structure: *length of the shortest arithmetic progression into which  $\mathbf{a}$  embeds* (exactly or approximately).
- LCD gives a stronger measure:  $\mathbf{a}$  clearly embeds into an arithmetic progression of length  $\text{LCD}(\mathbf{a}) \cdot \|\mathbf{a}\|_\infty$ .

## Littlewood-Offord Theory

$$S = \sum_{k=1}^N a_k \xi_k, \quad \mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}^N, \quad \|\mathbf{a}\|_2 = 1.$$

Small Ball Probability Thm [RUDELSON-V.'07], see [FRIEDLAND-SODIN'07]

$$\mathbb{P}(|S| \leq \varepsilon) \lesssim \varepsilon + \frac{1}{\text{LCD}_\alpha(\mathbf{a})} + e^{-\alpha^2}.$$

- Ignore for now the first and third terms in RHS. Theorem states that the **the small ball probability is inversely proportional to LCD**.
- So, the only obstacle to small ball probability is *structure*. If the small ball probability is large then  $\text{LCD}(\mathbf{a})$  is small, thus  $\mathbf{a}$  has strong additive structure (and in particular embeds into a short arithmetic progression).

# Littlewood-Offord Theory

$$S = \sum_{k=1}^N a_k \xi_k, \quad \mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}^N, \quad \|\mathbf{a}\|_2 = 1.$$

Small Ball Probability Thm [RUDELSON-V.'07], see [FRIEDLAND-SODIN'07]

$$\mathbb{P}(|S| \leq \varepsilon) \lesssim \varepsilon + \frac{1}{\text{LCD}_\alpha(\mathbf{a})} + e^{-\alpha^2}.$$

- The *first term* in RHS gives an “ideal” estimate, valid for continuous distributions. For example, for Gaussian  $\xi_k$ , we have  $\mathbb{P}(|S| \leq \varepsilon) \sim \varepsilon$ .
- The **third term** in RHS is typically *exponentially small* in  $N$ . Recall that in the definition of LCD, we require  $\text{dist}(t\mathbf{a}, \mathbb{Z}^N) \leq \alpha$ . So we would normally choose  $\alpha \sim 0.01\sqrt{N}$ ; this makes most coordinates of  $\mathbf{a}$  almost integers.

## Littlewood-Offord Theory: Examples

$$S = \sum_{k=1}^N a_k \xi_k, \quad \mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}^N, \quad \|\mathbf{a}\|_2 = 1.$$

Small Ball Probability Thm [RUDELSON-V'07], see [FRIEDLAND-SODIN'07]

$$\mathbb{P}(|S| \leq \varepsilon) \lesssim \varepsilon + \frac{1}{\text{LCD}_\alpha(\mathbf{a})} + e^{-\alpha^2}.$$

- For  $\mathbf{a} = (1, 1, 1, \dots, 1)$ , we have  $\|\mathbf{a}\|_2 = \sqrt{N}$ , so  $\text{LCD}(\frac{\mathbf{a}}{\|\mathbf{a}\|_2}) = \sqrt{N}$ . Then the Small Ball Probability Theorem gives

$$\mathbb{P}(S = 0) \lesssim N^{-1/2}.$$

(Recall [LITTLEWOOD-OFFORD '43, ERDÖS '45].)

- For  $\mathbf{a} = (1, 2, 3, \dots, N)$ ,  $\|\mathbf{a}\|_2 \sim N^{3/2}$ , so by a similar argument,

$$\mathbb{P}(S = 0) \lesssim N^{-3/2}.$$

(Recall [ERDÖS-MOSER '65, SÁRKÖZI-SZEMERÉDI '65, HÁLASZ '77].)

- For  $\mathbf{a} = (1^2, 2^2, 3^2, \dots, N^2)$ , we similarly get  $\mathbb{P}(S = 0) \lesssim N^{-5/2}$ .

## Littlewood-Offord Theory: Extensions

$$S = \sum_{k=1}^N a_k \xi_k, \quad \mathbf{a} = (\mathbf{a}_1, \dots, \mathbf{a}_N) \in \mathbb{R}^N, \quad \|\mathbf{a}\|_2 = 1.$$

Small Ball Probability Thm [RUDELSON-V.'07], see [FRIEDLAND-SODIN'07]

$$\mathbb{P}(|S| \leq \varepsilon) \lesssim \varepsilon + \frac{1}{\text{LCD}_\alpha(\mathbf{a})} + e^{-\alpha^2}.$$

- **Lévy concentration function:** the same bound for any small ball probability  $\mathbb{P}(|S - v| \leq \varepsilon)$  over  $v \in \mathbb{R}$ .
- **Higher dimensions:** Suppose  $\mathbf{a}_k \in \mathbb{R}^m$  are *genuinely*  $m$ -dimensional vectors (they are not too close to any lower-dimensional subspace). Then the small ball probability bound becomes *genuinely*  $m$ -dimensional:

$$\mathbb{P}(\|S\|_2 \leq \varepsilon\sqrt{m}) \lesssim \left( \varepsilon + \frac{\sqrt{m}}{\text{LCD}_\alpha(\mathbf{a})} \right)^m + e^{-\alpha^2}.$$

This strengthens the result of [HÁLASZ '77].

## Additive Structure of Random Vectors

$$\mathbf{S} = \sum_{k=1}^N \mathbf{a}_k \xi_k, \quad \mathbf{a} = (\mathbf{a}_1, \dots, \mathbf{a}_N) \in \mathbb{R}^N, \quad \|\mathbf{a}\|_2 = 1.$$

Small Ball Probability Thm [RUDELSON-V.'07], see [FRIEDLAND-SODIN'07]

$$\mathbb{P}(|\mathbf{S}| \leq \varepsilon) \lesssim \varepsilon + \frac{1}{\text{LCD}_\alpha(\mathbf{a})} + e^{-\alpha^2}.$$

- In order to successfully use this theorem, we need to know what coefficient vectors  $\mathbf{a}$  have *large LCD*. Specifically, we want LCD exponentially large in  $N$ . (This is largest possible).
- In other words, **what coefficient vectors have no additive structure?**
- Plausible answer: *random vectors*.
- Not always: Bernoulli random vectors ( $\pm 1$ ) have huge additive structure. So the randomness has to be used carefully.
- Correct answer: **normals to random hyperplanes**.

# Additive Structure of Random Vectors

## No Structure Theorem [RUDELSON-V. '07]

Let  $H$  be a hyperplane in  $\mathbb{R}^N$  spanned by  $N - 1$  independent vectors with i.i.d. random coefficients. Then the normal  $\mathbf{a}$  of  $H$  has no structure:

$$\text{LCD}(\mathbf{a}) \geq e^{cN}$$

with probability at least  $1 - e^{-cN}$ .

- **Nontrivial:** the coefficients of  $\mathbf{a}$  are not independent.

## No Structure in Higher Dimensions [RUDELSON-V. '08]

Let  $H$  be a hyperplane in  $\mathbb{R}^N$  spanned by  $N - m$  independent vectors with i.i.d. random coefficients. Then the the subspace  $H^\perp$  has no structure:

$$\text{LCD}(H^\perp) = \inf_{\mathbf{a} \in S(H^\perp)} \text{LCD}(\mathbf{a}) \geq \sqrt{N} e^{cN/m}$$

with probability at least  $1 - e^{-cN}$ .

# Applications to Distance Problems

Now we can answer the distance question asked in the beginning:

How far is a random vector  $X$  from a given subspace  $H$  in  $\mathbb{R}^N$ ?

- **Answer:** Expect the distance  $\sim \sqrt{\text{codim}(H)}$ . However, the probability depends on the amount of *additive structure* in  $H$ :

## Distance Theorem [RUDELSON-V. '08]

1. **Arbitrary subspaces.** Let  $H$  be a subspace of codimension  $m$ . Then, for  $\varepsilon \geq \sqrt{m}/\text{LCD}(H^\perp)$ ,

$$\mathbb{P}(\text{dist}(X, H) \leq \varepsilon\sqrt{m}) \lesssim \varepsilon^m + c^N.$$

2. **Random subspaces.**  $H$  is spanned by  $N - m$  independent random vectors. Then this bound holds for *arbitrary*  $\varepsilon > 0$ .

**Proof.** The first part is a direct application of Small Ball Probability Theorem in higher dimensions. For the second part, the No Structure Theorem says that LCD of random  $H^\perp$  is exponentially large.  $\square$

# Applications to Distance Problems

## Distance Theorem [RUDELSON-V. '08]

1. **Arbitrary subspaces.** Let  $H$  be a subspace of codimension  $m$ . Then, for  $\varepsilon \geq \sqrt{m}/\text{LCD}(H^\perp)$ ,

$$\mathbb{P}(\text{dist}(X, H) \leq \varepsilon\sqrt{m}) \lesssim \varepsilon^m + c^N.$$

2. **Random subspaces.**  $H$  is spanned by  $N - m$  independent random vectors. Then this bound holds for *arbitrary*  $\varepsilon > 0$ .

- Ignoring the term  $c^N$ , this matches our (simple) estimate for Gaussian vectors.
- The term  $c^N$  is necessary: for Bernoulli vectors,  $\mathbb{P}(X \in H) \geq 2^{-N}$ .
- The bound also holds for *affine* subspaces, i.e. for  $H + v$  for arbitrary translate  $v \in \mathbb{R}^N$ .
- Previous best bound by [TAO-VU'06]:

$$\mathbb{P}(\text{dist}(X, H) \leq \frac{1}{4N}) \lesssim \frac{1}{\sqrt{\log N}}.$$

## Invertibility of Random Matrices

- The distance bounds are essential to study the *invertibility of random matrices*. How likely is a random matrix to be invertible? What is the norm of the inverse?
- It is convenient to look at the *singular values*  $s_k(A)$ , which are the eigenvalues of  $|A| = \sqrt{A^*A}$  in the non-increasing order.
- Of particular importance are *the extreme* singular values

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

- We often want to bound the largest singular value  $s_1(A)$  above, the least singular value  $s_n(A)$  below. This would mean that  $A$  is a nice *isomorphic embedding*; it would not distort the norms of vectors too much.

# Invertibility of Random Matrices

**Problem.** Estimate the extreme sing. values of a random  $N \times n$  matrix.

**Model:** entries of a matrix  $A$  are random i.i.d. centered subgaussian random variables. (Weaker assumptions often suffice).

**Theorem (Asymptotics)** [YIN-BAI-KRISHNAIAH'88, SILVERSTEIN'85], [BAI'93]

Let  $N \rightarrow \infty$  and  $n/N \rightarrow \text{const}$ . Then

$$s_1(A) \rightarrow \sqrt{N} + \sqrt{n}, \quad s_n(A) \rightarrow \sqrt{N} - \sqrt{n} \quad \text{almost surely.}$$

( $a_n \rightarrow b_n$  means that  $a_n/b_n \rightarrow 1$ ).

**Shortcomings:**

- 1 No information about arbitrary *finite dimensions*  $N, n$ ;
- 2 No information about the least singular value  $s_n(A)$  when  $N/n \rightarrow 1$ , i.e. for *almost square or exactly square matrices*.

# Invertibility of Random Matrices

**Problem.** Estimate the extreme sing. values of a random  $N \times n$  matrix.

Bounds in finite dimensions?

- The largest singular value  $s_1(A)$  is well studied. A simple  $\varepsilon$ -net argument gives:

$$s_1(A) \sim \sqrt{N} \quad \text{with high probability.}$$

This is consistent with the asymptotic result  $s_1(A) \sim \sqrt{N} + \sqrt{n}$ .

- The probability is in fact exponentially large:

$$\mathbb{P}(s_1(A) \geq t\sqrt{N}) \leq e^{-ct^2N} \quad \text{for } t \geq C.$$

- The least singular value  $s_n(A)$  is harder to estimate. Generally, it is *less robust* than  $s_1(A)$ , more sensitive to perturbations.

## Invertibility of Random Matrices

We will now prove that the limiting result for the least singular value continues to hold in arbitrary fixed dimensions:

$$s_n(A) \gtrsim \sqrt{N} - \sqrt{n} \quad \text{with high probability.}$$

### Theorem (Least Singular Value) [RUDELSON-V.'08]

Let  $A$  be an  $N \times n$  matrix with i.i.d. centered subgaussian random entries of unit variance. Then, for every  $\varepsilon > 0$ ,

$$\mathbb{P}\left[s_n(A) \leq \varepsilon(\sqrt{N} - \sqrt{n-1})\right] \leq \varepsilon^{N-n+1} + e^{-cN}.$$

- This bound is *optimal in the whole range* of dimensions  $N \geq n$ .
- For an arbitrary *constant aspect ratio*  $n/N = \lambda > 1$ , this theorem recovers the result of [LITVAK-PAJOR-RUDELSON-TOMCZAK '05]:

$$\mathbb{P}(s_n(A) \leq c_\lambda \sqrt{N}) \leq c^N.$$

# Invertibility of Random Matrices

## Theorem (Least Singular Value) [RUDELSON-V.'08]

Let  $A$  be an  $N \times n$  matrix with i.i.d. centered subgaussian random entries of unit variance. Then, for every  $\varepsilon > 0$ ,

$$\mathbb{P}\left[s_n(A) \leq \varepsilon(\sqrt{N} - \sqrt{n-1})\right] \lesssim \varepsilon^{N-n+1} + c^N.$$

- A remarkable partial case is for *square matrices* ( $N = n$ ). Here  $\sqrt{N} - \sqrt{N-1} \sim 1/\sqrt{N}$ . So the theorem gives

$$s_N(A) \gtrsim \frac{1}{\sqrt{N}} \quad \text{with high probability.}$$

This was a prediction of [VON NEUMANN ET AL.'47, SMALE'85], verified previously only for Gaussian matrices [EDELMAN'88, SZAREK'90].

- More precisely, for square matrices the theorem reads as

$$\mathbb{P}(s_N(A) \leq \varepsilon/\sqrt{N}) \lesssim \varepsilon + c^N.$$

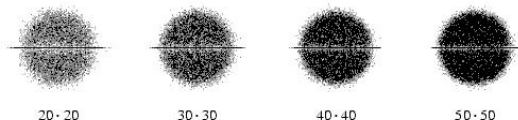
This was conjectured for Bernoulli matrices by [SPIELMAN-TENG'02].

- In particular,  $\mathbb{P}(s_N(A) = 0) \leq c^N$ . [KAHN-KOMLÓS-SZEMERÉDI'95]

# Application: Circular Law

## Circular Law

Let  $A$  be an  $n \times n$  random matrix with real independent entries. Then the joint distribution of the eigenvalues of  $\frac{1}{\sqrt{n}}A$  converges as  $n \rightarrow \infty$  to the uniform distribution on the unit disc in  $\mathbb{C}$ .



## History:

- Suggested by [PASTUR '73]. Claimed by [GIRKO '84].
- [EDELMAN '97] for Gaussian matrices
- [BAI '97] assuming density, 6-th moments of the entries
- [GÖTZE-TIKHOMIROV '07], [TAO-VU '07]: general matrices with  $2 + o(1)$  moments. **Deduced from the Least Singular Value Theorem.**

## Appendix 1: Proof of Small Ball Probability Theorem

Recall that we consider a sum of independent random variables

$$\mathbf{S} = \sum_{k=1}^N \mathbf{a}_k \xi_k, \quad \mathbf{a} = (\mathbf{a}_1, \dots, \mathbf{a}_N) \in \mathbb{R}^N, \quad \|\mathbf{a}\|_2 = 1.$$

We want to bound the small ball probability of  $\mathbf{S}$  in terms of the least common denominator of the coefficient vector  $\mathbf{a}$ , defined as

$$\text{LCD}_\alpha(\mathbf{a}) = \inf\{t > 0 : \text{dist}(t\mathbf{a}, \mathbb{Z}^N) \leq \alpha\}.$$

Small Ball Probability Thm [RUDELSON-V.'07], see [FRIEDLAND-SODIN'07]

$$\mathbb{P}(|\mathbf{S}| \leq \varepsilon) \lesssim \varepsilon + \frac{1}{\text{LCD}_\alpha(\mathbf{a})} + e^{-\alpha^2}.$$

## Appendix 1: Proof of Small Ball Probability Theorem

$$S = \sum_{k=1}^N a_k \xi_k, \quad \mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}^N, \quad \|\mathbf{a}\|_2 = 1.$$

- We will sketch the proof of the Small Ball Probability Theorem. There are two approaches.
- **Soft approach** [LITVAK-PAJOR-RUDELSON-TOMCZAK '05]: approximate the distribution of  $S$  by the standard Gaussian distribution, using the *Central Limit Theorem*.
- Berry-Esseen CLT:

$$\mathbb{P}(|S| \leq \varepsilon) \approx \mathbb{P}(|g| \leq \varepsilon) + N^{-1/2} \sim \varepsilon + N^{-1/2}.$$

- The error term  $N^{-1/2}$  that comes from CLT is *too big*. Finer approach is inspired by *ergodic theory*:

## Appendix 1: Proof of Small Ball Probability Theorem

We start with *Esseen inequality*, the method going back to [HALASZ '77]:

### Esseen Inequality

The small ball probability of a random variable  $S$  is bounded by the  $L^1$  norm of the characteristic function  $\phi(t) = \mathbb{E} \exp(iSt)$ :

$$\mathbb{P}(|S| \leq \varepsilon) \lesssim \int_{-1}^1 |\phi(t/\varepsilon)| dt.$$

- **Proof:** take Fourier transform.
- We shall use Esseen Inequality for the random sum  $S = \sum_1^N a_k \xi_k$ .
- Let us illustrate the argument on the example of Bernoulli  $\xi_k = \pm 1$ .  
By independence, the characteristic function of  $S$  factors as

$$\phi(t) = \prod_1^N \phi_k(t), \quad \phi_k(t) = \mathbb{E} \exp(ia_k \xi_k t) = \cos(a_k t).$$

## Appendix 1: Proof of Small Ball Probability Theorem

- Then

$$|\phi(t)| = \prod_1^N |\cos(a_k t)| \leq \exp(-f(t)),$$

where

$$f(t) = \sum_1^N \sin^2(a_k t).$$

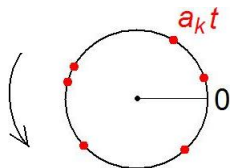
- Esséen's Inequality then yields

$$\begin{aligned} \mathbb{P}(|S| \leq \varepsilon) &\lesssim \int_{-1}^1 |\phi(t/\varepsilon)| dt \leq \int_{-1}^1 \exp(-f(t/\varepsilon)) dt \\ &\sim \varepsilon \int_{-1/\varepsilon}^{1/\varepsilon} \exp(-f(t)) dt. \end{aligned}$$

## Appendix 1: Proof of Small Ball Probability Theorem

$$\mathbb{P}(|S| \leq \varepsilon) \lesssim \varepsilon \int_{-1/\varepsilon}^{1/\varepsilon} \exp(-f(t)) dt, \quad \text{where } f(t) = \sum_1^N \sin^2(a_k t).$$

**Ergodic approach:** regard  $t$  as *time*,  $\varepsilon \int_{-1/\varepsilon}^{1/\varepsilon}$  as long term average. The system of  $n$  particles  $a_k t$  moves along  $\mathbb{T}$  with speeds  $a_k$ .



- We want that  $f(t)$  be large most of the time.
- How can this be *not* true?  $f(t)$  is the sum of  $\sin^2(ta_k)$ , so...
- If  $f(t)$  is small then  $\text{dist}(ta, \pi\mathbb{Z})$  is small (i.e. the particles return to the origin together).
- But if  $\text{dist}(ta, \pi\mathbb{Z})$  *frequently* becomes small, then **LCD(a) is small** (i.e. there is a lot of additive structure in  $a$ ).

This proves the Small Ball Probability Theorem. □

## Appendix 3: Proof of Least Singular Value Theorem

Will sketch the proof on the example of  $N \times N$  matrices. Want to show:

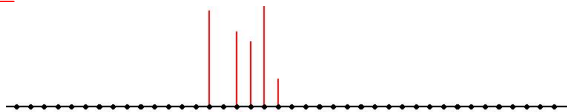
$$s_N(A) = \inf_{x \in S^{N-1}} \|Ax\| \gtrsim N^{-1/2} \quad \text{with high probability.}$$

- How to bound  $\|Ax\|$  below for all  $x \in S^{N-1}$ ? We shall do this separately for two extreme types of vectors  $x$ : **sparse** and **spread**.

We start with *sparse* vectors...

## Appendix 3: Proof of Least Singular Value Theorem

- A vector  $x \in \mathcal{S}^{N-1}$  is *sparse* if it has few nonzero coordinates:  
 $|\text{supp}(x)| \leq \delta N$  ( $\delta \sim 0.01$ )



- Want to bound  $\|Ax\|$  below for all sparse vectors  $x$ .

**Observation.** Since only  $\delta N$  coordinates of  $x$  are nonzero,  $A$  is effectively acting as a  $N \times \delta N$  *rectangular matrix* (restriction of  $A$  onto the support of  $x$ ).

- But the smallest singular value of tall matrices is well known (and simple to compute): it is  $\sim \sqrt{N}$ . Taking the union bound over all supports gives

$$\inf_{x \in \text{Sparse}} \|Ax\| \gtrsim \sqrt{N}.$$

This is much better than we need!

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

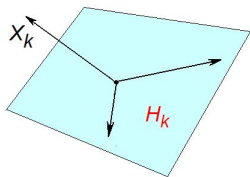
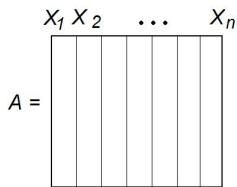
## Appendix 3: Proof of Least Singular Value Theorem

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



- **Difficulty.** Spread vectors contain *more information* than sparse. For example, they do not have a small  $\varepsilon$ -net.
- We develop a completely different *geometric argument* to prove a lower bound for  $\|Ax\|$  for spread vectors.

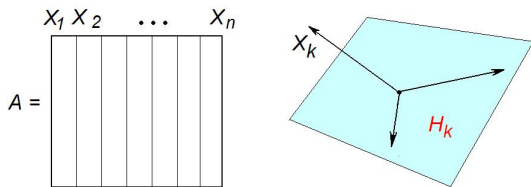
## Appendix 3: Proof of Least Singular Value Theorem



- The trivial rank argument:  $A$  is nonsingular  $\Leftrightarrow$  each column  $X_k$  does not lie in the span  $H_k$  of the others.
- A quantitative version should be:

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

## Appendix 3: Proof of Least Singular Value Theorem



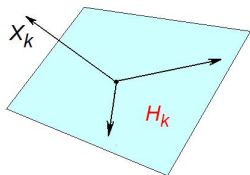
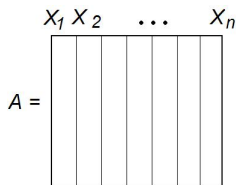
Here is such a geometric argument for spread  $x$ :

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

The right hand side does not depend on  $x$ ! Thus

$$\inf_{x \in \text{Spread}} \|Ax\| \gtrsim \frac{1}{\sqrt{N}} \text{dist}(X_k, H_k).$$

## Appendix 3: Proof of Least Singular Value Theorem



$$\inf_{x \in \text{Spread}} \|Ax\| \gtrsim \frac{1}{\sqrt{N}} \text{dist}(X_k, H_k).$$

Reall that the distance  $\text{dist}(X_k, H_k) \sim \text{const}$  with high probability:

$$\mathbb{P}(\text{dist}(X_k, H) < \varepsilon) \lesssim \varepsilon + c^N. \quad (\text{Distance Thm})$$

This gives

$$\inf_{x \in \text{Spread}} \|Ax\| \gtrsim \frac{1}{\sqrt{N}}.$$

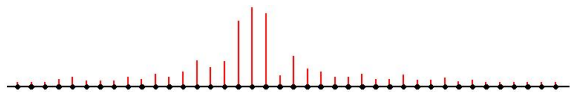
This settles the invertibility for spread vectors, So we now have...

## Appendix 3: Proof of Least Singular Value Theorem

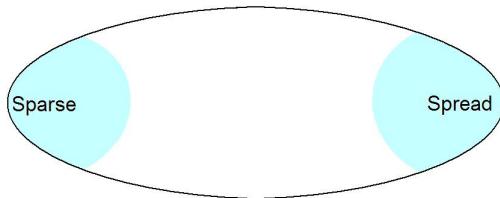
**Conclusion.**  $A$  is nicely invertible on both sparse and spread vectors:

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

- Unfortunately, most vectors are *neither sparse nor spread*:



- Question:** how to fill the gap – how to truly decompose  $S^{n-1}$ ?

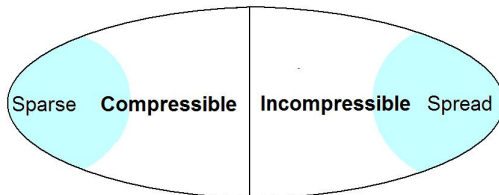


## Appendix 3: Proof of Least Singular Value Theorem

**Conclusion.**  $A$  is nicely invertible on both sparse and spread vectors:

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

- **Question:** how to fill the gap – how to truly decompose  $S^{n-1}$ ?



- **Answer:** according to the distance. If a vector  $x$  is within  $\varepsilon = 0.01$  to sparse vectors, treat it as a sparse vector (by approximation). Otherwise, treat it as a spread vector ( $x$  will have a *lot* of coordinates of order  $1/\sqrt{N}$ ).
- This leads to completion of the proof for square matrices. □

## Recap: Invertibility of Random Matrices

A version of the limiting result for the least singular value continues to hold in arbitrary fixed dimensions:

$$s_n(A) \gtrsim \sqrt{N} - \sqrt{n} \quad \text{with high probability.}$$

### Theorem (Least Singular Value) [RUDELSON-V.'08]

Let  $A$  be an  $N \times n$  matrix with i.i.d. centered subgaussian random entries of unit variance. Then, for every  $\varepsilon > 0$ ,

$$\mathbb{P}\left[s_n(A) \leq \varepsilon(\sqrt{N} - \sqrt{n-1})\right] \leq \varepsilon^{N-n+1} + e^{-cN}.$$