Singularity of random Bernoulli matrices

Konstantin Tikhomirov
Georgia Institute of Technology

July 3rd, 2019

Problem

Let $X_1, X_2, \ldots, X_n$ be independent random vectors uniformly distributed on vertices of the n–dimensional cube $[-1, 1]^n$. What is the probability that the vectors are linearly dependent?

Problem

Let $X_1, X_2, \ldots, X_n$ be independent random vectors uniformly distributed on vertices of the n–dimensional cube $[-1, 1]^n$. What is the probability that the vectors are linearly dependent?

The problem can be restated in terms of random matrices. Let B_n be an $n \times n$ random matrix with i.i.d ± 1 entries. What is the probability that the matrix is singular:

$$\mathbb{P}\{s_{\min}(B_n) = 0\} = ?$$

$s_{\min}(B_n) = \inf_{x \in S^{n-1}} \|B_n x\|_2$ — smallest singular value of B (i.e. smallest eigenvalue of positive semidefinite matrix $(B_n B_n^\top)^{1/2}$)

Motivation from numerical analysis

Let A be an $n \times n$ matrix, $s_{\max}(A) = \|A\| = \sup_{x \in S^{n-1}} \|Ax\|_2$ — the largest singular value of A. The condition number

$$\kappa(A) = \frac{s_{\max}(A)}{s_{\min}(A)}.$$

The condition number serves as measure of loss of precision when solving systems of linear equations.

Motivation from numerical analysis

Let A be an $n \times n$ matrix, $s_{\max}(A) = \|A\| = \sup_{x \in S^{n-1}} \|Ax\|_2$ — the largest singular value of A. The condition number

$$\kappa(A) = \frac{s_{\max}(A)}{s_{\min}(A)}.$$

The condition number serves as measure of loss of precision when solving systems of linear equations. Assume we look for solution of a system

$$Ax = b,$$

but the coefficient vector b is given with an error δb. Thus, we are solving the system

$$Ay = b + \delta b, \text{ where } y = x + \delta x.$$

Motivation from numerical analysis

Let A be an $n \times n$ matrix, $s_{\max}(A) = \|A\| = \sup_{x \in S^{n-1}} \|Ax\|_2$ — the largest singular value of A. The condition number

$$\kappa(A) = \frac{s_{\max}(A)}{s_{\min}(A)}.$$

The condition number serves as measure of loss of precision when solving systems of linear equations. Assume we look for solution of a system

$$Ax = b,$$

but the coefficient vector b is given with an error δb. Thus, we are solving the system

$$Ay = b + \delta b, \text{ where } y = x + \delta x.$$

We clearly have

$$\frac{\|\delta x\|_2}{\|x\|_2} = \frac{\|A^{-1}\delta b\|_2}{\|A^{-1}b\|_2} = \frac{\|A^{-1}\delta b\|_2}{\|\delta b\|_2} \frac{\|\delta b\|_2}{\|b\|_2} \frac{\|b\|_2}{\|A^{-1}b\|_2} \leq \kappa(A) \frac{\|\delta b\|_2}{\|b\|_2}.$$

Motivation from numerical analysis

Let A be an $n \times n$ matrix, $s_{\max}(A) = \|A\| = \sup_{x \in S^{n-1}} \|Ax\|_2$ — the largest singular value of A. The condition number

$$\kappa(A) = \frac{s_{\max}(A)}{s_{\min}(A)}.$$

The condition number serves as measure of loss of precision when solving systems of linear equations. Assume we look for solution of a system

$$Ax = b,$$

but the coefficient vector b is given with an error δb. Thus, we are solving the system

$$Ay = b + \delta b, \text{ where } y = x + \delta x.$$

We clearly have

$$\frac{\|\delta x\|_2}{\|x\|_2} = \frac{\|A^{-1}\delta b\|_2}{\|A^{-1}b\|_2} = \frac{\|A^{-1}\delta b\|_2}{\|\delta b\|_2} \frac{\|\delta b\|_2}{\|b\|_2} \frac{\|b\|_2}{\|A^{-1}b\|_2} \leq \kappa(A) \frac{\|\delta b\|_2}{\|b\|_2}.$$

A typical coefficient matrix can be modeled as a random matrix with some distribution (determined by the nature of the specific problem). Then estimating $\kappa(A)$, hence $s_{\min}(A)$, becomes important.

History

In the 1940-es–1950-es, the condition number of random matrices was studied by von Neumann and Goldstine using numerical simulations. In particular, they conjectured that, for an $n \times n$ random matrix G_n with i.i.d standard normal entries, the condition number $\kappa(G_n) = O(n)$ with probability close to one.

History

In the 1940-es–1950-es, the condition number of random matrices was studied by von Neumann and Goldstine using numerical simulations. In particular, they conjectured that, for an $n \times n$ random matrix G_n with i.i.d standard normal entries, the condition number $\kappa(G_n) = O(n)$ with probability close to one.

The limiting distribution of the condition number, and the smallest singular value, of Gaussian random matrices was only computed on 1980-es by Edelman, using a formula for the joint distribution of its singular values. Edelman proved that

$$\mathbb{P}\{s_{\min}(G_n) \leq t n^{-1/2}\} = 1 - \exp\left(-t^2/2 - t\right) + o(1), \qquad t > 0.$$

So, typically $s_{\min}(G_n)$ is of order $n^{-1/2}$. There are various arguments which show that $s_{\max}(G_n) = (2 + o(1))\sqrt{n}$ with high probability.

History

In the 1940-es–1950-es, the condition number of random matrices was studied by von Neumann and Goldstine using numerical simulations. In particular, they conjectured that, for an $n \times n$ random matrix G_n with i.i.d standard normal entries, the condition number $\kappa(G_n) = O(n)$ with probability close to one.

The limiting distribution of the condition number, and the smallest singular value, of Gaussian random matrices was only computed on 1980-es by Edelman, using a formula for the joint distribution of its singular values. Edelman proved that

$$\mathbb{P}\{s_{\min}(G_n) \leq tn^{-1/2}\} = 1 - \exp\left(-t^2/2 - t\right) + o(1), \qquad t > 0.$$

So, typically $s_{\min}(G_n)$ is of order $n^{-1/2}$. There are various arguments which show that $s_{\max}(G_n) = (2 + o(1))\sqrt{n}$ with high probability.

Corresponding results for non-Gaussian random matrices were obtained much later. Even the problem of showing $\mathbb{P}\{A \text{ is singular}\} = o_n(1)$ for a discrete random matrix A with i.i.d entries is not trivial.

History (continued)

In 1960-es, Komlós showed that for $n \times n$ random matrix B_n with i.i.d ± 1 entries,

$$\mathbb{P}\{B_n \text{ is singular}\} = o(1).$$

History (continued)

In 1960-es, Komlós showed that for $n \times n$ random matrix B_n with i.i.d ± 1 entries,

$$\mathbb{P}\{B_n \text{ is singular}\} = o(1).$$

The estimate was greatly improved about 30 years later by Kahn, Komlós and Szemerédi (1995), who showed that

$$\mathbb{P}\{B_n \text{ is singular}\} \leq 0.999^n,$$

i.e. the singularity probability is exponentially small in dimension.

History (continued)

In 1960-es, Komlós showed that for $n \times n$ random matrix B_n with i.i.d ± 1 entries,

$$\mathbb{P}\{B_n \text{ is singular}\} = o(1).$$

The estimate was greatly improved about 30 years later by Kahn, Komlós and Szemerédi (1995), who showed that

$$\mathbb{P}\{B_n \text{ is singular}\} \leq 0.999^n,$$

i.e. the singularity probability is exponentially small in dimension.

The trivial bound

$$\mathbb{P}\{B_n \text{ is singular}\} \geq \mathbb{P}\{\text{Two rows/columns of } B_n \text{ are equal up to a sign}\}$$

implies that

$$\mathbb{P}\{B_n \text{ is singular}\} \geq (1 - o(1))n^2 2^{1-n}.$$

It is natural to expect that equality of two rows or columns is the main contribution to singularity which leads to

History (continued)

Strong conjecture

$$\mathbb{P}\{B_n \text{ is singular}\} = (1 + o(1))n^2 2^{1-n}.$$

Weak conjecture

$$\mathbb{P}\{B_n \text{ is singular}\} = \left(\frac{1}{2} + o(1)\right)^n.$$

Both conjectures are folklore and have been restated many times in the literature. On can distinguish two existing approaches to these problems.

History (continued)

Strong conjecture

$$\mathbb{P}\{B_n \text{ is singular}\} = (1 + o(1))n^2 2^{1-n}.$$

Weak conjecture

$$\mathbb{P}\{B_n \text{ is singular}\} = \left(\frac{1}{2} + o(1)\right)^n.$$

Both conjectures are folklore and have been restated many times in the literature. On can distinguish two existing approaches to these problems.

The first one is a development of the argument of Kahn–Komlós–Szemerédi, and is based on the notion of *combinatorial dimension* and on replacing the original distribution of the entries with a "lazy" one. This approach is designed to work for discrete distributions.

History (continued)

Strong conjecture

$$\mathbb{P}\{B_n \text{ is singular}\} = (1 + o(1))n^2 2^{1-n}.$$

Weak conjecture

$$\mathbb{P}\{B_n \text{ is singular}\} = \left(\frac{1}{2} + o(1)\right)^n.$$

Both conjectures are folklore and have been restated many times in the literature. On can distinguish two existing approaches to these problems.

The first one is a development of the argument of Kahn–Komlós–Szemerédi, and is based on the notion of *combinatorial dimension* and on replacing the original distribution of the entries with a "lazy" one. This approach is designed to work for discrete distributions.

The second is based on decomposition of the Euclidean sphere and special *covering arguments* and gives small ball probability estimates for the smallest singular value for much broader class of random matrices.

History: argument of Kahn–Komlós–Szemerédi

The argument described below is a development of the original Kahn–Komlós–Szemerédi due to Tao and Vu.

Let B_n be the $n \times n$ Bernoulli (± 1) random matrix. The argument of Kahn, Komlós and Szemerédi of proving $\mathbb{P}\{B_n \text{ is singular}\} \leq 0.999^n$ starts by writing

$$\mathbb{P}\{B_n \text{ is singular}\} \leq 2^{o(n)} \mathbb{P}\{X_1, X_2, \ldots, X_n \text{ span a hyperplane}\},$$

where $X_1, X_2, \ldots, X_n$ are columns of B_n.

History: argument of Kahn–Komlós–Szemerédi

The argument described below is a development of the original Kahn–Komlós–Szemerédi due to Tao and Vu.

Let B_n be the $n \times n$ Bernoulli (± 1) random matrix. The argument of Kahn, Komlós and Szemerédi of proving $\mathbb{P}\{B_n \text{ is singular}\} \leq 0.999^n$ starts by writing

$$\mathbb{P}\{B_n \text{ is singular}\} \leq 2^{o(n)} \mathbb{P}\{X_1, X_2, \ldots, X_n \text{ span a hyperplane}\},$$

where $X_1, X_2, \ldots, X_n$ are columns of B_n. For any $d \in \frac{1}{n}\mathbb{N}$, let Ω_d be the set of all hyperplanes V such that

$$2^{-\varepsilon d/n - \varepsilon/n^2} \leq \mathbb{P}\{X_1 \in V\} \leq 2^{-\varepsilon d/n} \text{ (} V \text{ is of combinatorial dimension } d\text{)}.$$

History: argument of Kahn–Komlós–Szemerédi

The argument described below is a development of the original Kahn–Komlós–Szemerédi due to Tao and Vu.

Let B_n be the $n \times n$ Bernoulli (± 1) random matrix. The argument of Kahn, Komlós and Szemerédi of proving $\mathbb{P}\{B_n \text{ is singular}\} \leq 0.999^n$ starts by writing

$$\mathbb{P}\{B_n \text{ is singular}\} \leq 2^{o(n)} \mathbb{P}\{X_1, X_2, \ldots, X_n \text{ span a hyperplane}\},$$

where $X_1, X_2, \ldots, X_n$ are columns of B_n. For any $d \in \frac{1}{n}\mathbb{N}$, let Ω_d be the set of all hyperplanes V such that

$$2^{-\varepsilon d/n - \varepsilon/n^2} \leq \mathbb{P}\{X_1 \in V\} \leq 2^{-\varepsilon d/n} \text{ (} V \text{ is of combinatorial dimension } d\text{)}.$$

To prove the estimate for $\mathbb{P}\{B_n \text{ is singular}\}$, it is enough to verify that

$$\sum_{V \in \Omega_d} \mathbb{P}\{X_1, X_2, \ldots, X_n \text{ span } V\} \leq 2^{-\varepsilon n + o(n)}$$

for all $1 \leq d \leq n$. Here, $\varepsilon > 0$ is a small constant.

History: argument of Kahn–Komlós–Szemerédi

A crucial point of the argument is to estimate $\mathbb{P}\{X_1, X_2, \dots, X_n \text{ span } V\}$ in terms of $\mathbb{P}\{Y_1, Y_2, \dots, Y_n \text{ span } V\}$ for some special collection of random vectors $Y_1, Y_2, \dots, Y_n$. Assume for a moment that there are random vectors $Y_1, Y_2, \dots, Y_n$ such that for any hyperplane $V \in \Omega_d$ we have

$$\mathbb{P}\{X_1, X_2, \dots, X_n \text{ span } V\} \leq 0.99^n \, \mathbb{P}\{Y_1, Y_2, \dots, Y_n \text{ span } V\}.$$

History: argument of Kahn–Komlós–Szemerédi

A crucial point of the argument is to estimate $\mathbb{P}\{X_1, X_2, \dots, X_n \text{ span } V\}$ in terms of $\mathbb{P}\{Y_1, Y_2, \dots, Y_n \text{ span } V\}$ for some special collection of random vectors $Y_1, Y_2, \dots, Y_n$. Assume for a moment that there are random vectors $Y_1, Y_2, \dots, Y_n$ such that for any hyperplane $V \in \Omega_d$ we have

$$\mathbb{P}\{X_1, X_2, \dots, X_n \text{ span } V\} \leq 0.99^n \, \mathbb{P}\{Y_1, Y_2, \dots, Y_n \text{ span } V\}.$$

Then the trivial identity

$$\sum_{V \in \Omega_d} \mathbb{P}\{Y_1, Y_2, \dots, Y_n \text{ span } V\} \leq 1$$

would imply

$$\sum_{V \in \Omega_d} \mathbb{P}\{X_1, X_2, \dots, X_n \text{ span } V\} \leq 0.99^n.$$

History: argument of Kahn–Komlós–Szemerédi

A crucial point of the argument is to estimate $\mathbb{P}\{X_1, X_2, \ldots, X_n \text{ span } V\}$ in terms of $\mathbb{P}\{Y_1, Y_2, \ldots, Y_n \text{ span } V\}$ for some special collection of random vectors $Y_1, Y_2, \ldots, Y_n$. Assume for a moment that there are random vectors $Y_1, Y_2, \ldots, Y_n$ such that for any hyperplane $V \in \Omega_d$ we have

$$\mathbb{P}\{X_1, X_2, \ldots, X_n \text{ span } V\} \leq 0.99^n \, \mathbb{P}\{Y_1, Y_2, \ldots, Y_n \text{ span } V\}.$$

Then the trivial identity

$$\sum_{V \in \Omega_d} \mathbb{P}\{Y_1, Y_2, \ldots, Y_n \text{ span } V\} \leq 1$$

would imply

$$\sum_{V \in \Omega_d} \mathbb{P}\{X_1, X_2, \ldots, X_n \text{ span } V\} \leq 0.99^n.$$

It turns out that if we replace X_i with vectors $\widetilde{Y}_i$ of "lazy" variables taking values $\{-1, 0, 1\}$ then for any hyperplane V we have

$$\mathbb{P}\{X_1 \in V\} \leq 2^{-\varepsilon'} \mathbb{P}\{\widetilde{Y}_i \in V\}.$$

Then $\widetilde{Y}_1, \ldots, \widetilde{Y}_n$ can act as "approximately" as vectors $Y_1, \ldots, Y_n$ in the above reasoning (the actual argument is somewhat more complicated).

History: argument of Kahn–Komlós–Szemerédi

The original approach of Kahn–Komlós–Szemerédi was later improved by Tao and Vu who showed that

$$\mathbb{P}\{B_n \text{ is singular}\} \leq \left(\frac{3}{4} + o(1)\right)^n.$$

History: argument of Kahn–Komlós–Szemerédi

The original approach of Kahn–Komlós–Szemerédi was later improved by Tao and Vu who showed that

$$\mathbb{P}\{B_n \text{ is singular}\} \leq \left(\frac{3}{4} + o(1)\right)^n.$$

Further improvement was obtained by Bourgain–Vu–Wood about 10 years ago. Following the established strategy, with some important optimizations of the argument, they showed that

$$\mathbb{P}\{B_n \text{ is singular}\} \leq \left(\frac{1}{\sqrt{2}} + o(1)\right)^n,$$

and also obtained asymptotically optimal estimates for some special models of discrete random matrices. In particular, they showed that for $n \times n$ matrix M_n with i.i.d three-valued entries taking values $+1$ and -1 with probability $1/4$ and zero with probability a half,

$$\mathbb{P}\{M_n \text{ is singular}\} = \left(\frac{1}{2} + o(1)\right)^n.$$

History: quantitative argument of Tao–Vu

It is hard to extract any quantitative information about the smallest singular value of square random matrices from the theorem of Kahn–Komlós–Szemerédi and its refinements. An important step in this direction was made independently by Tao–Vu and Rudelson.

History: quantitative argument of Tao–Vu

It is hard to extract any quantitative information about the smallest singular value of square random matrices from the theorem of Kahn–Komlós–Szemerédi and its refinements. An important step in this direction was made independently by Tao–Vu and Rudelson.

Tao and Vu (2007) estimated the smallest singular value of discrete random matrices by studying arithmetic structure of "potential" almost null vectors of the matrix. In particular, their result implies that for any $K > 0$ there is $L > 0$ depending only on K such that for all sufficiently large n

$$\mathbb{P}\{s_{\min}(B_n) \leq n^{-L}\} \leq n^{-K}.$$

History: quantitative argument of Tao–Vu

It is hard to extract any quantitative information about the smallest singular value of square random matrices from the theorem of Kahn–Komlós–Szemerédi and its refinements. An important step in this direction was made independently by Tao–Vu and Rudelson.

Tao and Vu (2007) estimated the smallest singular value of discrete random matrices by studying arithmetic structure of "potential" almost null vectors of the matrix. In particular, their result implies that for any $K > 0$ there is $L > 0$ depending only on K such that for all sufficiently large n

$$\mathbb{P}\{s_{\min}(B_n) \le n^{-L}\} \le n^{-K}.$$

The proof is based on a theorem which asserts that any fixed integer vector $v = (v_1, \dots, v_n)$ with $\sup_{r \in \mathbb{R}} \mathbb{P}\{\sum_i b_i v_i = r\} \ge n^{-R}$, almost all coordinates of v are contained in a generalized arithmetic progression with some special properties. This allows to essentially bound the probability $\mathbb{P}\{s_{\min}(B_n) \le n^{-L}\}$ by a sum of probabilities of the form $\mathbb{P}\{\|B_n v\|_2 \le n^{-L''}\}$ for some special set of integer vectors v.

Approach of Rudelson and Vershynin

The famous result of Rudelson–Vershynin (2008) strengthens the theorem of Kahn–Komlós–Szemerédi simultaneously in two directions: by providing strong quantitative estimates on $s_{\min}$ and by taking a much broader class of distributions.

Approach of Rudelson and Vershynin

The famous result of Rudelson–Vershynin (2008) strengthens the theorem of Kahn–Komlós–Szemerédi simultaneously in two directions: by providing strong quantitative estimates on $s_{\min}$ and by taking a much broader class of distributions.

Let A_n be $n \times n$ random matrix with i.i.d entries of zero mean, unit variance and with bounded subgaussian moment. Then for any $t > 0$

$$\mathbb{P}\{s_{\min}(A_n) \leq t/\sqrt{n}\} \leq Ct + c^n,$$

where $C > 0$ and $c \in (0, 1)$ may only depend on the subgaussian moment.

Approach of Rudelson and Vershynin

The famous result of Rudelson–Vershynin (2008) strengthens the theorem of Kahn–Komlós–Szemerédi simultaneously in two directions: by providing strong quantitative estimates on $s_{\min}$ and by taking a much broader class of distributions.

Let A_n be $n \times n$ random matrix with i.i.d entries of zero mean, unit variance and with bounded subgaussian moment. Then for any $t > 0$

$$\mathbb{P}\{s_{\min}(A_n) \leq t/\sqrt{n}\} \leq Ct + c^n,$$

where $C > 0$ and $c \in (0, 1)$ may only depend on the subgaussian moment.

The argument of Rudelson and Vershynin is based on three key components:

- Compressible and incompressible vectors;
- Reduction to structural properties of random normals;
- The notion of the *Least Common Denominator*, and an extension of the classical Erdős–Littlewood–Offord lemma.

Rudelson–Vershynin: compressible/incompressible vectors

A vector $x \in S^{n-1}$ is called (δ, ρ)–*compressible* if the Euclidean distance to the set of δn–sparse vectors is at most ρ. For example, the unit vector $(0.5, 0.8, 0.2, 0.2, 0.1, -0.1, -0.1) \in S^6$ is $(4/7, \sqrt{0.03})$–compressible. The remaining vectors are called (δ, ρ)–*incompressible*.

Rudelson–Vershynin: compressible/incompressible vectors

A vector $x \in S^{n-1}$ is called (δ, ρ)–*compressible* if the Euclidean distance to the set of δn–sparse vectors is at most ρ. For example, the unit vector $(0.5, 0.8, 0.2, 0.2, 0.1, -0.1, -0.1) \in S^6$ is $(4/7, \sqrt{0.03})$–compressible. The remaining vectors are called (δ, ρ)–*incompressible*.

We have

$$\begin{aligned}\mathbb{P}\{s_{\min}(A_n) \le t/\sqrt{n}\} \le{}& \mathbb{P}\big\{\|A_n x\|_2 \le t/\sqrt{n} \text{ for some } x \in \mathrm{Comp}_n(\delta,\rho)\big\} \\ &+ \mathbb{P}\big\{\|A_n x\|_2 \le t/\sqrt{n} \text{ for some } x \in \mathrm{Incomp}_n(\delta,\rho)\big\}.\end{aligned}$$

Rudelson–Vershynin: compressible/incompressible vectors

A vector $x \in S^{n-1}$ is called (δ, ρ)–*compressible* if the Euclidean distance to the set of δn–sparse vectors is at most ρ. For example, the unit vector $(0.5, 0.8, 0.2, 0.2, 0.1, -0.1, -0.1) \in S^6$ is $(4/7, \sqrt{0.03})$–compressible. The remaining vectors are called (δ, ρ)–*incompressible*.

We have

$$\mathbb{P}\{s_{\min}(A_n) \le t/\sqrt{n}\} \le \mathbb{P}\big\{\|A_n x\|_2 \le t/\sqrt{n} \text{ for some } x \in \mathrm{Comp}_n(\delta,\rho)\big\} + \mathbb{P}\big\{\|A_n x\|_2 \le t/\sqrt{n} \text{ for some } x \in \mathrm{Incomp}_n(\delta,\rho)\big\}.$$

The first probability can be easily estimated using a standard *covering argument*: one can define a Euclidean ε–net $\mathcal{N}$ on the set of compressible vectors and apply the relation

$$\mathbb{P}\big\{\|A_n x\|_2 \le c\sqrt{n} \text{ for some } x \in \mathrm{Comp}_n(\delta,\rho)\big\} \le |\mathcal{N}| \sup_{x\in\mathcal{N}} \mathbb{P}\big\{\|A_n x\|_2 \le c\sqrt{n} + \varepsilon C\sqrt{n}\big\} + \mathbb{P}\{\|A_n\| \ge C\sqrt{n}\} \le e^{-c'n},$$

where $\mathbb{P}\big\{\|A_n x\|_2 \le c\sqrt{n} + \varepsilon C\sqrt{n}\big\}$ is estimated using that $A_n x$ is a vector with independent subgaussian components.

Rudelson–Vershynin: reduction to random normals

It is not difficult to show that *incompressible vectors are flat*. This means that every unit vector which has a considerable distance to the set of δn–sparse vectors, must have a proportional to n number of components of absolute value $\approx 1/\sqrt{n}$. This observation, together with a special averaging argument, implies

$$\begin{aligned}\mathbb{P}\big\{\|A_n x\|_2 \le t/\sqrt{n} \text{ for some } x \in \mathrm{Incomp}_n(\delta,\rho)\big\} \\ \le \frac{1}{\delta}\mathbb{P}\big\{|\langle \mathrm{col}_n(A_n), Y_n\rangle| \le t/\rho\big\}\big\},\end{aligned}$$

where Y_n is the random unit vector orthogonal to the first $n-1$ columns of A_n.

Rudelson–Vershynin: reduction to random normals

It is not difficult to show that *incompressible vectors are flat*. This means that every unit vector which has a considerable distance to the set of δn–sparse vectors, must have a proportional to n number of components of absolute value $\approx 1/\sqrt{n}$. This observation, together with a special averaging argument, implies

$$\begin{aligned}\mathbb{P}\big\{\|A_n x\|_2 \le t/\sqrt{n} \text{ for some } x \in \mathrm{Incomp}_n(\delta,\rho)\big\} \\ \le \frac{1}{\delta}\mathbb{P}\big\{|\langle \mathrm{col}_n(A_n), Y_n\rangle| \le t/\rho\big\}\big\},\end{aligned}$$

where Y_n is the random unit vector orthogonal to the first $n-1$ columns of A_n.

The *Lévy concentration function* of a random variable ξ is defined as

$$\mathcal{L}(\xi, s) := \sup_{r\in\mathbb{R}} \mathbb{P}\big\{|\xi - r| \le s\big\}, \quad s \ge 0.$$

Then, in view of the above, the theorem of Rudelson–Vershynin is implied by the estimate

$$\mathbb{P}_{Y_n}\big\{\mathcal{L}_{\mathrm{col}_n(A_n)}\big(\langle \mathrm{col}_n(A_n), Y_n\rangle, s\big) \le Cs, \quad s \ge e^{-c'n}\big\} \ge 1 - e^{-c'n}.$$

Rudelson–Vershynin: The Least Common Denominator

The *least common denominator* of a unit vector x is defined as

$$\mathrm{LCD}(x) := \inf\left\{\theta > 0 : \ \mathrm{dist}(\theta x, \mathbb{Z}^n) < \min(c_1\|\theta x\|_2, c_2\sqrt{n})\right\},$$

where $c_1, c_2 > 0$ are two small constants. The least common denominator characterizes "unstructuredness" of the vector.

Rudelson–Vershynin: The Least Common Denominator

The *least common denominator* of a unit vector x is defined as

$$\mathrm{LCD}(x) := \inf\left\{\theta > 0 : \ \mathrm{dist}(\theta x, \mathbb{Z}^n) < \min(c_1\|\theta x\|_2, c_2\sqrt{n})\right\},$$

where $c_1, c_2 > 0$ are two small constants. The least common denominator characterizes "unstructuredness" of the vector. The key element of the Rudelson–Vershynin argument is the relation between the least common denominator and properties of the random sums. If $a_1, \ldots, a_n$ are i.i.d random variables of unit variance then

$$\mathcal{L}\Big(\sum_{i=1}^n a_i x_i, t\Big) \le Ct + Ce^{-c'n} \quad \text{for all } t \ge \frac{1}{\mathrm{LCD}(x)}.$$

Rudelson–Vershynin: The Least Common Denominator

The *least common denominator* of a unit vector x is defined as

$$\mathrm{LCD}(x) := \inf\left\{\theta > 0 : \ \mathrm{dist}(\theta x, \mathbb{Z}^n) < \min(c_1\|\theta x\|_2, c_2\sqrt{n})\right\},$$

where $c_1, c_2 > 0$ are two small constants. The least common denominator characterizes "unstructuredness" of the vector. The key element of the Rudelson–Vershynin argument is the relation between the least common denominator and properties of the random sums. If $a_1, \ldots, a_n$ are i.i.d random variables of unit variance then

$$\mathcal{L}\Big(\sum_{i=1}^n a_i x_i, t\Big) \le Ct + Ce^{-c'n} \quad \text{for all } t \ge \frac{1}{\mathrm{LCD}(x)}.$$

Then the main theorem for $s_{\min}(A_n)$ follows by proving that for the random unit normal Y_n,

$$\mathbb{P}_{Y_n}\left\{\mathrm{LCD}(Y_n) \ge e^{c'n}\right\} \ge 1 - e^{-c'n}.$$

This result is proved using an elaborate covering argument, by ruling out the possibility that the least common denominator is subexponential.

Summary of earlier results

Let B_n be the $n \times n$ Bernoulli random matrix with i.i.d ± 1 entries.

- The result of Rudelson–Vershynin implies the estimate

$$\mathbb{P}\{s_{\min}(B_n) \leq t/\sqrt{n}\} \leq Ct + c^n, \quad t > 0,$$

for some constants $C > 0$, $c \in (0, 1)$. In particular, B_n is non-singular with probability at least $1 - c^n$.

Summary of earlier results

Let B_n be the $n \times n$ Bernoulli random matrix with i.i.d ± 1 entries.

- The result of Rudelson–Vershynin implies the estimate

$$\mathbb{P}\{s_{\min}(B_n) \leq t/\sqrt{n}\} \leq Ct + c^n, \quad t > 0,$$

for some constants $C > 0$, $c \in (0, 1)$. In particular, B_n is non-singular with probability at least $1 - c^n$.

- The argument of Kahn–Komlós–Szemerédi and its development by Tao–Vu and Bourgain–Vu–Wood provides stronger singularity probability estimates:

$$\mathbb{P}\{B_n \text{ is singular}\} \leq \Big(\frac{1}{\sqrt{2}} + o(1)\Big)^n,$$

although the method does not seem to imply strong small ball probability estimates for $s_{\min}(B_n)$.

Summary of earlier results

Let B_n be the $n \times n$ Bernoulli random matrix with i.i.d ± 1 entries.

- The result of Rudelson–Vershynin implies the estimate

$$\mathbb{P}\{s_{\min}(B_n) \leq t/\sqrt{n}\} \leq Ct + c^n, \quad t > 0,$$

for some constants $C > 0$, $c \in (0, 1)$. In particular, B_n is non-singular with probability at least $1 - c^n$.

- The argument of Kahn–Komlós–Szemerédi and its development by Tao–Vu and Bourgain–Vu–Wood provides stronger singularity probability estimates:

$$\mathbb{P}\{B_n \text{ is singular}\} \leq \left(\frac{1}{\sqrt{2}} + o(1)\right)^n,$$

although the method does not seem to imply strong small ball probability estimates for $s_{\min}(B_n)$.

- The folklore conjecture in the field is that

$$\mathbb{P}\{B_n \text{ is singular}\} = \left(\frac{1}{2} + o(1)\right)^n,$$

that is, considerable contribution to the matrix singularity comes from the event that two rows or columns of the matrix are equal.

Theorem

Theorem (T.'18)

Let B_n be an $n \times n$ random matrix with i.i.d ± 1 entries. Then for any $\varepsilon > 0$ there is $C > 0$ depending only on ε such that

$$\mathbb{P}\{s_{\min}(B_n) \leq t/\sqrt{n}\} \leq Ct + C\big(1/2 + \varepsilon\big)^n, \quad t > 0.$$

In particular,

$$\mathbb{P}\{B_n \text{ is singular}\} = \Big(\frac{1}{2} + o_n(1)\Big)^n.$$

Proof: preliminary reductions

Repeating an argument of Rudelson–Vershynin, we get

$$\begin{aligned}
&\mathbb{P}\big\{s_{\min}(B_n) \le t/\sqrt{n}\big\} \\
&\quad \le \mathbb{P}\big\{\|B_n x\|_2 \le t/\sqrt{n} \text{ for some } x \in \mathrm{Comp}_n(\delta,\rho)\big\} \\
&\qquad\qquad + \frac{1}{\delta}\mathbb{P}\big\{|\langle \mathrm{col}_n(B_n), Y_n\rangle| \le t/\rho\big\} \\
&\quad \le \left(\frac{1}{2}+\varepsilon\right)^n + C\,\mathbb{P}\big\{|\langle \mathrm{col}_n(B_n), Y_n\rangle| \le Ct\big\},
\end{aligned}$$

Proof: preliminary reductions

Repeating an argument of Rudelson–Vershynin, we get

$$\begin{aligned}
&\mathbb{P}\big\{s_{\min}(B_n) \le t/\sqrt{n}\big\} \\
&\quad \le \mathbb{P}\big\{\|B_n x\|_2 \le t/\sqrt{n} \text{ for some } x \in \mathrm{Comp}_n(\delta,\rho)\big\} \\
&\qquad + \frac{1}{\delta}\mathbb{P}\big\{|\langle \mathrm{col}_n(B_n), Y_n\rangle| \le t/\rho\big\} \\
&\quad \le \left(\frac{1}{2}+\varepsilon\right)^n + C\,\mathbb{P}\big\{|\langle \mathrm{col}_n(B_n), Y_n\rangle| \le Ct\big\},
\end{aligned}$$

Hence, to prove the theorem, it is enough to show that with probability $\mathbb{P}_{Y_n}$ at least $1-(1/2+\varepsilon)^n$ we have

$$\mathcal{L}_{\mathrm{col}_n(B_n)}\big(\langle \mathrm{col}_n(B_n), Y_n\rangle, t\big) \le Ct \quad \text{for all } t \ge \left(\frac{1}{2}+\varepsilon\right)^n.$$

Here, Y_n is the unit vector orthogonal to the first $n-1$ columns of B_n, and $\mathcal{L}(\xi,t) = \sup_r \mathbb{P}\{|\xi - r| \le t\}$.

In other words, we need to show that the normal vector Y_n is typically "very unstructured", so that its scalar product with the column $\mathrm{col}_n(B_n)$ behaves as a random variable with a bounded density up to the scale $\left(\frac{1}{2}+\varepsilon\right)^n$.

Proof: discretization

Fix a large constant $L > 0$, take some $N \ll \left(2 - \varepsilon\right)^n$ and define a set of unit vectors

$$Q_N := \Big\{ x : \ \sup\big\{ t \in [0,1] : \ \mathcal{L}\big(\langle \mathrm{col}_n(B_n), x\rangle, t\big) > Lt \big\} \in \Big[\frac{1}{2N}, \frac{1}{N}\Big] \Big\}.$$

Then the proof of the main result amounts to showing that

$$\mathbb{P}\{ Y_n \in Q_N \} \leq \left(\frac{1}{2} + \varepsilon\right)^n.$$

Proof: discretization

Fix a large constant $L > 0$, take some $N \ll \left(2 - \varepsilon\right)^n$ and define a set of unit vectors

$$Q_N := \left\{ x : \ \sup \left\{ t \in [0,1] : \ \mathcal{L}\big(\langle \mathrm{col}_n(B_n), x \rangle, t\big) > Lt \right\} \in \left[\frac{1}{2N}, \frac{1}{N} \right] \right\}.$$

Then the proof of the main result amounts to showing that

$$\mathbb{P}\{ Y_n \in Q_N \} \leq \left(\frac{1}{2} + \varepsilon \right)^n.$$

It can be shown that on the event $\{ Y_n \in Q_N \}$ the vector Y_n can be approximated by a vector $\mathbf{Y} \in \left(\frac{1}{N} \mathbb{Z} \right)^n$ such that

- (distance to Y_n) $\| Y_n - \mathbf{Y} \|_\infty \leq \frac{1}{N}$;

Proof: discretization

Fix a large constant $L > 0$, take some $N \ll (2-\varepsilon)^n$ and define a set of unit vectors

$$Q_N := \Big\{ x : \ \sup\big\{ t \in [0,1] : \ \mathcal{L}\big(\langle \mathrm{col}_n(B_n), x\rangle, t\big) > Lt \big\} \in \Big[\frac{1}{2N}, \frac{1}{N}\Big] \Big\}.$$

Then the proof of the main result amounts to showing that

$$\mathbb{P}\{Y_n \in Q_N\} \le \Big(\frac{1}{2} + \varepsilon\Big)^n.$$

It can be shown that on the event $\{Y_n \in Q_N\}$ the vector Y_n can be approximated by a vector $\mathbf{Y} \in \big(\frac{1}{N}\mathbb{Z}\big)^n$ such that

- (distance to Y_n) $\|Y_n - \mathbf{Y}\|_\infty \le \frac{1}{N}$;
- (anti-concentration) $\mathbb{P}\big\{\big|\sum_{i=1}^n b_i \mathbf{Y}_i\big| \le t\big\} \le C\,Lt$ for all $t \ge 1/N$;

Proof: discretization

Fix a large constant $L > 0$, take some $N \ll (2-\varepsilon)^n$ and define a set of unit vectors

$$Q_N := \Big\{ x : \ \sup\big\{ t \in [0,1] : \ \mathcal{L}\big(\langle \mathrm{col}_n(B_n), x\rangle, t\big) > Lt \big\} \in \Big[\frac{1}{2N}, \frac{1}{N}\Big] \Big\}.$$

Then the proof of the main result amounts to showing that

$$\mathbb{P}\{Y_n \in Q_N\} \le \Big(\frac{1}{2} + \varepsilon\Big)^n.$$

It can be shown that on the event $\{Y_n \in Q_N\}$ the vector Y_n can be approximated by a vector $\mathbf{Y} \in \big(\frac{1}{N}\mathbb{Z}\big)^n$ such that

- (distance to Y_n) $\|Y_n - \mathbf{Y}\|_\infty \le \frac{1}{N}$;
- (anti-concentration) $\mathbb{P}\big\{\big|\sum_{i=1}^n b_i \mathbf{Y}_i\big| \le t\big\} \le C\,Lt$ for all $t \ge 1/N$;
- (concentration) $\mathcal{L}\big(\sum_{i=1}^n b_i \mathbf{Y}_i, 1/N\big) \ge c\,\mathcal{L}\big(\sum_{i=1}^n b_i (Y_n)_i, 1/N\big)$;

Proof: discretization

Fix a large constant $L > 0$, take some $N \ll (2-\varepsilon)^n$ and define a set of unit vectors

$$Q_N := \Big\{ x : \ \sup\big\{ t \in [0,1] : \ \mathcal{L}\big(\langle \mathrm{col}_n(B_n), x\rangle, t\big) > Lt \big\} \in \Big[\frac{1}{2N}, \frac{1}{N}\Big] \Big\}.$$

Then the proof of the main result amounts to showing that

$$\mathbb{P}\{Y_n \in Q_N\} \le \Big(\frac{1}{2} + \varepsilon\Big)^n.$$

It can be shown that on the event $\{Y_n \in Q_N\}$ the vector Y_n can be approximated by a vector $\mathbf{Y} \in \big(\frac{1}{N}\mathbb{Z}\big)^n$ such that

- (distance to Y_n) $\|Y_n - \mathbf{Y}\|_\infty \le \frac{1}{N}$;
- (anti-concentration) $\mathbb{P}\big\{\big|\sum_{i=1}^n b_i \mathbf{Y}_i\big| \le t\big\} \le C\,Lt$ for all $t \ge 1/N$;
- (concentration) $\mathcal{L}\big(\sum_{i=1}^n b_i \mathbf{Y}_i, 1/N\big) \ge c\,\mathcal{L}\big(\sum_{i=1}^n b_i (Y_n)_i, 1/N\big)$;
- (distance to the column span) $\|B_{1..n-1}^\top \mathbf{Y}\|_2 \le Cn/N$.

Here, $B_{1..n-1}$ is the first $n-1$ columns of B_n.

How the vector $\mathbf{Y}$ is constructed:

Discretization (continued)

The procedure is called "random rounding" in the literature. It was used by

- Alon–Klartag, Klartag–Livshyts, and, more recently,
- Livshyts (2018) when estimating $s_{\min}$ for inhomogeneous random matrices.

The use of the random rounding in the Bernoulli setting is inspired by these papers.

Discretization (continued)

The procedure is called "random rounding" in the literature. It was used by

- Alon–Klartag, Klartag–Livshyts, and, more recently,
- Livshyts (2018) when estimating $s_{\min}$ for inhomogeneous random matrices.

The use of the random rounding in the Bernoulli setting is inspired by these papers.
To construct the approximation $\mathbf{Y}$ of the vector Y_n, satisfying the conditions mentioned above, we replace each component $(Y_n)_i$ with a random variable $\mathbf{Y}_i$ distributed on the set $\left\{\lfloor N(Y_n)_i \rfloor / N, \lfloor N(Y_n)_i \rfloor / N + 1/N\right\}$, and such that $\mathbb{E}_{\mathbf{Y}_i} \mathbf{Y}_i = (Y_n)_i$. Then with high probability $\mathbf{Y}$ satisfies the needed properties.

Thus, for each "bad" realization of the normal Y_n we can construct an appropriate discrete approximation $\mathbf{Y}$. Next:

Competition of net cardinality and the small ball probability

Our goal is to show that $\mathbb{P}\{Y_n \in Q_N\} \leq \left(\frac{1}{2} + \varepsilon\right)^n$, where

$$Q_N = \left\{x : \sup\left\{t \in [0,1] : \mathcal{L}\big(\langle \mathrm{col}_n(B_n), x\rangle, t\big) > Lt\right\} \in \left[\frac{1}{2N}, \frac{1}{N}\right]\right\}$$

(and $N \ll \big(2 - \varepsilon\big)^n$).

Competition of net cardinality and the small ball probability

Our goal is to show that $\mathbb{P}\{Y_n \in Q_N\} \le \left(\frac{1}{2}+\varepsilon\right)^n$, where

$$Q_N = \left\{x : \sup\left\{t \in [0,1] : \mathcal{L}\big(\langle \mathrm{col}_n(B_n), x\rangle, t\big) > Lt\right\} \in \left[\frac{1}{2N}, \frac{1}{N}\right]\right\}$$

(and $N \ll \left(2-\varepsilon\right)^n$).
For each "bad" realization $Y_n \in Q_N$, we constructed corresponding approximation $\mathbf{Y}$. The set of all these approximations $\mathcal{N} \subset \left(\frac{1}{N}\mathbb{Z}\right)^n$. Then we get

$$\mathbb{P}\{Y_n \in Q_N\} \le |\mathcal{N}| \sup_{y\in\mathcal{N}} \mathbb{P}\left\{\|B^{\top}_{1..n-1} y\|_2 \le Cn/N\right\} \le |\mathcal{N}|\,(\widetilde{C}n/N)^n.$$

Competition of net cardinality and the small ball probability

Our goal is to show that $\mathbb{P}\{Y_n \in Q_N\} \le \left(\frac{1}{2}+\varepsilon\right)^n$, where

$$Q_N = \left\{x:\ \sup\left\{t\in[0,1]:\ \mathcal{L}\big(\langle \mathrm{col}_n(B_n), x\rangle, t\big) > Lt\right\} \in \left[\frac{1}{2N}, \frac{1}{N}\right]\right\}$$

(and $N \ll \left(2-\varepsilon\right)^n$).
For each "bad" realization $Y_n \in Q_N$, we constructed corresponding approximation $\mathbf{Y}$. The set of all these approximations $\mathcal{N} \subset \left(\frac{1}{N}\mathbb{Z}\right)^n$. Then we get

$$\mathbb{P}\{Y_n \in Q_N\} \le |\mathcal{N}| \sup_{y\in\mathcal{N}} \mathbb{P}\left\{\|B^{\top}_{1..n-1} y\|_2 \le Cn/N\right\} \le |\mathcal{N}|\,(\widetilde{C}n/N)^n.$$

The estimate is satisfactory as long as $|\mathcal{N}| \ll (\widetilde{C}n/N)^{-n}$. So, cardinality of the discretization $\mathcal{N}$ must be relatively small. The estimate for the cardinality is the main element of the proof, and is based on *double counting*:

Cardinality of the net

Theorem (Double Counting)

Let $\delta \in (0,1]$, $M \geq 1$. There exist $L_{\text{B}} = L_{\text{B}}(\delta) > 0$ depending **only** on δ (and not on M) with the following property. Take a large n, and $1 \leq N \leq (2-\varepsilon)^n$, and let

$$\mathcal{A} := \{-2N, \dots, -N-1, N+1, \dots, 2N\}^{\delta n} \times \{-N, -N+1, \dots, N\}^{n-\delta n}.$$

Further, assume that a random vector $(\xi_1, \dots, \xi_n)$ is uniform on $\mathcal{A}$. Then

$$\mathbb{P}\left\{ 2^{-n} \sup_{\lambda \in \mathbb{R}} \sum_{(v_j)_{j=1}^n \in \{-1,1\}^n} \mathbf{1}_{[-\sqrt{n},\sqrt{n}]}\left(\lambda + v_1\xi_1 + \dots + v_n\xi_n\right) > \frac{L_{\text{B}}}{N} \right\} \leq e^{-M n}.$$

Cardinality of the net

Theorem (Double Counting)

Let $\delta \in (0,1]$, $M \geq 1$. There exist $L_{\text{B}} = L_{\text{B}}(\delta) > 0$ depending **only** on δ (and not on M) with the following property. Take a large n, and $1 \leq N \leq (2-\varepsilon)^n$, and let

$$\mathcal{A} := \{-2N, \dots, -N-1, N+1, \dots, 2N\}^{\delta n} \times \{-N, -N+1, \dots, N\}^{n-\delta n}.$$

Further, assume that a random vector $(\xi_1, \dots, \xi_n)$ is uniform on $\mathcal{A}$. Then

$$\mathbb{P}\left\{ 2^{-n} \sup_{\lambda \in \mathbb{R}} \sum_{(v_j)_{j=1}^n \in \{-1,1\}^n} \mathbf{1}_{[-\sqrt{n},\sqrt{n}]}\left(\lambda + v_1\xi_1 + \dots + v_n\xi_n\right) > \frac{L_{\text{B}}}{N} \right\} \leq e^{-M n}.$$

The crucial point of the statement is that L_{B} does not depend on M, so one can make M arbitrarily small as long as n is large. This statement (in fact, a little more technical version) is translated into the cardinality estimates for the net $\mathcal{N}$ discussed in the previous slide.

The Double Counting theorem: a "dynamical" viewpoint

It is important to get a convenient interpretation of the quantity

$$\sup_{\lambda\in\mathbb{R}} \sum_{(v_j)_{j=1}^n\in\{-1,1\}^n} \mathbf{1}_{[-\sqrt{n},\sqrt{n}]}\big(\lambda + v_1\xi_1 + \cdots + v_n\xi_n\big)$$

from the last theorem. Define $f_0(t) := \mathbf{1}_{[-\sqrt{n},\sqrt{n}]}(t)$;

$$f_i(t) := \frac{1}{2}f_{i-1}(t-\xi_i) + \frac{1}{2}f_{i-1}(t+\xi_i), \quad i = 1, 2, \ldots, n.$$

Then

$$2^{-n} \sum_{(v_j)_{j=1}^n\in\{-1,1\}^n} \mathbf{1}_{[-\sqrt{n},\sqrt{n}]}\big(\lambda + v_1\xi_1 + \cdots + v_n\xi_n\big) = f_n(\lambda).$$

Recall that $\xi_1, \ldots, \xi_n$ are independent random variables; with $(\xi_1, \ldots, \xi_n)$ uniformly distributed on

$$\mathcal{A} := \{-2N, \ldots, -N-1, N+1, \ldots, 2N\}^{\delta n} \times \{-N, -N+1, \ldots, N\}^{n-\delta n}.$$

The goal is then to bound the $\|\cdot\|_\infty$–norm of $f_n : \mathbb{Z} \to \mathbb{R}_+$.

The Double Counting theorem: a "dynamical" viewpoint

Example of the evolution of f_i's: