

Random discrete matrices

Van H. Vu

Department of Mathematics

Rutgers

vanvu@math.rutgers.edu

The models.

M_n (non-symmetric): n by n matrix with i.i.d (in many consideration this can be significantly weakened) entries: ξ_{ij} .

Q_n (symmetric): $\xi_{ij} = \xi_{ji}$.

Continuous models: ξ_{ij} have continuous distribution. Representative example: Gaussian.

Discrete models: ξ_{ij} have discrete distribution. Representative example: Bernoulli (± 1 with probability $1/2$).

The questions:

- (1) Rank of M_n . Probability of Singularity.
- (2) Determinant.
- (3) Spectral norm ($\|M_n\|_{spec}$ = the largest singular value)
- (4) Condition number ($\|M_n\|_{spec}\|M_n^{-1}\|_{spec}$).
- (5) Limiting distribution of the spectra.

Continuous models. Precise answers.

(1) Joint distribution of the eigenvalues:

$$p(\lambda_1, \dots, \lambda_n) = c_n \prod_{[i < j]} |\lambda_i - \lambda_j|^2 \prod_{i=1}^n e^{-n|\lambda_i|^2}.$$

(2) Moment method:

$$\sum_{i=1}^n \lambda_i^k = \text{Trace} M_n^k.$$

(3) Stieltjes transform:

$$s_n(z) := \frac{1}{n} \text{Trace} \left(\frac{1}{\sqrt{n}} M_n - z I_n \right)^{-1} = \frac{1}{n} \sum_{i=1}^n \frac{1}{n^{-1/2} \lambda_i - z}.$$

Discrete models.

(1) Not available; Use of (2)(3) are more limited.

New method

(4) Concentration function: $v = (a_1, \dots, a_n$

$$P_v := \max_x \mathbf{P}\left(\sum_{i=1}^n \xi_i a_i = x\right).$$

Littlewood-Offord (1943) (Erdos) If $a_i \neq 0$, then $P_v = O(n^{-1/2})$.

We developed a small theory around this theorem, focusing on Inverse statements. The main tool is [additive combinatorics](#).

The rank problem.

Theorem. (Komlos 1967) Almost surely M_n has full rank (or is non-singular).

Conjecture. (Weiss) The same holds for symmetric matrices, i.e., almost surely Q_n has full rank.

Theorem. (Costello-Tao-V., 2004) Almost surely Q_n has full rank.

Q_n can be seen as the adjacency matrix of the random graph $G(n, 1/2)$ (switching -1 to 0 does not affect the rank). So the above theorem can be rewritten as

Theorem. Almost surely $G(n, 1/2)$ has full rank.

Theorem. (Costello-V. 2006) For $p > (1 + o(1)) \log n/n$, $G(n, p)$ a.s. has full rank.

Quadratic Littlewood-Offord theorem. Consider the quadratic form

$$Q(\xi_i) := \sum_{1 \leq i, j \leq n} a_{ij} \xi_i \xi_j.$$

Theorem. (Costello-Tao-V.) If $a_{ij} \neq 0$, then for any x

$$P(Q = x) = O(n^{-1/4}).$$

Conjecture. $O(n^{-1/4})$ can be replaced by $O(n^{-1/2})$.

If true, it is sharp, as one can take $Q = (\xi_1 + \cdots + \xi_n)^2$.

Conjecture. Almost surely a random regular graph with degree at least 3 has full rank.

The singularity problem. Estimate p_n , the probability that M_n is singular.

By considering the probability that there are two equal rows

$$p_n \geq (1/2 + o(1))^n.$$

Conjecture. $p_n = (1/2 + o(1))^n$.

Theorem. (Komlos 1967) $p_n = o(1)$.

Theorem. (Komlos 1976) $p(n) = O(n^{-1/2})$.

Theorem. (Kahn-Komlos-Szemerédi 1995) $p_n = O(.999^n)$.

Theorem. (Tao-V. 2005) $p_n = (3/4 + o(1))^n$.

Theorem. (V.-Wood 2007) $p_n = (\sqrt{1/2} + o(1))^n$.

Some other models.

Instead of M_n we consider the following "lazy" model M_n^{lazy} . The entries of M_n^{lazy} are i.i.d random variables which equal zero with probability half and 1 and -1 with probability one quarter. (If one thinks of the entries of M_n as fair coin flips, then in the "lazy" model about half of the time we are lazy and simply write zero instead flipping a coin.) It is clear that for the lazy model the singular probability p_n^{lazy} is again at least $(1/2 + o(1))^n$ (which is the probability that there is a zero row).

Theorem. (V. Wood 2007) $p_n^{lazy} = (1/2 + o(1))^n$.

For a general result that covers the last two theorems, see Friday's lecture.

Random walks and Lazy random walks.

Let

$$S := \sum_{i=1}^n a_i \xi_i$$

and

$$S^\mu := \sum_{i=1}^n a_i \xi_i^\mu.$$

Consider $P(S = 0)$ and $P(S^\mu = 0)$. Intuitively, the second probability is much larger. However, there are cases where the two probabilities are comparable. For example, if all $a_i = 1$, then both probabilities are $\Theta(n^{-1/2})$.

The core of our method for the singularity problem is a theorem that characterizes all sets a_i where the two probabilities are comparable. (Even when both of them are exponentially small.) It relies on a method

of Halasz (1975) and many arguments from additive combinatorics.

Determinant. $|DetM_n|$.

Fact 1. Komlos (1967) result implies that a.s. $|DetM_n|$ is positive. In fact, since $|DetM_n|$ is divisible by 2^{n-1} , it is at least 2^{n-1} .

Fact 2. (Hadamard bound) $|DetM_n| \leq n^{n/2}$.

Fact 3. Turan (1940s) observed that $E(Det^2 M_n) = n! = n^{(1+o(1))n}$.

Conjecture. A.s. $|DetM_n| = n^{(1/2+o(1))n}$.

Theorem. (Tao-V. 2003) A.s. $|DetM_n| = n^{(1/2+o(1))n}$. In fact, a.s.

$$|DetM_n| \geq \sqrt{n!} \exp(-29\sqrt{n \log n}).$$

The main idea here is to view the determinant as the volume of the parallelepiped spanned by the row vectors of the matrix. Now expose the matrix row by row and compute the volume as the product of the distance from the i th vector to the plane spanned by the first $i - 1$ vectors.

Lemma. With very high probability $d_i \approx \sqrt{i}$.

Conjecture. A.s. $|DetQ_n| = n^{(1/2+o(1))n}$.

Main trouble here is that the rows are no longer dependent.

The condition number problem.

Let M be an $n \times n$ matrix, the *condition number* $\kappa(M)$ is defined as

$$\kappa(M) := \|M\| \|M^{-1}\|.$$

where $\|\cdot\|$ is the spectral norm. If M is singular, $\kappa(M) = \infty$.

The condition number plays a crucial role in numerical linear algebra. For example, the accuracy and stability of most algorithms used to solve the equation $Mx = b$ depend on $\kappa(M)$.

$\|M_n\|$ is, with very high probability, $\Theta(\sqrt{n})$. It is much harder to estimate $\|M_n^{-1}\|$.

A more general and more practical problem is to estimate $\kappa(A + M_n)$, where A is a fixed matrix.

Motivation. Why the simplex algorithm runs fast ? Smooth Analysis (Spielman-Teng 2000).

Key point: Noise helps !! κA may be large, but $\kappa(A + Noise)$ is almost surely small.

Spielman-Teng assumes that Noise is a random Gaussian and proved that

Theorem. (Spielman-Teng) Assuming that $\|A\| = n^{O(1)}$, then a.s. $\kappa(A + Noise) = n^{O(1)}$.

Question. (S-T) What happens with discrete noise ?

Toy case. $A = 0$, $Noise = M_n$. Rudelson (2005), Tao-V. (2005) proved Spielman-Teng statement for this case. Recently Rudelson-Vershynin (2007) obtained a very precise estimate.

Real case. Any A , any discrete noise:

Theorem. (Tao-V. 2006) Assuming that $\|A\| = n^{O(1)}$, then a.s. $\kappa(A + Noise) = n^{O(1)}$.

Main tool. Inverse Littlewood-Offord theorem. Recall the definition

$$P_v := \max_x \mathbf{P}\left(\sum_{i=1}^n \xi_i a_i = x\right).$$

Littlewood-Offord (1943) (Erdos) If $a_i \neq 0$, then $P_v = O(n^{-1/2})$.

The bound is sharp: $a_i = 1$.

If one forbids the a_i be the same, then the bound jumps quite a bit

Theorem. (Erdos-Moser, Sarkozi-Szemerédi, Stanley 80s)

$P_v = O(n^{-3/2})$.

Again, this bound is sharp. One can take $a_i = i$

Question. When is P_v large ?

For instance, if we know $P_v \geq n^{-5}$, what can we say about the a_i . We obtain a complete answer for this question.

Limiting distribution of the spectra.

Theorem. (Wigner 50s) The distribution of the eigenvalues of Q_n follows the semi-circle law.

Wigner proof introduced the Trace method.

Conjecture. The distribution of the eigenvalues of M_n follows the circular law.

Let A_n be a random matrix with i.i.d. entries having mean 0 and variance 1.

Conjecture. (Circular law) The distribution of the eigenvalues of A_n follows the circular law.

The statement is true for Gaussian (Ginibre, Mehta). For general continuous models, Girko (1984), Bai (1997) (moment slightly higher than 2)

Tao-V. (2007), Gotze-Tikhomirov (2007)

Theorem. The distribution of the eigenvalues of M_n follows the circular law.

Gotze-Tikhomirov proved CL for entries with mean 0 having sub-gaussian tails (exponential decay, in particular all moments are bounded).

Tao-V. needs to control a moment slightly higher than 2 (so very close to the main conjecture), but may need an additional condition.

The exact solution $x = M^{-1}b$, **in theory**, can be computed quickly (by Gaussian elimination, say).

However, **in practice** computers can only present a finite subset of real numbers. This leads to two difficulties: The represented numbers cannot be arbitrary large or small, and there are gaps between them.

A quantity which is frequently used in numerical analysis is $\epsilon_{\text{machine}}$ which is half of the distance from 1 to the nearest represented number. A fundamental result in numerical analysis asserts that if one denotes by \tilde{x} the result computed by computers, then the relative error $\frac{\|\tilde{x} - x\|}{\|x\|}$ satisfies

$$\frac{\|\tilde{x} - x\|}{\|x\|} = O(\epsilon_{\text{machine}} \kappa(M))$$

We call M *well conditioned* if $\kappa(M)$ is small. For quantitative purposes, we say that an n by n matrix M is *well conditioned* if its condition number is polynomially bounded in n ($\kappa(M) \leq n^C$ for some constant C independent of n).

Recall that

$$\kappa(M) := \|M\| \|M^{-1}\|.$$

It is easy to bound $\|M\|$:

$$\|M\|^2 \leq \sum_{ij} a_{ij}^2.$$

So if the entries of M are polynomially bounded, then $\|M\|$ is polynomially bounded.

However, a matrix with tiny entries can have **huge inverse**. There is a ± 1 matrix M whose inverse has entries as large as $n^{(1/2+o(1))n}$ (Alon-V. 96). The condition number of this M is super-exponential.

While ill-conditioned matrices do exist, one hardly encounter them in practice.

In fact, linear algebraic algorithms frequently runs faster (and gives higher accuracy) than the worst case analysis predict (for example the simplex method).

The "positive" effect of noise.

It happens frequently that while we are interested in solving a certain equation, because of the noise the computer actually ends up with solving a slightly perturbed version of it.

Spielman and Teng ([smooth analysis](#)) proposed the following general explanation

(P) *Let M be an arbitrary n by n matrix and N_n a random n by n matrix. Then with high probability $M + N_n$ is well conditioned.*

The crucial point here is that M itself may have a large condition number. For example, M can be singular, in which case the condition number is ∞ .

Demmel proved **(P)** for the case $M = 0$ and N_n is Gaussian. Recently, S-T proved **(P)** for arbitrary M and N_n is Gaussian.

Main open problem (S-T):

Verify **(P)** for discrete noise N_n .

$$\kappa(M + N_n) = \|M + N_n\| \|(M + N_n)^{-1}\|.$$

Since

$$\|M\| - \|N_n\| \leq \|M + N_n\| \leq \|M\| + \|N_n\|$$

and $\|N_n\|$ is, with very high probability, $O(\sqrt{n})$ for most natural models of random matrices, in order to make $\|M + N_n\|$ small, it is necessary to assume that the entries of M are bounded by n^C . This assumption takes care of $\|M + N_n\|$. **The critical issue is to bound $\|(M + N_n)^{-1}\|$.**

Theorem. (S-T) Let N_n be the random Gaussian matrix. Then for any $x > 0$,

$$\mathbf{P}(\|(M + N_n)^{-1}\| \geq n^B) = O(n^{-B+1/2}).$$

(geometrical interpretation; sketch of the proof)

In general $\|A^{-1}\|$ is proportional to **the inverse of *thickness* of the simplex spanned by the row vectors of A** . For example, if A is singular, this simplex is flat (contained in hyperplane) and the thickness is zero. The thickness is the minimum distance from one vector to the opposite hyperplane.

Consider the toy case when $M = 0$, so $M + N_n = N_n$. If N_n is Gaussian, **then one can fix the hyperplane**. The distance from a random gaussian vector to a fixed hyperplane is well-understood. It is typically $\Theta(n^{-1/2})$ (polynomially large).

Main difficulty for the discrete case: Assume that N_n is Bernoulli (random ± 1). We talk about a hyperplane spanned by $n - 1$ random ± 1 vectors. **One can no longer fix this hyperplane**. In fact, the distribution of the distance **depends very much on the position of the plane**.

Example. $H_1 := \{(x_1, \dots, x_n) | x_1 + x_2 = 0\}$, half of the hypercube has distance 0 to H_1 . $H_2 := \{(x_1, \dots, x_n) | x_1 + \dots + x_n = 0\}$, only $O(n^{-1/2})$ -fraction of the hypercube has distance 0 to H_2 .

Understanding *bad* discrete hyperplanes (where a large proportion of the cube has small distance to H) is the key to our study. And this leads to a beautiful connection to [additive combinatorics](#).

Theorem.(Tao-V. 06) For any constants A and C there is a constant B such that the following holds. Let M be an integer n by n matrix whose entries (in absolute values) are bounded from above by n^C and N_n be the n by n random Bernoulli matrix. Then

$$\mathbf{P}(\|(M + N_n)^{-1}\| \geq n^B) \leq n^{-A}.$$

In fact, the result holds under much more general (and more applicable) assumption.

- The distributions of the entries of N_n can be fairly arbitrary. The only essential condition is that they are **not concentrated on one point**.
- Different entries can have different distributions. In practice, a noise occurring to a large entry of M should have **larger variance** than the one occurring to a small entry.
- Not every entry needs to be random. We allow each rows and columns to have up to $n^{.99}$ *frozen* entries. In practice, 0 entries (which are zero by definition, not by measurement) are **frequently not effected by noise**.
- Only row and column independence are essential. ($O(n)$ random bits instead of $\Theta(n^2)$.)

Understanding "BAD" hyperplanes

A hyperplane H is bad if a "large" fraction of the ± 1 cube has "small" distance to H .

Large is $n^{-O(1)}$; Small is $n^{-\Omega(1)}$.

Simplified: A hyperplane H is bad if a "large" fraction of the ± 1 cube has zero distance to H .

Let $v = (a_1, \dots, a_n)$ be the normal vector of H . The probability that a random ± 1 vector (ξ_1, \dots, ξ_n) lies in H is

$$\mathbf{P}(a_1\xi_1 + a_2\xi_2 + \dots + a_n\xi_n = 0)$$

where ξ_1, \dots, ξ_n are i.i.d. Bernoulli random variables (± 1).

The problem of bounding

$$\mathbf{P}(a_1\xi_1 + a_2\xi_2 + \dots + a_n\xi_n = 0)$$

is a fundamental one in combinatorics and has a long history.

Littlewood-Offord (1943) If all $a_i \neq 0$, $\mathbf{P} = O(n^{-1/2} \log n)$.

Erdős (1945) $O(n^{-1/2})$. Sharp if take $a_1 = a_2 = \dots = a_n$, n even.

Erdős-Moser (1963) If the a_i are different: $\mathbf{P} = O(n^{-3/2} \log n)$.

Conjectured that log can be removed.

Sár”zy-Szemerédi (1965) Removed the log. The bound is sharp if take $a_i = i$.

Stanley (1980s) Show that $a_i = i$ is the extremal case.

Extensions to higher dimensions (a_i are vectors in \mathbf{R}^3 , say): Katona, Kleitman, Griggs et. al., Frankl-Füredi.

Halász (1977) Harmonic analysis proofs.