Random matrices: Distribution of the least singular value (via Property Testing)

Van H. Vu

Department of Mathematics
Rutgers
vanvu@math.rutgers.edu

(joint work with T. Tao, UCLA)

Let ξ be a real or complex-valued random variable and $M_n(\xi)$ denote the random $n \times n$ matrix whose entries are i.i.d. copies of ξ:

- ($\mathbf{R}$-normalization) ξ is real-valued with $\mathbf{E}\xi = 0$ and $\mathbf{E}\xi^2 = 1$.
- ($\mathbf{C}$-normalization) ξ is complex-valued with $\mathbf{E}\xi = 0$, $\mathbf{E}\Re(\xi)^2 = \mathbf{E}\Im(\xi)^2 = \frac{1}{2}$, and $\mathbf{E}\Re(\xi)\Im(\xi) = 0$.

In both cases ξ has mean zero and variance one.

Examples. real gaussian, complex gaussian, Bernoulli (± 1 with probability 1/2).

Numerical Algebra.

von Neumann-Goldstine (1940s): What is the condition number and the least singular value of a random matrix ?

Prediction. With high probability, $\sigma_n = \Theta(\sqrt{n}), \kappa = \Theta(n)$.

Smale (1980s), Demmel (1980s): Typical complexity of a numerical problem.

Spielman-Teng (2000s): Smooth analysis.

Probability/Mathematical Physics. A basic problem in Random Matrix Theory is to understand the distributions of the eigenvalues and singular values.

- Limiting distribution of the whole spectrum (such as Wigner semi-circle law).
- Limiting distribution of extremal eigenvalues/singular values (such as Tracy-Widom law).

A special case: Gaussian models. Explicit formulae for the joint distributions of the eigenvalues of $\frac{1}{\sqrt{n}}M_n$

$$(Real\ \ Gaussian)\ \ c_1(n)\prod_{1\le i<j\le n}|\lambda_i-\lambda_j|\exp(-\sum_{i=1}^n\lambda_i^2/2). \qquad (1)$$

$$(Complex\ \ Gaussian)\ \ c_2(n)\prod_{1\le i<j\le n}|\lambda_i-\lambda_j|^2\exp(-\sum_{i=1}^n\lambda_i^2/2). \qquad (2)$$

Explicit formulae for the joint distributions of the eigenvalues of $\frac{1}{n}M_nM_n^*$ (or the singular values of $\frac{1}{\sqrt{n}}M_n$)

$$(\textit{Real Gaussian})\ \ c_3(n)\prod_{1\le i<j\le n}(\lambda_i-\lambda_j)\prod_{i=1}^{n}\lambda_i^{-1/2}\exp(-\sum_{i=1}^{n}\lambda_i/2). \quad (3)$$

$$(\textit{Complex Gaussian})\ \ c_4(n)\prod_{1\le i<j\le n}|\lambda_i-\lambda_j|^2\exp(-\sum_{i=1}^{n}\lambda_i/2). \quad (4)$$

The limiting distributions for Gaussian matrices can be computed directly from these explicit formulae.

Universality Principle. The same results must hold for general normalized random variables.

Informally: The limiting distributions of the spectrum should not depend too much on the distribution of the entries.

Same spirit: Central limit theorem.

Bulk Distributions.

Circular Law. The limiting distribution of the eigenvalues of $\frac{1}{\sqrt{n}} M_n$ is uniform in the unit circle. (Proved for complex gaussian by Mehta 1960s, real gaussian by Edelman 1980s, Girko, Bai, Götze-Tykhomiro, Pan-Zhu, Tao-Vu (2000s). Full generality: Tao-Vu 2008.)

Marchenko-Patur Law. The limiting distribution of the eigenvalues of $\frac{1}{n} M_n M_n^*$ has density $\frac{1}{2\pi} \int_0^{\min(t,4)} \sqrt{\frac{4}{x} - 1}\, dx$. (Marchenko-Pastur 1967).

The singular values of M_n are often viewed as the (square roots) of the eigenvalues of $M_n M_n^*$ (Wishart or sample covariance random matrices).

Distributions of the extremal singular values.

Distribution at the soft-edge of the spectrum. Distribution of the largest singular value (or more generally the joint distribution of the k largest singular values).

Johansson (2000), Johnstone (2000) Gaussian case:

$$\frac{\sigma_n^2 - 4}{2^{4/3} n^{-2/3}} \to TW.$$

Soshnikov (2008): The result holds for all ξ with exponential tail.

Wigner's trace method. For all even k

$$\sigma_1(M)^k + \ldots + \sigma_n(M)^k = \text{Trace } (MM^*)^{k/2}.$$

Notice that if k is large, the left hand side is dominated by the largest term $\sigma_1(M)^k$. Thus, if one can estimate $\mathbf{E}\text{Trace } M^k$ for very large k, one could, in principle, get a good control on $\sigma_1(M)$.

$$\text{Trace } (MM^*)^l := \sum_{i_1,\ldots,i_l} m_{i_1 i_2} m^*_{i_2 i_3} \ldots m_{i_{l-1} i_l} m^*_{i_l i_1}.$$

$$\mathbf{E} m_{i_1 i_2} m^*_{i_2 i_3} \ldots m_{i_{l-1} i_l} m^*_{i_l i_1} = 0$$

unless $i_1 \ldots i_l i_1$ forms a special closed walk in K_n, thanks to the independence of the entries. (Füredi-Komlós, Soshnikov, V., Soshnikov-Peche etc).

Distribution at the hard-edge of the spectrum. Distribution of the least singular value (or more generally the joint distribution of the k smallest singular values).

Edelman (1988) Gaussian case:

Real Gaussian

$$\mathbf{P}(n\sigma_n(M_n(\mathbf{g_R}))^2 \le t) = 1 - e^{-t/2-\sqrt{t}} + o(1).$$

Complex Gaussian

$$\mathbf{P}(n\sigma_n(M_n(\mathbf{g_C}))^2 \le t) = 1 - e^{-t}.$$

Forrester (1994) Joint distribution of the least k singular values.

Ben Arous-Peche (2007) Gaussian divisible random variables.

What about general entries ?

The proofs for Gaussian cases relied on special properties of the Gaussian distribution and cannot be extended.

One can view $\sigma_n(M)$ as the **largest** singular value of M^{-1}. However, the trace method does apply as the entries of M^{-1} are not independent.

Property testing

Given a large, complex, structure S, we would like to study some parameter P of S. It has been observed that quite often one can obtain some good estimates about P by just looking at the small substructure of S, sampled randomly.

In our case, the large structure is our matrix $S := M_n^{-1}$, and the parameter in question is its largest singular value. It has turned out that this largest singular value can be estimated quite precisely (and with high probability) by sampling a few rows (say s) from S and considering the submatrix S' formed by these rows.

Sampling.

Assume, for simplicity, that $|\xi|$ is bounded and M_n is invertible with probability one.

$$\mathbf{P}(n\sigma_n(M_n(\xi))^2 \le t) = \mathbf{P}(\sigma_1(M_n(\xi)^{-1})^2 \ge n/t).$$

Let $R_1(\xi), \ldots, R_n(\xi)$ denote the rows of $M_n(\xi)^{-1}$.

Lemma [Random sampling] Let $1 \le s \le n$ be integers. A be an $n \times n$ real or complex matrix with rows $R_1, \ldots, R_n$. Let $k_1, \ldots, k_s \in \{1, \ldots, n\}$ be selected independently and uniformly at random, and let B be the $s \times n$ matrix with rows $R_{k_1}, \ldots, R_{k_s}$. Then

$$\mathbf{E}\|A^*A - \frac{n}{s}B^*B\|_F^2 \le \frac{n}{s}\sum_{k=1}^{n} |R_k|^4.$$

(special case of Frieze-Kannan-Vempala.)

$R_i = (a_{i1}, \ldots, a_{in})$. For $1 \le i \le j$, the ij entry of $A^*A - \frac{n}{s} B^*B$ is given by

$$\sum_{k=1}^{n} \overline{a_{ki}} a_{kj} - \frac{n}{s} \sum_{l=1}^{s} \overline{a_{k_l i}} a_{k_l j}. \tag{5}$$

For $l = 1, \ldots, s$, the random variables $\overline{a_{k_l i}} a_{k_l j}$ are iid with mean $\frac{1}{n} \sum_{k=1}^{n} \overline{a_{ki}} a_{kj}$ and variance

$$V_{ij} := \frac{1}{n} \sum_{k=1}^{n} |a_{ki}|^2 |a_{kj}|^2 - |\frac{1}{n} \sum_{k=1}^{n} \overline{a_{ki}} a_{kj}|^2, \tag{6}$$

and so the random variable (5) has mean zero and variance $\frac{n^2}{s} V_{ij}$.

Summing over i, j, we conclude that

$$\mathbf{E}\|A^*A - \frac{n}{s}B^*B\|_F^2 = \frac{n^2}{s}\sum_{i=1}^n \sum_{j=1}^n V_{ij}.$$

Discarding the second term in V_{ij}, we conclude

$$\mathbf{E}\|A^*A - \frac{n}{s}B^*B\|_F^2 \leq \frac{n}{s}\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n |a_{ki}|^2 |a_{kj}|^2.$$

Performing the i, j summations, we obtain the claim.

Bounding the error term

The expectation $\mathbf{E}|R_i(\xi)|$ is infinity. However, we have the following tail bound

Lemma. [Tail bound on $|R_i(\xi)|$] Let $R_1, \dots, R_n$ be the rows of $M_n(\xi)^{-1}$. Then

$$\mathbf{P}(\max_{1 \le i \le n} |R_i(\xi)| \ge n^{100/C_0}) \ll n^{-1/C_0}.$$

Inverting and Projecting

One dimensional case. Let A be an invertible matrix with columns $X_1, \ldots, X_n$. Let R_i be the rows of A^{-1}.

Fact. R_1 is the *reciprocal* of the projection of X_1 onto the normal direction of the hyperplane spanned by $X_2, \ldots, X_n$.

Proof. Consider the identity $A^{-1}A = I$. So R_1 is orthogonal with $X_2, \ldots, X_n$ and $R_1 \cdot X_1 = 1$.

Inverting and Projecting, continue

High dimensional case.

Lemma. [Projection lemma] Let V be the s-dimensional subspace formed as the orthogonal complement of the span of $X_{s+1}, \ldots, X_n$, which we identify with F^s (F is either real or complex) via an orthonormal basis, and let $\pi : F^n \to F^s$ be the orthogonal projection to $V \equiv F^s$. Let M be the $s \times s$ matrix with columns $\pi(X_1), \ldots, \pi(X_s)$. Then M is invertible, and we have

$$BB^* = M^{-1}(M^{-1})^*.$$

In particular, we have

$$\sigma_j(B) = \sigma_{s-j+1}(M)^{-1}$$

for all $1 \leq j \leq s$.

Most importantly, this means the **largest** singular value of B is the **smallest** singular value of M.

Together with the Sampling lemma and the Tail bound lemma, this reduces the study of the smallest singular value of an $n \times n$ matrix to that of an $s \times s$ matrix.

The key point of the argument is that the orthogonal projection onto a small dimensional subspace has an *averaging* effect that makes the image close to gaussian.

Similarity Dvoretzky theorem: A low dimensional random cross section of the n-dimensional unit cube looks like a ball with high probability.

One dimensional Berry-Esseen central limit theorem. Let $v_1, \ldots, v_n \in \mathbf{R}$ be real numbers with $v_1^2 + \ldots + v_n^2 = 1$ and let ξ be a $\mathbf{R}$-normalized random variable with finite third moment $\mathbf{E}|\xi|^3 < \infty$. Let $S \in \mathbf{R}$ denote the random variable

$$S = v_1\xi_1 + \ldots + v_n\xi_n$$

where $\xi_1, \ldots, \xi_n$ are iid copies of ξ. Then for any $t \in \mathbf{R}$ we have

$$\mathbf{P}(S \leq t) = \mathbf{P}(\mathbf{g_R} \leq t) + O(\sum_{j=1}^n |v_j|^3),$$

where the implied constant depends on the third moment $\mathbf{E}|\xi|^3$ of ξ. In particular, we have

$$\mathbf{P}(S \leq t) = \mathbf{P}(\mathbf{g_R} \leq t) + O(\max_{1 \leq j \leq n} |v_j|).$$

Morality. Sum of real iid random variables with non-degereated coefficients is asymptotically gaussian.

[Berry-Esséen-type central limit theorem for frames] Let $1 \leq N \leq n$, let F be the real or complex field, and let ξ be F-normalized and have finite third moment $\mathbf{E}|\xi|^3 < \infty$. Let $v_1, \ldots, v_n \in F^N$ be a *normalized tight frame* for F^N, or in other words

$$v_1 v_1^* + \ldots + v_n v_n^* = I_N, \tag{7}$$

where I_N is the identity matrix on F^N. Let $S \in F^N$ denote the random variable

$$S = \xi_1 v_1 + \ldots + \xi_n v_n,$$

where $\xi_1, \ldots, \xi_n$ are iid copies of ξ. Let G be the gaussian counterpart. Then for any measurable set $\Omega \subset F^N$ and any $\epsilon > 0$, one has

$$\mathbf{P}(S \in \Omega) \geq \mathbf{P}(G \in \Omega \backslash \partial_\epsilon \Omega) - O(N^{5/2} \epsilon^{-3} (\max_{1 \leq j \leq n} |v_j|))$$

and

$$\mathbf{P}(S \in \Omega) \leq \mathbf{P}(G \in \Omega \cup \partial_\epsilon \Omega) + O(N^{5/2} \epsilon^{-3} (\max_{1 \leq j \leq n} |v_j|)).$$

Morality. S behave like G on sets with nice boundary.

By Hoffman-Weilandt bound

$$\sum_{i=1}^{s} |\sigma_i(A) - \sigma_i(B)|^2 \le \|A - B\|_F^2.$$

Thus, if one view the matrix as a point in F^{s^2}, the set $\{x|\sigma_n(M(x)) \le t\}$ has nice boundary. So, with proper choice of parameters, $\mathbf{P}(G \in \Omega \backslash \partial_\epsilon \Omega)$ is approximately the same as $\mathbf{P}(\Omega)$. This means

$$\mathbf{P}(n\sigma_n^2(M_n(\xi) \le t) \approx \mathbf{P}(s^2 \sigma_s M_s(\mathbf{g}) \le t)$$

proving the Universality.

Theorem. [Universality for the least singular value](Tao-V. 09) Let ξ be $\mathbf{R}$- or $\mathbf{C}$-normalized, and suppose $\mathbf{E}|\xi|^{C_0} < \infty$ for some sufficiently large absolute constant C_0. Then for all $t > 0$, we have

$$\mathbf{P}(n\sigma_n(M_n(\xi))^2 \leq t) = \int_0^t \frac{1+\sqrt{x}}{2\sqrt{x}} e^{-(x/2+\sqrt{x})}\, dx + O(n^{-c}) \qquad (8)$$

if ξ is $\mathbf{R}$-normalized, and

$$\mathbf{P}(n\sigma_n(M_n(\xi))^2 \leq t) = \int_0^t e^{-x}\, dx + O(n^{-c})$$

if ξ is $\mathbf{C}$-normalized, where $c > 0$ is an absolute constant. The implied constants in the $O(.)$ notation depend on $\mathbf{E}|\xi|^{C_0}$ but are uniform in t.

Conjecture (Spielman-Teng 2002) Let ξ be the Bernoulli random variable. Then there is a constant $0 < b < 1$ such that for all $t \geq 0$

$$\mathbf{P}(\sqrt{n}\sigma_n(M_n(\xi)) \leq t) \leq t + b^n. \tag{9}$$

As $\int_0^t \frac{1+\sqrt{x}}{2\sqrt{x}} e^{-(x/2+\sqrt{x})}\, dx \approx t - t^3/3$, our result implies that this conjecture holds for $t \geq n^{-c}$.

For smaller t, it suggests a stronger bound must hold. (In other words, the term t in the conjectured bound is only the first order approximation of the truth.)

This theorem can be extended in several directions:

- joint distribution of the bottom k singular values of $M_n(\xi)$, for bounded k (and even when k is a small power of n).
- rectangular matrixes where the difference between the two dimensions is not too large.
- all results hold if we drop the condition that the entries have identical distribution. (It is important that they are all normalized, independent and their C_0-moments are uniformly bounded.)

The main technical steps

Tail bound lemma.

Non-degeneracy of normal vectors of a large dimension random subspace.

Berry-Esseen theorem for frames.

The tail bound lemma.

Lemma. [Tail bound on $|R_i(\xi)|$] Let $R_1, \dots, R_n$ be the rows of $M_n(\xi)^{-1}$. Then

$$\mathbf{P}(\max_{1 \le i \le n} |R_i(\xi)| \ge n^{100/C_0}) \ll n^{-1/C_0}.$$

Recall that R_1 is orthogonal to $X_2, \dots, X_n$ and $R_1 \cdot X_1 = 1$, where X_i are the rows of $M_n(\xi)$. Thus, $|R_1|$ is the reciprocal of d_1, the distance from X_1 onto the hyperplane spanned by $X_2, \dots, X_n$. So basically we need to understand the d_i.

It is easy to see that the distance from a random gaussian vector to a random hyperplane has guassian distribution.

It has turned out that this extends to other distributions. (As a toy example, one can consider ± 1 case.)

Lemma. [Random distance is gaussian] Let $X_1, \dots, X_n$ be random vectors whose entries are iid copies of ξ. Then the distribution of the distance d_1 from X_1 to $Span(X_2, \dots, X_n)$ is approximately gaussian, in the sense that

$$\mathbf{P}(d_1 \le t) = \mathbf{P}(|\mathbf{g}_F| \le t) + O(n^{-c}),$$

for some small constant c.

A naive application of the union bound is clearly insufficient, as it gives

$$\mathbf{P}(\min_{1 \le i \le n} d_i \le t) \ll n(t + n^{-c}).$$

The key fact that enables us to overcome the ineffectiveness of the union bound is that the distances d_i are correlated. They tend to be large or small at the same time. Quantitatively, we have

Lemma. [Correlation between distances] Let $n \geq 1$, let F be the real or complex field, let A be an $n \times n$ F-valued invertible matrix with columns $X_1, \dots, X_n$, and let $d_i := dist(X_i, V_i)$ denote the distance from X_i to the hyperplane V_i spanned by $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$. Let $1 \leq L < j \leq n$, let $V_{L,j}$ denote the orthogonal complement of the span of $X_{L+1}, \dots, X_{j-1}, X_{j+1}, \dots, X_n$, and let $\pi_{L,j} : F^n \to V_{L,j}$ denote the orthogonal projection onto $V_{L,j}$. Then

$$d_j \geq \frac{|\pi_{L,j}(X_j)|}{1 + \sum_{i=1}^{L} \frac{|\pi_{L,j}(X_i)|}{d_i}}.$$

Consider

$$distance := |X \cdot v|$$

where $X = (\xi_1, \ldots, \xi_m)$ is the random vector and $v = (a_1, \ldots, a_n)$ is the normal vector of the random hyperplane.

Claim. The normal vector of a random hyperplane, with high probability, looks *normal* (non-degenerate).

Tools: Sharp concentration inequalities.

Then use the one-dimensional Berry-Esseen theorem.