Singularity and determinant of random matrices

Lewis Memorial Lecture / DIMACS minicourse

March 17, 2008

Terence Tao (UCLA)

iid matrix ensembles

Let n be a large integer, and let $M = (a_{ij})_{1 \leq i,j \leq n}$ be a random matrix in which all the coefficients $a_{ij} \equiv a$ are independent and have the same distribution a, which can be either continuous or discrete, and either real or complex. Note that M is not assumed to be symmetric.

[The condition that a_{ij} are identical can be relaxed, but the independence assumption is vital to our current arguments.]

Model examples:

- **Gaussian ensemble** Each $a_{ij} \equiv N(0,1)$ is normally distributed with variance 1.
- **Bernoulli ensemble** Each $a_{ij} \in \{-1, +1\}$ with equal probability of each.
- **Sparse Bernoulli ensemble** Each $a_{ij} \in \{-1, 0, +1\}$ with probability $\mu/2$ of $+1, -1$ and probability $1 - \mu$ of 0.

The Gaussian case enjoys a very strong algebraic structure (with many useful identities); for instance, it is invariant under the action of left or right multiplication by the orthogonal group $O(n)$. Because of this structure, this case is almost completely understood.

These algebraic techniques do not seem to extend to ensembles such as the Bernoulli ensemble. Nevertheless, we expect these ensembles to have analogous behaviour (after we normalise the mean and variance of the underlying distribution a) - i.e. we believe in a "universality principle".

There are two sequences associated to the random matrix M that we wish to study:

- The (generalised) eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbf{C}$ (the roots of $\det(M - \lambda I) = 0$);
- The singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$ (the roots of $\det(MM^* - \sigma^2 I) = 0$).

Many other interesting characteristics of a random matrix are related to these numbers, e.g.

- The determinant $\det(M)$ is equal to $\prod_{i=1}^{n} \lambda_i$, and its magnitude is equal to $\prod_{i=1}^{n} \sigma_i$.
- In particular, the matrix M is invertible or non-singular iff $\det(M) \neq 0$ iff $\lambda_i \neq 0$ for all i, iff $\sigma_n > 0$.
- The trace $\operatorname{tr}(M)$ is equal to $\sum_i \lambda_i$.
- More generally, the moments $\operatorname{tr}(M^k)$ and $\operatorname{tr}((MM^*)^k) = \operatorname{tr}((M^*M)^k)$ are equal to $\sum_i \lambda_i^k$ and $\sum_i \sigma_i^{2k}$ respectively.

- The operator norm $\|M\| := \sup_{\|x\|\leq 1} \|Mx\|$ is equal to σ_1.
- The inverse operator norm $\|M^{-1}\|$ is equal to $1/\sigma_n$.
- The condition number $\|M\|\|M^{-1}\|$ is equal to σ_1/σ_n.
- The resolvent norms $\|(M - zI)^{-1}\|$ are related not to the singular values of M, but to the singular values of $M - zI$.

A basic question in random matrix theory is to estimate the probability distributions of the above quantities for a given random matrix ensemble M. In this lecture we focus on two specific questions:

- What is the singularity probability $\mathbf{P}(\det(M) = 0) = \mathbf{P}(\sigma_n = 0) = \mathbf{P}(\prod_i \lambda_i = 0)$?
- Can we estimate the probability distribution of $\det(M)$?

A popular and effective technique in random matrix theory is the moment method. In this method one first computes moments such as $\mathrm{tr}(M^k)$ and $\mathrm{tr}((MM^*)^k)$ for various values of k. One then uses these computations, together with the identities

$$\mathrm{tr}(M^k) = \sum_i \lambda_i^k; \quad \mathrm{tr}((MM^*)^k) = \sum_i \sigma_i^{2k}$$

to try to control the eigenvalues λ_i and singular values σ_i.

- Example: for the Bernoulli ensemble one has $\mathrm{tr}(MM^*) = n^2$, thus $\sum_i \sigma_i^2 = n^2$, and so the root mean square of σ_i is $\sqrt{n}$.
- By more sophisticated moment methods one can also conclude that $\sigma_1 = O(\sqrt{n})$ with high probability.
- From the elementary inequality $\sup_i |\lambda_i| \leq \|M\| = \sigma_1$ we also conclude that $\lambda_i = O(\sqrt{n})$ for all i w.h.p..
- This method also gives a limiting distribution of the σ_i (but not the λ_i!).

Unfortunately, moment methods are not good at detecting singularity (in which just one of the n eigenvalues or singular values vanishes). If M is well-conditioned (i.e. σ_1/σ_n bounded), then one can use the identity

$$\log|\det M| = \sum_i \log|\lambda_i| = \sum_i \log \sigma_i$$

and the moment method (approximating the log function by polynomials) to understand $\det M$. Unfortunately, for the matrices of interest, M can be quite ill-conditioned (and even singular).

A trivial case: non-singularity of the Gaussian ensemble

As a warm up, let us consider the following trivial result:

Proposition. A Gaussian random matrix M is non-singular with probability 1.

Trivial proof: The set $\{M : \det M = 0\}$ of singular matrices in $M_n(\mathbf{R})$ has positive codimension and thus has measure zero. Since the probability distribution of the Gaussian ensemble in $M_n(\mathbf{R})$ is absolutely continuous, the claim follows. $\square$

The trivial proof does not generalise well to the discrete case. Hence we will need to find a less trivial proof.

Let $X_1, \ldots, X_n \in \mathbf{R}^n$ be the rows of a Gaussian random matrix M, thus $X_1, \ldots, X_n$ are iid Gaussian random vectors. Observe that M is singular if and only if $X_1, \ldots, X_n$ are linearly dependent, or equivalently that X_i lies in the span of $X_1, \ldots, X_{i-1}$ for some i. Thus:

$$\mathbf{P}(\det(M) = 0) \leq \sum_{i=1}^{n} \mathbf{P}(X_i \in V_i)$$

where $V_i := \mathrm{span}(X_1, \ldots, X_{i-1})$.

Slightly less trivial proof: For each $1 \leq i \leq n$, we condition (fix) the vectors $X_1, \ldots, X_{i-1}$. Then the vector space $V_i = \text{span}(X_1, \ldots, X_{i-1})$ is fixed, has positive codimension and thus has measure zero. Since the distribution of X_i is absolutely continuous (and is independent of $X_1, \ldots, X_{i-1}$), we have

$$\mathbf{P}(X_i \in V_i | X_1, \ldots, X_{i-1}) = 0$$

for all $X_1, \ldots, X_{i-1}$.

Integrating, we get

$$\mathbf{P}(X_i \in V_i) = 0$$

and thus

$$\mathbf{P}(\det(M) = 0) \leq \sum_i \mathbf{P}(X_i \in V_i) = 0$$

as desired. $\square$

This proof will discretise more easily than the previous proof, because the geometry of the vector space $V_i \subset \mathbf{R}^n$ is easier to understand than the geometry of the singular set $\{M : \det(M) = 0\} \subset M_n(\mathbf{R})$.

We also see for the first time the very useful conditioning trick to fix some (but not all) of the randomness in the ensemble.

Determinant of the Gaussian ensemble

We now modify the above "less trivial" proof to control the determinant of a Gaussian matrix.

Theorem Let M be a Gaussian matrix. With probability $1 - o(1)$, we have $\log|\det M| \sim n \log n$.

[Much more precise bounds are possible, but we focus on this crude bound for simplicity.]

Once again, we introduce the random rows $X_1, \ldots, X_n \in \mathbf{R}^n$. Observe that $|\det M|$ is the volume of the parallelopiped generated by $X_1, \ldots, X_n$.

Using the elementary "base $\times$ height" formula for parallelopipeds repeatedly, we obtain

$$|\det M| = \prod_{i=1}^{n} \operatorname{dist}(X_i, \operatorname{span}(X_1, \ldots, X_{i-1})) = \prod_{i=1}^{n} \operatorname{dist}(X_i, V_i)$$

and thus

$$\log |\det M| = \sum_{i=1}^{n} \log \operatorname{dist}(X_i, V_i).$$

So we need to understand $\log \operatorname{dist}(X_i, V_i)$ for each i.

As before, we fix i and condition on $X_1, \ldots, X_{i-1}$ to fix V_i. Let $\pi_{V_i^\perp} : \mathbf{R}^n \to V_i^\perp$ be the projection to the orthogonal complement $V_i^\perp$ of V_i, then

$$\operatorname{dist}(X_i, V_i) = |\pi_{V_i^\perp}(X_i)|.$$

But as X_i is an n-dimensional Gaussian vector (with each coordinate $\equiv N(0,1)$), $\pi_{V_i^\perp}(X_i)$ is an $n-i+1$-dimensional Gaussian vector (with each coordinate $\equiv N(0,1)$).

In particular, we have

$$\log \operatorname{dist}(X_i, V_i) = \log |\pi_{V_i^\perp}(X_i)| \sim \log \sqrt{n-i}$$

with high probability (e.g. $1 - O(n^{-10})$), as long as $n - i$ is reasonably large (e.g. $n - i > 100 \log n$). In the remaining cases $i = n - O(\log n)$, we can use a crude bound such as

$$|\log \operatorname{dist}(X_i, V_i)| \leq 100 \log n$$

with probability $1 - O(n^{-10})$.

Summing this we have

$$\begin{aligned}\log|\det(M)| &= \sum_{i=1}^{n} \log \operatorname{dist}(X_i, V_i) \\ &\sim \sum_{i=1}^{n-100\log n} \log\sqrt{n-i} \\ &\quad + \sum_{i=n-100\log n}^{n} O(\log n) \\ &\sim n\log n\end{aligned}$$

with probability $1-o(1)$, as required. $\square$

Singularity of the Bernoulli ensemble

The singularity probability of a Bernoulli matrix is non-zero (e.g. two rows could be identical). We in fact have

Conjecture. A Bernoulli random matrix M is singular with probability $(\frac{1}{2} + o(1))^n$.

In 1967, Komlós showed:

Proposition. A Bernoulli random matrix M is singular with probability $o(1)$.

We prove this by an argument of Vu. We again consider the random rows $X_1, \ldots, X_n \in \mathbf{R}^n$ and the spaces $V_i := \mathrm{span}(X_1, \ldots, X_{i-1})$. This time, though, the X_i are not Gaussian, but are instead distributed uniformly in the discrete cube $\{-1, +1\}^n$.

Nevertheless, we still have the inequality

$$\mathbf{P}(\det(M) = 0) \leq \sum_{i=1}^{n} \mathbf{P}(X_i \in V_i).$$

In the low dimensional case $i \neq n - O(1)$, we can use the following simple bound:

Odlyzko bound: (1988) $\mathbf{P}(X_i \in V_i) \leq 2^{-(n-i+1)}$.

Proof: We condition on $X_1, \ldots, X_{i-1}$ to fix V_i, which has dimension at most $i-1$. Thus, there exist $i-1$ coordinates $a_{j_1}, \ldots, a_{j_{i-1}}$ which determine all the other $n-i+1$ coordinates of a vector $a = (a_1, \ldots, a_n) \in V_i$. Thus, if we condition on the coordinates $a_{j_1}, \ldots, a_{j_{i-1}}$ of X_i, the probability $X_i \in V_i$ is at most $2^{-(n-i+1)}$ as desired. $\square$

It remains to show that $\mathbf{P}(X_i \in V_i) = o(1)$ when $i = n - O(1)$. The hardest case is $i = n$.

We condition to fix V_n, and let v be a unit normal vector to V_n. Then

$$\mathbf{P}(X_n \in V_n) \leq \mathbf{P}(X_n \cdot v = 0).$$

Writing $X_n = (a_1, \ldots, a_n)$ and $v = (v_1, \ldots, v_n)$, we need to understand the collision probability

$$\mathbf{P}(a_1 v_1 + \ldots + a_n v_n = 0)$$

for the random walk $\pm v_1 \pm v_2 \ldots \pm v_n$.

The task of computing this collision probability is known as the Littlewood-Offord problem. A basic bound is:

Erdős' Littlewood-Offord inequality (1945) Suppose at least k of the $v_1, \ldots, v_n$ are non-zero. Then

$$\mathbf{P}(a_1 v_1 + \ldots + a_n v_n = 0) \ll \frac{1}{\sqrt{k}}.$$

Notice that this bound is tight if all the non-zero coefficients of v_i are of equal magnitude.

So we fix a parameter $1 \lll k \lll n$ and divide into two cases, depending on the sparsity of the fixed vector v.

- **Incompressible case:** At least k coefficients of v are non-zero. Then from Erdős' Littlewood-Offord inequality, $\mathbf{P}(X_n \in V_n) \leq 1/\sqrt{k}$, which is OK as long as $k \ggg 1$.
- **Compressible case:** Fewer than k coefficients of v are non-zero. But (as v is orthogonal to $X_1, \ldots, X_n$) this means that k of the columns $Y_1, \ldots, Y_n$ of M are linearly dependent. Direct counting shows that the probability of this is exponentially small in n if k is not too large (e.g. $k = O(\log \log n)$ is enough).

Putting all the above arguments together gives the bound

$$\mathbf{P}(\sigma_n = 0) = \mathbf{P}(\det(M) = 0) = o(1)$$

as desired. □

In the next lecture, we will see that a similar strategy also allows us to obtain a non-trivial lower bound on the singular value σ_n.

Determinant of the Bernoulli ensemble

A modification of the above argument also lets us control the determinant:

Theorem (T.-Vu 2005) Let M be a Bernoulli matrix. With probability $1 - o(1)$, we have $\log |\det M| = (1 - o(1))\frac{1}{2} n \log n$.

[More precise bounds are available.]

Some context for this bound:

- By symmetry of the Bernoulli distribution, $\det M$ is distributed symmetrically around the origin.
- The moment method gives $\mathbf{E} \det M^2 = n! = \exp(n \log n - O(n))$ (Turán, 1940). Higher moments can be computed but rapidly become difficult (and not too useful).
- We have the crude bound $|\det M| \leq |X_1| \ldots |X_n| = n^{n/2} = \exp(\frac{1}{2} n \log n)$, which is attained for Hadamard matrices.
- By row reduction, $\det M$ is an integer multiple of 2^{n-1}. The mod p distribution is also understood.

Sketch of proof:

- As before, we rely on the identity

$$\log\det M = \sum_{i=1}^{n} \log \operatorname{dist}(X_i, V_i) = \sum_{i=1}^{n} \log |\pi_{V_i^\perp} X_i|.$$

- In the low-dimensional case $i \le n - O(\log n)$, the vector $\pi_{V_i^\perp} X_i$ is no longer Gaussian. Nevertheless, $\operatorname{dist}(X_i, V_i)$ is still a Lipschitz function of X_i, and one can use Talagrand's inequality (1996) to show that $\log \operatorname{dist}(X_i, V_i)$ concentrates near $\log \sqrt{n-i}$ as in the Gaussian case.

- In the high dimensional case $i > n - O(\log n)$, one again has to consider a unit normal vector v to V_i, thus $\text{dist}(X_i, V_i) \geq |X_i \cdot v|$. We split into two cases.
- In the **compressed case** in which almost all the coefficients of v are very small (and so a few columns $Y_1, \ldots, Y_n$ are almost dependent), one can use direct counting arguments to show that this case is very unlikely.
- In the **uncompressed case** in which there are many coefficients that are not too small, one can use a version of Erdős' Littlewood-Offord bound to obtain a crude lower bound on $|X_i \cdot v|$. $\square$

Exponential bounds

The above arguments, combined with more advanced Littlewood-Offord (and inverse Littlewood-Offord) theorems, eventually give $\mathbf{P}(\det(M) = 0) = O_A(n^{-A})$ for any A. (More on this in the next lecture.) But one can do even better:

Theorem (Kahn-Komlós-Szemerédi 1995) For the Bernoulli ensemble, we have $\mathbf{P}(\det(M) = 0) \leq (c + o(1))^n$ for some $0 < c < 1$.

$c \leq 0.999$	Kahn-Komlós-Szemerédi (1995)
$c \leq 0.939\ldots$	T.-Vu (2005)
$c \leq \frac{3}{4} = 0.75$	T.-Vu (2006)
$c \leq \frac{1}{\sqrt{2}} = 0.707\ldots$	Bourgain-Vu-Wood (2008)
$c = \frac{1}{2}$?	Conjectured

Similar results are known for more general ensembles (Rudelson-Vershynin, 2007; Bourgain-Vu-Wood, 2008). For certain discrete ensembles, the optimal value of c is attained (Bourgain-Vu-Wood, 2008).

Now we sketch a proof. Instead of using the inequality

$$\mathbf{P}(\det(M) = 0) \leq \sum_{i=1}^{n} \mathbf{P}(X_i \in V_i)$$

as before, it is more convenient to use the variant

$$\mathbf{P}(\det(M) = 0) \leq \sum_{V} \mathbf{P}(X_1, \ldots, X_n \text{ span } V)$$

where V ranges over proper subspaces. Note that

$$\mathbf{P}(X_1, \ldots, X_n \text{ span } V) \leq \mathbf{P}(X \in V)^n$$

where X is uniformly distributed in $\{-1, +1\}^n$.

- If the space V is poor (so $\mathbf{P}(X \in V) \leq c^n$), a conditioning argument lets one bound this case easily (using $n-1$ of the vectors $X_1, \ldots, X_n$ to span V).
- If the space V is very rich (so $\mathbf{P}(X \in V) > o(1)$), the Erdős Littlewood-Offord inequality lets us conclude that a normal vector of V is compressed, and counting arguments let us bound this case easily.
- The difficult case is when V is only somewhat rich: $c^n \leq \mathbf{P}(X \in V) \leq o(1)$.

The key new idea here is a swapping argument: one finds another random vector Y with the concentration property

$$\mathbf{P}(X \in V) \leq c\mathbf{P}(Y \in V).$$

Morally speaking, this implies that

$$\mathbf{P}(X_1, \ldots, X_n \text{ span } V) \leq c^n \mathbf{P}(Y_1, \ldots, Y_n \text{ span } V)$$

and so

$$\mathbf{P}(\det(M) = 0) \leq c^n \mathbf{P}(\det(M') = 0) \leq c^n$$

where M' is formed using the rows $Y_1, \ldots, Y_n$. (The precise swapping argument is more technical.)

One good choice of random vector $Y = (b_1, \ldots, b_n)$ is a sparse Bernoulli vector, in which each b_i equals ± 1 with probability $\mu/2$ and 0 with probability $1 - \mu$.

Intuitively, this vector has more zeroes in it and so is more likely to lie in the vector space V. (For instance, it is identically zero with probability $(1 - \mu)^n$.)

To prove $\mathbf{P}(X \in V) \leq c\mathbf{P}(Y \in V)$, the main task is to show that

$$\mathbf{P}(X \cdot v = 0) \leq c\mathbf{P}(Y \cdot v = 0)$$

for certain unit vectors v. "Lazy random walk is more concentrated than random walk".

Key ingredients: (Halász, 1975)

- The Fourier identities

$$\mathbf{P}(X \cdot v = 0) = \int_{\mathbf{R}/\mathbf{Z}} \prod_{i=1}^{n} \cos(2\pi v_i t)\ dt$$

$$\mathbf{P}(Y \cdot v = 0) = \int_{\mathbf{R}/\mathbf{Z}} \prod_{i=1}^{n} (1-\mu) + \mu \cos(4\pi v_i t)\ dt.$$

- The elementary inequalities

$$|\cos(\theta)| \leq [(1-\mu) + \mu \cos(2\theta)]^{\sigma}$$

$$|\cos(\alpha)||\cos(\beta)| \leq [(1-\mu) + \mu \cos(2(\alpha+\beta))]^{2\sigma}$$

for various $1/4 < \mu < 1$ and $\sigma > 1$.

- The Mann-Kneser-Macbeath inequality (1953)

$$\mathrm{mes}(A+B) \geq \min(\mathrm{mes}(A)+\mathrm{mes}(B), 1)$$

from additive combinatorics, where $A, B \subset \mathbf{R}/\mathbf{Z}$.

This is already enough to get $c \leq 0.939\ldots$. To get $c \leq 3/4$ or $c \leq 1/\sqrt{2}$ requires more advanced inverse theorems of Freiman type from additive combinatorics, in order to classify the exceptional vectors v in which $\mathbf{P}(X \in V)$ is not extremely small compared to $\mathbf{P}(Y \in V)$.