Random Matrices: Invertibility, Structure, and Applications

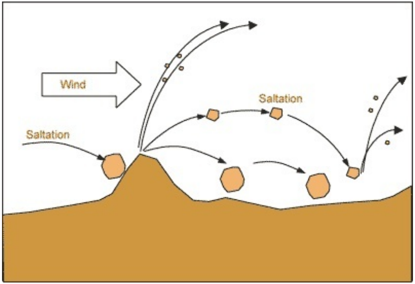
Roman Vershynin

University of Michigan

2011 Canadian Mathematical Society Summer Meeting
June 3, University of Alberta, Edmonton

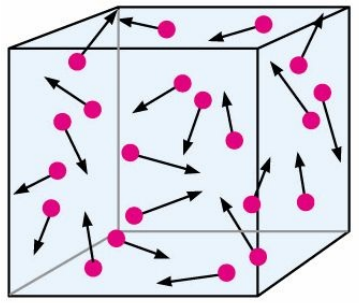
Chaos and Order

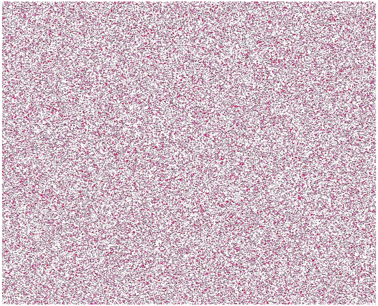
Many complex systems that occur in nature and society exhibit chaos on the **microscopic** level and order on the **macroscopic** level.



Chaos and Order

Gas molecules:





Statistical mechanics: randomness at the microscopic level averages out at the macroscopic level.

Probability Theory

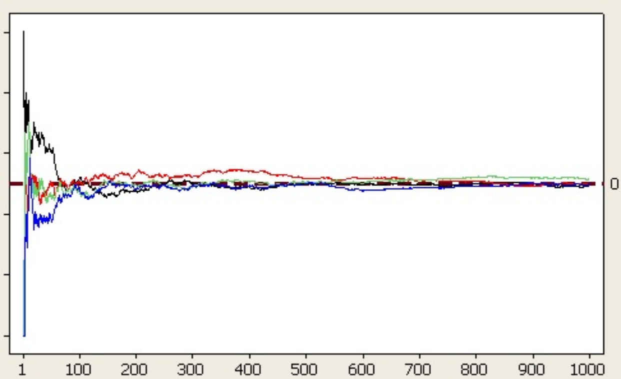
- **Microscopic:** independent random variables $X_1, X_2, \dots$
- **Macroscopic:** function $f(X_1, \dots, X_n)$ where n is large.
- Example: Bernoulli r.v's $X_i = \pm 1$ with probabilities $\frac{1}{2}$.
  At each game, gain \$1 or lose \$1 independently.
  Macroscopic quantity: average gain

$$f(X_1, \dots, X_n) = \frac{X_1 + \cdots + X_n}{n}.$$

Probability Theory

Limit theorems describe the macroscopic picture as $n \to \infty$.

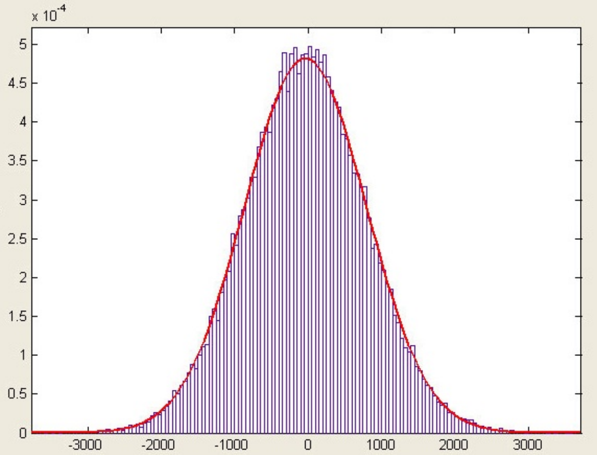
Law of Large Numbers:



$$\frac{X_1 + \cdots + X_n}{n} \to 0 \quad \text{almost surely}$$

Probability Theory

Central Limit Theorem:



$$X_1 + \cdots + X_n \approx N(0, \sqrt{n}) \quad \text{in distribution}$$

Probability Theory

- **Microscopic:** independent random variables $X_1, X_2, \ldots$
- **Macroscopic:** function $f(X_1, \ldots, X_n)$.
- Functions may be **more complex** than the sum $X_1 + \cdots + X_n$.
- Example: **random matrix theory**.

Random Matrix Theory

- **Microscopic:** independent random variables X_{ij}, arranged in a matrix

$$H = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \multicolumn{4}{c}{\dotfill} \\ X_{n1} & X_{n2} & \cdots & X_{nn} \end{pmatrix}$$

- **Macroscopic:** the eigenvalues of H

$$\lambda_1(H), \ldots, \lambda_n(H).$$

Random Matrix Theory

One can make H **symmetric** by placing independent rv's above the diagonal and reflecting:

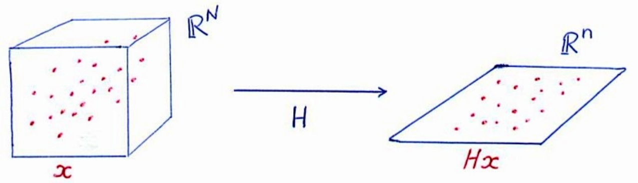
$$X_{ij} = X_{ji}$$

This is a Wigner random matrix:

$$H = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{12} & X_{22} & \cdots & X_{2n} \\ \hdotsfor{4} \\ X_{1n} & X_{2n} & \cdots & X_{nn} \end{pmatrix}$$

Why Random Matrices?

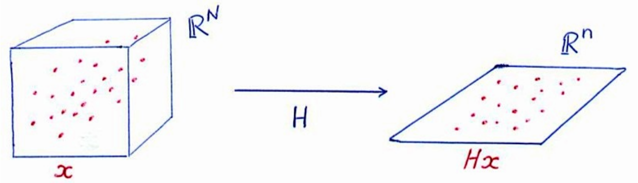
- Computer Science, Information Theory (1990's+): random matrices provide a mechanism for **dimension reduction**.
- Data points $x \in \mathbb{R}^N$ (high dimension) need to be mapped into $\mathbb{R}^n$ (low dimension) while preserving the essential information in the data.



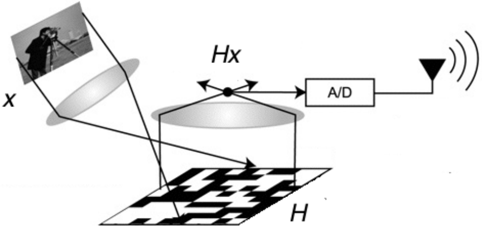
- Use a random linear transformation, given by an $n \times N$ **random matrix** H with independent entries.

Johnson-Lindenstrauss Lemma '84: *Given m data points in $\mathbb{R}^N$, one can reduce the dimension to $n \sim \log m$ while approximately preserving all pairwise distances between the points.*

Why Random Matrices?



Compressed Sensing (2004+): allows one to exactly **recover the data** $x \in \mathbb{R}^N$ from its random measurement $Hx \in \mathbb{R}^n$, provided the data x has "low information content", i.e. x is a **sparse** vector. In polynomial time.



Why Random Matrices?

- Numerical Analysis [Von Neumann et al. 40's]: analysis of algorithms for solving large **linear equations**

$$Ax = b.$$

- Use a **random matrix** A to test the quality (speed and accuracy) of a linear solver.
- Here one models a **"typical"** input A of an algorithm as a **random** input. Average analysis of algorithms.
- Many algorithms perform better when A is well conditioned, i.e. the condition number

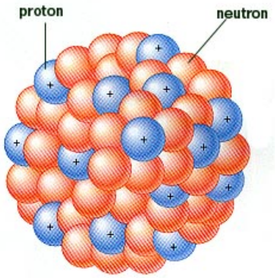
$$\kappa(A) = \|A\| \|A^{-1}\|$$

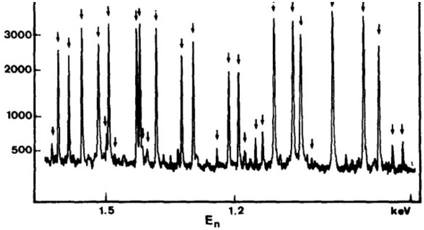
is not too large.

- Question: *Are random matrices well conditioned?*

Why Random Matrices?

- Physics: Excitation spectrum of heavy nuclei, e.g. U_{238}. **Excitation spectrum** = the energy levels for which a neutron will bounce off the nucleus (scattering resonances).





- Protons and neutrons in the nucleus of U_{238} interact with each other in a complicated way. The Hamiltonian is too complex. Its spectrum is **difficult to compute** either theoretically or by simulation.

Why Random Matrices?

- Wigner 50's: One models the complicated Hamiltonian as an $n \times n$ **symmetric random matrix**

$$H = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{12} & X_{22} & \cdots & X_{2n} \\ \multicolumn{4}{c}{\dotfill} \\ X_{1n} & X_{2n} & \cdots & X_{nn} \end{pmatrix}$$

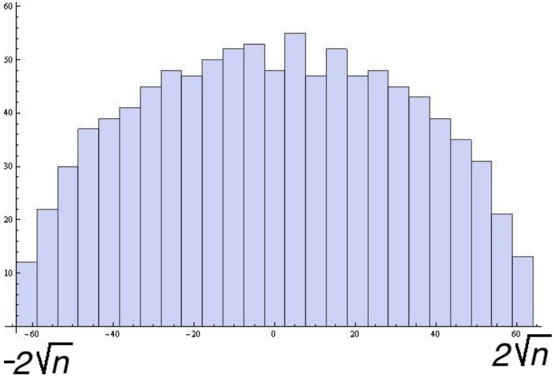
- The excitation spectrum = the eigenvalues

$$\lambda_1(H), \ldots, \lambda_n(H).$$

- The **distribution of the eigenvalues** now becomes computable. So, what is it?

Semicircle Law

The **histogram** of the eigenvalues of a 1000 × 1000 symmetric matrix with independent $N(0,1)$ entries:



Benedek Valkó's course on random matrices `http://www.math.wisc.edu/~valko/courses/833/833.html`

After rescaling...

Semicircle Law

Semicircle law [Wigner '55]: *Let H be a symmetric random matrix with $N(0,1)$ entries. Then the eigenvalue histogram of $\frac{1}{\sqrt{n}}H$ (i.e. the "empirical spectral distribution") converges to the semi-circle supported in $[-2,2]$.*

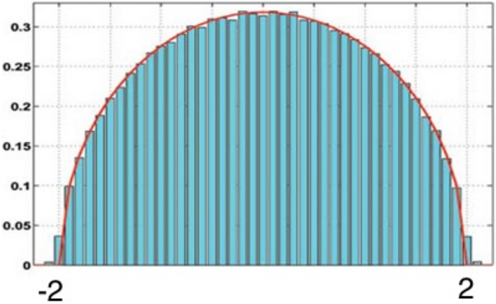
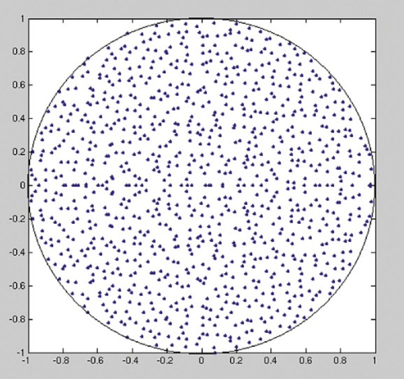


Image by Alan Edelman, MIT open courseware 18.996 / 16.399 Random Matrix Theory and Its Applications

Circular Law

Circular law [Mehta '67]: *Let H be a random matrix with all independent $N(0,1)$ entries. Then the empirical spectral distribution of $\frac{1}{\sqrt{n}}H$ converges to the uniform measure on the unit disc in $\mathbb{C}$.*



Universality

- The limit laws of random matrix theory (semicircle, circular) are **the same for different distributions** of entries X_{ij}, e.g. normal $N(0,1)$, Bernoulli ± 1 etc.
- **Microscopic** laws may be different (and even unknown), but **macroscopic** picture is the same. Importance: one can replace the unknown distribution by normal.

- The same phenomenon as in the Central Limit Theorem:

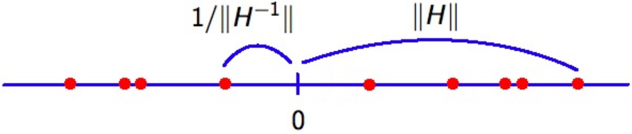
$$X_1 + \cdots + X_n \approx N(0, \sqrt{n}).$$

  The same limit **regardless of the distribution** of X_i.
- For semicircle law, universality was proved by [Pastur'73], see [Bai-Silverstein'10]. For circular law, universality was established by [Girko'84, Edelman'97, Bai'97, Götze-Tikhomirov'07, Pan-Zhou'07, Tao-Vu'07-08].

Local Regime

- The limit laws are **global**; they state something for the **bulk** of the eigenvalues (say, for 10% or 1% of eigenvalues).
- Where are **individual eigenvalues**? Local regime.

  There is extensive recent work, with many questions answered [Tao-Vu'05+, Rudelson-V'07+, V, L. Erdös-Schlein-Yau'08+].
- Why local regime? The eigenvalue **nearest** 0 determines the **invertibility** properties of H. The eigenvalue farthest from 0 determines the operator norm of H:



- If there is an eigenvalue at 0, then H is **singular**. Otherwise H has **full rank**.
- The limit laws do not preclude **one** eigenvalue to stick to 0 almost surely.

Invertibility

Invertibility Problem: *Are random matrices H likely singular or full rank?*

- Answer: likely to have full rank.
- 1. For $n \times n$ matrices with **all independent entries**.

Conjecture [P. Erdös]: *For Bernoulli matrices with ± 1 entries,*

$$\mathbb{P}\{H \text{ is singular}\} = \Big(\frac{1}{2} + o(1)\Big)^n$$

$\approx \mathbb{P}\{$*two rows or two columns of H are equal up to a sign*$\}$.

- Best known result: $\big(\frac{1}{\sqrt{2}} + o(1)\big)^n$ [Bourgain-Wood-Vu'10].
- For **general distributions** of entries, one still has [Rudelson-V'08]:

$$\mathbb{P}\{H \text{ is singular}\} \leq \exp(-cn).$$

Invertibility

- 2. For **symmetric matrices**, the invertibility conjecture is the same. For **Bernoulli** symmetric matrices with ±1 entries,

$$\mathbb{P}\{H \text{ is singular}\} = \Big(\frac{1}{2} + o(1)\Big)^n ?$$

- Best known result [V'11]:

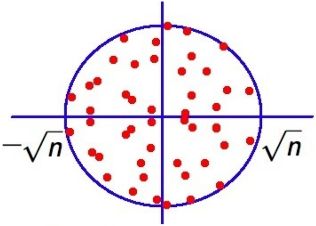
$$\mathbb{P}\{H \text{ is singular}\} \le \exp(-n^c).$$

This also holds for **general** distributions of entries.

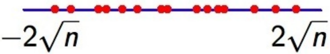
Delocalization

More general phenomenon:

The spectrum of a random matrix H is **delocalized**.



All independent entries

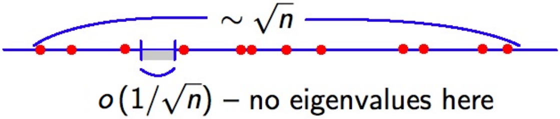


Symmetric

- 1. Eigenvalues of H **do not stick to any particular point.** The probability that the spectrum **hits** a particular point is $\exp(-cn)$ for matrices H with all independent entries [Rudelson-V'08].
- Similarly for symmetric matrices H: $\exp(-n^c)$ [V'11].

Delocalization

- 2. Moreover, the eigenvalues of H **do not stick to small intervals**.
- The specturm of a symmetric random matrix **misses** any fixed interval smaller than the **average eigenvalue gap** (which is $1/\sqrt{n}$). [Erdös-Schlein-Yau, Tao-Vu, V'11].



- In particular, eigenvalues are **separated from** 0 by $1/\sqrt{n}$. So

$$\|H^{-1}\| = O(\sqrt{n}), \qquad \|H\| = O(\sqrt{n}).$$

- Therefore **the condition number is linear** in n:

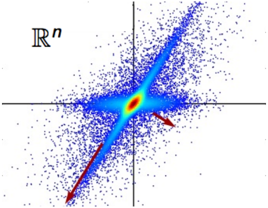
$$\kappa(H) = \|H\|\|H^{-1}\| = O(n).$$

Same if H has all independent entries [Rudelson-V'08].

- Thus: **Random matrices are well conditioned**.
  This addresses a problem of Von Neumann et al. 40's.

Random Matrices in Statistics: Covariance Estimation

- Statistics: Principal Component Analysis (PCA): determine the axes along which **most correlation occurs**. This is the **covariance structure** of the distribution.

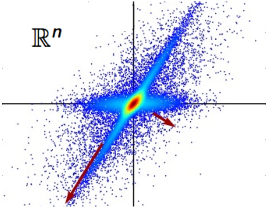


$$H = \begin{bmatrix} — X_1 — \\ — X_2 — \\ \cdots\cdots \\ — X_N — \end{bmatrix}$$

- We sample a few **data points** $X_1, \ldots, X_N \in \mathbb{R}^n$ independently from the distribution. We organize them as an $N \times n$ random matrix H with **independent rows.** Warning: not independent **entries**!
- Compute the $n \times n$ matrix $H^T H$, the Wishart random matrix. Its eigenvectors are the **principal components**.

Problem: *How many sample points $N = N(n)$ are needed to estimate the covariance structure of a distribution in $\mathbb{R}^n$?*

Random Matrices in Statistics: Covariance Estimation



$$H = \begin{bmatrix} - X_1 - \\ - X_2 - \\ \cdots\cdots \\ - X_N - \end{bmatrix}$$

- A different look at the Wishart matrix:

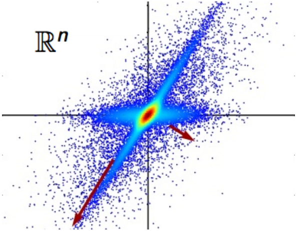
$$\Sigma_N = \frac{1}{N} H^T H = \frac{1}{N} \sum_{i=1}^{N} X_i X_i^T$$

This is the sample covariance matrix, $n \times n$ symmetric random matrix.

- Our hope: Σ_N is a good estimate for the population covariance matrix

$$\Sigma = \mathbb{E}\, X_i X_i^T.$$

Random Matrices in Statistics: Covariance Estimation



Sample and population covariance matrices:

$$\Sigma_N = \frac{1}{N}\sum_{i=1}^{N} X_i X_i^T, \qquad \Sigma = \mathbb{E}\, X_i X_i^T.$$

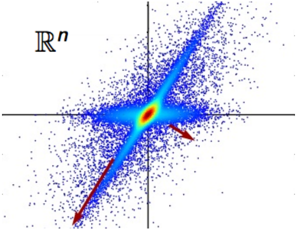
- Key: Σ_N is a **sum of independent random matrices** $X_i X_i^T$.
- Law of Large Numbers in higher dimensions implies:

$$\Sigma_N \to \Sigma \quad \text{as } N \to \infty, \quad n \text{ fixed.}$$

- But we need a **small** sample size N!
- What is the **smallest sample size** $N = N(n)$ so that $\Sigma_N \approx \Sigma$?
- $N \geq n$ is needed (for the full rank).[1]

[1] For structured data, one can have $N \ll n$, see e.g. [Levina-V'10].

Random Matrices in Statistics: Covariance Estimation



Sample and population covariance matrices:

$$\Sigma_N = \frac{1}{N} \sum_{i=1}^{N} X_i X_i^T, \qquad \Sigma = \mathbb{E}\, X_i X_i^T.$$

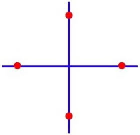
- Use **quantitative form of Law of Large Numbers** – classical deviation inequalities for sums of independent random variables (Khinchine, Bernstein, Chernoff, . . .)
- For matrices, one uses **non-commutative** versions of deviation inequalities. One obtains (for general distributions!) that

$$N = O(n \log n)$$

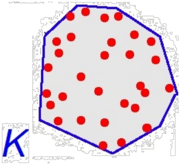
suffices for $\Sigma_N \approx \Sigma$ in the operator norm. [Rudelson'99]

Random Matrices in Statistics: Covariance Estimation

- $N = O(n \log n)$ sample points always suffice.
- In general, $\log n$ oversampling is needed (for very discrete distributions).



Problem [Kannan-Lovasz-Simonovits'97]:
$N = O(n)$ sample points should suffice for covariance estimation of the uniform distribution in an arbitrary **convex set** *K in $\mathbb{R}^n$.*



Important for **volume estimation** of K.

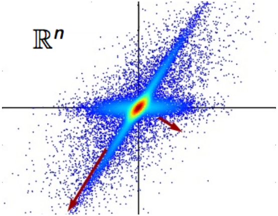
Theorem [Adamczak-Litvak-Pajor-Tomczak'09]: (UofA)
The KLS Conjecture is true.

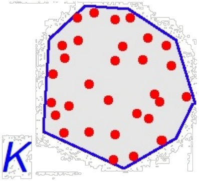
Conjecture [V'10]: $N = O(n)$ suffices for **most** distributions.

Random Matrices in Statistics: Covariance Estimation

Theorem [Srivastava-V'11]: $N = O(n)$ sample points suffice for covariance estimation for all distributions satisfying mild **regularity** assumptions.

In particular, this holds for **convex sets**, yielding an alternative approach to KLS problem.

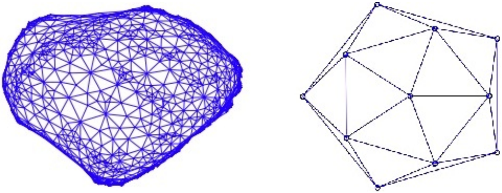




Regularity assumption: $2 + \varepsilon$ moments of k-dimensional marginals outside the ball of radius $O(\sqrt{k})$.

Covariance Estimation and the Spectral Sparsifier

- The new method: randomizing the spectral sparsifier of [Batson-Spielman-Srivastava'08].
- Spectral sparsification is a deterministic method that allows one to approximate a given **dense graph** by a **sparse graph:**



Daniel Spielman FOCS'07 tutorial on spectral graph theory

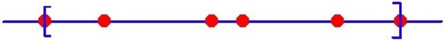
- Randomization makes the spectral sparsifier appear as a natural **method in Random Matrix Theory**.

Covariance Estimation and the Spectral Sparsifier

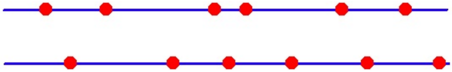
- Goal: Control the whole **spectrum** of the Wishart matrix

$$W = \sum_{i=1}^{N} X_i X_i^T,$$

i.e. the left and right **spectral edges**:



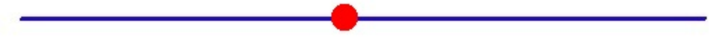
- Method: Add $X_i X_i^T$ **one at a time**, and keep track how the spectrum of W evolves.
- Eigenvalues interlace (**Cauchy interlacing theorem**):



Covariance Estimation and the Spectral Sparsifier

Evolution of the spectrum of the Wishart matrix in $\mathbb{R}^n$:

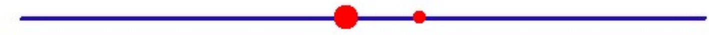
$$W = \sum_{i=1}^{N} X_i X_i^T \qquad N = 1, 2, \ldots$$



Covariance Estimation and the Spectral Sparsifier

Evolution of the spectrum of the Wishart matrix in $\mathbb{R}^n$:

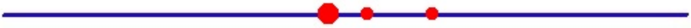
$$W = \sum_{i=1}^{N} X_i X_i^T \qquad N = 1, 2, \ldots$$



Covariance Estimation and the Spectral Sparsifier

Evolution of the spectrum of the Wishart matrix in $\mathbb{R}^n$:

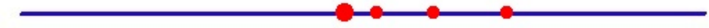
$$W = \sum_{i=1}^{N} X_i X_i^T \qquad N = 1, 2, \ldots$$



Covariance Estimation and the Spectral Sparsifier

Evolution of the spectrum of the Wishart matrix in $\mathbb{R}^n$:

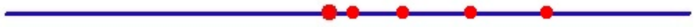
$$W = \sum_{i=1}^{N} X_i X_i^T \qquad N = 1, 2, \ldots$$



Covariance Estimation and the Spectral Sparsifier

Evolution of the spectrum of the Wishart matrix in $\mathbb{R}^n$:

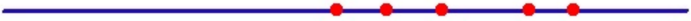
$$W = \sum_{i=1}^{N} X_i X_i^T \qquad N = 1, 2, \ldots$$



Covariance Estimation and the Spectral Sparsifier

Evolution of the spectrum of the Wishart matrix in $\mathbb{R}^n$:

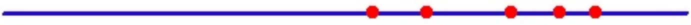
$$W = \sum_{i=1}^{N} X_i X_i^T \qquad N = 1, 2, \ldots$$



Covariance Estimation and the Spectral Sparsifier

Evolution of the spectrum of the Wishart matrix in $\mathbb{R}^n$:

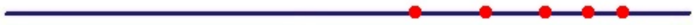
$$W = \sum_{i=1}^{N} X_i X_i^T \qquad N = 1, 2, \ldots$$



Covariance Estimation and the Spectral Sparsifier

Evolution of the spectrum of the Wishart matrix in $\mathbb{R}^n$:

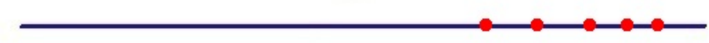
$$W = \sum_{i=1}^{N} X_i X_i^T \qquad N = 1, 2, \ldots$$



Covariance Estimation and the Spectral Sparsifier

Evolution of the spectrum of the Wishart matrix in $\mathbb{R}^n$:

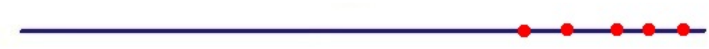
$$W = \sum_{i=1}^{N} X_i X_i^T \qquad N = 1, 2, \ldots$$



Covariance Estimation and the Spectral Sparsifier

Evolution of the spectrum of the Wishart matrix in $\mathbb{R}^n$:

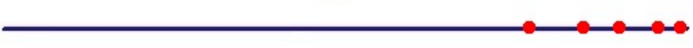
$$W = \sum_{i=1}^{N} X_i X_i^T \qquad N = 1, 2, \dots$$



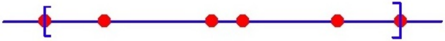
Covariance Estimation and the Spectral Sparsifier

Evolution of the spectrum of the Wishart matrix in $\mathbb{R}^n$:

$$W = \sum_{i=1}^{N} X_i X_i^T \qquad N = 1, 2, \dots$$



Covariance Estimation and the Spectral Sparsifier



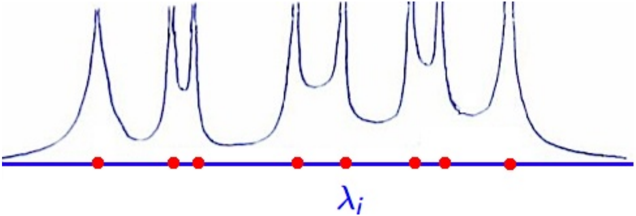
- Difficulty: The **spectral edges** (the extreme eigenvalues) are not controlled by interlacing, they are **free** on one side. They are difficult to compute.
- Solution: **Soften** the spectral edges:

Covariance Estimation via Stieltjes Transform

- **Stieltjes Transform** of the spectrum of W is the function

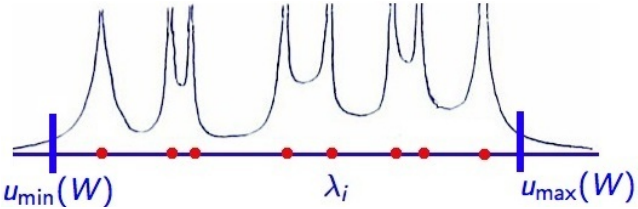
$$m_W(u) = \mathrm{trace}(uI - W)^{-1} = \sum_{i=1}^{N} \frac{1}{u - \lambda_i} \qquad u \in \mathbb{R}.$$

- Ignoring the sign, $m_W(u)$ looks like this:



- Physical interpretation: Put unit **electric charges** at points λ_i. The **electric potential** measured at u equals $m_W(u)$.

Covariance Estimation via Stieltjes Transform

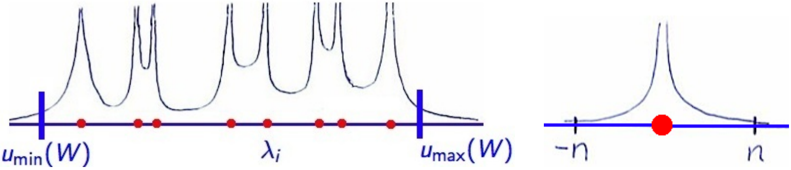


- Find the leftmost/rightmost locations $u_{\min}(W)$, $u_{\max}(W)$ where the electric **potential** is some fixed **constant**:

$$m_W(u) = \phi \qquad \text{(say, } \phi = 1000\text{).}$$

- These locations are **soft** proxies of the **spectral edges**. They "harden" as $\phi \to \infty$.

Covariance Estimation via Stieltjes Transform



- Key: As opposed to the usual spectral edges, the soft edges $u_{\min}(W), u_{\max}(W)$ **are computable**.
- Why? They are determined by the Stieltjes transform of $W = \sum_{i=1}^N X_i X_i^T$, which can be **recomputed** by adding one term at a time. (Sherman-Morrison formula).[2]
- One shows that the proxies **increase by** $1 \pm o(1)$ **at every step**.
- After N steps, they are $\approx N \pm n$. QED.

[2] For $W = V + XX^T$, one has $m_W(u) = m_V(u) + \frac{X^T(uI-V)^{-2}X}{1-X^T(uI-V)^{-1}X}$

References

- Tutorial: R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, 2010.
- Survey: M. Rudelson, R. Vershynin, *Non-asymptotic theory of random matrices: extreme singular values*, 2010.
- Invertibility of Symmetric Matrices: R. Vershynin, *Invertibility of symmetric random matrices*, 2011.
- Covariance Estimation: N. Srivastava, R. Vershynin, *Covariance estimation for distributions with* $2+\varepsilon$ *moments*, 2011 (TBA).