Plan and Bibliography

INTRODUCTION TO THE NON-ASYMPTOTIC ANALYSIS OF RANDOM MATRICES

ROMAN VERSHYNIN

Spring-Summer School "Random matrices - Stochastic geometry - Compressed sensing"
IHP, Paris, June 20–22, 2011

This 9-hour course develops some non-asymptotic methods for the analysis of the extreme singular values of random matrices with independent rows. Two applications are given, for the problem of estimating covariance matrices in statistics, and for validating probabilistic constructions of measurement matrices in compressed sensing. Much of the material is taken from the tutorial [V]. Other references are given in the text.

Contents

1. Random matrices and random vectors 2
1.1. Marchenko-Pastur Law 2
1.2. Singular values and operator norms 3
1.3. Bai-Yin Law 3
1.4. Goals, methods 3
1.5. Isotropic random vectors 3
1.6. Gordon's theorem 4
2. Sums of independent random variables 4
2.1. Sub-gaussian random variables 4
2.2. Sub-exponential random variables 4
3. Sums of sub-exponential random matrices 4
3.1. Ahlswede-Winter's method 5
3.2. Rudelson's inequality 6
4. Matrices with independent sub-gaussian rows 6
4.1. Sub-gaussian random vectors 7
4.2. Nets 7
4.3. Extreme singular values 7
5. Matrices with independent heavy-tailed rows 8
5.1. Heavy-tailed distributions 8
5.2. Extreme singular values 8
6. Application: covariance estimation 9
6.1. Sub-gaussian distributions 9
6.2. Heavy-tailed distributions 10
6.3. Low-dimensional distributions: PCA 11
7. Application: RIP for sub-gaussian matrices 12

Partially supported by NSF grants FRG DMS 0918623 and DMS 1001829.

7.1. Definition and spectral characterization 12
7.2. Sub-gaussian RIP 12
8. RIP for heavy-tailed matrices 13
8.1. Statement and reduction to a uniform LLN 14
8.2. Symmetrization 14
8.3. Uniform Rudelson's inequality 15
8.4. Uniform Rudelson's inequality implies RIP 16
8.5. Dudley's inequality 16
8.6. Proof of uniform Rudelson's inequality 17
9. Extreme eigenvalues via spectral sparsifiers 18
9.1. Covariance estimation without logarithmic oversampling 18
9.2. Soft spectral edges and Stieltjes transform 19
9.3. Updating the soft spectral edges 20
9.4. Conclusion for the upper spectral edge 21
9.5. The fault and its correction: regularity of the distribution 22
References 23

DAY 1

1. Random matrices and random vectors

References to classical random matrix theory: [Anderson, Guionnet, Zeitouni], [Bai, Silverstein]. The book [Anderson, Guionnet, Zeitouni] can be downloaded from Ofer Zeitouni's webpage.

1.1. **Marchenko-Pastur Law.** See e.g. [Götze, Tikhomirov].

Consider an $N \times n$ random matrix A whose entries are iid, mean zero, unit variance.

Wishart matrix: $p \times p$ symmetric random matrix

$$W = \frac{1}{N} A^T A.$$

This normalization by $A \mapsto \frac{1}{\sqrt{N}} A$ is meant to make the columns of A unit norm on average.

W has p eigenvalues $\lambda_i(W)$, which are non-negatibe real numbers.

Asymptotic, or limiting, regime: $N, n \to \infty$ while $n/N \to y \in (0,1)$.

Theorem 1.1 (Marchenko-Pastur Law). *Consider the counting function of the eigenvalues*

$$F_W(x) = \frac{1}{p} \Big| \{ i : \lambda_i(W) \le x \} \Big|, \qquad x \in \mathbb{R}.$$

It is called the empirical distribution function. Then for every x,

$$\mathbb{E}\, F_W(x) \to F(x)$$

where $F(x)$ is the distribution function of a distribution on $\mathbb{R}$ with density

$$F'(x) = \frac{1}{2\pi x y}\sqrt{(b-x)(a-x)}\,\mathbf{1}_{[a,b]}(x), \qquad a = (1-\sqrt{y})^2, \quad b = (1+\sqrt{y})^2.$$

Make a drawing.

Note: the spectrum is compactly supported. The spectral edges a, b have special meaning.

1.2. **Singular values and operator norms.** [V, Section 2.1]

Definition of singular values $s_i(A)$.

Extreme singular values: $s_{\min}(A)$, $s_{\max}(A)$.

Extreme singular values and geometric distortion. Approximate isometries.

Extreme singular values and the operator norms: $s_{\max}(A) = \|A\|$, $s_{\min}(A) = 1/\|A^\dagger\|$.

1.3. **Bai-Yin Law.** See [Bai-Yin].

Theorem 1.2 (Bai-Yin's Law). *Assume, in addition to Marchenko-Pastur, that the entries of A have finite fourth moment. Then the singular values of the normalized matrix $\bar{A} = \frac{1}{\sqrt{N}}A$ satisfy*

$$s_{\min}(\bar{A}) \to 1-\sqrt{y}, \quad s_{\max}(\bar{A}) = 1+\sqrt{y} \quad \textit{almost surely.}$$

Thus: tall random matrices ($y = n/N \approx 0$) should be approximate isometries.

1.4. **Goals, methods.** Goals: study the extreme singular values of $N \times n$ random matrices A.

(a) In the *non-asymptotic* regime where N, n are fixed;

(b) A has just *independent rows* (or columns) rather than independent entries;

(c) the rows of A will be *sampled from some distribution* in R^n;

(d) the distribution may be highly non-Gaussian, perhaps discrete, and often *heavy-tailed.*

Distributions of interest:

1.5. **Isotropic random vectors.** [V, Section 2.5]

X: random vector in $\mathbb{R}^n$. *Covariance matrix:* $\Sigma = \Sigma(X) = \mathbb{E}\, X \otimes X = \mathbb{E}\, XX^T$.

X is *isotropic* if $\Sigma(X) = I$. Equivalently, $\mathbb{E}\langle X, x\rangle^2 = \|x\|_2^2$ for all $x \in \mathbb{R}^n$.

X arbitrary $\Rightarrow \Sigma^{-1/2}X$ is isotropic.

X isotropic $\Rightarrow \mathbb{E}\,\|X\|_2^2 = n$.

Examples of isotropic random vectors: Gaussian, Bernoulli, product, coordinate, spherical, uniform on a convex set

An extra example: *random Fourier* X = a row of an orthogonal $n \times n$ matrix chosen uniformly at random. In particular, from DFT matrix with entries

$$W_{\omega,t} = \exp\Big(-\frac{2\pi i \omega t}{n}\Big), \quad \omega, t \in \{0, \dots, n-1\}.$$

X corresponds to a random frequency.

1.6. **Gordon's theorem.** Time permitting. [V, Theorem 32, Corollary 35].

2. Sums of independent random variables

The material is from [V, Sections 2.3, 2.4]. See the book [Petrov] for much more.

2.1. **Sub-gaussian random variables.** [V, Section 2.3]

Gaussian and sub-gaussian random variables. [V, Lemma 5].

Sub-gaussian norm $\|X\|_{\psi_2}$.

Examples: Gaussian, Bernoulli, bounded.

Rotation invariance: [V, Lemma 9].

Hoeffding-type inequality: [V, Proposition 10]. CLT.

2.2. **Sub-exponential random variables.** [V, Section 2.4]

Sub-exponential random variables; norm $\|\cdot\|_{\psi_1}$.

Example: Sub-gaussian squared = sub-exponential: [V, Lemma 14].

MGF of sub-exponentials: [V, Lemma 15].

Bernstein-type inequality: [V, Proposition 16].

Sums of iid random variables: [V, Corollary 17].

3. Sums of sub-exponential random matrices

[V, Section 2.6]. We study

$$S_N = \sum_{i=1}^{N} X_i$$

where X_i are $n \times n$ independent random matrices, self-adjoint, zero mean. We seek *"non-commutative"* versions of deviation inequalities.

3.1. **Ahlswede-Winter's method.** Imitate arguments for scalar-valued case. [Ahlswede-Winter].

Matrix calculus. Order: $A \succeq B$ if $A - B$ is positive semi-definite.

Matrix exponential: $e^A = \sum_{k=1}^N A^k/k!$

Caveat: $e^{A+B} \neq e^A e^B$. Two ways around:

1. *Golden-Thompson inequality:*

$$\operatorname{tr} e^{A+B} \leq \operatorname{tr}(e^A e^B)$$

2. *Lieb's theorem:* For every self-adjoint H, the function

$$A \mapsto \operatorname{tr} \exp(H + \log A)$$

is concave on the positive-semidefinite cone.

Let's see how Lieb's theorem works, following [Tropp]. One can use Golden-Thompson instead, see [Oliveira]. Since $\|S_N\| \leq t$ is equivalent to $-tI \leq S_N \leq tI$, we have

$$\mathbb{P}\{\|S_N\| > t\} = \mathbb{P}\{S_N \not\preceq tI \text{ or } S_N \not\succeq -tI\}.$$

Repeat Bernstein's trick:

$$\mathbb{P}\{S_N \not\preceq tI\} = \mathbb{P}\{e^{\lambda S_N} \not\preceq e^{\lambda t I}\} \leq \mathbb{P}\{\operatorname{tr} e^{\lambda S_N} > e^{\lambda t}\} \leq e^{-\lambda t}\, \mathbb{E} \operatorname{tr} e^{\lambda S_N}.$$

$$\mathbb{E} \operatorname{tr} e^{\lambda S_N} = \mathbb{E}_N \operatorname{tr} \exp(\lambda S_{N-1} + \lambda X_N) = \mathbb{E}_N \operatorname{tr} \exp(\lambda S_{N-1} + \log e^{\lambda X_N}).$$

Condition on $X_1, \dots, X_{N-1}$. Using Lieb's theorem, we can apply Jensen's inequality to move the exectation w.r.to X_N inside:

$$\leq \mathbb{E}_{N-1} \operatorname{tr} \exp(\lambda S_{N-1} + \log \mathbb{E}\, e^{\lambda X_N}).$$

Repeat this for X_{N-1}, X_{N-2} etc. Eventually

$$\leq \operatorname{tr} \exp\Big(\sum_{i=1}^N \log \mathbb{E}\, e^{\lambda X_i}\Big) \leq n \cdot \Big\| \exp\Big(\sum_{i=1}^N \log \mathbb{E}\, e^{\lambda X_i}\Big)\Big\| = n \cdot \exp\Big(\Big\| \sum_{i=1}^N \log \mathbb{E}\, e^{\lambda X_i}\Big\|\Big)$$

The last identity holds because the eigenvalues of e^A are $e^{\lambda_i(A)}$.

The MGF of each matrix $\mathbb{E}\, e^{\lambda X_i}$ is easy to estimate. Finish as in the proof of Bernstein's inequality, optimizing λ.

Working out this argument yields the following non-comutative versions of Hoeffding and Bernstein:

Theorem 3.1 (Non-commutative Hoeffding, see [Tropp]). *Consider $n \times n$ self-adjoint matrices A_i and independent standard normal (or Bernoulli) r.v.'s g_i. Then*

$$\mathbb{P}\Big\{\Big\| \sum_{i=1}^N g_i A_i \Big\| \geq t\Big\} \leq n \cdot \exp\Big(-\frac{t^2}{2\sigma^2}\Big), \qquad \textit{where} \quad \sigma^2 = \Big\| \sum_{i=1}^N A_i^2 \Big\|.$$

Compare with classical Hoeffding: note (a) $\sigma \sim \|a\|_2$, (b) dimension factor n appears.

Theorem 3.2 (Non-commutative Bernstein, see [Tropp]). *Consider independent self-adjoint mean zero $n \times n$ random matrices X_i, and assume that $\|X_i\| \le K$. Then*

$$\mathbb{P}\Big\{\Big\| \sum_{i=1}^N X_i \Big\| \ge t\Big\} \le n \cdot \exp\Big[- c \min\Big(\frac{t^2}{\sigma^2}, \frac{t}{K} \Big) \Big].$$

3.2. **Rudelson's inequality.** See [V, Section 2.6]

Go back to non-commutative Hoeffding. Even the *expected value* of $\| \sum_i g_i A_i \|$ is non-trivial!

The expected value $\sim$ median = the value of t such that $n \cdot \exp(-t^2/2\sigma^2) = 1/2$. Solving for t yields $t \sim \sqrt{\log n} \cdot \sigma$. Hence

$$\mathbb{E} \Big\| \sum_{i=1}^N g_i A_i \Big\| \sim \sqrt{\log n} \cdot \Big\| \sum_{i=1}^N A_i^2 \Big\|^{1/2}. \tag{3.1}$$

Exercise: deduce (3.1) from Theorem 3.2 rigorously.

Example: rank-one matrices $A_i = x_i x_i^T$ where $x_i \in \mathbb{R}^n$ are some vectors. A quick computation of RHS of (3.1) yields:

Corollary 3.3 (Rudelson's inequality). *Let $x_i \in \mathbb{R}^n$ be vectors and g_i be independent standard normal (or Bernoulli) random variables. Then*

$$\mathbb{E} \Big\| \sum_{i=1}^N g_i \, x_i x_i^T \Big\| \le C \sqrt{\log n} \cdot \max_{i \le N} \|x_i\|_2 \cdot \Big\| \sum_{i=1}^N x_i x_i^T \Big\|^{1/2}.$$

Remarks:

(a) $\log n$ is needed in general; will come back to this later.

(b) Non-commutative Hoeffding, (3.1), and thus Rudelson's inequality can also be derived from a non-commutative form of *Khintchine inequality*, see [V, Section 2.6].

DAY 2

4. Matrices with independent sub-gaussian rows

[V, Section 4.1]. We study the extreme singular values of $N \times n$ matrices A with independent rows. The rows are random vectors chosen from some distribution in $\mathbb{R}^n$. First we consider very regular distributions: sub-gaussian.

4.1. **Sub-gaussian random vectors.** [V, end of Section 2.5]

[V, Definition 22].

Example: product of scalar sub-gaussian distributions, [V, Lemma 24].

Examples: Gaussian, Bernoulli, spherical, **not** coordinate, uniform on **some** convex sets – [V, Example 25].

4.2. **Nets.** [V, Section 2.2]

$\|A\| = \sup_{x \in S^{n-1}} \|Ax\|_2$. Need a uniform estimate on all $\|Ax\|$ for all $x \in S^{n-1}$. Individual estimate on each $\|Ax\|_2$ will follow from concentration, then union bound. Difficult to take union bound over infinite set S^{n-1}. So we discretize the sphere.

[V, Definition 1].

Nets of a sphere [V, Lemma 2].

Computing the operator norm on a net [V, Lemma 3].

Exercise: for symmetric operators, [V, Lemma 4].

4.3. **Extreme singular values.** [V, Section 4.1].

Main theorem, see [V, Theorem 39]:

Theorem 4.1 (Sub-gaussian rows)**.** *Let A be an $N \times n$ matrix whose rows are independent sub-gaussian isotropic random vectors in $\mathbb{R}^n$. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-ct^2)$ one has*

$$\sqrt{N} - C\sqrt{n} - t \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + C\sqrt{n} + t.$$

The constants $c, C > 0$ depend only on the maximal sub-gaussian norm of the rows.

Divide by $\sqrt{N}$. Theorem states that all singular values of $\bar{A} = \frac{1}{\sqrt{N}}A$ are

$$s_i(\bar{A}) = 1 \pm C\sqrt{\frac{n}{N}} \quad \text{with high probability.}$$

So $\bar{A}$ is an approximate isometry if $N \gg n$. In other words, *"tall random matrices are approximate isometries"*.

Compare to Bai-Yin Law: non-asymptotic, no independence of entries required, but C is present.

Proof. Follows from Bernstein and a covering argument. We want to show an equivalent statement:

$$|s_i(\bar{A})^2 - 1| \leq \varepsilon \quad \text{small}, \quad \varepsilon \sim \sqrt{\frac{n}{N}}.$$

In terms of eigenvalues, $s_i(\bar{A})^2 = \lambda_i(\bar{A}^T\bar{A}) = \lambda_i(\frac{1}{N}A^TA)$. So we want to show:

$$\|\frac{1}{N}A^TA - I\| \leq \varepsilon.$$

(i.e. that Wishart matrix is close to identity).

Now follow the argument in [V, Theorem 39] ... □

Remark: a similar result holds for matrices with independent sub-gaussian *columns*, [V, Theorem 58]. The argument in this case is different, it uses *decoupling.* Study it.

5. Matrices with independent heavy-tailed rows

[V, Section 4.2]

5.1. Heavy-tailed distributions.

We want to do away with *all* regularity assumptions like sub-gaussian, sub-exponential etc.

Why? *Discrete distributions in $\mathbb{R}^n$ are usually heavy-tailed.*

Example: an isotropic distribution uniformly distributed on a set of polynomial number of points $\{x_i\}$ in $\mathbb{R}^n$, say n^{10} points. Since $\mathbb{E}\,\|X\|_2^2 = n$, there is a point with $\|x_i\|_2 \geq \sqrt{n}$, and with probability mass n^{-10}. Then this is a bad direction for X: the random variable $\langle X, x_i \rangle$ takes value $\langle x_i, x_i \rangle \geq n$ with probability n^{-10}, so it is not sub-gaussian or sub-exponential. It has a polynomially heavy tail.

Example of heavy-tailed random vectors: coordinate, Fourier.

5.2. Extreme singular values.

Remarkably general main theorem, see [V, Theorem 41]:

Theorem 5.1 (Heavy-tailed rows). *Let A be an $N \times n$ matrix whose rows A_i are independent isotropic random vectors in $\mathbb{R}^n$. Suppose $\|A_i\|_2 \leq \sqrt{m}$ almost surely, for some number m. Then for every $t \geq 0$, one has*

$$\sqrt{N} - t\sqrt{m} \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + t\sqrt{m} \tag{5.1}$$

with probability at least $1 - 2n \cdot \exp(-ct^2)$. Here $c > 0$ is an absolute constant.

Since $\mathbb{E}\,\|A_i\|_2^2 = n$, the theorem is usually applied with $m \sim n$. Then the theorem may be compared with Bai-Yin law and the Theorem 4.1 for sub-gaussian rows.

Proof (with $m \sim n$). Follows from non-commutative Bernstein. We again want to show

$$\|\frac{1}{N}A^T A - I\| \leq \varepsilon \quad \text{small}, \quad \varepsilon \sim t\sqrt{\frac{n}{N}}.$$

Now follow the argument in [V, Theorem 41] ... □

A version of Theorem 5.1 can also be derived from Rudelson's inequality, Corollary 3.3, see [V, Section 4.2].

The main price for heavy-tailed rows, compared with sub-gaussian rows (Theorem 4.1) is the dimensional factor n that appears in the probability estimate $1 - 2n \cdot \exp(-ct^2)$. For this to be, say, 1/2, one needs to take $t \sim \log n$. So in reality Theorem 5.1 says that

$$\sqrt{N} - C\sqrt{n \log n} \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + C\sqrt{n \log n} \qquad \text{w.h.p.}$$

The lower bound is only non-trivial for $N \gtrsim n \log n$, i.e. for *logarithmically tall* matries.

The logarithmic factor is generally needed. Example: coordinate distribution (coupon collector's problem), see [V, Remark 43]. However, this is about the only example when log is needed for the lower bound; see the last lecture.

6. Application: covariance estimation

[V, Section 4.3]

6.1. **Sub-gaussian distributions.**

X: random vector in $\mathbb{R}^n$, for simplicity $\mathbb{E}\, X = 0$. Recall that the covariance matrix of X is

$$\Sigma = \Sigma(X) = \mathbb{E}\, XX^T.$$

Goal: estimate Σ in the operator norm from a sample of N independent points $X_1, \ldots, X_N$.

Sample covariance matrix:

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

By LLN,

$$\|\Sigma_N - \Sigma\| \to 0 \qquad \text{as } N \to \infty.$$

How large should the sample size $N = N(n, \varepsilon)$ be for accurate estimation? Say, for

$$\|\Sigma_N - \Sigma\| \le \varepsilon \|\Sigma\| \text{ ?}$$

Let's start with sub-gaussian distributions. Form an $N \times n$ matrix A with rows $A_i = \Sigma^{-1/2} X_i$. Thus indepednent, isotropic rows. Apply Theorem 4.1 for A.

In its proof, we showed that w.h.p.

$$\|\frac{1}{N} A^T A - I\| \le \varepsilon_0, \quad \varepsilon_0 \sim \sqrt{\frac{n}{N}}.$$

Multiplying by $\Sigma^{1/2}$ on the left and on the right, we obtain

$$\Big\| \frac{1}{N} \sum_{i=1}^{N} X_i X_i^T - \Sigma \Big\| = \|\Sigma_N - \Sigma\| \le \varepsilon_0 \|\Sigma\|.$$

We need that $\varepsilon_0 \le \varepsilon$. This holds when

$$N \gtrsim \varepsilon^{-2} n.$$

We have proved [V, Corollary 50]:

Corollary 6.1 (Covariance estimation for sub-gaussian distributions)**.** *Consider a sub-gaussian distribution*[1] *in* $\mathbb{R}^n$ *with covariance matrix* Σ*, and let* $\varepsilon \in (0,1)$*,* $t \geq 1$*. Then with probability at least* $1 - 2\exp(-t^2 n)$ *one has*

$$\text{If } N \geq C(t/\varepsilon)^2 n \quad \text{then } \|\Sigma_N - \Sigma\| \leq \varepsilon \|\Sigma\|.$$

Here C *depends only on the sub-gaussian norm of the distribution.*

In words: *the sample size* $N = O(n)$ *suffices for covariance estimation of sub-gaussian distributions.*

6.2. **Heavy-tailed distributions.**

Let us try a similar method for general (heavy-tailed) rows, using Theorem 5.1. Recall that in its proof we showed that if A has independent isotropic rows A_i with $\|A_i\| \leq m$, then

$$\|\frac{1}{N}A^T A - I\| \leq \varepsilon_0, \quad \varepsilon_0 \sim t\sqrt{\frac{m}{N}}$$

with probability at least $1 - 2n \cdot \exp(-ct^2)$.

If the rows are not isotropic but rather have covariance matrix Σ then a straightforward modification of the argument yields

$$\|\frac{1}{N}A^T A - \Sigma\| \leq \varepsilon_0 \, \|\Sigma\|^{1/2}, \quad \varepsilon_0 \sim t\sqrt{\frac{m}{N}}$$

(exercise).

Let us apply this for $A_i = X_i$, thus for $\Sigma_N = \frac{1}{N}A^T A$. We choose $t \sim \log n$ so that the probability is at least, say, 0.99. We again want the error to be $\varepsilon_0 \, \|\Sigma\|^{1/2} \leq \varepsilon$. Substituting ε_0 and solving for N shows that

$$N \gtrsim \varepsilon^{-2} \|\Sigma\|^{-1} m \log n.$$

Such sample size N guarantees the accurate covariance estimation:

$$\|\Sigma_N - \Sigma\| \leq \varepsilon \|\Sigma\|.$$

See [V, Corollary 52] for a full statement of this result.

How large is $m = \max \|X\|_2$ in practice? Since $\Sigma = \mathbb{E}\, X X^T$,

$$\mathbb{E}\, \|X\|_2^2 = \operatorname{tr}(\Sigma) \leq n\|\Sigma\|.$$

So for "most of the distribution",

$$m \lesssim n\|\Sigma\|.$$

To make this rigorous, we can *truncate* the distribution to make sure that $\|X\|_2 \lesssim n\|\Sigma\|$ a.s. In practice, we can estimate $\mathbb{E}\, \|X\|_2^2$ and *reject* the larger points.

[1] Technically, we assume that the isotropic vector $\Sigma^{-1/2}X$ is sub-gaussian. If, instead, one insists on X itself being sub-gaussian, then one gets $\|\Sigma_N - \Sigma\| \leq \varepsilon\|\Sigma\|$, see [V, Corollary 50].

Substituting such m, we have

$$N \gtrsim \varepsilon^{-2} n \log n.$$

So the general result is the same as for sub-gaussian distributions, except there is a *logarithmic oversampling.*

In words, *the sample size $N = O(n \log n)$ suffices for covariance estimation of general distributions.*

This was first proved (for isotropic distributions) by Mark Rudelson.

6.3. **Low-dimensional distributions: PCA.**

Principal Component Analysis (PCA): determine the covariance structure of a distribution, i.e. the eigenvectors and eigenvalues of Σ. Typical assumption: few large eigenvalues, i.e. the distribution is essentially low-dimensional.

The *full* covariance structure can be inferred within ε error from Σ_N once

$$\|\Sigma_N - \Sigma\| \le \varepsilon \|\Sigma\|.$$

As we know, this can be done with $N = O(n \log n)$ samples. But if the distribution is essentially low-dimensional, we can do better, possibly with $N \ll n$.

The intrinsic dimension = *effective rank* of the matrix Σ:

$$r(\Sigma) = \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|}.$$

Example: suppose the distribution is supported on some r-dimensional subspace E of $\mathbb{R}^n$. Then

$$r(\Sigma) \le r.$$

Moreover, if such distribution is isotropic in E then $r(\Sigma) = r$.

The effective rank is *stable* compared with the usual rank. Approximately low-dimenional $\Rightarrow r(\Sigma)$ is small.

We can repeat the argument in the previous section, but now with $r = r(\Sigma)$:

$$\mathbb{E}\, \|X\|_2^2 = \operatorname{tr}(\Sigma) = r\|\Sigma\|.$$

We recover the covariance estimation with $N \gtrsim \varepsilon^{-2} r \log n$.

In words, *the sample size $N = O(r \log n)$ suffices for covariance estimation of approximately r-dimensional distributions.*

So, we may have $N \ll n$.

Connections to compressed sensing.

Other structural models for which $N \ll n$ may be possible:

(a) sparse covariance matrices [Levina-V],

(b) graphical models.

7. Application: RIP for sub-gaussian matrices

7.1. **Definition and spectral characterization.** [V, Section 6]

Restricted Isometry Property (RIP) is a property of measurement matrices in compressed sensing, see Jared's course.

Definition (RIP)**.** An $m \times n$ matrix A satisfies the *restricted isometry property* for sparsity level $k \geq 1$ if there exists $\delta \in (0,1)$ such that the inequality

$$(1-\delta)\|x\|_2 \leq \|Ax\|_2 \leq (1+\delta)\|x\|_2$$

holds for all $x \in \mathbb{R}^n$ with $|\operatorname{supp}(x)| \leq k$. The smallest number $\delta = \delta(A,k)$ is called the *restricted isometry constant* of A.

In practice, $\delta = 0.1$ suffices.

Make a drawing.

Goal: construct a RIP matrix with small m (few "measurements"), e.g. with

$$m \sim k \log n.$$

Only random constructions are known. Random matrices.

Spectral characterization of RIP. From the geometric meaning of the extreme singular values, we see that $\delta(A,k) \leq \delta$ is equivalent to:

$$1-\delta \leq s_{\min}(A_J) \leq s_{\max}(A_J) \leq 1+\delta \tag{7.1}$$

for all $J \subseteq [n]$, $|J| = k$. Here A_J stands for the minor of A formed by the columns indexed by J.

7.2. **Sub-gaussian RIP.** Let's show that *sub-gaussian matrices satisfy RIP*. So let A be a matrix with independent, isotropic, sub-gaussian rows A_i.

Fix J for a moment, and apply Theorem 4.1 for the $m \times k$ matrix A_J. We get:

$$\sqrt{m} - C\sqrt{k} - t \leq s_{\min}(A_J) \leq s_{\max}(A_J) \leq \sqrt{m} + C\sqrt{k} + t \tag{7.2}$$

with probability $1 - 2\exp(-ct^2)$.

Now take union bound over all subsets $J \subseteq [n]$, $|J| = k$. There are

$$\binom{n}{k} \leq \Big(\frac{en}{k}\Big)^k$$

such subsets. So 7.2 holds simultaneously for all J with probability

$$1 - \Big(\frac{en}{k}\Big)^k \cdot 2\exp(-ct^2) \gtrsim 1 - \exp\big(k\log(n/k) - t^2\big).$$

We want the probability to be, say 0.99, which holds if we choose $t \sim \sqrt{k\log(n/k)}$.

Putting this value of t into (7.2), we see that

$$(1-\delta)\sqrt{m} \leq s_{\min}(A_J) \leq s_{\max}(A_J) \leq (1+\delta)\sqrt{m}$$

if we choose $m \gtrsim \delta^{-2} k \log(n/k)$. Dividing both sides by$\sqrt{m}$, we recover the spectral form of RIP (7.1) for the normalized matrix $\frac{1}{\sqrt{m}}A$.

We have proved a version of a result of Mendelson, Pajor and Tomczak-Jaegermann, see [V, Theorem 65]:

Theorem 7.1 (Sub-gaussian RIP). *Let A be an $m \times n$ sub-gaussian random matrix whose rows are independent sub-gaussian isotropic random vectors. Then the normalized matrix $\bar{A} = \frac{1}{\sqrt{m}}A$ satisfies the following for every sparsity level $k \leq n$ and every number $\delta \in (0,1)$:*

$$\textit{if } m \gtrsim \delta^{-2} k \log(en/k) \quad \textit{then } \delta(\bar{A}, k) \leq \delta$$

with probability at least $1 - 2\exp(-c\delta^2 m)$.

In words, for every dimension n and sparsity level k the following holds. *Random $m \times n$ matrices with independent sub-gaussian rows satisfy RIP for the "number of measurements"*

$$m \sim k \log(n/k).$$

Optimal (see Jared).

A similar result holds for independent sub-gaussian *columns* [V, Theorem 65].

DAY 3

8. RIP for heavy-tailed matrices

[V, Section 6.2]

RIP: Recall that an $m \times n$ matrix A satisfies RIP for sparsity level $k \geq 1$ if there exists $\delta \in (0,1)$ such that[2]

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2$$

holds for all $x \in \mathbb{R}^n$ with $|\operatorname{supp}(x)| \leq k$. The smallest number $\delta = \delta(A,k)$ is called the *restricted isometry constant* of A.

Goal: prove that random matrices with *general* independent rows (not just sub-gaussian as on Day 2) satisfy RIP with

$$m \sim k \log^4 n.$$

Motivation: Partial Fourier matrices. Rows of A are sampled uniformly from $n \times n$ DFT matrix. Measurements Ax consists of m random frequencies of the signal x.

[2] We now use a squared version of RIP. This is equivalent to the version given on Day 2 since $(1+\delta)^2 \sim 1+2d$ for small δ.

Implication in Compressed Sensing: *one can effectively recover a k-sparse signal x from $m \sim k \log^4 n$ random frequencies of x.*

8.1. **Statement and reduction to a uniform LLN.**

Theorem 8.1 (Heavy-tailed RIP)**.** *Let A be an $m \times n$ matrix whose rows A_i are independent isotropic random vectors, and with uniformly bounded entries: $|A_{ij}| = O(1)$ a.s. Then the normalized matrix $\bar{A} = \frac{1}{\sqrt{m}} A$ satisfies the following for every sparsity level $k \le n$ and every number $\delta \in (0,1)$:*

$$\text{if } m \ge C\delta^{-2} k \log^4 n \quad \text{then } \mathbb{E}\,\delta(\bar{A},k) \le \delta.$$

Here C depends only on the bound on the entries.

Argument: the result won't follow by union bound from Theorem 5.1 applied to every k-column minor of A. The probability is too weak; this approach leads to $m \sim k^2$ (*"quadratic bottleneck"*).

Reformulation of RIP for A:

$$\|A_J^T A_J - I_J\| \le \delta \quad \text{for every} J \subseteq [n], \ |J| = k.$$

The conclusion of Theorem 5.1 is equivalent to:

$$E := \mathbb{E} \max_{|J|=k} \Big\| \frac{1}{m} A_J^T A_J - I_J \Big\| \le \delta.$$

Since $A_J^T A_J = \sum_{i=1}^m (A_i)_J (A_i)_J^T$,

$$E = \mathbb{E} \max_{|J|=k} \Big\| \frac{1}{m} \sum_{i=1}^m (A_i)_J (A_i)_J^T - I_J \Big\|. \tag{8.1}$$

For each J, the $(A_i)_J (A_i)_J^T$ are independent random $k \times k$ matrices with mean I_J (by isotropy).

So "E is small" is a uniform version of LLN – uniform over the subsets J.

How to prove it?

8.2. **Symmetrization.**

Lemma 8.2 (Symmetrization)**.** [V, Lemma 46] *Let (X_i) be a finite sequence of independent random vectors valued in some Banach space, and (ε_i) and (g_i) be independent symmetric Bernoulli (resp. standard normal) random variables. Then*

$$\mathbb{E} \Big\| \sum_i (X_i - \mathbb{E} X_i) \Big\| \le 2\,\mathbb{E} \Big\| \sum_i \varepsilon_i X_i \Big\| \lesssim \mathbb{E} \Big\| \sum_i g_i X_i \Big\|.$$

Proof. Consider r.v's $\tilde{X}_i = X_i - X_i'$ where (X_i') is an independent copy of (X_i). Then $\tilde{X}_i$ are independent symmetric random variables, i.e. the sequence $(\tilde{X}_i)$ is distributed identically with $(-\tilde{X}_i)$ and thus also with $(\varepsilon_i \tilde{X}_i)$. Replacing $\mathbb{E}\, X_i$ by $\mathbb{E}\, X_i'$ and using Jensen's inequality, symmetry, and triangle inequality, we obtain the required inequality

$$\begin{aligned} \mathbb{E}\Big\| \sum_i (X_i - \mathbb{E}\, X_i) \Big\| &\le \mathbb{E}\Big\| \sum_i \tilde{X}_i \Big\| = \mathbb{E}\Big\| \sum_i \varepsilon_i \tilde{X}_i \Big\| \\ &\le \mathbb{E}\Big\| \sum_i \varepsilon_i X_i \Big\| + \mathbb{E}\Big\| \sum_i \varepsilon_i X_i' \Big\| = 2\,\mathbb{E}\Big\| \sum_i \varepsilon_i X_i \Big\|. \end{aligned}$$

For the second part of the lemma, since $\mathbb{E}\, |g_i| = \text{const}$,

$$\begin{aligned} \mathbb{E}\Big\| \sum_i \varepsilon_i X_i \Big\| &\sim \mathbb{E}\Big\| \sum_i \varepsilon_i\, \mathbb{E}\, |g_i| X_i \Big\| \\ &\le \mathbb{E}\Big\| \sum_i \varepsilon_i |g_i| X_i \Big\| \qquad \text{(by Jensen)} \\ &= \mathbb{E}\Big\| \sum_i g_i X_i \Big\| \qquad \text{(by symmetry of } g_i,\ \varepsilon_i|g_i| \equiv g_i). \end{aligned}$$

□

We apply symmetrization to the sum in E (Exercise: prove a uniform version of symmetrization with $\max_{|J| \le k}$, [V, Lemma 70].) This yields

$$E \lesssim \frac{1}{m}\, \mathbb{E} \max_{|J|=k} \Big\| \sum_{i=1}^{m} g_i\, (A_i)_J (A_i)_J^T \Big\|. \tag{8.2}$$

Now condition on A_i; we have a Gaussian sum.

How to bound this?

8.3. Uniform Rudelson's inequality.

Without $\max_{|J| \le k}$, we would just apply Rudelson's inequality, Corollary 3.3: for vectors $x_i \in \mathbb{R}^k$

$$\mathbb{E}\Big\| \sum_{i=1}^{m} g_i\, x_i x_i^T \Big\| \le C\sqrt{\log k} \cdot \max_{i \le m} \|x_i\|_2 \cdot \Big\| \sum_{i=1}^{m} x_i x_i^T \Big\|^{1/2}.$$

and would quickly finish the proof.

So now we know what we need – a uniform Rudelson's inequality over subsets J.

Proposition 8.3 (Uniform Rudelson's inequality)**.** [Rudelson-V], [V, Proposition 68] *Let $x_i \in \mathbb{R}^n$ be vectors such that $\|x_i\|_\infty = O(1)$. Let g_i be independent standard normal random variables. Then for every $k \le n$ one has*

$$\mathbb{E} \max_{|J|=k} \Big\| \sum_{i=1}^{m} g_i\, (x_i)_J (x_i)_J^T \Big\| \overset{*}{\lesssim} \sqrt{k} \cdot \max_{|J|=k} \Big\| \sum_{i=1}^{m} (x_i)_J (x_i)_J^T \Big\|^{1/2}$$

where $*$ hides some logarithmic factors.

Compare with classical Rudelson's inequality: by Hölder and the assumption,

$$\|(x_i)_J\|_2 \le \sqrt{k}\|(x_i)_J\|_\infty \lesssim \sqrt{k}.$$

So the uniform inequality contains the classical one (modulo log factors).

Will discuss the proof of uniform Rudelson later. Now:

8.4. **Uniform Rudelson's inequality implies RIP.** Apply uniform Rudelson to (8.2) conditionally on A_i, and then take the expectation w.r.to A_i:

$$E \overset{*}{\lesssim} \frac{\sqrt{k}}{m} \cdot \mathbb{E} \max_{|J|=k} \Big\| \sum_{i=1}^{m} (A_i)_J (A_i)_J^T \Big\|^{1/2}.$$

Compare the RHS with the definition (8.1) of E; so make E appear there:

$$E \overset{*}{\lesssim} \sqrt{\frac{k}{m}} \cdot \mathbb{E} \max_{|J|=k} \Big\| \frac{1}{m} \sum_{i=1}^{m} (A_i)_J (A_i)_J^T \Big\|^{1/2} \leq \sqrt{\frac{k}{m}}\, (E+1)^{1/2}.$$

Solving this quadratic equation yields

$$E \overset{*}{\lesssim} \sqrt{\frac{k}{m}}$$

We need $E \leq \delta$, this holds if

$$m \overset{*}{\gtrsim} \delta^{-2} k$$

as required. (An accurate computation gives $m \gtrsim \delta^{-2} k \log^4 n$ as proised in RIP.)

Theorem 8.1 is proved. □

We will now outline a proof of uniform Rudelson's inequality following [Rudelson-V]. The reader who wishes to understand this method may choose to study the proof of Lemma 3.2 in [Rudelson – early work] first, and then study the argument in [Rudelson-V].

8.5. **Dudley's inequality.** The argument is based on *stochastic processes.* We shall realize the LHS of uniform Rudelson as the maximum of a Gaussian process, and we use Dudley's inequality to estimate it.

Theorem 8.4 (Dudley's inequality)**.** *See e.g.* [Talagrand, (0.3), (1.18)] *Let* $(Z_t)_{t \in T}$ *be a Gaussian process. Define a metric* d *on the set* T *by*

$$d(s,t) := \|Z_s - Z_t\|_{L_2} = \big(\mathbb{E}\,|Z_s - Z_t|^2\big)^{1/2}.$$

Then

$$\mathbb{E} \sup_{t \in T} |Z_t| \leq C \int_0^\infty \sqrt{\log N(T,d,u)}\; du$$

where $N(T,d,u)$ *is the* covering number *of* T *in metric* d*, which is the minimal number of balls of radius* u *which cover* T*.*

$\log N(T,d,u)$ is called the *metric entropy* of T.

Example: g is a standard normal vector in $\mathbb{R}^n$, and

$$T \subset \mathbb{R}^n, \quad Z_t = \langle g, t \rangle = \sum_{i=1}^{n} g_i t_i.$$

(Essentially all Gaussian processes can be represented as in this example). Then

$$d(s,t)^2 = \mathbb{E}\Big|\sum_{i=1}^n g_i(s_i - t_i)\Big|^2 = \sum_{i=1}^n (s_i - t_i)^2 = \|s-t\|_2^2.$$

So $d =$ the *Euclidean distance* in $\mathbb{R}^n$.

8.6. **Proof of uniform Rudelson's inequality.** We want to bound

$$E := \mathbb{E}\max_{|J|=k}\Big\|\sum_{i=1}^m g_i\,(x_i)_J(x_i)_J^T\Big\| = \mathbb{E}\max_{\substack{|J|=k\\ x\in B_2^J}}\Big|\sum_{i=1}^m g_i\langle x_i, x\rangle^2\Big|$$

where B_2^J denotes the unit Euclidean ball in $\mathbb{R}^J$. The RHS is the maximum of a Gaussian process, taken over

$$x \in T := \bigcup_{|J|=k} B_2^J$$

(the union of k-dimensional discs in $\mathbb{R}^n$.)

We apply Dudley's inequality as in the example above with $t_i = \langle x_i, x\rangle^2$:

$$E \lesssim \int_0^\infty \sqrt{\log N(T,d,u)}\,du$$

where

$$d(x,y)^2 = \sum_{i=1}^m \Big(\langle x_i, x\rangle^2 - \langle x_i, y\rangle^2\Big)^2.$$

We simplify $d(x,y)$ using the identity $a^2 - b^2 = (a+b)(a-b)$:

$$d(x,y)^2 \le \sum_{i=1}^m \big(\langle x_i, x\rangle + \langle x_i, y\rangle\big)^2 \cdot \max_{i\le m}|\langle x_i, x-y\rangle|^2$$

The first factor simplifies using Minkowski inequality:

$$\sum_{i=1}^m \big(\langle x_i, x\rangle + \langle x_i, y\rangle\big)^2 \le 4\max_{z\in T}\sum_{i=1}^m \langle x_i, z\rangle^2 = 4\max_{|J|=k}\Big\|\sum_{i=1}^m (x_i)_J(x_i)_J^T\Big\|,$$

which is the RHS in uniform Rudelson's inequality, so that's fine.

The second factor $\max_{i\le m}|\langle x_i, x-y\rangle|$ looks like the ℓ_∞ distance in $\mathbb{R}^m$. If x_i were the coordinate basis vectors, that would be precisely the ℓ_∞ distance. In our case, x_i are arbitrary vectors with $\|x_i\|_\infty = O(1)$.

So the proof reduces to a geometric problem – estimate the covering number of the set T in the ℓ_∞-type distance

$$d'(x,y) = \max_{i\le m}|\langle x_i, x-y\rangle|.$$

We skip this step, see [Rudelson-V]. One substitutes the estimate on the covering number into Dudley's inequality, and finishes the proof. □

9. Extreme eigenvalues via spectral sparsifiers

The recent work [Batson-Spielman-Srivastava] on spectral sparsifiers: approximate a dense graph by a sparse graph.

The argument in [Batson-Spielman-Srivastava] suggests a new method in RMT for the extreme eigenvalues. This idea is pursued in [Srivastava-V].

9.1. **Covariance estimation without logarithmic oversampling.** Go back to *covariance estimation*: X mean zero random vector in $\mathbb{R}^n$,

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

are the covariance and sample covariance matrices.

By applying a linear transformation ($X \mapsto \Sigma^{-1/2}X$), we can assume that $\Sigma = I$, i.e. the distribution is isotropic.

Question: what is the minimal number of samples $N = N(n, \varepsilon)$ which guarantees

$$\|\Sigma_N - I\| \le \varepsilon\ ?$$

Let's say, $\varepsilon = 0.01$ fixed.

Answers:

(a) $N \sim n$ for sub-gaussian distributions, see Section 6.1.

(b) $N \sim n \log n$ for general distributions, see Section 6.2.

The logarithmic oversampling is needed in general:

Example: the *coordinate random vector* X, uniformly distributed in the set of $2n$ points $(\pm e_k)$ where $(e_k)_{k=1}^n$ is an orthonormal basis in $\mathbb{R}^n$. In order that

$$\lambda_{\min}(\Sigma_n) > 0$$

one needs Σ_N to have full rank, for which all n basis vectors e_k need be present in the sample $X_1, \dots, X_N$. By the coupon collector's problem, this can only happen if

$$N \gtrsim n \log n.$$

Question: for what distributions is logarithmic oversampling needed?

It is quite difficult to weaken the sub-gaussian assumption while keeping $N \sim n$. Recent solution to a problem of Kannan, Lovasz and Simonovits:

Theorem 9.1. [Adamczak, Litvak, Pajor, Tomczak] *$N \sim n$ for all sub-exponential distributions (supported in a ball of radius $\sqrt{N}$.) In particular, $N \sim n$ for the uniform distribution on an arbitrary convex body.*

Goal: using the method of spectral sparsifiers, we show that $N \sim n$ all regular distributions. The coordinate distribution is about the only "irregular" example when the logarithmic oversampling is needed.

Theorem 9.2. [Srivastava-V] *Assume X has $2+\eta$ moments for some $\eta > 0$, i.e.*

$$\sup_{x \in S^{n-1}} \mathbb{E}\,|\langle X_i, x\rangle|^{2+\eta} = O(1).$$

Then for $N \geq cn$,

$$\lambda_{\min}(\Sigma_N) \geq 1 - \varepsilon.$$

A similar upper bound $\lambda_{\max}(\Sigma_N) \leq 1 + \varepsilon$ holds under a somewhat stronger assumption (on all marginals, not just one-dimensional).[3] *Combining the upper and the lower bounds yields*

$$\|\Sigma_N - I\| \leq \varepsilon.$$

The constant $c > 0$ depends only on the moment bound, η and ε.

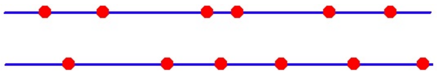
9.2. **Soft spectral edges and Stieltjes transform.** The proof is based on a randomization of the method of [Batson-Spielman-Srivastava]. We present the latter now.

Goal: Control the spectral edges $\lambda_{\min}(W)$ and $\lambda_{\max}(W)$ of the Wishart matrix

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

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*):



Difficulty: The spectral edges are not controlled by interlacing, they are *free* on one side. Thus they are difficult to compute.

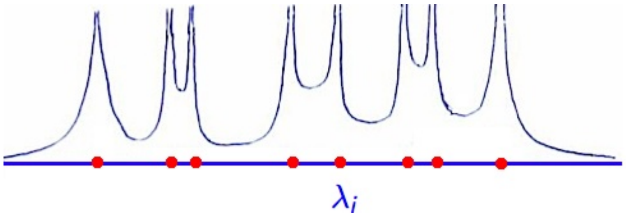
Solution: soften the spectral edges as follows:

Stieltjes Transform of the spectrum of W is the function

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

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

[3] Assumtion for the upper bound: All k-dimensional marginals of X have uniformly bounded $2+\eta$ moments outside the ball of radius $O(\sqrt{k})$.

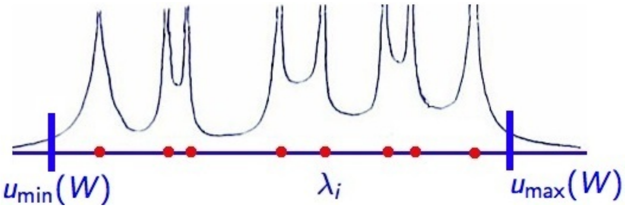


Physical interpretation: Put unit *electric charges* at points λ_i. Then $m_W(u)$ is the *electric potential* measured at u.

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

$$m_W(u) = \phi \qquad \text{(say, } \phi = 120\text{V or } 220\text{V).}$$

These locations act as *soft spectral edges.* They "harden" as $\phi \to \infty$ and "soften" as $\phi \to 0$:



9.3. **Updating the soft spectral edges.** 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 $W = \sum_{i=1}^N X_i X_i^T$. It can be updated by adding one term $X_i X_i^T$ at a time.

Sherman-Morrison formula for updating the inverse: for an invertible matrix A and a vector x,

$$(A + xy^T)^{-1} = A^{-1} - \frac{A^{-1}xy^TA^{-1}}{1 + y^TA^{-1}x}.$$

Let's see how this works, say for the upper soft edge. Suppose it is currently $u = u_{\max}(W)$, so

$$m_W(u) = \phi.$$

Add a term xx^T to W; how far will the soft edge move (to the right)? To answer this, we need to find a shift δ for which

$$m_{W+xx^T}(u + \delta) \le m_W(u) = \phi. \tag{9.1}$$

(This will imply that the soft spectral edge has moved by at most δ.)

By Sherman-Morrison:

$$\begin{aligned} m_{W+xx^T}(u+\delta) &= \operatorname{tr}(u+\delta-W-xx^T)^{-1} \\ &= m_W(u+\delta) + \frac{x^T(u+\delta-W)^{-2}x}{1-x^T(u+\delta-W)^{-1}x} \\ &= m_W(u) - \big(m_W(u) - m_W(u+\delta)\big) + \frac{x^T(u+\delta-W)^{-2}x}{1-x^T(u+\delta-W)^{-1}x} \end{aligned}$$

Further,[4]

$$m_W(u) - m_W(u+\delta) = \operatorname{tr}\left[(u-W)^{-1} - (u+\delta-W)^{-1}\right] \geq \delta \operatorname{tr}(u+\delta-W)^{-2}.$$

Substituting this and rearranging the terms, we see that if

$$\frac{x^T(u+\delta-W)^{-2}x}{\delta \operatorname{tr}(u+\delta-W)^{-2}} + x^T(u+\delta-W)^{-1}x \leq 1 \tag{9.2}$$

then (9.1) holds, and therefore the shift of the soft edge is $\leq \delta$.

9.4. **Conclusion for the upper spectral edge.** In our case, $x = X$ is a random isotropic vector. Take expectation in the sufficient condition (9.2). Since $\mathbb{E}\, X^T A X = \operatorname{tr}(A)$, the average of (9.2) becomes

$$\frac{1}{\delta} + \operatorname{tr}(u+\delta-W)^{-1} \leq 1.$$

Now, $\operatorname{tr}(u+\delta-W)^{-1} = m_W(u+\delta) \leq m_W(u) = \phi$, so the sufficient condition becomes

$$\frac{1}{\delta} + \phi \leq 1, \quad \text{or equivalently:} \quad \delta \geq \frac{1}{1-\phi}.$$

This simply says that the shift of the soft edge is always at most

$$\frac{1}{1-\phi}.$$

Induction argument: start with $N = 0$, $W = 0$, for which the initial soft spectral edge is

$$u = \frac{n}{\phi}.$$

because $m_0(u) = \operatorname{tr}(uI)^{-1} = n/u = \phi$.

Add N terms $X_i X_i^T$ to W one at a time, each time moving the soft spectral edge by at most $\frac{1}{1-\phi}$. At the end, the soft (and thus the hard) spectral edge of $W = \sum_{i=1}^N X_i X_i^T$ is at most

$$\frac{n}{\phi} + \frac{N}{1-\phi}.$$

[4] This follows from $\frac{1}{u-\lambda_i} - \frac{1}{u+\delta-\lambda_i} = \frac{\delta}{(u-\lambda_i)(u+\delta-\lambda_i)} \geq \frac{\delta}{(u+\delta-\lambda_i)^2}$.

Divide by N, thus for sample covariance matrix $\Sigma_N = \frac{1}{N} X_i X_i^T$,

$$\lambda_{\max}(\Sigma_N) \le \frac{y}{\phi} + \frac{1}{1-\phi}, \qquad \text{where } y = \frac{n}{N}.$$

Optimize in ϕ:

$$\lambda_{\max}(\Sigma_N) \le (1+\sqrt{y})^2.$$

This is exactly *the upper spectral edge b in Marchenko-Pastur Law* (and Bai-Yin)! – see Section 1.

9.5. **The fault and its correction: regularity of the distribution.** This argument can't be completely correct.[5] Otherwise the logarithmic oversampling is not needed for any distribution, which is wrong.

The fault: the sufficient condition (9.2) has to be *always* satisfied, even for a random vector $x = X$. One can't take expectation there.

Correction: Define the shift $\delta = \delta(x)$ to be the minimal number which makes (9.2) true. For $x = X$ a random vector, *the shift $\delta(X)$ is random.*

Prove that

$$\mathbb{E}\,\delta(X) \le \frac{1}{1-\phi} \tag{9.3}$$

or a similar upper bound.

This would correct the argument. But for (9.3) to hold, one needs *a bit of regularity* of the distribution, like $2+\eta$ moments. Otherwise there might be huge bursts of δ, making the expectation large.

See [Srivastava-V] for details.

[5] The argument can't be correct for *random* vectors. But if one is allowed to *choose* a value of X_i at every step, this is a correct argument, due to [Batson-Spielman-Srivastava].

References

[Adamczak, Litvak, Pajor, Tomczak] R. Adamczak, A. Litvak, A. Pajor, N. Tomczak-Jaegermann, *Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles*, J. Amer. Math. Soc. 23 (2010), 535–561.

[Anderson, Guionnet, Zeitouni] G. Anderson, A. Guionnet, O. Zeitouni, *An introduction to random matrices.* Cambridge Studies in Advanced Mathematics, 118. Cambridge University Press, Cambridge, 2010.

[Ahlswede-Winter] R. Ahlswede, A. Winter, *Strong converse for identication via quantum channels*, IEEE Trans. Information Theory 48 (2002), 568–579.

[Bai-Yin] Z. D. Bai,Y. Q. Yin, *Limit of the smal lest eigenvalue of a large-dimensional sample covariance matrix*, Ann. Probab. 21 (1993), 1275–1294.

[Bai, Silverstein] Z. D. Bai,Y. J. Silverstein, *Spectral analysis of large dimensional random matrices.* Second edition. Springer Series in Statistics. Springer, New York, 2010.

[Batson-Spielman-Srivastava] J. Batson, D. Spielman, N. Srivastava, *Twice-Ramanujan Sparsifiers*, STOC 2009. SICOMP, to appear.

[Götze, Tikhomirov] F. Götze, A. Tikhomirov, *Rate of convergence in probability to the Marchenko-Pastur law*, Bernoulli 10 (2004), 503–548.

[Levina-V] E. Levina, R. Vershynin, *Partial estimation of covariance matrices*, Probability Theory and Related Fields, to appear.

[Oliveira] R. Oliveira, *Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges*, preprint at arXiv:0911.0600

[Petrov] V. V. Petrov, *Sums of independent random variables.* Translated from the Russian by A. A. Brown. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 82. Springer-Verlag, New York-Heidelberg, 1975.

[Rudelson – early work] M. Rudelson, *Contact points of convex bodies*, Israel J. of Math. 101 (1997), 93–124.

[Rudelson] M. Rudelson, *Random vectors in the isotropic position*, J. Funct. Anal. 164 (1999), 60–72.

[Rudelson-V] M. Rudelson, R. Vershynin, R. *On sparse reconstruction from Fourier and Gaussian measurements*, Comm. Pure Appl. Math. 61 (2008), 1025–1045.

[Srivastava-V] N. Srivastava, R. Vershynin, *Covariance estimation for distributions with* $2 + \varepsilon$ *moments*, submitted.

[Talagrand] M. Talagrand, *The generic chaining. Upper and lower bounds of stochastic processes.* Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005.

[Tropp] J. Tropp, *User-friendly tail bounds for sums of random matrices*, submitted.

[V] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices.* In: Compressed Sensing: Theory and Applications, eds. Yonina Eldar and Gitta Kutyniok. Cambridge University Press, to appear (2010).

Department of Mathematics, University of Michigan, 530 Church St., Ann Arbor, MI 48109, U.S.A.

E-mail address: romanv@umich.edu