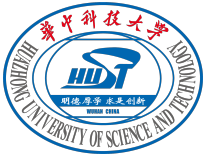
Random Matrix Theory for Modern Machine Learning: New Intuitions, Improved Methods, and Beyond: Part 1

Short Course @ Institut de Mathématiques de Toulouse, France

Zhenyu Liao

School of Electronic Information and Communications
Huazhong University of Science and Technology

July 1, 2024

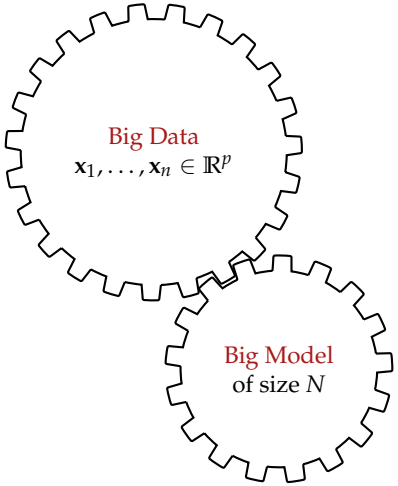


Schedule of the mini-course

1. Monday, July 1st (**today**): Motivation and Mathematical Background (concentration, resolvent-based approach to eigenspectral analysis, etc.)
2. Tuesday, July 2nd (**afternoon**): Four Ways to Characterize Sample Covariance Matrices and Some More Random Matrix Models (Wigner semicircle law, generalized sample covariance model, and separable covariance model)
3. Wednesday, July 3rd: Linear Master Theorem (information-plus-noise and additive spiked models) and RMT for Linear Machine Learning (Low-rank approximation, classification, and linear least squares)
4. Thursday, July 4th: Linearization of Nonlinear Models (Taylor expansion and Orthogonal Polynomial) and Nonlinear ML models via linearization: Kernel Methods in the Proportional Regime

Outline

Motivation: understanding large-dimensional machine learning



- **Big Data era**: exploit large n, p, N
- counterintuitive phenomena different from classical asymptotics statistics
- complete change of understanding of many methods in statistics, machine learning, signal processing, and wireless communications
- **Random Matrix Theory (RMT)** provides the tools!

Sample covariance matrix in the large n, p regime

- **Problem**: estimate covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$ from n data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$,
- Maximum likelihood sample covariance matrix with entry-wise convergence

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} \in \mathbb{R}^{p\times p}, \quad [\hat{\mathbf{C}}]_{ij} \to [\mathbf{C}]_{ij}$$

  almost surely as $n \to \infty$: optimal for $n \gg p$ (or, for p "small").

- In the regime $n \sim p$, conventional wisdom breaks down: for $\mathbf{C} = \mathbf{I}_p$ with $n < p$, $\hat{\mathbf{C}}$ has at least $p - n$ zero eigenvalues:

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty} \Rightarrow \text{ eigenvalue mismatch and not consistent!}$$

- due to **loss of matrix norm "equivalence"**: $\|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\| \leq p\|\mathbf{A}\|_{\max}$ for $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\|\mathbf{A}\|_{\max} \equiv \max_{ij} |\mathbf{A}_{ij}|$.

When is one in the random matrix regime? Almost always!

What about $n = 100p$? For $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$: MP law

$$\mu(dx) = (1 - c^{-1})^+ \delta(x) + \frac{1}{2\pi c x} \sqrt{(x - E_-)^+ (E_+ - x)^+} dx$$

where $E_- = (1 - \sqrt{c})^2$, $E_+ = (1 + \sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$. Close match!

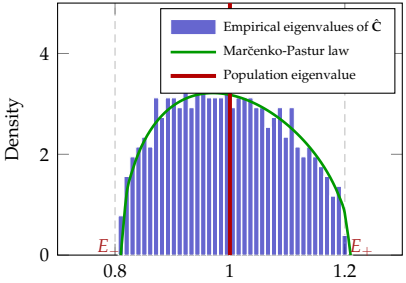


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

- eigenvalues span on $[E_- = (1 - \sqrt{\mathbf{c}})^2, E_+ = (1 + \sqrt{\mathbf{c}})^2]$.
- for $\mathbf{n = 100p}$, on a range of $\pm\mathbf{2}\sqrt{\mathbf{c}} = \pm\mathbf{0.2}$ around the population eigenvalue **1**.

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in biology, finance, signal processing, telecommunication, and machine learning, must **fail** even with $n = 100p$!
- **RMT** as a flexible and powerful tool to **understand** and **recreate** these methods
- in essence: large-scale system with increasing complexity in need of low complexity analysis
- as an motivating example, how RMT can be applied to assess kernel spectral clustering in **machine learning**

"Curse of dimensionality": loss of relevance of Euclidean distance

- Binary Gaussian mixture classification $\mathbf{x} \in \mathbb{R}^p$:

$$\mathcal{C}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1), \text{ versus } \mathcal{C}_2 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2);$$

- Neyman-Pearson test: classification is possible only when

$$\boxed{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \geq C_{\boldsymbol{\mu}}, \text{ or } \|\mathbf{C}_1 - \mathbf{C}_2\| \geq C_{\mathbf{C}} \cdot p^{-1/2}}$$

  for some constants $C_{\boldsymbol{\mu}}, C_{\mathbf{C}} > 0$ [CLM18].
- In this non-trivial setting, for $\mathbf{x}_i \in \mathcal{C}_a, \mathbf{x}_j \in \mathcal{C}_b$:

$$\boxed{\max_{1 \leq i \neq j \leq n} \left\{ \left| \frac{1}{p} \|\mathbf{x}_i - \mathbf{x}_j\|^2 - \frac{2}{p} \operatorname{tr} \mathbf{C}^\circ \right| \right\} \xrightarrow{a.s.} 0}$$

  as $n, p \to \infty$ (i.e., $n \sim p$), for $\mathbf{C}^\circ \equiv \frac{1}{2}(\mathbf{C}_1 + \mathbf{C}_2)$, regardless of the classes $\mathcal{C}_a, \mathcal{C}_b$!

[1] Romain Couillet, Zhenyu Liao, and Xiaoyi Mai. "Classification asymptotics in the random matrix regime". In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE. 2018, pp. 1875–1879

Loss of relevance of Euclidean distance: visual representation

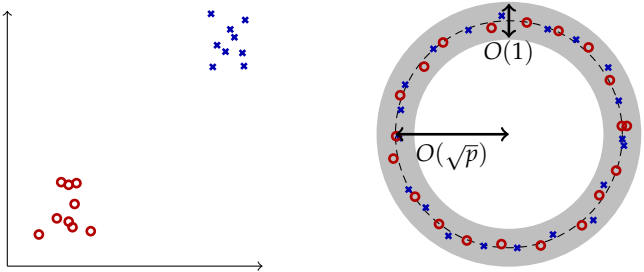
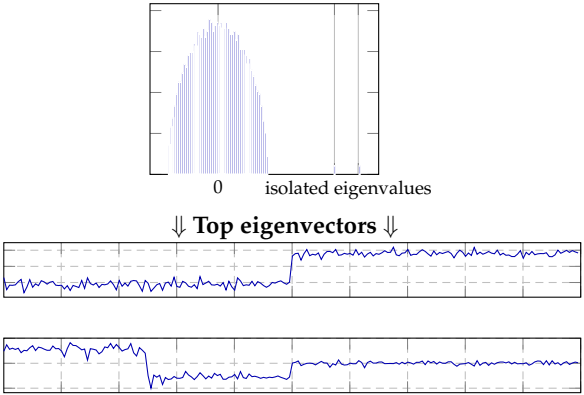


Figure: Visual representation of classification in (**left**) small and (**right**) large dimensions.

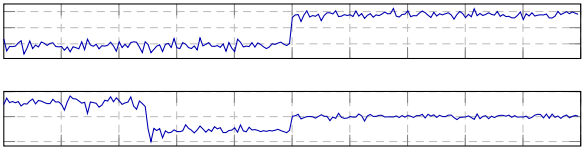
$\Rightarrow$ Direct consequence to various distance-based machine learning methods (e.g., kernel spectral clustering)!

Reminder on kernel spectral clustering

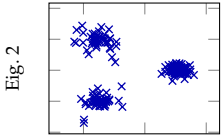
Two-step classification of n data points with distance kernel $\mathbf{K} \equiv \{f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)\}_{i,j=1}^n$:



Reminder on kernel spectral clustering



$\Downarrow$ **K-dimensional representation** $\Downarrow$

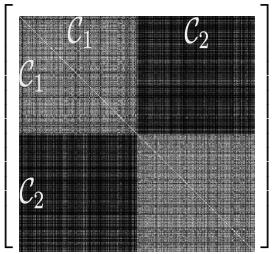


$\Downarrow$

EM or k-means clustering

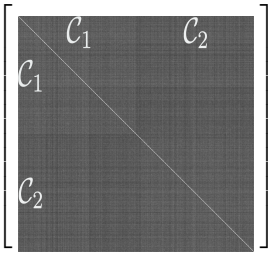
Cluster Gaussian data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$, with second top eigenvectors $\mathbf{v}_2$ of heat kernel $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$, small and large dimensional data.

(a) $p = 5, n = 500$

$\mathbf{K} =$ 

$\mathbf{v}_2 =$ 

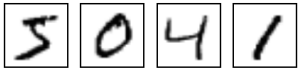
(b) $p = 250, n = 500$

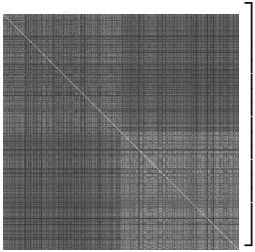
$\mathbf{K} =$ 

$\mathbf{v}_2 =$ 

Kernel matrices for large dimensional real-world data

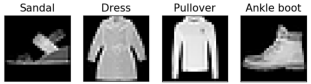
(a) MNIST

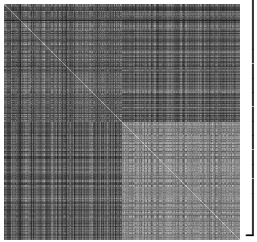


$$\mathbf{K} = \left[\right.$$  $$\left. \right]$$

$$\mathbf{v}_2 = \left[\right.$$  $$\left. \right]$$

(b) Fashion-MNIST



$$\mathbf{K} = \left[\right.$$  $$\left. \right]$$

$$\mathbf{v}_2 = \left[\right.$$  $$\left. \right]$$

A RMT viewpoint of large kernel matrices

- "local" linearization of **nonlinear** kernel matrices in large dimensions, e.g., Gaussian kernel matrix $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$ with $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$ (e.g., $\mathcal{C}_1 : \mathbf{x}_i = \boldsymbol{\mu}_1 + \mathbf{z}_i$ versus $\mathcal{C}_2 : \mathbf{x}_j = \boldsymbol{\mu}_2 + \mathbf{z}_j$) so that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2/p \xrightarrow{a.s.} 2, \text{ and } \mathbf{K} = \exp\left(-\frac{2}{2}\right)\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right) + g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T} + * + o_{\|\cdot\|}(1)$$

  with Gaussian $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n}$ and class-information $\mathbf{j} = [\mathbf{1}_{n/2}; -\mathbf{1}_{n/2}]$,
- accumulated effect of small "hidden" statistical information ($\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ in this case)

A RMT viewpoint of large kernel matrices

Therefore

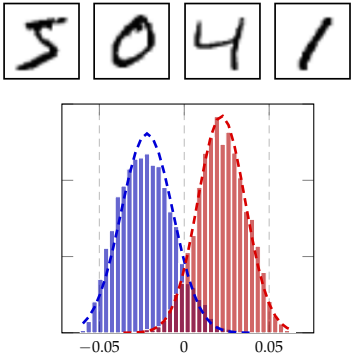
- **entry-wise**:

$$\mathbf{K}_{ij} = \exp(-1)\Big(1 + \underbrace{\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}_{O(p^{-1/2})}\Big) \pm \underbrace{\frac{1}{p}g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)}_{O(p^{-1})} + *, \text{ so that } \frac{1}{p}g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|) \ll \frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j,$$

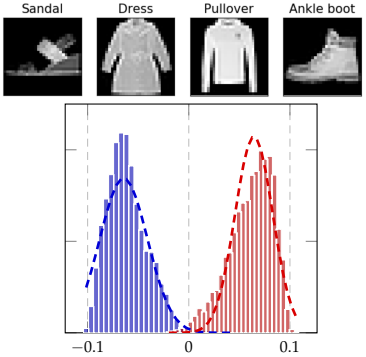
- **spectrum-wise**:
  - $\|\mathbf{K} - \exp(-1)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\| \not\to 0$;
  - $\|\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\| = O(1)$ and $\|g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T}\| = O(1)$!
- **Same** phenomenon as the sample covariance example: $[\hat{\mathbf{C}} - \mathbf{C}]_{ij} \to 0 \not\Rightarrow \|\hat{\mathbf{C}} - \mathbf{C}\| \to 0$!

$\Rightarrow$ With **RMT**, we understand kernel spectral clustering for large dimensional data!

Some more numerical results



(a) MNIST



(b) Fashion-MNIST

Figure: Empirical histogram of LS-SVM soft output versus RMT prediction, $n = 2\,048$, $p = 784$, $\gamma = 1$ with Gaussian kernel, for MINST (**left**, 7 versus 9) and Fashion-MNIST (**right**, 8 versus 9) data. Results averaged over 30 runs.

[2] **Zhenyu Liao** and Romain Couillet. "A Large Dimensional Analysis of Least Squares Support Vector Machines". In: *IEEE Transactions on Signal Processing* 67.4 (2019), pp. 1065–1074

Take-away of this section

- sample covariance matrix $\hat{\mathbf{C}}$ have different behavior in the large n, p regime
- loss of matrix norm "equivalence" for large matrices $\|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\| \leq p\|\mathbf{A}\|_{\max}$ for $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\|\mathbf{A}\|_{\max} \equiv \max_{ij} |\mathbf{A}_{ij}|$
- in the non-trivial classification regime: **loss of relevance of Euclidean distance**
- direct consequence in **all distance-based** ML methods, e.g., kernel spectral clustering
- **RMT** provides **an answer**

Characterization of scalar random variables: from moments to tails

Definition (Moments and moment generating function, MGF)

For a scalar random variable x defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we denote

- $\mathbb{E}[x]$ the *expectation* of x;
- $\mathrm{Var}[x] = \mathbb{E}[(x - \mathbb{E}[x])^2]$ the *variance* of x;
- for $p > 0$, $\mathbb{E}[x^p]$ the p^{th} *moment* of x, and $\mathbb{E}[|x|^p]$ the p^{th} *absolute moment*;
- for $\lambda \in \mathbb{R}$, $M_x(\lambda) = \mathbb{E}[e^{\lambda x}] = \sum_{p=0}^{\infty} \frac{\lambda^p}{p!} \mathbb{E}[x^p]$ the *moment generating function (MGF)* of x.

- the (absolute) moment of x writes as an integral of the tail of x
- characterization of the probability that x differs from a deterministic value by more than $t > 0$.

Lemma (Moments versus tails)

For a scalar random variable x and fixed $p > 0$, we have

1. $\mathbb{E}[|x|^p] = \int_0^{\infty} pt^{p-1} \mathbb{P}(|x| \geq t)\, dt$
2. $\mathbb{P}(|x| \geq t) \leq \exp(-\lambda t) M_x(\lambda)$, *for $t > 0$ and MGF $M_x(\lambda)$*

Definition (Sub-gaussian and sub-exponential distributions)

For a standard Gaussian random variable $x \sim \mathcal{N}(0,1)$, its law given by $\mu(dt) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$, so that $\mathbb{P}(x \geq X) = \mu([X,\infty)) = \frac{1}{\sqrt{2\pi}} \int_X^\infty \exp(-t^2/2)\, dt \leq \exp(-X^2/2)$.

- We say y is a *sub-gaussian random variable* if it has a tail that decays *as fast as* standard Gaussian random variables, that is
$$\mathbb{P}\left(|y| \geq t\right) \leq \exp(-t^2/\sigma_{\mathcal{N}}^2), \tag{1}$$
for some $\sigma_{\mathcal{N}} > 0$ (known as the *sub-gaussian norm* of y) for all $t > 0$.
- We can define a *sub-exponential random variable* z similarly via $\mathbb{P}(|z| \geq t) \leq \exp(-t/\sigma_{\mathcal{N}})$.

- for a sub-gaussian random variable x of mean $\mu = \mathbb{E}[x]$ and sub-gaussian norm $\sigma_{\mathcal{N}}$ that
$$\mathbb{P}\left(|x - \mu| \geq t\sigma_{\mathcal{N}}\right) \leq \exp(-t^2), \tag{2}$$
for all $t > 0$, in which the sub-gaussian norm $\sigma_{\mathcal{N}}$ of x acts as a scale parameter (that is similar, in spirit, to the variance parameter of Gaussian distribution).

A collection of scalar random variables: from LLN to CLT

For a collection of independent and identically distributed (i.i.d.) random variables $x_1, \ldots, x_n$ of mean μ and variance σ^2, we have, by independence, that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \mu, \quad \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[x_i] = \frac{\sigma^2}{n}. \tag{3}$$

- for μ, σ^2 do *not* scale with n, the (random) sample mean strongly concentrates around its expectation μ.

Theorem (Weak and strong law of large numbers, LLN)

For a sequence of i.i.d. random variables $x_1, \ldots, x_n$ with finite expectation $\mathbb{E}[x_i] = \mu < \infty$, we have

- *the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i \to \mu$ in probability as $n \to \infty$, known as the **weak law of large numbers (WLLN)**;*
- *the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i \to \mu$ almost surely as $n \to \infty$, known as the **strong law of large numbers (SLLN)**.*

Theorem (Central limit theorem, CLT)

For a sequence of i.i.d. random variables $x_1, \dots, x_n$ with $\mathbb{E}[x_i] = \mu$ and $\mathrm{Var}[x_i] = \sigma^2$, we have, for every $t \in \mathbb{R}$ that

$$\mathbb{P}\left(\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu) \geq t\right) \to \frac{1}{\sqrt{2\pi}}\int_t^{\infty} e^{-x^2/2}\,dx \tag{4}$$

as $n \to \infty$. That is, as $n \to \infty$, the random variable $\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu) \to \mathcal{N}(0,1)$ in distribution.

Remark: the results of LLN and CLT can be compactly written as

$$\frac{1}{n}\sum_{i=1}^{n} x_i \simeq \underbrace{\mu}_{O(1)} + \underbrace{\mathcal{N}(0,1)\cdot\sigma/\sqrt{n}}_{O(n^{-1/2})}, \tag{5}$$

as $n \to \infty$, for μ, σ both of order $O(1)$.

(i) In the first order (of magnitude $O(1)$), it has an asymptotically deterministic behavior around the expectation μ; and

(ii) in the second order (of magnitude $O(n^{-1/2})$), it strongly concentrates around this deterministic quantity with a universal Gaussian fluctuation, regardless of the distribution of the component of x_i.

Concentration of random vectors in high dimensions?

- "concentration" for a random vector $\mathbf{x} \in \mathbb{R}^n$?

Observation (Random vectors do not "concentrate" around their means)

For two *independent* random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, having i.i.d. entries with zero mean and unit variance (that is, $\mu = 0$ and $\sigma = 1$), we have that

$$\mathbb{E}[\|\mathbf{x} - \mathbf{0}\|_2^2] = \mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{x}] = \mathrm{tr}(\mathbb{E}[\mathbf{x}\mathbf{x}^\mathsf{T}]) = n, \tag{6}$$

and further by independence that

$$\mathbb{E}[\|\mathbf{x} - \mathbf{y}\|_2^2] = \mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{x} + \mathbf{y}^\mathsf{T}\mathbf{y}] = 2n. \tag{7}$$

- the origin $\mathbf{0}$ (and *mean* of $\mathbf{x}$) is always, in expectation, at the midpoint of two independent draws of random vectors in $\mathbb{R}^n$
- any random vector $\mathbf{x} \in \mathbb{R}^n$ with n large is not close to its mean
- $\mathbf{x}$ does not itself "concentrate" around any n-dimensional deterministic vector in any traditional sense.

Concentration of random vectors and their linear scalar observations

- In spite of this, from the LLN and CLT one expects that some types of "observations" of $\mathbf{x} \in \mathbb{R}^n$ (e.g., averages over all the entries of $\mathbf{x}$, to retrieve the sample mean), must concentrate in some sense for n large
- we "interpret" the sample mean as a linear scalar observation of a vector $\mathbf{x} \in \mathbb{R}^n$.

Remark (Sample mean as a linear scalar observation)

Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector having i.i.d. entries, then the sample mean of the entries of $\mathbf{x}$ can be rewritten as the following linear scalar observation $f : \mathbb{R}^n \to \mathbb{R}$ of $\mathbf{x}$ defined as

$$f(\mathbf{x}) = \mathbf{1}_n^\mathsf{T} \mathbf{x}/n = \frac{1}{n} \sum_{i=1}^{n} x_i, \text{ or } f(\cdot) = \mathbf{1}_n^\mathsf{T}(\cdot)/n. \tag{8}$$

- LLN and CLT are nothing but asymptotic characterization of the concentration behavior of the linear scalar observation $f(\mathbf{x})$ of the random vector $\mathbf{x} \in \mathbb{R}^n$
- we can say things non-asymptotically as well, under two different assumptions on the tail of $\mathbf{x}$.
  - (i) are only assumed to have finite variance σ^2 (but nothing on its tail behavior or higher-order moments); and
  - (ii) have sub-gaussian tails with sub-gaussian norm $\sigma_{\mathcal{N}}$.

Asymptotic and non-asymptotic concentration of random vectors

Table: Different types of characterizations of the linear scalar observation $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_n/n$ for $\mathbf{x} \in \mathbb{R}^n$, having i.i.d. entries with mean $\mathbb{E}[x_i] = \mu$ and variance σ^2 or sub-gaussian norm $\sigma_{\mathcal{N}}$.

	First-order behavior	Second-order behavior
Asymptotic	$f(\mathbf{x}) \to \mu$ via Law of Large Numbers	$\frac{\sqrt{n}}{\sigma}(f(\mathbf{x}) - \mu) \to \mathcal{N}(0,1)$ in law Central Limit Theorem
Non-asymptotic under finite variance	$\mathbb{E}[f(\mathbf{x})] = \mu$	$\mathbb{P}\left(\lvert f(\mathbf{x}) - \mu\rvert \geq t\sigma/\sqrt{n}\right) \leq t^{-2}$ via variance computation and Chebyshev's inequality
Non-asymptotic under sub-gaussianity	$\mathbb{E}[f(\mathbf{x})] = \mu$	$\mathbb{P}\left(\lvert f(\mathbf{x}) - \mu\rvert \geq t\sigma_{\mathcal{N}}/\sqrt{n}\right) \leq \exp(-Ct^2)$ via sub-gaussian tail bound

Remark (Concentration of scalar observation of large random vectors: asymptotic and non-asymptotics): A random vector $\mathbf{x} \in \mathbb{R}^n$, when "observed" via the linear scalar observation $f(\mathbf{x}) = \mathbf{1}_n^\mathsf{T}\mathbf{x}/n$:

$$f(\mathbf{x}) \simeq \underbrace{\mu}_{O(1)} + \underbrace{X/\sqrt{n}}_{O(n^{-1/2})}, \tag{9}$$

for n large, with some random X of order $O(1)$ that:

(i-i) has a tail that decays (at least) as t^{-2}, for finite n and $\mathbf{x}$ having entries of bounded variance;

(i-ii) has a sub-gaussian tail (at least) as $\exp(-t^2)$, for finite n and $\mathbf{x}$ having sub-gaussian entries;

(ii) has a precise Gaussian tail *independent* of the law of (the entries of) $\mathbf{x}$, but in the limit of $n \to \infty$ via CLT.

The concentration properties extend beyond the specific *linear* observation, $f(\mathbf{x}) = \mathbf{1}_n^\mathsf{T}\mathbf{x}/n$, to many types of (possibly) nonlinear observations.

Definition (Observation maps)

For random vector $\mathbf{x} \in \mathbb{R}^n$, we say $f(\mathbf{x}) \in \mathbb{R}$ is a scalar observation of $\mathbf{x}$ with observation map $f\colon \mathbb{R}^n \to \mathbb{R}$.

Table: Different types of scalar observations $f(\mathbf{x})$ of random vector $\mathbf{x} \in \mathbb{R}^n$, having independent entries.

	Scalar observation	Characterization
Linear	sample mean $f(\mathbf{x}) = \mathbf{1}_n^\mathbf{T}\mathbf{x}/n$, and $f(\mathbf{x}) = \mathbf{a}^\mathbf{T}\mathbf{x}$ for $\mathbf{a} \in \mathbb{R}^n$	Table in last slide
Lipschitz	$f(\mathbf{x})$ for a Lipschitz map $f\colon \mathbb{R}^n \to \mathbb{R}$	Lipschitz concentration
Quadratic form	$f(\mathbf{x}) = \mathbf{x}^\mathbf{T}\mathbf{A}\mathbf{x}$ for some $\mathbf{A} \in \mathbb{R}^{n\times n}$	Hanson–Wright inequality
Nonlinear quadratic form	$f(\mathbf{x}) = \sigma(\mathbf{x}^\mathbf{T}\mathbf{Y})\mathbf{A}\sigma(\mathbf{Y}^\mathbf{T}\mathbf{x})$ for entry-wise $\sigma\colon \mathbb{R} \to \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{Y} \in \mathbb{R}^{p\times n}$	Nonlinear quadratic concentration, of direct use in NN

Theorem (Concentration of Lipschitz map of Gaussian random vectors, [Ver18, Theorem 5.2.2])

For a standard Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ *and a Lipschitz function* $f\colon \mathbb{R}^n \to \mathbb{R}$ *that satisfies* $|f(\mathbf{y}_1) - f(\mathbf{y}_2)| \leq K_f \|\mathbf{y}_1 - \mathbf{y}_2\|_2$ *for any* $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^n$*, we have, for all* $t > 0$ *that*

$$\mathbb{P}\left(|f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})]| \geq t\right) \leq \exp(-Ct^2/K_f^2), \tag{10}$$

for some universal constant $C > 0$*, with* $K_f > 0$ *known as the Lipschitz constant of* f*.*

Remark (Concentration of Lipschitz observation of large random vectors): The Lipschitz scalar observations $f(\mathbf{x})$ of the random vector $\mathbf{x} \in \mathbb{R}^n$ behave as

$$f(\mathbf{x}) \simeq \mathbb{E}[f(\mathbf{x})] + K_f, \tag{11}$$

for n large, where K_f is the Lipschitz constant of f (that is, in general, of order $O(n^{-1/2})$, for example for $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_n/n$). This leads to first- and second-order behaviors:

(i) In the first order, $f(\mathbf{x})$ fluctuate around the deterministic quantity $\mathbb{E}[f(\mathbf{x})]$; and

(ii) in the second order, it concentrates around this deterministic quantity with a fluctuation/deviation that is proportional to K_f (and or order $O(n^{-1/2})$) and has a sub-gaussian tail

[3] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018

Concentration of quadratic forms

- intuitively expect that non-Lipschitz observation $f(\mathbf{x})$ still concentrates in some way, but "less so"
- important special case of quadratic forms, $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ for some given $\mathbf{A} \in \mathbb{R}^{n\times n}$

Theorem (Hanson–Wright inequality for quadratic forms, [Ver18, Theorem 6.2.1])

For a random vector $\mathbf{x} \in \mathbb{R}^n$ having independent, zero-mean, unit-variance, sub-gaussian entries with sub-gaussian norm bounded by $\sigma_\mathcal{N}$, and deterministic matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, we have, for every $t > 0$, that

$$\mathbb{P}\left(\left|\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} - \operatorname{tr}\mathbf{A}\right| \geq t\right) \leq \exp\left(-\frac{C}{\sigma_\mathcal{N}^2}\min\left(\frac{t^2}{\sigma_\mathcal{N}^2\|\mathbf{A}\|_F^2}, \frac{t}{\|\mathbf{A}\|_2}\right)\right), \tag{12}$$

for some universal constant $C > 0$.

- depending on the interplay between the "range" t and the deterministic matrix $\mathbf{A}$, the random quadratic form $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ swings between a sub-gaussian ($\exp(-t^2)$) and a sub-exponential ($\exp(-t)$) tail
- **Remark**: squared norm $\|\mathbf{x}\|_2^2$ as quadratic observation of $\mathbf{x} \in \mathbb{R}^n$: $\frac{1}{n}\|\mathbf{x}\|_2^2 \simeq 1 + O(n^{-1/2})$ for n large,
  - (i) In the first order, $\|\mathbf{x}\|_2^2/n$ fluctuate around the deterministic quantity one; and
  - (ii) in the second order, it concentrates around this deterministic quantity with a fluctuation/deviation that grows with $\sigma_\mathcal{N}^2$ and of order $O(n^{-1/2})$ with a sub-gaussian tail when close to the deterministic quantity, and with a sub-exponential tail (so with a fluctuation with heavier tail and concentrates "less" than the Lipschitz case) when far away.

Concentration of nonlinear quadratic forms

- nonlinear quadratic forms $\frac{1}{n} f(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A} f(\mathbf{Y}^\mathsf{T}\mathbf{x})$ for Gaussian $\mathbf{x} \in \mathbb{R}^p$ and deterministic $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{Y} \in \mathbb{R}^{p\times n}$

Theorem (Concentration of nonlinear quadratic forms, [LtC18, Lemma 1])

For a standard Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and deterministic $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{Y} \in \mathbb{R}^{p\times n}$ such that $\|\mathbf{A}\|_2 \le 1, \|\mathbf{Y}\|_2 = 1$, we have, for Lipschitz function $f : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant K_f and any $t > 0$ that

$$\mathbb{P}\left(\left| \frac{1}{n} f(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A} f(\mathbf{Y}^\mathsf{T}\mathbf{x}) - \frac{1}{n} \operatorname{tr} \mathbf{A}\mathbf{K}_f(\mathbf{Y}) \right| \ge \frac{t}{\sqrt{n}} \right) \le \exp\left(-\frac{C}{K_f^2} \min\left(\frac{t^2}{(|f(0)| + K_f\sqrt{p/n})^2}, \sqrt{n}t \right) \right), \tag{13}$$

with $\mathbf{K}_f(\mathbf{Y}) = \mathbb{E}_\mathbf{x}[f(\mathbf{Y}^\mathsf{T}\mathbf{x}) f(\mathbf{x}^\mathsf{T}\mathbf{Y})] \in \mathbb{R}^{n\times n}$, for some universal constant $C > 0$.

- a nonlinear extension of the Hanson–Wright inequality (consider, e.g., $\mathbf{Y} = \mathbf{I}_n$ with $p = n$)

Remark (Concentration of nonlinear quadratic form observation of large random vectors):

$$\frac{1}{n} f(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A} f(\mathbf{Y}^\mathsf{T}\mathbf{x}) \simeq \frac{1}{n} \operatorname{tr} \mathbf{A}\mathbf{K}_f(\mathbf{Y}) + O(n^{-1/2}), \tag{14}$$

for n large, with $\max\{f(0), K_f, p/n\} = O(1)$, and similar first and second order behavior as above.

[4] Cosme Louart, **Zhenyu Liao**, and Romain Couillet. "A random matrix approach to neural networks". In: *Annals of Applied Probability* 28.2 (2018), pp. 1190–1248

Lemma (Polarization identity)

For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have $\mathbf{x}^\mathsf{T}\mathbf{y} = \frac{1}{2}\left(\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2\right)$.

Observation (Different scaling for inner products and Euclidean norms of large random vectors)

Consider a random vector $\mathbf{x} \in \mathbb{R}^n$, so that $\sqrt{n}\mathbf{x}$ has i.i.d. entries with zero mean, unit variance, and finite fourth order moment $m_4 < \infty$ (the scaling by $\sqrt{n}$ is so that $\mathbb{E}[\|\mathbf{x}\|_2^2] = 1$), and a deterministic vector $\mathbf{y} \in \mathbb{R}^n$ of unit norm $\|\mathbf{y}\|_2 = 1$. Then, by LLN and CLT

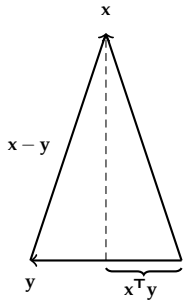
$$\mathbf{x}^\mathsf{T}\mathbf{y} \simeq 0 + \mathcal{N}(0,1)/\sqrt{n}, \tag{15}$$

for n large, so inner product $\mathbf{x}^\mathsf{T}\mathbf{y} = O(n^{-1/2})$. On the other hand, $\mathbb{E}[(\mathbf{x}^\mathsf{T}\mathbf{x})^2] = \frac{n+m_4-1}{n}$ and

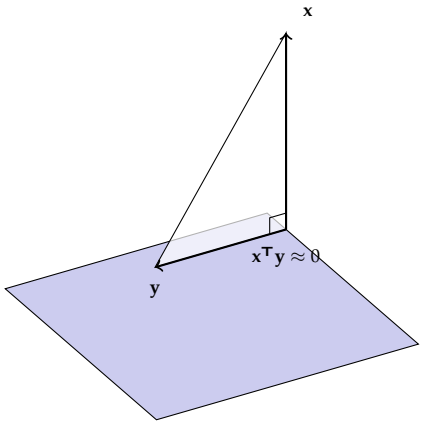
$$\|\mathbf{x}\|_2^2 = \mathbf{x}^\mathsf{T}\mathbf{x} \simeq 1 + \mathcal{N}(0, m_4 - 1)/\sqrt{n}, \quad \|\mathbf{x}-\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 + O(n^{-1/2}) = 2 + O(n^{-1/2}), \tag{16}$$

so that the Euclidean distance between $\mathbf{x}$ and any fixed $\mathbf{y}$ (or their norms) is much larger (in fact by a factor of $\sqrt{n}$) than their inner product.

Numerical illustration



(a) Deterministic



(b) Large-dimensional random

Figure: Visualization of the polarization identity or (a) *deterministic* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and (b) large-dimensional random vector $\mathbf{x} \in \mathbb{R}^n$ and deterministic $\mathbf{y} \in \mathbb{R}^n$.

A quick recap on linear algebra: matrices

Definition (Matrix inner product and Frobenius norm)

Given matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times n}$,

- $\mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{Y}) = \sum_{i=1}^{n} [\mathbf{X}^\mathsf{T}\mathbf{Y}]_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ji}Y_{ji}$ is the matrix inner product between $\mathbf{X}$ and $\mathbf{Y}$, where $\mathrm{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$; and
- $\|\mathbf{X}\|_F^2 = \mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{X}) = \sum_{i=1}^{n} [\mathbf{X}^\mathsf{T}\mathbf{X}]_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ji}^2$ denotes the (squared) Frobenius norm of $\mathbf{X}$, which is also the sum of the squared entries of $\mathbf{X}$.

Definition (Matrix norm)

For $\mathbf{X} \in \mathbb{R}^{p \times n}$, the following "entry-wise" extension of the p-norms of vectors.

1. matrix Frobenius norm $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} X_{ij}^2} = \|\mathrm{vec}(\mathbf{X})\|_2$ that extends the vector ℓ_2 Euclidean norm; and
2. matrix maximum norm $\|\mathbf{X}\|_{\max} = \max_{i,j} |X_{ij}| = \|\mathrm{vec}(\mathbf{X})\|_\infty$ that extends the vector ℓ_∞ norm.

and also matrix norm induced by vectors: $\|\mathbf{X}\|_p \equiv \sup_{\|\mathbf{v}\|_p = 1} \|\mathbf{X}\mathbf{v}\|_p$.

- taking $p = 2$ is the spectral norm: $\|\mathbf{X}\|_2 = \sqrt{\lambda_{\max}(\mathbf{X}\mathbf{X}^\mathsf{T})} = \sigma_{\max}(\mathbf{X})$, with $\lambda_{\max}(\mathbf{X}\mathbf{X}^\mathsf{T})$ and $\sigma_{\max}(\mathbf{X})$ the maximum eigenvalue and singular of $\mathbf{X}\mathbf{X}^\mathsf{T}$ and $\mathbf{X}$, respectively.

A quick recap on linear algebra: matrices

- Frobenius norm and spectral norm are matrix Schatten norms (that applies the vector p-norms on the vector of singular values of the matrix)
- are known to be unitarily invariant, that is $\|\mathbf{X}\| = \|\mathbf{UXV}\|$ for all matrices $\mathbf{X}$ and unitary matrices $\mathbf{U}, \mathbf{V}$ of appropriate dimensions

Remark (Matrix norm "equivalence")

For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, one has the following

1. $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{\operatorname{rank}(\mathbf{A})} \cdot \|\mathbf{A}\|_2 \leq \sqrt{\max(m,n)} \cdot \|\mathbf{A}\|_2$, *so that the control of the spectral norm via the Frobenius norm can be particularly loose for matrices of large rank; and*
2. $\|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\|_2 \leq \sqrt{mn} \cdot \|\mathbf{A}\|_{\max}$, *with* $\|\mathbf{A}\|_{\max} \equiv \max_{i,j} |A_{ij}|$ *the max norm of* $\mathbf{A}$, *so that the max and spectral norm can be significantly different for matrices of large size.*

- The fact that matrix norm "equivalence" holds only up to **dimensional factors** (e.g., rank and size) is crucial in large-dimensional data analysis and ML, as we have seen in the examples of SCM and kernel spectral clustering above.

Definition (Eigen-decomposition of symmetric matrices)

A symmetric real matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ admits the following eigen-decomposition

$$\mathbf{X} = \mathbf{U}_{\mathbf{X}}\mathbf{\Lambda}_{\mathbf{X}}\mathbf{U}_{\mathbf{X}}^{\top} = \sum_{i=1}^{n} \lambda_i(\mathbf{X})\mathbf{u}_i\mathbf{u}_i^{\top}, \tag{17}$$

for diagonal $\mathbf{\Lambda}_{\mathbf{X}} = \mathrm{diag}\{\lambda_i(\mathbf{X})\}_{i=1}^{n}$ containing $\lambda_1(\mathbf{X}), \ldots, \lambda_n(\mathbf{X})$ the real eigenvalues of $\mathbf{X}$, and orthonormal $\mathbf{U}_{\mathbf{X}} = [\mathbf{u}_1, \ldots, \mathbf{u}_n] \in \mathbb{R}^{n\times n}$ containing the corresponding eigenvectors. In particular,

$$\mathbf{X}\mathbf{u}_i = \lambda_i(\mathbf{X})\mathbf{u}_i. \tag{18}$$

- interested in a single eigenvalue of a symmetric real matrix, $\mathbf{X} \in \mathbb{R}^{n\times n}$, one may either resort to the eigenvalue-eigenvector equation in (18) or the determinant equation $\det(\mathbf{X} - \lambda\mathbf{I}_n) = 0$
- classical RMT is interested in the *joint* behavior of all eigenvalues $\lambda_1(\mathbf{X}), \ldots, \lambda_n(\mathbf{X})$, e.g., the (empirical) eigenvalue distribution of $\mathbf{X}$

Definition (Empirical Spectral Distribution, ESD)

For a real symmetric matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$, the *empirical spectral distribution (ESD)* or *empirical spectral measure* $\mu_{\mathbf{X}}$ of $\mathbf{X}$ is defined as the normalized counting measure of the eigenvalues $\lambda_1(\mathbf{X}), \ldots, \lambda_n(\mathbf{X})$ of $\mathbf{X}$,

$$\mu_{\mathbf{X}} \equiv \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{X})}, \tag{19}$$

where δ_x represents the Dirac measure at x. Since $\int \mu_{\mathbf{X}}(dx) = 1$, the spectral measure $\mu_{\mathbf{X}}$ of a matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ (which may be random or not) is a probability measure.

- $\int t\mu_{\mathbf{X}}(dt) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i(\mathbf{X})$ is the first moment of $\mu_{\mathbf{X}}$, and gives the average of all eigenvalues of $\mathbf{X}$; and
- $\int t^2\mu_{\mathbf{X}}(dt) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i^2(\mathbf{X})$ is the second moment of $\mu_{\mathbf{X}}$, so that $\int t^2\mu_{\mathbf{X}}(dt) - \left(\int t\mu_{\mathbf{X}}(dt)\right)^2$ gives the variance of the eigenvalues of $\mathbf{X}$.

Connection between linear equation and spectral decomposition

Consider the linear equation

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{20}$$

with $\mathbf{A} \in \mathbb{R}^{p\times n}$ and $\mathbf{b} \in \mathbb{R}^p$, we aim to solve for $\mathbf{x} \in \mathbb{R}^n$ solution to Equation (20).

- for square $\mathbf{A}$ with $p = n$, then Equation (20) admits a unique solution if and only if $\mathbf{A}$ is invertible, that is, 0 is not an eigenvalue of $\mathbf{A}$, and the solution is given by

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \tag{21}$$

- in the general case with $p \neq n$, $\mathbf{A}$ can be a fat ($p < n$) or tail ($p > n$) matrix, and is not invertible in either case, we use the Moore–Penrose pseudoinverse.

Definition (Moore–Penrose pseudoinverse)

For a real matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$, we say $\mathbf{X}^+ \in \mathbb{R}^{n\times p}$ is a (Moore–Penrose) pseudoinverse of $\mathbf{X}$ if it satisfies $\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$, $\mathbf{X}^+\mathbf{X}\mathbf{X}^+ = \mathbf{X}^+$, and both $\mathbf{X}\mathbf{X}^+$ and $\mathbf{X}^+\mathbf{X}$ are symmetric. In particular, for $\mathbf{X} = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^\mathsf{T}$ the SVD of $\mathbf{X}$, with orthonormal $\mathbf{U}_\mathbf{X} \in \mathbb{R}^{p\times p}$ and $\mathbf{V}_\mathbf{X} \in \mathbb{R}^{n\times n}$, the pseudoinverse of $\mathbf{X}$ can be written as

$$\mathbf{X}^+ = \mathbf{V}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}^{-1}\mathbf{U}_\mathbf{X}, \tag{22}$$

with $\mathbf{\Sigma}_\mathbf{X}^{-1}$ inverting all positive values in $\mathbf{\Sigma}_\mathbf{X}$ and leaving zeros unchanged.

Regularized inverse

The pseudoinverse "solves" the linear equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ in the following sense:

- The solutions to Equation (20) exist if and only if $\mathbf{A}\mathbf{A}^+\mathbf{b} = \mathbf{b}$, and all its solutions are given by

$$\mathbf{x} = \mathbf{A}^+\mathbf{b} + (\mathbf{I}_n - \mathbf{A}^+\mathbf{A})\mathbf{y}, \tag{23}$$

  for arbitrary $\mathbf{y} \in \mathbb{R}^n$. The solution is unique if and only if $\mathbf{I}_n - \mathbf{A}^+\mathbf{A} = \mathbf{0}$ and that $\mathbf{A}$ has full column rank.
- As a consequence, the solution $\hat{\mathbf{x}} = \mathbf{A}^+\mathbf{b}$ provides the least squares solution to Equation (20), as

$$\arg\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 = \mathbf{A}^+\mathbf{b}. \tag{24}$$

- however, can be numerically unstable as it inverts all singular values $\sigma(\mathbf{X})$ of $\mathbf{X}$ to $1/\sigma(\mathbf{X})$, see later (e.g., Part 3) for a manifestation of this under the (modern) name of double descent
- in the case of square $\mathbf{X}$, an alternative is the regularized inverse of $\mathbf{X}$,

$$\boxed{\mathbf{Q}_{\mathbf{X}}(\gamma) = (\mathbf{X} + \gamma\mathbf{I})^{-1},} \tag{25}$$

  for some regularization parameter $\gamma > 0$, with $\lambda_i(\mathbf{Q}_{\mathbf{X}}(\gamma)) = \frac{1}{\lambda_i(\mathbf{X})+\gamma}$, and $\|\mathbf{Q}_{\mathbf{X}}\| \leq 1/\gamma$.
- solves the regularized linear equation (i.e., ridge regression) as

$$\arg\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 + \gamma\|\mathbf{x}\|_2 = \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \gamma\mathbf{I}_p)^{-1}\mathbf{b} = (\mathbf{A}^\mathsf{T}\mathbf{A} + \gamma\mathbf{I}_n)^{-1}\mathbf{A}^\mathsf{T}\mathbf{b}. \tag{26}$$

- two solutions equivalent for any $\gamma > 0$, taking $\gamma \to 0$ is the "ridgeless" least squares solution $\mathbf{A}^+\mathbf{b}$.

A unified spectral analysis approach via the resolvent

- **Note**: here everything hold deterministically, not necessarily random **yet**
- combined with **deterministic equivalent** technique to be discussed in Part 2, gives the whole picture

Definition (Resolvent)

For a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$, the resolvent $\mathbf{Q}_{\mathbf{X}}(z)$ of $\mathbf{X}$ is defined, for $z \in \mathbb{C}$ not an eigenvalue of $\mathbf{X}$, as

$$\mathbf{Q}_{\mathbf{X}}(z) \equiv \left(\mathbf{X} - z\mathbf{I}_p\right)^{-1}. \tag{27}$$

Proposition (Properties of resolvent)

For $\mathbf{Q}_{\mathbf{X}}(z)$ the resolvent of a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$ with ESD $\mu_{\mathbf{X}}$ with supported on $\mathrm{supp}(\mu_{\mathbf{X}})$, then

(i) $\mathbf{Q}_{\mathbf{X}}(z)$ is complex analytic on its domain of definition $\mathbb{C} \setminus \mathrm{supp}(\mu_{\mathbf{X}})$;

(ii) it is bounded in the sense that $\|\mathbf{Q}_{\mathbf{X}}(z)\|_2 \leq 1/\operatorname{dist}(z, \mathrm{supp}(\mu_{\mathbf{X}}))$;

(iii) $x \mapsto \mathbf{Q}_{\mathbf{X}}(x)$ for $x \in \mathbb{R} \setminus \mathrm{supp}(\mu_{\mathbf{X}})$ is an increasing matrix-valued function with respect to symmetric matrix partial ordering (i.e., $\mathbf{A} \succeq \mathbf{B}$ whenever $\mathbf{z}^\mathsf{T}(\mathbf{A} - \mathbf{B})\mathbf{z} \geq 0$ for all $\mathbf{z}$).

A unified spectral analysis approach via the resolvent

- for real z, the resolvent $\mathbf{Q}_{\mathbf{X}}(z)$ is nothing but a regularized inverse of $\mathbf{X}$
- when interested in the eigenvalues and eigenvectors of $\mathbf{X} \in \mathbb{R}^{p \times p}$, consider the eigenvalue and eigenvector equation

$$\mathbf{X}\mathbf{v} = \lambda \mathbf{v} \Leftrightarrow (\mathbf{X} - \lambda \mathbf{I}_p)\mathbf{v} = \mathbf{0}, \quad \lambda \in \mathbb{R}, \mathbf{v} \in \mathbb{R}^p, \tag{28}$$

  for an eigenvalue-eigenvector pair $(\lambda, \mathbf{v})$ of $\mathbf{X}$ with $\mathbf{v} \neq \mathbf{0}$
- again a linear system, but solving for a pair of eigenvalue and eigenvector $(\lambda, \mathbf{v})$ for which the inverse/resolvent $(\mathbf{X} - \lambda \mathbf{I}_p)^{-1}$ does not exist
- while seemingly less convenient at first sight, turns out to be very efficient in providing a unified assess to general spectral functionals of $\mathbf{X}$, by taking z to be complex and exploiting tools from **complex analysis**

Theorem (Cauchy's integral formula)

For $\Gamma \subset \mathbb{C}$ a positively (i.e., counterclockwise) oriented simple closed curve and a complex function $f(z)$ analytic in a region containing Γ and its inside, then

(i) *if $z_0 \in \mathbb{C}$ is enclosed by Γ, $f(z_0) = -\frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z_0 - z}\, dz$;*

(ii) *if not, $\frac{1}{2\pi i} \oint_\Gamma \frac{f(z)}{z_0 - z}\, dz = 0$.*

A resolvent approach to spectral analysis

$$(\mathbf{X} - \lambda \mathbf{I}_p)\mathbf{v} = \mathbf{0} \Rightarrow \mathbf{Q}_{\mathbf{X}}(z) = (\mathbf{X} - z\mathbf{I}_n)^{-1} \tag{29}$$

- let $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$ be the spectral decomposition of $\mathbf{X}$, with $\mathbf{\Lambda} = \{\lambda_i(\mathbf{X})\}_{i=1}^p$ eigenvalues and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p] \in \mathbb{R}^{p \times p}$ the associated eigenvectors, then

$$\mathbf{Q}(z) = \mathbf{U}(\mathbf{\Lambda} - z\mathbf{I}_p)^{-1}\mathbf{U}^\mathsf{T} = \sum_{i=1}^{p} \frac{\mathbf{u}_i \mathbf{u}_i^\mathsf{T}}{\lambda_i(\mathbf{X}) - z}. \tag{30}$$

- thus, same eigenspace as $\mathbf{X}$, but maps the eigenvalues $\lambda_i(\mathbf{X})$ of $\mathbf{X}$ to $1/(\lambda_i(\mathbf{X}) - z)$.

Applying Cauchy's integral formula to the resolvent matrix $\mathbf{Q}_{\mathbf{X}}(z)$ allows one to (somewhat magically!) assess the **eigenvalue** and **eigenvector** behavior of $\mathbf{X}$:

- characterize the eigenvalues of $\mathbf{X}$, one needs to determine a $z \in \mathbb{R}$ such that $\mathbf{Q}_{\mathbf{X}}(z)$ does *not* exist.
- can be done by directly calling the Cauchy's integral formula, which allows to determine the value of a (sufficiently nice) function f at a point of interest $z_0 \in \mathbb{R}$, by integrating its "inverse" $g_f(z) = f(z)/(z_0 - z)$ on the complex plane.
- this "inverse" $g_f(z)$ is akin to the resolvent and does not, *by design*, exist at the point of interest z_0.
- in the following example, we compare the two approaches of

(i) **directly solving** the determinantal equation; and
(ii) use resolvent + Cauchy's integral formula.

A resolvent approach to spectral analysis: an example

Consider the following two-by-two real symmetric random matrix

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 \\ x_2 & x_3 \end{bmatrix} \in \mathbb{R}^{2\times 2}, \tag{31}$$

for (say independent) random variables x_1, x_2, x_3. For $\lambda_1(\mathbf{X})$ and $\lambda_2(\mathbf{X})$ the two (random) eigenvalues of $\mathbf{X}$ with associated (random) eigenvectors $\mathbf{u}_1(\mathbf{X}), \mathbf{u}_2(\mathbf{X}) \in \mathbb{R}^2$, we are interested in

$$f_{\mathbf{X}} = \mathbb{E}\left[f(\lambda_1(\mathbf{X})) + f(\lambda_2(\mathbf{X}))\right], \qquad g_{i,\mathbf{X}} = \mathbf{a}^\mathsf{T}\mathbb{E}[\mathbf{u}_i(\mathbf{X})\mathbf{u}_i(\mathbf{X})^\mathsf{T}]\mathbf{b},\ i \in \{1,2\}, \tag{32}$$

for some function $f\colon \mathbb{R} \to \mathbb{R}$ and deterministic $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$.

(i) **Directly solve** for the eigenvalues from the determinantal equation as

$$0 = \det(\mathbf{X} - \lambda \mathbf{I}_2) \Leftrightarrow \lambda(\mathbf{X}) = \frac{1}{2}\left(x_1 + x_3 \pm \sqrt{(x_1+x_3)^2 - 4(x_1x_3 - x_2^2)}\right), \tag{33}$$

and the associated eigenvectors from $\mathbf{X}\mathbf{u}_i(\mathbf{X}) = \lambda_i(\mathbf{X})\mathbf{u}_i(\mathbf{X})$, $i \in \{1,2\}$. Then compute $f_{\mathbf{X}} = \mathbb{E}\left[f(\lambda_1(\mathbf{X})) + f(\lambda_2(\mathbf{X}))\right]$, $g_{i,\mathbf{X}} = \mathbf{a}^\mathsf{T}\mathbb{E}[\mathbf{u}_i(\mathbf{X})\mathbf{u}_i(\mathbf{X})^\mathsf{T}]\mathbf{b}$

- needs to **re-compute** of the expectation for a different choice of function f and the eigen-pair $(\lambda_1(\mathbf{X}), \mathbf{u}_1(\mathbf{X}))$ or $(\lambda_2(\mathbf{X}), \mathbf{u}_2(\mathbf{X}))$ of interest.

(ii) The **resolvent** approach:

$$\begin{aligned} f_{\mathbf{X}} &= \mathbb{E}\left[f(\lambda_1(\mathbf{X})) + f(\lambda_2(\mathbf{X}))\right] \\ &= \mathbb{E}\left[-\frac{1}{2\pi\imath}\oint_\Gamma \left(\frac{f(z)}{\lambda_1(\mathbf{X})-z} + \frac{f(z)}{\lambda_2(\mathbf{X})-z}\right) dz\right] \\ &= -\frac{1}{2\pi\imath}\oint_\Gamma \mathbb{E}\left[f(z)\,\mathrm{tr}\,\mathbf{Q}_{\mathbf{X}}(z)dz\right] = -\frac{1}{2\pi\imath}\oint_\Gamma f(z)\,\mathrm{tr}\left(\mathbb{E}[\mathbf{Q}_{\mathbf{X}}(z)]\right) dz, \end{aligned}$$

for Γ a positively-oriented contour that circles around both (random) eigenvalues of $\mathbf{X}$.

- a much more **unified approach** to the quantity $f_{\mathbf{X}}$ for different choices of f
- compute the expected resolvent once (which is much simpler in the case of large random matrices)
- then perform **contour integration** with the function f of interest.
- similarly, for $g_{i,\mathbf{X}}$, it follows that

$$g_{i,\mathbf{X}} = \mathbf{a}^\mathsf{T}\mathbb{E}[\mathbf{u}_i(\mathbf{X})\mathbf{u}_i(\mathbf{X})^\mathsf{T}]\mathbf{b} = -\frac{1}{2\pi\imath}\oint_{\Gamma_i} \mathbf{a}^\mathsf{T}\mathbb{E}[\mathbf{Q}_{\mathbf{X}}(z)]\mathbf{b}\,dz \tag{34}$$

for some contour Γ_i that circles around only $\lambda_i(\mathbf{X}), i \in \{1,2\}$

- given the expected resolvent $\mathbb{E}[\mathbf{Q}(z)]$, it suffices to choose the specific contour Γ_i to get the different expressions of $g_{1,\mathbf{X}}$ and $g_{2,\mathbf{X}}$

Resolvent as the core object

Objects of interest	Functionals of resolvent $\mathbf{Q}_{\mathbf{X}}(z)$
ESD $\mu_{\mathbf{X}}$ of $\mathbf{X}$	Stieltjes transform $m_{\mu_{\mathbf{X}}}(z) = \frac{1}{p} \operatorname{tr} \mathbf{Q}_{\mathbf{X}}(z)$
Linear spectral statistics (LSS): $f(\mathbf{X}) \equiv \frac{1}{p} \sum_i f(\lambda_i(\mathbf{X}))$	Integration of trace of $\mathbf{Q}_{\mathbf{X}}(z)$: $-\frac{1}{2\pi i} \oint_\Gamma f(z) \frac{1}{p} \operatorname{tr} \mathbf{Q}_{\mathbf{X}}(z)\, dz$ (via Cauchy's integral)
Projections of eigenvectors $\mathbf{v}^\mathsf{T}\mathbf{u}(\mathbf{X})$ and $\mathbf{v}^\mathsf{T}\mathbf{U}(\mathbf{X})$ onto some given vector $\mathbf{v} \in \mathbb{R}^p$	Bilinear form $\mathbf{v}^\mathsf{T}\mathbf{Q}_{\mathbf{X}}(z)\mathbf{v}$ of $\mathbf{Q}_{\mathbf{X}}$
General matrix functional $F(\mathbf{X}) = \sum_i f(\lambda_i(\mathbf{X}))\mathbf{v}_1^\mathsf{T}\mathbf{u}_i(\mathbf{X})\mathbf{u}_i(\mathbf{X})^\mathsf{T}\mathbf{v}_2$ involving both eigenvalues and eigenvectors	Integration of bilinear form of $\mathbf{Q}_{\mathbf{X}}(z)$: $-\frac{1}{2\pi i} \oint_\Gamma f(z)\mathbf{v}_1^\mathsf{T}\mathbf{Q}_{\mathbf{X}}(z)\mathbf{v}_2\, dz$

Using the resolvent to access eigenvalue distribution

Definition (Resolvent)

For a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$, the resolvent $\mathbf{Q}_{\mathbf{X}}(z)$ of $\mathbf{X}$ is defined, for $z \in \mathbb{C}$ not an eigenvalue of $\mathbf{X}$, as

$$\mathbf{Q}_{\mathbf{X}}(z) \equiv \left(\mathbf{X} - z\mathbf{I}_p\right)^{-1}. \tag{35}$$

- let $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}}$ be the spectral decomposition of $\mathbf{X}$, with $\mathbf{\Lambda} = \{\lambda_i(\mathbf{X})\}_{i=1}^{p}$ eigenvalues and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p] \in \mathbb{R}^{p\times p}$ the associated eigenvectors, then

$$\mathbf{Q}(z) = \mathbf{U}(\mathbf{\Lambda} - z\mathbf{I}_p)^{-1}\mathbf{U}^{\mathsf{T}} = \sum_{i=1}^{p} \frac{\mathbf{u}_i\mathbf{u}_i^{\mathsf{T}}}{\lambda_i(\mathbf{X}) - z}. \tag{36}$$

- thus, same eigenspace as $\mathbf{X}$, but maps the eigenvalues $\lambda_i(\mathbf{X})$ of $\mathbf{X}$ to $1/(\lambda_i(\mathbf{X}) - z)$.
- eigenvalue of $\mathbf{Q}_{\mathbf{X}}(z)$, and the resolvent matrix itself, must explode as z approaches any eigenvalue of $\mathbf{X}$.
- take the trace $\operatorname{tr}\mathbf{Q}_{\mathbf{X}}(z)$ of $\mathbf{Q}_{\mathbf{X}}(z)$ as the quantity to "locate" the eigenvalues of the matrix $\mathbf{X}$ of interest
- for $\mu_{\mathbf{X}} \equiv \frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i(\mathbf{X})}$ the ESD of $\mathbf{X}$,

$$\boxed{\frac{1}{p}\operatorname{tr}\mathbf{Q}(z) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i(\mathbf{X}) - z} = \int \frac{\mu_{\mathbf{X}}(dt)}{t - z} \equiv m_{\mu_{\mathbf{X}}}(z)}. \tag{37}$$

Definition (Stieltjes transform)

For a real probability measure μ with support $\mathrm{supp}(\mu)$, the *Stieltjes transform* $m_\mu(z)$ is defined, for all $z \in \mathbb{C} \setminus \mathrm{supp}(\mu)$, as

$$m_\mu(z) \equiv \int \frac{\mu(dt)}{t - z}. \tag{38}$$

Proposition (Properties of Stieltjes transform, [HLN07])

For m_μ the Stieltjes transform of a probability measure μ, it holds that

(i) m_μ is complex analytic on its domain of definition $\mathbb{C} \setminus \mathrm{supp}(\mu)$;

(ii) it is bounded $|m_\mu(z)| \leq 1/\operatorname{dist}(z, \mathrm{supp}(\mu))$;

(iii) it is an increasing function on all connected components of its restriction to $\mathbb{R} \setminus \mathrm{supp}(\mu)$ (since $m'_\mu(x) = \int (t-x)^{-2} \mu(dt) > 0$) with $\lim_{x \to \pm\infty} m_\mu(x) = 0$ if $\mathrm{supp}(\mu)$ is bounded; and

(iv) $m_\mu(z) > 0$ for $z < \inf \mathrm{supp}(\mu)$, $m_\mu(z) < 0$ for $z > \sup \mathrm{supp}(\mu)$ and $\Im[z] \cdot \Im[m_\mu(z)] > 0$ if $z \in \mathbb{C} \setminus \mathbb{R}$; and

BTW, for any $\mathbf{u} \in \mathbb{R}^p$ and matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ so that $\mathrm{tr}(\mathbf{A}) = 1$, $\mathbf{u}^\mathsf{T} \mathbf{Q}_\mathbf{X}(z) \mathbf{u}$, $\mathrm{tr}(\mathbf{A}\mathbf{Q}_\mathbf{X}(z))$ are STs.

[5] Walid Hachem, Philippe Loubaton, and Jamal Najim. "Deterministic equivalents for certain functionals of large random matrices". In: *The Annals of Applied Probability* 17.3 (2007), pp. 875–930

Definition (Inverse Stieltjes transform)

For a, b continuity points of the probability measure μ, we have

$$\mu([a,b]) = \frac{1}{\pi} \lim_{y \downarrow 0} \int_a^b \Im \left[m_\mu(x + \imath y) \right] \, dx. \tag{39}$$

Besides, if μ admits a density f at x (i.e., $\mu(x)$ is differentiable in a neighborhood of x and $\lim_{\epsilon \to 0} (2\epsilon)^{-1} \mu([x-\epsilon, x+\epsilon]) = f(x)$),

$$f(x) = \frac{1}{\pi} \lim_{y \downarrow 0} \Im \left[m_\mu(x + \imath y) \right]. \tag{40}$$

Definition (Linear Spectral Statistic, LSS)

For a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$, the *linear spectral statistics* (LSS) $f_{\mathbf{X}}$ of $\mathbf{X}$ is defined as the averaged statistics of the eigenvalues $\lambda_1(\mathbf{X}), \ldots, \lambda_p(\mathbf{X})$ of $\mathbf{X}$ via some function $f\colon \mathbb{R} \to \mathbb{R}$, that is

$$f(\mathbf{X}) = \frac{1}{p}\sum_{i=1}^{p} f(\lambda_i(\mathbf{X})). \tag{41}$$

In particular, we have $= \int f(t)\mu_{\mathbf{X}}(dt)$, for $\mu_{\mathbf{X}}$ the ESD of $\mathbf{X}$.

LSS via contour integration: For $\lambda_1(\mathbf{X}), \ldots, \lambda_p(\mathbf{X})$ eigenvalues of a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$, some function $f\colon \mathbb{R} \to \mathbb{R}$ that is complex analytic in a compact neighborhood of the support $\mathrm{supp}(\mu_{\mathbf{X}})$ (of the ESD $\mu_{\mathbf{X}}$ of $\mathbf{X}$), then

$$f(\mathbf{X}) = \int f(t)\mu_{\mathbf{X}}(dt) = -\int \frac{1}{2\pi\imath}\oint_{\Gamma} \frac{f(z)\,dz}{t-z}\mu_{\mathbf{X}}(dt) = -\frac{1}{2\pi\imath}\oint_{\Gamma} f(z) m_{\mu_{\mathbf{X}}}(z)\,dz, \tag{42}$$

for *any* contour Γ that encloses $\mathrm{supp}(\mu_{\mathbf{X}})$, i.e., all the eigenvalues $\lambda_i(\mathbf{X})$.

Remark (LSS to retrieve the inverse Stieltjes transform formula):

$$\frac{1}{p}\sum_{\lambda_i(\mathbf{X})\in[a,b]}\delta_{\lambda_i(\mathbf{X})} = -\frac{1}{2\pi\imath}\oint_\Gamma 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz$$
$$= -\frac{1}{2\pi\imath}\int_{a-\varepsilon_x-\imath\varepsilon_y}^{b+\varepsilon_x-\imath\varepsilon_y} 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz - \frac{1}{2\pi\imath}\int_{b+\varepsilon_x+\imath\varepsilon_y}^{a-\varepsilon_x+\imath\varepsilon_y} 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz$$
$$-\frac{1}{2\pi\imath}\int_{a-\varepsilon_x+\imath\varepsilon_y}^{a-\varepsilon_x-\imath\varepsilon_y} 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz - \frac{1}{2\pi\imath}\int_{b+\varepsilon_x-\imath\varepsilon_y}^{b+\varepsilon_x+\imath\varepsilon_y} 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz.$$

- Since $\Re[m(x+\imath y)] = \Re[m(x-\imath y)], \Im[m(x+\imath y)] = -\Im[m(x-\imath y)]$;
- we have $\int_{a-\varepsilon_x}^{b+\varepsilon_x} m_{\mu_\mathbf{X}}(x-\imath\varepsilon_y)\,dx + \int_{b+\varepsilon_x}^{a-\varepsilon_x} m_{\mu_\mathbf{X}}(x+\imath\varepsilon_y)dx = -2\imath\int_{a-\varepsilon_x}^{b+\varepsilon_x}\Im[m_{\mu_\mathbf{X}}(x+\imath\varepsilon_y)]\,dx$;
- and consequently $\mu([a,b]) = \frac{1}{p}\sum_{\lambda_i(\mathbf{X})\in[a,b]}\lambda_i(\mathbf{X}) = \frac{1}{\pi}\lim_{\varepsilon_y\downarrow 0}\int_a^b \Im[m_{\mu_\mathbf{X}}(x+\imath\varepsilon_y)]\,dx$.

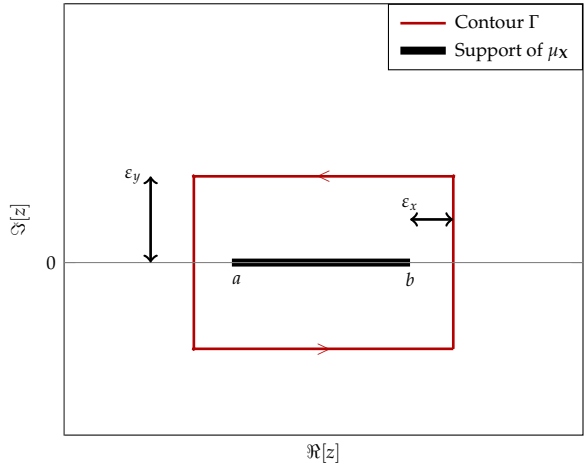


Figure: Illustration of a rectangular contour Γ and support of $\mu_{\mathbf{X}}$ on the complex plane.

Spectral functionals via resolvent

Definition (Matrix spectral functionals)

For a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$, we say $F\colon \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ is a matrix spectral functional of $\mathbf{X}$,

$$F(\mathbf{X}) = \sum_{i\in\mathcal{I}\subseteq\{1,\dots,p\}} f(\lambda_i(\mathbf{X}))\mathbf{u}_i\mathbf{u}_i^\mathsf{T}, \quad \mathbf{X} = \sum_{i=1}^{p} \lambda_i(\mathbf{X})\mathbf{u}_i\mathbf{u}_i^\mathsf{T}. \tag{43}$$

Spectral functional via contour integration: For $\mathbf{X} \in \mathbb{R}^{p\times p}$, resolvent $\mathbf{Q}_\mathbf{X}(z) = (\mathbf{X} - z\mathbf{I}_p)^{-1}$, $z \in \mathbb{C}$, and $f\colon \mathbb{R} \to \mathbb{R}$ analytic in a neighborhood of the contour $\Gamma_\mathcal{I}$ that circles around the eigenvalues $\lambda_i(\mathbf{X})$ of $\mathbf{X}$ with their indices in the set $\mathcal{I} \subseteq \{1,\dots,p\}$,

$$F(\mathbf{X}) = -\frac{1}{2\pi\imath} \oint_{\Gamma_\mathcal{I}} f(z)\mathbf{Q}_\mathbf{X}(z)\,dz. \tag{44}$$

Example: access to the i-th eigenvector $\mathbf{u}_i$ of $\mathbf{X}$ through

$$\mathbf{u}_i\mathbf{u}_i^\mathsf{T} = -\frac{1}{2\pi\imath} \oint_{\Gamma_{\lambda_i(\mathbf{X})}} \mathbf{Q}_\mathbf{X}(z)\,dz, \tag{45}$$

for $\Gamma_{\lambda_i(\mathbf{X})}$ a contour circling around $\lambda_i(\mathbf{X})$ only, so eigenvector projection $(\mathbf{v}^\mathsf{T}\mathbf{u}_i)^2 = -\frac{1}{2\pi\imath}\oint_{\Gamma_{\lambda_i(\mathbf{X})}} \mathbf{v}^\mathsf{T}\mathbf{Q}_\mathbf{X}(z)\mathbf{v}\,dz$.

Take-away messages of this section

- "basic" probability: **concentration** of scalar observations of large random vectors: simple and involved, linear and nonlinear objects
- boils down to **expectation computation/evaluation**
- **same** holds for scalar observations of large random matrices
- linear algebra: matrix norm "equivalence" but up to **dimensional factors**
- **resolvent** (i.e., regularized inverse) naturally appears in eigenvalue/eigenvector assessment
- a unified **resolvent**-based to eigenspectral analysis of (not necessarily random) matrices: Cauchy's integral formula, Stieltjes transform (and its inverse), Linear Spectral Statistic, and generic matrix spectral functionals, etc.

Thank you! Q & A?