Mathematics of Deep Learning
Lecture 5: Introduction to the asymptotic analysis of random matrices

Bruno Loureiro
Département d'Informatique, École Normale Supérieure - PSL & CNRS, France

14/02/2025

Get in touch at: bruno.loureiro@di.ens.fr

Notation

We denote finite discrete sets as $[n] \coloneqq \{1, \dots, n\}$. We denote vectors by bold lower letters $\boldsymbol{v} \in \mathbb{R}$ and matrices by bold capital letters $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, and their elements by v_i and A_{ij}. The spectrum of a symmetric square matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ as $\mathrm{spec}(\boldsymbol{M}) \coloneqq \{\lambda_1, \dots, \lambda_d\} \subset \mathbb{R}$ which we always assume is ordered $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. We denote by $||\boldsymbol{A}||_{\mathtt{op}}$, $||\boldsymbol{A}||_*$ and $||\boldsymbol{A}||_{\mathtt{F}}$ the operator, nuclear and Frobenius norm, respectively.

1 Motivation

Random matrix theory is the area of mathematics concerned with the properties of matrices with random entries. This is a vast field, and here we will deliberately limit the discussion to the aspects which are relevant to asymptotic analysis of ridge regression.

We have seen that the analysis of the risk of ridge regression boils down to the investigation of the following bias and variance terms:

$$B(\boldsymbol{\theta}_\star, \boldsymbol{X}, \lambda) = \lambda^2 \,\mathrm{Tr}\left\{\boldsymbol{\theta}_\star\boldsymbol{\theta}_\star^\top \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{\Sigma} \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1}\right\} \tag{1.1}$$

$$V(\boldsymbol{X}, \lambda, \sigma^2) = \frac{\sigma^2}{n}\,\mathrm{Tr}\left\{\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1}\right\} - \frac{\lambda\sigma^2}{n}\,\mathrm{Tr}\left\{\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-2}\right\} \tag{1.2}$$

where $\hat{\boldsymbol{\Sigma}}_n = {}^1\!/_n \boldsymbol{X}^\top \boldsymbol{X}$ is the sample covariance matrix and we assumed the data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ has i.i.d. rows $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. The expressions above take one of the following two forms:

$$\mathrm{Tr}\left\{\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{M}}(z)\right\}, \qquad \mathrm{Tr}\left\{\boldsymbol{A}\boldsymbol{R}_{\boldsymbol{M}}(z)\boldsymbol{B}\boldsymbol{R}_{\boldsymbol{M}}(z)\right\} \tag{1.3}$$

where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$ are deterministic matrices and $\boldsymbol{R}(z) \in \mathbb{R}^{d\times d}$ is the resolvent matrix associated to the symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$:

$$\boldsymbol{R}_{\boldsymbol{M}}(z) = (\boldsymbol{M} - z\boldsymbol{I}_d)^{-1}, \qquad z \in \mathbb{C} \setminus \mathrm{spec}(\boldsymbol{M}) \tag{1.4}$$

Note the resolvent is a symmetric matrix defined everywhere in the complex plane, except at the values in the real axis where the eigenvalues of $\boldsymbol{M}$ lie. It contains useful information on both the spectrum

of $\boldsymbol{M}$ and its eigenvectors, since it can be diagonalised on the same basis of $\boldsymbol{M}$:

$$\boldsymbol{R}_{\boldsymbol{M}}(z) = \sum_{i=1}^{d} \frac{\boldsymbol{u}_i \boldsymbol{u}_i^\top}{\lambda_i - z} \tag{1.5}$$

where $\boldsymbol{u}_i$ are the eigenvectors of $\boldsymbol{M}$. Therefore, by the Cauchy integral formula we have:

$$f(\lambda_j) = -\int_{\gamma_j} \frac{\mathrm{d}z}{2\pi i} f(z) \boldsymbol{R}_{\boldsymbol{M}}(z), \qquad \boldsymbol{u}_j \boldsymbol{u}_j^\top = -\int_{\gamma_j} \frac{\mathrm{d}z}{2\pi i} \boldsymbol{R}_{\boldsymbol{M}}(z) \tag{1.6}$$

where γ_j is a contour circling the eigenvalue λ_j and f is any analytic function. Therefore, the traces in eq. (1.3) can be seen as extracting some statistics about eigenvalues or eigenvectors of $\boldsymbol{M}$ through the resolvent.

Our goal in this lecture is to introduce the main concepts and tools used to study the asymptotic behaviour of these quantities when $d \to \infty$.

2 Main notions

Let $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ denote a symmetric random matrix and $\mathrm{spec}(\boldsymbol{M}) = \{\lambda_1, \dots, \lambda_d\} \subset \mathbb{R}$ its real eigenvalues. These are random variables, and one of the central goals of random matrix theory is to characterise their statistics. Studying the statistics of single eigenvalues can be challenging, so instead we turn to studying large aggregates of eigenvalues.

Definition 1 (Empirical spectral measure)**.** Let $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ denote a symmetric matrix with eigenvalues $\mathrm{spec}(\boldsymbol{M}) = \{\lambda_1, \dots, \lambda_d\} \subset \mathbb{R}$. The empirical spectral measure of $\boldsymbol{M}$ is defined as:

$$\hat{\mu}_{\boldsymbol{M}} = \frac{1}{d} \sum_{i=1}^{d} \delta_{\lambda_i} \tag{2.1}$$

Note that $\int \hat{\mu}_{\boldsymbol{M}}(\mathrm{d}x) = 1$ and therefore $\hat{\mu}_{\boldsymbol{M}}$ is also probability measure.

Remark 1. The empirical spectral measure is the normalised counting measure of how many eigenvalues lie in an interval:

$$\hat{\mu}_{\boldsymbol{M}}([a, b]) = \frac{1}{d}\{\#\text{ eigenvalues } \lambda_i \in [a, b]\}. \tag{2.2}$$

Since $\hat{\mu}$ is a measure, studying its limiting behaviour is mathematically tricky. Instead, it is convenient to study its *Stieltjes transform*:

Definition 2 (Stieltjes transform)**.** Let μ denote a probability measure over $\mathbb{R}$. The *Stieltjes transform* of μ is defined as:

$$s_\mu(z) = \int \frac{\mu(\mathrm{d}\lambda)}{\lambda - z}, \qquad z \in \mathbb{C} \setminus \mathrm{supp}(\mu) \tag{2.3}$$

The Stieltjes transform can be thought as an "integrated version" of the measure. It satisfies some useful properties:

Properties 1. The Stieltjes transform satisfies the following properties:

(a) **Boundness:** For all $z \in \mathbb{C} \setminus \mathrm{supp}(\mu)$:

$$|s_\mu(z)| \leq \frac{1}{\mathrm{dist}(z, \mathrm{supp}(\mu))} \leq \frac{1}{|\operatorname{Im}(z)|} \tag{2.4}$$

This is straightforward to show. Letting $z = x + i\eta$ with $\eta > 0$:

$$|s_\mu(z)| \leq \int \mu_{\boldsymbol{M}}(\mathrm{d}\lambda) \left|\frac{1}{\lambda - z}\right| = \int \mu_{\boldsymbol{M}}(\mathrm{d}\lambda) \frac{1}{\sqrt{(\lambda - x)^2 + \eta^2}} \leq \int \mu_{\boldsymbol{M}}(\mathrm{d}\lambda) \frac{1}{\eta} = \frac{1}{\eta} \tag{2.5}$$

(b) **Sign:** We have that $\operatorname{Im}(z)\operatorname{Im}(s_\mu(z)) \geq 0$ since:

$$\eta \operatorname{Im}[s_\mu(x+i\eta)] = \eta \int \mu(\mathrm{d}\lambda) \operatorname{Im}\left[\frac{1}{\lambda-(x+i\eta)}\right] = \frac{1}{\pi}\int \mu(\mathrm{d}\lambda)\frac{\eta^2}{(\lambda-x)^2+\eta^2} \geq 0 \tag{2.6}$$

(c) **Derivatives:** The Stieltjes transform is an analytic function on $\mathbb{C}\setminus\operatorname{supp}(\mu)$. In particular, it is infinitely differentiable, and the derivatives can be similarly bounded:

$$|s^{(k)}(z)| \leq \frac{(k-1)!}{|\operatorname{Im}(z)|^{k+1}} \tag{2.7}$$

In particular, this implies $s_\mu(x)$ is an increasing function on all connected components of $x \in \mathbb{R}\setminus\operatorname{supp}(\mu)$ since $s'_\mu(x) > 0$.

(d) **Moments:** Suppose $\operatorname{supp}(\mu) = [-M, M]$ has bounded support. Then, for any $|z| > M$ we can Taylor expand:

$$\begin{aligned} s_\mu(z) = \int \frac{\mu(\mathrm{d}\lambda)}{\lambda - z} &= -\frac{1}{z}\int \left(1-\frac{\lambda}{z}\right)^{-1}\mu(\mathrm{d}\lambda) = -\frac{1}{z}\int \mu(\mathrm{d}\lambda)\sum_{k=0}^{\infty}\left(\frac{\lambda}{z}\right)^k \\ &= -\sum_{k=0}^{\infty} z^{-(k+1)}\mathbb{E}_{X\sim\mu}[X^k] \end{aligned} \tag{2.8}$$

Therefore, the moments $\mathbb{E}_{X\sim\mu}[X^k]$ of the distribution μ are given by the coefficients of the Taylor expansion of $\tilde{s}_\mu(u) = s_\mu(1/z)$. This means the Stieltjes transform can also be seen as a moment generating function for μ.

(e) **Large z asymptotics:** As a corollary of the previous point, we have:

$$s_\mu(z) = -\frac{1}{z} + O(|z|^{-2}), \quad \text{as } |z| \to \infty \tag{2.9}$$

Since the moments of μ can be computed from s_μ, property (d) suggests that we can fully reconstruct the probability measure from the Stieltjes transform. This intuition is indeed the case, and is formalised by the *Stieltjes-Perron inversion formula*:

Proposition 1 (Stieltjes-Perron inversion formula)**.** Let a, b denote continuity points of the probability measure μ. Then:

$$\mu([a,b]) = \frac{1}{\pi}\lim_{\eta\to 0^+}\int_a^b \operatorname{Im}[s_\mu(x+i\eta)]\mathrm{d}x. \tag{2.10}$$

In the case where μ admits a density f at x, this is simply given by:

$$f(x) = \lim_{\eta\to 0^+}\operatorname{Im}[s_\mu(x+i\eta)]. \tag{2.11}$$

Sketch of the proof. The main idea of the proof is to note that:

$$\frac{1}{\pi}\operatorname{Im}[s_\mu(x+i\eta)] = \frac{1}{\pi}\int \mu(\mathrm{d}\lambda)\operatorname{Im}\left[\frac{1}{\lambda-(x+i\eta)}\right] = \frac{1}{\pi}\int \mu(\mathrm{d}\lambda)\frac{\eta}{(\lambda-x)^2+\eta^2} \tag{2.12}$$

This has an interesting interpretation: the imaginary part of the Stieltjes transform can be seen as a convolution between μ and the Cauchy distribution[1]. To see this, let $X \sim \mu$ and $Y \sim$ Cauchy. Then, $X + tY$ has density exactly given by $\frac{1}{\pi}\operatorname{Im}[s_\mu(x+i\eta)]$, and for any $f : \mathbb{R} \to \mathbb{R}$:

$$\mathbb{E}[f(X+Y)] = \int \frac{\mathrm{d}x}{\pi} f(x)\operatorname{Im}[s_\mu(x+i\eta)] \tag{2.13}$$

As a consequence, the density $\frac{1}{\pi}\operatorname{Im}[s_\mu(x+i\eta)]$ converges weakly to the density μ as $\eta \to 0^+$. □

[1] Recall the Cauchy distribution is a continuous probability distribution with probability density function $\text{Cauchy}(x) = \frac{1}{\pi(1+x^2)}$ supported on $\mathbb{R}$.

The last property we will need from the Stieltjes transform is that it not only captures everything we need about the measure, but this also transfer under limits.

Proposition 2. Let μ_n denote a sequence of probability measures on $\mathbb{R}$. Then, μ_n converges weakly to μ if and only if s_{μ_n} converges point-wise to s_μ in the upper-half complex plane $\mathcal{C}_+ = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$.

Remark 2. Proposition 2 also holds for random sequences of probability measures μ_n under almost surely convergence and convergence in probability in both sides of the "if and only if" statement.

⚠ We can have sequences of probability measures μ_n that do not converge, but for which $s_{\mu_n}(z) \to s(z)$ not corresponding to the Stieltjes transform of a probability measure. For example:

$$\mu_n = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_n \tag{2.14}$$

We have:

$$s_{\mu_n}(z) = -\frac{1}{2z} + \frac{1}{2}\frac{1}{n-z} \to -\frac{1}{2z} \tag{2.15}$$

So far, our discussion of the Stieltjes transform has been for a generic probability distribution over $\mathbb{R}$. We now turn our attention to the particular case when this is the empirical spectral measure $\hat{\mu}$ of a symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$. In this case, note the Stieltjes transform can also be written as:

$$s_{\hat{\mu}}(z) = \int \frac{\mathrm{d}\hat{\mu}(\lambda)}{\lambda - z} = \frac{1}{d}\sum_{i=1}^{d}\frac{1}{\lambda_i - z} = \frac{1}{d}\,\mathrm{Tr}(\boldsymbol{M} - z\boldsymbol{I}_d)^{-1} =: \frac{1}{d}\,\mathrm{Tr}\,\boldsymbol{R}_{\boldsymbol{M}}(z) \tag{2.16}$$

which shows that the Stieltjes transform of the empirical spectral measure is nothing but the empirical average diagonal of the resolvent matrix, connecting back to the discussion in section 1. In particular, if $\boldsymbol{M}$ is a random matrix, $\hat{\mu}$ is a random probability measure and $s_{\hat{\mu}}(z)$ a random complex function. The fact that the Stieltjes transform can be seen as the empirical average suggests it is prompt to concentration in the high-dimensional limit $d \to \infty$. This will play a key role in the analysis that follow.

3 The Wigner semi-circle law

So far, our discussion has been general. We now turn our attention to a particular example: the Gaussian Orthogonal Ensemble (GOE).

Definition 3 (Gaussian Orthogonal Ensemble)**.** The GOE is the ensemble of random symmetric matrices $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ with independent upper-triangular Gaussian entries:

$$\boldsymbol{W} \sim \texttt{GOE}(d) \quad \Leftrightarrow \quad \begin{cases} W_{ii} \sim \mathcal{N}(0, 2/n), & 1 \le i \le d \\ W_{ij} \sim \mathcal{N}(0, 1/n), & 1 \le i < j \le d \end{cases} \tag{3.1}$$

Remark 3. An alternative and equivalent characterisation of $\boldsymbol{W} \sim \boldsymbol{W} \sim \texttt{GOE}(d)$ is:

$$\boldsymbol{W} = \frac{1}{\sqrt{2d}}(\boldsymbol{G} + \boldsymbol{G}^\top) \tag{3.2}$$

where $\boldsymbol{G} \in \mathbb{R}^{d\times d}$ is a matrix with Gaussian i.i.d. entries $G_{ij} \sim \mathcal{N}(0, 1)$.

Properties 2. GOE matrices $\boldsymbol{W} \sim \texttt{GOE}(d)$ satisfy the following useful properties:

(a) **Normalisation:**

$$\frac{1}{d}\operatorname{Tr}\boldsymbol{W}^2 \xrightarrow{a.s.} 1, \qquad d \to \infty \tag{3.3}$$

Since:

$$\begin{aligned}\frac{1}{d}\mathbb{E}\left[\operatorname{Tr}\boldsymbol{W}^2\right] &= \frac{1}{d}\mathbb{E}\left[\sum_{i,j=1}^{d} W_{ij}W_{ji}\right] = \frac{1}{d}\sum_{i,j=1}^{d}\mathbb{E}[W_{ij}^2] = \frac{1}{n}\sum_{i=1}^{d}\mathbb{E}[W_{ii}^2] + \frac{1}{n}\sum_{i\neq j}^{d}\mathbb{E}[W_{ij}^2] \\ &= \frac{2}{n} + \frac{n-1}{n} = 1 + \frac{1}{n} \xrightarrow{d\to\infty} 1\end{aligned} \tag{3.4}$$

(b) **Norms:** An entry of $\boldsymbol{W} \sim \texttt{GOE}(d)$ has a typical magnitude $W_{ij} = O(1/\sqrt{n})$. From property (a), we have:

$$||\boldsymbol{W}||_{\texttt{F}} \coloneqq \sqrt{\operatorname{Tr}[\boldsymbol{W}^\top \boldsymbol{W}]} = O(\sqrt{d}) \tag{3.5}$$

Similarly, since the eigenvalues must sum to $\operatorname{Tr}\boldsymbol{W} = \sum_{i=1}^{d}\lambda_i = O(d)$ and their average square distance from the origin $1/d \operatorname{Tr}\boldsymbol{W}^2 = 1/d\sum_{i=1}^{d}\lambda_i^2 = O(1)$, it suggests that we have:

$$||\boldsymbol{W}||_{\texttt{op}} = O(1). \tag{3.6}$$

Although this heuristic is correct, proving it is not trivial. See Chapter 2.3 in Terence Tao's notes. Together, this implies that the typical spacing between eigenvalues is $O(1/d)$.

(c) **Rotation invariance:** Let $\boldsymbol{U} \in O(d)$ be an orthogonal matrix $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_d$ and $\boldsymbol{W} \sim \texttt{GOE}(d)$. Then, $\boldsymbol{U}\boldsymbol{W}\boldsymbol{U}^\top \sim \texttt{GOE}(d)$ (i.e. $\boldsymbol{U}\boldsymbol{W}\boldsymbol{U}^\top \overset{d}{=} \boldsymbol{W}$). This follows from eq. (3.2) the rotational invariance of Gaussian matrices.

The asymptotic behaviour of the empirical spectral measure of GOE matrices was derived by Physicist Eugene Wigner in (Wigner, 1955), who used GOE matrices to model the Hamiltonian of complex nuclei. This is given by the celebrated *Wigner semi-circle law*, see fig. 1 for an illustration.

Theorem 1 (Wigner semi-circle law)**.** Let $\boldsymbol{W} \sim \texttt{GOE}(d)$. Then, in the limit $d \to \infty$ the Stieltjes transform $s_{\hat{\mu}}$ of the empirical spectral density $\hat{\mu}_{\boldsymbol{W}}$ converges almost surely to a deterministic function:

$$s_{\hat{\mu}}(z) = \frac{1}{d}\operatorname{Tr}\left(\boldsymbol{W} - z\boldsymbol{I}_d\right)^{-1} \xrightarrow{a.s.} s_\mu(z) = \frac{\sqrt{z^2-4}-z}{2}, \quad z \in \mathbb{C}\setminus[-2,2]. \tag{3.7}$$

In particular, s_μ corresponds to the Stieltjes transform of a probability measure with a density supported at $[-2,2]$, known as the Wigner semi-circle law:

$$\mu(\mathrm{d}x) = \frac{\sqrt{4-x^2}}{2\pi}\mathbf{1}_{[-2,2]}(x)\mathrm{d}x \tag{3.8}$$

Which thanks to Proposition 2 implies the almost sure weak convergence of $\hat{\mu}_{\boldsymbol{W}}$ to μ as $d \to \infty$.

Sketch of the proof. The proof is separated in two steps:

(I) Show that the Stieltjes transform of the empirical spectral measure $s_{\hat{\mu}}$ concentrates almost surely to its expectation:

$$s_{\hat{\mu}}(z) \xrightarrow{a.s.} \mathbb{E}[s_{\hat{\mu}}(z)], \qquad \text{as } d \to \infty \tag{3.9}$$

(II) Show that:

$$\mathbb{E}[s_{\hat{\mu}}(z)] \to s_\mu(z), \qquad \text{as } d \to \infty \tag{3.10}$$

where s_μ can be exactly characterised as the solution of a quadratic equation.

Part I: concentration — To show concentration, we start by showing that the Stieltjes transform $s_{\hat{\mu}}(z)$ is a Lipschitz continous function of the Gaussian matrix $\sqrt{2d}\boldsymbol{G}$:

Lemma 1. Let $\boldsymbol{W} \sim \texttt{GOE}(d)$ denote a GOE matrix:

$$\boldsymbol{W} = \frac{\boldsymbol{G} + \boldsymbol{G}^\top}{\sqrt{2d}} \tag{3.11}$$

where $\boldsymbol{G}$ is a Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Then, for any $z \in \mathbb{C} \setminus \text{spec}(\boldsymbol{W})$, the Stieltjes transform $s_{\hat{\mu}_{\boldsymbol{W}}}(z)$ of the the empirical spectral distribution of $\boldsymbol{W}$ is a Lipschitz continuous function of $\sqrt{2d}\boldsymbol{G}$ with Lipschitz constant given by:

$$L_d(z) \coloneqq \frac{1}{d\,\text{Im}(z)^2} \tag{3.12}$$

Sketch of the proof. This follows from a combination of matrix identities and inequalities for the different matrix norms:

$$\begin{aligned}
|s_{\hat{\mu}_{\boldsymbol{W}}}(z) - s_{\hat{\mu}_{\boldsymbol{W}'}}(z)| &= \left| \frac{1}{d} \text{Tr}(\boldsymbol{W} - z\boldsymbol{I}_d)^{-1} - \frac{1}{d} \text{Tr}\big(\boldsymbol{W}' - z\boldsymbol{I}_d\big)^{-1} \right| \\
&\overset{(a)}{=} \left| \frac{1}{d} \text{Tr}\left\{ (\boldsymbol{W} - z\boldsymbol{I}_d)^{-1} (\boldsymbol{W} - \boldsymbol{W}') (\boldsymbol{W}' - z\boldsymbol{I}_d)^{-1} \right\} \right| \\
&\overset{(b)}{\leq} \frac{1}{d} ||(\boldsymbol{W} - z\boldsymbol{I}_d)^{-1}||_{\texttt{op}} ||(\boldsymbol{W}' - z\boldsymbol{I}_d)^{-1}||_{\texttt{op}} ||\boldsymbol{W} - \boldsymbol{W}||_* \\
&\overset{(d)}{\leq} \frac{2}{\sqrt{d}} \frac{1}{\text{Im}(z)^2} ||\boldsymbol{G} - \boldsymbol{G}'||_{\texttt{F}} \\
&= \frac{1}{d} \frac{1}{\text{Im}(z)^2} ||\sqrt{2d}\boldsymbol{G} - \sqrt{2d}\boldsymbol{G}'||_{\texttt{F}}
\end{aligned} \tag{3.13}$$

where in (a) we used the resolvent identity:

$$\boldsymbol{A}^{-1} - \boldsymbol{B}^{-1} = \boldsymbol{A}^{-1}(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{B}^{-1}, \tag{3.14}$$

in (b) we used that:

$$\text{Tr}(\boldsymbol{A}\boldsymbol{B}) \leq ||\boldsymbol{A}||_* \cdot ||\boldsymbol{B}||_{\texttt{op}} \tag{3.15}$$

and finally in (c) we used the fact that the operator norm is exactly the spectral radius, i.e. for $z \in \mathbb{C} \setminus \text{spec}(\boldsymbol{W})$:

$$\begin{aligned}
||\boldsymbol{R}_{\boldsymbol{W}}(z)||_{\texttt{op}} &\coloneqq \sup_{\xi \in \text{spec}((\boldsymbol{W} - z\boldsymbol{I}_d)^{-1})} |\xi| \\
&= \sup_{\lambda \in \text{spec}(\boldsymbol{W})} \frac{1}{|\lambda - z|} \\
&= \frac{1}{\text{dist}(z, \text{spec}(\boldsymbol{W}))} \leq \frac{1}{\text{Im}(z)}
\end{aligned} \tag{3.16}$$

together with the following bound:

$$||\boldsymbol{W}||_* \leq \sqrt{d} ||\boldsymbol{W}||_{\texttt{F}} \tag{3.17}$$

□

This allow us to apply standard tail bounds for Lipschitz functions of Gaussian variables:

Lemma 2 (Theorem 5.6 in (Boucheron et al., 2013))**.** Let $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ be a vector with i.i.d. standard Gaussian entries, and consider $f : \mathbb{R}^d \to \mathbb{R}$ an L-Lipschitz function. Then, for all $t > 0$:

$$\mathbb{P}[|f(\boldsymbol{g}) - \mathbb{E}[f(\boldsymbol{g})]| \geq t] \leq e^{-\frac{t^2}{2L^2}} \tag{3.18}$$

Applying this result to the Stieltjes transform, for any $t > 0$ and $z \in \mathbb{C} \setminus \text{supp}(\boldsymbol{W})$:

$$\mathbb{P}\left[|s_{\hat{\mu}}(z) - \mathbb{E}[s_{\hat{\mu}}(z)]| \geq t\right] \leq e^{-\frac{t^2}{2L_d(z)^2}} = e^{-\frac{d^2 \,\text{Im}(z)^4 t^2}{2}} \tag{3.19}$$

In particular, this implies convergence in probability as $d \to \infty$ for $|z| = O(1)$. This can be made stronger by using the following lemma:

Lemma 3. Let $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ denote two independent random vectors from the same distribution with finite variance, and consider a L-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$. Then:

$$\text{Var}(f(\boldsymbol{x})) \leq L^2 \text{Var}(\boldsymbol{x}) \tag{3.20}$$

Sketch of the proof. This is a consequence of a simple calculation:

$$\begin{aligned}
2\text{Var}(f(\boldsymbol{x})) &\overset{(a)}{=} 2\text{Var}\left(f(\boldsymbol{x}) - f(\boldsymbol{x}')\right) \\
&\overset{\text{def.}}{=} \mathbb{E}\left[(f(\boldsymbol{x}) - f(\boldsymbol{x}'))^2\right] - \mathbb{E}\left[f(\boldsymbol{x}) - f(\boldsymbol{x}')\right]^2 \\
&\overset{(b)}{=} \mathbb{E}\left[(f(\boldsymbol{x}) - f(\boldsymbol{x}'))^2\right] \\
&\overset{(c)}{\leq} L^2 \mathbb{E}\left[||\boldsymbol{x} - \boldsymbol{x}'||_2^2\right] \\
&= 2L^2 \text{Var}(\boldsymbol{x})
\end{aligned} \tag{3.21}$$

where in (a) we used that $\boldsymbol{x}, \boldsymbol{x}'$ are independent (and therefore uncorrelated) to write $\text{Var}(f(\boldsymbol{x}) - f(\boldsymbol{x}')) = \text{Var}(f(\boldsymbol{x})) + \text{Var}(-f(\boldsymbol{x}')) = 2\text{Var}(f(x))$, in (b) we used that $\boldsymbol{x}, \boldsymbol{x}'$ are identically distributed and in (c) the fact f is Lipschitz. □

Applying lemma 3 to the Stieltjes transform:

$$\text{Var}(s_{\hat{\mu}}(z)) \leq L_d(z)^2 = \frac{1}{d^2} \frac{1}{\text{Im}(z)^4} \tag{3.22}$$

which implies almost sure convergence:

$$s_{\hat{\mu}}(z) \xrightarrow{a.s.} \mathbb{E}[s_{\hat{\mu}}(z)], \qquad \text{as } d \to \infty \tag{3.23}$$

at fixed $|z| = O(1)$.

Part II: asymptotic characterisation — We now turn to the second part of the proof, which consists of finding an exact characterisation of the limit $\mathbb{E}[s_{\hat{\mu}}(z)] \to s_\mu(z)$. We will follow a proof scheme known as *leave-one-out*. The key idea is to note that since the entries of $\boldsymbol{W}$ are i.i.d., any sub-matrix of size $(d-1) \times (d-1)$ is still a GOE matrix. Moreover, for large d the properties of the submatrix should be almost the same as the ones of the full matrix. In mathematical terms, we fix an $i \in [d]$ and define the square matrix $\boldsymbol{W}^{(i)} \in \mathbb{R}^{(d-1)\times(d-1)}$ obtained by deleting the corresponding row and column:

$$\boldsymbol{W}^{(i)}_{jk} = (W_{jk})_{j,k \neq i} \tag{3.24}$$

As we said before, this is also a GOE matrix up to an adjustment of the normalisation:

$$\sqrt{\frac{d}{d-1}}\boldsymbol{W}^{(i)} \sim \texttt{GOE}(d-1). \tag{3.25}$$

The key idea now is to relate the Stieltjes transform of $\boldsymbol{W}$ to that of $\boldsymbol{W}^{(i)}$. To ease the notation, let s_d and s_{d-1} denote the Stieltjes transforms associated to $\boldsymbol{W}$ and $\boldsymbol{W}^{(i)}$, respectively. By definition, we have:

$$\mathbb{E}[s_d(z)] = \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d}(\boldsymbol{W}-z\boldsymbol{I}_d)^{-1}_{ii}\right] = \mathbb{E}\left[(\boldsymbol{W}-z\boldsymbol{I}_d)^{-1}_{ii}\right] \tag{3.26}$$

where in the last equality we used all elements in the diagonal have the same distribution (i.e. the expectation is independent of i). Now, we can use the expression for the inverse of a block matrix:

$$\begin{bmatrix} W_{11} & \boldsymbol{w}_1 \\ \boldsymbol{w}_1^\top & \boldsymbol{W}^{(1)} \end{bmatrix}^{-1} = \frac{1}{W_{11} - z - \boldsymbol{w}_1^\top(\boldsymbol{W}^{(1)}-z\boldsymbol{I}_d)^{-1}\boldsymbol{w}_1} \tag{3.27}$$

where for concreteness we wrote for $i=1$, but the above is true for any minor of $\boldsymbol{W}$. Therefore:

$$\mathbb{E}[s_d(z)] = \mathbb{E}\left[\frac{1}{W_{11} - z - \boldsymbol{w}_1^\top(\boldsymbol{W}^{(1)}-z\boldsymbol{I}_d)^{-1}\boldsymbol{w}_1}\right] \overset{(a)}{=} \mathbb{E}\left[\frac{1}{-z - \boldsymbol{w}_1^\top(\boldsymbol{W}^{(1)}-z\boldsymbol{I}_d)^{-1}\boldsymbol{w}_1}\right] + o(1) \tag{3.28}$$

where in (a) we used the fact that $W_{11} = O(1/\sqrt{d})$ is subleading with respect to the remaining terms, which are $O(1)$. Note that the row $\boldsymbol{w}_1$ is a random vector with covariance $1/d\boldsymbol{I}_d$, and is independent from $\boldsymbol{W}^{(1)}$. Therefore, we should expect that:

$$\boldsymbol{w}_1^\top(\boldsymbol{W}^{(1)}-z\boldsymbol{I}_d)^{-1}\boldsymbol{w}_1 = \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\Big(\boldsymbol{W}^{(1)}-z\boldsymbol{I}_d\Big)^{-1} + o(1) =: \mathbb{E}[s_{d-1}(z)] + o(1) \tag{3.29}$$

Finally, we also expect both s_{d-1}, s_d to concentrate (due to Part I) to the same value when $d\to\infty$, in other words:

$$s_d(z) = s_{d-1}(z) + o(1) \xrightarrow{a.s.} s_\mu(z) \tag{3.30}$$

Putting together, we expect $s_\mu(z)$ to satisfy:

$$s(z) = \frac{1}{-z - s(z)}. \tag{3.31}$$

This quadratic equation has two solutions:

$$s_\pm(z) = \frac{-z \pm \sqrt{z^2-4}}{2}. \tag{3.32}$$

and it is easy to check that only $s_\mu(z) = s_+(z)$ satisfy the properties 1 of the Stieltjes transform, as claimed in eq. (3.7).

Now, to make the informal argument above rigorous, we need to justify both eq. (3.29) and eq. (3.30). The first is a consequence of Hanson-Wright inequality:

Theorem 2 (Hanson-Wright inequality)**.** Let $\boldsymbol{x}\in\mathbb{R}^d$ be a random vector with independent, mean zero sub-Gaussian coordinates, and let $\boldsymbol{A}\in\mathbb{R}^{d\times d}$ denote a deterministic matrix. Then, for every $t\geq 0$ we have:

$$\mathbb{P}\Big[|\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}]| \geq t\Big] \leq 2e^{-c\min\left\{\frac{t^2}{K^4||\boldsymbol{A}||_{\mathrm{F}}^2}, \frac{t}{K^2||\boldsymbol{A}||_{\mathrm{op}}}\right\}} \tag{3.33}$$

where $K = \max_i ||x_i||_{\psi_2}$.

See Theorem 6.2.1 in Vershynin (2018) for a detailed discussion. Recalling that in our case we have:

$$||\boldsymbol{R}_{\boldsymbol{W}}(z)||_{\mathtt{op}} \leq \frac{1}{\mathrm{Im}(z)}, \qquad ||\boldsymbol{R}_{\boldsymbol{W}}(z)||_{\mathtt{F}} \leq \frac{\sqrt{d}}{\mathrm{Im}(z)} \tag{3.34}$$

This give us:

$$\left|\boldsymbol{w}_1^\top (\boldsymbol{W}^{(1)} - z\boldsymbol{I}_d)^{-1}\boldsymbol{w}_1 - \mathbb{E}[s_{d-1}(z)]\right| = O\left(\sqrt{\frac{\log d}{d}}\right) \tag{3.35}$$

Finally, eq. (3.30) is justified by Weyl's interlacing lemma:

Lemma 4 (Weyl's interlacing lemma)**.** Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$ be two symmetric matrices with eigenvalues $(\lambda_i(\boldsymbol{A}))_{i\in[d]}$, $(\lambda_i(\boldsymbol{B}))_{i\in[d]}$ arranged in nonincreasing order. The, for all $i \in [n]$:

$$\lambda_i(\boldsymbol{A}+\boldsymbol{B}) \leq \lambda_{i+j}(\boldsymbol{A}) + \lambda_{d-j}(\boldsymbol{B}), \quad j = 0, 1, \ldots, d-i \tag{3.36}$$

$$\lambda_{i-j+1}(\boldsymbol{A}) + \lambda_j(\boldsymbol{B}) \leq \lambda_i(\boldsymbol{A}+\boldsymbol{B}), \quad j = 1, \ldots, i \tag{3.37}$$

$$\tag{3.38}$$

In particular, taking $i = 1$ in the first equation and $i = d$ in the second, together with the fact $\lambda_j(\boldsymbol{B}) = -\lambda_{d+1-j}(-\boldsymbol{B})$ for $j = 1, \ldots, d$ implies that:

$$\max_{j\in[d]}\{\lambda_j(\boldsymbol{A}) - \lambda_j(\boldsymbol{B})\} \leq ||\boldsymbol{A}-\boldsymbol{B}||_{\mathtt{op}} \tag{3.39}$$

Applying this for $\boldsymbol{A} = \boldsymbol{W}$ and $\boldsymbol{B} = \boldsymbol{W}^{(i)}$ implies that the difference between their eigenvalues is of $O(1)$, which implies:

$$s_d(z) = s_{d-1}(z) + O(1/d) \tag{3.40}$$

Therefore, together we can state that:

$$\mathbb{E}[s_d(z)] = \mathbb{E}\left[\frac{1}{-z - \mathbb{E}[s_d(z)] + o(1)}\right] \overset{(a)}{=} \frac{1}{-z - \mathbb{E}[s_d(z)]} + o(1) \tag{3.41}$$

where in (a) we used that

$$\frac{1}{-z - \mathbb{E}[s_d(z)] + o(1)} - \frac{1}{-z - \mathbb{E}[s_d(z)} = \frac{o(1)}{(-z - \mathbb{E}[s_d(z)] + o(1))(-z - \mathbb{E}[s_d(z))} \tag{3.42}$$

as long as the denominator is bounded away from zero (which is the case, since it is lower bounded by $|\,\mathrm{Im}(z)| > 0$).

Equation (3.41) is enough to provide a characterisation of $\mathbb{E}[s_d(z)]$ at large d. But to pass to the limit, it remains to show that the self-consistent equation is robust to $o(1)$ changes, i.e. that the solution of eq. (3.41) converge to $s_\mu(z)$. To do so, we cam track closer the error by noting that eq. (3.41) can also be written as:

$$\mathbb{E}[s_d(z)] = \frac{1}{-z - \mathbb{E}[s_d(z)] + \epsilon_d(z)} \tag{3.43}$$

with $\epsilon_d(z) \xrightarrow{a.s.} 0$ as $d \to \infty$ with fixed $|z| = O(1)$. This has solution:

$$s_d^{\pm}(z) = \frac{-(z+\epsilon_d(z)) \pm \sqrt{(z+\epsilon_d(z))^2 - 4}}{2} \tag{3.44}$$

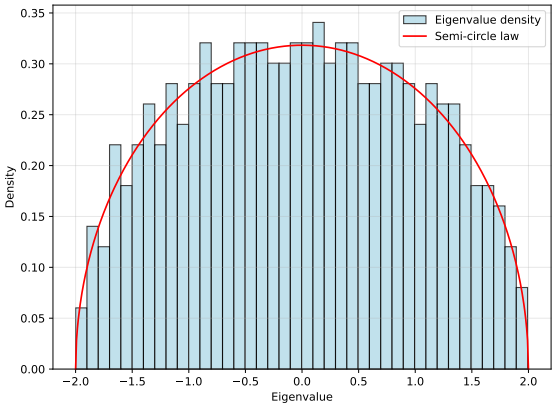

Figure 1: Histogram of eigenvalues of a GOE matrix of dimension $d = 500$ with 40 bins. The red solid curve denotes de Wigner semi-circle law $\mu(\mathrm{d}x) = \frac{\sqrt{4-x^2}}{2\pi}\mathbf{1}_{[-2,2]}\mathrm{d}x$.

To decide the correct branch, we can look at the $|z| \to \infty$ asymptotics in eq. (2.9), for which $s(z)$ needs to satisfy:

$$s_\mu(z) = -\frac{1}{z} + O(|z|^{-2}), \text{ as } |z| \to \infty \tag{3.45}$$

It is easy to check that the positive branch is the only one consistent with this. Therefore, at $d \to \infty$ we can conclude:

$$s_d^+(z) \xrightarrow{a.s.} s_\mu(z) \coloneqq \frac{-z + \sqrt{z^2-4}}{2}, \text{ as } d \to \infty \tag{3.46}$$

□

Remark 4 (Local vs. global laws)**.** The result we proved in Theorem 1 is known as a *global law*, since we assumed $|z| = O(1)$ throughout the proof. Indeed, from the inversion formula in proposition 1, controlling the rate of convergence of the Stieltjes transform in eq. (3.7) at a point $z = x + i\eta$ requires the control of $O(\eta d)$ eigenvalues around the point x. Therefore, since we assumed $\eta = O(1)$, our global law is only valid on a *macroscopic* scale $\eta d \gg 1$. A careful handling of the concentration rates can lead to finer results known as *local laws*, valid on an *mesoscopic scale* $1 \ll \eta d \ll d$ (or $\eta d = o(d)$). For Wigner matrices, such local results were first established by Erdős et al. (2009). Finally, note that we cannot hope to do better than that. The typical separation of eigenvalues of GOE matrix inside the bulk is of $O(d^{-1})$. Therefore, at the *microscopic scale* $\eta = \Theta(1/d)$ we we have to deal with statistical properties of single eigenvalues, rather than extensive aggregates. At the edge of the spectrum, eigenvalues have stronger fluctuations of $O(d^{-2/3})$, which are characterised by the celebrated Tracy-Widom law (Tracy and Widom, 2000) — see Figure 2 for an illustration.

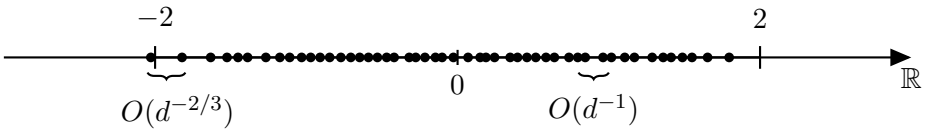

Figure 2: Scale of eigenvalue fluctuations for GOE matrices.

Remark 5 (Universality)**.** Gaussian orthogonal matrices are not the only ensemble of random matrices that have an asymptotic spectral density given by the semi-circle law. Indeed, this is true for a large ensemble of random matrices with i.i.d. entries, known as *Wigner matrices*. For instance, the ensemble of adjacency matrices of random Erdös-Rényi graphs, which can be fairly sparse, also have asymptotic semi-circle law (Erdős et al., 2012). See Tao and Vu (2011) and references therein to go deeper.

4 Sample covariance matrices and anisotropic laws

Now that we have studied in detail one of the classical random matrix theory result, we turn our attention to our original motivation, which is to understand limits of traces of the type eq. (1.3) that appear in the study of ridge regression. There are two main differences with respect to the case studied in Section 3:

- The random matrices appearing in eqs. (1.1) and (1.2) do not have independent entries, and therefore are not Wigner matrices. Indeed, they take the form of *sample covariance matrices*:

$$\hat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{d\times d} \tag{4.1}$$

  where $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ are random rectangular matrices with i.i.d. rows $\boldsymbol{x}_i \in \mathbb{R}^d$. In the context of random matrix theory, these are also known as *Wishart matrices*. Wishart matrices are widespread in statistics, where one is often interested in estimating the covariance $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top]$ from n i.i.d. samples of the covariates $\boldsymbol{x}_i$, $i \in [n]$ (note we will always assume the covariates are centred $\mathbb{E}[\boldsymbol{x}_i] = 0$).

- Differently from the discussion in Section 3 which revolved around the asymptotic limit of traces of the resolvent (a.k.a. Stieltjes transform), the expression in eq. (1.3) involve products of the resolvent with deterministic matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$, which in regression problems will either be rank-one matrices of the signal $\boldsymbol{A} = \boldsymbol{\theta}_\star \boldsymbol{\theta}_\star^\top$ or the population covariance matrix itself $\boldsymbol{B} = \boldsymbol{\Sigma}$ — see the expression of the bias eq. (1.1) for a concrete example. Asymptotic limits of these type of expressions are known as *anisotropic laws*, in contrast to theorem 1 which are isotropic.

4.1 Classical statistical regime

The large d limit of Wishart matrices notably depends on the interplay with the number of samples n. For instance, the *classical regime* that is often studied in statistics textbooks considers the limit of large sample sizes $n \to \infty$ at fixed covariate dimension $d = O(1)$, and for which the following guarantees hold:

(a) Entrywise consistency:

$$\hat{\boldsymbol{\Sigma}}_n \xrightarrow{a.s.} \boldsymbol{\Sigma}, \qquad \text{as } n \to \infty, \quad d = O(1) \tag{4.2}$$

which is a consequence of the strong law of large numbers for individual entries.

(b) Entrywise Gaussian fluctuations:

$$\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}) \xrightarrow{d} \mathcal{N}(0, V) \qquad \text{as } n \to \infty \text{ with } d = O(1) \tag{4.3}$$

where V is a covariance tensor with entries:

$$V_{jklm} = \mathbb{E}[(X_{ij}X_{ik} - \Sigma_{jk})(X_{il}X_{im} - \Sigma_{lm})] \tag{4.4}$$

Note this depends on fourth moments of the underlying distribution. This is a consequence of the Central Limit Theorem applied to $\hat{\boldsymbol{\Sigma}}_n$ when viewed as a vector in $\mathbb{R}^{d^2}$.

(c) Let $\text{spec}(\boldsymbol{\Sigma}) = \{\lambda_1, \ldots, \lambda_d\}$ and $\text{spec}(\hat{\boldsymbol{\Sigma}}_n) = \{\hat{\lambda}_1, \ldots, \hat{\lambda}_d\}$. Then, almost sure convergence entry-wise implies point-wise almost sure convergence of the individual eigenvalues:

$$\hat{\lambda}_k \xrightarrow{a.s.} \lambda_k, \qquad \text{as } n \to \infty \text{ with } d = O(1) \tag{4.5}$$

Moreover, we also have Gaussian fluctuations for the eigenvalues:

$$\sqrt{n}(\hat{\lambda}_k - \lambda_k) \xrightarrow{d} \mathcal{N}(0, \sigma_k^2) \qquad \text{as } n \to \infty \text{ with } d = O(1) \tag{4.6}$$

For a variance σ_k^2 depending on fourth moments of the underlying distribution.

Remark 6. The sample covariance matrix $\hat{\boldsymbol{\Sigma}}_n$ is the maximum likelihood estimator for $\boldsymbol{\Sigma}$. Therefore, the above results also follow directly from consistency of the MLE in the limit $n \to \infty$ at fixed $d = O(1)$.

These results suffice to characterise the traces of interest in eq. (1.3) in the classical limit. For instance, they imply that:

$$\text{Tr}\{\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1}\} \xrightarrow{a.s.} \text{Tr}\{\boldsymbol{A}(\boldsymbol{\Sigma} - z\boldsymbol{I}_d)^{-1}\}, \qquad \text{as } n \to \infty \text{ with } d = O(1) \tag{4.7}$$

4.2 High-dimensional regime

The results derived in the classical regime are weakly dependent on the distribution of the covariates. Indeed, they only require mild condition on the existence of fourth moments, which we deliberately omited. This is a common trend in classical statistics: as data is abundant $n \gg 1$, the details of the underlying distribution is not very important for consistent estimation. However, in real life limits don't exist, and "large" begs the question: *how large does n needs to be with respect to d?* Unfortunately, the above results do not allow us to answer this question as they do not track the explicit dependence on the dimension $d = O(1)$. A simple refinement of the above results can be derived under an assumption on the tails of the covariates:

Theorem 3 (Vershynin (2018), Theorem 4.7.1)**.** Let $\boldsymbol{x}_i \in \mathbb{R}^d$, $i \in [n]$ denote independent sub-Gaussian vectors with zero mean and covariance $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top] \in \mathbb{R}^{d \times d}$. Then, with probability $1 - 2e^{-t}$:

$$||\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}||_{\text{op}} \leq C \left(\sqrt{\frac{d+t}{n}} + \frac{d+t}{n} \right) ||\boldsymbol{\Sigma}||_{\text{op}} \tag{4.8}$$

Theorem 3 tell us that the sample covariance matrix is operator norm consistent as long as $d/n \to 0$. Our goal now will be to use RMT to quantify the differences between $\hat{\boldsymbol{\Sigma}}_n$ and $\boldsymbol{\Sigma}$ when both $n, d \to \infty$ with $n = \Theta(d)$, i.e. $d/n = \gamma_n \to \gamma = O(1)$.

Theorem 4 (Anisotropic law for Wishart matrices)**.** Let $\hat{\boldsymbol{\Sigma}}_n = 1/n \boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{d \times d}$ with $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{Z}$ is a sub-Gaussian matrix with zero mean and unit variance and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is a positive-definite matrix with eigenvalues $\text{spec}(\boldsymbol{\Sigma}) = \{\lambda_k : k \in [d]\} \subset \mathbb{R}_+$ and bounded operator norm $||\boldsymbol{\Sigma}||_{\text{op}} < C$. Assume that the empirical measure of eigenvalues $\hat{\mu}_n = 1/d \sum_{i \in [d]} \delta_{\lambda_i}$ converges (weakly) to a probability distribution μ on $\mathbb{R}_+$ with compact support as $d \to \infty$. Then, for any $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ with bounded operator norm:

$$\frac{1}{d} \text{Tr}\{\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1}\} \xrightarrow{a.s.} -\frac{1}{z\tilde{s}(z)} \frac{1}{d} \text{Tr} \left\{ \boldsymbol{A} \left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(z)} \boldsymbol{I}_d \right)^{-1} \right\} \text{ as } d \to \infty \text{ and } d/n \to \gamma., \tag{4.9}$$

where $\tilde{s}(z)$ is the unique solution of the following self-consistent equation:

$$\frac{1}{\tilde{s}(z)} + z = \gamma \int_0^\infty \mu(\mathrm{d}\lambda) \frac{\lambda}{1 + \lambda \tilde{s}(z)} \tag{4.10}$$

Remark 7. Before we move to the proof of this result, a few remarks are in order.

- Note that when $\gamma \to 0^+$, $\tilde{s}(z) \to -1/z$ and eq. (4.9) gives the classical result in eq. (4.7), as expected. For $\gamma > 0$, this provides a remarkable result: the high-dimensional limit of trace statistics of the empirical covariance resolvent is equal to the population, up to a renormalisation of the spectral parameter $z \mapsto 1/\tilde{s}(z)$.

- The quantity $\tilde{s}(z)$ is known as the companion Stieltjes transform, and can be shown to be given by:

$$\tilde{s}_d(z) \coloneqq \frac{1}{n} \operatorname{Tr} \left(1/n \boldsymbol{X}\boldsymbol{X}^\top - z\boldsymbol{I}_n \right)^{-1} \xrightarrow{a.s.} \tilde{s}(z) \quad \text{as} \quad n, d \to \infty \tag{4.11}$$

  i.e. it is the Stieltjes transform of the data Gram matrix $1/n \boldsymbol{X}\boldsymbol{X}^\top$. This is related to the standard Stieltjes transform for the sample covariance matrix:

$$s_d(z) \coloneqq \frac{1}{d} \operatorname{Tr} \left(1/n \boldsymbol{X}^\top\boldsymbol{X} - z\boldsymbol{I}_d \right)^{-1} \xrightarrow{a.s.} s(z) \quad \text{as} \quad n, d \to \infty \tag{4.12}$$

  by the following relation:

$$s_d(z) = \frac{1}{\gamma} \tilde{s}_d(z) + \frac{1-\gamma}{\gamma z} \tag{4.13}$$

  which is a simple consequence of the fact that the matrices $\boldsymbol{X}\boldsymbol{X}^\top$ and $\boldsymbol{X}^\top\boldsymbol{X}$ have the same spectrum up to the zero eigenvalues.

- The anisotropic law in eq. (4.9) is sometimes referred in the literature as a *deterministic equivalent*. More precise, we say that a random matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ has deterministic equivalent $\bar{\boldsymbol{M}} \in \mathbb{R}$, denoted $\boldsymbol{M} \leftrightarrow \bar{\boldsymbol{M}}$, if for any deterministic matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ with bounded operator norm:

$$\frac{1}{d} \operatorname{Tr} \boldsymbol{A}\boldsymbol{M} \to \frac{1}{d} \operatorname{Tr} \boldsymbol{A}\bar{\boldsymbol{M}}, \qquad \text{as } d \to \infty \tag{4.14}$$

  with the convergence being sometimes almost surely or in probability. In other words, measuring any scalar or "trace-like" statistics of the random matrix is the same as doing the same measurement on the deterministic matrix. See (Couillet and Liao, 2022) for a reference that employs this notion.

Sketch of the proof. We now sketch the main ideas in the proof of theorem 4 using a similar leave-one-out argument as in the the Wigner case, without aiming at the same level of rigour. As in the Wigner case, the proof can be separated in two steps: (a) the almost sure convergence of the (companion) Stieltjes transform; (b) finding an exact asymptotic characterisation of the trace of interest:

$$\frac{1}{d} \operatorname{Tr} \boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1} \overset{(b)}{=} \operatorname{Tr} \boldsymbol{A}(-z\tilde{s}_d(z)\boldsymbol{\Sigma} - z\boldsymbol{I}_d)^{-1} + o(1) \overset{(a)}{=} \operatorname{Tr} \boldsymbol{A}(-z\tilde{s}(z)\boldsymbol{\Sigma} - z\boldsymbol{I}_d)^{-1} + o(1) \tag{4.15}$$

One could proceed at directly proving the approximations above. However, this would assume that we know the asymptotic limit of the trace in advance. Instead, here we focus on how to derive (b) from scratch, using a leave-one-out argument over the independent rows of the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_n$. For that, define:

$$\hat{\boldsymbol{\Sigma}}_n^{(i)} = \frac{1}{n} \sum_{j \neq i} \boldsymbol{x}_j \boldsymbol{x}_j^\top \tag{4.16}$$

such that $\hat{\boldsymbol{\Sigma}}_n = \hat{\boldsymbol{\Sigma}}_n^{(i)} + 1/n \boldsymbol{x}_i \boldsymbol{x}_i^\top$. Similarly, define the leave-one-out resolvent matrix:

$$\boldsymbol{R}^{(i)}(z) \coloneqq \boldsymbol{R}_{\hat{\boldsymbol{\Sigma}}_n^{(i)}}(z) = \left(\hat{\boldsymbol{\Sigma}}_n^{(i)} - z\boldsymbol{I}_d \right)^{-1}. \tag{4.17}$$

This can be related to the sample covariance resolvent via the Sherman-Morrison lemma:

Lemma 5 (Sherman-Morrison lemma)**.** Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ denote an invertible matrix and $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{d\times d}$ two vectors such that $\boldsymbol{v}^\top \boldsymbol{A}^{-1}\boldsymbol{u} \neq -1$. Then:

$$(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v})^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top \boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^\top \boldsymbol{A}^{-1}\boldsymbol{u}} \tag{4.18}$$

Applying it to the sample covariance resolvent:

$$\begin{aligned}\boldsymbol{R}(z) \coloneqq (\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1} &= \left(\frac{1}{n}\sum_{j=1}^n \boldsymbol{x}_j\boldsymbol{x}_j^\top - z\boldsymbol{I}_d\right)^{-1} = \left(\frac{1}{n}\sum_{j\neq i}^n \boldsymbol{x}_j\boldsymbol{x}_j^\top + \frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{n} - z\boldsymbol{I}_d\right)^{-1} \\ &= \left(\hat{\boldsymbol{\Sigma}}_n^{(i)} - z\boldsymbol{I}_d\right)^{-1}\left[\boldsymbol{I}_d - \frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top\left(\hat{\boldsymbol{\Sigma}}_n^{(i)} - z\boldsymbol{I}_d\right)^{-1}}{n + \boldsymbol{x}_i^\top\left(\hat{\boldsymbol{\Sigma}}_n^{(i)} - z\boldsymbol{I}_d\right)^{-1}\boldsymbol{x}_i}\right]\end{aligned} \tag{4.19}$$

Hence:

$$\boldsymbol{R}(z) - \boldsymbol{R}^{(i)}(z) = -\frac{\boldsymbol{R}^{(i)}(z)\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{R}^{(i)}(z)}{n + \boldsymbol{x}_i^\top \boldsymbol{R}^{(i)}(z)\boldsymbol{x}_i}, \tag{4.20}$$

and we expect that for any deterministic matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ with bounded spectral norm and $z \in \mathbb{C}_+$:

$$\operatorname{Tr}\boldsymbol{A}\left(\boldsymbol{R}(z) - \boldsymbol{R}^{(i)}(z)\right) = o(1) \tag{4.21}$$

Note in particular that:

$$\begin{aligned}\operatorname{Tr}\boldsymbol{A} &= \operatorname{Tr}\boldsymbol{R}(z)^{-1}\boldsymbol{R}(z)\boldsymbol{A} \\ &= \operatorname{Tr}\boldsymbol{R}(z)\hat{\boldsymbol{\Sigma}}_n\boldsymbol{A} - z\operatorname{Tr}\boldsymbol{R}(z)\boldsymbol{A}\end{aligned} \tag{4.22}$$

The first term can be written as:

$$\begin{aligned}\operatorname{Tr}\{\boldsymbol{R}(z)\hat{\boldsymbol{\Sigma}}_n\boldsymbol{A}\} &= \frac{1}{n}\sum_{i=1}^n \operatorname{Tr}\left\{\boldsymbol{R}(z)\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{A}\right\} \\ &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^\top\boldsymbol{A}\boldsymbol{R}(z)\boldsymbol{x}_i\end{aligned} \tag{4.23}$$

we would like to apply the Hanson-Wright inequality theorem 2 to this term. However, $\boldsymbol{R}$ is correlated with $\boldsymbol{x}_i$, and as in the Wigner case we need to first remove this correlation. From eq. (4.19) it follows that:

$$\boldsymbol{R}(z)\boldsymbol{x}_i = \frac{\boldsymbol{R}^{(i)}(z)\boldsymbol{x}_i}{1 + {}^1\!/_n\boldsymbol{x}_i^\top\boldsymbol{R}^{(i)}(z)\boldsymbol{x}_i} \tag{4.24}$$

and hence:

$$\operatorname{Tr}\{\boldsymbol{R}(z)\hat{\boldsymbol{\Sigma}}_n\boldsymbol{A}\} = \sum_{i=1}^n \frac{\boldsymbol{x}_i^\top\boldsymbol{A}\boldsymbol{R}^{(i)}(z)\boldsymbol{x}_i}{n + \boldsymbol{x}_i^\top\boldsymbol{R}^{(i)}(z)\boldsymbol{x}_i} \tag{4.25}$$

By theorem 2, we have almost surely that for $|z| = O(1)$:

$$\boldsymbol{x}_i^\top\boldsymbol{A}\boldsymbol{R}^{(i)}(z)\boldsymbol{x}_i = \operatorname{Tr}\{\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{R}^{(i)}(z)\} + o(d), \qquad \boldsymbol{x}_i^\top\boldsymbol{R}^{(i)}(z)\boldsymbol{x}_i = \operatorname{Tr}\{\boldsymbol{\Sigma}\boldsymbol{R}^{(i)}(z)\} + o(d) \tag{4.26}$$

Hence:

$$\operatorname{Tr}\{\boldsymbol{R}(z)\hat{\boldsymbol{\Sigma}}_n\boldsymbol{A}\} = \sum_{i=1}^{n} \frac{\operatorname{Tr}\{\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{R}^{(i)}(z)\}}{n + \operatorname{Tr}\{\boldsymbol{\Sigma}\boldsymbol{R}^{(i)}(z)\}} + o(d) = n\frac{\operatorname{Tr}\{\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{R}(z)\}}{n + \operatorname{Tr}\{\boldsymbol{\Sigma}\boldsymbol{R}(z)\}} + o(d) \tag{4.27}$$

where in the last equality we used that $\boldsymbol{R}^{(i)} = \boldsymbol{R} + o(1)$. Now inserting this back into eq. (4.22), we have that:

$$\begin{aligned}\operatorname{Tr}\boldsymbol{A} &= n\frac{\operatorname{Tr}\{\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{R}(z)\}}{n + \operatorname{Tr}\{\boldsymbol{\Sigma}\boldsymbol{R}(z)\}} - z\operatorname{Tr}\boldsymbol{R}(z)\boldsymbol{A} + o(d) \\ &= \operatorname{Tr}\left\{\boldsymbol{R}(z)\left(\frac{\boldsymbol{\Sigma}}{1 + \frac{1}{n}\operatorname{Tr}\boldsymbol{\Sigma}\boldsymbol{R}(z)} - z\boldsymbol{I}_d\right)\boldsymbol{A}\right\} + o(d)\end{aligned} \tag{4.28}$$

which is valid for any deterministic $\boldsymbol{A} \in \mathbb{R}^{d\times d}$. Rearranging this equation:

$$\operatorname{Tr}\left\{\left[\frac{1}{z}\left(\frac{\boldsymbol{R}(z)\boldsymbol{\Sigma}}{1 + \frac{1}{n}\operatorname{Tr}\boldsymbol{R}(z)\boldsymbol{\Sigma}} - \boldsymbol{I}_d\right) - \boldsymbol{R}(z)\right]\boldsymbol{A}\right\} = o(d). \tag{4.29}$$

and comparing with the definition of the deterministic equivalent $\bar{R}(z)$:

$$\operatorname{Tr}\boldsymbol{A}(\boldsymbol{R}(z) - \bar{\boldsymbol{R}}(z)) = o(d) \tag{4.30}$$

one would like to identify:

$$\bar{\boldsymbol{R}}(z) = \frac{1}{z}\left(\frac{\boldsymbol{R}(z)\boldsymbol{\Sigma}}{1 + \frac{1}{n}\operatorname{Tr}\boldsymbol{R}(z)\boldsymbol{\Sigma}} - \boldsymbol{I}_d\right) + o(d) \tag{4.31}$$

However, this still depends on the resolvent $\boldsymbol{R}(z)$, a random quantity. To get rid of this dependence, we evaluate eq. (4.28) at $\boldsymbol{A} = \boldsymbol{I}_d$, which after recognising the Stieltjes transform $s_d(z) = {}^1\!/_d \operatorname{Tr}\boldsymbol{R}(z)$ give us:

$$d = \frac{\operatorname{Tr}\{\boldsymbol{R}(z)\boldsymbol{\Sigma}\}}{1 + \frac{1}{n}\operatorname{Tr}\{\boldsymbol{\Sigma}\boldsymbol{R}(z)\}} - zds_d(z). \tag{4.32}$$

This can be rearranged to give:

$$1 + zs_d(z) = \frac{\frac{1}{d}\operatorname{Tr}\{\boldsymbol{R}(z)\boldsymbol{\Sigma}\}}{1 + \frac{\gamma}{d}\operatorname{Tr}\{\boldsymbol{R}(z)\boldsymbol{\Sigma}\}} = \frac{1}{\gamma}\left(1 - \frac{1}{1 + \gamma\operatorname{Tr}\{\boldsymbol{R}(z)\boldsymbol{\Sigma}\}}\right) \tag{4.33}$$

Now solving for $\operatorname{Tr}\{\boldsymbol{R}(z)\boldsymbol{\Sigma}\}$ give us:

$$\frac{1}{1 + \gamma\operatorname{Tr}\{\boldsymbol{R}(z)\boldsymbol{\Sigma}\}} = 1 - \gamma(1 + zs_d(z)) = -z\tilde{s}_d(z) \tag{4.34}$$

where we recognised the companion Stieltjes transform (4.11) from the relation in eq. (4.13). This shows that actually the terms depending on $\boldsymbol{R}$ are actually only a function of $\tilde{s}_d(z)$, a quantity which concentrate. Inserting this back in eq. (4.28)

$$\operatorname{Tr}\boldsymbol{A} = \operatorname{Tr}\left\{\boldsymbol{R}(z)\left(-z\tilde{s}_d(z)\boldsymbol{\Sigma} - z\boldsymbol{I}_d\right)\boldsymbol{A}\right\} + o(d) \tag{4.35}$$

$$= \operatorname{Tr}\left\{\boldsymbol{R}(z)\left(-z\tilde{s}(z)\boldsymbol{\Sigma} - z\boldsymbol{I}_d\right)\boldsymbol{A}\right\} + o(d) \tag{4.36}$$

where in the second equality we used that $\tilde{s}_d(z) = \tilde{s}(z) + o(1)$. Finally, we can evaluate this at the (deterministic) value $\boldsymbol{A} = (-z\tilde{s}(z)\boldsymbol{\Sigma} - z\boldsymbol{I}_d)^{-1}$ to get:

$$\frac{1}{d}\operatorname{Tr}\{(-z\tilde{s}(z)\boldsymbol{\Sigma} - z\boldsymbol{I}_d)^{-1}\} = \frac{1}{d}\operatorname{Tr}\boldsymbol{R}(z) + o(1) = s(z) + o(1) \tag{4.37}$$

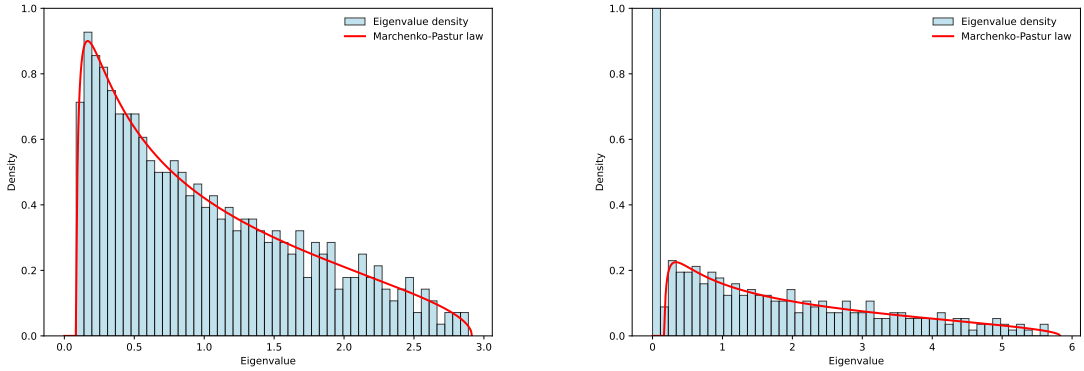

Figure 3: Histogram of eigenvalues of a Wishart matrix of dimension $d = 500$ with 40 bins for $\gamma = 0.5$ (**left**) and $\gamma = 2$ (**right**). The red solid curve denotes de Marchenko-Pastur law given by eq. (4.43)

This is exactly the result stated in eq. (4.9). Using again eq. (4.13) allow us to get the self-consistent equation on $\tilde{s}(z)$ stated in eq. (4.10):

$$\frac{1}{\tilde{s}(z)} + z = \lim_{d\to\infty} \frac{\gamma}{d} \operatorname{Tr}\big(\boldsymbol{\Sigma}(1 + \tilde{s}(z)\boldsymbol{\Sigma})^{-1}\big) = \int_0^\infty \mu(\mathrm{d}\lambda) \frac{\lambda}{1 + \lambda\tilde{s}(z)} \tag{4.38}$$

The argument above can be made rigorous by justifying the different approximations we have taken. We now briefly sketch how to proceed.

Using the resolvent identity in eq. (3.14) we can write the difference between the resolvent and its (finite d) equivalent:

$$(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1} = (-z\tilde{s}_d(z)\boldsymbol{\Sigma} - z\boldsymbol{I}_d)^{-1}(\boldsymbol{I}_d - \boldsymbol{\Delta}) \tag{4.39}$$

where:

$$\boldsymbol{\Delta} := \hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1} - z\tilde{s}_d(z)\boldsymbol{\Sigma}(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1} \tag{4.40}$$

Therefore, proving eq. (4.15) amounts to showing that $\operatorname{Tr}(-z\tilde{s}_d(z)\boldsymbol{\Sigma} - z\boldsymbol{I}_d)^{-1}\boldsymbol{\Delta} = o(d)$. Using the leaving-one-out argument as in eq. (4.19), we can show that $\boldsymbol{\Delta}$ can be re-written as:

$$\boldsymbol{\Delta} = \sum_{i=1}^{n} \frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top(\hat{\boldsymbol{\Sigma}}_n^{(i)} - z\boldsymbol{I}_d)^{-1} - \boldsymbol{\Sigma}(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1}}{n + \boldsymbol{x}_i^\top(\hat{\boldsymbol{\Sigma}}_n^{(i)} - z\boldsymbol{I}_d)^{-1}\boldsymbol{x}_i} \tag{4.41}$$

writing $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_i$ for independent sub-Gaussian vectors $\boldsymbol{u}_i \in \mathbb{R}^d$ with zero mean and unit variance, controlling $\operatorname{Tr}\boldsymbol{\Delta}$ involves controlling traces of the type $\operatorname{Tr}\boldsymbol{A}(\boldsymbol{u}_i\boldsymbol{u}_i - \boldsymbol{I}_d)$ for deterministic $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ with bounded operator norm. See (Silverstein, 1995) for a detailed proof. □

4.3 The Marchenko-Pastur law

In the isotropic case $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, the solution of the self-consistent equation eq. (4.10) together with eq. (4.13) give us the celebrated result by Marchenko and Pastur for the asymptotic Stieltjes transform of Wishart matrices (Pastur and Martchenko, 1967):

$$s(z) = \frac{1 - \gamma - z + \sqrt{(1 - \gamma + z)^2 - 4z}}{2\gamma z} \tag{4.42}$$

This is the Stieltjes transform of the celebrated *Marchenko-Pastur law*:

$$\mu_{\mathrm{mp}}(\mathrm{d}x) = \left(1 - \frac{1}{\gamma}\right)_+ \delta_0 + \frac{\sqrt{(\gamma_+ - x)(x - \gamma_-)}}{2\pi\gamma x} \mathbf{1}_{[\gamma_-, \gamma_+]}(x)\mathrm{d}x \tag{4.43}$$

where the edges of the distribution are $\gamma_\pm = (1 \pm \sqrt{\gamma})^2$. Figure 3 illustrates this result. Recall that $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, and therefore the spectral measure of the population covariance is simply a point mass:

$$\mu_{\boldsymbol{\Sigma}} = \delta_1 \tag{4.44}$$

This is strikingly different from μ_{mp}, where the eigenvalues are spread over a full interval $[\gamma_-, \gamma_+] \subset \mathbb{R}_+$. In particular, when $\gamma > 1$ ($d > n$), there are $d - n$ zero eigenvalues (since $\mathrm{rank}(\hat{\boldsymbol{\Sigma}}_n) = n$ almost surely). Therefore, in this high-dimensional regime a naive statistician might be mistakenly led to believe that the true data has a low-dimensional structure where in fact it is completely isotropic.

4.4 Other useful results

To conclude of exposition of random matrix theory, we present an additional result concerning bi-product of traces that are relevant to our discussion of ridge regression. These can be derived and proven following exactly the same argument as in Theorem 4. However, their derivation is more cumbersome, and we refer the interested reader to (Bach, 2024) for a detailed discussion.

Theorem 5 (Proposition 1 in (Bach, 2024))**.** Under the same assumptions as in Theorem 5, for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$ with bounded operator norm we have:

$$\begin{aligned}\mathrm{Tr}\{\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1}\boldsymbol{B}(\hat{\boldsymbol{\Sigma}}_n - z\boldsymbol{I}_d)^{-1}\} \xrightarrow{a.s.} & \frac{1}{z^2\tilde{s}(z)^2}\mathrm{Tr}\left\{\boldsymbol{A}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(z)}\boldsymbol{I}_d\right)^{-1}\boldsymbol{B}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(z)}\boldsymbol{I}_d\right)^{-1}\right\} \\ & + \frac{1}{z^2\tilde{s}(z)^2}\frac{\mathrm{Tr}\left\{\boldsymbol{A}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(z)}\boldsymbol{I}_d\right)^{-2}\boldsymbol{\Sigma}\right\}\cdot\mathrm{Tr}\left\{\boldsymbol{B}\left(\boldsymbol{\Sigma} + \frac{1}{\tilde{s}(z)}\boldsymbol{I}_d\right)^{-2}\boldsymbol{\Sigma}\right\}}{n - \mathrm{Tr}\left\{\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + 1/\tilde{s}(z)\boldsymbol{I}_d)^{-2}\right\}}\end{aligned} \tag{4.45}$$

where $\tilde{s}(z)$ is the unique solution of the following self-consistent equation eq. (4.10).

5 To go further

A nice list of useful lecture notes on random matrix theory.

- Terence Tao's lecture notes on random matrix theory taught at UCLA.
- Lecture 17 of Song Mei's STAT260 course taught at Berkeley.
- Florent Benaych-Georges and Antti Knowles lecture notes on local laws for Wigner matrices.
- Chapter 6 of Djalil Chafai lecture notes for the course "*Phénomènes de grande dimension*" taught at ENS (in French).
- Charles Bordenave lecture notes on random matrix theory for a course taught at IMPA.

Books:

- Zhidong Bai and Jack W. Silverstein "*Spectral Analysis of Large Dimensional Random Matrices* ", available online at the editor's webpage. Classical reference by the people behind many of the results in RMT.

- Romain Couillet and Zhenyu Liao book "*Random Matrix Methods for Machine Learning*", available online at the author's webpage. A good recent reference with detailed proofs and a discussion on the applications to machine learning.
- Marc Potters and Jean-Philippe Bouchaud book "*A First Course in Random Matrix Theory*". Very good reference written by physicists, with plenty of intuition.

References

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024.

S Boucheron, G Lugosi, and P Massart. Concentration inequalities: A nonasymptotic theory of independence. univ. press, 2013.

Romain Couillet and Zhenyu Liao. *Random matrix methods for machine learning*. Cambridge University Press, 2022.

László Erdős, Benjamin Schlein, and Horng-Tzer Yau. Local semicircle law and complete delocalization for wigner random matrices. *Communications in Mathematical Physics*, 287(2):641–655, 2009.

László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin. Spectral statistics of erdős-rényi graphs ii: Eigenvalue spacing and the extreme eigenvalues. *Communications in Mathematical Physics*, 314 (3):587–640, 2012.

LA Pastur and VA Martchenko. The distribution of eigenvalues in certain sets of random matrices. *Math. USSR-Sbornik*, 1(4):457–483, 1967.

Jack W Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995.

Terence Tao and Van Vu. Random matrices: Universality of local eigenvalue statistics. *Acta Mathematica*, 206(1):127–204, 2011. doi: 10.1007/s11511-011-0061-3. URL https://doi.org/10.1007/s11511-011-0061-3.

Craig A Tracy and Harold Widom. *The distribution of the largest eigenvalue in the Gaussian ensembles: $\beta = 1, 2, 4$*. Springer, 2000.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Eugene P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564, 1955. ISSN 0003486X, 19398980. URL http://www.jstor.org/stable/1970079.