LARGE DEVIATIONS FOR THE LARGEST EIGENVALUE OF RADEMACHER MATRICES

Alice Guionnet, Jonathan Husson

LARGE DEVIATIONS FOR THE LARGEST EIGENVALUE OF RADEMACHER MATRICES

ALICE GUIONNET AND JONATHAN HUSSON

Abstract. In this article, we consider random Wigner matrices, that is symmetric matrices such that the subdiagonal entries of X_n are independent, centered, and with variance one except on the diagonal where the entries have variance two. We prove that, under some suitable hypotheses on the laws of the entries, the law of the largest eigenvalue satisfies a large deviation principle with the same rate function as in the Gaussian case. The crucial assumption is that the Laplace transform of the entries must be bounded above by the Laplace transform of a Gaussian variable with same variance. This is satisfied by the Rademacher law and the uniform law on $[-\sqrt{3}, \sqrt{3}]$. We extend our result to complex entries Wigner matrices and Wishart matrices.

1. Introduction

Very few large deviation principles could be proved so far in random matrix theory. Indeed, the natural quantities of interest such as the spectrum and the eigenvectors are complicated functions of the entries. Hence, even if one considers the simplest model of Wigner matrices which are self-adjoint with independent identically distributed entries above the diagonal, the probability that the empirical measure of the eigenvalues or the largest eigenvalue deviates towards an unlikely value is very difficult to estimate. A well known case where probabilities of large deviations can be estimated is the case where the entries are Gaussian, centered and well chosen covariances, the so-called Gaussian ensembles. In this case, the joint law of the eigenvalues has an explicit form, independent of the eigenvectors, displaying a strong Coulomb gas interaction. This formula could be used to prove a large deviations principle for the empirical measure in [8] and for the largest eigenvalue [7] (see also [19] for further discussions of the Wishart case, and [12]). More recently, in a breakthrough paper, Bordenave and Caputo [11] could tackle the case of matrices with heavy tails, that is Wigner matrices with entries with stretched exponential tails, going to zero at infinity more slowly than a Gaussian tail. The driving idea to approach this question was to show that large deviations are in this case created by a few large entries, so that the empirical measure deviates towards the free convolution of the semi-circle law and the limiting spectral measure of the matrix created by these few large entries. This idea could be also used to grasp the large deviations of the largest eigenvalue [2]. In the Wishart case, [13] considered the large deviations for the largest eigenvalue of Wishart matrices $W = XX^*$, but in the regime where the matrix X is $L \times M$ with L much smaller than M. Hence large deviations for bounded entries, or simply entries with sub-Gaussian tails, remained mysterious in the case of Wigner matrices or Wishart matrices with L of order M. In this article we analyze the large deviations of the

This work was supported in part by Labex MILYON.

largest eigenvalue of Wigner matrices with Rademacher or uniformly distributed random variables. More precisely our result holds for any independent identically distributed entries with distribution with Laplace transform bounded above by the Laplace transform of the Gaussian law with the same variance. We then prove a large deviation principle with the same rate function than in the Gaussian case: large deviations are universal in this class of measures. We show that this result generalizes to complex entries Wigner matrices as well as to Wishart matrices. We are considering the case of general sub-Gaussian entries in a companion paper with F. Augeri. We show in particular that the rate function may be different from the rate function of the Gaussian case, at least for deviations towards very large values.

1.1. **Statement of the results.** We consider a family of independent random variables $(a_{i,j}^{(1)})_{0\le i\le j\le N}$, such that the variables $a_{i,j}^{(1)}$ are distributed according to the laws $\mu_{i,j}^N$. We moreover assume that the $\mu_{i,j}^N$ are centered :

$$\mu_{i,j}^N(x) = \int x d\mu_{i,j}^N(x) = 0$$

and with covariance:

$$\mu_{i,j}^N(x^2) = \int x^2 d\mu_{i,j}^N(x) = 1, \forall 1 \le i < j \le N, \qquad \mu_{i,i}^N(x^2) = 2, \quad \forall 1 \le i \le N\,.$$

We say that a probability measure μ has a sharp sub-Gaussian Laplace transform iff

$$\forall t \in \mathbb{R}, T_\mu(t) = \int \exp\{tx\} d\mu(x) \le \exp\Big\{\frac{t^2\mu(x^2)}{2}\Big\}\,. \tag{1}$$

The terminology "sharp" comes from the fact that for t small, we must have

$$T_\mu(t) \ge \exp\{\frac{t^2\mu(x^2)}{2}(1+o(t))\}\,.$$

Then we assume that

Assumption 1.1 (A0)**.** *We assume that the $\mu_{i,j}^N$ satisfy a sharp Gaussian Laplace transform in the sense that*

- *$(\mu_{i,j}^N)_{i\le j}$ have a sharp sub-Gaussian Laplace transform,*
- *The $\mu_{i,j}^N$ have a uniform lower bounded Laplace transform: For any $\delta > 0$ there exists $\varepsilon(\delta) > 0$ such that for any $|t| \le \varepsilon(\delta)$, any $1 \le i \le j \le N$, any $N \in \mathbb{N}$,*

$$T_{\mu_{i,j}^N}(t) \ge \exp\{\frac{(1-\delta)t^2\mu_{i,j}^N(x^2)}{2}\}\,.$$

Moreover, we assume that the $T_{\mu_{i,j}^N}$ are uniformly C^3 in a neighborhood of the origin: for $\epsilon > 0$ small enough $\sup_{|t|\le\epsilon}\sup_{i,j,N}|\partial_t^3 \ln T_{\mu_{i,j}^N}(t)|$ is finite.

Observe that the $\mu_{i,j}^N$ have a uniform lower bounded Laplace transform as soon as they do not depend on N and there are finitely many different of them.

Remark 1.1. *We could assume a weaker upper bound on the Laplace transform for the diagonal entries such as the existence of A finite such that*

$$\int e^{tx} d\mu_{i,i}^N(x) \leq \exp\{t^2 + A|t|\}, \quad \forall 1 \leq i \leq N,$$

see the proof of Theorem 1.17.

Example 1.2. (1) *Clearly a centered Gaussian variable has a sharp sub-Gaussian Laplace transform.*

(2) *The Rademacher law $B = \frac{1}{2}(\delta_{-1} + \delta_1)$ satisfies a sharp sub-Gaussian Laplace transform since for all t*

$$T_B(t) = \cosh(t) \leq e^{t^2/2}\,.$$

(3) *U, the uniform law on the interval $[-\sqrt{3}, \sqrt{3}]$, satisfies a sharp sub-Gaussian Laplace transform since we have*

$$\int x^2 dU(x) = 1\,,$$

and

$$T_U(t) = \frac{1}{t\sqrt{3}} \sinh(t\sqrt{3}) = \sum_{n\geq 0} \frac{t^{2n} 3^n}{(2n+1)!}\,.$$

Since for all $n \geq 0$, $\frac{3^n}{(2n+1)!} \leq \frac{1}{2^n n!}$, it follows that $T_U(t) \leq e^{\frac{t^2}{2}}$.

(4) *More generally if μ is a symmetric measure on $\mathbb{R}$ (i.e. such as $\mu(-A) = \mu(A)$ for any Borel subset A of $\mathbb{R}$) such that*

$$\int x^2 d\mu(x) = 1, \quad \int x^{2n} d\mu(x) \leq \frac{(2n)(2n-1)\cdots(n+1)}{2^n} \quad \forall n \geq 2$$

then μ satisfies a sharp sub-Gaussian Laplace transform.

(5) *If X, Y are two independent variables with distribution μ and μ', two probability measures which have a sharp sub-Gaussian Laplace transform, for any $a \in [0,1]$, the distribution of $\sqrt{a}X + \sqrt{1-a}Y$ has a sharp sub-Gaussian Laplace transform.*

(6) *If $\mu_{i,j}^N = \mu$ for all i, j, then they satisfy a uniform lower bound on the Laplace transform.*

Note that many measures do not have a sharp sub-Gaussian Laplace transform, e.g. the sparse Gaussian law obtained by multiplying a Gaussian variable by a Bernoulli variable, or the sum of most Rademacher laws. We will also make classical assumptions to use standard concentration of measure tools:

Assumption 1.2. *There exists a compact set K such that the support of all $\mu_{i,j}^N$ is included in K for all $i, j \in \{1, \ldots, N\}$and all integer number N, or all $\mu_{i,j}^N$ satisfy a log-Sobolev inequality with the same constant c independent of N.*

Remark 1.3. *All the examples of Example 1.2 satisfy Assumption 1.2, except possibly for sums of Gaussian variables and bounded entries.*

We then construct for all $N \in \mathbb{N}$, a real Wigner matrix $N \times N$ $X_N^{(1)}$ by setting :

$$X_N^{(1)}(i,j) = \begin{cases} \frac{a_{i,j}^{(1)}}{\sqrt{N}} \text{ when } i \leq j, \\ \frac{a_{j,i}^{(1)}}{\sqrt{N}} \text{ when } i > j\,. \end{cases}$$

We denote $\lambda_{min}(X_N^{(1)}) = \lambda_1 \leq \lambda_2 \cdots \leq \lambda_N = \lambda_{\max}(X_N^{(1)})$ the eigenvalues of $X_N^{(1)}$. It is well known [21] that under our hypotheses the empirical distribution of the eigenvalues $\hat{\mu}^N_{X_N^{(1)}} = \frac{1}{N}\sum_{i=1}^N \delta_{\lambda_i}$ converges weakly towards the semi-circle distribution σ: for all bounded continuous function f

$$\lim_{N\to\infty} \int f(x) d\hat{\mu}^N_{X_N^{(1)}}(x) = \int f(x) d\sigma(x) = \frac{1}{2\pi}\int_{-2}^2 f(x)\sqrt{4-x^2}dx \qquad a.s.$$

It is also well known that the eigenvalues stick to the bulk since we assumed the entries have sub-Gaussian moments [14, 1] :

$$\lim_{N\to\infty} \lambda_{min}(X_N^{(1)}) = -2 \qquad \lim_{N\to\infty} \lambda_{\max}(X_N^{(1)}) = 2, \qquad a.s$$

Our main result is a large deviation principle from this convergence.

Theorem 1.4. *Suppose Assumptions 1.1 and 1.2 hold. Then, the law of the largest eigenvalue $\lambda_{\max}(X_N^{(1)})$ of $X_N^{(1)}$ satisfies a large deviation principle with speed N and good rate function $I^{(1)}$ which is infinite on $(-\infty, 2)$ and otherwise given by*

$$I^{(1)}(\rho) = \int_2^\rho \sqrt{x^2-4}dx\,.$$

In other words, for any closed subset F of $\mathbb{R}$,

$$\limsup_{N\to\infty} \frac{1}{N} \ln P\left(\lambda_{\max}(X_N^{(1)}) \in F\right) \leq -\inf_F I^{(1)}\,,$$

whereas for any open subset O of $\mathbb{R}$

$$\liminf_{N\to\infty} \frac{1}{N} \ln P\left(\lambda_{\max}(X_N^{(1)}) \in O\right) \geq -\inf_O I^{(1)}\,.$$

The same result holds for the opposite of the smallest eigenvalue $-\lambda_{min}(X_N^{(1)})$.

Therefore, the large deviations principles are the same as in the case of Gaussian entries as soon as the entries have a sharp sub-Gaussian Laplace transforms and are bounded, for instance for Rademacher variables or uniformly distributed variables. Hereafter we show how this result generalizes to other settings. First, this result generalizes to the case of Wigner matrices with complex entries as follows. We now consider a family of independent random variables $(a_{i,j}^{(2)})_{1\leq i\leq j\leq N}$, such that the variables $a_{i,j}^{(2)}$ are distributed according to a law $\mu_{i,j}^N$ when $i \leq j$, which are centered probability measures on $\mathbb{C}$ (and on $\mathbb{R}$ if $i = j$). We write $a_{i,j}^{(2)} = x_{i,j} + iy_{i,j}$ where $x_{i,j} = \Re(a_{i,j}^{(2)})$ and $y_{i,j} = \Im(a_{i,j}^{(2)})$. We suppose that for all $i \in [1, N]$, $y_{i,i} = 0$. In this context, for a probability measure on $\mathbb{C}$, we will consider its Laplace transform to be the function

$$T_\mu(z) := \int \exp\{\Re(a\bar{z})\} d\mu(a)\,.$$

We assume that

Assumption 1.3 (A0c). *For all $i < j$*

$$\forall t \in \mathbb{C}, T_{\mu_{i,j}^N}(t) \leq \exp(|t|^2/4)$$

and for all i

$$\forall t \in \mathbb{R}, T_{\mu_{i,i}^N}(t) \leq \exp(t^2/2)\,.$$

We assume that for all $\delta > 0$ there exists $\varepsilon(\delta) > 0$ so that for all complex number t with modulus bounded by $\varepsilon(\delta)$

$$T_{\mu_{i,j}^N}(t) \geq \exp\{\frac{|t|_2^2}{4}(1-\delta)\}, i < j, \qquad T_{\mu_{i,i}^N}(t) \geq \exp\{\frac{(1-\delta)t^2}{2}\}\,.$$

Moreover, for $\epsilon > 0$ small enough $\sup_{|t|\leq\epsilon} \sup_{i,j,N} |\partial_t^3 \ln T_{\mu_{i,j}^N}(t)|$ is finite.

Observe that the above hypothesis implies that for all $i < j$, $2\mathbb{E}[x_{i,j}^2] = 2\mathbb{E}[y_{i,j}^2] = \mathbb{E}[x_{i,i}^2] = 1$ and $\mathbb{E}[x_{i,j}y_{i,j}] = 0$. Examples of distributions satisfying Assumption 1.3 are given by taking $(x_{i,j}, y_{i,j})$ centered independent variables with law satisfying a sharp sub-Gaussian Laplace transform. Hereafter, we extend naturally Assumption 1.2 by assuming that the compact K is a compact subset of $\mathbb{C}$ or log-Sobolev inequality holds in the complex setting.

We then construct for all $N \in \mathbb{N}$, $X_N^{(2)}$ a complex Wigner matrix $N \times N$ by letting :

$$X_N^{(2)}(i,j) = \begin{cases} \frac{a_{i,j}^{(2)}}{\sqrt{N}} \text{ when } i \leq j \\ \\ \frac{\overline{a_{j,i}^{(2)}}}{\sqrt{N}} \text{ when } i > j \end{cases}$$

Again, it is well known that the spectral measure of $X_N^{(2)}$ converges towards the semi-circle distribution σ and that the eigenvalues stick to the bulk [1].

Theorem 1.5. *Assume that Assumptions 1.3 and 1.2 hold. Then, the law of the largest eigenvalue $\lambda_{\max}(X_N^{(2)})$ of $X_N^{(2)}$ satisfies a large deviation principle with speed N and good rate function $I^{(2)}$ which is infinite on $(-\infty, 2)$ and otherwise given by*

$$I^{(2)}(\rho) = 2I^{(1)}(\rho) = 2\int_2^\rho \sqrt{x^2-4}dx\,.$$

We finally generalize our result to the case of Wishart matrices. We let L, M be two integers with $N = L + M$. Let $G_{L,M}^{(\beta)}$ be an $L \times M$ matrix with independent entries $(a_{i,j}^{(\beta)})_{\substack{1\leq i\leq L \\ 1\leq j\leq M}}$ with laws $\mu_{i,j}^{L,M}$ on the real line if $\beta = 1$ and on the complex plane if $\beta = 2$. The $\mu_{i,j}^{L,M}$ satisfy a sharp sub-Gaussian Laplace transform (with real or complex values) for all $i, j \in [1, M] \times [1, N]$, and its complementary uniform lower bound (Assumption 1.1, or Assumption 1.3), are centered and have covariance one. We set $W_{L,M}^{(\beta)} = \frac{1}{L} G_{L,M}^{(\beta)} (G_{L,M}^{(\beta)})^*$. When M/L converges towards α, the spectral distribution of $W_{L,M}^{(\beta)}$ converges towards the Pastur-Marchenko law [18]: for any bounded continuous function f

$$\lim_{N\to\infty} \int f(x) d\hat{\mu}_{W_{L,M}^{(\beta)}}^L(x) = \int f(x) d\pi_\alpha(x) \qquad a.s$$

where if $\alpha \geq 1$ and $a_\alpha = (1-\sqrt{\alpha})^2, b_\alpha = (1+\sqrt{\alpha})^2$,

$$\pi_\alpha(dx) = \frac{\sqrt{(b_\alpha - x)(x-a_\alpha)}}{2\pi x} \mathbb{1}_{[a_\alpha, b_\alpha]} dx\,.$$

When $\alpha < 1$, the limiting spectral measure has aditionnally a Dirac mass at the origin with mass $1-\alpha$. We hereafter concentrates on the case $M \geq L$ up to replace $W_{L,M}^{(\beta)}$ by $(G_{L,M}^{(\beta)})^* G_{L,M}^{(\beta)}/M$. Again, the extreme eigenvalues were shown to stick to the bulk [5]. We prove a large deviation principle from this convergence:

Theorem 1.6. *Assume that the $\mu_{i,j}^N$ satisfy Assumption 1.2. Assume they satisfy a sharp Gaussian Laplace transform 1.1 when $\beta = 1$ or 1.3 when $\beta = 2$, and a uniform lower bounded Laplace transform 1.1 when $\beta = 1$ or 1.3 when $\beta = 2$. Assume that there exists $\alpha \geq 1$ and $\kappa > 0$ so that $\frac{M}{L} - \alpha = o(N^{-\kappa})$. Then, the law of the largest eigenvalue $\lambda_{\max}(W_{L,M}^{(\beta)})$ of $W_{L,M}^{(\beta)}$ satisfies a large deviation principle with speed N and good rate function $J^{(\beta)}$ which is infinite on $(-\infty, b_\alpha)$ and otherwise given by*

$$J^{(\beta)}(x) = \frac{\beta}{4(1+\alpha)} \int_{b_\alpha}^x \frac{\sqrt{(b_\alpha - y)(y-a_\alpha)}}{y} \mathbb{1}_{[a_\alpha, b_\alpha]} dy\,.$$

where $\beta = 1$ in the case of real entries, and $\beta = 2$ in the case of complex entries.

This problem can be seen as a generalization of the previous cases since if we consider the $N \times N$ matrix

$$X_N^{(w_\beta)} = \begin{pmatrix} 0 & \frac{1}{\sqrt{N}} G_{L,M}^{(\beta)} \\ \frac{1}{\sqrt{N}} (G_{L,M}^{(\beta)})^* & 0 \end{pmatrix}$$

the spectrum of the $N \times N$ matrix $X_N^{(w_\beta)}$ is given by L eigenvalues $\sqrt{\frac{L}{N}\lambda}$, L eigenvalues $-\sqrt{\frac{L}{N}\lambda}$, where λ are the eigenvalues of $W_{L,M}^{(\beta)}$, and $M-L$ vanishing eigenvalues. Hence, the largest eigenvalue of $W_{L,M}^{(\beta)}$ is the square of the largest eigenvalue of $X_N^{(w_\beta)}$ multiplied by N/L. It is therefore equivalent to show a large deviation principle for the largest eigenvalue of $X_N^{(w_\beta)}$, with rate function

$$I^{(w_\beta)}(x) = J^{(\beta)}((1+\alpha)x^2)\,.$$

This amounts to consider a Wigner matrix with some entries set to zero. We denote $a_{i,j}^{(w_\beta)}$ the entries of $\sqrt{N} X_N^{(w_\beta)}$:

$$\begin{aligned} a_{i,j}^{(w_\beta)} &= 0, \qquad \text{if } i,j \leq M \text{ or } i,j \geq M+1, \\ a_{i,j}^{(w_\beta)} &= a_{i-M,j}^{(\beta)}, \quad i \geq M+1, j \leq M, \\ a_{i,j}^{(w_\beta)} &= \overline{a}_{j-M,i}^{(\beta)}, \quad j \geq M+1, i \leq N. \end{aligned}$$

Again, we denote by $\mu_{i,j}^N$ the law of the i,jth entry of this matrix. Hereafter, we denote by σ_w the limiting spectral distribution of $X_N^{(w_\beta)}$ given for any test function f by

$$\int f(x) d\sigma_w(x) = \frac{1}{1+\alpha} \left(\int f(\sqrt{\frac{x}{1+\alpha}}) d\pi_\alpha(x) + \int f(-\sqrt{\frac{x}{1+\alpha}}) d\pi_\alpha(x) \right) + \frac{\alpha - 1}{\alpha + 1} f(0)\,.$$

Therefore, we shall prove Theorem 1.6 by showing that

Theorem 1.7. *Assume that the $\mu_{i,j}^N$ satisfy Assumption 1.2. Assume they satisfy a sharp Gaussian Laplace transform 1.1 when $\beta = 1$ or 1.3 when $\beta = 2$, and a uniform lower bounded Laplace transform 1.1 when $\beta = 1$ or 1.3 when $\beta = 2$. Assume that there exists $\alpha \geq 1$ and $\kappa > 0$ so that $\frac{M}{L} - \alpha = o(N^{-\kappa})$. Then, the law of the largest eigenvalue $\lambda_{\max}(X_N^{(w_\beta)})$ of $X_N^{(w_\beta)}$ satisfies a large deviation principle with speed N and good rate function $I^{(w_\beta)}$ which is infinite on $(-\infty, \tilde{b}_\alpha)$, $\tilde{b}_\alpha = \sqrt{(1+\alpha)^{-1} b_\alpha})$ and otherwise given by*

$$I^{(w_\beta)}(x) = \frac{\beta}{1+\alpha} \int_{\tilde{b}_\alpha}^x \frac{1}{y} \sqrt{(1+\alpha)^2 (y^2-1)^2 - 4\alpha} dy \, .$$

where $\beta = 1$ in the case of real entries, and two in the case of complex entries.

Acknowledgments: Alice Guionnet wishes to thank A. Dembo for long discussions about large deviations for the largest eigenvalue for sub-Gaussian matrices in Abu Dhabi in 2011. The idea to tilt measures by the spherical integral came out magically from a discussion with M. Potters in UCLA in 2017 and we wish to thank him for this beautiful inspiration. We also benifitted from many discussions with M. Maida with whom one of the author is working on a companion paper on unitarily invariant ensembles, as well as with Fanny Augeri with whom we are working on a follow up paper tackling the general sub-Gaussian case. Finally, we are very grateful for stimulating discussions with O. Zeitouni and N. Cook.

This work was supported by the LABEX MILYON (ANR-10-LABX-0070) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX- 0007) operated by the French National Research Agency (ANR).

1.2. **Scheme of the proof.** The idea of the proof is reminiscent of Cramer's approach to large deviations: we appropriately tilt measures to make the desired deviations likely. The point is to realize that it is enough to shift the measure in a random direction and use estimates on spherical intergrals obtained by one of the author and M. Maida [15]. To be more precise, we shall follow the usual scheme to prove first exponential tightness:

Lemma 1.8. *For $\beta = 1, 2, w_1, w_2$, assume that the distribution of the entries $a_{i,j}^{(\beta)}$ satisfy Assumption 1.2 for $\beta = 1, w_1$ and Assumption 1.3 for $\beta = 2, w_2$. Then:*

$$\lim_{K \to +\infty} \limsup_{N \to \infty} \frac{1}{N} \ln \mathbb{P}[\lambda_{\max}(X_N^{(\beta)}) > K] = -\infty$$

Similar results hold for $\lambda_{\min}(X_N^{(\beta)})$.

This result is proved in Section 2. Therefore it is enough to prove a weak large deviation principle.

In the following we summarize the assumptions on the distribution of the entries as follows :

Assumption 1.4. *Either the $\mu_{i,j}^N$ are uniformly compactly supported in the sense that there exists a compact set K such that the support of all $\mu_{i,j}^N$ is included in K, or the*

$\mu_{i,j}^N$ satisfy a uniform log-Sobolev inequality in the sense that there exists a constant c independent of N such that for all smooth function f

$$\int f^2 \ln \frac{f^2}{\mu_{i,j}^N(f^2)} d\mu_{i,j}^N \leq c\mu_{i,j}^N(\|\nabla f\|_2^2)\,.$$

When $\beta = 1, w_1$ $\mu_{i,j}^N$ satisfy Assumption 1.1, when $\beta = 2, w_2$, they satisfy Assumption 1.3. In the case of Wishart matrices, $\beta = w_1$ or w_2, we assume that there exists $\alpha > 1$ and $\kappa > 0$ so that $|\frac{M}{L} - \alpha| \leq N^{-\kappa}$ for N large enough. Moreover, for $\epsilon > 0$ small enough $\sup_{|t|\leq\epsilon} \sup_{i,j,N} |\partial_t^3 \ln T_{\mu_{i,j}^N}(t)|$ is finite.

We shall first prove that we have a weak large deviation upper bound:

Theorem 1.9. *Assume that Assumption 1.4 holds. Let $\beta = 1, 2, w_1, w_2$. Then, for any real number x,*

$$\limsup_{\delta\to 0} \limsup_{N\to\infty} \frac{1}{N} \ln \mathbb{P}\left(\left|\lambda_{\max}(X_N^{(\beta)}) - x\right| \leq \delta\right) \leq -I_\beta(x)$$

We shall then obtain the large deviation lower bound.

Theorem 1.10. *Assume that Assumption 1.4 holds. Let $\beta = 1, 2, w_1, w_2$. Then, for any real number x,*

$$\liminf_{\delta\to 0} \liminf_{N\to\infty} \frac{1}{N} \ln \mathbb{P}\left(\left|\lambda_{\max}(X_N^{(\beta)}) - x\right| < \delta\right) \geq -I_\beta(x)$$

To prove Theorem 1.9, we first show that the rate function is infinite below the right edge of the support of the limiting spectral distribution. To this end, we use that the spectral measure $\hat{\mu}_N$ converges towards its limit which much larger probability. We denote this limit σ_β: $\sigma_1 = \sigma_2 = \sigma$ and $\sigma_{w_1} = \sigma_{w_2} = \sigma_w$. We let d denote the Dudley distance:

$$d(\mu, \nu) = \sup_{\|f\|_L \leq 1} \left| \int f(x) d\mu(x) - \int f(x) d\nu(x) \right|,$$

where $\|f\|_L = \sup_{x\neq y} \left|\frac{f(x)-f(y)}{x-y}\right| + \sup_x |f(x)|$.

Lemma 1.11. *Assume that the $\mu_{i,j}^N$ are uniformly compactly supported or satisfy a uniform log-Sobolev inequality, as well as, in the case w_1, w_2, that there exists $\kappa > 0$ such that $|\frac{M}{N} - \alpha| \leq N^{-\kappa}$. Then, for $\beta = 1, 2, w_1, w_2$, there exists $\kappa' \in (0, \frac{1}{10} \wedge \kappa)$ such that*

$$\limsup_{N\to\infty} \frac{1}{N} \ln \mathbb{P}\left(d(\hat{\mu}_{X_N^{(\beta)}}^N, \sigma_\beta) > N^{-\kappa'}\right) = -\infty\,.$$

The proof of this lemma is given in the appendix. As a consequence, we deduce that the extreme eigenvalues can not deviate towards a point inside the support of the limiting spectral measure with probability greater than $e^{-N^{1+\kappa}}$ and therefore

Corollary 1.12. *Under the assumption of Lemma 1.11, For $\beta = 1, 2$ let x be a real number in $(-\infty, 2)$ or, for $\beta = w_1, w_2$, take $x < \tilde{b}_\alpha$. Then, for $\delta > 0$ small enough,*

$$\limsup_{N\to\infty} \frac{1}{N} \ln \mathbb{P}\left(|\lambda_{\max}(X_N^{(\beta)}) - x| \leq \delta\right) = -\infty\,.$$

Indeed, as soon $\delta > 0$ is small enough so that $x+\delta$ is smaller than $2-\delta$ for $\beta = 1, 2$ (resp $b_\alpha - \delta$ for $\beta = w_1, w_2$), $d(\hat{\mu}_N, \sigma_\beta)$ is bounded below by some $\kappa(\delta) > 0$ on $|\lambda_{\max}(X_N^{(\beta)}) - x| \leq \delta$. Hence, Lemma 1.11 implies the Corollary.

In order to prove the weak large deviation bounds for the remaining x's, we shall tilt the measure by using spherical integrals:

$$I_N(X, \theta) = \mathbb{E}_e[e^{\theta N \langle e, Xe \rangle}]$$

where the expectation holds over e which follows the uniform measure on the sphere $\mathbb{S}^{N-1}$ with radius one. The asymptotics of

$$J_N(X, \theta) = \frac{1}{N} \ln I_N(X, \theta)$$

were studied in [15] where it was proved that

Theorem 1.13. [15, Theorem 6]

If $(E_N)_{N\in\mathbb{N}}$ is a sequence of $N \times N$ real symmetric matrices when $\beta = 1$ and complex Gaussian matrices when $\beta = 2$ such that :

- *The sequence of empirical measures $\hat{\mu}_{E_N}^N$ weakly converges to a compactly supported measure μ,*
- *There are two reals $\lambda_{\min}(E), \lambda_{\max}(E)$ such that $\lim_{N\to\infty} \lambda_{\min}(E_N) = \lambda_{\min}(E)$ and $\lim_{N\to\infty} \lambda_{\max}(E_N) = \lambda_{\max}(E)$,*

and $\theta \geq 0$, then :

$$\lim_{N\to\infty} J_N(E_N, \theta) = J(\mu, \theta, \lambda_{\max}(E))$$

The limit J is defined as follows. For a compactly supported probability measure we define its Stieltjes transform G_μ by

$$G_\mu(z) := \int_{\mathbb{R}} \frac{1}{z - t} d\mu(t)$$

We assume hereafter that μ is supported on a compact $[a, b]$. Then G_μ is a bijection from $\mathbb{R} \setminus [a, b]$ to $]G_\mu(a), G_\mu(b)[\setminus\{0\}$ where $G_\mu(a), G_\mu(b)$ are taken as the limits of $G_\mu(t)$ when $t \to a^-$ and $t \to b^+$. We denote by K_μ its inverse and let $R_\mu(z) := K_\mu(z) - 1/z$ be its R-transform as defined by Voiculescu in [20] (defined on $]G_\mu(a), G_\mu(b)[$). In the sequel, for any compactly supported probability measure μ, we denote by $r(\mu)$ the right edge of the support of μ. In order to define the rate function, we now introduce, for any $\theta \geq 0$, and $\lambda \geq r(\mu)$,

$$J(\mu, \theta, \lambda) := \theta v(\theta, \mu, \lambda) - \frac{\beta}{2} \int \log\left(1 + \frac{2}{\beta}\theta v(\theta, \mu, \lambda) - \frac{2}{\beta}\theta y\right) d\mu(y),$$

with

$$v(\theta, \mu, \lambda) := \begin{cases} R_\mu(\frac{2}{\beta}\theta), & \text{if } 0 \leq \frac{2\theta}{\beta} \leq H_{\max}(\mu, \lambda) := \lim_{z\downarrow\lambda} \int \frac{1}{z-y} d\mu(y), \\ \lambda - \frac{\beta}{2\theta}, & \text{if } \frac{2\theta}{\beta} > H_{\max}(\mu, \lambda), \end{cases}$$

We shall later use that spherical integrals are continuous. We recall here Proposition 2.1 from [17] and Theorem 6. from [15]. We denote by $\|A\|$ the operator norm of the matrix A given by $\|A\| = \sup_{\|u\|_2=1} \|Au\|_2$ where $\|u\|_2 = \sqrt{\sum |u_i|^2}$.

Proposition 1.14. *For every $\theta > 0$, every $\kappa \in]0, 1/2[$, every $M > 0$, there exist a function $g_\kappa : \mathbb{R}^+ \to \mathbb{R}^+$ going to 0 at 0 such that for any $\delta > 0$ and N large enough, with B_N and B'_N such that $d(\hat{\mu}^N_{B_N}, \hat{\mu}^N_{B'_N}) < N^{-\kappa}$, $|\lambda_{\max}(B_N) - \lambda_{\max}(B_N)| < \delta$ and $\sup_N ||B_N|| \leq M$, $\sup_N ||B'_N|| \leq M$:*

$$|J_N(B_N, \theta) - J_N(B'_N, \theta)| < g_\kappa(\delta)\,.$$

From Theorem 1.13 and Proposition 1.14, we deduce that :

Corollary 1.15. *For every $\theta > 0$, every $\kappa \in]0, 1/2[$, every $M > 0$, for any $\delta > 0$ and μ a probability measure supported in $[-M, M]$, if we denote by $\mathcal{B}_N$ the set of symmetric matrices B_N such that $d(\mu_{B_N}, \mu) < N^{-\kappa}$, $|\lambda_{\max}(B_N) - \rho| < \delta$, and $\sup_N ||B_N|| \leq M$, for N large enough, we have :*

$$\limsup_{N\to\infty} \sup_{B_N \in \mathcal{B}_N} |J_N(B_N, \theta) - J(\mu, \theta, \rho)| \leq 2g_\kappa(\delta)$$

where g_κ is the function in Proposition 1.14.

By Lemma 1.8 and Lemma 1.11, it is enough to study the probability of deviations on the set where J_N is continuous:

Corollary 1.16. *Suppose Assumption 1.2 holds. For $\delta > 0$, take a real number x and set for M large (larger than $x + \delta$ in particular), $\mathcal{A}^M_{x,\delta}$ to be the set of $N \times N$ self-adjoint matrices given by*

$$\mathcal{A}^M_{x,\delta} = \{X : |\lambda_{\max}(X) - x| < \delta\} \cap \{X : d(\hat{\mu}^N_X, \sigma_\beta) < N^{-\kappa'}\} \cap \{X : \|X\| \leq M\}\,,$$

where κ' is chosen as in Lemma 1.11 . Let x be a real number, $\delta > 0$ and κ' as in Lemma 1.11. Then, for any $L > 0$, for M large enough

$$\mathbb{P}\left(\left|\lambda_{\max}(X^{(\beta)}_N) - x\right| < \delta\right) = \mathbb{P}\left(X^{(\beta)}_N \in \mathcal{A}^M_{x,\delta}\right) + O(e^{-NL})\,.$$

We are now in position to get an upper bound for $\mathbb{P}\left(X^{(\beta)}_N \in \mathcal{A}^M_{x,\delta}\right)$. In fact, by Tchebyshev inequality, for any $\theta \geq 0$,

$$\begin{aligned}\mathbb{P}\left(X^\delta_N \in \mathcal{A}^M_{x,\delta}\right) &\leq \mathbb{E}[I_N(X^{(\beta)}_N, \theta)] \exp\{-N \inf_{X \in \mathcal{A}^M_{x,\delta}} J_N(X, \theta)\} \\ &\leq \mathbb{E}[I_N(X^{(\beta)}_N, \theta)] \exp\{N(2g_\kappa(\delta) - J(\sigma_\beta, \theta, x) + o(\delta))\} \qquad (2)\end{aligned}$$

where we used that $x \to J(\sigma_\beta, \theta, x)$ is continuous. It is therefore central to derive the asymptotics of

$$F_N(\theta, \beta) = \frac{1}{N} \ln \mathbb{E}[I_N(X^{(\beta)}_N, \theta)]$$

and we shall prove in section 3 that

Theorem 1.17. *Suppose Assumption 1.4 holds. For $\beta = 1, 2, w_1, w_2$ and $\theta \geq 0$,*

$$\lim_{N\to\infty} F_N(\theta, \beta) = F(\theta, \beta)$$

with $F(\theta, \beta) = \theta^2/\beta$ if $\beta = 1, 2$ and when $\beta = w_i$, $i = 1, 2$:

$$F(\theta, w_i) = \sup_{x\in[0,1]} \{\frac{2\theta^2}{i} x(1-x) + \frac{i}{2(1+\alpha)} \ln(1-x) + \frac{i\alpha}{2(1+\alpha)} \ln x\} - iC_\alpha\,,$$

where $C_\alpha = \frac{1}{2(1+\alpha)} \ln(\frac{1}{1+\alpha}) + \frac{\alpha}{2(1+\alpha)} \ln \frac{\alpha}{1+\alpha}$

We therefore deduce from (2), Corollaries 1.16 and 1.15 , and Theorem 1.17, by first letting N going to infinity, then δ to zero and finally M to infinity, that

$$\limsup_{\delta\to 0} \limsup_{N\to\infty} \frac{1}{N} \ln \mathbb{P}\left(\left|\lambda_{\max}(X_N^{(\beta)}) - x\right| < \delta\right) \leq F(\theta,\beta) - J(\sigma_\beta,\theta,x)\,.$$

We next optimize over θ to derive the upper bound:

$$\limsup_{\delta\to 0} \limsup_{N\to\infty} \frac{1}{N} \ln \mathbb{P}\left(\left|\lambda_{\max}(X_N^{(\beta)}) - x\right| < \delta\right) \leq -\sup_{\theta\geq 0}\{J(\sigma_\beta,\theta,x) - F(\theta,\beta)\}\,. \quad (3)$$

To complete the proof of Theorem 1.9, we show in section 4 that, with the notations of Theorems 1.6,1.5, and 1.7,

Proposition 1.18. *For* $\beta = 1, 2, w_1, w_2$,

$$I_\beta(x) = \sup_{\theta\geq 0}\{J(\sigma_\beta,\theta,x) - F(\theta,\beta)\}\,.$$

To prove the complementary lower bound, we shall prove that

Lemma 1.19. *For* $\beta = 1, 2$, *for any* $x > 2$ *and for* $\beta = w_1, w_2$ *for any* $x > \tilde{b}_\alpha$, *there exists* $\theta = \theta_x \geq 0$ *such that for any* $\delta > 0$ *and* M *large enough,*

$$\liminf_{N\to\infty} \frac{1}{N} \ln \frac{\mathbb{E}[\mathbb{1}_{X_N^{(\beta)}\in\mathcal{A}_{x,\delta}^M} I_N(X_N^{(\beta)},\theta)]}{\mathbb{E}[I_N(X_N^{(\beta)},\theta)]} \geq 0\,.$$

This lemma is proved by showing that the matrix whose law has been tilted by the spherical integral is approximately a rank one perturbation of a Wigner matrix, from which we can use the techniques developped to study the famous BBP transition [6]. The conclusion follows since then

$$\begin{aligned}
\mathbb{P}\left(X_N^{(\beta)} \in \mathcal{A}_{x,\delta}^M\right) &\geq \frac{\mathbb{E}[\mathbb{1}_{X_N^\delta\in\mathcal{A}_{x,\delta}^M} I_N(X_N^{(\beta)},\theta_x)]}{\mathbb{E}[I_N(X_N^{(\beta)},\theta_x)]} \mathbb{E}[I_N(X_N^{(\beta)},\theta_x)] \exp\{-N \sup_{X\in\mathcal{A}_{x,\delta}^M} J_N(X,\theta_x)\} \\
&\geq \exp\{N(g_\kappa(\delta) + F(\theta_x,\beta) - J(\sigma_\beta,\theta_x,x) + o(\delta))\} \\
&\geq \exp\{-NI_\beta(x) - No(\delta)\}
\end{aligned}$$

where we finally used Theorem 1.17 and Lemma 1.19.

2. Exponential tightness

In this section we prove Lemma 1.8. We will use a standard net argument that we recall for completness. For $N \in \mathbb{N}$, let R_N be a 1/2-net of the sphere (i.e. a subset of the sphere $\mathbb{S}_{N-1}$ such as for all $u \in \mathbb{S}_{N-1}$ there is $v \in R_N$ such that $||u-v||_2 \leq 1/2$. Here the sphere is inside $\mathbb{R}^N$ for $\beta = 1, w_1$ and $\mathbb{C}^N$ for $\beta = 2, w_2$). We know that we can take R_N with cardinality smaller than 3^N. We notice that for $M > 0$

$$\mathbb{P}[||X_N^{(\beta)}|| \geq 4K] \leq 9^N \sup_{u,v\in R_N} \mathbb{P}[\langle X_N^{(\beta)} u, v\rangle \geq K] \quad (4)$$

Indeed, if we denote, for $v \in \mathbb{S}^{N-1}$, u_v to be an element of R_N such that $\|u_v - v\|_2 \leq 1/2$,

$$\|X_N^{(\beta)}\| = \sup_{v\in\mathbb{S}^{N-1}} \|X_N^{(\beta)} v\|_2 \leq \sup_{v\in\mathbb{S}^{N-1}} (\|X_N^{(\beta)} u_v\|_2 + \frac{1}{2}\|X_N^{(\beta)}\|)$$

so that

$$\|X_N^{(\beta)}\| \leq 2 \sup_{u\in R_N} \|X_N^{(\beta)} u\|_2 \tag{5}$$

Similarly, taking $v = \frac{X_N^{(\beta)} u}{\|X_N^{(\beta)} u\|_2}$, we find

$$\|X_N^{(\beta)} u\|_2 = \langle v, X_N^{(\beta)} u\rangle \leq \langle u_v, X_N^{(\beta)} u\rangle + \|v - u_v\|_2 \|X_N^{(\beta)} v\|_2$$

from which we deduce that

$$\|X_\beta^N\| \leq 4 \sup_{u,v\in R_N} \langle X_N^{(\beta)} u, v\rangle$$

and (4) follows. We next bound the probability of deviations of $\langle X_N^{(\beta)} v, u\rangle$ by using Tchebychev's inequality. For $\theta \geq 0$ we indeed have

$$\begin{aligned}
\mathbb{P}[\langle X_N^{(\beta)} u, v\rangle \geq K] &\leq \exp\{-\theta N K\}\mathbb{E}[\exp\{N\theta\langle X_N^{(\beta)} u, v\rangle\}] \\
&\leq \exp\{-\theta N K\}\mathbb{E}[\exp\left\{\sqrt{N}\left(2\sum_{i<j} \Re(a_{i,j}^{(\beta)} u_i \bar{v}_j) + \sum_i a_{i,i} u_i v_i\right)\right\}] \\
&\leq \exp\{-\theta N K\} \exp\left(\frac{\theta^2 N}{\beta'}(2\sum_{i<j} |u_i|^2|v_j|^2 + \sum_i |u_i|^2|v_i|^2)\right) \qquad (6)
\end{aligned}$$

where we used that the entries have a sharp sub-Gaussian Laplace transform. In the case of Wishart matrices, we bounded above some vanishing contributions by a non-negative term. When $\beta = w_i$, $\beta' = i$, otherwise $\beta' = \beta$. We can now complete the upper bound:

$$\begin{aligned}
\mathbb{P}[\langle X_N^{(\beta)} u, v\rangle \geq K] &\leq \exp\left(\frac{\theta^2 N}{\beta'} \frac{||u||_2^2||v||_2^2 + \langle u, v\rangle^2}{2} - \theta N K\right) \\
&\leq \exp\left(N\left(\frac{1}{\beta'} - K\right)\right)
\end{aligned}$$

where we took $\theta = 1$. We conclude that :

$$\mathbb{P}[\langle X_N^{(\beta)} u, v\rangle \geq K] \leq \exp\left(N(1-K)\right)$$

This complete the proof of the Lemma with (4).

3. Proof of Theorem 1.17

We consider in this section a random unitary vector e taken uniformly on the sphere $\mathbb{S}^{N-1}$ and independent of $X_N^{(\beta)}$. We define F_N by setting, for $\theta > 0$:

$$F_N(\theta, \beta) = \frac{1}{N} \ln \mathbb{E}_{X_N^{(\beta)}} \mathbb{E}_e[\exp(N\theta\langle e, X_N^{(\beta)} e\rangle)]$$

where we take both the expectation $\mathbb{E}_e$ over e and the expectation $\mathbb{E}_{X_N^{(\beta)}}$ over $X_N^{(\beta)}$. In this section we derive the asymptotics of $F_N(\theta,\beta)$. $F(\theta,\beta)$ is as in Theorem 1.17. We prove a refinment of Theorem 1.17, which shows that under our assumption of sharp sub-Gaussian tails, the random vector e stays delocalized under the tilted measure.

Proposition 3.1. *Suppose Assumption 1.1 holds if $\beta = 1, w_1$ and Assumption 1.3 holds if $\beta = 2, w_2$. Denote by $V_N^\epsilon = \{e \in \mathbb{S}^{N-1} : \forall i, |e_i| \leq N^{-1/4-\epsilon}\}$. Then, for $\epsilon \in (0, \frac{1}{4})$,*

$$F(\theta,\beta) = \lim_{N\to+\infty} F_N(\theta,\beta) = \lim_{N\to\infty} \frac{1}{N} \ln \mathbb{E}_e[\mathbb{1}_{e\in V_N^\epsilon} \mathbb{E}_{X_N^{(\beta)}}[\exp(N\theta\langle e, X_N^{(\beta)} e\rangle)]]$$

We first consider the case of Wigner matrices and then the case of Wishart matrices: in both cases the proof shows that the above delocalization holds (i.e we can restrict ourselves to vectors e in V_N^ϵ) and we shall not mention it in the following statements.

3.1. **Wigner matrices.** In this section we prove Theorem 1.17 in the case of Wigner matrices, namely:

Lemma 3.2. *Suppose Assumption 1.1 holds if $\beta = 1$ and Assumption 1.3 holds if $\beta = 2$. Then for any $\theta \geq 0$*

$$\lim_{N\to+\infty} F_N(\theta,\beta) = F(\theta,\beta) = \frac{\theta^2}{\beta}\,.$$

Proof. By denoting $L_\mu = \ln T_\mu$, we have :

$$\begin{aligned}
\mathbb{E}_{X_N^{(\beta)}}[\exp(N\theta\langle e, X_N^{(\beta)} e\rangle)] &= \mathbb{E}_{X_N^{(\beta)}}[\exp\{\sqrt{N}\theta(2\sum_{i<j} \Re(a_{i,j}^{(\beta)} e_j \bar{e}_i) + \sum_i a_{i,i}^{(\beta)} |e_i|^2)\}] \\
&= \exp\{\sum_{i<j} L_{\mu_{i,j}^N}(2\theta \bar{e}_i e_j \sqrt{N}) + \sum_i L_{\mu_{i,i}^N}(\theta |e_i|^2 \sqrt{N})\}
\end{aligned}$$

where we used the independence of the $(a_{i,j}^{(\beta)})_{i\leq j}$. Using that the entries have a sharp sub-Gaussian Laplace transform (using on the diagonal the weaker bound $L_{\mu_{i,i}^N}(t) \leq \frac{1}{\beta} t^2 + A|t|$) and $\sum e_i^2 = 1$, we deduce that:

$$\begin{aligned}
\mathbb{E}_{X_N^{(\beta)}}[\exp(N\theta\langle e, X_N^{(\beta)} e\rangle)] &\leq \mathbb{E}_e[\exp\{\frac{2N\theta^2}{\beta} \sum_{i<j} |e_i|^2 |e_j|^2 + \frac{N\theta^2}{\beta} \sum_i |e_i|^4 + A\sqrt{N}\theta \sum_i e_i^2\}] \\
&\leq \exp(N\frac{\theta^2}{\beta} + A\sqrt{N}\theta)
\end{aligned}$$

So that we have proved the upper bound that

$$\limsup_{N\to\infty} F_N(\theta,\beta) \leq \limsup_{N\to\infty} \sup_{e\in\mathbb{S}^{N-1}} \frac{1}{N} \ln \mathbb{E}_{X_N^{(\beta)}}[\exp(N\theta\langle e, X_N^{(\beta)} e\rangle)] \leq \frac{\theta^2}{\beta} \tag{7}$$

We next prove the corresponding lower bound. The idea is that the expectation over the vector e concentrates on delocalized eigenvectors with entries so that $\sqrt{N} e_i \bar{e}_j$ is going to zero for all i, j. As a consequence we will be able to use the uniform lower bound on the Laplace transform to lower bound $F_N(\theta,\beta)$.

Let $V_N^\epsilon = \{e \in \mathbb{S}^{N-1} : \forall i, |e_i| \leq N^{-1/4-\epsilon}\}$ be the subset of the sphere $\mathbb{S}^{N-1}$ with entries smaller than $N^{-1/4-\epsilon}$ for some $\epsilon \in (0, \frac{1}{4})$. We have that :

$$\mathbb{E}[\exp(N\theta\langle e, X_\beta^N e\rangle)] \geq \mathbb{E}_e[\mathbb{1}_{e\in V_N^\epsilon} \prod_{i<j} \exp\{L_{\mu^{N_{i,j}}}(2\sqrt{N}\theta\bar{e}_i e_j)\} \prod_i \exp\{L_{\mu_{i,i}^N}(\sqrt{N}\theta|e_i|^2)\}]$$

For $e \in V_N^\epsilon$, $2\sqrt{N}\theta|e_i e_j| \leq 2\theta N^{-\epsilon}$ so that :

$$\lim_{N\to+\infty} \sup_{e\in V_N^\epsilon} |2\sqrt{N}\theta e_i e_j| = 0$$

By the uniform lower bound on the Laplace transform of Assumptions 1.1 or 1.3, we deduce that for any $\delta > 0$

$$\mathbb{E}[\exp(N\theta\langle e, X_\beta^N e\rangle)] \geq \mathbb{P}_e[V_N^\epsilon] e^{N\frac{\theta^2}{\beta}(1-\delta)}. \tag{8}$$

We shall use that

Lemma 3.3. *For any $\epsilon \in (0, 1/4)$ we have*

$$\lim_{N\to\infty} \mathbb{P}_e[e \in V_N^\epsilon] = 1$$

.

As a consequence, we deduce from (8) that for any $\delta > 0$ and N large enough

$$\liminf_{N\to\infty} F_N(\theta, \beta) \geq (1-\delta)\frac{\theta^2}{\beta}$$

So that together with (7) we have proved the announced limit

$$\lim_{N\to\infty} F_N(\theta, \beta) = \frac{\theta^2}{\beta}$$

which completes the proof of Lemma 3.2.

Finally we prove Lemma 3.3. To this end we use the well known representation of the vector e as a renormalized (real or complex) Gaussian vector:

$$e = \frac{g}{\|g\|_2}$$

where $g = (g_1, ..., g_N)$ is a Gaussian vector of covariance matrix I_N. By the law of large numbers, we have the following almost sure limit :

$$\lim_{N\to\infty} \frac{||g||_2}{\sqrt{N}} = 1$$

We also have by the union bound

$$\mathbb{P}[\exists i \in [1, N], |g_i| > N^{1/4-\epsilon}/2] \leq N\mathbb{P}[|g_1| > N^{1/4-\epsilon}/2] \leq N\exp\{-\frac{1}{4}N^{1/2-2\epsilon}\}$$

from which the result follows.

□

3.2. **Wishart matrices.** In this subsection we prove Theorem 1.17 in the case of Wishart matrices, namely:

Lemma 3.4. *Let $\beta = w_1$ or w_2. Suppose Assumption 1.4 holds. Then for any $\theta \geq 0$, for $i = 1, 2$*

$$\lim_{N\to\infty} F_N(\theta, w_i) = F(\theta, w_i) = \sup_{x\in[0,1]} \{\frac{2\theta^2}{i} x(1-x) + \frac{i}{2(1+\alpha)} \ln(1-x) + \frac{i\alpha}{2(1+\alpha)} \ln x\} - iC_\alpha \,,$$

where $C_\alpha = \frac{1}{2(1+\alpha)} \ln(\frac{1}{1+\alpha}) + \frac{\alpha}{2(1+\alpha)} \ln \frac{\alpha}{1+\alpha}$. Moreover, the supremum is achieved at a unique $x_{\theta,\alpha}$ in $[0, 1]$ (as it maximizes a strictly concave function). $x_{\theta,\alpha}$ is the almost sure limit of $\|e_1\|_2^2$, the norm of the first M entries of e, under the tilted law

$$d\mathbb{P}^\theta(e) = \frac{\mathbb{E}_X[\exp\{\theta N \langle e, X_\beta^N e\rangle\}] d\mathbb{P}(e)}{\mathbb{E}_e[\mathbb{E}_X[\exp\{\theta N \langle e, X_\beta^N e\rangle\}]]} \,.$$

Proof. We have, with the same notations than in the previous case :

$$\mathbb{E}_{X_N^{w_i}}[\exp(N\theta\langle X_N^{w_i} e, e\rangle)] \;=\; \exp\left\{ \sum_{1\leq i\leq M, 1\leq j\leq L} L_{\mu_{i,j}^N}(\sqrt{N} 2\theta e_i^{(1)} \bar{e}_j^{(2)}) \right\}$$

where $e = (e^{(1)}, e^{(2)})$, that is $e^{(1)}$ is the vector made of the M first entries of e and $e^{(2)}$ the vector made of the L last entries of e. Using that the $\mu_{i,j}^N$ have a sharp sub-Gaussian Laplace transform and a uniform lower bounded Laplace transform, we deduce that with $V_N^\epsilon = \{e \in \mathbb{S}^{N-1} \,:\, |e_i| \leq N^{-1/4-\epsilon}\}$ we find that for any $\delta > 0$ and N large enough

$$\mathbb{E}_e[\mathbb{1}_{V_N^\epsilon} \exp\{(1-\delta)\frac{2\theta^2}{i} N \|e^{(1)}\|_2^2 \|e^{(2)}\|_2^2\}] \leq \mathbb{E}_{X_N^{w_i}}[I_N(\theta, w_i)] \leq \mathbb{E}_e[\exp\{\frac{2\theta^2}{i} N \|e^{(1)}\|_2^2 \|e^{(2)}\|_2^2\}] \tag{9}$$

where $\|e^{(1)}\|_2^2 = 1 - \|e^{(2)}\|_2^2$ follows a Beta law with parameters $(iM/2, iL/2)$, so its distribution is given by

$$\mathrm{Beta}_{iM/2,iL/2}(dx) = C_{M,L} x^{iM/2} (1-x)^{iL/2} \mathbb{1}_{x\in[0,1]} dx \,,$$

with $C_{M,L} = \Gamma(iN/2)/\Gamma(iM/2)\Gamma(iL/2)$. Therefore, Laplace method implies that

$$\begin{aligned} &\lim_{N\to\infty} \frac{1}{N} \ln \mathbb{E}_e[\exp\{\frac{2\theta^2}{i} N \|e^{(1)}\|_2^2 \|e^{(2)}\|_2\}] \qquad (10) \\ &= \sup_{x\in[0,1]} \{\frac{2\theta^2}{i} x(1-x) + \frac{i\alpha}{2(1+\alpha)} \ln(1-x) + \frac{i}{2(1+\alpha)} \ln x\} - iC_\alpha \,. \end{aligned}$$

(10) thus yields the expected upper bound. To get the lower bound in (9), observe that conditionning by $\|e^{(1)}\|_2$, the entries of $e^{(1)}$ and e^2 follow uniform laws on the sphere so that Lemma 3.3 applies. Hence, V_N^ϵ has probability going to one under this conditionnal measure and we can remove its indicator function in the lower bound of (9). We then apply Laplace method under the Beta law to conclude. Finally, we see from the above that for any set A, any $\delta > 0$

$$\mathbb{P}^\theta(\|e^{(1)}\|_2^2 \in A) \leq \exp\{-NF(\theta, w_i) + N\delta\} \int_A x^{iM/2} (1-x)^{iL/2} \exp\{\frac{2\theta^2}{i} Nx(1-x)\} dx$$

from which it follows by Laplace method that the law of $\|e^{(1)}\|_2^2$ follows a large deviation upper bound with speed N and good rate function which is infinite outside $[0,1]$ and otherwise given by

$$-\frac{2\theta^2}{i}x(1-x) - \frac{i}{2(1+\alpha)}\ln(1-x) - \frac{i\alpha}{2(1+\alpha)}\ln x + F(\theta, w_i)\,.$$

In particular $\|e^{(1)}\|_2^2$ converges almost surely towards the unique minimizer $x_{\theta,\alpha}$ of this strictly convex function.

□

4. Identification of the rate function

To complete the proof of the large deviation upper bound of Theorem 1.9, we need to identify the rate function, that is prove Proposition 1.18. This could a priori be done by saying that the rate function corresponds to the one that is well known for the Gaussian case. But for the sake of completness, we verify directly that we have the same result.

4.1. **Wigner matrices.** We first consider the case of Wigner matrices. Recall that we found for $\beta = 1, 2$

$$I_\beta(x) = \max_{\theta>0}\left(J(\sigma,\theta,x) - \frac{\theta^2}{\beta}\right)$$

where

$$J(\mu,\theta,\lambda) := \theta v(\theta,\mu,\lambda) - \frac{\beta}{2}\int \log\left(1 + \frac{2}{\beta}\theta v(\theta,\mu,\lambda) - \frac{2}{\beta}\theta y\right) d\mu(y),$$

with

$$v(\theta,\mu,\lambda) := \begin{cases} R_\mu(\frac{2}{\beta}\theta), & \text{if } 0 \le \frac{2\theta}{\beta} \le H_{\max}(\mu,\lambda) := \lim_{z\downarrow\lambda}\int \frac{1}{z-y}d\mu(y), \\ \lambda - \frac{\beta}{2\theta}, & \text{if } \frac{2\theta}{\beta} > H_{\max}(\mu,\lambda), \end{cases}$$

When $\mu = \sigma$, $R_\sigma(x) = x$ and $G_\sigma(\lambda) = \frac{1}{2}(\lambda - \sqrt{\lambda^2-4})$.

The critical points of $\varphi(\theta,x) = J(\sigma,\theta,x) - \frac{\theta^2}{\beta}$ for fixed x satisfy

$$\frac{2\theta}{\beta} = \partial_\theta J(\sigma,\theta,x)\,.$$

- For $\frac{2\theta}{\beta} \le G_\sigma(x)$, $\varphi(\theta)$ vanishes uniformly as $J(\sigma,\theta,x) = .\frac{\beta}{2}\int_0^{\frac{2}{\beta}\theta} R_\sigma(u)du = \frac{\theta^2}{\beta}$.
- For $\frac{2\theta}{\beta} > G_\sigma(x)$, the maximum is achieved at a solution of

$$\frac{2\theta_x}{\beta} = x - \frac{\beta}{2\theta_x}$$

which gives

$$\frac{2\theta_x}{\beta} = \frac{1}{2}(x + \sqrt{x^2-4}) = \frac{1}{G_\sigma(x)}\,.$$

Hence, $I_\beta(x) = \varphi(\theta_x, x)$. We can compute its derivative and since θ_x is a critical point of φ, we find

$$\partial_x I_\beta(x) = \partial_x \varphi(\theta_x, x) = \theta_x - \frac{\beta}{2} G_\sigma(x) = \beta\sqrt{x^2-4}$$

which proves the claim since $I_\beta(2) = 0$.

4.2. **Wishart matrices.** Let us now consider Wishart matrices and compute

$$I_{w_\beta}(x) = \max_{\theta>0} \left(J(\sigma_w, \theta, x) - F(\theta, w_\beta)\right) .$$

As in the previous proof we try to compute

$$\partial_x I_{w_\beta}(x) = \theta_x - \frac{\beta}{2} G_{\sigma_w}(x)$$

where θ_x is the argmax of $\varphi(\theta, x) = J(\sigma_w, \theta, x) - F(\theta, w_\beta)$. Note that the latter exists as φ is continuous in θ, going to $-\infty$ at infinity. To identify θ_x we remark that when it is larger than $\frac{\beta}{2} G_{\sigma_w}(x)$, it must satisfy, as a critical point of φ,

$$x = \partial_\theta F(\theta, w_\beta) + \frac{\beta}{2\theta} =: K(\theta) .$$

Our goal is therefore to identify K and in fact its inverse. Now, we claim that $\theta \to F(\theta, w_\beta)$ is analytic in a neighborhood of $\mathbb{R}^{+*}$. We recall that it is given in terms of $x_{\theta,\alpha}$, see Lemma 3.4. $x_{\theta,\alpha}$ is a maximizer, and therefore as a critical point it is solution of

$$\psi(x, \theta) = \frac{1}{\beta^2}\theta^2(1-2x) + \frac{\alpha}{(1+\alpha)x} - \frac{1}{(1+\alpha)(1-x)} = 0 .$$

Clearly $x \to \psi(x, \theta)$ takes its zeroes away from $0, 1$ and is analytic in a complex neighborhood of $[\epsilon, 1-\epsilon]$ for any $\epsilon > 0$. Moreover, at $\theta = \infty$, ψ vanishes at $x = 1/2$ only. But for $\Re(\theta) > \delta$, the real part of $-\partial_x \psi(\theta, x)$ is bounded below uniformly by some $c(\epsilon) > 0$ uniformly a complex neighborhood U_ϵ of $[\epsilon, 1-\epsilon]$ provided the imaginary part of θ is smaller than some $\kappa_{\epsilon,\delta} > 0$. Hence, the implicit function theorem implies that $\theta \to x_{\theta,\alpha}$, and so $F(., w_\beta)$, is analytic in a complex neighborhood of $\Re(\theta) \geq \delta$. We next show that for θ small enough,

$$F(\theta, w_\beta) = \frac{\beta}{2} \int_0^{\frac{2}{\beta}\theta} R_{\sigma_w}(u) du . \tag{11}$$

It is clearly lower bounded by this value as for any M

$$F(\theta, w_\beta) \geq \liminf_{N\to\infty} \frac{1}{N} \ln \mathbb{E}_{X_N^{(w_\beta)}}[1_{|\lambda_{\max}(X_N^{(w_\beta)})| \leq M} I_N(X_N^{(w_\beta)}, \theta)]$$

so that for $\frac{2\theta}{\beta} \leq G_{\sigma_w}(M)$,[15, Theorem 1.6] gives the lower bound. The upper bound is obtained similarly by using the exponential tightness which permits to restrict oneself to $\{|\lambda_{\max}| \leq M\}$. Therefore, we conclude that K is analytic in $\Re(\theta) > \delta$ and equals $K_{\sigma_w}(\frac{2\theta}{\beta})$ for small θ. We want to find the inverse of K. We thus look for an analytic extension of K_{σ_w}. But in fact K_{σ_w} satisfies an algebraic equation. Indeed, observe that

$$G_{\sigma_w}(x) = 2x G_{\pi_\alpha}((1+\alpha)x^2) + \frac{\alpha - 1}{(1+\alpha)x}$$

where it is well known that G_{π_α}, the Stieltjes transform of the Wishart matries, is solution of

$$(2z)^2 G_{\pi_\alpha}(z)^2 - 4z(z+1-\alpha)G_{\pi_\alpha}(z) + 4z - 8\alpha = 0\,.$$

We deduce that at least for small x, K_{σ_w} is solution of

$$((1+\alpha)K_{\sigma_w}(x)x+1-\alpha)^2-2(K_{\sigma_w}(x)+1-\alpha)((1+\alpha)xK_{\sigma_w}(x)+1-\alpha)+4(1+\alpha)K_{\sigma_w}(x)^2-8\alpha = 0\,.$$

As a consequence, K is also solution of this equation for all x, by analyticity. Now, we are looking for the inverse of K and so we deduce θ_x is solution of the equation

$$(\frac{2}{\beta}(1+\alpha)x\theta_x + 1 - \alpha)^2 - 2(x+1-\alpha)(\frac{2}{\beta}(1+\alpha)x\theta_x + 1 - \alpha) + 4(1+\alpha)x^2 - 8\alpha = 0\,.$$

For $\frac{2\theta_x}{\beta} \leq G_{\sigma_w}(x)$, the solution is

$$\frac{2}{\beta}\theta_x = \frac{2\alpha}{1+\alpha}\frac{x^2+1-\alpha-\sqrt{(x^2-1-\alpha)^2-4\alpha}}{2x^2} + \frac{1-\alpha}{1+\alpha}\frac{1}{x} = G_{\sigma_w}(x)\,.$$

but when $\frac{2\theta_x}{\beta} > G_{\sigma_w}(x)$ we have to take the other solution of the quadratic equation

$$\frac{2}{\beta}\theta_x = \frac{2\alpha}{1+\alpha}\frac{x^2+1-\alpha+\sqrt{(x^2-1-\alpha)^2-4\alpha}}{2x^2} + \frac{1-\alpha}{1+\alpha}\frac{1}{x}$$

As a result, we then have

$$\partial_x I_{w_\beta}(x) = \theta_x - \frac{\beta}{2}G_{\sigma_w}(x) = \frac{\beta\alpha}{1+\alpha}\frac{\sqrt{(x^2-1-\alpha)^2-4\alpha}}{x^2}$$

5. Large deviation lower bounds

Recall that we need to prove Lemma 1.19, that is find for any $x > 2$ (or $\tilde{b}_\alpha$ for Wishart matrices) a $\theta = \theta_x \geq 0$ such that for any $\delta > 0$ and M large enough,

$$\liminf_{N\to\infty} \frac{1}{N} \ln \frac{\mathbb{E}[\mathbb{1}_{X_N^{(\beta)}\in\mathcal{A}_{x,\delta}^M} I_N(X_N^{(\beta)},\theta)]}{\mathbb{E}[I_N(X_N^{(\beta)},\theta)]} \geq 0\,,$$

where we recall that

$$\mathcal{A}_{x,\delta}^M = \{X : |\lambda_{\max}(X) - x| < \delta\} \cap \{d(\hat{\mu}_X^N, \sigma_\beta) < N^{-\kappa'}\} \cap \{\|X\| \leq M\}\,,$$

For a vector e of the sphere $\mathbb{S}^{N-1}$ and X a random symmetric matrix, we denote by $\mathbb{P}_N^{(e,\theta)}$ the probability measure defined by :

$$d\mathbb{P}_N^{(e,\theta)}(X) = \frac{\exp(N\theta\langle Xe, e\rangle)}{\mathbb{E}_X[\exp(N\theta\langle Xe, e\rangle)]} d\mathbb{P}_N(X)$$

where $\mathbb{P}_N$ is the law of $X_N^{(\beta)}$. We have

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{X_N^{(\beta)}\in\mathcal{A}_{x,\delta}^M} I_N(X_N^{(\beta)},\theta)] &= \mathbb{E}_e[\mathbb{P}_N^{(e,\theta)}(\mathcal{A}_{x,\delta}^M)\mathbb{E}_X[\exp(N\theta\langle Xe,e\rangle)]] \\ &\geq \mathbb{E}_e[1_{e\in V_N^\epsilon}\mathbb{P}_N^{(e,\theta)}(\mathcal{A}_{x,\delta}^M)\mathbb{E}_X[\exp(N\theta\langle Xe,e\rangle)]] \qquad (12)\end{aligned}$$

where we recall that $V_N^\epsilon = \{e \in \mathbb{S}^{N-1} : |e_i| \leq N^{-1/4-\epsilon}\}$. The main point to prove the lower bound will be to show that $\mathbb{P}_N^{(e,\theta)}(\mathcal{A}_{x,\delta}^M)$ is close to one for delocalized vectors $e \in V_N^\epsilon$ and then proceed as before to show that V_N^ϵ has probability close to one under the tilted measure. More precisely, we will show that for $\epsilon \in (\frac{1}{8}, \frac{1}{4})$, we can find θ so that for any $x > 2$ (resp $x > \tilde{b}_\alpha$) and $\delta > 0$ we can find $\theta_x \geq 0$ so that for M large enough,

$$\lim_{N\to\infty} \inf_{e\in V_N^\epsilon} \mathbb{P}_N^{(e,\theta_x)}(\mathcal{A}_{x,\delta}^M) = 1 \tag{13}$$

This gives the desired estimate once we show that $\mathbb{P}_e(V_N^\epsilon)$ goes to one since we then deduce from (12) that for N large enough so that the above is greater than $1/2$

$$\mathbb{E}[\mathbb{1}_{X_N^{(\beta)}\in\mathcal{A}_{x,\delta}^M} I_N(X_N^{(\beta)},\theta)] \;\geq\; \frac{1}{2}\mathbb{E}_e[1_{e\in V_N^\epsilon}\mathbb{E}_{X_N^{(\beta)}}[\exp(N\theta\langle X_N^{(\beta)}e, e\rangle)]]$$

so that the desired estimate follows from Proposition 3.1. The first point is to show that

Lemma 5.1. *Take $\epsilon \in (0, \frac{1}{4})$. There exists $\kappa > 0$, for $\epsilon > 0$, for any θ,*

- *for K large enough:*

$$\lim_{N\to\infty} \sup_{e\in V_N^\epsilon} \mathbb{P}_N^{(e,\theta)}\left(\lambda_{\max}(X_N^{(\beta)}) \geq K\right) = 0$$

- $$\limsup_{N\to\infty} \sup_{e\in V_N^\epsilon} \mathbb{P}_N^{(e,\theta)}\left(d(\hat{\mu}^N_{X_N^{(\beta)}}, \sigma_\beta) > N^{-\kappa'}\right) = 0\,.$$

Proof. We hereafter fix a vector e on the sphere. The proof of the exponential tightness is exactly the same as for Lemma 1.8. Indeed, by Jensen's inequality, we have

$$\mathbb{E}_X[\exp(N\theta\langle X_N^{(\beta)}e, e\rangle)] \geq \exp\{N\theta\mathbb{E}_X[\langle X_N^{(\beta)}e, e\rangle]\} = 1$$

Moreover, by Tchebychev's inequality, for any $u, v, e \in \mathbb{S}^{N-1}$, we have

$$\begin{aligned}\int \mathbb{1}_{\langle X_N^{(\beta)}u,v\rangle\geq K} \exp(N\theta\langle X_N^{(\beta)}e, e\rangle) d\mathbb{P}_N &\leq \exp\{-NK\}\mathbb{E}_X[\exp(N\theta\langle X_N^{(\beta)}e, e\rangle + N\langle X_N^{(\beta)}u, v\rangle)] \\ &\leq \exp\{-NK\}\exp\{N\theta^2\sum_{i,j}|e_i\bar{e}_j + u_i\bar{v}_j|^2\} \\ &\leq \exp\{-NK + 4\theta^2 N\}\end{aligned}$$

from which we deduce after taking u, v on a δ-net as in Lemma 1.8 that

$$\mathbb{P}_N^{(e,\theta)}\left(\lambda_{\max}(X) \geq K\right) \leq 9^N \exp\{-\frac{1}{4}NK + 4\theta^2 N\}$$

which proves the first point. The second is a direct consequence of Lemma 1.11 and the fact that the log density of $\mathbb{P}_N^{(e,\theta)}$ with respect to $\mathbb{P}_N$ is bounded by $\theta N(|\lambda_{\max}(X)| + |\lambda_{min}(E)|)$ which is bounded by θKN with overwhelming probability by the previous point (recall that $\lambda_{\min}(X)$ satisfies the same bounds than $\lambda_{\max}(X)$).

□

Hence, the main point of the proof is to show that

Lemma 5.2. *Pick $\epsilon \in]\frac{1}{8}, \frac{1}{4}[$. For any $x > 2$ if $\beta = 1, 2$ and $x > \tilde{b}_\alpha$ if $\beta = w_1, w_2$, there exists θ_x such that for every $\eta > 0$,*

$$\lim_{N\to\infty} \sup_{e\in V_N^\epsilon} \mathbb{P}_N^{(e,\theta_x)}[|\lambda_{\max} - x| \geq \eta] = 0$$

Again, we first consider the simpler Wigner matrix case and then the case of Wishart matrices.

5.1. **Proof of Lemma 5.2 for Wigner matrices.** For $e \in V_N^\epsilon$ fixed, let $X^{(e),N}$ be a matrix with law $\mathbb{P}_N^{(e,\theta)}$. We have :

$$X^{(e),N} = \mathbb{E}[X^{(e),N}] + (X^{(e),N} - \mathbb{E}[X^{(e),N}])$$

where $\mathbb{E}[X]$ denotes the matrix with entries given by the expectation of the entries of the matrix X. We first show that $\mathbb{E}[X^{(e),N}]$ is approximately a rank one matrix.

Lemma 5.3. *For $\epsilon \in]\frac{1}{8}, \frac{1}{4}[$, there exists $\kappa(\epsilon) > 0$ so that for $e \in V_N^\epsilon$:*

$$\mathbb{E}[X^{(e),N}] = 2\theta e\, e^* + \Delta^{(e),N}$$

where the spectral radius of $\Delta^{(e),N}$ is bounded by $N^{-\kappa(\epsilon)}$ uniformly on $e \in V_N^\epsilon$.

Proof of the lemma. We can express the density of $\mathbb{P}_N^{(e,\theta)}$ as the following product :

$$\frac{d\mathbb{P}_N^{(e,\theta)}}{d\mathbb{P}_{X_N}}(X) = \prod_{i<j} \exp(2\sqrt{N}\Re(e_i\bar{e_j}a_{i,j}) - L_{\mu_{i,j}^N}(2\sqrt{N}e_i\bar{e}_j)) \prod_i \exp(\sqrt{N}|e_i|^2 a_{i,i} - L_{\mu_{i,i}^N}(\sqrt{N}|e_i|^2))$$

where the $a_{i,j}$ are defined as in the introduction, basically a rescaling of the entries by multiplication by $\sqrt{N}$.

So since we took our $a_{i,j}$ independent (for $i \leq j$), the entries $X_{i,j}^{(e),N}$ remain independent and their mean is given in function of the Taylor expansion of L as follows :

$$(\mathbb{E}[X^{(e),N})])_{i,j} = \frac{L'_{\mu_{i,j}^N}(2\sqrt{N}\theta e_i\bar{e}_j)}{\sqrt{N}} = \frac{2\theta}{\beta} e_i\bar{e}_j + \frac{\delta_{i,j}(2\sqrt{N}\theta e_i\bar{e}_j)N\theta^2|e_i|^2|e_j|^2}{\sqrt{N}}$$

if $i \neq j$, and if $i = j$

$$(\mathbb{E}[X^{(e),N}])_{i,i} = \frac{L'_{\mu_{i,i}^N}(\sqrt{N}\theta|e_i|^2)}{\sqrt{N}} = \frac{2\theta}{\beta} e_i\bar{e}_i + \frac{\delta_{i,i}(2\sqrt{N}\theta|e_i|^2)N\theta^2|e_i|^4}{\sqrt{N}}$$

where we used that by centering and variance one, $L'_{\mu_{i,j}^N}(0) = 0$, $HessL_{\mu_{i,j}^N}(0) = \frac{1}{\beta}Id$ for all $i \neq j, N$, $L''_{\mu_{i,i}^N}(0) = \frac{2}{\beta}$ for all i, N, and where

$$|\delta_{i,j}(t)| \leq 4 \sup_{|u|<t} \max_{i,j,N} \{|L^{(3)}_{\mu_{i,j}^N}(u)|\}.$$

In order to bound the spectral radius of this remainder term, we use the following lemma :

Lemma 5.4. *Let A be an Hermitian matrix and B a real symmetric matrix such that :*

$$\forall i,j, |A_{i,j}| \leq B_{i,j}$$

Then the spectral radius of A is smaller than the spectral radius of B.

Proof. Indeed, if we take u on the sphere such that $||Au||_2 = ||A||$, then, by denoting A' the matrix $(|A_{i,j}|)$ and u' the vector $(|u_i|)$, we have by the triangular inequality

$$||A|| = ||Au||_2 \leq ||A'u'||_2 \leq ||Bu'||_2 \leq ||B||\,.$$

□

Therefore, if we choose C so that $C \geq \sup_{N,i,j} \delta_{i,j}(2\sqrt{N}\theta e_i \bar{e}_j)\theta^2$ and set $|e|^2$ to be the vector with entries $(|e_i|^2)_{1\leq i\leq N}$, we have

$$||\Delta^{(e),N}|| \leq C\sqrt{N}|||e|^2(|e|^2)^t||$$

Since $|||e|^2(|e|^2)^*|| = |||e|^2||_2^2 = \sum_i e_i^4 \leq N^{-4\epsilon}$, we deduce that if we take $\epsilon' \in]1/8, 1/4[$ we have with $\kappa(\epsilon) = 1/2 - 4\epsilon$:

$$||\Delta^{(e),N}|| = N^{-\kappa(\epsilon)}\,.$$

□

Remark 5.5. *F. Augeri noticed that a maybe more elegant proof of this point would be to use Latala's etsimate:*

$$\mathbb{E}[|||Y|||] \leq C \sup_j \left(\sum_i |Y_{i,j}|^2\right)^{\frac{1}{2}}\,.$$

Now we denote :

$$\overline{X^{(e),N}} := X^{(e),N} - \mathbb{E}[X^{(e),N}]$$

The entries of $\overline{X^{(e),N}}$ are independent, centered of variance $\partial_z\partial_{\bar{z}} L_{\mu_{i,j}^N}(\theta e_i\bar{e}_j\sqrt{N})/N$. Recall that $\partial_z\partial_{\bar{z}} L_{\mu_{i,j}^N}(0) = 1$ and that the third derivative of the Laplace transform of the entries are uniformly bounded so that

$$\partial_z\partial_{\bar{z}} L_{\mu_{i,j}^N}(\theta e_i\bar{e}_j\sqrt{N}) = 1 + \delta_{i,j}(\sqrt{N}|e_ie_j|) = 1 + O(N^{-2\epsilon})$$

uniformly on V_N^ϵ.

We can then consider $\widetilde{X}^{(e),N}$ defined by : :

$$\widetilde{X}_{i,j}^{(e),N} = \frac{\overline{X}_{i,j}^{(e),N}}{\sqrt{\partial_z\partial_{\bar{z}} L_{\mu_{i,j}^N}(\theta e_i\bar{e}_j\sqrt{N})}}$$

Set $Y^{(e),N} = \overline{X}^{(e),N} - \widetilde{X}^{(e),N}$. So, we have

$$(Y^{(e),N})_{i,j} = \overline{X}_{i,j}^{(e),N}\left(1 - \frac{1}{\sqrt{\partial_z\partial_{\bar{z}} L_{\mu_{i,j}^N}(\theta e_i\bar{e}_j\sqrt{N})}}\right)$$

We next show that for all $\delta > 0$:

$$\lim_{N\to+\infty} \sup_{e\in V_N^\epsilon} \mathbb{P}[||Y^{(e),N}|| > \delta] = 0 \tag{14}$$

Indeed, we have the following lemma which is a variant of [1, Theorem 2.1.22] :

Lemma 5.6. *Consider for all $N \in \mathbb{N}$ a random Hermitian matrix A^N with independent subdiagonal entries which are centered and for all $k \in \mathbb{N}$:*

$$r_k^N = \max_{i,j} N^{-k/2}\mathbb{E}[|A_{i,j}^N|^k]$$

Suppose that there exists $N_0 \in \mathbb{N}, C > 0$ such that for $N \geq N_0$:

$$r_2^N \leq 1, \qquad r_k^N \leq k^{Ck}$$

Then for all $\delta > 0$, $\mathbb{P}[\lambda_{\max}(A^N) > 2 + \delta]$ goes to zero as N goes to infinity.

The proof of this lemma is strictly identical to Theorem 2.1.22 in [1] as we only need to estimate large moments of the matrix, which only requires upper bounds on moments of the entries (and not equality as assumed in [1]) as soon as the entries are centered. We apply this lemma to the matrices $Y^{(e),N}/\delta$ to derive (14): note that the hypothesis on the upper bound on moments is a clear consequence of our bounds on Laplace transform.

Hence, since

$$X^{(e),N} = \widetilde{X}^{(e),N} + \frac{2\theta}{\beta} ee^* + \Delta^{(e),N} + Y^{(e),N}\,,$$

we conclude by combining (14) and Lemma 5.3 that for $\epsilon \in]1/4, 1/8[$ and all $\delta > 0$

$$\lim_{N\to\infty} \sup_{e\in V_N^\epsilon} \mathbb{P}_N^{(e,\theta)}[||X^{(e),N} - (\widetilde{X}^{(e),N} + \frac{2\theta}{\beta} ee^*)|| > \delta] = 0 \tag{15}$$

since all estimates were clearly uniform on $e \in V_N^\epsilon$.

And so, to conclude we need only to identify the limit of $\lambda_{\max}(\widetilde{X}^{(e),N} + \frac{2\theta}{\beta} ee^*)$. It is given by the well known BBP transition. We however collect the main elements of the argument. To identify this limit, we easily see as in [9] that the eigenvalues of $\widetilde{X}^{(e),N} + \frac{2\theta}{\beta} ee^*$ satisfy

$$0 = \det(z - \widetilde{X}^{(e),N} - \frac{2\theta}{\beta} ee^*) = \det(z - \widetilde{X}^{(e),N}) \det(1 - \frac{2\theta}{\beta}(z - \widetilde{X}^{(e),N})^{-1} ee^*)$$

and therefore z is an eigenvalue away from the spectrum of $\widetilde{X}^{(e),N}$ iff

$$\langle e, (z - \widetilde{X}^{(e),N})^{-1} e\rangle = \frac{\beta}{2\theta}\,.$$

But it was shown in Theorem 2.15 of [10] that for all $z > 2$, all $v \in \mathbb{S}^{N-1}$, $\langle v, (z - \widetilde{X}^{(e),N})^{-1} v\rangle$ converges almost surely towards $G_\sigma(z)$ and therefore we conclude that the largest eigenvalue $\lambda_{\max}(\widetilde{X}^{(e),N} + \frac{2\theta}{\beta} ee^*)$, must converge towards the solution ρ_θ to

$$G_\sigma(\rho_\theta) = \frac{\beta}{2\theta}$$

as soon as it is strictly greater than 2. We find a unique solution to this equation: it is given by

$$\rho_\theta = \frac{2\theta}{\beta} + \frac{\beta}{2\theta}\,.$$

Reciprocally, for any $x > 2$, we can find $\theta_x = \frac{\beta}{2}(x + \sqrt{x^2-4})$ so that $x = \rho_{\theta_x}$. Hence, we have proved that for any sequence of vectors $e \in V_N^\epsilon$ we have the desired estimate for any $\eta > 0$

$$\lim_{N\to\infty} \sup_{e\in V_N^\epsilon} \mathbb{P}_N^{(e,\theta_x)}[|\lambda_{\max} - x| \geq \eta] = 0$$

which also entails the convergence of the supremum over V_N^ϵ and thus the Lemma.

5.2. **Proof of Lemma 5.2 for Wishart matrices.** We next prove Lemma 5.2 for Wishart matrices and fix $e = (e^{(1)}, e^2) \in V_N^\epsilon$. We decompose as in the previous proof

$$X^{(e),N} = \widetilde{X}^{(e),N} + \mathbb{E}[X^{(e),N}] + Y^{(e),N}\,,$$

where the entries of $\widetilde{X}^{(e),N}$ are centered and with covariance $1/N$ and $Y^{(e),N}$ goes to zero in norm. We then find by the same argument that

$$\mathbb{E}[X^{(e),N}] = \frac{2\theta}{i}\begin{pmatrix} 0 & e^{(1)}(e^2)^* \\ e^2(e^{(1)})^* & 0 \end{pmatrix} + \Delta^{(e),N}$$

where $\|\Delta^{(e),N}\| \leq N^{-\kappa(\epsilon)}$. Letting

$$S^{(e)} = \begin{pmatrix} e^{(1)} & 0 \\ 0 & e^2 \end{pmatrix} \quad \text{and } T^{(e)} = \begin{pmatrix} 0 & (e^2)^* \\ (e^{(1)})^* & 0 \end{pmatrix}$$

we notice that

$$\begin{pmatrix} 0 & e^{(1)}(e^2)^* \\ e^2(e^{(1)})^* & 0 \end{pmatrix} = S^{(e)}T^{(e)}\,.$$

Therefore, we need to find $z > \tilde{b}_\alpha$ such that

$$0 = \det(z - \widetilde{X}^{N,(e)} - \frac{2\theta}{i}S^{(e)}T^{(e)}) = \det(z - \widetilde{X}^{N,(e)})\det(1 - \frac{2\theta}{i}T^{(e)}(z - \widetilde{X}^{N,(e)})^{-1}S^{(e)}) \quad (16)$$

By writing $G_{\widetilde{X}^{N,(e)}}(z) = (z - \widetilde{X}^{N,(e)})^{-1}$ by blocks :

$$G_{\widetilde{X}^{N,(e)}}(z) = \begin{pmatrix} G_{1,1}(z) & G_{1,2}(z) \\ G_{2,1}(z) & G_{2,2}(z) \end{pmatrix} = \begin{pmatrix} zG_{\widetilde{X}^{N,(e)}(\widetilde{X}^{N,(e)})^*}(z^2) & \widetilde{X}^{N,(e)}G_{(\widetilde{X}^{N,(e)})^*\widetilde{X}^{N,(e)}}(z^2) \\ G_{(\widetilde{X}^{N,(e)})^*\widetilde{X}^{N,(e)}}(z^2)(\widetilde{X}^{N,(e)})^* & zG_{(\widetilde{X}^{N,(e)})^*\widetilde{X}^{N,(e)}}(z^2) \end{pmatrix}$$

where $G_{1,1}$ is $M \times M$, $G_{1,2}$ $N \times M$, $G_{2,2}$ $N \times N$, we get the simpler equation

$$\det\left(I - \frac{2\theta}{i}\begin{pmatrix} \langle e^2, G_{2,1}(z)e^{(1)}\rangle & \langle e^2, G_{2,2}(z)e^2\rangle \\ \langle e^{(1)}, G_{1,1}(z)e^{(1)}\rangle & \langle e^{(1)}, G_{1,2}(z)e^2\rangle \end{pmatrix}\right) = 0$$

Therefore, we need to find z such that

$$|1 - \frac{2\theta}{i}\langle e^2, G_{2,1}(z)e^{(1)}\rangle|^2 - \frac{4\theta^2}{i^2}\langle e^2, G_{2,2}(z)e^2\rangle\langle e^{(1)}, G_{1,1}(z)e^{(1)}\rangle = 0 \quad (17)$$

We are going to prove that

Lemma 5.7. *For any $\delta > 0$*

$$\begin{aligned}
\limsup_{N\to\infty} \sup_{e\in V_N^\epsilon} \mathbb{P}_N^{(e,\theta)} \left(|\langle e^{(1)}, G_{1,1}(z) e^{(1)} \rangle - z(1+\alpha)||e^{(1)}||_2^2 G_{MP(\alpha)}((1+\alpha)z^2)| > \delta \right) &= 0 \\
\limsup_{N\to\infty} \sup_{e\in V_N^\epsilon} \mathbb{P}_N^{(e,\theta)} \left(|\langle e^{2}, G_{2,2}(z) e^{2} \rangle - z(1+\alpha)||e^{2}||_2^2 G_{MP(1/\alpha)}((1+\alpha)z^2)| > \delta \right) &= 0 \\
\limsup_{N\to\infty} \sup_{e\in V_N^\epsilon} \mathbb{P}_N^{(e,\theta)} \left(|\langle e^{2}, G_{2,1}(z) e^{(1)} \rangle| > \delta \right) &= 0
\end{aligned}$$

where $G_{MP(\alpha)}$ is the Stieltjes transform of a Pastur Marchenko law with parameter α.

We first derive Lemma 5.2 assuming that Lemma 5.7 holds. We have seen in Lemma 3.4 that $\|e^{(1)}\|_2$ converges towards $x_{\theta,\alpha}$ almost surely. Therefore, we arrive to the limiting equation

$$(1+\alpha)^2 z^2 G_{MP(\alpha)}((1+\alpha)z^2) G_{MP(1/\alpha)}((1+\alpha)z^2) = \frac{i^2}{4\theta^2 x_{\theta,\alpha}(1-x_{\theta,\alpha})}$$

Now, we claim that $\varphi(\theta) = \theta^2 x_{\theta,\alpha}(1-x_{\theta,\alpha})$ is continuous, increasing, going from 0 to $+\infty$. As $x_{\theta,\alpha}$ is a complicated solution of θ (solution of a degree three polynomial equation), we use the following asymptotic characterization which easily follows from the previous large deviation considerations, see Lemma 3.4:

$$\frac{4\theta}{i} x_{\theta,\alpha}(1-x_{\theta,\alpha}) = \partial_\theta F(\theta, w_i)\,,$$

where we use that the derivatives of $x_{\theta,\alpha}$ vanishes as it is a critical point of the maximum. We moreover notice that $G(\theta) = F(\sqrt{\theta}, w_i)$ is convex in θ (as a supremum of convex functions). Hence,

$$\varphi(\theta) = \frac{i}{4}\theta \partial_\theta F(\theta, w_i) = \frac{i}{2}\theta^2 G'(\theta^2)$$

It follows that φ is smooth as F is and moreover

$$\varphi'(\theta) = i(\theta G'(\theta^2) + \theta^3 G''(\theta))\,.$$

But since φ is non negative, G' is non negative and so φ' is non negative for all $\theta \geq 0$. The fact that φ goes to infinity at infinity is clear as $x_{\theta,\alpha}$ then goes to $1/2$. Moreover, for $z > \tilde{b}_\alpha$, $z \mapsto zG_{MP(\alpha)}((1+\alpha)z^2)$ and $z \mapsto zG_{MP(1/\alpha)}((1+\alpha)z^2)$ are positive and decreasing, and therefore so are their product. Hence, there exist a $\theta_\alpha > 0$ so that for every $\theta \geq \theta_\alpha$, the equation above has a unique solution on $[\tilde{b}_\alpha, +\infty[$. Moreover, if we denote ρ_θ this solution, $\theta \mapsto \rho_\theta$ is a bijection from $[\theta_\alpha, +\infty[$ onto $[\tilde{b}_\alpha, +\infty[$.

Proof of Lemma 5.7. We recall that X a $M \times L$ Wishart matrix with centered entries with covariance one and sub-Gaussian tails, $e = (e^{(1)}, e^2)$ a unit vector and

$$G_{1,1}(z) = (z - XX^*)^{-1}, G_{22}(z) = (z - X^*X)^{-1}, G_{1,2}(z) = X(z - X^*X)^{-1}.$$

The first two points of the Lemma are direct consequences of [10, Theorem 2.5]. It remains to see that $\langle e^2, G_{2,1}(z) e^{(1)} \rangle$ goes to 0 as N goes to infinity. Because $G_{2,1}(z) = X(z - X^*X)^{-1}$ is not the resolvent of the Wishart matrix, but its multiplication by X,

we can not apply directly [10, Theorem 2.5]. We will give an elementary proof of this result,based on classical moment computations. Indeed, for $\varepsilon > 0$, on the set where $\{\|X^*X\| \leq b_\alpha + \varepsilon\}$, for $z^2 > b_\alpha + 2\varepsilon$ we can expand

$$\langle e^2, G_{2,1}(z)e^{(1)}\rangle = -\sum \frac{\langle e^{(1)}, X(X^*X)^k e^2\rangle}{z^{2k+1}} = -\sum_{k=1}^{K} \frac{\langle e^{(1)}, X(X^*X)^k e^2\rangle}{z^{2k+1}} + O\left(\frac{1}{\varepsilon}\left(\frac{b_\alpha+\varepsilon}{b_\alpha+2\varepsilon}\right)^{K+1}\right).$$

and hence it is enough to get the convergence in probability of K moments with $K \geq 2\varepsilon^{-1}\ln\varepsilon^{-1}$:

$$\lim_{N\to\infty} \langle e^{(1)}, X(X^*X)^k e^2\rangle = 0,\ k \leq K\,.$$

To this end we first prove that

$$\lim_{N\to\infty} \mathbb{E}[\langle e^{(1)} X(X^*X)^k e^2\rangle] = 0 \tag{18}$$

and then

$$\lim_{N\to\infty} Var(\langle e^{(1)}, X(X^*X)^k e^2\rangle) = 0\,. \tag{19}$$

We first prove (18). It is clearly true for $k = 0$ by centering of the entries and so we consider $k \geq 1$. Let's call $\mathcal{W}_k$ the set of words $(v_1, ..., v_{2k+1})$ of length $k+1$ so that $v_{2j} \in \{1, ..., L\}$ and $v_{2j+1} \in \{1, ..., M\}$. We use the following notation :

$$E_v = \mathbb{E}[a_{v_1,v_2} a_{v_2,v_3} ... a_{v_{2k+1},v_{2k+2}}]$$

We have

$$\mathbb{E}[\langle e^{(1)}, X(X^*X)^k e^2\rangle] = \frac{1}{N^{k+1/2}} \sum_{v\in\mathcal{W}_{2k+1}} e^{(1)}_{v_1} E_v e^2_{v_{2k+2}}$$

Given a word v, we can construct a bipartite graph G_v whose vertices are the $\{v_1, v_3, ...\} \cup \{L + v_2, L + v_4, ...\}$ of whose edges (occasionally multiple) are the $(L + v_{2i}, v_{2i-1})$ and $(L + v_{2i}, v_{2i+1})$. We denote $V^{(1)}(v)$ the number of vertices in G_v lying in $\{1, ..., L\}$, $V^2(v)$ the number of vertices in G_v lying in $\{L+1, ..., L+M\}$ and $V(v) = V^{(1)}(v) + V^2(v)$ and $A(v)$ the number of edges of G_v. If e is an edge of G_v, we denote $n_v(e)$ the multiplicity of this edge.

Let's recall that here the $a_{i,j}$ are independant but not identically distributed. Nevertheless their variance are 1 and their moments are bounded uniformly i.e. for every k there exists $C_k < +\infty$ such that :

$$\sup_{N,i,j} \mathbb{E}[|a_{i,j}|^k] \leq C_k$$

For every word v of length k, we can define $C_v = \prod_{j\leq k} C_j^{l(v,j)}$ where $l(v, j)$ is the number of edge of multiplicity j in G_v. we then have

$$|E_v| \leq C_v$$

We say that two words v, w are equivalent if there exists a bijection $\phi : \{1, ..., L\} \to \{1, ..., M\}$ and a bijection $\psi : \{1, ..., M\} \to \{1, ..., M\}$ such that $v_{2j} = \phi(w_{2j})$ and $v_{2j+1} = \psi(w_{2j+1})$. If two words v and w are equivalent then $C_v = C_w$.

Let $\mathcal{T}_k$ be a system of representants of words of length $k+1$ for this equivalency relationship. We have

$$\mathbb{E}[\langle e^{(1)}, X(X^*X)^k e^2\rangle] = \frac{1}{N^{k+1/2}} \sum_{j=2}^{2k+2} \sum_{t\in\mathcal{T}_{2k+1}, V(v)=j} \sum_{v|v\sim t} e^{(1)}_{v_1} E_v e^2_{v_{2k+2}}$$

Let's notice that if G_v has an edge of multiplicity 1, then $E_v = 0$ (since the $a_{i,j}$ are independant and centered). So for E_v to be non zero we need that $A(v) \leq (2k+1)/2$ so $A(v) \leq k$. Since G_v is connected $V(v) \leq A(v) + 1 \leq k + 1$. If $v \in \mathcal{W}_{2k+1}$, there exists $N_v := (L-1)...(L-V^{(1)}(v)+1)(M-1)...(M-V^2(v)+1) \leq N^{V(v)-2}$ equivalent words w_1 provided we fix v_1 and v_{2k+2} so we have the following bound :

$$\mathbb{E}[\langle e^{(1)}, X(X^*X)^k e^2\rangle] \leq \frac{1}{N^{k+1/2}} \sum_{j=2}^{k+1} \sum_{t\in\mathcal{T}_{2k+1}, V(t)=j} C_t N_t \sum_{1\leq v_1\leq L, 1\leq v_{2k+2}\leq M} |e^{(1)}_{v_1} e^2_{v_{2k+2}}|$$

By using the Cauchy Schwartz inequality, we have that :

$$\sum_{1\leq i\leq L, 1\leq j\leq M} |e^{(1)}_i e^2_j| \leq N\|e^{(1)}\|_2 \times \|e^2\|_2 \leq N$$

which yields

$$\mathbb{E}[\langle e^{(1)}, X(X^*X)^k e^2\rangle] \leq \frac{1}{N^{k-1/2}} \sum_{j=2}^{k+1} \sum_{t\in\mathcal{T}_{2k+1}, V(t)=j} C_t N^{j-2}$$

The leading order term here is in $N^{-1/2}$ for $k \geq 1$ and so

$$\lim_{N\to\infty} \sup_{\|e\|_2=1} |\mathbb{E}[\langle e^{(1)}, X(X^*X)^k e^2\rangle]| = 0\,.$$

We proceed similarly for the covariance (19):

$$\mathbb{V}\mathrm{ar}(\langle e^{(1)}, X(X^*X)^k e^2\rangle) = \frac{1}{N^{2k+1}} \sum_{v\in\mathcal{W}_{2k+1}, w\in\mathcal{W}_{2k+1}} e^{(1)}_{v_1} e^{(1)}_{w_1} T_{v,w} e^2_{v_{2k+2}} e^2_{w_{2k+2}}$$

Where $T_{v,w} = E_{v,w} - E_v E_w$ and $E_{v,w} = \mathbb{E}[a_{v_1,v_2} a_{v_2,v_3} ... a_{v_k,v_{k+1}} a_{w_1,w_2} a_{w_2,w_3} ... a_{w_k,w_{k+1}}]$ We extend naturally the previous definitions to couples of words. Let us now do the same analysis than before with couples of words. Let's take $\tilde{\mathcal{T}}_{2k+1}$ a system of representant for the equivalency relationship for couples of words. Let $(v, w) \in \tilde{\mathcal{T}}_{2k+1}$

First, if $G_{v,w}$ is not connected, since it is the union of two connected graphs G_v and G_w, we have that G_v and G_w don't have any edges in common and so, by independence of the entries $T_{v,w} = 0$. So we can assume that $G_{v,w}$ is connected.

Then several cases arise :

First, if $v_1 \neq w_1$ and $v_{2k+2} \neq w_{2k+2}$, then if one edge of $G_{v,w}$ is of multiplicity 1, then $T_{v,w} = 0$. So we can assume that all edges are of multiplicity at least 2. We deduce

that $A(v,w) \leq 2k+1$ and $V(v,w) \leq 2k+2$. Let $N_{v,w}$ be the number of couple of words equivalent to (v,w) provided $(v_1, w_1, v_{2k+2}, w_{2k+2})$ are fixed, we have $N_{v,w} \leq N^{2k-2}$. Hence

$$\sum_{(u,t)\sim(v,w)} e^{(1)}_{u_1} e^{(1)}_{t_1} T_{v,w} e^2_{u_{2k+2}} e^2_{t_{2k+2}} \leq N^{2k}(C_{v,w} - C_v C_w)$$

Then, if $v_1 = w_1$ and $v_{2k+2} \neq w_{2k+2}$ or if $v_1 \neq w_1$ and $v_{2k+2} = w_{2k+2}$, the same reasoning concerning the edges holds. So, we have $V(v,w) \leq 2k+2$ and if $N_{v,w}$ is the number of couple of words equivalent to (v,w) provided $(v_1, w_1, v_{2k+2}, w_{2k+2})$ are fixed, we have $N_{v,w} \leq N^{2k-1}$. If we are in the case $v_1 = w_1$:

$$\sum_{(u,t)\sim(v,w)} e^{(1)}_{u_1} e^{(1)}_{t_1} T_{v,w} e^2_{u_{2k+2}} e^2_{t_{2k+2}} \leq N^{2k} ||e^{(1)}||^2 (C_{v,w} - C_v C_w)$$

And lastly, $v_1 = w_1$ and $v_{2k+2} = w_{2k+2}$ we have again $N_{v,w} \leq N^{2k}$ and

$$\sum_{(u,t)\sim(v,w)} e^{(1)}_{u_1} e^{(1)}_{t_1} T_{v,w} e^2_{u_{2k+2}} e^2_{t_{2k+2}} \leq N^{2k} ||e^{(1)}||^2 ||e^2||^2 (C_{v,w} - C_v C_w)$$

So we have

$$\mathbb{V}\mathrm{ar}(\langle e^{(1)}, X(X^*X)^k e^2 \rangle) = O\left(\frac{1}{N}\right)$$

□

6. Appendix: Proof of Lemma 1.11

In this section, we want to prove that the assumptions 1.1 and 1.3 are verified if $\mu_{i,j}$ are supported inside a common compact K or satisfy a log-Sobolev inequality with a uniformly bounded constant c for the matrices $X_N^{(1)}, X_N^2, X_N^{w_1}, X_N^{w_2}$.

Lemma 6.1. *There exists $\kappa \in (0, \frac{1}{10})$ such that*

$$\lim_{N\to\infty} \frac{1}{N} \ln \mathbb{P}[d(\mu_{X_N^{(\beta)}}, \sigma_\beta) > N^{-\kappa}] = -\infty$$

for $\beta = 1, 2, w_1, w_2$, where σ_β is the semi-circle law when $\beta = 1, 2$ and the Pastur-Marchenko law with index α if $\beta = w$ (in the latter case we assume $M/N - \alpha = o(N^{-\kappa})$).

For this, we will use two concentration results respectively from [16] and [3].

Theorem 6.2. *By* [16, Theorem 1.4)] *(for the compact case) and* [16, Corollary 1.4 b)] *(for the logarithmic Sobolev case), we have for N large enough*

$$\limsup_{N\to\infty} \frac{1}{N^{7/6}} \ln \mathbb{P}[d(\mu_{X_N}, \mathbb{E}[\mu_{X_N}]) > N^{-1/6}] < 0$$

where d is the Dudley distance.

We therefore only need

Theorem 6.3. *([3, Theorem 4.1]) If we let for every N :*

$$F_{X_N^{(1)}}(x) = \mu_{X_N^{(1)}}(]-\infty, x])$$
$$F_{\sigma_1}(x) = \sigma_1(]-\infty, x])$$

we have that

$$\sup_{x\in\mathbb{R}} |F_{\sigma_1}(x) - \mathbb{E}[F_{X_N^{(1)}}(x)]| = O(N^{-1/4}) .$$

In order to conclude, we need only to use Lemma 1.8 to see that $F_N(-M)$ and $1-F_N(M)$ decay exponentially fast in N for some fixed M so that

$$d(\mathbb{E}[\mu_{X_N}], \sigma_1) \leq 4e^{-N}\|f\|_\infty + 2M\|f\|_L \sup_{x\in\mathbb{R}} |F(x) - \mathbb{E}[F_N(x)]| = o(N^{-1/6}) .$$

The same results hold in the complex case. For Wishart matrices, we rely on [4, Theorem w.1 and w.2]. Recall that $W_N = G_N G_N^*$.

Theorem 6.4. *([3, Theorem 4.1]) Assume that $M/N \in (1, \epsilon^{-1})$ for some fixed ϵ and M/N converges towards α. Then*

$$\sup_{x\in\mathbb{R}} |F_{\pi_\alpha}(x) - \mathbb{E}[F_{W_N}(x)]| = O(N^{-1/10}) .$$

We can as well use Lemma 1.8 to conclude that $1 - F_{W_N}(M)$ goes to zero like e^{-N} for M large enough. Finally, we conclude by noticing that since

$$\int f(x) d\mathbb{E}[\hat{\mu}_{X_N^w}](x) = \frac{N}{N+M}\int (f(\sqrt{\lambda}) + f(-\sqrt{\lambda}))d\hat{\mu}_{W_N}(\lambda) + \frac{M-N}{N} f(0),$$

we have

$$\begin{aligned}\left|\int f(x) d(\mathbb{E}[\hat{\mu}_{X_N^w}] - \sigma_w)(x)\right| &\leq \|f\|_\infty(|\frac{M}{N} - \alpha| + e^{-N}) + \int_0^M |\partial_\lambda f(\sqrt{\lambda})||F_{\pi_\alpha}(\lambda) - \mathbb{E}[F_{W_N}(\lambda)]|d\lambda \\ &\leq \|f\|_L(N^{-\kappa} + e^{-N} + 2MN^{-\frac{1}{10}})\end{aligned}$$

References

[1] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

[2] Fanny Augeri. Large deviations principle for the largest eigenvalue of Wigner matrices without Gaussian tails. *Electron. J. Probab.*, 21:Paper No. 32, 49, 2016.

[3] Z. D. Bai. Convergence rate of expected spectral distributions of large random matrices. I. Wigner matrices. *Ann. Probab.*, 21(2):625–648, 1993.

[4] Z. D. Bai. Convergence rate of expected spectral distributions of large random matrices. II. Sample covariance matrices. *Ann. Probab.*, 21(2):649–672, 1993.

[5] Z. D. Bai and Jack W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.*, 26(1):316–345, 1998.

[6] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697, 2005.

[7] G. Ben Arous, A. Dembo, and A. Guionnet. Aging of spherical spin glasses. *Probab. Theory Related Fields*, 120(1):1–67, 2001.

[8] G. Ben Arous and A. Guionnet. Large deviations for Wigner's law and Voiculescu's non-commutative entropy. *Probab. Theory Related Fields*, 108(4):517–542, 1997.
[9] F. Benaych-Georges, A. Guionnet, and M. Maida. Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *ArXiv e-prints*, September 2010.
[10] A. Bloemendal, L. Erdos, A. Knowles, H.-T. Yau, and J. Yin. Isotropic Local Laws for Sample Covariance and Generalized Wigner Matrices. *ArXiv e-prints*, August 2013.
[11] Charles Bordenave and Pietro Caputo. A large deviation principle for Wigner matrices without Gaussian tails. *Ann. Probab.*, 42(6):2454–2496, 2014.
[12] David S. Dean and Satya N. Majumdar. Large deviations of extreme eigenvalues of random matrices. *Phys. Rev. Lett.*, 97(16):160201, 4, 2006.
[13] Anne Fey, Remco van der Hofstad, and Marten J. Klok. Large deviations for eigenvalues of sample covariance matrices, with applications to mobile communication systems. *Adv. in Appl. Probab.*, 40(4):1048–1071, 2008.
[14] Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1(3):233–241, 1981.
[15] A. Guionnet and M. Maï da. A Fourier view on the R-transform and related asymptotics of spherical integrals. *J. Funct. Anal.*, 222(2):435–490, 2005.
[16] A. Guionnet and O. Zeitouni. Concentration of the spectral measure for large matrices. *Electron. Comm. Probab.*, 5:119–136, 2000.
[17] Mylène Maï da. Large deviations for the largest eigenvalue of rank one deformations of Gaussian ensembles. *Electron. J. Probab.*, 12:1131–1150, 2007.
[18] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Math. USSR Sb.*, 1:457–483, 1967. English translation of Mat. Sbornik 72 507–536.
[19] Pierpaolo Vivo, Satya N. Majumdar, and Oriol Bohigas. Large deviations of the maximum eigenvalue in Wishart random matrices. *J. Phys. A*, 40(16):4317–4337, 2007.
[20] Dan Voiculescu. The analogues of entropy and of Fisher's information measure in free probability theory. V. Noncommutative Hilbert transforms. *Invent. Math.*, 132(1):189–227, 1998.
[21] E. P. Wigner. On the distribution of the roots of certain symmetric matrices. *Annals Math.*, 67:325–327, 1958.

(Alice Guionnet) Université de Lyon, ENSL, CNRS, France
E-mail address: Alice.Guionnet@umpa.ens-lyon.fr

(Jonathan Husson) Université de Lyon, ENSL, CNRS, France
E-mail address: Jonathan.Husson@umpa.ens-lyon.fr