Electronic Journal of Probability

Electron. J. Probab. **25** (2020), no. 14, 1–24.
ISSN: 1083-6489 https://doi.org/10.1214/19-EJP405

Large deviations for the largest eigenvalue of the sum of two random matrices*

Alice Guionnet† Mylène Maïda‡

Abstract

In this paper, we consider the addition of two matrices in generic position, namely $A+UBU^*$, where U is drawn under the Haar measure on the unitary or the orthogonal group. We show that, under mild conditions on the empirical spectral measures of the deterministic matrices A and B, the law of the largest eigenvalue satisfies a large deviation principle, in the scale N, with an explicit rate function involving the limit of spherical integrals. We cover in particular the case when A and B have no outliers.

Keywords: random matrix; large deviations; extreme eigenvalues; free convolution.
AMS MSC 2010: 15A52; 60F10; 46L54.
Submitted to EJP on November 23, 2018, final version accepted on December 22, 2019.
Supersedes arXiv:1810.02538.
Supersedes HAL:hal-01887673.

1 Introduction

Understanding the spectrum of the sum $A+B$ of two Hermitian matrices knowing the spectra of A and B respectively is a classical and difficult problem. Since the pioneering works of Voiculescu [27], we know that free probability provides efficient tools to describe, at least asymptotically, the spectrum of the sum of two large Hermitian matrices in generic position from one another. More precisely, if A_N and B_N are two deterministic $N \times N$ Hermitian matrices and U_N is a unitary random matrix distributed according to the Haar measure, then, in the large N limit, A_N and $U_N B_N U_N^*$ are asymptotically free and the spectral distribution of $H_N := A_N + U_N B_N U_N^*$ is given by the free convolution of the spectral distributions of A_N and B_N. This global law, that is the convergence of the spectral distribution of H_N at macroscopic scale, has been studied in details in [26, 25] among others. The local law, that is the comparison of the spectral

*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) and by the Labex CEMPI (ANR-11-LABX-0007-01) of Université de Lille operated by the French National Research Agency (ANR).
†École Normale Supérieure de Lyon, France. E-mail: aguionne@umpa.ens-lyon.fr
‡Université de Lille, France. E-mail: mylene.maida@univ-lille.fr

distribution of H_N with the free additive convolution of the spectral distributions of A_N and B_N below the macroscopic scale was then investigated in [22] and [4]. In this paper, we will be interested in the behavior of the largest eigenvalue of H_N. As a corollary of the results of [14] on strong asymptotic freeness, we know that if A_N and B_N have no outliers, then the largest eigenvalue of H_N converges to the right edge of the support of the free convolution of the spectral distributions of A_N and B_N. In this work, we investigate the large deviations of this extreme eigenvalue.

In the framework of random matrix theory, there are very few large deviation results known about the spectrum, basically because the eigenvalues are complicated functions of the entries. A notable exception is given by the Gaussian invariant ensembles for which the joint law of the eigenvalues can be explicitly written as a Coulomb gas. Based on this explicit formula, large deviation principles for the empirical spectral distribution at global scale have been established in [9] and for the largest eigenvalue in [8]. Another special case is given by the sum of a deterministic matrix and a Gaussian invariant ensemble. Then, the spectrum can be constructed as the realization at time one of a Hermitian (or symmetric) Brownian motion starting from a given deterministic matrix. This point of view was used by [21] to study the large deviations of the empirical measure, and the large deviations for the process of the largest eigenvalue starting from the origin were derived in [16]. One of the applications of the present paper is to provide the large deviations for the largest eigenvalue of this sum by using another approach based on spherical integrals. Beyond these cases where specific tools are available, it was observed by [12] that deviations of the spectrum of Wigner matrices for which the distribution of the entries has a tail which is heavier than Gaussian are naturally created by large entries. This key remark allowed to obtain the large deviations for the empirical measure in [12] (see also [18] for the counterpart for covariance matrices) and for the largest eigenvalue in [2]. Large deviations for the spectrum of Wigner matrices with sub-Gaussian entries is still completely open as far as the empirical measure is concerned. One can mention the deviations results of [3] for the moments of the empirical spectral distribution in several models. Concerning the deviations of the largest eigenvalue, beyond the works [8, 16, 2] already cited above, the following models have been so far studied : Gaussian ensembles plus a rank one perturbation by [23], very thin covariance matrices by [17], finite rank perturbations of deterministic matrices or unitarily invariant ensembles by [10]. In a companion paper [19], Guionnet and Husson have established a large deviation principle for the largest eigenvalue of Wigner matrices with entries having sharp sub-Gaussian tails, such as Rademacher matrices. They show that the speed and the rate function of this large deviation principle are the same as in the Gaussian case.

2 Statement of the results

Let $(A_N)_{N\geq 1}$ and $(B_N)_{N\geq 1}$ be two sequences of deterministic real diagonal matrices, with A_N and B_N of size $N\times N$. We denote by $\lambda_1^{(A_N)} \geq \ldots \geq \lambda_N^{(A_N)}$ and $\lambda_1^{(B_N)} \geq \ldots \geq \lambda_N^{(B_N)}$ their respective eigenvalues in non increasing order, by

$$\|A_N\| := \max(|\lambda_1^{(A_N)}|, |\lambda_N^{(A_N)}|) \text{ and } \|B_N\| := \max(|\lambda_1^{(B_N)}|, |\lambda_N^{(B_N)}|)$$

their respective spectral radius. We define by

$$\hat{\mu}_{A_N} := \frac{1}{N}\sum_{j=1}^N \delta_{\lambda_j^{(A_N)}} \text{ and } \hat{\mu}_{B_N} := \frac{1}{N}\sum_{j=1}^N \delta_{\lambda_j^{(B_N)}}$$

their respective empirical spectral distributions.

For $\beta = 1$ or 2, we denote by m_N^β the Haar measure on the orthogonal group $\mathcal{O}_N$ if $\beta = 1$ and on the unitary group $\mathcal{U}_N$ if $\beta = 2$. For any $N \times N$ unitary matrix U, we denote by $H_N(U) := A_N + UB_NU^*$ and by $\lambda_{\max}^N$ the largest eigenvalue of $H_N(U)$. The goal of the present work is to establish a large deviation principle for the law of $\lambda_{\max}^N$ under the Haar measure m_N^β. This large deviation principle holds under mild assumptions that we now detail.

Assumption 2.1.

(H_{bulk}) *The sequences of empirical spectral distributions $(\hat{\mu}_{A_N})_{N\geq 1}$ and $(\hat{\mu}_{B_N})_{N\geq 1}$ converge weakly as N grows to infinity respectively to μ_a and μ_b, compactly supported on $\mathbb{R}$. Moreover, $\sup_{N\geq 1}(\|A_N\| + \|B_N\|) < \infty$.*

(H_{edge}) *The largest eigenvalues $\lambda_1^{(A_N)}$ and $\lambda_1^{(B_N)}$ converge as N grows to infinity to ρ_a and ρ_b respectively.*

A key argument of the proof will be a tilt of the measure by a rank one spherical integral. The rank one spherical integral is defined as follows: for any $\theta \geq 0$ and M_N an Hermitian matrix of size N,

$$I_N^\beta(\theta, M_N) := \int \mathrm{e}^{N\theta(UM_NU^*)_{11}} m_N^\beta(\mathrm{d}U) \quad \text{and} \quad J_N^\beta(\theta, M_N) := \frac{1}{N}\log I_N^\beta(\theta, M_N).$$

The rate function of our large deviation principle will crucially involve the limit $J_\mu^\beta(\theta,\rho)$ of $J_N^\beta(\theta, H_N)$ as N grows to infinity, which we now describe. For μ a compactly supported probability measure on $\mathbb{R}$, we denote by $\mathsf{r}(\mu)$ the right edge of the support of μ and by G_μ the Stieltjes transform of μ given for $\lambda > \mathsf{r}(\mu)$ by

$$G_\mu(\lambda) := \int \frac{1}{\lambda - y}\mu(\mathrm{d}y).$$

It is decreasing on the interval $(\mathsf{r}(\mu), \infty)$. By taking the limit as λ decreases to $\mathsf{r}(\mu)$, one can also define $G_\mu(\mathsf{r}(\mu)) \in \mathbb{R}_+ \cup \infty$. As G_μ is bijective from $(\mathsf{r}(\mu), \infty)$ to $(0, G_\mu(\mathsf{r}(\mu)))$, one can define its inverse on the latter interval, that we denote by K_μ. Then, for any $z \in (0, G_\mu(\mathsf{r}(\mu)))$, we define

$$R_\mu(z) := K_\mu(z) - \frac{1}{z}.$$

The function R_μ is called the R-transform of μ. One can check that R_μ is an increasing function and that $\lim_{z\to 0} R_\mu(z) = \int \lambda\mu(\mathrm{d}\lambda)$: R_μ is bijective from $(0, G_\mu(\mathsf{r}(\mu)))$ to $\left(\int \lambda\mu(\mathrm{d}\lambda), \mathsf{r}(\mu) - \frac{1}{G_\mu(\mathsf{r}(\mu))}\right)$. We denote by Q_μ its inverse on this interval. We can now define, for $\beta = 1$ or 2, $\theta \geq 0$, μ a compactly supported probability measure and $\rho \geq \mathsf{r}(\mu)$:

$$J_\mu^\beta(\theta,\rho) := \begin{cases} \frac{\beta}{2}\int_0^{\frac{2\theta}{\beta}} R_\mu(u)\mathrm{d}u, & \text{if } 0 \leq \frac{2\theta}{\beta} \leq G_\mu(\rho), \\ \theta\rho - \frac{\beta}{2}\log\theta - \frac{\beta}{2}\int\log(\rho - y)\mu(\mathrm{d}y) + \frac{\beta}{2}\left(\log\frac{\beta}{2} - 1\right), & \text{if } \frac{2\theta}{\beta} > G_\mu(\rho). \end{cases}$$

The convergence of $J_N^\beta(\theta, M_N)$ towards $J_\mu^\beta(\theta,\rho)$, obtained by the authors in [20], will be stated precisely in Lemma 3.4. At this point, we want to emphasize that, for θ large enough, the limit depends not only on the limiting spectral distribution μ but also of the limit ρ of the largest eigenvalue of M_N: this observation is crucial in our use of the spherical integral to produce an interesting tilt. If μ_1 and μ_2 are two probability measures compactly supported on $\mathbb{R}$, we denote by $\mu_1 \boxplus \mu_2$ the free convolution of μ_1 and μ_2. It is uniquely determined as the unique probability measure with R-transform equal to the sum of the R-transforms of μ_1 and μ_2 (see [27]). For any $\theta \geq 0$ and $x \geq \mathsf{r}(\mu_a \boxplus \mu_b)$, we denote by

$$I^\beta(\theta, x) := J_{\mu_a\boxplus\mu_b}^\beta(\theta, x) - J_{\mu_a}^\beta(\theta,\rho_a) - J_{\mu_b}^\beta(\theta,\rho_b),$$

and

$$I^{\beta}(x) := \begin{cases} \sup_{\theta\geq 0} I^{\beta}(\theta, x), & \text{if } x \geq \mathsf{r}(\mu_a \boxplus \mu_b), \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.1}$$

It is easy to check the following:

Lemma 2.2. *Let μ_a, μ_b, ρ_a and ρ_b be given as in Assumption 2.1. For $\beta = 1$ or 2, the function I^{β} is a good rate function, that is for any $\alpha \in \mathbb{R}$, the level set $\{I^{\beta} \leq \alpha\}$ is a compact subset of $\mathbb{R}$. Moreover, for any $x > \rho_a + \rho_b$, $I^{\beta}(x) = +\infty$.*

The proof will be given at the beginning of Section 4. We can now state the main results of this paper. The first result is the following large deviation upper bound:

Proposition 2.3. *Under Assumption 2.1, for $\beta = 1$ or 2, for any $x \in \mathbb{R}$ such that*

$$G_{\mu_a \boxplus \mu_b}(x) \leq \min\left(G_{\mu_a}(\rho_a), G_{\mu_b}(\rho_b)\right), \tag{2.2}$$

we have

$$\lim_{\delta\downarrow 0} \limsup_{N\to+\infty} \frac{1}{N} \log m_N^{\beta}\left(\lambda_{\max}^N \in [x-\delta, x+\delta]\right) \leq -I^{\beta}(x).$$

We will then derive the following large deviation lower bound:

Proposition 2.4. *Under Assumption 2.1, for $\beta = 1$ or 2, for any $x \in \mathbb{R}$ such that*

$$G_{\mu_a \boxplus \mu_b}(x) \leq \min\left(G_{\mu_a}(\rho_a), G_{\mu_b}(\rho_b)\right), \tag{2.3}$$

we have

$$\lim_{\delta\downarrow 0} \liminf_{N\to+\infty} \frac{1}{N} \log m_N^{\beta}\left(\lambda_{\max}^N \in [x-\delta, x+\delta]\right) \geq -I^{\beta}(x).$$

This leads to the following important corollary:

Theorem 2.5. *Under Assumption 2.1 and if moreover,*

$$G_{\mu_a \boxplus \mu_b}(\mathsf{r}(\mu_a \boxplus \mu_b)) \leq \min\left(G_{\mu_a}(\rho_a), G_{\mu_b}(\rho_b)\right), \tag{NoOut}$$

then, for $\beta = 1$ or 2, the law of $\lambda_{\max}^N$ under m_N^{β} satisfies a large deviation principle in the scale N with good rate function I^{β}. More precisely, for any F closed Borel subset of $\mathbb{R}$,

$$\limsup_{N\to+\infty} \frac{1}{N} \log m_N^{\beta}\left(\lambda_{\max}^N \in F\right) \leq -\inf_F I^{\beta},$$

and for any O open Borel subset of $\mathbb{R}$,

$$\liminf_{N\to+\infty} \frac{1}{N} \log m_N^{\beta}\left(\lambda_{\max}^N \in 0\right) \geq -\inf_O I^{\beta}.$$

A few remarks have to be made on the condition (NoOut). Under assumptions that are slightly stronger than Assumption (H_{bulk}), [7] established that, whenever (NoOut) is satisfied, $A_N + UB_NU^*$ has no outlier, that is, its largest eigenvalue converges to $\mathsf{r}(\mu_a \boxplus \mu_b)$. Another related remark is that, if A_N and B_N have no outliers, namely $\rho_a = \mathsf{r}(\mu_a)$ and $\rho_b = \mathsf{r}(\mu_b)$, then the condition (NoOut) is automatically satisfied. This will be stated in Lemma 6.1 and leads to the following corollary

Corollary 2.6. *Under the assumption (H_{bulk}), if A_N and B_N have no outliers, then for $\beta = 1$ or 2, the law of $\lambda_{\max}^N$ under m_N^{β} satisfies a large deviation principle in the scale N with good rate function I^{β}.*

From there, one can recover partly Theorem 3.2. in [23].

Remark 2.7. If we choose A_N to be a rank one deterministic matrix with eigenvalue $\rho_a > 0$ and UB_NU^* to be a random matrix from the Gaussian Unitary (or Orthogonal) Ensemble, one can study the largest eigenvalue of $A_N + UB_NU^*$ by conditioning on the deviations of the largest eigenvalue of UB_NU^*. These large deviations were obtained in [8] and we denote by J^β its rate function. If $\rho_a \leq \sqrt{\frac{\beta}{2}}$, we know that the deformed model has no outliers and one can apply Corollary 5. For any $x \geq \sqrt{2\beta}$, the rate function of the deformed model is given by $K^\beta(x) := \inf_{\sqrt{2\beta} \leq y \leq x}(J^\beta(y) + I^\beta(x))$, where I^β corresponds to $\mu_a = \delta_0$, $\mu_b = \sigma_\beta$ and $\rho_b = y$. Standard computations allow to identify the rate function as the funtion $K^\beta_{\rho_a}$ in [23].

To get a taste of what happens in the case with outliers, we also consider in Appendix A the following model: let $(U_1^{(1)}, \ldots, U_1^{(p)})$ be independent random vectors uniformly distributed on the unit sphere (in $\mathbb{R}^N$ if $\beta = 1$ and $\mathbb{C}^N$ if $\beta = 2$) and $\gamma_1, \ldots, \gamma_p$ be nonnegative real numbers. We consider the following deformed model :

$$X_N := A_N + UB_NU^* + \sum_{i=1}^{p} \gamma_i U_1^{(i)} (U_1^{(i)})^*. \tag{2.4}$$

We show in Theorem A.5 that we still have a large deviation principle, for which the rate function will depend on the γ_i's. The rest of the paper will be organized as follows: in the next section, we will first prove a more general result than Proposition 2.3, that holds not only for m_N^β but also for a whole family of tilted measures. This will be helpful in the proof of Proposition 2.4, that will be developed in Section 5. Before getting there, we will study in Section 4 some properties of the rate function I^β. The last section will be devoted to the proof of Theorem 2.5 and Corollary 2.6, with Lemma 6.1 as prerequisite. At the end of the paper, in Appendix A, we will study the deviations of the largest eigenvalue of X_N for the deformed model (2.4).

3 Large deviation upper bound for tilted measures

For $\theta \geq 0$, $\beta = 1$ or 2, we define a tilted measure on $\mathcal{O}_N$ if $\beta = 1$ and $\mathcal{U}_N$ if $\beta = 2$ as follows

$$m_N^{\beta,\theta}(\mathrm{d}U) := \frac{I_N^\beta(\theta, A_N + UB_NU^*)}{I_N^\beta(\theta, A_N) I_N^\beta(\theta, B_N)} m_N^\beta(\mathrm{d}U).$$

It is easy to check that $m_N^{\beta,\theta}$ is a probability measure: indeed, for any U, we have that $I_N^\beta(\theta, A_N + UB_NU^*) \geq 0$ and $\mathbb{E}_{m_N^\beta}(I_N^\beta(\theta, A_N + UB_NU^*)) = I_N^\beta(\theta, A_N) I_N^\beta(\theta, B_N)$. For these tilted measures, we have the following weak large deviation upper bound:

Proposition 3.1. *Under Assumption 2.1, for $\beta = 1$ or 2, for any $\theta \geq 0$, for any $x < r(\mu_a \boxplus \mu_b)$,*

$$\lim_{\delta \downarrow 0} \limsup_{N \to +\infty} \frac{1}{N} \log m_N^{\beta,\theta}\left(\lambda_{\max}^N \in [x - \delta, x + \delta]\right) = -\infty, \tag{3.1}$$

and for any $x \geq r(\mu_a \boxplus \mu_b)$ such that

$$G_{\mu_a \boxplus \mu_b}(x) \leq \min\left(G_{\mu_a}(\rho_a), G_{\mu_b}(\rho_b)\right), \tag{3.2}$$

we have,

$$\lim_{\delta \downarrow 0} \limsup_{N \to +\infty} \frac{1}{N} \log m_N^{\beta,\theta}\left(\lambda_{\max}^N \in [x - \delta, x + \delta]\right) \leq -\left[I^\beta(x) - I^\beta(\theta, x)\right]. \tag{3.3}$$

Remark 3.2. Applying this proposition with $\theta = 0$ gives Proposition 2.3.

As we will see in Section 5, establishing an upper bound for any $\theta \geq 0$ will be useful in the proof of Proposition 2.4. To prove Proposition 3.1, and in particular its first statement, we will need to check that, under $m_N^{\beta,\theta}$ the empirical spectral distribution

$$\hat{\mu}_N := \frac{1}{N} \sum_{j=1}^{N} \delta_{\lambda_j^{(H_N(U))}}$$

of $H_N(U) = A_N + UB_NU^*$ concentrates around the deterministic probability measure given by its expectation $\mathbb{E}_{m_N^\beta} \hat{\mu}_N$ much faster than e^{-N}. More precisely, we equip the set $\mathcal{P}(\mathbb{R})$ of probability measures on $\mathbb{R}$ with the bounded Lipschitz distance d: for any Lipschitz function $f : \mathbb{R} \to \mathbb{R}$, we define $\|f\|_\infty := \sup_{x\in\mathbb{R}} |f(x)|$ and $\|f\|_{\mathrm{Lip}} := \sup_{x\neq y} \frac{|f(x)-f(y)|}{|x-y|}$, then for any μ and ν in $\mathcal{P}(\mathbb{R})$,

$$\mathrm{d}(\mu, \nu) := \sup_{\substack{\|f\|_\infty \leq 1 \\ \|f\|_{\mathrm{Lip}} \leq 1}} \left(\int f \mathrm{d}\mu - \int f \mathrm{d}\nu \right).$$

We then have the following concentration result:

Lemma 3.3. *Under Assumption (H_{bulk}), for $\beta = 1$ or 2 and any $\theta \geq 0$,*

$$\limsup_{N\to\infty} \frac{1}{N} \log m_N^{\beta,\theta} \left(\mathrm{d}(\hat{\mu}_N, \mathbb{E}_{m_N^\beta} \hat{\mu}_N) > N^{-1/4} \right) = -\infty.$$

Proof. Let $\beta = 1$ or 2 and $\theta \geq 0$ be fixed. Observe that for any Hermitian matrix M_N bounded by K in operator norm we have

$$e^{-\theta KN} \leq I_N^\beta(\theta, A_N) \leq e^{\theta KN}.$$

As a consequence, for any Borel subset A of $\mathcal{O}_N$ if $\beta = 1$ and $\mathcal{U}_N$ if $\beta = 2$, we have:

$$\begin{aligned} m_N^{\beta,\theta}(A) &= \frac{1}{I_N^\beta(\theta, A_N) I_N^\beta(\theta, B_N)} \int_A I_N^\beta(\theta, A_N + UB_NU^*) m_N^\beta(\mathrm{d}U) \\ &\leq \mathrm{e}^{2N\theta K} \, m_N^\beta(A), \end{aligned}$$

with $K := \sup_{N\geq 1}(\|A_N\| + \|B_N\|)$, which is assumed to be finite. Therefore it is enough to prove Lemma 3.3 for $\theta = 0$, that is

$$\limsup_{N\to\infty} \frac{1}{N} \log m_N^\beta \left(\mathrm{d}(\hat{\mu}_N, \mathbb{E}_{m_N^\beta} \hat{\mu}_N) > N^{-1/4} \right) = -\infty.$$

For $\beta = 2$, Theorem 3.8 in [24] states that there exist $c, C > 0$ such that

$$m_N^2 \left(\mathrm{d}(\hat{\mu}_N, \mathbb{E}_{m_N^2} \hat{\mu}_N) > N^{-1/4} \right) \leq C\mathrm{e}^{-cN^{3/2}}, \tag{3.4}$$

from which the lemma follows. A careful reading of [24] shows that the exact same result as (3.4) also holds for $\beta = 1$. □

Before proving Proposition 3.1, we will recall some results about the convergence and the continuity of spherical integrals.

Lemma 3.4 (Proposition 2.1 in [23] and Theorem 6 in [20])**.** *For any $\theta \geq 0$, there exists a continuous function g_θ with $g_\theta(0) = 0$ such that for any $\delta > 0$, if the sequences $(G_N)_{N\geq 1}$ and $(G'_N)_{N\geq 1}$ are such that $\sup_N(\|G_N\| + \|G'_N\|) < \infty$, and for N large enough, $\mathrm{d}(\hat{\mu}_{G_N}, \hat{\mu}_{G'_N}) \leq N^{-1/4}$ and $|\lambda_1(G_N) - \lambda_1(G'_N)| \leq \delta$, then,*

$$\left| \frac{1}{N} \log I_N^\beta(\theta, G_N) - \frac{1}{N} \log I_N^\beta(\theta, G'_N) \right| \leq g_\theta(\delta).$$

If moreover, $\hat{\mu}_{G_N}$ converges weakly, as N goes to infinity to μ and $\lambda_1(G_N)$ converges to ρ, then

$$\frac{1}{N}\log I_N^{\beta}(\theta, G_N) \xrightarrow[N\to\infty]{} J_\mu(\theta,\rho).$$

We can now prove Proposition 3.1. In the sequel, we will denote by $\nu_N^{\beta} := \mathbb{E}_{m_N^{\beta}}\hat{\mu}_N$.

Proof of Proposition 3.1. The first claim (3.1) is a direct consequence of the previous lemma. Indeed, let $x < \mathsf{r}(\mu_a \boxplus \mu_b)$ and $\delta_0 := \frac{\mathsf{r}(\mu_a\boxplus\mu_b)-x}{2}$. Then, for any $\delta \le \delta_0$, there exists $\varepsilon(\delta) > 0$,

$$\{\lambda_{\max}^N \in [x-\delta, x+\delta]\} \subset \{\mathrm{d}(\hat{\mu}_N, \mu_a \boxplus \mu_b) > \varepsilon(\delta)\}. \tag{3.5}$$

Using Corollary 5.4.11 for $\beta = 2$ and Exercise 5.4.18 for $\beta = 1$ in [1], we know that ν_N^{β} converges weakly to $\mu_a \boxplus \mu_b$ as N goes to infinity. As the distance d metrizes the weak convergence, for N large enough,

$$\{\lambda_{\max}^N \in [x-\delta, x+\delta]\} \subset \{\mathrm{d}(\hat{\mu}_N, \nu_N^{\beta}) > \varepsilon(\delta)/2\}$$

so that, by Lemma 3.3, for any $\delta \le \delta_0$,

$$\limsup_{N\to\infty} \frac{1}{N}\log m_N^{\beta,\theta}\left(\lambda_{\max}^N \in [x-\delta, x+\delta]\right) = -\infty.$$

We now prove (3.3). Let $\delta > 0$ and $x \ge \mathsf{r}(\mu_a \boxplus \mu_b)$ be fixed and define the following event:

$$\mathsf{E}_{N,\delta}^x := \left\{\lambda_{\max}^N \in [x-\delta, x+\delta], \mathrm{d}(\hat{\mu}_N, \nu_N^{\beta}) \le N^{-1/4}\right\}. \tag{3.6}$$

Then we have,

$$m_N^{\beta,\theta}\left(\lambda_{\max}^N \in [x-\delta, x+\delta]\right) \le m_N^{\beta,\theta}(\mathsf{E}_{N,\delta}^x) + m_N^{\beta,\theta}(\mathrm{d}(\hat{\mu}_N, \nu_N^{\beta}) > N^{-1/4}).$$

By Lemma 3.3, it is therefore enough to show that

$$\lim_{\delta\downarrow 0}\limsup_{N\to\infty} \frac{1}{N}\log m_N^{\beta,\theta}\left(\mathsf{E}_{N,\delta}^x\right) \le -\left[I^{\beta}(x) - I^{\beta}(\theta, x)\right].$$

To lighten a bit the notations we write A, B and H for A_N, B_N and $H_N = A_N + UB_NU^*$ respectively. For any $\theta, \theta' \ge 0$, we have

$$\begin{aligned}
m_N^{\beta,\theta}(\mathsf{E}_{N,\delta}^x) &= \frac{1}{I_N^{\beta}(\theta,A)I_N^{\beta}(\theta,B)}\mathbb{E}_{m_N^{\beta}}\left(1_{\mathsf{E}_{N,\delta}^x} I_N^{\beta}(\theta,H)\frac{I_N^{\beta}(\theta',H)}{I_N^{\beta}(\theta',H)}\right)\\
&\le \frac{\mathbb{E}_{m_N^{\beta}}(I_N^{\beta}(\theta',H))}{I_N^{\beta}(\theta,A)I_N^{\beta}(\theta,B)} \sup_{U\in\mathsf{E}_{N,\delta}^x} \frac{I_N^{\beta}(\theta, A+UBU^*)}{I_N^{\beta}(\theta', A+UBU^*)}\\
&= \frac{I_N^{\beta}(\theta',A)I_N^{\beta}(\theta',B)}{I_N^{\beta}(\theta,A)I_N^{\beta}(\theta,B)} \sup_{U\in\mathsf{E}_{N,\delta}^x} \frac{I_N^{\beta}(\theta, A+UBU^*)}{I_N^{\beta}(\theta', A+UBU^*)}
\end{aligned}$$

We now have to estimate $\sup_{U\in\mathsf{E}_{N,\delta}^x} I_N^{\beta}(\theta, A+UBU^*)$ and $\inf_{U\in\mathsf{E}_{N,\delta}^x} I_N^{\beta}(\theta', A+UBU^*)$ respectively. We detail the first term, the second being similar. According to [7] (see also Section 4.1.2 in [13]), if $x \ge \mathsf{r}(\mu_a \boxplus \mu_b)$ satisfies $G_{\mu_a\boxplus\mu_b}(x) \le \min\left(G_{\mu_a}(\rho_a), G_{\mu_b}(\rho_b)\right)$, then m_N^{β} almost surely, $\limsup_{N\to\infty} \lambda_{\max}^N \le x$ and therefore $\limsup_{N\to\infty} \mathsf{r}(\nu_N^{\beta}) \le x$. It is therefore possible to build a sequence $(G_N)_{N\ge 1}$ of deterministic matrices such that $\sup_{N\ge 1}\|G_N\| < \infty$ and for any $N \ge 1$, $\lambda_1^{(G_N)}$ converges to x and $\mathrm{d}(\hat{\mu}_{G_N}, \nu_N^{\beta}) \le N^{-1/4}$. With the notations of Lemma 3.4, for N large enough and for any $U \in \mathsf{E}_{N,\delta}^x$,

$$\left|\frac{1}{N}\log I_N^{\beta}(\theta, A+UBU^*) - \frac{1}{N}\log I_N^{\beta}(\theta, G_N)\right| \le g_\theta(\delta).$$

Therefore,

$$\begin{aligned}\limsup_{N\to\infty}\frac{1}{N}\log m_N^{\beta;\theta}(\mathsf{E}^x_{N,\delta}) &\le \lim_{N\to\infty}(J_N^\beta(\theta',A)+J_N^\beta(\theta',B)-J_N^\beta(\theta,A)-J_N^\beta(\theta,B))\\ &\quad+\lim_{N\to\infty}(J_N^\beta(\theta,G_N)-J_N^\beta(\theta',G_N))+g_\theta(\delta)+g_{\theta'}(\delta),\\ &\le -(I^\beta(\theta',x)-I^\beta(\theta,x))+g_\theta(\delta)+g_{\theta'}(\delta),\end{aligned}$$

where at the last line, we have used the second part of Lemma 3.4. Letting δ going to zero and then optimizing over $\theta'\ge 0$, we get the required upper bound. □

4 Properties of the rate function I^β

We now check the properties of the rate function I^β defined in (2.1).

Proof of Lemma 2.2. An ingredient for the proof is the following: for any compactly supported μ, for any $\theta\ge 0$ and $\rho\ge \mathsf{r}(\mu)$ such that $\theta\le G_\mu(\rho)$, we have

$$\rho-\frac{1}{\theta}\le R_\mu(\theta)\le \rho-\frac{1}{G_\mu(\rho)}. \tag{4.1}$$

Indeed, as K_μ is a decreasing function, we have $R_\mu(\theta)=K_\mu(\theta)-\frac{1}{\theta}\ge\rho-\frac{1}{\theta}$. On the other hand, the limit of $R_\mu(\theta)$ as θ grows to $G_\mu(\rho)$ is $\rho-\frac{1}{G_\mu(\rho)}$. As R_μ is nondecreasing, we get the upper bound. Moreover, it is easy to check that, for any $x\ge 0$, there exist $C,C'\in\mathbb{R}$ (depending on μ and x but not on θ) such that, for θ large enough, we have

$$\theta x-\frac{\beta}{2}\log\theta+C\le J_\mu^\beta(\theta,x)\le\theta x+C',$$

so that, for any $x\ge 0$, there exists $c,c'\in\mathbb{R}$ such that, for θ large enough,

$$\theta(x-\rho_a-\rho_b)-\frac{\beta}{2}\log\theta+c\le I^\beta(\theta,x)\le\theta(x-\rho_a-\rho_b)+\beta\log\theta+c'.$$

If $x>\rho_a+\rho_b$, letting θ grow to infinity, we obtain that $I^\beta(x)=+\infty$.

If $\theta\ge 0$ is small enough,

$$I^\beta(\theta,x)=\frac{\beta}{2}\int_0^{\frac{2\theta}{\beta}}(R_{\mu_a\boxplus\mu_b}(u)-R_{\mu_a}(u)-R_{\mu_b}(u))\mathrm{d}u=0,$$

by the properties of the R-transform. The function I^β is therefore nonnegative. If we denote by g the lower semi-continuous function which is equal to $-\infty$ on $[\mathsf{r}(\mu_a\boxplus\mu_b),+\infty)$ and $+\infty$ outside, then $I^\beta=\sup(g,\sup_\theta I^\beta(\theta,\cdot))$ is lower semi-continuous as a supremum of lower semi-continuous functions. As it is infinite outside the interval $[\mathsf{r}(\mu_a\boxplus\mu_b),\rho_a+\rho_b]$, it is a good rate function. □

We will now turn to the proof of the lower bound of our large deviation principle, stated in Proposition 2.4. To complete its proof, we will need to further study the properties of the function I^β. First, let us remark that the cases when μ_a is a Dirac mass at ρ_a (or μ_b is a Dirac mass at ρ_b) are not very interesting. In this case, the free convolution $\mu_a\boxplus\mu_b$ is just a shift of μ_b by ρ_a (or respectively of μ_a by ρ_b) and $\lambda^N_{\max}$ converges with probability one to $\rho_a+\rho_b$. Hence, the large deviations have an infinite rate function in the scale N except at $\rho_a+\rho_b$ where it vanishes.

Consequently, in the sequel, one can assume without loss of generality that

Assumption 4.1. *μ_a is not a Dirac mass at ρ_a and μ_b is not a Dirac mass at ρ_b.*

We have the following:

Lemma 4.2. *Under Assumptions 2.1 and 4.1, for any* $\mathsf{r}(\mu_a \boxplus \mu_b) \le x < \rho_a + \rho_b$ *such that*

$$G_{\mu_a \boxplus \mu_b}(x) \le \min(G_{\mu_a}(\rho_a), G_{\mu_b}(\rho_b)),$$

then, for $\beta = 1$ *or* 2*, there exists a unique* $\theta \ge 0$ *such that*

$$I^\beta(\theta, x) = \sup_{\theta' \ge 0} I^\beta(\theta', x).$$

We denote by $\theta_x^\beta := \mathrm{argmax}_{\theta \ge 0} I^\beta(\theta, x)$*. For any* $\mathsf{r}(\mu_a \boxplus \mu_b) \le y \le \rho_a + \rho_b$ *such that* $x \neq y$*,*

$$\sup_{\theta \ge 0} I^\beta(\theta, y) > I^\beta(\theta_x^\beta, y).$$

Proof of Lemma 4.2. Let $\mathsf{r}(\mu_a \boxplus \mu_b) \le x < \rho_a + \rho_b$ such that

$$G_{\mu_a \boxplus \mu_b}(x) \le \min(G_{\mu_a}(\rho_a), G_{\mu_b}(\rho_b)).$$

The first remark is that if $G_{\mu_a}(\rho_a)$ and $G_{\mu_b}(\rho_b)$ are infinite, then $\mathsf{r}(\mu_a \boxplus \mu_b) = \rho_a + \rho_b$ and there is nothing to check since $[r(\mu_a \boxplus \mu_b), \rho_a + \rho_b)$ is empty. Indeed, if $G_{\mu_a}(\rho_a) = G_{\mu_b}(\rho_b) = \infty$, we see by the inequalities (4.1), that

$$\lim_{x \to \infty} R_{\mu_a}(x) = \rho_a \quad \text{and} \quad \lim_{x \to \infty} R_{\mu_b}(x) = \rho_b,$$

so that

$$\lim_{x \to \infty} K_{\mu_a \boxplus \mu_b}(x) = \rho_a + \rho_b \quad \text{and} \quad \lim_{x \to \rho_a + \rho_b} G_{\mu_a \boxplus \mu_b}(x) = \infty,$$

leading to $\mathsf{r}(\mu_a \boxplus \mu_b) \ge \rho_a + \rho_b$. By symmetry of the problem, without loss of generality, one can now assume that $G_{\mu_a}(\rho_a) \le G_{\mu_b}(\rho_b)$ and $G_{\mu_a}(\rho_a) < \infty$.

With the function I^β defined in (2.1), if we denote by I_x^β the function $\theta \mapsto I^\beta(\theta, x)$, then there exist some constants C_1, C_2 and C_3 (that may depend on μ_a, ρ_a, μ_b, ρ_b and x but not on θ) such that

$$I_x^\beta(\theta) = \begin{cases} 0, & \text{if } 0 \le \frac{2\theta}{\beta} \le G_{\mu_a \boxplus \mu_b}(x), \\ \theta x - \frac{\beta}{2} \log \theta - \frac{\beta}{2} \int_0^{\frac{2\theta}{\beta}} (R_{\mu_a} + R_{\mu_b})(u) \mathrm{d}u + C_1, & \text{if } G_{\mu_a \boxplus \mu_b}(x) \le \frac{2\theta}{\beta} \le G_{\mu_a}(\rho_a), \\ \theta(x - \rho_a) - \frac{\beta}{2} \int_0^{\frac{2\theta}{\beta}} R_{\mu_b}(u) \mathrm{d}u + C_2, & \text{if } G_{\mu_a}(\rho_a) \le \frac{2\theta}{\beta} \le G_{\mu_b}(\rho_b), \\ \theta(x - \rho_a - \rho_b) + \frac{\beta}{2} \log \theta + C_3, & \text{if } \frac{2\theta}{\beta} \ge G_{\mu_b}(\rho_b), \end{cases}$$

where the last line does not occur if $G_{\mu_b}(\rho_b) = \infty$. In the computation, we have used the well known fact that $R_{\mu_a \boxplus \mu_b} = R_{\mu_a} + R_{\mu_b}$ when the three functions are well defined. Therefore, one can check that the function I_x^β is continuously differentiable and its derivative is given by:

$$(I_x^\beta)'(\theta) = \begin{cases} 0, & \text{if } 0 \le \frac{2\theta}{\beta} \le G_{\mu_a \boxplus \mu_b}(x), \\ x - K_{\mu_a \boxplus \mu_b}\left(\frac{2\theta}{\beta}\right), & \text{if } G_{\mu_a \boxplus \mu_b}(x) \le \frac{2\theta}{\beta} \le G_{\mu_a}(\rho_a), \\ x - \rho_a - R_{\mu_b}\left(\frac{2\theta}{\beta}\right), & \text{if } G_{\mu_a}(\rho_a) \le \frac{2\theta}{\beta} \le G_{\mu_b}(\rho_b), \\ x - \rho_a - \rho_b + \frac{\beta}{2\theta}, & \text{if } \frac{2\theta}{\beta} \ge G_{\mu_b}(\rho_b). \end{cases}$$

We now set $\alpha_x := \frac{1}{\rho_a + \rho_b - x}$. We claim that

$$\alpha_x \ge G_{\mu_a}(\rho_a).$$

Indeed, K_{μ_b} is well defined on the interval $(0, G_{\mu_b}(\rho_b))$, so that $K_{\mu_b}(G_{\mu_a}(\rho_a))$ and therefore $K_{\mu_a \boxplus \mu_b}(G_{\mu_a}(\rho_a))$ are well defined. As $K_{\mu_a \boxplus \mu_b}$ is a decreasing function and

$$G_{\mu_a \boxplus \mu_b}(x) \leq G_{\mu_a}(\rho_a),$$

we have

$$x \geq K_{\mu_a \boxplus \mu_b}(G_{\mu_a}(\rho_a)) = K_{\mu_a}(G_{\mu_a}(\rho_a)) + K_{\mu_b}(G_{\mu_a}(\rho_a)) - \frac{1}{G_{\mu_a}(\rho_a)}$$

As K_{μ_b} is also a decreasing function, this yields:

$$x \geq K_{\mu_a}(G_{\mu_a}(\rho_a)) + K_{\mu_b}(G_{\mu_b}(\rho_b)) - \frac{1}{G_{\mu_a}(\rho_a)} = \rho_a + \rho_b - \frac{1}{G_{\mu_a}(\rho_a)},$$

which is equivalent to $\alpha_x \geq G_{\mu_a}(\rho_a)$. There are therefore two cases to consider and we claim that:

Case 1: If $G_{\mu_a}(\rho_a) \leq \alpha_x < G_{\mu_b}(\rho_b)$, then I_x^β reaches its maximum at

$$\theta_x^\beta := \frac{\beta}{2} Q_{\mu_b}(x - \rho_a),$$

where Q_{μ_b} is the inverse of R_{μ_b} as defined in Section 2;

Case 2: If $\alpha_x \geq G_{\mu_b}(\rho_b)$, then I_x^β reaches its maximum at $\theta_x^\beta := \frac{\beta}{2}\alpha_x$.

Let us now prove this claim. On the interval $\left[0, \frac{\beta}{2} G_{\mu_a}(\rho_a)\right]$, the function $(I_x^\beta)'$ is nondecreasing and it vanishes at zero, it is therefore nonnegative so that I_x^β is nondecreasing on this interval. We have

$$(I_x^\beta)'\left(\frac{\beta}{2} G_{\mu_a}(\rho_a)\right) \geq 0 \quad \text{and} \quad (I_x^\beta)'\left(\frac{\beta}{2} G_{\mu_b}(\rho_b)\right) = -\frac{1}{\alpha_x} + \frac{1}{G_{\mu_b}(\rho_b)}.$$

Moreover, as R_{μ_b} is increasing, $(I_x^\beta)'$ is decreasing on the interval $\left[\frac{\beta}{2} G_{\mu_a}(\rho_a), \frac{\beta}{2} G_{\mu_b}(\rho_b)\right]$. We now distinguish the two cases.

In Case 1, $(I_x^\beta)'\left(\frac{\beta}{2} G_{\mu_b}(\rho_b)\right) < 0$, and therefore there exists

$$\theta_x^\beta \in \left[\frac{\beta}{2} G_{\mu_a}(\rho_a), \frac{\beta}{2} G_{\mu_b}(\rho_b)\right)$$

such that I_x^β is increasing on $\left[\frac{\beta}{2} G_{\mu_a}(\rho_a), \theta_x^\beta\right]$ and then decreasing. One can check that the point where $(I_x^\beta)'$ cancels is given by $\frac{\beta}{2} Q_{\mu_b}(x - \rho_a)$. Moreover, $(I_x^\beta)'$ is decreasing on $\left[\frac{\beta}{2} G_{\mu_b}(\rho_b), \infty\right)$ and negative at $\frac{\beta}{2} G_{\mu_b}(\rho_b)$ so it remains negative and I_x^β is decreasing on this interval. The first claim holds true.

In Case 2, $(I_x^\beta)'\left(\frac{\beta}{2} G_{\mu_b}(\rho_b)\right) \geq 0$, and therefore I_x^β is increasing on $\left[\frac{\beta}{2} G_{\mu_a}(\rho_a), \frac{\beta}{2} G_{\mu_b}(\rho_b)\right]$. But $(I_x^\beta)'$ is nonnegative at $\frac{\beta}{2} G_{\mu_b}(\rho_b)$, decreasing on $\left[\frac{\beta}{2} G_{\mu_b}(\rho_b), \infty\right)$ and converges to $x - \rho_a - \rho_b < 0$ as θ grows to ∞. Therefore, there exists $\theta_x^\beta \in \left(\frac{\beta}{2} G_{\mu_b}(\rho_b), \infty\right)$ such that I_x^β is increasing on $\left(\frac{\beta}{2} G_{\mu_b}(\rho_b), \theta_x^\beta\right]$ and then decreasing. One can check that the point where $(I_x^\beta)'$ cancels is given by $\frac{\beta}{2}\alpha_x$ and the second claim holds true. This concludes the proof of the uniqueness of θ.

Moreover, looking carefully at the definition of θ_x^β in Case 1 and Case 2, one can see that it is an increasing function of x. Indeed, it is increasing on the intervals $\{x \in$

$\mathbb{R}/\alpha_x \geq G_{\mu_b}(\rho_b)\}$ and $\{x \in \mathbb{R}/G_{\mu_a}(\rho_a) \leq \alpha_x < G_{\mu_b}(\rho_b)\}$ respectively and it is continuous at $x = G_{\mu_b}(\rho_b)$. As a consequence, for $x \neq y$ such that $\mathsf{r}(\mu_a \boxplus \mu_b) \leq x, y < \rho_a + \rho_b$, $\theta_x^\beta \neq \theta_y^\beta$ and therefore $\sup_{\theta \geq 0} I^\beta(\theta, y) > I^\beta(\theta_x^\beta, y)$.

We now have to deal with the case when $y = \rho_a + \rho_b$, that is to show that:

$$\sup_{\theta \geq 0} I^\beta(\theta, \rho_a + \rho_b) > I^\beta(\theta_x^\beta, \rho_a + \rho_b). \tag{4.2}$$

If $G_{\mu_b}(\rho_b)$ is finite, for $\theta > \frac{\beta}{2} G_{\mu_b}(\rho_b)$,

$$I^\beta(\theta, \rho_a + \rho_b) = \frac{\beta}{2} \log \theta + C_3$$

and therefore the supremum is infinite and (4.2) holds. Assume now that $G_{\mu_b}(\rho_b) = \infty$. As $\mu_b \neq \delta_{\rho_b}$, then, there exists $\alpha \in (0, 1]$ and M finite such that, for any $x \geq \rho_b$,

$$G_{\mu_b}(x) \leq \frac{1 - \alpha}{x - \rho_b} + M.$$

From there, we get that, for any $u > G_{\mu_a}(\rho_a) \vee \frac{2M}{\alpha}$,

$$u \leq \frac{1 - \alpha}{K_{\mu_b}(u) - \rho_b} + M \quad \text{so that} \quad R_{\mu_b}(u) \leq \rho_b + \frac{1 - \alpha}{(1 - \frac{\alpha}{2})u} - \frac{1}{u} \leq \rho_b - \frac{\alpha}{2u}.$$

Therefore, there exist $c, c' \in \mathbb{R}$, such that for any $\theta \geq G_{\mu_a}(\rho_a) \vee \frac{2M}{\alpha}$,

$$I^\beta(\theta, \rho_a + \rho_b) \geq \theta \rho_b - \frac{\beta}{2} \int_{\frac{2M}{\alpha}}^{\frac{2\theta}{\beta}} \left(\rho_b - \frac{\alpha}{2u}\right) du + c = \frac{\beta\alpha}{4} \log \theta + c'$$

so that, letting θ grow to infinity, we get again that $I^\beta(\rho_a + \rho_b) = \infty$ and (4.2) holds. This concludes the proof of Lemma 4.2. □

5 Large deviation lower bound

The goal of this section is to show Proposition 2.4. A classical strategy to get a large deviation lower bound is to tilt the measure in such a way that the rare event $\{\lambda_{\max}^N \in [x - \delta, x + \delta]\}$ becomes typical under the tilted measure. We now check that it is possible to make such a tilt[1]:

Lemma 5.1. *Under Assumptions 2.1, for any $x \in [\mathsf{r}(\mu_a \boxplus \mu_b), \rho_a + \rho_b)$ such that*

$$G_{\mu_a \boxplus \mu_b}(x) \leq \min(G_{\mu_a}(\rho_a), G_{\mu_b}(\rho_b)),$$

for $\beta = 1$ or 2, we have

$$\lim_{\delta \downarrow 0} \liminf_{N \to \infty} \frac{1}{N} \log m_N^{\beta, \theta_x^\beta} \left(\mathsf{E}_{N,\delta}^x\right) \geq 0,$$

where $\mathsf{E}_{N,\delta}^x$ was defined in (3.6) and θ_x^β in Lemma 4.2.

Proof of Lemma 5.1. Let $\beta = 1$ or 2. The first remark is that, almost surely, $|\lambda_{\max}^N| \leq K$, where we recall that $K := \sup_{N \geq 1}(\|A_N\| + \|B_N\|)$.

Let $\mathsf{r}(\mu_a \boxplus \mu_b) \leq x < \rho_a + \rho_b$ be fixed. If we denote by

$$L_x^\beta(y) := \begin{cases} \sup_{\theta \geq 0} I^\beta(\theta, y) - I^\beta(\theta_x^\beta, y), & \text{if } \mathsf{r}(\mu_a \boxplus \mu_b) \leq y \leq \rho_a + \rho_b, \\ \infty, & \text{otherwise,} \end{cases}$$

[1] As for Lemma 4.2, we want to mention that Lemma 5.1 holds without Assumption 4.1, that we add to simplify the proof.

we know from Proposition 3.1 that, for any $y \in \mathbb{R}$,

$$\lim_{\delta \downarrow 0} \limsup_{N\to\infty} \frac{1}{N} \log m_N^{\beta,\theta_x^\beta} \left(\lambda_{\max}^N \in [y-\delta, y+\delta]\right) \le -L_x^\beta(y). \tag{5.1}$$

Let $\delta > 0$ be fixed. We denote by F_δ the compact set $[-K,K] \setminus (x-\delta, x+\delta)$. For any $\eta > 0, y \in F_\delta$, we also denote by

$$L_{x,\eta}^\delta(y) := \min\left(L_x^\beta(y) - \eta, \frac{1}{\eta}\right).$$

From (5.1), for any $y \in F_\delta$, there exists $\gamma_{y,\eta} > 0$ such that

$$\limsup_{N\to\infty} \frac{1}{N} \log m_N^{\beta,\theta_x^\beta} \left(\lambda_{\max}^N \in [y-\gamma_{y,\eta}, y+\gamma_{y,\eta}]\right) \le -L_{x,\eta}^\delta(y).$$

As F_δ is a compact set, one can extract from the family $([y-\gamma_{y,\eta}, y+\gamma_{y,\eta}])_{y\in F_\delta}$ a finite covering $F_\delta = \cup_{i=1}^r [y_i - \gamma_{y_i,\eta}, y_i + \gamma_{y_i,\eta}]$. From there, we get that

$$\limsup_{N\to\infty} \frac{1}{N} \log m_N^{\beta,\theta_x^\beta} \left(\lambda_{\max}^N \in F\right) \le \max_{1\le i\le r} -L_{x,\eta}^\beta(y_i) \le -\inf_{y\in F} L_{x,\eta}^\beta(y).$$

Letting η going to zero, we deduce that

$$\limsup_{N\to\infty} \frac{1}{N} \log m_N^{\beta,\theta_x^\beta} \left(\lambda_{\max}^N \in F\right) \le -\inf_{y\in F_\delta} L_x^\beta(y).$$

By Lemma 4.2, we know that L_x^β is nonnegative and vanishes only at x, so that, $\inf_{y\in F_\delta} L_x^\delta(y) > 0$. Therefore, we deduce that, for N large enough,

$$m_N^{\beta,\theta_x^\beta} \left(\lambda_{\max}^N \in [x-\delta, x+\delta]\right) \ge \frac{3}{4}.$$

But, in virtue of Lemma 3.3, for N large enough, we also have

$$m_N^{\beta,\theta_x^\beta} \left(\mathrm{d}(\hat{\mu}_N, \nu_N^\beta) \le N^{-1/4}\right) \ge \frac{3}{4}$$

so that

$$m_N^{\beta,\theta_x^\beta} \left(\mathsf{E}_{N,\delta}^x\right) \ge \frac{1}{2},$$

and Lemma 5.1 follows. □

From there, one can easily get the large deviation lower bound.

Proof of Proposition 2.4. As mentioned in Section 4, without loss of generality, one can assume Assumption 4.1. Let $\beta = 1$ or 2 and $x \in \mathbb{R}$ be fixed. If $x > \rho_a + \rho_b$ or $x < \mathsf{r}(\mu_a \boxplus \mu_b)$, Lemma 2.2 gives that $I^\beta(x) = \infty$, so that the lower bound obviously holds. Moreover, as we have seen at the end of the proof of Lemma 4.2, as μ_b is not a Dirac mass at ρ_b, then $I^\beta(\rho_a + \rho_b) = \infty$ and the lower bound also holds for $x = \rho_a + \rho_b$.

Let us now assume that $\mathsf{r}(\mu_a \boxplus \mu_b) \le x < \rho_a + \rho_b$ and let θ_x^β be the corresponding shift defined in Lemma 4.2. Then, with $\mathsf{E}_{N,\delta}^x$ defined in (3.6) and recalling that $A = A_N, B = B_N$ and $H = A + UBU^*$, we have:

$$\begin{aligned} m_N^\beta(\lambda_{\max}^N \in [x-\delta, x+\delta]) \ge m_N^\beta(\mathsf{E}_{N,\delta}^x) &= \mathbb{E}_{m_N^\beta}\left(1_{\mathsf{E}_{N,\delta}^x} \frac{I_N^\beta(\theta_x^\beta, H)}{I_N^\beta(\theta_x^\beta, H)}\right) \\ &\ge \inf_{U\in \mathsf{E}_{N,\delta}^x} \frac{1}{I_N^\beta(\theta_x^\beta, A+UBU^*)} \\ &\qquad \times I_N^\beta(\theta_x^\beta, A) I_N^\beta(\theta_x^\beta, B) m_N^{\beta,\theta_x^\beta}(\mathsf{E}_{N,\delta}^x) \end{aligned}$$

so that, using again Lemma 3.4, we get:

$$\liminf_{N\to\infty} \frac{1}{N} \log m_N^\beta \left(\lambda_{\max}^N \in [x-\delta, x+\delta]\right) \geq -I^\beta(\theta_x^\beta, x) - g_{\theta_x^\beta}(\delta) + \liminf_{N\to\infty} \frac{1}{N} \log m_N^{\beta,\theta_x^\beta} \left(\mathsf{E}_{N,\delta}^x\right).$$

Letting δ going to zero and using Lemma 5.1, we get that

$$\lim_{\delta\downarrow 0} \liminf_{N\to\infty} \frac{1}{N} m_N^\beta(\lambda_{\max}^N \in [x-\delta, x+\delta]) \geq -I^\beta(\theta_x^\beta, x) \geq -I^\beta(x).$$

This concludes the proof. □

6 Proof of the main theorem and its corollary

Proof of Theorem 2.5. Assume that Assumption 2.1 and the condition (NoOut) are satisfied. Without loss of generality, one can add Assumption 4.1. As already stated in the proof of Lemma 5.1, almost surely, $|\lambda_{\max}^N| \leq K$, where we recall that $K := \sup_{N\geq 1}(\|A_N\| + \|B_N\|)$.

In particular,

$$\limsup_{N\to\infty} \frac{1}{N} \log m_N^\beta \left(\lambda_{\max}^N \in [-K, K]^c\right) = -\infty.$$

Using e.g. Theorem D.4(a) and Corollary D.6 in [1], it is enough to show that, for any $x \in \mathbb{R}$,

$$\lim_{\delta\downarrow 0} \limsup_{N\to\infty} \frac{1}{N} \log m_N^\beta \left(\lambda_{\max}^N \in [x-\delta, x+\delta]\right) = \lim_{\delta\downarrow 0} \liminf_{N\to\infty} \frac{1}{N} \log m_N^\beta \left(\lambda_{\max}^N \in [x-\delta, x+\delta]\right) = -I^\beta(x).$$

The upper bound is nothing but Proposition 2.3, obtained from Proposition 3.1 for $\theta = 0$ and the lower bound is given by Proposition 2.4. □

We now prove Corollary 2.6. Our goal is to show that if A_N and B_N have no outliers, then the condition (NoOut) is automatically satisfied. Indeed, if A_N and B_N have no outliers, it means that their respective largest eigenvalues converge to the edge of the support of the limiting measure, that is to say $\rho_a = \mathsf{r}(\mu_a)$ and $\rho_b = \mathsf{r}(\mu_b)$. Therefore, Corollary 2.6 is a direct consequence of the following lemma:

Lemma 6.1. *For any probability measures μ and ν compactly supported on $\mathbb{R}$, we have*

$$G_{\mu\boxplus\nu}(\mathsf{r}(\mu\boxplus\nu)) \leq \min(G_\mu(\mathsf{r}(\mu)), G_\nu(\mathsf{r}(\nu))).$$

Proof. If one of the measures μ or ν is a single point mass, the additive free convolution is just a translation and we have equality. We now assume that none of them is a single point mass. In general, we know (see e.g. [6]) that there exists a function ω, called the subordination function, which is analytic on $\mathbb{C}^+ := \{z \in \mathbb{C}, \mathfrak{Im}\, z > 0\}$ such that, for all $z \in \mathbb{C}^+$,

$$G_{\mu\boxplus\nu}(z) = G_\mu(\omega(z)).$$

This gives immediately that for any $z \in \mathbb{C}^+$,

$$\mathfrak{Im}\, G_{\mu\boxplus\nu}(z) = -\mathfrak{Im}\,\omega(z). \int \frac{\mathrm{d}\mu(t)}{|t-\omega(z)|^2}. \tag{6.1}$$

By [5, Theorem 2.3], as μ or ν are not a single point mass, $G_{\mu\boxplus\nu}$ can be continuously extended to $\mathbb{C}^+ \cup \mathbb{R}$ with values in $\overline{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$. Moreover, as μ and ν are compactly supported, by [6, Theorem 3.3(3)], ω can also be continuously extended to $\mathbb{C}^+ \cup \mathbb{R}$. Let u be a real number in the interval $(\mathsf{r}(\mu \boxplus \nu), \infty)$. Then $\lim_{v \downarrow 0} \int \frac{\mathrm{d}\mu(t)}{|t-\omega(u+\mathrm{i}v)|^2} > 0$ and $\lim_{v\downarrow 0} \mathfrak{Im}\, G_{\mu\boxplus\nu}(u+\mathrm{i}v) = \mathfrak{Im}\, G_{\mu\boxplus\nu}(u) = 0$, so that, using (6.1), we get $\mathfrak{Im}\,\omega(u) = \lim_{v\downarrow 0} \mathfrak{Im}\, G_{\mu\boxplus\nu}(u+\mathrm{i}v) = 0$. Therefore, ω restricted to the interval $(\mathsf{r}(\mu \boxplus \nu), \infty)$ takes values in $\mathbb{R} \cup \{\infty\}$. Moreover, as ω is continuous and $\omega(z)$ goes to ∞ as z goes to ∞, we get that $\omega((\mathsf{r}(\mu \boxplus \nu), \infty))$ is an interval of the form (a, ∞).

We now want to show by contradiction that $a \geq \mathsf{r}(\mu)$. Let us assume that $a < \mathsf{r}(\mu)$. For any $y > 0$, we have

$$-\int_a^{\mathsf{r}(\mu)} \mathfrak{Im}\, G_\mu(x+\mathrm{i}y)\mathrm{d}x = \int_a^{\mathsf{r}(\mu)} \mathrm{d}\mu(t) \left(\arctan\left(\frac{r(\mu)-t}{y}\right) - \arctan\left(\frac{a-t}{y}\right)\right).$$

As y decreases to zero, the right hand-side goes to $\frac{\pi}{2}(\mu([a, \mathsf{r}(\mu)) + \mu((a, \mathsf{r}(\mu)]) > 0$. On the other hand, for any $x \in (a, \mathsf{r}(\mu)) \subset \omega((\mathsf{r}(\mu \boxplus \nu), \infty))$, there exists $x' > \mathsf{r}(\mu \boxplus \nu)$, such that $x = \omega(x')$ and ω is holomorphic from a neighborhood of x' to a neighborhood of x. As

$$\lim_{\tilde{x}\to x} \mathfrak{Im}\, G_\mu(\tilde{x}) = \lim_{\tilde{x}'\to x'} \mathfrak{Im}\, G_\mu(\omega(\tilde{x}')) = \mathfrak{Im}\, G_{\mu\boxplus\nu}(x') = 0,$$

by dominated convergence, we get that the left hand-side goes to zero, as y decreases to zero. This leads to a contradiction and we deduce that $\omega((\mathsf{r}(\mu \boxplus \nu), \infty)) \subset (\mathsf{r}(\mu), \infty)$, so that

$$\omega(\mathsf{r}(\mu \boxplus \nu)) \geq \mathsf{r}(\mu).$$

As G_μ is decreasing on $(\mathsf{r}(\mu), \infty)$, this gives

$$G_{\mu\boxplus\nu}(\mathsf{r}(\mu \boxplus \nu)) = G_\mu(\omega(\mathsf{r}(\mu \boxplus \nu))) \leq G_\mu(\mathsf{r}(\mu))\,.$$

As μ and ν play symmetric roles, this concludes the proof of Lemma 6.1. □

A Study of the deformed model (2.4)

A.1 Large deviations for the smallest eigenvalue of H_N

In order to study the deviations of the largest eigenvalue of the deformed model below its expected value, we will need a counterpart of Theorem 2.5 for the smallest eigenvalue of H_N. We first state the counterpart of the condition (NoOut). For any compactly supported probability measure μ, we denote by $\mathsf{l}(\mu)$ the left edge of the support of μ. One can extend the definitions of G_μ, K_μ, R_μ and Q_μ given in Section 2: for any $\lambda < \mathsf{l}(\mu)$, $G_\mu(\lambda) := \int \frac{1}{\lambda - y}\mu(\mathrm{d}y)$; G_μ is decreasing from $(-\infty, \mathsf{l}(\mu))$ into $(G_\mu(\mathsf{l}(\mu)), 0)$ so we denote again by K_μ its inverse. For any $z \in (G_\mu(\mathsf{l}(\mu)), 0)$ we set $R_\mu(z) := K_\mu(z) - \frac{1}{z}$, which is increasing with inverse Q_μ. We then introduce the following assumption:

(NoDown) *The smallest eigenvalues $\lambda_N^{(A_N)}$ and $\lambda_N^{(B_N)}$ converge as N grows to infinity to ℓ_a and ℓ_b respectively and $G_{\mu_a\boxplus\mu_b}(\mathsf{l}(\mu_a \boxplus \mu_b)) \geq \max(G_{\mu_a}(\ell_a), G_{\mu_b}(\ell_b))$.*

As in Lemma 6.1, one can check that this condition is satisfied if A_N and B_N have no outliers, this time in the sense that $\ell_a = \mathsf{l}(\mu_a)$ and $\ell_b = \mathsf{l}(\mu_b)$. We now extend the definition of the rate function I^β introduced in (2.1). For $\beta = 1$ or 2, $\theta \leq 0$, μ a compactly supported probability measure and $\ell \leq \mathsf{l}(\mu)$, we define:

$$J_\mu^\beta(\theta, \ell) := \begin{cases} \frac{\beta}{2}\int_0^{\frac{2\theta}{\beta}} R_\mu(u)\mathrm{d}u, & \text{if } G_\mu(\ell) \leq \frac{2\theta}{\beta} \leq 0, \\ \theta\ell - \frac{\beta}{2}\log(-\theta) - \frac{\beta}{2}\int \log(y-\ell)\mu(\mathrm{d}y) + \frac{\beta}{2}\left(\log\frac{\beta}{2} - 1\right), & \text{if } \frac{2\theta}{\beta} < G_\mu(\ell). \end{cases}$$

For any $\theta \leq 0$ and $x \leq \mathsf{l}(\mu_a \boxplus \mu_b)$, we denote by

$$I^\beta(\theta, x) := J^\beta_{\mu_a\boxplus\mu_b}(\theta, x) - J^\beta_{\mu_a}(\theta, \ell_a) - J^\beta_{\mu_b}(\theta, \ell_b),$$

and

$$I^{\beta}_{\min}(x) := \begin{cases} \sup_{\theta \le 0} I^{\beta}(\theta, x), & \text{if } x \le \mathsf{l}(\mu_a \boxplus \mu_b), \\ \infty, & \text{otherwise.} \end{cases} \tag{A.1}$$

Applying Theorem 2.5 to $-A_N$ and $-B_N$, one can get a large deviation principle for the smallest eigenvalue $\lambda^N_{\min}$ of H_N :

Corollary A.1. *Under the assumptions (H_{bulk}) and (NoDown), for $\beta = 1$ or 2, the law of $\lambda^N_{\min}$ under m^{β}_N satisfies a large deviation principle in the scale N with good rate function $I^{\beta}_{\min}$.*

A.2 Asymptotic independence of the deviations of $\lambda^N_{\min}$ and $\lambda^N_{\max}$

Before going to the study of the deformed model itself, we will need the following proposition:

Proposition A.2. *Let $(M_N)_{N \ge 1}$ be a sequence of deterministic matrices such that $M := \sup_{N \ge 1} \|M_N\| < \infty$, and, as N goes to infinity, $\lambda^{(M_N)}_1$ and $\lambda^{(M_N)}_N$ converge respectively to ρ and ℓ and $\hat{\mu}_{M_N}$ converges weakly to μ. Let e_1 and e_2 be two random vectors uniformly distributed on the unit sphere of $\mathbb{R}^N$ if $\beta = 1$ (respectively of $\mathbb{C}^N$ if $\beta = 2$), orthogonal to each other. Let $\theta \ge 0$ and $\theta' \le 0$ be fixed. Then*

$$\lim_{N \to \infty} \frac{1}{N} \log \frac{\mathbb{E}\left(\mathrm{e}^{N\theta \langle e_1, M_N e_1 \rangle + N\theta' \langle e_2, M_N e_2 \rangle} \right)}{I_N(\theta, M_N) I_N(\theta', M_N)} = 0. \tag{A.2}$$

In other words, when θ and θ' are of opposite sign, the rank two spherical integral asymptotically factorizes in the scale e^N. As an immediate corollary, we find that the large deviations of $\lambda_{\min}$ and $\lambda_{\max}$ are asymptotically independent.

Corollary A.3. *Under the assumptions (H_{bulk}), (NoDown) and* (NoOut)*, for $\beta = 1$ or 2, the law of $(\lambda^N_{\min}, \lambda^N_{\max})$ under m^{β}_N satisfies a large deviation principle in the scale N and with good rate function $I^{\beta}_{\min}(x) + I^{\beta}(y)$.*

Proof. The proof is to tilt the measure by the rank two spherical integral of Proposition A.2 which implies that for $\theta > 0$ and $\theta' < 0$

$$m_N(|\lambda^N_{\min} - x| + |\lambda^N_{\max} - y| \le \delta)$$

$$\le e^{-N(J^{\beta}_{\mu_A \boxplus \mu_B}(\theta', x) + J^{\beta}_{\mu_A \boxplus \mu_B}(\theta, y) + o(1))} \mathbb{E}_U \left[\mathbb{E}_e \left(\mathrm{e}^{N\theta \langle e_1, (A+UBU^*) e_1 \rangle + N\theta' \langle e_2, (A+UBU^*) e_2 \rangle} \right) \right]$$

Now, since the law of (e_1, e_2) and (Ue_1, Ue_2) are independent and equidistributed, we deduce the upper bound as before. The proof of the lower bound is the same since for any (x, y) we find a unique couple (θ'_x, θ_y) which optimizes the rate function. □

Moreover, it is easy to deduce the following corollary, which is the extension of Proposition 2.3 to $\theta < 0$:

Corollary A.4. *Under Assumption 2.1, for $\beta = 1$ or 2, for any $\theta < 0$, for any $x < r(\mu_a \boxplus \mu_b)$,*

$$\lim_{\delta \downarrow 0} \limsup_{N \to +\infty} \frac{1}{N} \log m^{\beta,\theta}_N \left(\lambda^N_{\max} \in [x - \delta, x + \delta] \right) = -\infty, \tag{A.3}$$

and for any $x \ge r(\mu_a \boxplus \mu_b)$,

$$\lim_{\delta \downarrow 0} \limsup_{N \to +\infty} \frac{1}{N} \log m^{\beta,\theta}_N \left(\lambda^N_{\max} \in [x - \delta, x + \delta] \right) \le -I^{\beta}(x). \tag{A.4}$$

With Proposition A.2 in hand, the proof of Corollary A.4 follows the same lines as the proof of Proposition 2.3. We do not detail it and go directly to the proof of Proposition A.2. Note that this kind of factorization property has been already shown for θ and θ' not too far from zero, we refer the reader to [20, Theorem 7] or [15]. Our goal here is to extend this result to any pair of (θ, θ') of opposite sign.

Proof of Proposition A.2. For the sake of simplicity, we will stick to the case $\beta = 1$. Let g and g' be two independent standard Gaussian vectors in $\mathbb{R}^N$. If we denote by $\| \cdot \|_2$ the Euclidean norm and set

$$e_1 := \frac{g}{\|g\|_2}, h := g' - \frac{\langle g, g' \rangle}{\|g\|_2^2} g \text{ and } e_2 := \frac{h}{\|h\|_2},$$

then it is well known that (e_1, e_2) are two random vectors uniform on the unit sphere in $\mathbb{R}^N$, orthogonal to each other. Moreover, (e_1, e_2) is independent from $(\|g\|_2, \|g'\|_2, \langle g, g' \rangle)$. Indeed, one can use the following system of coordinates: $r := \|g\|_2$, $\gamma_1, \ldots, \gamma_{N-1}$ are the polar coordinates of g, $r' := \|g'\|_2$, η is the angle between g and g, and $\gamma'_1, \ldots, \gamma'_{N-2}$ are the angles needed to spot g' on the cone of angle η around g. One can check that the Gaussian measure decomposes as a product measure in these coordinates, $(\|g\|_2, \|g'\|_2, \langle g, g' \rangle)$ is a function of r, r', η whereas (e_1, e_2) is a function of the γ's and γ's. In particular, for any ε, if we let

$$A_N^\varepsilon := \{|\langle g, g' \rangle| \leq \varepsilon \|g\|_2 \|g'\|_2\},$$

then A_N^ε is independent of (e_1, e_2). Moreover, on A_N^ε, we have $\|h\|_2^2 \geq \|g'\|_2^2 (1 - \varepsilon^2)$ so that, for $\varepsilon < 1/2$,

$$\left| \theta \langle e_1, M_N e_1 \rangle + \theta' \langle e_2, M_N e_2 \rangle - \theta \frac{1}{\|g\|_2} \langle g, M_N g \rangle - \theta' \frac{1}{\|g'\|_2} \langle g', M_N g' \rangle \right| \leq 4M\varepsilon,$$

and

$$\begin{aligned} \mathbb{E}\left(\mathrm{e}^{N\theta \langle e_1, M_N e_1 \rangle + N\theta' \langle e_2, M_N e_2 \rangle} \right) &= \frac{1}{\mathbb{P}(A_N^\varepsilon)} \mathbb{E}\left(1_{A_N^\varepsilon} \mathrm{e}^{N\theta \langle e_1, M_N e_1 \rangle + N\theta' \langle e_2, M_N e_2 \rangle} \right) \\ &\leq \frac{\mathrm{e}^{4NM\varepsilon}}{\mathbb{P}(A_N^\varepsilon)} I_N(\theta, M_N) I_N(\theta', M_N). \end{aligned}$$

Because of the law of large numbers, for any $\varepsilon > 0$, $\mathbb{P}(A_N^\varepsilon)$ converges to 1 as N goes to infinity. Hence, letting N go to infinity and then ε going to zero, we get the upper bound in (A.2).

We now prove the lower bound. If O is an orthogonal matrix, the law of (Oe_1, Oe_2) is the same as the law of (e_1, e_2) so that we can assume without loss of generality that M_N is real diagonal, with eigenvalues that we denote by $\lambda_1^N \geq \lambda_2^N \geq \ldots \geq \lambda_N^N$. We refer the reader to Proposition 16 and Lemmas 18 to 21 in [20], in particular the proof of Lemma 19. We recall that T is the rate function for the large deviations of $\frac{\sum_{i=1}^N \lambda_i^N g_i^2}{\sum_{i=1}^N g_i^2}$. Let α_1^* be such that

$$\theta \alpha_1^* - T(\alpha_1^*) = \sup_\alpha (\theta \alpha - T(\alpha)).$$

As $\theta \geq 0$, one can check that $\alpha_{\min} \leq \alpha_1^* \leq \rho$. If $\alpha_1^* \in [\alpha_{\min}, \alpha_{\max}]$, we set $x_1 := 0$, whereas if $\alpha_1^* \in (\alpha_{\max}, \rho)$, we set $x_1 := (\rho - \alpha_1^*)((\rho - \alpha_1^*) H_{\max} - 1)$. Similarly, let α_2^* be such that

$$\theta' \alpha_2^* - T(\alpha_2^*) = \sup_\alpha (\theta' \alpha - T(\alpha)).$$

As $\theta' < 0$, one can check that $\ell \le \alpha_2^* \le \alpha_{\max}$. If $\alpha_2^* \in [\alpha_{\min}, \alpha_{\max}]$, we set $x_2 := 0$, whereas if $\alpha_2^* \in (\ell, \alpha_{\min})$, we set $x_2 := (\ell - \alpha_1^*)((\ell - \alpha_1^*)H_{\min} - 1)$. We now define, for any $\delta > 0$,

$$\begin{aligned} B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} \quad &:= \quad \Bigg\{ \left| (\lambda_1^N - \alpha_1^*) \frac{g_1^2}{N} + x_1 \right| \le \delta, \left| \frac{1}{N} \sum_{i=2}^{N} (\lambda_i^N - \alpha_1^*) g_i^2 - x_1 \right| \le \delta, \\ &\qquad \left| (\lambda_N^N - \alpha_2^*) \frac{(g_N')^2}{N} + x_2 \right| \le \delta, \left| \frac{1}{N} \sum_{i=1}^{N-1} (\lambda_i^N - \alpha_2^*)(g_i')^2 - x_2 \right| \le \delta \Bigg\}, \\ C \quad &:= \quad \left\{ \forall i \ge 2, g_i^2 \le N^{1/4}, \forall i \le N-1, (g_i')^2 \le N^{1/4} \right\}, \\ E_\delta \quad &:= \quad \left\{ \|g\|^2 \ge \sqrt{\delta} N, \|g'\|^2 \ge \sqrt{\delta} N \right\} \end{aligned}$$

We have

$$\begin{aligned} &\mathbb{E}\left(\mathrm{e}^{N\theta \langle e_1, M_N e_1 \rangle + N\theta' \langle e_2, M_N e_2 \rangle} \right) \ge \mathbb{E}\left(\mathbf{1}_{A_N^\varepsilon \cap B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} \cap C \cap E_\delta} \mathrm{e}^{N\theta \langle e_1, M_N e_1 \rangle + N\theta' \langle e_2, M_N e_2 \rangle} \right) \\ &\quad \ge E\left(\mathbf{1}_{A_N^\varepsilon \cap B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} \cap C \cap E_\delta} \mathrm{e}^{\theta \frac{1}{\|g\|_2} \langle g, M_N g \rangle - \theta' \frac{1}{\|g'\|_2} \langle g', M_N g' \rangle} \right) \mathrm{e}^{-4NM\varepsilon} \\ &\quad \ge \mathbb{P}(A_N^\varepsilon \cap B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} \cap C \cap E_\delta) \mathrm{e}^{-4NM\varepsilon} \mathrm{e}^{N\theta\alpha_1^* + N\theta'\alpha_2^* - 2N(\theta+\theta')\sqrt{\delta}} \end{aligned} \tag{A.5}$$

Now, if $\sigma_1, \dots, \sigma_N$ are N independent Rademacher random variables, independent of g and g', then $\langle g, g' \rangle$ and $\sum_{i=1}^N \sigma_i g_i g_i'$ have the same law. Therefore, since the sets $B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2}$, C and E_δ are independent of the sign of the g_i's,

$$\mathbb{P}(A_N^\varepsilon | B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} \cap C \cap E_\delta) = \mathbb{P}\left(\mathbb{P}(|\sum_{i=1}^N \sigma_i g_i g_i'| \le \varepsilon \|g\|_2 \|g'\|_2 \Big| (g, g')) \Big| B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} \cap C \cap E_\delta \right)$$

where the second expectation holds on the σ's only. Using the concentration properties of the Rademacher random variables (or the Azuma Hoeffding inequality), one gets that

$$\mathbb{P}\left(\left| \sum_{i=1}^N \sigma_i g_i g_i' \right| \ge \varepsilon \|g\|_2 \|g'\|_2 \Big| (g, g') \right) \le \mathrm{e}^{-\frac{\varepsilon^2 \|g\|_2^2 \|g'\|_2^2}{\sum_{i=1}^N g_i^2 (g_i')^2}}.$$

On $B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} \cap C \cap E_\delta$, the right hand side is bounded above by $\mathrm{e}^{-4\sqrt{N}\varepsilon^2\delta}$, so that we can conclude that, for any $\varepsilon, \delta > 0$, $\mathbb{P}(A_N^\varepsilon | B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} \cap C \cap E_\delta)$ converges to one as N goes to infinity.

Furthermore, we have that

$$\mathbb{P}(B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} \cap C \cap E_\delta) \ge \mathbb{P}(B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} \cap C) - \mathbb{P}(E_\delta^c). \tag{A.6}$$

Since it is well known that

$$\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}(E_\delta^c) \le -2(\sqrt{\delta} - 1 - \log(\sqrt{\delta}))$$

where the above right hand side goes to $-\infty$ as δ goes to zero, we only need to estimate the first term in the right hand side of (A.6) for small enough δ. Now

$$\mathbb{P}(B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} \cap C) \ge \mathbb{P}(B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} | C) \mathbb{P}(C),$$

where the last term goes to one as N goes to infinity. The last thing to check is that

$$\lim_{\delta \downarrow 0} \liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}(B^{\delta}_{\alpha_1^*, x_1, \alpha_2^*, x_2} | C) \ge -T(\alpha_1^*) - T(\alpha_2^*).$$

Indeed, going back to the proofs of Lemmas 18 and 19 in [20] (see also [11]), one can check that if $\alpha_1^* \in [\alpha_{\min}, \alpha_{\max}]$,

$$\lim_{\delta\downarrow 0}\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=2}^{N}(\lambda_i^N-\alpha_1^*)g_i^2\right|\le\delta|C\right)\ge -L^{\alpha_1^*}(0).$$

The proof is the same except that in the computation of the log-Laplace the integral will go from $-N^{1/4}$ to $N^{1/4}$ instead of running on $\mathbb{R}$ and this will not change the limit. Similarly, if $\alpha_1^* > \alpha_{\max}$,

$$\lim_{\delta\downarrow 0}\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(\left|(\lambda_1^N-\alpha_1^*)\frac{g_1^2}{N}+x_1\right|\le\delta\right)\ge -\frac{x_1}{2(\rho-\alpha_1^*)},$$

and

$$\lim_{\delta\downarrow 0}\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=2}^{N}(\lambda_i^N-\alpha_1^*)g_i^2-x_1\right|\le\delta|C\right)\ge -L^{\alpha_1^*}(x_1).$$

Putting everything together in (A.5), taking the limit as N then δ and ε go to zero, we get the factorization property. □

A.3 Large deviations for the largest eigenvalue in the deformed model

For the sake of simplicity, when treating the deformed model, we will stick to the case $\beta=1$. For any $x > \mathsf{r}(\mu_a \boxplus \mu_b)$, we denote by μ_x the measure defined as follows: for any bounded measurable function f,

$$\int f(\lambda)\mu_x(\mathrm{d}\lambda)=\int f\left(\frac{1}{x-\lambda}\right)\mu_a\boxplus\mu_b(\mathrm{d}\lambda).$$

In particular, for any $x > \mathsf{r}(\mu_a \boxplus \mu_b)$, $\int \lambda\mu_x(\mathrm{d}\lambda) = G_{\mu_a\boxplus\mu_b}(x)$.

For any $x > \rho \ge \mathsf{r}(\mu_a \boxplus \mu_b)$ and $\ell \le \mathsf{l}(\mu_a \boxplus \mu_b)$ we define

$$\alpha_{x,+}(\rho) := \begin{cases} \frac{G_{\mu_a\boxplus\mu_b}(\rho)}{1+(x-\rho)G_{\mu_a\boxplus\mu_b}(\rho)}, & \text{if } G_{\mu_a\boxplus\mu_b}(\rho)<\infty,\\ \infty, & \text{otherwise}\end{cases}$$

and

$$\alpha_{x,-}(\ell) := \begin{cases} \frac{G_{\mu_a\boxplus\mu_b}(\ell)}{1+(x-\ell)G_{\mu_a\boxplus\mu_b}(\ell)}, & \text{if } G_{\mu_a\boxplus\mu_b}(\ell)>-\infty,\\ -\infty, & \text{otherwise.}\end{cases}$$

For $\alpha \in \left(\frac{1}{x-\ell}, \frac{1}{x-\rho}\right)$ and $\kappa \notin \left(\frac{1}{x-\ell}, \frac{1}{x-\rho}\right)$, we set

$$h_{\alpha,x}(\kappa) := \frac{1}{2}\int \log\left(\frac{\kappa-\lambda}{\kappa-\alpha}\right)\mu_x(\mathrm{d}\lambda).$$

and we also extend it to

$$h_{\alpha,x}\left(\frac{1}{x-\rho}\right) := \lim_{y\downarrow\rho} h_{\alpha,x}\left(\frac{1}{x-y}\right) \quad \text{and} \quad h_{\alpha,x}\left(\frac{1}{x-\ell}\right) := \lim_{y\uparrow\ell} h_{\alpha,x}\left(\frac{1}{x-y}\right)$$

We set, for $x > \rho \ge \mathsf{r}(\mu_a \boxplus \mu_b)$ and $\ell \le \mathsf{l}(\mu_a \boxplus \mu_b)$,

$$T^+_{x,\rho}(\alpha) := \begin{cases} h_{\alpha,x}(K_{\mu_x}(Q_{\mu_x}(\alpha))), & \text{if } \alpha\in[G_{\mu_a\boxplus\mu_b}(x),\alpha_{x,+}(\rho)],\\ h_{\alpha,x}\left(\frac{1}{x-\rho}\right), & \text{if } \alpha\in\left(\alpha_{x,+}(\rho),\frac{1}{x-\rho}\right),\\ \infty, & \text{if } \alpha>\frac{1}{x-\rho}\end{cases} \tag{A.7}$$

and

$$T^-_{x,\ell}(\alpha) := \begin{cases} h_{\alpha,x}(K_{\mu_x}(Q_{\mu_x}(\alpha))), & \text{if } \alpha \in [\alpha_{x,-}(\ell), G_{\mu_a \boxplus \mu_b}(x)], \\ h_{\alpha,x}\left(\frac{1}{x-\ell}\right), & \text{if } \alpha \in \left(\frac{1}{x-\ell}, \alpha_{x,-}(\ell)\right) \\ \infty, & \text{if } \alpha < \frac{1}{x-\ell}. \end{cases} \tag{A.8}$$

The quantities above can easily be extended to the case $x = \rho > \mathsf{r}(\mu_a \boxplus \mu_b)$ (only the first line of (A.7) will be relevant). For $x = \rho = \mathsf{r}(\mu_a \boxplus \mu_b)$, we set

$$T^+_{x,x}(\alpha) := 0 \quad \text{if } \alpha \geq G_{\mu_a \boxplus \mu_b}(x)$$

and

$$T^-_{x,\ell}(\alpha) := \infty \quad \text{if } \alpha < G_{\mu_a \boxplus \mu_b}(x)$$

For $\gamma := (\gamma_1, \ldots, \gamma_p)$ a p-uplet of nonnegative real numbers, we now define,

$$L^{(0)}_\gamma(y) := I^1(y), \quad \text{if } y \geq \mathsf{r}(\mu_a \boxplus \mu_b)$$

and, for any $1 \leq i \leq p$,

$$L^{(i)}_\gamma(x) := \begin{cases} \inf\limits_{y \leq \mathsf{l}(\mu_a \boxplus \mu_b)} \left\{ T^-_{x,y}\left(\frac{1}{\gamma_i}\right) + I^1_{\min}(y) \right\}, & \text{if } \mathsf{r}(\mu_a \boxplus \mu_b) \leq x < K_{\mu_a \boxplus \mu_b}\left(\frac{1}{\gamma_i}\right), \\ \inf\limits_{\mathsf{r}(\mu_a \boxplus \mu_b) \leq y \leq x} \left\{ T^+_{x,y}\left(\frac{1}{\gamma_i}\right) + L^{(i-1)}_\gamma(y) \right\}, & \text{if } x \geq K_{\mu_a \boxplus \mu_b}\left(\frac{1}{\gamma_i}\right), \\ \infty, & \text{if } x < \mathsf{r}(\mu_a \boxplus \mu_b), \end{cases}$$

with the convention that

$$K_{\mu_a \boxplus \mu_b}\left(\frac{1}{\gamma_i}\right) = \mathsf{r}(\mu_a \boxplus \mu_b) \quad \text{if } G_{\mu_a \boxplus \mu_b}(\mathsf{r}(\mu_a \boxplus \mu_b)) \leq \frac{1}{\gamma_i}.$$

Note that this rate function should not depend on the ordering of the γ_i's, which is far from obvious on the formula above.

We can now state our main result. We recall that $(U_1^{(1)}, \ldots, U_1^{(p)})$ are independent random vectors uniformly distributed on the unit sphere. To simplify the notations, they can be viewed as respective first column vectors of p independent matrices distributed according to m^1_N.

Theorem A.5. *Under the assumptions (H_{bulk}), (NoOut) and (NoDown), for any $p \in \mathbb{N}^*$ and any $\gamma \in (\mathbb{R}_+)^p$, the law of the largest eigenvalue $\widetilde{\lambda^N_{\max}}$ of the matrix $X_N := A_N + UB_NU^* + \sum_{i=1}^p \gamma_i U_1^{(i)}(U_1^{(i)})^*$, defined in (2.4), under $(m^1_N)^{\otimes(p+1)}$ satisfies a large deviation principle in the scale N with good rate function $L^{(p)}_\gamma$.*

Before proving Theorem A.5, we need to state a variant of Proposition 16 in [20]. We denote by P the standard Gaussian measure on $\mathbb{R}$ and we assume that $(g_1, \ldots, g_N)$ follows the law $P^{\otimes N}$. For any N-tuple of real numbers $\lambda := (\lambda_1, \ldots, \lambda_N)$ and $x \notin \{\lambda_1, \ldots, \lambda_N\}$, we denote by $v_{N,\lambda}(x) := \frac{\sum_{i=1}^N \frac{1}{x-\lambda_i} g_i^2}{\sum_{i=1}^N g_i^2}$.

Proposition A.6. *Let $(\lambda^N_i)_{N \in \mathbb{N}^*, 1 \leq i \leq N}$ be a triangular array of real numbers such that $\frac{1}{N}\sum_{i=1}^N \delta_{\lambda^N_i}$ converges to $\mu_a \boxplus \mu_b$ as N grows to ∞. We denote by $\lambda^N := (\lambda^N_1, \ldots, \lambda^N_N)$. Assume that $\max_{i=1}^N \lambda^N_i$ converges, as N grows to ∞, to $\rho \geq \mathsf{r}(\mu_a \boxplus \mu_b)$. Let x be a real number such that, for N large enough, $x > \max_{i=1}^N \lambda^N_i$. Then, for any $\alpha \in \mathbb{R}$ such that $\alpha \geq G_{\mu_a \boxplus \mu_b}(x)$, we have*

$$\lim_{\delta \downarrow 0} \limsup_{N \to \infty} \frac{1}{N} \log P^{\otimes N}\left(v_{N,\lambda^N}(x) \in [\alpha - \delta, \alpha + \delta]\right)$$
$$= \lim_{\delta \downarrow 0} \liminf_{N \to \infty} \frac{1}{N} \log P^{\otimes N}\left(v_{N,\lambda^N}(x) \in [\alpha - \delta, \alpha + \delta]\right) = -T^+_{x,\rho}(\alpha). \tag{A.9}$$

Assume that $\min_{i=1}^N \lambda_i^N$ *converges, as* N *grows to* ∞*, to* $\ell \leq \mathsf{l}(\mu_a \boxplus \mu_b)$*. Then, for any* $\alpha \in \mathbb{R}$ *such that* $\alpha < G_{\mu_a \boxplus \mu_b}(x)$*, we have*

$$\lim_{\delta \downarrow 0} \limsup_{N \to \infty} \frac{1}{N} \log P^{\otimes N} \left(v_{N,\lambda^N}(x) \in [\alpha - \delta, \alpha + \delta] \right)$$
$$= \lim_{\delta \downarrow 0} \liminf_{N \to \infty} \frac{1}{N} \log P^{\otimes N} \left(v_{N,\lambda^N}(x) \in [\alpha - \delta, \alpha + \delta] \right) = -T^-_{x,\ell}(\alpha) . \quad \text{(A.10)}$$

We will not give a full proof of Proposition A.6. This follows from an adaptation of Lemma 18 and Proposition 16 in [20]. In Lemma 18 in particular, one can check that the deviations above the mean (which is $G_{\mu_a \boxplus \mu_b}(x)$ in the present case) may involve not only the limiting empirical distribution but also the limit as N grows to ∞ of the largest particle (denoted by $\max_{i=1}^N \gamma_i$ there and equal to $\frac{1}{x - \max_{i=1}^N \lambda_i^N}$ in the present case), whereas the deviations below the mean may depend on the limiting smallest particle, equal to $\frac{1}{\min_{i=1}^N \lambda_i^N - x}$ here.

The rest of this section is devoted to the proof of Theorem A.5 in the case $p = 1$. For $p > 1$, the proof is very similar, except that instead of conditioning by the deviations of the extreme eigenvalues of H_N, we will condition on the deviations of extreme eigenvalues of the model at step $p - 1$.

Proof of Theorem A.5 in the case $p = 1$*.* We recall that we stick to the case $\beta = 1$. Let $\gamma_1 > 0$ be fixed. As in the proof of Theorem 2.5, the exponential tightness is straightforward : for any $N \geq 1$,

$$|\widetilde{\lambda^N_{\max}}| \leq K + \gamma_1.$$

Again, using e.g. Theorem D.4(a) and Corollary D.6 in [1], it is enough to show that, for any $x \in \mathbb{R}$,

$$\lim_{\delta \downarrow 0} \limsup_{N \to \infty} \frac{1}{N} \log (m_N^1)^{\otimes 2} \left(\widetilde{\lambda^N_{\max}} \in [x - \delta, x + \delta] \right)$$
$$= \lim_{\delta \downarrow 0} \liminf_{N \to \infty} \frac{1}{N} \log (m_N^1)^{\otimes 2} \left(\widetilde{\lambda^N_{\max}} \in [x - \delta, x + \delta] \right) = -L^{(1)}_\gamma(x).$$

For any z which does not belong to the spectrum of H_N, one can write

$$\det(zI_N - X_N) = \det(zI_N - H_N) \gamma_1 \left(\frac{1}{\gamma_1} - (U_1^{(1)})^* (zI_N - H_N)^{-1} U_1^{(1)} \right).$$

Therefore, z is an eigenvalue of X_N which is not an eigenvalue of H_N if and only if

$$(U_1^{(1)})^* (zI_N - H_N)^{-1} U_1^{(1)} = \frac{1}{\gamma_1}.$$

The Haar measure is invariant by unitary conjugation, so that if $H_N = U_N D_N U_N^*$, then $(U_1^{(1)})^* (zI_N - H_N)^{-1} U_1^{(1)}$ and $(U_1^{(1)})^* (zI_N - D_N)^{-1} U_1^{(1)}$ have the same law and one can assume in the sequel that H_N is diagonal. Moreover, as $U_1^{(1)}$ is a column vector of a matrix distributed according to the Haar measure, the equation now reads

$$v_N(z) := \sum_{i=1}^N \frac{1}{z - \lambda_i^{(H_N)}} \frac{g_i^2}{\sum\limits_{i=1}^N g_i^2} = \frac{1}{\gamma_1},$$

with $(g_1, \ldots, g_N)$ having distribution $P^{\otimes N}$.

For any $(\lambda_1,\dots,\lambda_N)$ and $(v_1,\dots,v_N)$ such that $\sum_{i=1}^N v_i^2 = 1$ fixed, we define on $(\max_{i=1}^N \lambda_i, \infty)$, the function

$$f_{\lambda,v} : z \mapsto \sum_{i=1}^N \frac{1}{z-\lambda_i} v_i^2.$$

This function is decreasing and continuous on $(\max_{i=1}^N \lambda_i,\infty)$, uniformly on $(v_1,\dots,v_N)$ such that $\sum_{i=1}^N v_i^2 = 1$. Therefore, there exists a function ε_λ going to zero at zero, such that for $z \in (\max_{i=1}^N \lambda_i,\infty)$, $f_\lambda(z) = \frac{1}{\gamma_1}$ if and only if, for any $\delta > 0$ small enough, for any $x \in [z-\delta, z+\delta]$, $f_\lambda(x) \in \left[\frac{1}{\gamma_1} - \varepsilon_\lambda(\delta), \frac{1}{\gamma_1} + \varepsilon_\lambda(\delta)\right]$. Let $x > \mathsf{r}(\mu_a \boxplus \mu_b)$ be fixed. Let y such that $\mathsf{r}(\mu_a \boxplus \mu_b) \le y < x$ and set $\eta_0 := \frac{x-y}{4}$. For any $\eta < \eta_0$, similarly to the definition of $E^y_{N,\eta}$ in (3.6), we introduce

$$\widetilde{E^y_{N,\eta}} := \left\{ \lambda^N_{\max} \in [y, y+\eta], \mathrm{d}(\hat\mu_N, \nu_N^\beta) \le N^{-1/4} \right\}$$

The analysis will be the same, except possibly for $y = \mathsf{r}(\mu_a \boxplus \mu_b)$. We have that for any $U \in \widetilde{E^y_{N,\eta}}$, $\lambda^N_{\max} = \lambda_1^{(H_N)} \in [y, y+\eta]$. Therefore, if we denote by

$$\widetilde{v_N}(x,y) := \frac{1}{x-y} \frac{g_1^2}{\sum\limits_{i=1}^N g_i^2} + \sum_{i=2}^N \frac{1}{x-\lambda_i^{(H_N)}} \frac{g_i^2}{\sum\limits_{i=1}^N g_i^2}$$

then

$$|\widetilde{v_N}(x,y) - v_N(x)| \le \frac{|y-\lambda_1^{(H_N)}|}{(x-y)(x-\lambda_1^{(H_N)})} \le \frac{\eta}{\eta_0^2}.$$

Therefore, for any $\eta < \eta_0$, there exists a continuous function ε_η going to zero at zero such that, for any $\delta \le \eta$ and N large enough,

$$\begin{aligned}
&(m_N^1)^{\otimes 2}\left(\left\{\widetilde{\lambda^N_{\max}} \in [x-\delta, x+\delta]\right\} \cap \widetilde{\mathsf{E}^y_{N,\eta}}\right) \\
&\qquad = P^{\otimes N} \otimes m_N^1 \left(\left\{\widetilde{v^N}(x,y) \in \left[\frac{1}{\gamma_1} - \varepsilon(\delta), \frac{1}{\gamma_1} + \varepsilon(\delta)\right]\right\} \cap \widetilde{\mathsf{E}^y_{N,\eta}}\right) \\
&\qquad = P^{\otimes N} \otimes m_N^1 \left(\widetilde{v^N}(x,y) \in \left[\frac{1}{\gamma_1} - \varepsilon(\delta), \frac{1}{\gamma_1} + \varepsilon(\delta)\right] | \widetilde{\mathsf{E}^y_{N,\eta}}\right) m_N^1(\widetilde{\mathsf{E}^y_{N,\eta}}) \qquad \text{(A.11)}
\end{aligned}$$

The probability measure on the right handside is $P^{\otimes N} \otimes m_N^1$ because $\widetilde{v^N}(x,y)$ can be seen as a function of U of law m_N^1 and of $(g_1,\dots,g_N)$ of law $P^{\otimes N}$.

If we assume that $\eta, \delta < \frac{|x-y|}{4}$ and for all $i \in \mathbb{N}^*$, $\lambda_i \le y+\eta$, one can choose ε_λ uniformly in $(\lambda_1,\dots,\lambda_N)$. Now, let $U \in \widetilde{\mathsf{E}^y_{N,\eta}}$ be chosen. We denote by $\lambda_1^N := y$ and for any $2 \le i \le N$, $\lambda_i^N := \lambda_i^{(H_N)}$. Then $\frac{1}{N}\sum_{i=1}^N \delta_{\lambda_i^N}$ converges to $\mu_a \boxplus \mu_b$ and $\max_{i=1}^N \lambda_i^N = y$. By Proposition A.6, if $G_{\mu_a \boxplus \mu_b}(x) \le \frac{1}{\gamma_1}$, we have

$$\lim_{\delta \downarrow 0} \liminf_{N\to\infty} \frac{1}{N} \log P^{\otimes N}\left(\widetilde{v_N}(x) \in \left[\frac{1}{\gamma_1} - \varepsilon_\eta(\delta), \frac{1}{\gamma_1} + \varepsilon_\eta(\delta)\right] | \widetilde{\mathsf{E}^y_{N,\eta}}\right) = -T^+_{x,y}\left(\frac{1}{\gamma_i}\right),$$

so that

$$\lim_{\delta\downarrow 0} \liminf_{N\to\infty} \frac{1}{N} \log (m_N^1)^{\otimes 2}(\widetilde{\lambda^N_{\max}} \in [x-\delta, x+\delta]) \ge -T^+_{x,y}\left(\frac{1}{\gamma_1}\right) + \lim_{N\to\infty} \frac{1}{N} \log m_N^1(\widetilde{\mathsf{E}^y_{N,\eta}}).$$

Taking the limit of the right hand-side as η goes to zero, we get using Theorem 2.5 that

$$\lim_{\delta\downarrow 0} \lim_{N\to\infty} \frac{1}{N} \log (m_N^1)^{\otimes 2}(\widetilde{\lambda^N_{\max}} \in [x-\delta, x+\delta]) \ge -T^+_{x,y}\left(\frac{1}{\gamma_1}\right) - I^1(y) \ge -L^{(1)}_\gamma(x),$$

where the last inequality was obtained by optimizing on y.

Assume now that $\mathsf{r} := \mathsf{r}(\mu_a \boxplus \mu_b) < x < K_{\mu_a \boxplus \mu_b}\left(\frac{1}{\gamma_1}\right)$. Similarly to (3.6), we define, for $y < \mathsf{l}(\mu_a \boxplus \mu_b)$

$$\mathsf{E}^{y,-}_{N,\eta} := \left\{ \lambda^N_{\min} \in [y, y+\eta], \lambda^N_{\max} \in [\mathsf{r}-\eta, \mathsf{r}+\eta], \mathrm{d}(\hat{\mu}_N, \nu^1_N) \leq N^{-1/4} \right\},$$

and

$$v_{N,-}(x) := \sum_{i=1}^{N-1} \frac{1}{x - \lambda_i^{(H_N)}} \frac{g_i^2}{\sum\limits_{i=1}^{N} g_i^2} + \frac{1}{x-y} \frac{g_N^2}{\sum\limits_{i=1}^{N} g_i^2}.$$

For η small enough and $\delta \leq \eta$, we can then write as above

$$\begin{aligned}(m^1_N)^{\otimes 2}(\widetilde{\lambda^N_{\max}} \in [x-\delta, x+\delta]) &\geq (m^1_N)^{\otimes 2}(\widetilde{\lambda^N_{\max}} \in [x-\delta, x+\delta] \cap \mathsf{E}^{y,-}_{N,\eta}) \\ &= P^{\otimes n} \otimes m^1_N \left(v_{N,-}(x) \in \left[\frac{1}{\gamma_1} - \varepsilon_\eta(\delta), \frac{1}{\gamma_1} + \varepsilon_\eta(\delta) \right] \cap E^{y,-}_{N,\eta} \right) \\ &= P^{\otimes n} \otimes m^1_N \left(v_{N,-}(x) \in \left[\frac{1}{\gamma_1} - \varepsilon_\eta(\delta), \frac{1}{\gamma_1} + \varepsilon_\eta(\delta) \right] | E^{y,-}_{N,\eta} \right) m^1_N(E^{y,-}_{N,\eta}). \quad \text{(A.12)}\end{aligned}$$

In this case, by Proposition A.6,

$$\lim_{\delta \downarrow 0} \liminf_{N\to\infty} \frac{1}{N} \log P^{\otimes N} \left(v_{N,-}(x) \in \left[\frac{1}{\gamma_1} - \varepsilon_\eta(\delta), \frac{1}{\gamma_1} + \varepsilon_\eta(\delta) \right] | \mathsf{E}^{y,-}_{N,\eta} \right) = -T^-_{x,y}\left(\frac{1}{\gamma_1}\right),$$

so that

$$\lim_{\delta \downarrow 0} \liminf_{N\to\infty} \frac{1}{N} \log (m^1_N)^{\otimes 2}(\widetilde{\lambda^N_{\max}} \in [x-\delta, x+\delta]) = -T^-_{x,y}\left(\frac{1}{\gamma_1}\right) + \liminf_{N\to\infty} \frac{1}{N} \log m^1_N(E^{y,-}_{N,\eta}). \quad \text{(A.13)}$$

The last step to prove the lower bound in this case is to check

$$\lim_{\eta \downarrow 0} \liminf_{N\to\infty} \frac{1}{N} \log m^1_N(E^{y,-}_{N,\eta}) \geq -I^1_{\min}(y). \quad \text{(A.14)}$$

Then, taking the limit as η goes to zero in (A.13) and optimizing in y gives the required lower bound.

We now prove (A.14). Similarly to Lemma 4.2 and 5.1 (by symmetry between the smallest and largest eigenvalue), one can show that there exists a unique $\theta_y \leq 0$ such that, for any $\eta > 0$ and N large enough,

$$m_N^{1,\theta_y}\left(\lambda^N_{\min} \in [y-\eta, y+\eta], \mathrm{d}(\hat{\mu}_N, \nu^1_N) \leq N^{-1/4} \right) \geq \frac{2}{3}.$$

Applying Corollary A.4, as $\theta_y \leq 0$, we have that, for any $\eta > 0$ and N large enough,

$$m_N^{1,\theta_y}(\lambda^N_{\max} \in [\mathsf{r}-\eta, \mathsf{r}+\eta]) \geq \frac{2}{3}, \quad \text{(A.15)}$$

so that, for any $\eta > 0$ and N large enough,

$$m_N^{1,\theta_y}(\mathsf{E}^{y,-}_{N,\eta}) \geq \frac{1}{3}.$$

With this ingredient, the proof of (A.14) goes as in the proof of Proposition 2.4:

$$\begin{aligned}m^1_N(\mathsf{E}^{y,-}_{N,\eta}) &= \mathbb{E}_{m^1_N}\left(1_{\mathsf{E}^{y,-}_{N,\eta}} \frac{I^1_N(\theta_y, H)}{I^1_N(\theta_y, H)} \right) \\ &\geq \inf_{U \in \mathsf{E}^{y,-}_{N,\eta}} \frac{1}{I^1_N(\theta_y, A + UBU^*)} I^1_N(\theta_y, A) I^1_N(\theta_y, B) m_N^{1,\theta_y}(\mathsf{E}^{y,-}_{N,\eta}),\end{aligned}$$

so that, using again Lemma 3.4, we get:

$$\lim_{\eta\downarrow 0}\liminf_{N\to\infty}\frac{1}{N}\log m_N^1\left(\mathsf{E}_{N,\eta}^{y,-}\right)\geq -I_{\min}^1(\theta_y,y)-\lim_{\eta\downarrow 0}g_{\theta_y}(\eta)=-I_{\min}^1(y).$$

The strategy to get the upper bound is similar: we know that, for $N\geq 1$, $\lambda_{\max}^N\in[-K,K]$ and $\lambda_{\min}^N\in[-K,K]$. For any $\delta>0$, there exists $p\in\mathbb{N}^*$ and $\rho_1,\dots,\rho_p$ such that

$$[-K,K]\subset\cup_{i=1}^p[\rho_i-\delta,\rho_i+\delta].$$

Assume that $G_{\mu_a\boxplus\mu_b}(x)\leq\frac{1}{\gamma_1}$.

$$\begin{aligned}(m_N^1)^{\otimes 2}(\widetilde{\lambda_{\max}^N}\in[x-\delta,x+\delta])&\leq(m_N^1)^{\otimes 2}(\widetilde{\lambda_{\max}^N}\in[x-\delta,x+\delta]\cap\{\mathrm{d}(\hat\mu_N,\nu_N^1)\leq N^{-1/4}\})\\&\quad+m_N^1(\mathrm{d}(\hat\mu_N,\nu_N^1)>N^{-1/4})\\&\leq\sum_{i=1}^p(m_N^1)^{\otimes 2}(\widetilde{\lambda_{\max}^N}\in[x-\delta,x+\delta]\cap\mathsf{E}_{N,\delta}^{\rho_i})\\&\quad+m_N^1(\mathrm{d}(\hat\mu_N,\nu_N^1)>N^{-1/4}).\end{aligned}$$

We then use Lemma 3.3 to get rid of the last term and apply the same strategy as before, combining the relation (A.12) and Proposition A.6 for the main term.

Assume now that $G_{\mu_a\boxplus\mu_b}(x)>\frac{1}{\gamma_1}$. We apply the very same strategy with $\mathsf{E}_{N,\delta}^{\rho_i,-}$ instead of $\mathsf{E}_{N,\delta}^{\rho_i}$ and the bound (A.14). □

References

[1] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni, *An introduction to random matrices*, Cambridge Studies in Advanced Mathematics, vol. 118, Cambridge University Press, Cambridge, 2010. MR-2760897

[2] Fanny Augeri, *Large deviations principle for the largest eigenvalue of Wigner matrices without Gaussian tails*, Electron. J. Probab. **21** (2016), Paper No. 32, 49. MR-3492936

[3] Fanny Augeri, *On the large deviations of traces of random matrices*, Ann. Inst. Henri Poincaré Probab. Stat. **54** (2018), no. 4, 2239–2285. MR-3865672

[4] Zhigang Bao, László Erdős, and Kevin Schnelli, *Local law of addition of random matrices on optimal scale*, Comm. Math. Phys. **349** (2017), no. 3, 947–990. MR-3602820

[5] Serban Teodor Belinschi, *A note on regularity for free convolutions*, Ann. Inst. H. Poincaré Probab. Statist. **42** (2006), no. 5, 635–648. MR-2259979

[6] Serban Teodor Belinschi, *The Lebesgue decomposition of the free additive convolution of two probability distributions*, Probab. Theory Related Fields **142** (2008), no. 1–2, 125–150. MR-2413268

[7] Serban Teodor Belinschi, Hari Bercovici, Mireille Capitaine, and Maxime Février, *Outliers in the spectrum of large deformed unitarily invariant models*, Ann. Probab. **45** (2017), no. 6A, 3571–3625. MR-3729610

[8] Gérard Ben Arous, Amir Dembo, and Alice Guionnet, *Aging of spherical spin glasses*, Probab. Theory Related Fields **120** (2001), no. 1, 1–67. MR-1856194

[9] Gérard Ben Arous and Alice Guionnet, *Large deviations for Wigner's law and Voiculescu's non-commutative entropy*, Probab. Theory Related Fields **108** (1997), no. 4, 517–542. MR-1465640

[10] Florent Benaych-Georges, Alice Guionnet, and Mylène Maïda, *Large deviations of the extreme eigenvalues of random deformations of matrices*, Probab. Theory Related Fields **154** (2012), no. 3-4, 703–751. MR-3000560

[11] Bernard Bercu, Fabrice Gamboa, and Alain Rouault, *Large deviations for quadratic forms of stationary Gaussian processes*, Stochastic Process. Appl. **71** (1997), no. 1, 75–90. MR-1480640

[12] Charles Bordenave and Pietro Caputo, *A large deviation principle for Wigner matrices without Gaussian tails*, Ann. Probab. **42** (2014), no. 6, 2454–2496. MR-3265172

[13] Mireille Capitaine and Catherine Donati-Martin, *Spectrum of deformed random matrices and free probability*, Advanced topics in random matrices, Panor. Synthèses, vol. 53, Soc. Math. France, Paris, 2017, pp. 151–190. MR-3792626

[14] Benoît Collins and Camille Male, *The strong asymptotic freeness of Haar and deterministic matrices*, Ann. Sci. Éc. Norm. Supér. (4) **47** (2014), no. 1, 147–163. MR-3205602

[15] Benoît Collins and Piotr Śniady, *New scaling of Itzykson-Zuber integrals*, Ann. Inst. H. Poincaré Probab. Statist. **43** (2007), no. 2, 139–146. MR-2303115

[16] Catherine Donati-Martin and Mylène Maïda, *Large deviations for the largest eigenvalue of an Hermitian Brownian motion*, ALEA Lat. Am. J. Probab. Math. Stat. **9** (2012), no. 2, 501–530. MR-3069375

[17] 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** (2008), no. 4, 1048–1071. MR-2488532

[18] Benjamin Groux, *Asymptotic freeness for rectangular random matrices and large deviations for sample convariance matrices with sub-Gaussian tails*, Electron. J. Probab. **22** (2017), Paper No. 53, 40. MR-3666016

[19] Alice Guionnet and Jonathan Husson, *Large deviations for the largest eigenvalue of Rademacher matrices*, to appear in Annals of Probab. (2020), https://arxiv.org/abs/1810.01188.

[20] Alice Guionnet and Mylène Maïda, *A Fourier view on the R-transform and related asymptotics of spherical integrals*, J. Funct. Anal. **222** (2005), no. 2, 435–490. MR-2132396

[21] Alice Guionnet and Ofer Zeitouni, *Large deviations asymptotics for spherical integrals*, J. Funct. Anal. **188** (2002), no. 2, 461–515. MR-1883414

[22] Vladislav Kargin, *A concentration inequality and a local law for the sum of two random matrices*, Probab. Theory Related Fields **154** (2012), no. 3–4, 677–702. MR-3000559

[23] Mylène Maïda, *Large deviations for the largest eigenvalue of rank one deformations of Gaussian ensembles*, Electron. J. Probab. **12** (2007), 1131–1150. MR-2336602

[24] Elizabeth S. Meckes and Mark W. Meckes, *Concentration and convergence rates for spectral measures of random matrices*, Probab. Theory Related Fields **156** (2013), no. 1–2, 145–164. MR-3055255

[25] L. Pastur and V. Vasilchuk, *On the law of addition of random matrices*, Comm. Math. Phys. **214** (2000), no. 2, 249–286. MR-1796022

[26] Roland Speicher, *Free convolution and the random sum of matrices*, Publ. Res. Inst. Math. Sci. **29** (1993), no. 5, 731–744. MR-1245015

[27] Dan Voiculescu, *Limit laws for random matrices and free products*, Invent. Math. **104** (1991), no. 1, 201–220. MR-1094052

Acknowledgments. 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 benefited from many discussions with J. Husson and F. Augeri with whom one of the author is working on a companion project on Wigner matrices. We also thank Benjamin McKenna for pointing out a gap in the proof of Proposition 3.1 and anonymous referees for very useful remarks in particular on the Appendix. Finally, we are very grateful for stimulating discussions with O. Zeitouni and N. Cook.