Electronic Journal of Probability

Electron. J. Probab. **29** (2024), article no. 187, 1–48.
ISSN: 1083-6489 https://doi.org/10.1214/24-EJP1228

# Large deviations for the largest eigenvalue of generalized sample covariance matrices[*]

Jonathan Husson[†] Benjamin M^cKenna[‡]

**Abstract**

We establish a large-deviations principle for the largest eigenvalue of a generalized sample covariance matrix, meaning a matrix proportional to $Z^T \Gamma Z$, where $Z$ has i.i.d. real or complex entries and $\Gamma$ is not necessarily the identity. We treat the classical case when $Z$ is Gaussian and $\Gamma$ is positive definite, but we also cover two orthogonal extensions: Either the entries of $Z$ can instead be sharp sub-Gaussian, a class including Rademacher and uniform distributions, where we find the same rate function as for the Gaussian model; or $\Gamma$ can have negative eigenvalues if $Z$ remains Gaussian. The latter case confirms formulas of Maillard in the physics literature.

We also apply our techniques to the largest eigenvalue of a deformed Wigner matrix, real or complex, where we upgrade previous large-deviations estimates to a full large-deviations principle. Finally, we remove several technical assumptions present in previous related works.

**Keywords:** large deviations; sample covariance matrices; Wishart matrices; deformed Wigner matrices.
**MSC2020 subject classifications:** 60B20; 60F10; 15B52.
Submitted to EJP on March 8, 2023, final version accepted on October 12, 2024.
Supersedes arXiv:2302.02847v1.

## 1 Introduction

### 1.1 Our results

In this paper, we give a large-deviations principle at speed $N$ for the largest eigenvalue $\lambda_{\max}$ of a *generalized sample covariance matrix*, meaning a random matrix of the form

$$H_N = \frac{1}{M} Z^T \Gamma Z = \frac{1}{M} \sum_{i=1}^{M} d_i z_i z_i^T, \tag{1.1}$$

[*] This material is based upon work supported by the National Science Foundation under Grant No. DMS-1928930 while the authors were in residence at the Mathematical Sciences Research Institute (now the Simons Laufer Mathematical Sciences Institute) in Berkeley, California, during the Fall 2021 semester.

[†] University of Michigan, United States of America. E-mail: jonathan.husson@uca.fr

[‡] Harvard University, Center of Mathematical Sciences and Applications, United States of America.
E-mail: mckenna@math.gatech.edu

where all the notation is defined below. Informally, this means that we find a function $I$ such that

$$\mathbb{P}(\lambda_{\max}(H_N) \approx x) \approx \exp(-NI(x)). \tag{1.2}$$

Here $Z \in \mathbb{R}^{M\times N}$ has i.i.d. entries distributed according to some centered probability measure $\mu$ with unit variance; its rows, which are independent vectors of length $N$, are written in column form (in order to match the literature) as $\{z_i\}_{i=1}^M$; *generalized* means that the deterministic matrix $\Gamma = \operatorname{diag}(d_1,\dots,d_M) = \operatorname{diag}(d_1^{(M)},\dots,d_M^{(M)})$ may not be the identity; and $M$ and $N$ are parameters both tending to infinity in such a way that their ratio is order one.

Our results also hold for the complex case $H_N = M^{-1}Z^*\Gamma Z$, but for the sake of simplicity we limit our notation to the real case for most of the paper.

When all $d_i$'s are positive, the matrix $H_N$ is sometimes called an *inhomogeneous* sample covariance matrix (or an inhomogeneous Wishart matrix). In this paper, we can allow for negative $d_i$'s, in the special case when $\mu$ is Gaussian. This part of our result matches a recent result of Maillard in the physics literature, which inspired our work [27]. We also allow non-Gaussian $\mu$'s, but we require them to lie in a special class of measures called *sharp sub-Gaussian*, which includes as important special cases the Rademacher and uniform distributions, and here we require all the $d_i$'s to be positive; in this case we show that the rate function is the same as in the corresponding Gaussian case. (The rate function probably should *not* match the Gaussian one when some $d_i$'s are negative; see Remark 2.19.)

The case where $\Gamma$ is positive definite is better known. In that case, $H_N$ has the same distribution as $\frac{1}{M}\sum_{i=1}^M x_i x_i^T$, where the vectors of $x_i$ are independent, but the components of each $x_i$ have possibly nontrivial covariance matrix $\Gamma$. In this case, $H_N$ of course also has the same nonzero eigenvalues as the variant $\frac{1}{M}\Gamma^{1/2}ZZ^T\Gamma^{1/2}$, and the literature is sometimes phrased in this way. Our results are stated and proved for diagonal $\Gamma$, but in the Gaussian case, they apply automatically to non-diagonal $\Gamma$ just by rotational invariance of Gaussian measure. This means that, in the Gaussian case, when $\Gamma$ is positive definite, $H_N$ is a sample covariance matrix of *independent* observations of possibly *correlated* data. (The case of non-diagonal $\Gamma$ and non-Gaussian $Z$ is interesting, but we do not consider it in this paper.) But we emphasize that the case where $\Gamma$ has positive and negative eigenvalues also has genuine statistical interest: it appears, for example, in MANOVA estimation of the above scenario when the observations $x_i$ are no longer independent of one another. For a discussion of this case, we direct the readers to [14].

### 1.2 Previous results on large deviations for random matrices

Our work fits into a recent lineage of papers, starting with [16] and [18], which establish large-deviations principles for random matrices using the method of *tilting by spherical integrals* (see Section 2.2 for a brief introduction to spherical integrals). Papers in this line include [6, 1, 31, 17, 22, 4] in the math literature, and [27, 32] in the physics literature. Compared to these previous works, the main technical novelty in our work is the establishment of rigorous non-Gaussian results beyond the so-called $x_c$ *threshold*. That is, in the model (1.1), it turns out that there exists some $x_c = x_c(\Gamma)$, which may be finite or infinite depending on $\Gamma$, such that if $x_c < \infty$ then the function $I$ from (1.2) is non-analytic at $x_c$. A similar phenomenon was partially shown for additively deformed Wigner matrices in [31]; we will define this model properly below, but informally speaking, it comes in both a Gaussian variety and a sharp sub-Gaussian variety (and other varieties not considered). In [31], one of the present authors showed the analogue of (1.2) for all $x$ in the Gaussian case, but only for $x < x_c$ in the sharp sub-Gaussian case. Dealing with the regime $x \geq x_c$ in the sharp sub-Gaussian case remained a challenge.

Following these works, Maillard considered the present model (1.1), in the Gaussian case and at the physics level of rigor. He also found an $x_c$ threshold, and proposed an interesting new tilting for establishing (1.2) in the regime $x \geq x_c$, which motivated our work. We are able to verify his results, plus add new ones of our own, but without using his techniques. Instead, when considering a model with $x_c < +\infty$, we find an approximating sequence of models which all have $x_c = +\infty$. In the context of recent results on large deviations for random matrices, a similar argument first appeared in [22], but to handle a problem other than the $x_c$ threshold. Here we show it can also be useful for $x_c$ thresholds, in a variety of models. Indeed, although the bulk of the paper treats the sample-covariance model (1.1), in Appendix D we consider the deformed-Wigner case and complete the analysis started by [31].

We now remark on the interpretation of the $x_c$ threshold. Typically, points of non-analyticity in rate functions appear from a change in the underlying mechanism of deviation (here, meaning that the cheapest strategy to make $\{\lambda_{\max}(H_N) \approx x\}$ would be qualitatively different if $x < x_c$ or $x > x_c$). For these models, one wonders if this phase transition appears in the eigenvector corresponding to the largest eigenvalue: one might imagine that, conditioned on the event $\{\lambda_{\max}(H_N) \approx x\}$, this eigenvector is localized for $x > x_c$ (reflecting a localized strategy, such as making one matrix entry very large) and delocalized for $x < x_c$ (reflecting a delocalized strategy, such as making every matrix entry a little larger than normal). In the sharp sub-Gaussian case, this would be coherent with our proof strategy, which involves integrating quadratic forms $\langle e, H_N e\rangle$ over $e$ in the unit sphere; we show that this integral is dominated by delocalized $e$ for $x < x_c$, but this argument breaks as $x \uparrow x_c$ (see the proof of Lemma 3.11 below). Understanding this better is a major motivation for us to consider this model; we also direct readers to related recent work of Cook, Ducatez, and Guionnet [10], which shows for a different model that the conditional top eigenvector is localized for $x$ large enough. However, in our model, the $x_c$ threshold is sometimes finite and sometimes infinite, depending roughly on whether the limiting measure of $\Gamma$ has a continuous density at its right edge, and we do not have a heuristic explanation for why this edge behavior of $\Gamma$ would allow/forbid this sort of (de)localization transition in the top eigenvector of $H_N$.

As special cases of our main theorem, we recover earlier results in the homogeneous case $\Gamma = \mathrm{Id}$, i.e., for usual Wishart matrices. Majumdar and Vergassola computed the rate function in the Gaussian case [28]; later, recent work by Guionnet and the first author extended this sharp sub-Gaussian Wishart matrices [16]. The computation that our rate function matches theirs was carried out by Maillard [27].

We also develop techniques to remove a technical assumption present in the works cited above. Typically, the results of these works can be written informally as “Suppose $X$ is a random matrix built from samples of $\mu$, a sharp-sub-Gaussian measure, which is also either compactly supported or satisfies the log-Sobolev inequality. Then $\lambda_{\max}(X)$ satisfies an LDP.” We show how to remove the “compact-or-log-Sobolev” assumption, ending with theorems of the form “Suppose $X$ is a random matrix built from samples of $\mu$, a sharp sub-Gaussian measure. Then $\lambda_{\max}(X)$ satisfies an LDP.” For example, our main result is written in this latter way, as is our deformed-Wigner result just mentioned. In Appendix C, we give one more example, refining a result of Guionnet and the first author [16]. We show that this extension is nontrivial, in Remark C.2, by giving an example of a sharp sub-Gaussian law that is neither compactly supported nor satisfies the log-Sobolev inequality. As we explain there, our proof also allows one to bypass the use of local laws.

Finally, we make a historical remark on the sharp-sub-Gaussian condition. Of course the notion of “sub-Gaussian random variable” is quite classical, but the works in the lineage discussed above are all based on the additional ideas that (a) a particular subclass of this family, called “sharp-sub-Gaussian” in Definition 2.1 below, is special, and (b) what makes this subclass special, is that random matrices constructed from its members satisfy bounds with the *same constants* as for Gaussian measure itself. To the best of our knowledge, these LDP works were interested in finding, but not aware of, any other place in the probability literature where these ideas appeared. We recently learned of such a place, namely a remark in recent work of Zhivotovskiy on the operator norm of a sum of independent random matrices [40, Remark 2.9]. Zhivotovskiy remarks that these same ideas previously appeared implicitly in works of Catoni and collaborators.

### 1.3 Previous results on generalized sample covariance matrices

The model $H_N$ and its variants are quite classical. Their first appearance, to the best of our knowledge, is in a 1967 paper of Marčenko and Pastur, which proved a global law for the variant $A + \frac{1}{M} Z^T \Gamma Z$, where $A$ is a sequence of deterministic matrices whose empirical measures have some limit, and where $\Gamma$ is random and diagonal with i.i.d. entries. The global law for our precise model appears in (more general) work of Silverstein and Bai [36], which also gives a good summary of the literature on global laws for variants of $H_N$.

In our work we assume that $\Gamma$ has no outlier eigenvalues; for example we do not allow $\Gamma$ to be a finite-rank perturbation of the identity. This restriction is generally believed, and under very mild restrictions proved, to prevent $H_N$ from typically having outlier eigenvalues. That is, we are not in the regime of the Baik-Ben Arous-Péché (BBP) transition.

Thus the largest eigenvalue tends to the right endpoint of the asymptotic support of $H_N$. For some variants of $H_N$, it is known to have Tracy-Widom fluctuations around this point. For example, in the complex Gaussian case, which is determinantal, El Karoui [12] introduced an edge regularity condition and treated all $\Gamma$’s satisfying this condition, in the regime $\frac{M}{N} \geq 1$; Onatski [33] extended this to $\frac{M}{N} < 1$. In the direction of universality, Bao, Pan, and Zhou [3] allowed $Z$ to have complex non-Gaussian entries; Lee and Schnelli [26] allowed $Z$ to have real non-Gaussian entries, as long as $\Gamma$ is diagonal; Knowles and Yin [25] gave the (non-trivial) extension of this to non-diagonal $\Gamma$. Although we only treat the largest eigenvalue, it is also of statistical interest to consider the case when $\Gamma$ is such that $H_N$ has asymptotically disconnected support, and show that the rightmost eigenvalues for *each connected component* of the support each have Tracy-Widom fluctuations, which are independent of each other. Hachem, Hardy, and Najim [21] obtained such a result in the complex Gaussian case, followed by recent work of Fan and Johnstone [14] in the real Gaussian case. Finally, Wang [39] recently obtained the first non-trivial speed of convergence to the Tracy-Widom distribution, of order $N^{-1/57}$ (in the case $\Gamma = \mathrm{Id}$, Wang found a much better speed of $N^{-2/9}$, which Schnelli and Xu [35] recently improved to almost $N^{-1/3}$).[1]

### 1.4 Numerical examples

Although all the objects will be formally introduced later, we take a moment to give some numerical examples of our result. Later, we will assume that the empirical spectral measures of $\Gamma$ tend to some probability measure $\rho$, and recall that the empirical spectral measures of $H_N$ then tend to some probability measure $\sigma$ (see Theorem 2.4). We need two “branches” of the Stieltjes transform of this measure, both the ordinary one $G_\sigma$ and a less-common “second branch” $\widetilde{G}_\sigma$ (see Lemma 2.8). Ultimately our rate function is given as one half of the integral of the difference of these functions starting from the

[1] Many of the cited works are written for the model $\frac{1}{M}\Gamma^{1/2} Z^T Z \Gamma^{1/2}$, which of course has the same non-zero eigenvalues as our model, but requires $\Gamma \geq 0$.

right endpoint $r(\sigma)$, that is $I_\sigma(x) = \frac{1}{2}\int_{r(\sigma)}^{x} [\widetilde{G}_\sigma(u) - G_\sigma(u)]\,\mathrm{d}u$ (see Definition 2.14), which can fail to be analytic at the previously-mentioned $x_c$ threshold according to whether $\widetilde{G}_\sigma$ attains some value (in which case $x_c$ is this point) or is only asymptotic to it (in which case $x_c = +\infty$). For various choices of $M$, $N$, and $\Gamma$, Figures 1, 2, and 3 each show (i) a histogram of the eigenvalues of a single numerical sample of $H_N$ with Gaussian data; (ii) plots of $G_\sigma$ and $\widetilde{G}_\sigma$; and (iii) $x_c$ and $\theta_{\max}$ if present.

We direct the reader interested in more figures to the work of Maillard [27]; his Figure 1 is a roughly similar schematic to our figures, but where $\Gamma$ is a discretization of the Marčenko–Pastur law, and his Figure 2 contains various plots of rate functions.

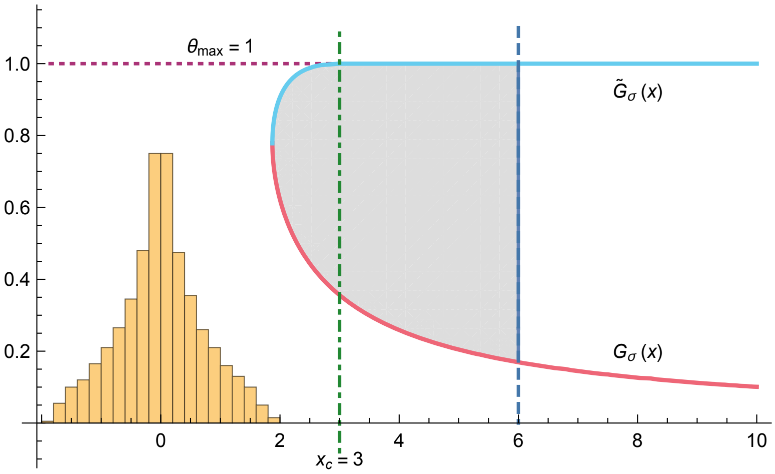

Figure 1: A plot when $N = 1000$, $M = 2000$, and $\Gamma$ is a discretization of the semicircle distribution supported on $[-2, 2]$. The histogram is the eigenvalues of one sample of $H_N$; the red curve is the Stieltjes transform $G_\sigma$; the cyan curve is the "second branch" $\widetilde{G}_\sigma$; the dot-dashed green line is the $x_c$ value, where $\widetilde{G}_\sigma$ attains its asymptote $\theta_{\max} = 1$ (dotted purple line). The rate function in our theorem is (one half) the integral of $\widetilde{G}_\sigma - G_\sigma$ from the right endpoint of the measure (here approximately $1.8$) up to $x$, i.e., at the sample value $x = 6$ (dashed blue line), the rate function is (one half) the area of the grey region. From this geometric description one can easily read off basic properties of the rate function, including that it is convex, increasing, and grows linearly at infinity. Made with Mathematica.

### 1.5 Organization

The paper is organized as follows: In Section 2, we define our model and the associated $x_c$ threshold, and state our main results. The proof is given in two stages: First for models with $x_c = +\infty$ (in Section 3), then for models with $x_c < +\infty$ (in Section 4) by approximation using $x_c = +\infty$ models. In Section 5 we outline the minor adjustments necessary if $H_N$ is complex Hermitian instead of real symmetric.

In Appendix A, we provide for completeness an extension of classical random-matrix-concentration results of Guionnet and Zeitouni to the setting of generalized sample covariance matrices. In Appendix B, we give a straightforward extension of Talagrand's classic results on concentration for product measures, but one that we were unable to find in the literature: His results are written for independent random variables valued in $[-1, 1]$, and we extend his results to independent random variables valued in the $d$-dimensional unit ball for any $d$. The quality of the estimate does not depend on $d$ (essentially because the radius of this ball does not depend on $d$), which may be of

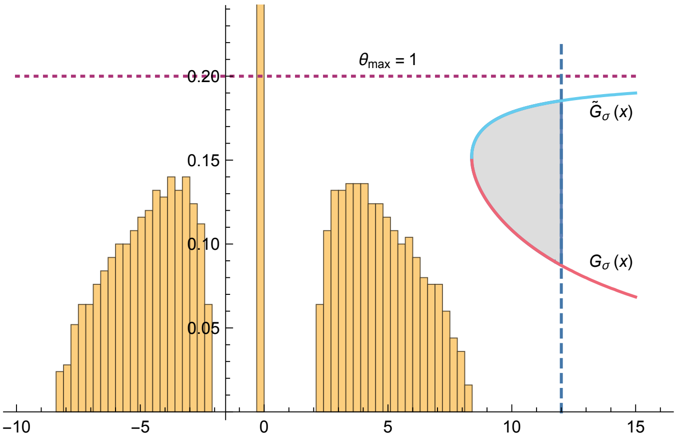

Figure 2: A plot when $N = 5000$, $M = 1000$, and $\Gamma$ is a discretization of $\frac{1}{2}(\delta_{-1} + \delta_{+1})$. This gives an example of the case when the limiting distribution is multicut (it is supported on two intervals), and when $x_c = +\infty$ (here, $\widetilde{G}_\sigma$ is asymptotic to $\theta_{\max}$ but does not reach it, unlike in Figure 1). Since the matrix is rank-deficient, there are quite a lot of eigenvalues at zero; for this visualization, we truncate the full height of that bar, which would otherwise extend well beyond the top of the figure. We also rescale the height of the histogram so that it is comparable to the values of $G_\sigma$ and $\widetilde{G}_\sigma$. Made with Mathematica.

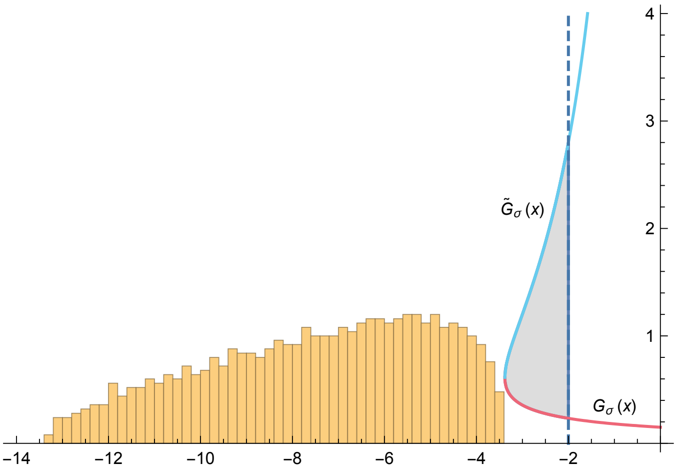

Figure 3: A plot when $N = 1000$, $M = 10000$, and $\Gamma$ is a discretization of $\frac{1}{2}(\delta_{-10} + \delta_{-5})$. This gives an example of the case when $r(\sigma) < 0$, where the rate function is finite only on $[r(\sigma), 0)$, and tends to infinity at zero. We also rescale the height of the histogram so that it is comparable to the values of $G_\sigma$ and $\widetilde{G}_\sigma$. Made with Mathematica.

independent interest. The particular case $d = 2$ allows us to consider complex random matrices, whose upper-triangular entries are independent of one another, but whose real and imaginary parts are simply *uncorrelated*, rather than *independent* as required in previous results. Finally, in Appendix C we show by example how to remove the “compact-or-log-Sobolev” assumption from works in the literature, and in Appendix D we state and prove our results on deformed Wigner matrices.

### 1.6 Notation

We write the Lipschitz and bounded-Lipschitz norms of a function $f : \mathbb{R} \to \mathbb{R}$, respectively, as

$$\|f\|_{\mathrm{Lip}} = \sup_{x \neq y} \left| \frac{f(x) - f(y)}{x - y} \right|,$$
$$\|f\|_{\mathcal{L}} = \|f\|_{\mathrm{Lip}} + \|f\|_{\infty}.$$

We will need the function classes

$$\mathcal{F}_{\mathrm{Lip}} = \{f : \mathbb{R} \to \mathbb{R} : \|f\|_{\mathcal{L}} \leq 1\}$$

and, for a given compact set $\mathcal{K} \subset \mathbb{R}$,

$$\mathcal{F}_{\mathrm{Lip},\mathcal{K}} = \{f \in \mathcal{F}_{\mathrm{Lip}} : \mathrm{supp}(f) \subset \mathcal{K}\}.$$

We will also need the bounded-Lipschitz, Wasserstein-1, and Kolmogorov-Smirnov distances between probability measures on $\mathbb{R}$, defined respectively by

$$\begin{aligned} d_{\mathrm{BL}}(\mu, \nu) &= \sup_{f \in \mathcal{F}_{\mathrm{Lip}}} \left| \int_{\mathbb{R}} f(x)(\mu - \nu)(\mathrm{d}x) \right|, \\ \mathrm{W}_1(\mu, \nu) &= \sup_{f : \|f\|_{\mathrm{Lip}} \leq 1} \left| \int_{\mathbb{R}} f(x)(\mu - \nu)(\mathrm{d}x) \right|, \\ d_{\mathrm{KS}}(\mu, \nu) &= \sup\{|\mu((-\infty, x]) - \nu((-\infty, x])| : x \in \mathbb{R}\}, \end{aligned} \tag{1.3}$$

the first of which metrizes weak convergence. If the probability measure $\mu$ is compactly supported, we write $G_\mu$ for its Stieltjes transform with the sign convention

$$G_\mu(z) = \int \frac{\mu(\mathrm{d}\lambda)}{z - \lambda},$$

and write $\ell(\mu)$ and $r(\mu)$ for its left and right endpoints, respectively.

We introduce the Dyson index $\beta$ as a shorthand for the symmetry class under consideration, either real-symmetric ($\beta = 1$) or complex-Hermitian ($\beta = 2$).

If $T$ is a matrix, we write its operator norm (with respect to Euclidean distance) as $\|T\|$, its Frobenius norm as $\|T\|_F^2 = \sum_{i,j} |T_{ij}|^2$, and its trace norm as $\|T\|_* = \sum_i \sigma_i(T)$. If $T$ is square and $N \times N$, we number its eigenvalues as

$$\lambda_{\min}(T) = \lambda_1(T) \leq \lambda_2(T) \leq \cdots \lambda_N(T) = \lambda_{\max}(T),$$

sometimes dropping the dependence on $T$ from the notation, and write its empirical spectral measure as

$$\hat{\mu}_T = \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i(T)}. \tag{1.4}$$

We write $\mathrm{Leb}(\cdot)$ for the Lebesgue measure, and in the appendices we will need the semicircle law normalized as

$$\rho_{\mathrm{sc}}(\mathrm{d}x) = \frac{\sqrt{(4 - x^2)_+}}{2\pi} \,\mathrm{d}x.$$

## 2 Results

### 2.1 Set-up

Theorem 2.15, our main result in the real case, is stated in Section 2.5; the complex analogue, Theorem 2.21, is given in Section 2.6. In order to state them properly, we need to precisely define both our model, in the present Section 2.1, as well as the limit $\sigma$ of its empirical measure and a particular function $\widetilde{G}_\sigma$, which is a sort of "second branch" of the Stieltjes transform needed to define the rate function for our LDP, in Section 2.3. In Section 2.4 we discuss certain degenerate cases.

We recall from (1.1) our fundamental random matrix

$$H_N = \frac{1}{M} Z^T \Gamma Z = \frac{1}{M} \sum_{i=1}^{M} d_i z_i z_i^T,$$

which we now define precisely. Fix $\alpha > 0$, and choose a sequence $M = (M_N)_{N=1}^{\infty}$ such that

$$\lim_{N\to\infty} \frac{M_N}{N} = \alpha.$$

For each $M$, we consider a deterministic real $M \times M$ matrix

$$\Gamma = \Gamma_M = \operatorname{diag}(d_1, \dots, d_M),$$

where $(d_1, \dots, d_M) = (d_1^{(M)}, \dots, d_M^{(M)})$ are ordered without loss of generality as $\lambda_{\min}(\Gamma) = d_1 \leq \dots \leq d_M = \lambda_{\max}(\Gamma)$. We suppose there exists a compactly supported probability measure $\rho$ such that

$$\hat{\mu}_{\Gamma_M} \to \rho \quad \text{weakly.} \tag{2.1}$$

We need the following definition.

**Definition 2.1.** *A centered probability measure $\mu$ on $\mathbb{R}$ is called* sharp sub-Gaussian *if it has unit variance and*

$$\int_{\mathbb{R}} e^{tx} \mu(\mathrm{d}x) \leq e^{\frac{t^2}{2}} \quad \textit{for all } t \in \mathbb{R}.$$

Standard Gaussian measure is sharp sub-Gaussian, but so are the Rademacher law $\frac{1}{2}(\delta_{-1} + \delta_{+1})$ and the uniform law on $[-\sqrt{3}, \sqrt{3}]$. We fix some sharp sub-Gaussian measure $\mu$, and let $Z \in \mathbb{R}^{M \times N}$ be a matrix with i.i.d. entries distributed according to $\mu$. To match the notation in the literature, we take the *rows* of $Z$, which are independent vectors of length $N$, and write them in *column* form as $\{z_i\}_{i=1}^{M}$.

The following assumption will be made throughout the paper, although we will only state it explicitly in the main theorem.

**Assumption 2.2.** *We assume that $\Gamma$ has asymptotically no (external) outliers, in the sense that*

$$\lambda_{\min}(\Gamma) \to \ell(\rho),$$
$$\lambda_{\max}(\Gamma) \to r(\rho).$$

We allow internal outliers, in the sense that, if $\rho$ has disconnected support, $\Gamma$ can have eigenvalues in the gaps between the components.

### 2.2 Spherical integrals

One fundamental object in this paper is a *(rank-one) spherical integral*, defined for (deterministic) real-symmetric $N \times N$ matrices $A$ and parameters $\theta \geq 0$ by

$$\mathbb{E}_e[e^{N\theta\langle e, Ae\rangle}],$$

where $\mathbb{E}_e$ means that $e$ should be taken uniform on the unit sphere in $\mathbb{R}^N$. Often we will take $A$ random, in which case we will always take $e$ to be independent of $A$. There is also a complex analogue, where $e$ is uniform on the unit sphere in $\mathbb{C}^N$; we will mostly abuse notation by writing $\mathbb{E}_e$ for both the real and complex case, but the underlying field will always match that of the matrix argument (i.e., $e$ is real/complex exactly when $A$ is real/complex); in particular, we will mostly discuss the real case until Section 5.

A rank-one spherical integral is a special case of the Harish-Chandra/Itzykson/Zuber (HCIZ) integral, defined for real-symmetric $N \times N$ matrices $A$ and $B$ by

$$\int_{\mathcal{O}_N} e^{N \operatorname{tr}(AOBO^T)} \, \mathrm{d}O,$$

where $\mathrm{d}O$ is the Haar measure on the group $\mathcal{O}_N$ of $N \times N$ orthogonal matrices, when one takes $B = \operatorname{diag}(\theta, 0, \dots, 0)$. The HCIZ integral is so named because Harish-Chandra, and independently Itzykson and Zuber, gave exact formulas to evaluate it at finite $N$. However, these beautiful formulas are not particularly suitable for large-$N$ asymptotics when $A = A_N$ and $B = B_N$. When $A_N$ and $B_N$ have full rank, these asymptotics were first predicted by Matytsin in the physics literature [30], then established in the mathematical literature by Guionnet and Zeitouni [20]. The full-rank asymptotics are quite complicated, but it turns out that the rank-one special case has much simpler asymptotics: Informally $\mathbb{E}_e[e^{N\theta\langle e, A_N e\rangle}] \asymp e^{NJ(\hat{\mu}_{A_N}, \theta, \lambda_{\max}(A_N))}$, where $J$ is a simple special function given in Definition 3.1 below. This result is originally due to Guionnet and Maïda [15], but for the proof it will be more convenient to use a technical strengthening of their result (roughly, with better uniformity properties in the arguments) given recently by Guionnet and the first author [17].

### 2.3 The Dyson equation and the limit of the empirical measure

The following two thresholds will be important:

**Definition 2.3.** *Let*

$$\theta_{\max} := \begin{cases} \frac{\alpha}{r(\rho)} & \text{if } r(\rho) > 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{2.2}$$

$$x_c(\rho) := \begin{cases} r(\rho)^2 G_\rho(r(\rho)) + \left(\frac{1}{\alpha} - 1\right) r(\rho) & \text{if } r(\rho) > 0 \text{ and } G_\rho(r(\rho)) < \infty, \\ +\infty & \text{otherwise} \end{cases} \tag{2.3}$$

With such a setup, the following convergence theorem for the empirical measure $\hat{\mu}_{H_N}$ is essentially classical:

**Theorem 2.4.** *With $H_N$ as above, the sequence of empirical measures $(\hat{\mu}_{H_N})_{N=1}^\infty$ converges weakly in probability toward a compactly supported measure $\sigma$. Furthermore, the Stieltjes transform of $\sigma$ on $(r(\sigma), +\infty)$ satisfies*

$$H_\rho(G_\sigma(x)) = x, \tag{2.4}$$

*where $H_\rho$ is defined on $(0, \theta_{\max})$ by*

$$H_\rho(y) := \frac{1}{y} + \int_{\mathbb{R}} \frac{\alpha u}{\alpha - yu} \rho(\mathrm{d}u). \tag{2.5}$$

*If $r(\rho) \leq 0$, then $r(\sigma) \leq 0$. (However, it is possible to have $r(\sigma) < 0$ when $r(\rho) > 0$; see Remark 2.7 for an example.)*

**Remark 2.5.** Although we will not need it, we remark that the restriction of $G_\sigma$ to $(r(\sigma), +\infty)$ has range exactly $(0, \theta_c)$, where $\theta_c$ defined in Section 3.7 below satisfies $\theta_c \leq \theta_{\max}$.

*Proof.* Theorem 1.1 in [36] gives every claim, other than that the measure $\sigma$ is compactly supported, and that $r(\sigma) \leq 0$ when $r(\rho) \leq 0$. To show the former, note that, since $\rho$ is compactly supported, then $H_\rho$ is meromorphic in a neighborhood of zero, with a simple pole at zero. From the inverse function theorem, $z \mapsto \frac{1}{H_\rho(z)}$ is therefore holomorphic in a neighborhood of zero, so (via (2.5)) $G_\sigma$ is holomorphic in a neighborhood of infinity, meaning $\sigma$ is indeed compactly supported. To show the latter, note that we can always choose $\Gamma$ without outliers, meaning in such a way that $\ell(\hat{\mu}_\Gamma) = \ell(\rho)$ and $r(\hat{\mu}_\Gamma) = r(\rho)$ at finite $N$, and that for $H_N$ defined with such $\Gamma$ we have $H_N \leq r(\rho) \leq 0$, so that $r(\sigma) \leq 0$. □

**Remark 2.6.** We will not need it, but the measure $\sigma$ can be interpreted in the language of free probability. Indeed, when $\rho$ is supported on $\mathbb{R}^+$ or $\mathbb{R}^-$, then $\sigma$ is just the free multiplicative convolution of $\rho$ and the corresponding Marčenko-Pastur law. The free multiplicative convolution is usually defined only for measures supported on a half-line, but when $\rho$ has two nontrivial components $\rho^\pm(A) = \rho(A \cap \mathbb{R}^\pm)$, we can write $\sigma$ as the free *additive* convolution of the measures $\sigma^\pm$, which are respectively the free *multiplicative* convolutions of $\rho^\pm$ with the Marčenko-Pastur law.

Indeed, writing $\rho = \rho^+ + \rho^-$ induces a decomposition $\Gamma = \Gamma^+ + \Gamma^-$ into positive and negative $d_i$'s, and thus a decomposition $H_N = H_N^+ + H_N^-$ with $H_N^\pm = \frac{1}{M} Z^T \Gamma^\pm Z$. The matrices $H_N^+$ and $H_N^-$ are independent, and their empirical measures tend to $\sigma^\pm$, respectively. This is true regardless of the underlying law of $Z$, which we will choose to be Gaussian; then, due to rotation invariance of the Gaussian law, they have the same joint distribution as $H_N^+$ and $OH_N^-O^T$, where $O$ is a Haar orthogonal/unitary matrix independent of everything else; and since asymptotically these matrices are *freely* independent, the empirical measure of their sum tends to the free *additive* convolution of the empirical measures.

**Remark 2.7.** Note that we can have $r(\rho) > 0$ but $r(\sigma) < 0$. For instance, let us choose $M_N = 4N$ and let us take

$$\Gamma_N = \text{diag}(\underbrace{2, \dots, 2}_{2N \text{ times}}, \underbrace{-2K, \dots, -2K}_{2N \text{ times}})$$

where $K$ is a constant we will fix later. Then we have for this model $\rho = \frac{1}{2}(\delta_{-2K} + \delta_2)$. Remark 2.6 explains that, in this case, $\sigma$ is the free *additive* convolution of two measures, namely the free *multiplicative* convolutions of delta masses at 2 and $-2K$, respectively, with Marčenko-Pastur. Since the free multiplicative convolution in this case just rescales the measures, $\sigma$ is the free additive convolution of two stretched Marčenko-Pasturs, one near 2 and one near $-2K$. The factors of two here are so that the Marčenko-Pastur laws are gapped away from zero; since $r(\mu \boxplus \nu) \leq r(\mu) + r(\nu)$ in general, by taking $K$ sufficiently large, we can thus force $r(\sigma) < 0$.

The equation (2.4) is called a *Dyson equation*. The following lemma defines a certain function $\widetilde{G}_\sigma$, which should be thought of as a second branch of the Stieltjes transform. Its proof will be given in Section 3.7.

**Lemma 2.8.** *First, if $x_c(\rho)$ is finite then $x_c(\rho) = H_\rho(\theta_{\max}) := \lim_{y \uparrow \theta_{\max}} H_\rho(y)$, and $x_c(\rho) \geq r(\sigma)$. Second, except in the case when $r(\rho) \leq 0$ and $r(\sigma) = 0$ (this degenerate case is handled in Section 2.4), there exists a real-valued, continuous function $\widetilde{G}_\sigma$, defined on the domain*

$$D := \begin{cases} [r(\sigma), +\infty) & \textit{if } r(\rho) > 0, \\ [r(\sigma), 0) & \textit{if } r(\rho) \leq 0, \end{cases} \tag{2.6}$$

*with the following properties:*

1. *If $x \in D$ and $x \leq x_c(\rho)$, then $\{w : H_\rho(w) = x\} = \{G_\sigma(x), \widetilde{G}_\sigma(x)\}$ (in particular, $H_\rho(\widetilde{G}_\sigma(x)) = x$).*
2. *We have $\widetilde{G}_\sigma(x) > G_\sigma(x)$ for $x > r(\sigma)$, and $\widetilde{G}_\sigma(r(\sigma)) = G_\sigma(r(\sigma))$.*
3. *$\widetilde{G}_\sigma$ is real analytic on $D \setminus \{x_c(\rho)\}$.*
4. *$\widetilde{G}_\sigma$ is nondecreasing on $D$, and more specifically*
   - *If $r(\rho) > 0$ and $x_c(\rho) = +\infty$, then $\widetilde{G}_\sigma$ is strictly increasing on $D$, with $\lim_{x\to+\infty} \widetilde{G}_\sigma(x) = \theta_{\max}$.*
   - *If $r(\rho) > 0$ and $x_c(\rho) < +\infty$, then $\widetilde{G}_\sigma$ is strictly increasing on $(r(\sigma), x_c(\rho))$, with $\lim_{x\uparrow x_c(\rho)} \widetilde{G}_\sigma(x) = \theta_{\max}$, and $\widetilde{G}_\sigma(x) = \theta_{\max}$ for $x \geq x_c(\rho)$.*
   - *If $r(\rho) < 0$, then $\widetilde{G}_\sigma$ is strictly increasing on $D$, with $\lim_{x\uparrow 0} \widetilde{G}_\sigma(x) = +\infty$.*

### 2.4 Degenerate cases

For our purposes, there are two possibilities for degenerate behavior. We first explain them informally:

1. If $\rho(\{0\}) > 0$, this means that one is writing
$$H_N = \frac{1}{M}\sum_{i=1}^{M} d_i z_i z_i^T$$
where a macroscopic fraction of the $d_i$'s are (asymptotically) zero. Our results do apply to this case as written, but as a coherence check, we confirm in Remark 2.10 below that the rate function remains the same if one instead discards these zero $d_i$'s, which amounts to removing the zero atom from $\rho$, renormalizing to keep it a probability measure, and adjusting $\alpha$ correspondingly.

2. The case $r(\rho) \leq 0$ and $r(\sigma) = 0$ is degenerate, since then $\lambda_{\max}(H_N)$ is essentially trapped at zero: On the one hand, $\lambda_{\max}(H_N)$ cannot push into the bulk at this scale, so it must be asymptotically nonnegative. On the other hand $H_N$ is (asymptotically almost) negative semidefinite, so $\lambda_{\max}(H_N)$ must be asymptotically nonpositive. This is expressed precisely in a degenerate LDP.

We now formalize the results just explained. All the claims here are proved in Section 3.6.

**Lemma 2.9.** *If $\rho$ is a measure on $\mathbb{R}$ such that $r(\rho) \leq 0$ and $\rho(\{0\}) = 0$, then the following are equivalent:*

- *$r(\sigma) = 0$,*
- *$\alpha \leq 1$.*

**Remark 2.10.** If $\rho(\{0\}) > 0$, then we claim that the following two procedures give the same result (meaning the same $\sigma$ and the same rate function): (a) applying our results to $\rho$ as written, or (b) creating a new measure $\tau$ that removes the spike at zero, changes $\alpha$ to some $\alpha'$, and considers a corresponding $H'_N$.

Indeed, if $\rho(\{0\}) = \beta > 0$, we can write $\rho = (1-\beta)\rho' + \beta\delta_0$ where $\rho'$ is a measure with $\rho'(\{0\}) = 0$ (and, if $r(\rho) \leq 0$, then $r(\rho') \leq 0$). Then it is easy to see that $H_N = H'_N + H''_N$, where $H'_N = Z'^T\Gamma'_N Z'$ and $H''_N = Z''^T\Gamma''_N Z''$, where $\Gamma'_N$ is a $M'_N \times M'_N$ matrix with $\lim_N \frac{M'_N}{N} = \alpha(1-\beta)$, and where the empirical measure of $\Gamma'_N$ converges toward $\rho'$ and $||\Gamma''_N||$ converges to $0$. So for $\epsilon > 0$, $\lim \frac{1}{N}\log \mathbb{P}[||H''_N|| \geq \epsilon] = -\infty$. This reduces the problem to stating a large deviation principle for $H'_N$, and furthermore, if $r(\rho) \leq 0$, we proved that $r(\sigma) = 0$ if and only if $\alpha(1-\beta) \leq 1$ (we state this result in the following corollary). With $\Delta_N := \frac{M'_N}{M_N}\Gamma'_N$, we can rewrite $H'_N$ as $H'_N = \frac{1}{M'_N} Z'^T \Delta_N Z'$. Since the empirical measure of $\Delta_N$ converges toward $\tau := m_{1-\beta}\sharp\rho'$, where $m_{1-\beta}\sharp\rho'$ is the pushforward of $\rho'$ by multiplication by $1-\beta$ (i.e., $m_{1-\beta}(x) = (1-\beta)x$), we can apply our main result, Theorem 2.15, to $H'_N$ with $\tau$ instead of $\rho$ and $\alpha' := \alpha(1-\beta)$ instead of $\alpha$. It remains to show then that the statement of Theorem 2.15 remains unchanged when we change $\rho$ into $\rho'$ and $\alpha$ into $\alpha'$. For this we need only to show that the functions $H_\rho$ and $H_\tau$ are the same, since we will see that the rate function of the large deviation principle only depends on $\rho$ through $H_\rho$. Indeed,

$$\begin{aligned} H_\rho(y) &= \frac{1}{y} + \int \frac{\alpha u}{\alpha - yu}\rho(du) = \frac{1}{y} + (1-\beta)\int \frac{\alpha u}{\alpha - yu}\rho'(du) \\ &= \frac{1}{y} + (1-\beta)\int \frac{\alpha(1-\beta)^{-1}u}{\alpha - (1-\beta)^{-1}yu}\tau(du) = \frac{1}{y} + \int \frac{(1-\beta)\alpha u}{(1-\beta)\alpha - yu}\tau(du) = H_\tau(y). \end{aligned}$$

Therefore the rate function we obtain by applying Theorem 2.15 to $H'_N$ is the same as what we obtain by applying Theorem 2.15 to $H_N$.

**Corollary 2.11.** *If $\rho$ is a measure on $\mathbb{R}$ such that $r(\rho) \leq 0$, then the following are equivalent:*

- $r(\sigma) = 0$,
- $\alpha(1 - \rho(\{0\})) \leq 1$.

**Definition 2.12.** *We will summarize the (equivalent) conditions of Corollary 2.11 by saying that the pair $(\rho, \alpha)$ is* degenerate.

This definition is justified by the following (straightforward) result.

**Proposition 2.13.** *Define $I^{\text{degen}} : \mathbb{R} \to [0, +\infty]$ by*

$$I^{\text{degen}}(x) = \begin{cases} 0 & \text{if } x = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

*If $\rho(\{0\}) = 0$, and the pair $(\rho, \alpha)$ is degenerate, then $\lambda_{\max}(H_N)$ satisfies a large deviation principle at speed $N$ with the good rate function $I^{\text{degen}}$.*

In the following, we will always assume that the pair $(\rho, \alpha)$ is nondegenerate.

### 2.5 Main result in the real setting

**Definition 2.14.** *Suppose the pair $(\rho, \alpha)$ is nondegenerate. With $D$ as in* (2.6)*, $G_\sigma$ the Stieltjes transform as usual, and $\widetilde{G}_\sigma$ the "second branch" of the Stieltjes transform from Lemma 2.8, define $I_\sigma : D \to [0, +\infty]$ by*

$$I_\sigma(x) = \begin{cases} \frac{\beta}{2}\int_{r(\sigma)}^{x} [\widetilde{G}_\sigma(u) - G_\sigma(u)]\,\mathrm{d}u & \text{if } x \in D, \\ +\infty & \text{otherwise.} \end{cases}$$

Our main theorem holds under either of the following two assumptions, recalling that $\mu$ is the common distribution of the entries of $Z$.

**Assumption A.** The measure $\mu$ is sharp sub-Gaussian, and the support of $\rho$ is in $[0, \infty)$.

**Assumption B.** The measure $\mu$ is Gaussian.

**Theorem 2.15** (Main theorem, real version)**.** *If Assumption 2.2 holds, the pair $(\rho, \alpha)$ is nondegenerate (in the sense of Definition 2.12), and also either Assumption A or Assumption B holds, then $\lambda_{\max}(H_N)$ satisfies a large deviation principle at speed $N$*

*with the good rate function $I_\sigma$. This function is convex and strictly increasing on $D$ (in particular, it vanishes uniquely at $r(\sigma)$). If additionally $r(\rho) > 0$, then*

$$\lim_{x \to +\infty} \frac{I_\sigma(x)}{x} = \frac{\theta_{\max}}{2}.$$

*If $\mu$ is actually Gaussian (Assumption B), then by rotational invariance we do not need to assume that $\Gamma$ is diagonal; we can just assume it is real symmetric and satisfies eq.* (2.1) *and Assumption 2.2.*

We remark that the claimed properties of $I_\sigma$ follow immediately from its definition and Lemma 2.8, which write the rate function in the form $I_\sigma(x) = \frac{1}{2}\int_{r(\sigma)}^{x} g(y)\,\mathrm{d}y$, where $g$ is strictly increasing, positive for arbitrarily small arguments, and $\lim_{x\to+\infty} \frac{g(x)}{\theta_{\max}} = 1$.

Theorem 2.15 will follow from the following two results.

**Proposition 2.16.** *If Assumption 2.2 holds, the pair $(\rho, \alpha)$ is nondegenerate, and also either Assumption A or Assumption B holds, and*

$$x_c(\rho) = +\infty,$$

*then $\lambda_{\max}(H_N)$ satisfies a large deviation principle at speed $N$ with the good rate function $I_\sigma$.*

**Proposition 2.17.** *Proposition 2.16 implies Theorem 2.15.*

**Remark 2.18.** Propositions 2.16 and 2.17 have fairly different proofs from one another. Proposition 2.16 is proved by tilting with spherical integrals; for every $x < x_c(\rho)$, we are able to find appropriate tilt that makes the deviations $\{\lambda_{\max} \approx x\}$ likely. But Proposition 2.17 uses neither explicit tilting, nor almost anything else in the proof details of Proposition 2.16; instead it goes by approximating models with $x_c(\rho) < +\infty$ using models with $x_c(\rho) = +\infty$, and textbook results on obtaining LDPs by taking limits in a sequence of approximating LDPs.

**Remark 2.19.** In the non-Gaussian case, the requirement that $\rho$ be supported in $(0, \infty)$ is not just technical. When some $d_i$'s are negative, the rate function should likely be different from the Gaussian case. Indeed, suppose all the $d_i$'s are equal to $-d$ for some $d > 0$. Then of course $\lambda_{\max}(-\frac{d}{M}Z^TZ) = -\lambda_{\min}(\frac{d}{M}Z^TZ)$, but (except in degenerate cases) the left-hand deviations of the smallest eigenvalue of a Rademacher covariance matrix are not the same as the Gaussian analogue: On the one hand, in the Gaussian case, the limiting empirical measure of $\frac{d}{M}Z^TZ$ is a stretched version of the Marčenko–Pastur law, $\sigma(\mathrm{d}x) = \frac{\sqrt{((x-\ell_{d,\alpha})(r_{d,\alpha}-x))_+}}{2\pi\frac{d}{\alpha}x}$, where $\ell_{d,\alpha} = d(1-\frac{1}{\sqrt{\alpha}})^2$ and $r_{d,\alpha} = d(1+\frac{1}{\sqrt{\alpha}})^2$, and the rate function for the left tail of the smallest eigenvalue in this Gaussian case is given by

$$I_\sigma(x) = \frac{1}{2}\int_x^{\ell_{d,\alpha}} \frac{\sqrt{(\ell_{d,\alpha}-y)(r_{d,\alpha}-y)}}{\frac{d}{\alpha}y}\,\mathrm{d}y,$$

which tends to infinity as $x \downarrow 0$. (One can obtain this as a special case of our main result, or by Coulomb-gas arguments, see e.g. [24].) On the other hand, the Rademacher rate function must be finite at zero since $\mathbb{P}(\text{two columns of } Z \text{ agree}) \geq 2^{-M}$. (The rank-deficient case $\alpha \leq 1$ is degenerate in the sense of Definition 2.12; in this case, for both the Gaussian and Rademacher versions of $\frac{d}{M}Z^TZ$, the smallest eigenvalue satisfies an LDP at speed $N$ with the degenerate rate function $I^{\text{degen}}$ from Proposition 2.13, so the rate functions match but for a trivial reason.)

### 2.6 Main result in the complex setting

Our main result also translates to the complex setting. In this, though, we need to take the entries of $Z$ to be i.i.d. distributed according the some centered probability measure

$\mu$ on the complex plane such that $\int(\Im z)^2\mu(dz) = \int(\Re z)^2\mu(dz) = 1/2$ and $\int \Im z \Re z \mu(dz) = 0$. Our model then becomes

$$H_N = \frac{1}{M} Z^* \Gamma Z$$

where $Z^*$ denotes the Hermitian conjugate of $Z$. All the other assumptions on $M = M_N$ and $\Gamma$ remain the same.

We can extend Definition 2.1 to complex random variables:

**Definition 2.20.** *A centered probability measure $\mu$ on $\mathbb{C}$ is called* sharp sub-Gaussian in $\mathbb{C}$ *if for $X$ $\mu$-distributed, the random vector $(\Re X, \Im X)$ has covariance matrix $\frac{1}{2}\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)$ (the real and imaginary parts must be uncorrelated, but do not have to be independent) and*

$$\int_{\mathbb{R}} e^{\Re(w\overline{z})}\mu(\mathrm{d}w) \le e^{\frac{|z|^2}{4}} \quad \textit{for all } z \in \mathbb{C}.$$

We need then to slightly update Assumption A to our complex setting:

**Assumption C.** The measure $\mu$ is sharp sub-Gaussian in $\mathbb{C}$, and the support of $\rho$ is in $[0, \infty)$.

Then we have for this model an LDP exactly identical to the one we obtain in the real case, except the rate function we have here is twice the rate function of the real case.

**Theorem 2.21** (Main theorem, complex version)**.** *If Assumption 2.2 holds, the pair $(\rho, \alpha)$ is nondegenerate, and also either Assumption B or Assumption C holds, then $\lambda_{\max}(H_N)$ satisfies a large deviation principle at speed $N$ with the good rate function $2I_\sigma$.*

*If $\mu$ is actually Gaussian (Assumption B), then by rotational invariance we do not need to assume that $\Gamma$ is diagonal; we can just assume it is complex Hermitian and satisfies eq.* (2.1) *and Assumption 2.2.*

The proof remains mostly the same as in the real case, up to some tweaks in the computations of our spherical integrals; we describe the needed adjustments precisely in Section 5.

# 3 Proof for infinite $x_c$

## 3.1 Outline of the proof

In this section, we outline the proof of Proposition 2.16. The proof goes by introducing a variational formulation of the rate function, which is more opaque but technically more convenient, and then proving the LDP with this variational formulation via tilting by spherical integrals. The broad sketch of the proof in this section resembles previous works, but as explained in the introduction, new arguments allow us to bypass several technical assumptions present in previous works. The assumption $x_c = \infty$ is crucial for the large deviation lower bound (Proposition 3.9). Indeed, it allows one to choose, for each $x \ge r(\sigma)$, a parameter $\theta_x$ such that under the tilted measure $\mathbb{P}^{\theta_x}$ defined in Definition 3.7, $\lambda_{\max}(H_N)$ converges to $x$. Without this assumption, we would have to choose $\theta_x = \theta_{\max}$ for every $x \ge x_c$, and our proof strategy then breaks down.

**Definition 3.1.** *For $\mu$ a compactly supported probability measure on $\mathbb{R}$, $\theta \ge 0$, and $\lambda \ge r(\mu)$, let*

$$v(\mu, \theta, \lambda) := \begin{cases} \lambda - \frac{1}{2\theta} & \textit{if } G_\mu(\lambda) \le 2\theta, \\ G_\mu^{-1}(2\theta) - \frac{1}{2\theta} & \textit{if } G_\mu(\lambda) \ge 2\theta \ge 0, \end{cases}$$

$$J(\mu, \theta, \lambda) := \theta v(\mu, \theta, \lambda) - \frac{1}{2}\int_{\mathbb{R}} \log(1 + 2\theta v(\mu, \theta, \lambda) - 2\theta y)\mu(\mathrm{d}y),$$

**Definition 3.2.** *For* $0 \le \theta < \theta_{\max}$*, let*

$$F(\rho,\theta) := -\frac{\alpha}{2}\int_{\mathbb{R}} \log\left(1-\frac{\theta t}{\alpha}\right)\rho(\mathrm{d}t),$$
$$I_\sigma(x,\theta) := J\left(\sigma,\frac{\theta}{2},x\right) - F(\rho,\theta).$$

*Using these, define* $\widetilde{I_\sigma} : \mathbb{R} \to [0,+\infty]$ *by*

$$\widetilde{I_\sigma}(x) := \begin{cases} \sup_{0\le\theta<\theta_{\max}} I_\sigma(x,\theta) & \textit{if } x \in D, \\ +\infty & \textit{otherwise.} \end{cases}$$

**Lemma 3.3.** ***(Simplification of the rate function)*** *Assume* $(\rho,\alpha)$ *is nondegenerate. If* $x_c(\rho) = +\infty$*, we have*

$$I_\sigma(x) = \widetilde{I_\sigma}(x),$$

*and this function is convex on the set* $D$ *defined in* (2.6) *and vanishes uniquely at* $x = r(\sigma)$*. Furthermore, for every* $x \in D^\circ$ *(the interior of* $D$*) there exists a unique* $\theta_x$ *with* $0 \le \theta_x < \theta_{\max}$ *such that*

$$\widetilde{I_\sigma}(x) = \sup_{0\le\theta<\theta_{\max}} I_\sigma(x,\theta) = I_\sigma(x,\theta_x).$$

*In fact one can take*

$$\theta_x = \widetilde{G}_\sigma(x),$$

*and the map* $x \mapsto \theta_x$ *is injective.*

*Proof.* We compute

$$\frac{\partial}{\partial\theta} J\left(\sigma,\frac{\theta}{2},x\right) = \begin{cases} \frac{G_\sigma^{-1}(\theta)}{2} - \frac{1}{2\theta} & \text{if } \theta \le G_\sigma(x), \\ \frac{x}{2} - \frac{1}{2\theta} & \text{if } \theta \ge G_\sigma(x). \end{cases}$$

We also compute

$$\frac{\partial}{\partial\theta} F(\rho,\theta) = \frac{1}{2}\int_{\mathbb{R}} \frac{\alpha u}{\alpha-\theta u} d\rho(u) = \frac{H_\rho(\theta)}{2} - \frac{1}{2\theta}.$$

Using the Dyson equation (2.4), we have

$$\frac{\partial}{\partial\theta} I_\sigma(x,\theta) = \begin{cases} 0 & \text{if } \theta \le G_\sigma(x), \\ \frac{x}{2} - \frac{H_\rho(\theta)}{2} & \text{if } \theta \ge G_\sigma(x). \end{cases}$$

Thus the function $I_\sigma(x,\cdot)$ is increasing on $[G_\sigma(x), \widetilde{G}_\sigma(x)]$ and decreasing on $[\widetilde{G}_\sigma(x), +\infty)$. Therefore $\theta_x := \widetilde{G}_\sigma(x)$ is the optimizing $\theta$ value, i.e., $\widetilde{I_\sigma}(x) = I_\sigma(x,\theta_x)$. (When $x = r(\sigma)$, we define $\theta_{r(\sigma)} := \theta_c$ by convention, and note that this argument shows $\widetilde{I_\sigma}(r(\sigma)) = 0$.) Thus $\partial_\theta I_\sigma(x,\theta)|_{\theta=\theta_x} = 0$, which lets us compute the total derivative as

$$\frac{\mathrm{d}}{\mathrm{d}x}\widetilde{I_\sigma}(x) = \frac{\partial}{\partial x} I_\sigma(x,\theta)\bigg|_{\theta=\theta_x} = \frac{1}{2}(\theta_x - G_\sigma(x)) = \frac{1}{2}(\widetilde{G}_\sigma(x) - G_\sigma(x)).$$

(This is a general strategy to show that functions of the form $f(x) = \sup_{y\in\mathbb{R}} g(x,y) = g(x,y_x)$ have derivative $f'(x) = \partial_x g(x,y)|_{y=y_x}$.) Since $I_\sigma$ and $\widetilde{I_\sigma}$ have the same derivative and agree at $x = r(\sigma)$, they match as functions. As explained before, convexity follows since the derivative is strictly increasing, and injectivity follows from point 4 of Lemma 2.8. □

**Remark 3.4.** This is not necessary for our proof, but we note that when $x_c(\rho) < +\infty$, i.e. $r(\rho) > 0$ and $G_\rho(r(\rho)) < \infty$, one can check that $\int_{\mathbb{R}} \log(r(\rho) - t)\rho(\mathrm{d}t) > -\infty$, hence one can make sense of $F(\rho, \theta_{\max})$ and $I_\sigma(x, \theta_{\max})$. In this case, one can define $I_\sigma^\dagger : \mathbb{R} \to [0, +\infty]$ by

$$I_\sigma^\dagger(x) := \begin{cases} I_\sigma(x, \widetilde{G}_\sigma(x)) & \text{if } x \geq r(\sigma) \\ +\infty & \text{otherwise} \end{cases} = \begin{cases} \sup_{0 \leq \theta < \theta_{\max}} I_\sigma(x, \theta) & \text{if } r(\sigma) \leq x < x_c(\rho), \\ I_\sigma(x, \theta_{\max}) & \text{if } x \geq x_c(\rho), \\ +\infty & \text{otherwise.} \end{cases}$$

The point of this remark is that one can check

$$I_\sigma(x) = I_\sigma^\dagger(x).$$

Indeed, the proof of Lemma 3.3 already checked this for $x < x_c(\rho)$. For $x \geq x_c(\rho)$, one can compute

$$\partial_x I_\sigma(x, \theta_{\max}) = \partial_x J(\sigma, \theta_{\max}/2, x) = \frac{1}{2}(\theta_{\max} - G_\sigma(x)) = \frac{1}{2}(\widetilde{G}_\sigma(x) - G_\sigma(x)),$$

meaning that $I_\sigma$ and $I_\sigma^\dagger$ have the same derivative. However, the formulation of $I_\sigma$ in Definition 2.14 is technically more convenient, so in the proof we use $I_\sigma$ rather than $I_\sigma^\dagger$.

Assumptions 2.2 and A require $\mathrm{supp}(\rho) \subset [0, \infty)$ but permit a handful of negative $d_i$'s at finite $N$, which must tend to zero in the limit of large dimension. However, it is technically more convenient to work with the case when all $d_i$'s are nonnegative at finite $N$. The following result allows us to restrict to this case; we omit its proof, since it is essentially the same as that of Proposition 2.13.

**Lemma 3.5.** *Under Assumptions 2.2 and A, define*

$$\Gamma^+ = \mathrm{diag}(\max(d_1, 0), \ldots, \max(d_M, 0)),$$

*and $H_N^+ = M^{-1} Z^T \Gamma^+ Z$. Then $\lambda_{\max}(H_N)$ and $\lambda_{\max}(H_N^+)$ are exponentially equivalent. In particular, LDPs for one automatically hold for the other.*

In the remainder, we tacitly replace $H_N$ with $H_N^+$ when necessary.

**Lemma 3.6.** ***(Exponential tightness)*** *Under either Assumption A or Assumption B, we have*

$$\lim_{K \to \infty} \limsup_{N \to \infty} \frac{1}{N} \log \mathbb{P}(|\lambda_{\max}(H_N)| > K) = -\infty.$$

*Proof.* If $M$ is large enough that $d_i \in (\ell(\rho) - 1, r(\rho) + 1)$ for all $i$, then it is elementary that

$$|\lambda_{\max}(H_N)| \leq \max(|\ell(\rho) - 1|, |r(\rho) + 1|)\lambda_{\max}(M^{-1} Z^T Z).$$

Exponential tightness of $\lambda_{\max}(M^{-1} Z^T Z)$ was proved in [16, Lemma 1.9]. □

**Definition 3.7.** ***(Tilted measures)*** *For $\theta \geq 0$, consider the tilted measure $\mathbb{P}^\theta(Z)$ on $M \times N$ matrices with density*

$$\frac{\mathrm{d}\mathbb{P}^\theta}{\mathrm{d}\mathbb{P}}(Z) = \frac{\mathbb{E}_e[e^{N \frac{\theta}{2} \langle e, \frac{1}{M} Z^T \Gamma Z e \rangle}]}{\mathbb{E}_{e, H_N}[e^{N \frac{\theta}{2} \langle e, H_N e \rangle}]}.$$

**Proposition 3.8.** ***(Weak LDP upper bound for tilted measures)*** *For $0 \leq \theta < \theta_{\max}$,*

$$\limsup_{\delta \downarrow 0} \limsup_{N \to \infty} \frac{1}{N} \log \mathbb{P}^\theta(|\lambda_{\max}(H_N) - x| \leq \delta) \begin{cases} \leq -(\widetilde{I}_\sigma(x) - I_\sigma(x, \theta)) & \text{if } x \in D, \\ = -\infty & \text{otherwise.} \end{cases}$$

*Notice that $\mathbb{P}^0 = \mathbb{P}$, and $I_\sigma(x, 0) = 0$, so in particular we have the weak LDP upper bound for the measure we care about.*

**Proposition 3.9.** ***(Weak LDP lower bound)***

$$\liminf_{\delta\downarrow 0}\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}(|\lambda_{\max}(H_N)-x|<\delta)\geq -\widetilde{I_\sigma}(x).$$

Proposition 2.16 follows in the classical way[2] from Lemma 3.3, Lemma 3.6, Proposition 3.8, and Proposition 3.9.

**Remark 3.10.** Actually, the proofs of Propositions 3.8 and 3.9 are slightly more general than just stated: They work "up to $x_c(\rho)$" in the sense that they show

$$\begin{aligned}-I_\sigma(x)&\leq\liminf_{\delta\downarrow 0}\liminf_{N\to+\infty}\frac{1}{N}\log\mathbb{P}(|\lambda_{\max}(H_N)-x|\leq\delta)\\&\leq\limsup_{\delta\downarrow 0}\limsup_{N\to+\infty}\frac{1}{N}\log\mathbb{P}(|\lambda_{\max}(H_N)-x|\leq\delta)\leq -I_\sigma(x)\end{aligned}$$

whenever $x<x_c(\rho)$ (and actually the upper bound works even for $x>x_c(\rho)$), not just whenever $x_c(\rho)=+\infty$ and $x\in\mathbb{R}$ as stated. But since our final result holds regardless of $x_c(\rho)$, we will not need this level of generality here.

### 3.2 Annealed spherical integral

The goal of this subsection is to prove the following lemma.

**Lemma 3.11.** *Under either Assumption A or Assumption B, for every $0\leq\theta<\theta_{\max}$, we have*

$$\lim_{N\to\infty}\frac{1}{N}\log\mathbb{E}_{e,H_N}[e^{N\frac{\theta}{2}\langle e,H_N e\rangle}]=F(\rho,\theta). \tag{3.1}$$

*Proof.* For every unit vector $e$, we have

$$\begin{aligned}\mathbb{E}_{H_N}[e^{N\frac{\theta}{2}\langle e,H_N e\rangle}]&=\mathbb{E}_{H_N}[e^{N\frac{\theta}{2}\langle e,\left(\frac{1}{M}\sum_{k=1}^M d_k z_k z_k^T\right)e\rangle}]\\&=\mathbb{E}_{H_N}\left[\prod_{k=1}^M e^{\frac{N}{M}\frac{\theta}{2}d_k\langle z_k,e\rangle^2}\right]=\prod_{k=1}^M\mathbb{E}_{H_N}[e^{\frac{N}{M}\frac{\theta}{2}d_k\langle z_k,e\rangle^2}]\end{aligned}$$

since the $z_k$'s are independent (throughout, we use $\mathbb{E}_{H_N}$ for the expectation over the randomness in the $z_k$'s). Applying a Hubbard-Stratonovich transformation, we find

$$\mathbb{E}_{H_N}[e^{N\frac{\theta}{2}\langle e,H_N e\rangle}]=\prod_{k=1}^M\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}\mathbb{E}_{H_N}\left[e^{2x\langle z_k,e\rangle\sqrt{\frac{N}{M}\frac{\theta}{2}d_k}}\right]e^{-x^2}\,\mathrm{d}x,$$

where we interpret $\sqrt{d_k}=\mathrm{i}\sqrt{-d_k}$ for those $d_k$ which are negative.

In the Gaussian case (i.e., under Assumption B), the remainder of the proof is easy: We can exactly compute $\prod_{j=1}^N\mathbb{E}[e^{x\sqrt{2\frac{N}{M}\theta d_k}(z_k)_j e_j}]=e^{x^2\frac{N}{M}\theta d_k}$, giving

$$\frac{1}{N}\log\mathbb{E}_{e,H_N}[e^{N\frac{\theta}{2}\langle e,H_N e\rangle}]=-\frac{M}{2N}\int_{\mathbb{R}}\log\left(1-\frac{N}{M}\theta t\right)\hat{\mu}_\Gamma(\mathrm{d}t).$$

The limiting replacement of $\hat{\mu}_\Gamma$ with $\rho$, and of $\frac{M}{N}$ with $\alpha$, is routine, but it requires $\theta<\theta_{\max}$.

[2] The argument is that exponential tightness automatically upgrades what is called a "weak LDP" to a full LDP (see, e.g., [11, Section 1.2]), and that a weak LDP can be witnessed just by small-ball probabilities (as in Propositions 3.8 and 3.9). The latter follows essentially immediately from the definition of a weak LDP on pp. 6-7 of [11].

In the sharp sub-Gaussian case (i.e., under Assumption A), the upper bound is similar: For every *real* $c$ and every unit $e$ we have

$$\mathbb{E}_{H_N}[\exp(c\langle z_k, e\rangle)] = \prod_{j=1}^N \mathbb{E}_{H_N}[\exp(c(z_k)_j e_j)] \leq \prod_{j=1}^N \exp\left(\frac{c^2 e_j^2}{2}\right) = \exp\left(\frac{c^2}{2}\right).$$

Since each $d_k$ is positive, we can apply this with real $c = 2\sqrt{\frac{N}{M}\frac{\theta}{2}d_k}$ to find

$$\frac{1}{N}\log \mathbb{E}_{e,H_N}[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}] \leq -\frac{M}{2N}\int_{\mathbb{R}} \log\left(1 - \frac{N}{M}\theta t\right)\hat{\mu}_\Gamma(\mathrm{d}t),$$

which finishes the proof of the upper bound. For the lower bound, fix $0 < \epsilon < \frac{1}{4}$ and define

$$V_N^\epsilon = \{e : \|e\|_\infty \leq N^{-\frac{1}{4}-\epsilon}\} \subset \mathbb{S}^{N-1}.$$

Since $\mu$ is standardized and has finite moment generating function (everywhere, in particular near zero), we know that for every $\delta > 0$ there exists $\eta > 0$ such that, for any $|t| \leq \eta$,

$$\int e^{tx}\mu(\mathrm{d}x) \geq e^{\frac{(1-\delta)}{2}t^2}.$$

In particular, if we fix $\delta$ so small that $f(t) = 1 - \frac{(1-\delta)\theta t}{\alpha}$ is bounded below by something strictly positive on the support of $\rho$ (this is possible since $\theta < \theta_{\max}$) and write

$$\eta' = \frac{\eta}{\sqrt{2(\alpha+1)\theta(r(\rho)+1)}},$$

then whenever $e \in V_N^\epsilon$ and $|x| \leq \eta' N^{\frac{1}{4}+\epsilon}$, for each $k$ we have

$$\mathbb{E}_{H_N}\left[\exp\left(2x\langle z_k, e\rangle\sqrt{\frac{N}{M}\frac{\theta}{2}d_k}\right)\right] \geq \exp\left((1-\delta)x^2\frac{N}{M}\theta d_k\right).$$

Thus for such $e$ we have

$$\mathbb{E}_{H_N}[e^{\frac{N}{M}\frac{\theta}{2}d_k\langle z_k, e\rangle^2}] \geq \frac{1}{\sqrt{\pi}}\int_{-\eta' N^{\frac{1}{4}+\epsilon}}^{\eta' N^{\frac{1}{4}+\epsilon}} \exp\left(-\left(1 - (1-\delta)\frac{N}{M}\theta d_k\right)x^2\right)\mathrm{d}x.$$

From standard Gaussian tail bounds, if $c, d > 0$ are independent of $N$ then

$$\frac{1}{\sqrt{\pi}}\int_{dN^{\frac{1}{4}+\epsilon}}^\infty \exp(-cx^2)\,\mathrm{d}x \leq \sqrt{c}\exp(-2cd^2N^{\frac{1}{2}+2\epsilon})$$

so that

$$\begin{aligned}
&\mathbb{E}_{H_N}[e^{\frac{N}{M}\frac{\theta}{2}d_k\langle z_k, e\rangle^2}] \\
&\geq \left(1 - (1-\delta)\frac{N}{M}\theta d_k\right)^{-\frac{1}{2}} - 2\sqrt{1 - \frac{N}{M}\theta d_k}\exp\left(-2\left(1 - \frac{N}{M}\theta d_k\right)(\eta')^2 N^{\frac{1}{2}+2\epsilon}\right) \\
&\geq \left(1 - \frac{(1-\delta)\theta d_k}{\alpha}\right)^{-\frac{1}{2}}\left(1 - C\left|\frac{N}{M} - \frac{1}{\alpha}\right|\right) - C'\exp(-C''N^{\frac{1}{2}+2\epsilon})
\end{aligned}$$

for some constants $C, C', C''$ depending on $\delta$ through $\eta$, but not depending on $k$. Therefore

$$\begin{aligned}
&\frac{1}{N}\log \mathbb{E}_{e,H_N}[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}] \\
&\geq \frac{M}{N}\int_{\mathbb{R}} \underbrace{\log\left[\left(1 - \frac{(1-\delta)\theta t}{\alpha}\right)^{-\frac{1}{2}}\left(1 - C\left|\frac{N}{M} - \frac{1}{\alpha}\right|\right) - C'\exp(-C''N^{\frac{1}{2}+2\epsilon})\right]}_{=:f_N(t)}\rho(\mathrm{d}t) \\
&\quad + \frac{1}{N}\log \mathbb{P}(e \in V_N^\epsilon)
\end{aligned}$$

From [16, Lemma 3.3], we have $\mathbb{P}(e \in V_N^\epsilon) \to 1$. Furthermore, $f_N(t) \to -\frac{1}{2}\log(1-\frac{(1-\delta)\theta t}{\alpha})$ pointwise as $N \to \infty$, and is bounded above on the support of $\rho$ from our choice of $\delta$, so dominated convergence gives

$$\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{E}_{e,H_N}[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}] \geq -\frac{\alpha}{2}\int_{\mathbb{R}} \log\left(1 - \frac{(1-\delta)\theta t}{\alpha}\right)\rho(\mathrm{d}t).$$

Letting $\delta \downarrow 0$ with dominated convergence (possible since $\theta < \theta_{\max}$) again finishes the proof. □

### 3.3 Concentration of measure

The goal of this subsection is to prove the following proposition.

**Proposition 3.12.** *If either Assumption A or Assumption B holds, then for every* $\epsilon > 0$ *we have*

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N}, \sigma) > \epsilon) = -\infty. \tag{3.2}$$

**Remark 3.13.** By definition, control of $d_{\mathrm{BL}}$ is control of integrals with respect to bounded-Lipschitz test functions. Strictly speaking we do not need to control all such test functions; as we will see below, Proposition 3.12 is needed to apply Theorem 6.1 of [17], which is a continuity result for the function $J(\mu,\theta,\lambda)$ in each of its variables. Inside the proof of that result, we find that weak convergence of some sequence $(\mu_N)_{N=1}^\infty$ to some $\mu$ (as witnessed by $d_{\mathrm{BL}}(\mu_N, \mu) \to 0$) is needed only for convergence of certain derivatives of the special function $J$, as well as for certain integrals of translations of the logarithm (away from its singularity). Perhaps simpler arguments would apply for these specific test functions, but we think that the uniform control of test functions provided by Proposition 3.12 is of independent interest.

*Proof.* The structure of the proof is common between the different cases of Assumption A and Assumption B. However, one technical estimate is proved quite differently for the different cases; we will mention this at the appropriate moment below, and otherwise tacitly treat both cases simultaneously.

Consider the decomposition

$$Z = A + B,$$

where

$$A_{ij} = Z_{ij}\mathbb{1}_{|Z_{ij}|\leq N^\gamma} \tag{3.3}$$

for some $\gamma = \gamma(\epsilon) > 0$ to be chosen. (Recall the $Z_{ij}$'s are order one, so $B$ is typically sparse.) Define the matrix

$$H_N^A = \frac{1}{M}A^T\Gamma A$$

and, for large positive $L$, the event

$$\mathcal{E}_L = \{\mathrm{supp}(\hat{\mu}_{H_N}) \subset (-L, L)\} \cap \{\mathrm{supp}(\hat{\mu}_{H_N^A}) \subset (-L, L)\}.$$

To prove (3.2), it suffices to check

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A}) > \epsilon, \mathcal{E}_L) = -\infty \quad \text{for every } L > 100r(\sigma), \tag{3.4}$$

$$\lim_{L\to\infty} \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}(\mathcal{E}_L^c) = -\infty, \tag{3.5}$$

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^A}, \mathbb{E}[\hat{\mu}_{H_N^A}]) > \epsilon) = -\infty, \tag{3.6}$$

$$\lim_{N\to\infty} d_{\mathrm{BL}}(\mathbb{E}[\hat{\mu}_{H_N^A}], \sigma) = 0. \tag{3.7}$$

We start with (3.4), which we prove by adapting arguments of Bordenave, Caputo, and Chafaï (namely [8, Lemma C.2] and [7, Lemma 2.2]). Whenever $f$ is a $C^1$ test function with $\|f\|_{\text{Lip}} + \|f\|_\infty \leq 1$, integration by parts gives

$$\begin{aligned}\left|\int f(\lambda)(\hat{\mu}_{H_N} - \hat{\mu}_{H_N^A})(\mathrm{d}\lambda)\right| \mathbf{1}_{\mathcal{E}_L} &= \left|\int f'(\lambda)(F_{\hat{\mu}_{H_N}} - F_{\hat{\mu}_{H_N^A}})(\mathrm{d}\lambda)\right| \mathbf{1}_{\mathcal{E}_L} \\ &\leq \|f'\|_\infty \|F_{\hat{\mu}_{H_N}} - F_{\hat{\mu}_{H_N^A}}\|_1 \mathbf{1}_{\mathcal{E}_L} \leq 2L\|f\|_{\text{Lip}} \|F_{\hat{\mu}_{H_N}} - F_{\hat{\mu}_{H_N^A}}\|_\infty \\ &\leq 2L d_{\text{KS}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A}).\end{aligned}$$

If $f$ just has $\|f\|_{\text{Lip}} + \|f\|_\infty \leq 1$ but is not necessarily $C^1$, there is a $C^1$ function $g$ with $\|g\|_{\text{Lip}} + \|g\|_\infty \leq 1$ and $\|f - g\|_{L^\infty([-L,L])} \leq 2L d_{\text{KS}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A})$; thus

$$\mathbb{P}(d_{\text{BL}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A}) > \epsilon, \mathcal{E}_L) \leq \mathbb{P}\left(d_{\text{KS}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A}) > \frac{\epsilon}{4L}\right).$$

It is classical (a consequence of interlacing of singular values, see e.g. [2, Theorem A.44]) that

$$d_{\text{KS}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A}) \leq \frac{1}{N}\operatorname{rank}(Z - A) = \frac{1}{N}\operatorname{rank}(B) \leq \frac{1}{N}\sum_{i,j} \mathbb{1}_{|Z_{ij}| > N^\gamma}$$

for any $\Gamma$. Now $(\mathbb{1}_{|Z_{ij}|>N^\gamma})_{1 \leq i \leq M, 1 \leq j \leq N}$ is a collection of $NM$ i.i.d. Bernoulli variables with mean

$$p_N := \mathbb{P}(|Z_{ij}| > N^\gamma) \leq \exp(-cN^{2\gamma})$$

for some $c$ depending on the sub-Gaussian norm of $\mu$. Writing

$$\sigma^2 = NMp_N(1 - p_N) \leq \exp\left(-\frac{c}{2}N^{2\gamma}\right)$$

and $h(x) = (x+1)\log(x+1) - x$, Bennett's inequality [5] gives

$$\mathbb{P}\left(\sum_{i \leq j} \mathbb{1}_{|Z_{ij}|>N^\gamma} - NMp_N \geq t\right) \leq \exp\left(-\sigma^2 h\left(\frac{t}{\sigma^2}\right)\right)$$

for any $t > 0$. We will choose $t = N\frac{\epsilon}{4L} - NMp_N \geq N\frac{\epsilon}{8L}$, which has $\frac{t}{\sigma^2} \to +\infty$; since $h(x) \geq x \log x$ for sufficiently large arguments, we have

$$\mathbb{P}(d_{\text{BL}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A}) > \epsilon, \mathcal{E}_L) \leq \exp\left(-t\log\left(\frac{t}{\sigma^2}\right)\right).$$

Since $\gamma > 0$, this suffices for (3.4).

The estimate (3.5) is essentially a consequence of exponential tightness, Lemma 3.6, which controls $|\lambda_{\text{max}}(H_N)|$; the same proof controls $|\lambda_{\text{min}}(H_N)|$, as well as the extreme eigenvalues of $H_N^A$ (the proof of Lemma 3.6 references [16, Lemma 1.9], which goes through for $A$).

The verification of (3.6) is fairly different in the sharp sub-Gaussian case (Assumption A) vs. the Gaussian case (Assumption B). The latter case is Lemma 3.14 below, while the former case is Lemma 3.15 below (this is also where we select $\gamma$ as a function of $\epsilon$).

For (3.7), we first claim

$$\lim_{N \to \infty} d_{\text{BL}}(\mathbb{E}[\hat{\mu}_{H_N^A}], \mathbb{E}[\hat{\mu}_{H_N}]) = 0. \tag{3.8}$$

It suffices to show $\mathbb{E}[X_N] \to 0$, where $X_N = d_{\text{BL}}(\hat{\mu}_{H_N^A}, \hat{\mu}_{H_N})$. But $X_N$ is a random variable between zero and two, and (3.4) and (3.5) show that for every $\epsilon > 0$ and $N \geq N_0(\epsilon)$ we have $\mathbb{P}(X_N > \epsilon) \le \exp(-100N)$, which shows $\mathbb{E}[X_N] \to 0$ and hence (3.8). It remains only to show

$$\lim_{N\to\infty} d_{\mathrm{BL}}(\mathbb{E}[\hat{\mu}_{H_N}], \sigma) = 0.$$

This is equivalent to the claim $\mathbb{E}[\hat{\mu}_{H_N}] \to \sigma$. The original result of this type is due to Marčenko and Pastur [29], but for the model where the $d_i$'s are i.i.d. draws from $\rho$, instead of being deterministic with the property $\frac{1}{M}\sum \delta_{d_i} \to \rho$; the result for our, latter variant of this model is due to Silverstein and Bai [36] (actually, their result holds in greater generality). □

**Lemma 3.14.** *Under Assumption B, for each $\epsilon > 0$, there exists $\gamma = \gamma(\epsilon) > 0$ such that, if $H_N^A$ is defined in terms of $\gamma$ using* (3.3)*, then*

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^A}, \mathbb{E}[\hat{\mu}_{H_N^A}]) > \epsilon) = -\infty. \tag{3.9}$$

*Proof.* In this case, it is technically inconvenient that the truncation of $A$ is discontinuous; thus we further decompose

$$A = C + D,$$

where

$$C_{ij} = \begin{cases} Z_{ij} & \text{if } |Z_{ij}| \le N^\gamma, \\ N^\gamma \mathrm{sign}(Z_{ij}) & \text{if } |Z_{ij}| > N^\gamma \end{cases}, \qquad D_{ij} = \begin{cases} 0 & \text{if } |Z_{ij}| \le N^\gamma, \\ -N^\gamma \mathrm{sign}(Z_{ij}) & \text{if } |Z_{ij}| > N^\gamma \end{cases},$$

and define the matrix

$$H_N^C = \frac{1}{M} C^T \Gamma C.$$

To prove (3.9), it suffices to check

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^A}, \hat{\mu}_{H_N^C}) > \epsilon, \mathcal{E}_L) = -\infty \quad \text{for every } L > 100 r(\sigma), \tag{3.10}$$

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^C}, \mathbb{E}[\hat{\mu}_{H_N^C}]) > \epsilon) = -\infty, \tag{3.11}$$

$$\lim_{N\to\infty} d_{\mathrm{BL}}(\mathbb{E}[\hat{\mu}_{H_N^C}], \mathbb{E}[\hat{\mu}_{H_N^A}]) = 0. \tag{3.12}$$

The proof of (3.10) is a close copy of the proof of (3.4). Namely, one shows in the same way as before that

$$\mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^A}, \hat{\mu}_{H_N^C}) > \epsilon, \mathcal{E}_L) \le \mathbb{P}\left(d_{\mathrm{KS}}(\hat{\mu}_{H_N^A}, \hat{\mu}_{H_N^C}) > \frac{\epsilon}{4L}\right) \le \mathbb{P}\left(\frac{1}{N}\mathrm{rank}(D) > \frac{\epsilon}{4L}\right).$$

Earlier, we bounded the rank of $B$ by its number of nonzero entries. Here we do the same for $D$, but by construction $B$ and $D$ have the same number of nonzero entries, so the rest of the estimate is exactly the same. Similarly, the proof of (3.12) is analogous to the proof of (3.8).

So it remains only to show (3.11), and we start by showing

$$\sup_{f\in\mathcal{F}_{\mathrm{Lip}}} \mathbb{P}\left(\left|\int_{\mathbb{R}} f(\lambda)(\hat{\mu}_{H_N^C} - \mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda)\right| \ge \delta\right) \le \exp\left(-c\delta^2 N^{2-2\gamma}\right), \tag{3.13}$$

where $c$ is some constant depending on

$$d_{\max} = \sup_M \max_{i=1}^N \left| d_i^{(M)} \right|,$$

which is finite by assumption, and the aspect ratio $\alpha$ (and which can later change from line to line, but is always a function only of $d_{\max}$ and $\alpha$). Indeed, we shift perspective slightly by defining $C : \mathbb{R}^{M\times N} \to \mathbb{R}^{M\times N}$ as

$$C(Z)_{ij} = \begin{cases} Z_{ij} & \text{if } |Z_{ij}| \leq N^\gamma, \\ N^\gamma \text{sign}(Z_{ij}) & \text{if } |Z_{ij}| > N^\gamma, \end{cases}$$

and $H_N^{C(Z)}$ as $H_N^{C(Z)} = M^{-1}C(Z)^T\Gamma C(Z)$. Fix some $f \in \mathcal{F}_{\text{Lip}}$, and consider the map $h = h_f : \mathbb{R}^{M\times N} \to \mathbb{R}$ defined by

$$h(Z) = \int_{\mathbb{R}} f(\lambda)\hat{\mu}_{H_N^{C(Z)}}(\mathrm{d}\lambda).$$

We want to show that $h$ is Lipschitz. Using an $L^p$ version of the Hoffman-Wielandt inequality with $p = 1$ (see e.g. [23, Theorem II]), and writing $\pi \in S_N$ for a permutation in the permutation group on $N$ letters, we have

$$\begin{aligned} |h(Z_1) - h(Z_2)| &\leq \min_{\pi\in S_N} \frac{1}{N}\sum_{i=1}^N \left| f(\lambda_i(H_N^{C(Z_1)})) - f(\lambda_{\pi(i)}(H_N^{C(Z_2)}))\right| \\ &\leq \min_{\pi\in S_N} \frac{1}{N}\sum_{i=1}^N \left| \lambda_i(H_N^{C(Z_1)}) - \lambda_{\pi(i)}(H_N^{C(Z_1)})\right| \\ &\leq \frac{1}{N}\sum_{i=1}^N \left|\lambda_i(H_N^{C(Z_1)} - H_N^{C(Z_2)})\right| = \frac{1}{N}\|H_N^{C(Z_1)} - H_N^{C(Z_2)}\|_* \\ &\leq \frac{1}{NM}(\|C(Z_1)^T\Gamma(C(Z_1) - C(Z_2))\|_* + \|(C(Z_1) - C(Z_2))^T\Gamma C(Z_2)\|_*) \end{aligned}$$

Recalling that, for a matrix $M$, $\|M\|_F = \sqrt{\sum_{i,j}|M_{ij}|^2}$ denotes its Frobenius norm, we will now use that $\|T_1T_2\|_* \leq \|T_1\|_F\|T_2\|_F$; that $\|C(Z_1)^T\Gamma\|_F \leq \sqrt{NM}d_{\max}N^\gamma$, since the entries have magnitude at most $d_{\max}N^\gamma$; and that

$$\|C(Z_1) - C(Z_2)\|_F \leq \|Z_1 - Z_2\|_F$$

(this estimate is why we needed to define $C$; the analogue for $A$ is not true). These give

$$|h(Z_1) - h(Z_2)| \leq 2d_{\max}\frac{N^\gamma}{\sqrt{NM}}\|Z_1 - Z_2\|_F \leq \frac{4d_{\max}}{\sqrt{\alpha}}\frac{N^\gamma}{N}\|Z_1 - Z_2\|_F.$$

By Gaussian concentration for Lipschitz functions (see, e.g., [9, Theorem 5.6]), we have

$$\mathbb{P}\left(\left|\int_{\mathbb{R}} f(\lambda)(\hat{\mu}_{H_N^C} - \mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda)\right| \geq \delta\right) = \mathbb{P}(|h(Z) - \mathbb{E}[h(Z)]| \geq \delta) \leq \exp\left(-c\delta^2N^{2-2\gamma}\right)$$

which proves (3.13).

Now we want to upgrade by taking a supremum over $f$ inside the probability – at first just over $\mathcal{F}_{\text{Lip},\mathcal{K}}$, which we recall denotes the set of functions in $\mathcal{F}_{\text{Lip}}$ supported in some compact set $\mathcal{K}$ with diameter $\text{diam}(\mathcal{K})$. Guionnet and Zeitouni give a very useful construction for this purpose (which we will also mimic in the proof of the sub-Gaussian case below): For any $\Delta > 0$, they construct a set of $2(\lceil\frac{\text{diam}(\mathcal{K})}{\Delta}\rceil + 1)$ functions $(h_k)_{k=1}^{2(\lceil\frac{\text{diam}(\mathcal{K})}{\Delta}\rceil+1)}$ in $\mathcal{F}_{\text{Lip}}$ with the property that, for any given $f \in \mathcal{F}_{\text{Lip},\mathcal{K}}$, one can choose $\lceil\frac{\text{diam}(\mathcal{K})}{\Delta}\rceil + 1$ of the $h_k$'s whose sum, called $f_\Delta$, satisfies $\|f - f_\Delta\|_\infty \leq \Delta$. Since $f_\Delta$ is actually a sum of this finite count of functions, not a linear combination of them, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}_{\mathrm{Lip},\mathcal{K}}}\left|\int_{\mathbb{R}} f(\lambda)(\hat{\mu}_{H_N^C}-\mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda)\right|\geq\delta\right)\\
&\leq 2\left(\left\lceil\frac{\mathrm{diam}(\mathcal{K})}{\Delta}\right\rceil+1\right)\\
&\quad\times\max_{k=1}^{2(\lceil\frac{\mathrm{diam}(\mathcal{K})}{\Delta}\rceil+1)}\mathbb{P}\left(\left|\int_{\mathbb{R}} h_k(\lambda)(\hat{\mu}_{H_N^C}-\mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda)\right|\geq\frac{\delta-2\Delta}{2(\lceil\mathrm{diam}(\mathcal{K})/\Delta\rceil+1)}\right)\\
&\leq 2\left(\left\lceil\frac{\mathrm{diam}(\mathcal{K})}{\Delta}\right\rceil+1\right)\exp\left(-cN^{2-2\gamma}\left(\frac{\delta-2\Delta}{2\left(\lceil\mathrm{diam}(\mathcal{K})/\Delta\rceil+1\right)}\right)^2\right)
\end{aligned}$$

Like Guionnet and Zeitouni, we choose $\Delta=\delta/4$; then whenever $\delta<1$ and $\mathrm{diam}(\mathcal{K})>1$, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}_{\mathrm{Lip},\mathcal{K}}}\left|\int_{\mathbb{R}} f(\lambda)(\hat{\mu}_{H_N^C}-\mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda)\right|\geq\delta\right)\leq\frac{16\delta}{\mathrm{diam}(\mathcal{K})}\exp\left(-c\frac{N^{2-2\gamma}\delta^4}{\mathrm{diam}(\mathcal{K})^2}\right). \tag{3.14}$$

Now we define, for $L>0$, the event

$$\mathcal{E}_{C,L}=\{\mathrm{supp}(\hat{\mu}_{H_N^C})\subset(-L,L)\},$$

mimicking the event $\mathcal{E}_L$ from above. In the same way that we proved the $\mathcal{E}_L$ was likely in (3.5), one can prove

$$\lim_{L\to\infty}\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}(\mathcal{E}_{C,L}^c)=-\infty. \tag{3.15}$$

On the other hand, fix large $L$. For any $f\in\mathcal{F}_{\mathrm{Lip}}$, there exists $\widetilde{f}\in\mathcal{F}_{\mathrm{Lip}}$ that agrees with $f$ on $(-L,L)$ and vanishes outside $(-2L,2L)$, say. Furthermore, on the event $\mathcal{E}_{C,L}$, the empirical measure $\hat{\mu}_{H_N^C}$ is supported on $(-L,L)$, although its expectation is not; thus with $\mathcal{K}:=(-2L,2L)$ we have

$$\begin{aligned}
&\mathbb{P}\left(d_{\mathrm{BL}}(\hat{\mu}_{H_N^C},\mathbb{E}[\hat{\mu}_{H_N^C}])\geq\delta,\mathcal{E}_{C,L}\right)\\
&=\mathbb{P}\left(\sup_{f\in\mathcal{F}_{\mathrm{Lip}}}\left|\int_{\mathbb{R}} f(\lambda)(\hat{\mu}_{H_N^C}-\mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda)\right|\geq\delta,\mathcal{E}_{C,L}\right)\\
&\leq\mathbb{P}\left(\sup_{\widetilde{f}\in\mathcal{F}_{\mathrm{Lip},\mathcal{K}}}\left|\int_{\mathbb{R}}\widetilde{f}(\lambda)(\hat{\mu}_{H_N^C}-\mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda)\right|\geq\frac{\delta}{2},\mathcal{E}_{C,L}\right)\\
&\quad+\mathbb{P}\left(\sup_{f\in\mathcal{F}_{\mathrm{Lip}}}\left|\int_{\mathbb{R}}(f(\lambda)-\widetilde{f}(\lambda))\mathbb{E}[\hat{\mu}_{H_N^C}](\mathrm{d}\lambda)\right|\geq\frac{\delta}{2}\right)\\
&\leq\mathbb{P}\left(\sup_{\widetilde{f}\in\mathcal{F}_{\mathrm{Lip},\mathcal{K}}}\left|\int_{\mathbb{R}}\widetilde{f}(\lambda)(\hat{\mu}_{H_N^C}-\mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda)\right|\geq\frac{\delta}{2}\right)+\mathbf{1}\left\{\int_{\mathbb{R}}\mathbf{1}_{|\lambda|\geq L}\mathbb{E}[\hat{\mu}_{H_N^C}](\mathrm{d}\lambda)\geq\frac{\delta}{4}\right\}\\
&\leq\frac{4\delta}{L}\exp\left(-c\frac{N^{2-2\gamma}\delta^4}{L^2}\right)+\mathbf{1}\left\{\int_{\mathbb{R}}\mathbf{1}_{|\lambda|\geq L}\mathbb{E}[\hat{\mu}_{H_N^C}](\mathrm{d}\lambda)\geq\frac{\delta}{4}\right\}
\end{aligned}$$

where we used (3.14) in the last line. To handle the indicator, we note

$$\int_L^\infty\mathbb{E}[\hat{\mu}_{H_N^C}](\mathrm{d}\lambda)\leq\mathbb{P}(\lambda_{\max}(H_N^C)\geq L)\leq\mathbb{P}(\mathcal{E}_{C,N}^c)$$

and similarly for the left tail, (3.15) gives that the indicator vanishes for $L$ large enough depending on $\delta$; thus

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}_{\mathrm{Lip}}}\left|\int_{\mathbb{R}} f(\lambda)(\hat{\mu}_{H_N^C}-\mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda)\right|\geq\delta,\mathcal{E}_{C,L}\right)\leq\frac{4\delta}{L}\exp\left(-c\frac{N^{2-2\gamma}\delta^4}{L^2}\right)$$

for all $L > L_0(\delta)$. Combined with (3.15), this finishes the proof of (3.11), and thus finishes the proof of Lemma 3.14. □

**Lemma 3.15.** *Fix $\gamma < 1/19$. Under Assumption A, for every $\epsilon > 0$, if $H_N^A$ is defined in terms of $\gamma$ using* (3.3)*, then*

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^A}, \mathbb{E}[\hat{\mu}_{H_N^A}]) > \epsilon) = -\infty.$$

The proof of this lemma is given in Appendix A.

### 3.4 Proof of the weak LDP upper bound

**Definition 3.16.** *Consider the following deterministic set of $M \times N$ real matrices:*

$$\mathcal{A}^L_{x,\delta,\epsilon} = \left\{ Z : \text{ with } H_N = H_N(Z) = \frac{1}{M} Z^T \Gamma Z, \quad |\lambda_{\max}(H_N) - x| < \delta, \right.$$
$$\left. d_{\mathrm{BL}}(\hat{\mu}_{H_N}, \sigma) < \epsilon, \text{ and } \|H_N\| \le L \right\}.$$

**Lemma 3.17.** *For every $x \geq r(\sigma)$ and $0 \leq \theta < \theta_{\max}$, and every large enough $L$, we have*

$$\begin{aligned}&\lim_{\epsilon\downarrow 0} \limsup_{\delta\downarrow 0} \limsup_{N\to\infty} \sup_{Z\in\mathcal{A}^L_{x,\delta,\epsilon}} \left| \frac{1}{N} \log \mathbb{E}_e[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}] - J\left(\sigma, \frac{\theta}{2}, x\right) \right| \\ &= \lim_{\delta\downarrow 0} \limsup_{\epsilon\downarrow 0} \limsup_{N\to\infty} \sup_{Z\in\mathcal{A}^L_{x,\delta,\epsilon}} \left| \frac{1}{N} \log \mathbb{E}_e[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}] - J\left(\sigma, \frac{\theta}{2}, x\right) \right| = 0.\end{aligned}$$

*Proof.* This is an easy consequence of (stronger results from) [17]. Fix (small) $t > 0$ and (large) $L > 0$; their Theorem 6.2 shows that there exists $N_0(t, L)$ such that for $N \geq N_0(t, L)$ we have

$$\sup_{Z\in\mathcal{A}^L_{x,\delta,\epsilon}} \left| \frac{1}{N} \log \mathbb{E}[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}] - J\left(\hat{\mu}_{H_N}, \frac{\theta}{2}, \lambda_{\max}(H_N)\right) \right| \le t.$$

On the other hand, their Theorem 6.1 shows that $J(\mu, \frac{\theta}{2}, x)$ is a jointly continuous function of $\mu$ (in the set of probability measures compactly supported on $[-L, L]$) and $x$ (in the set $[-L, L]$); thus for $\max(\delta, \epsilon)$ below some threshold depending on $t$ we have

$$\sup_{Z\in\mathcal{A}^L_{x,\delta,\epsilon}} \left| J\left(\hat{\mu}_{H_N}, \frac{\theta}{2}, \lambda_{\max}(H_N)\right) - J\left(\sigma, \frac{\theta}{2}, x\right) \right| \le t.$$

We combine these two estimates with the quantifiers in the right order, then take $t \downarrow 0$. □

**Lemma 3.18.** *For each $x \geq r(\sigma)$, $0 \leq \theta < \theta_{\max}$, and $L$ large enough, we have*

$$\limsup_{\delta\downarrow 0} \limsup_{\epsilon\downarrow 0} \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}^\theta(\mathcal{A}^L_{x,\delta,\epsilon}) \le -(\widetilde{I_\sigma}(x) - I_\sigma(x, \theta)).$$

*Proof.* For any $0 \leq \theta' < \theta_{\max}$,

$$\begin{aligned}\mathbb{P}^\theta(\mathcal{A}^L_{x,\delta,\epsilon}) &= \frac{1}{\mathbb{E}_{e,H_N}[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}]} \mathbb{E}_{H_N}\left[ \mathbb{1}_{H_N\in\mathcal{A}^L_{x,\delta,\epsilon}} \mathbb{E}_e[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}] \frac{\mathbb{E}_e[e^{N\frac{\theta'}{2}\langle e, H_N e\rangle}]}{\mathbb{E}_e[e^{N\frac{\theta'}{2}\langle e, H_N e\rangle}]} \right] \\ &\le \frac{\mathbb{E}_{e,H_N}[e^{N\frac{\theta'}{2}\langle e, H_N e\rangle}]}{\mathbb{E}_{e,H_N}[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}]} \left( \sup_{Z\in\mathcal{A}^L_{x,\delta,\epsilon}} \mathbb{E}_e[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}] \right) \left( \sup_{Z\in\mathcal{A}^L_{x,\delta,\epsilon}} \frac{1}{\mathbb{E}_e[e^{N\frac{\theta'}{2}\langle e, H_N e\rangle}]} \right)\end{aligned}$$

By Lemmas 3.11 and 3.17, this shows

$$\limsup_{\delta\downarrow 0}\limsup_{\epsilon\downarrow 0}\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}^{\theta}(\mathcal{A}^{L}_{x,\delta,\epsilon})\leq I_{\sigma}(x,\theta)-I_{\sigma}(x,\theta').$$

Taking the infimum over $0\leq\theta'<\theta_{\max}$ on the right-hand side completes the proof. □

**Lemma 3.19.** *Suppose $A_{N,M}$ is a doubly-indexed sequence of events with*

$$\lim_{M\to\infty}\lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}(A_{N,M})=-\infty.$$

*Then for every $\theta<\theta_{\max}$ we have*

$$\lim_{M\to\infty}\lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}^{\theta}(A_{N,M})=-\infty.$$

*In particular (by taking $A_{N,M}$ independent of $M$), if $\lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}(A_N)=-\infty$, then $\lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}^{\theta}(A_N)=-\infty$.*

*Proof of Proposition 3.8.* If $x<r(\sigma)$, then for small enough $\delta$ we have $\{|\lambda_{\max}(H_N)-x|\leq\delta\}\subset\{d_{\mathrm{BL}}(\hat{\mu}_{H_N},\sigma)>\epsilon\}$ for some $\epsilon=\epsilon(\delta)$; this suffices by Lemma 3.19 and Proposition 3.12. In the remainder we assume $x\geq r(\sigma)$.

For large $L$ and arbitrary fixed $\delta,\epsilon>0$, we have

$$\mathbb{P}^{\theta}(|\lambda_{\max}(H_N)-x|\leq\delta)\leq\mathbb{P}^{\theta}(\mathcal{A}^{L}_{x,\delta,\epsilon})+\mathbb{P}^{\theta}(d_{\mathrm{BL}}(\hat{\mu}_{H_N},\sigma)>\epsilon)+\mathbb{P}^{\theta}(\|H_N\|>L).$$

Then we take the normalized log of both sides; by taking $N\to+\infty$, and applying again Lemma 3.19 and Proposition 3.12, we find

$$\begin{aligned}&\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}^{\theta}(|\lambda_{\max}(H_N)-x|\leq\delta)\\&\leq\max\left\{\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}^{\theta}(\mathcal{A}^{L}_{x,\delta,\epsilon}),\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}^{\theta}(\|H_N\|>L)\right\}.\end{aligned}$$

Then we take $\epsilon\downarrow 0$, then $\delta\downarrow 0$, and apply Lemma 3.18 to get

$$\begin{aligned}&\limsup_{\delta\downarrow 0}\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}^{\theta}(|\lambda_{\max}(H_N)-x|\leq\delta)\\&\leq\max\left\{-(\widetilde{I_{\sigma}}(x)-I_{\sigma}(x,\theta)),\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}^{\theta}(\|H_N\|>L)\right\}.\end{aligned}$$

By taking $L\to\infty$ and applying Lemmas 3.19 and 3.6, we finish the proof. □

*Proof of Lemma 3.19.* We claim that there exists $N_0$ (which depends on the sequence $(\Gamma_M)_{M=1}^{\infty}$, the speed of convergence in the limit $\frac{M_N}{N}\to\alpha$, the measure $\rho$, and on $\theta$) such that, for every unit vector $e$ and for $N\geq N_0$, we have

$$\mathbb{E}_{H_N}[e^{N\frac{\theta}{2}\langle e,H_Ne\rangle}]\geq e^{-2N|F(\rho,\theta)|}.$$

Indeed, in the Gaussian case, the proof of Lemma 3.11 shows that $\mathbb{E}_{H_N}[e^{N\frac{\theta}{2}\langle e,H_Ne\rangle}]=e^{Nf(M,N,\Gamma,\theta)}$ for some function $f$ with the property that $\lim_{N\to\infty}f(M_N,N,\Gamma_{M_N},\theta)=F(\rho,\theta)$, uniformly in the unit vector $e$. Thus for $N$ large enough we have $f(M,N,\Gamma,\theta)\geq-2\,|F(\rho,\theta)|$, say. In the sub-Gaussian case, we even have a sharper estimate: Since each $d_k$ is nonnegative, the function $x\mapsto e^{\frac{N}{M}\frac{\theta}{2}d_kx^2}$ is convex, so Jensen's inequality gives

$$\mathbb{E}_{H_N}[e^{N\frac{\theta}{2}\langle e,H_Ne\rangle}]=\prod_{k=1}^{M}\mathbb{E}_{H_N}[e^{\frac{N}{M}\frac{\theta}{2}\langle z_k,e\rangle^2}]\geq\prod_{k=1}^{M}e^{\frac{N}{M}\frac{\theta}{2}\mathbb{E}[\langle z_k,e\rangle]^2}=1,$$

since the underlying measure $\mu$ is centered. This means that, if $A$ is an event and $\epsilon > 0$, we have

$$\mathbb{P}^{\theta}(A) \leq e^{2N|F(\rho,\theta)|}\mathbb{E}_{H_N}[\mathbf{1}_A \mathbb{E}_e[e^{N\frac{\theta}{2}\langle e,H_N e\rangle}]] \leq e^{2N|F(\rho,\theta)|}\mathbb{P}(A)^{\frac{\epsilon}{1+\epsilon}}\mathbb{E}_{e,H_N}[e^{N\frac{(1+\epsilon)\theta}{2}\langle e,H_N e\rangle}]^{\frac{1}{1+\epsilon}}$$

If $\epsilon$ is so small that $(1+\epsilon)\theta < \theta_{\max}$, we apply Lemma 3.11 to finish the proof. □

### 3.5 Proof of the weak LDP lower bound

**Lemma 3.20.** *Let $x \geq r(\sigma)$, and let $\theta_x$ be as defined in Lemma 3.3. Then for all $L$ sufficiently large and $\delta, \epsilon$ sufficiently small (depending on $x$), we have*

$$\lim_{N\to\infty} \frac{1}{N}\log \mathbb{P}^{\theta_x}(\mathcal{A}^L_{x,\delta,\epsilon}) = 0.$$

*Proof.* We claim that actually $\mathbb{P}^{\theta_x}(\mathcal{A}^L_{x,\delta,\epsilon})$ tends to one. As shown in the proof of Proposition 3.8, for all $0 \leq \theta < \theta_{\max}$, the quantities $\mathbb{P}^{\theta}(d_{\mathrm{BL}}(\hat{\mu}_{H_N},\sigma) \geq \epsilon)$ and $\mathbb{P}^{\theta}(\|H_N\| \geq L)$ tend to zero (actually exponentially quickly), for $L$ large enough and all $\epsilon > 0$. Thus it suffices to show

$$\mathbb{P}^{\theta_x}(|\lambda_{\max}(H_N) - x| \geq \delta) = \mathrm{o}(1). \tag{3.16}$$

But Proposition 3.8 gives a weak LDP upper bound for the variable $\lambda_{\max}(H_N)$ under the measures $\mathbb{P}^{\theta_x}$, which are exponentially tight, with rate function $J_x(y)$ that is infinite for $y < r(\sigma)$ and otherwise equal to $J_x(y) = \widetilde{I_\sigma}(y) - I_\sigma(y,\theta_x)$. Lemma 3.3 shows that $J_x$ is nonnegative and vanishes uniquely at $x$; indeed, if $r(\sigma) \leq y \neq x$, then $J_x(y) = \sup_{0\leq\theta<\theta_{\max}} I_\sigma(y,\theta) - I_\sigma(x,\theta_x)$, and the supremum is achieved uniquely at $\theta_y$, which is different from $\theta_x$ since $y \neq x$. Combined with exponential tightness of $\mathbb{P}^{\theta_x}$ (see Lemmas 3.6 and 3.19), this gives (3.16). (We remark that the condition $x < x_c$ is crucial for this part of the proof.) □

*Proof of Proposition 3.9.* We have

$$\begin{aligned}
&\mathbb{P}(|\lambda_{\max}(H_N) - x| \leq \delta) \\
&\geq \mathbb{P}(\mathcal{A}^L_{x,\delta,\epsilon}) \\
&\geq \frac{\mathbb{E}[\mathbf{1}_{\mathcal{A}^L_{x,\delta,\epsilon}}\mathbb{E}_e[e^{N\frac{\theta_x}{2}\langle e,H_N e\rangle}]]}{\mathbb{E}[e^{N\frac{\theta_x}{2}\langle e,H_N e\rangle}]}\mathbb{E}[e^{N\frac{\theta_x}{2}\langle e,H_N e\rangle}]\left(\inf_{Z\in\mathcal{A}^L_{x,\delta,\epsilon}} \frac{1}{\mathbb{E}_e[e^{N\frac{\theta_x}{2}\langle e,H_N e\rangle}]}\right). \\
&= \mathbb{P}^{\theta_x}(\mathcal{A}^L_{x,\delta,\epsilon})\mathbb{E}[e^{N\frac{\theta_x}{2}\langle e,H_N e\rangle}]\left(\inf_{Z\in\mathcal{A}^L_{x,\delta,\epsilon}} \frac{1}{\mathbb{E}_e[e^{N\frac{\theta_x}{2}\langle e,H_N e\rangle}]}\right).
\end{aligned}$$

From Lemmas 3.11, 3.17, and 3.20, we have

$$\begin{aligned}
\liminf_{\delta\downarrow 0}\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}(|\lambda_{\max}(H_N) - x| \leq \delta) &\geq F(\rho,\theta_x) - J\left(\sigma,\frac{\theta_x}{2},x\right) \\
&= -I_\sigma(x,\theta_x) \geq -\widetilde{I_\sigma}(x).
\end{aligned}$$

□

### 3.6 Degenerate cases

*Proof of Lemma 2.9.* Let $\rho$ be a compactly supported measure on $\mathbb{R}$ such that $r(\rho) \leq 0$. Theorem 2.4 already shows that $r(\sigma) \leq 0$; as in the proof of that theorem, we consider a sequence $(\Gamma_M)_{M=1}^\infty$ chosen without outliers, and the matrices $H_N$ defined using these $\Gamma_M$'s. There are two cases:

1. If $\alpha \leq 1$, note that $H_N \geq -KZ^TZ$ where $K = \max_M(-\lambda_{\min}(\Gamma_M))$. The empirical measure of $Z^TZ$ converges toward the Marčenko-Pastur distribution $\mathrm{MP}_\alpha$ and

therefore the cumulative distribution function of $\sigma$ is everywhere smaller than the distribution function of $m_{-K}\sharp\mathrm{MP}_\alpha$, where $m_{-K}\sharp\mathrm{MP}_\alpha$ is the push-forward of $\mathrm{MP}_\alpha$ by multiplication by $-K$ (i.e., $m_{-K}(x) = -Kx$). Since $r(m_{-K}\sharp\mathrm{MP}_\alpha) = -K\ell(\mathrm{MP}_\alpha) = 0$, we have that $r(\sigma) \geq 0$.

2. If $\alpha > 1$, we can find a $\delta > 0$ such that $\alpha - \delta > 1$. If we call $\delta_M = \lfloor \delta M \rfloor$, we let $\Gamma'_M$ be the submatrix of $\Gamma_M$ with the first $M - \delta_M$ rows and columns, and $\Gamma''_M$ be the submatrix of $\Gamma_M$ with the last $\delta_M$ rows and columns. We also let $Z'$ be the submatrix of $Z$ with the first $M - \delta_M$ rows, and $Z''$ be the submatrix of $Z$ with the last $\delta_M$ rows. We have $H_N = H'_N + H''_N$ with $H'_N = Z'^T\Gamma'_M Z'$ and $H''_N = Z''^T\Gamma''_M Z''$. Since $\rho(\{0\}) = 0$, we have that $\limsup_N \lambda_{\max}(\Gamma'_M) = c < 0$ (we must have $c \leq 0$ since all entries of $\Gamma_M$ are nonpositive at finite $M$; since $\Gamma'_M$ contains the most negative entries of $\Gamma$ by our ordering, if we had $c = 0$, then all the entries of $\Gamma''_M$ would be asymptotically zero, meaning that $\rho$ would have an atom at zero of mass at least $\delta$, which we ruled out by assumption), and since $H''_N \leq 0$, we have for $N$ large enough $H_N \leq H'_N \leq cZ'^T Z'$. So $r(\sigma) \leq c\ell(\mathrm{MP}_{\alpha-\delta}) < 0$. □

*Proof of Proposition 2.13.* The same proof as in the nondegenerate case shows that $\lambda_{\max}$ cannot push into the bulk at this speed, which handles small-ball probabilities $\mathbb{P}(\lambda_{\max} \approx x)$ for $x < 0$.

If $\Gamma$ is negative semidefinite, then the rest of the argument is trivial, since $H_N \leq 0$ and thus $\lambda_{\max}(H_N) \leq 0$ deterministically.

So suppose that $\Gamma$ has a handful of positive eigenvalues at finite $N$, meaning that $H_N$ is not necessarily negative semidefinite. Define $\Gamma^{\mathrm{nsd}} := \mathrm{diag}(d_1^{(\mathrm{nsd})}, \ldots, d_M^{(\mathrm{nsd})})$, where $d_i^{(\mathrm{nsd})} = \min(d_i, 0)$, set $\epsilon_N = \|\Gamma - \Gamma^{(\mathrm{nsd})}\|$ which tends to zero by Assumption 2.2, and set $H_N^{(\mathrm{nsd})} := \frac{1}{M} Z^T \Gamma^{(\mathrm{nsd})} Z$, which we couple with $H_N$ by using the same noise $Z$ to define both. We note that $\|H_N - H_N^{(\mathrm{nsd})}\| \leq \frac{\epsilon_N}{M}\|Z\|^2$ and thus, for every $\delta > 0$,

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(|\lambda_{\max}(H_N^{(\mathrm{nsd})}) - \lambda_{\max}(H_N)| > \delta) \leq \lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(\sqrt{M}^{-1}\|Z\| > \delta/\epsilon_N) = -\infty$$

(the details of this are given in the proof of Lemma 4.3 below). This means that the sequences $(\lambda_{\max}(H_N))_{N=1}^\infty$ and $(\lambda_{\max}(H_N^{(\mathrm{nsd})}))_{N=1}^\infty$ are *exponentially equivalent*; since the latter sequence satisfies the desired LDP by the above argument, it is classical (see, e.g., [11, Theorem 4.2.13]) that the former does, as well. □

### 3.7 Second branch of the Stieltjes transform

*Proof of Lemma 2.8.* If $x_c(\rho)$ is finite, then

$$\begin{aligned} x_c(\rho) &= r(\rho)^2 G_\rho(r(\rho)) + \left(\frac{1}{\alpha} - 1\right) r(\rho) = \frac{1}{\theta_{\max}} + r(\rho)\left(\int_{\mathbb{R}} \frac{r(\rho) - (r(\rho) - u)}{r(\rho) - u} \rho(\mathrm{d}u)\right) \\ &= \frac{1}{\theta_{\max}} + \int_{\mathbb{R}} \frac{\alpha u}{\alpha - \frac{\alpha}{r(\rho)} u} \rho(\mathrm{d}u) = H_\rho(\theta_{\max}). \end{aligned}$$

The claim $x_c(\rho) \geq r(\sigma)$ will be shown along the course of the proof. We will eventually need three cases. Common to them is the computation of

$$f_\rho(\theta) := \theta^2 H'_\rho(\theta) = -1 + \alpha \int_{\mathbb{R}} \frac{u^2\theta^2}{(\alpha - u\theta)^2} \rho(\mathrm{d}u).$$

Notice that $\lim_{\theta\downarrow 0} f_\rho(\theta) = -1$. We claim $f_\rho$ is strictly increasing for $\theta \in (0, \theta_{\max})$. Indeed, it is (a constant plus) an average over $u$ of the functions $f_{u,\rho}(\theta) := \frac{u^2\theta^2}{(\alpha - u\theta)^2}$; the function $f_{0,\rho}$ is constant, and the functions $f_{u,\rho}$ are strictly increasing for each $u \neq 0$, since their derivatives $\frac{2(\frac{\alpha}{u})\theta(\frac{\alpha}{u}-\theta)}{(\frac{\alpha}{u}-\theta)^4}$ have the same sign as $\frac{\alpha}{u}(\frac{\alpha}{u}-\theta)$, and the sign of the latter can be checked by hand depending on the sign of $u$ (in the case $u > 0$, this relies on $\theta < \theta_{\max} = \frac{\alpha}{r(\rho)}$).

The three cases are:

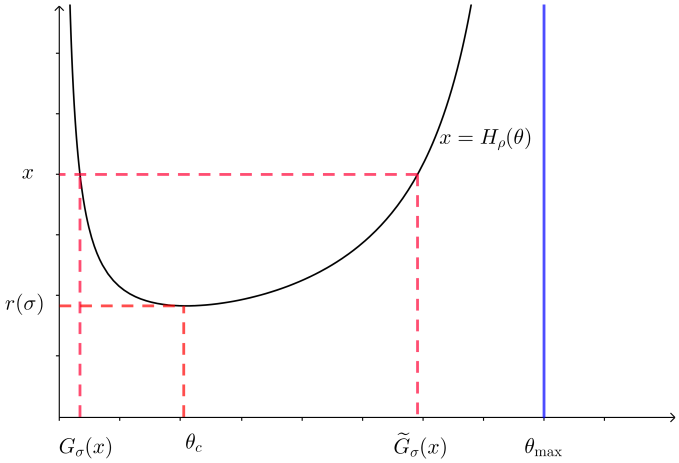

Figure 4: The graph of the function $H_\rho$ when $\rho = \delta_1$ and $\alpha = 8$, along with representations of $r(\sigma)$, $\theta_c$, $G_\sigma(x)$, and $\widetilde{G}_\sigma(x)$. This serves as an example of the case when $r(\rho) > 0$ and $G_\rho(r(\rho)) = +\infty$.

1. **Case 1 ($r(\rho) > 0$ and $G_\rho(r(\rho)) = +\infty$), shown in Figure 4:** It is easy to see that $H_\rho(\theta_{\max}) = +\infty$. Thus $f_\rho(\theta)$ is positive for some $\theta$; since it is also strictly increasing and tends to $-1$ at zero, there exists a unique $\theta_c \in (0, \theta_{\max})$ where it vanishes, i.e., there exists a unique $\theta_c \in (0, \theta_{\max})$ such that $H_\rho$ is decreasing on $(0, \theta_c)$ and increasing on $(\theta_c, \theta_{\max})$. Using the uniqueness of the analytic continuation, one can argue that $H_\rho(\theta_c) = r(\sigma)$ and $G_\sigma(r(\sigma)) = \theta_c$. Furthermore, one sees that the equation $H_\rho(y) = x$, considered as a function of $y \in (0, \theta_{\max})$ parametrized by $x \in \mathbb{R}$,

   (a) has no solution if $x < r(\sigma)$.

   (b) has one solution if $x = r(\sigma)$. That solution is $\theta_c$, and we set $\widetilde{G}_\sigma(r(\sigma)) := \theta_c$.

   (c) has two solutions $y_1$ and $y_2$ such that $0 < y_1 < \theta_c < y_2 < \theta_{\max}$ if $x > r(\sigma)$. Furthermore, due to the Dyson equation (2.4), we clearly have $y_1 = G_\sigma(x)$. We write $\widetilde{G}_\sigma(x)$ for the second solution $y_2$. In particular, $\widetilde{G}_\sigma$ defined this way on $[r(\sigma), +\infty)$ is analytic increasing and $\lim_{x\to\infty} \widetilde{G}_\sigma(x) = \theta_{\max}$.

2. **Case 2 ($r(\rho) > 0$ and $G_\rho(r(\rho)) < +\infty$), shown in Figure 5:** Once again, $H_\rho$ is decreasing on $(0, \theta_c)$ and increasing $(\theta_c, \theta_{\max})$, but $\theta_c = \theta_{\max}$ if and only if $x_c = r(\sigma)$. As before, the equation $H_\rho(y) = x$, considered as a function of $y \in (0, \theta_{\max})$ parametrized by $x \in \mathbb{R}$,

   (a) has no solution if $x < r(\sigma)$.

   (b) has one solution if $x = r(\sigma)$. That solution is $\theta_c$, and we set $\widetilde{G}_\sigma(r(\sigma)) := \theta_c$.

   (c) has two solutions $y_1$ and $y_2$ such that $0 < y_1 < \theta_c \le y_2 < \theta_{\max}$ if $r(\sigma) < x \le x_c$. Once again, we have $y_1 = G_\sigma(x)$ and we will set $\widetilde{G}_\sigma := y_2$.

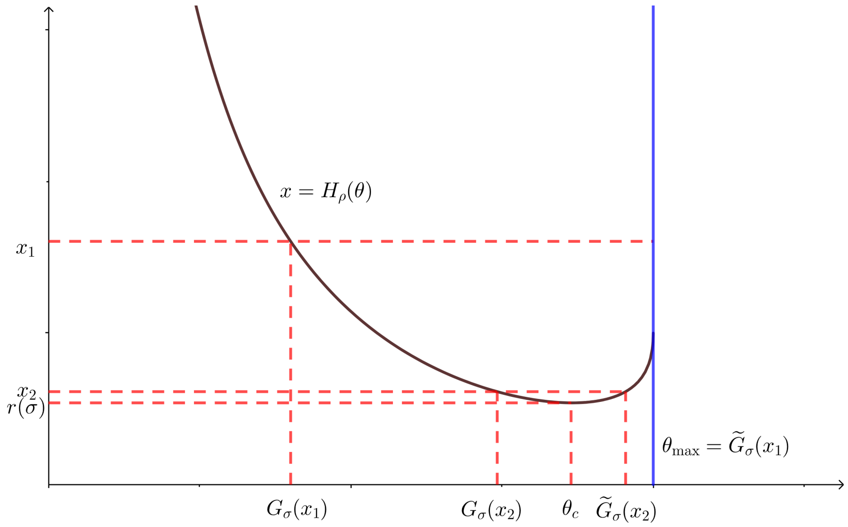

Figure 5: The graph of the function $H_\rho$ when $\rho$ is the semi-circular measure supported in $[-8, 8]$ and $\alpha = 1$, along with representations of $r(\sigma)$, $\theta_c$, $G_\sigma(x)$, and $\widetilde{G}_\sigma(x)$ for some $x_1 > x_c$ and $x_2 < x_c$. This serves as an example of the case $r(\rho) > 0$ and $G_\rho(r(\rho)) < +\infty$.

(d) has one solution $y$ such that $0 < y < \theta_c$ if $x > x_c(\rho)$. However, in this case we will define $\widetilde{G}_\sigma(x) := \theta_{\max}$.

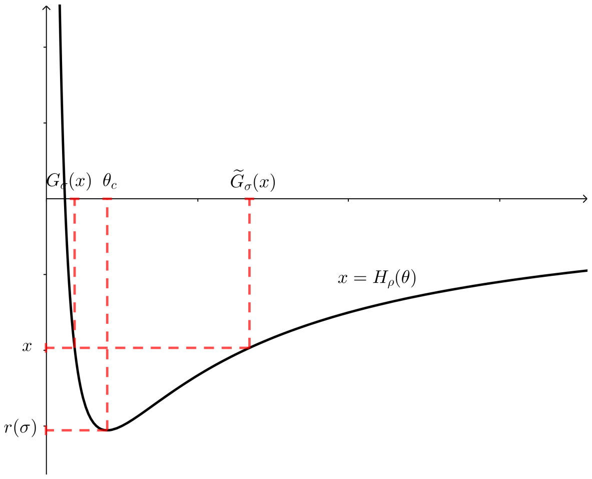

Figure 6: The graph of the function $H_\rho$ when $\rho = \delta_{-10}$ and $\alpha = 5$, along with representations of $r(\sigma)$, $\theta_c$, $G_\sigma(x)$, and $\widetilde{G}_\sigma(x)$ (with a 1:2 rescaling ratio between the $x$-axis and the $y$-axis for readability). It serves as an example of the case when $r(\rho) < 0$ and $\alpha > 1$.

3. **Case 3 ($r(\rho) \leq 0$):** Using Remark 2.10 and Lemma 2.9, we need only consider the case where $\rho(\{0\}) = 0$ and $\alpha > 1$ (see Figure 6). Since $f_\rho$ is strictly increasing, either $H'_\rho(\theta)$ is negative for all $\theta$, or there exists $\theta_c \in (0, \theta_{\max})$ such that $H'_\rho(\theta)$ is negative for $\theta \in (0, \theta_c)$ and positive for $\theta \in (\theta_c, \theta_{\max})$. But since

$$H_\rho(\theta) = (1 - \alpha)\theta^{-1} + \mathrm{o}_{\theta \to +\infty}(\theta^{-1}) \tag{3.17}$$

and $\alpha > 1$, we must be in the latter case, and indeed must have $H_\rho(\theta_c) < 0$ and $H_\rho < 0$ on $(\theta_c, +\infty)$. Then the equation $H_\rho(y) = x$, considered as a function of $y \in (0, +\infty)$ parametrized by $x \in \mathbb{R}$,

(a) has one solution for $x \geq 0$. This solution is equal to $G_\sigma(x)$.

(b) has two solutions $y_1, y_2$ for $0 > x > r(\sigma)$, with $0 < y_1 < \theta_c < y_2$. We have that $y_1 = G_\sigma(x)$ and we denote the second solution $y_2$ by $\widetilde{G}_\sigma(x)$. Once again $\widetilde{G}_\sigma$ is analytic increasing between $r(\sigma)$ and 0, and $\lim_{x \uparrow 0} \widetilde{G}_\sigma = +\infty$.

(c) has one solution for $x = r(\sigma)$, namely $y = \theta_c$.

(d) has no solutions for $x < r(\sigma)$. □

## 4 Proof for finite $x_c$

In this section, we prove Proposition 2.17. We remark that, from the definition (2.3), we can only have $x_c(\rho) < +\infty$ if $r(\rho) > 0$.

We fix once and for all some $\rho$ with $x_c(\rho) < +\infty$, and try to prove the associated LDP, assuming that we know the LDP for every model with $x_c = +\infty$. The proof goes by approximation. Precisely, we are going to discretize the right edge of $\rho$ by replacing the $d_i$'s greater than $r(\rho) - \epsilon$ by $r(\rho)$.

**Definition 4.1.** *For $\epsilon > 0$, we define $\Gamma^{(\epsilon)} := \mathrm{diag}(d_1^{(\epsilon)}, ..., d_M^{(\epsilon)})$ where $d_i^{(\epsilon)} = d_i$ if $d_i \leq r(\rho) - \epsilon$ and $d_i^{(\epsilon)} = r(\rho)$ for $d_i > r(\rho) - \epsilon$. The same way, we define $H_N^{(\epsilon)}$ as:*

$$H_N^{(\epsilon)} := \frac{1}{M} Z^T \Gamma^{(\epsilon)} Z.$$

*It will be important later that we couple $H_N^{(\epsilon)}$ with $H_N$, by using the same noise $Z$ to define both. We define also $\rho^{(\epsilon)}$ to be the probability measure on $\mathbb{R}$ given by*

$$\rho^{(\epsilon)}(A) := \rho(A \cap ]-\infty, r(\rho) - \epsilon]) + \rho(]r(\rho) - \epsilon, r(\rho)])\delta_{r(\rho)} \tag{4.1}$$

*for Borel $A$.*

Let us remark that

$$\lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \delta_{d_i^{(\epsilon)}} = \rho^{(\epsilon)}$$

as long as $\rho$ does not have an atom at $r(\rho) - \epsilon$. Since a probability measure can have at most countably many atoms, we can take $\epsilon \to 0$ along some $\rho$-dependent sequence avoiding such atoms, which we will do implicitly in the rest of the proof.

Then, if we assume we have avoided such atoms, the empirical measure of $H_N^{(\epsilon)}$ converges toward a measure $\sigma^{(\epsilon)}$ characterized by the fact that its Stieltjes transform is the inverse function of $H_{\rho^{(\epsilon)}}$.

Since $\rho^{(\epsilon)}$ has an atom at its right endpoint, we have $G_{\rho^{(\epsilon)}}(r(\rho^{(\epsilon)})) = +\infty$ and therefore we have that $\lambda_{\max}(H_N^{(\epsilon)})$ satisfies a large deviations principle with rate function $I^{(\epsilon)}$ defined as

$$I^{(\epsilon)}(x) = \begin{cases} \frac{1}{2} \int_{r(\sigma^{(\epsilon)})}^{x} \left( \widetilde{G_{\sigma^{(\epsilon)}}}(t) - G_{\sigma^{(\epsilon)}}(t) \right) \mathrm{d}t & \text{if } x \geq r(\sigma^{(\epsilon)}), \\ +\infty & \text{otherwise.} \end{cases}$$

To prove our result we will need the following three lemmas.

**Lemma 4.2.** *The function $\epsilon \mapsto r(\sigma^{(\epsilon)})$ is non-decreasing, and*

$$\lim_{\epsilon \to 0} r(\sigma^{(\epsilon)}) = r(\sigma). \tag{4.2}$$

*Furthermore, the functions $I^{(\epsilon)}$ converge uniformly on all compact subsets of $(r(\sigma), +\infty)$ toward $I$ as $\epsilon \to 0$.*

**Lemma 4.3.** *For every $K > 0$, if the ratio $\frac{\eta}{\epsilon}$ is large enough depending on $K$, then*

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}[\|H_N - H_N^{(\epsilon)}\| \geq \eta] \leq -K$$

*where we recall $\|\cdot\|$ is the operator norm (or spectral radius in this case).*

**Lemma 4.4.** *Define $J : \mathbb{R} \to \mathbb{R}$ by*

$$J(x) = \sup_{\delta>0} \liminf_{\epsilon\downarrow 0} \inf_{y\in(x-\delta,x+\delta)} I^{(\epsilon)}(y). \tag{4.3}$$

*Then $I = J$. Furthermore, $I$ is a good rate function, and for every closed set $F \subset \mathbb{R}$, we have*

$$\inf_{y\in F} I(y) \leq \limsup_{\epsilon\downarrow 0} \inf_{y\in F} I^{(\epsilon)}(y). \tag{4.4}$$

Let us assume these three lemmas momentarily, and prove that they imply the large deviation principle.

*Proof of Proposition 2.17.* This will be an immediate consequence of Theorem 4.2.16 of [11], which explains how to recover an LDP for $\lambda_{\max}(H_N)$ from LDPs for $\lambda_{\max}(H_N^{(\epsilon)})$ in the $\epsilon \downarrow 0$ limit.[3] The condition they define as "exponentially good approximations," translated into our notation, reads

$$\lim_{\epsilon\downarrow 0} \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(\left|\lambda_{\max}(H_N^{(\epsilon)}) - \lambda_{\max}(H_N)\right| > \delta\right) = -\infty,$$

which follows from Lemma 4.3. We checked the remaining conditions of their result in Lemma 4.4 above. □

*Proof of Lemma 4.2.* By construction, $r(\rho^{(\epsilon)}) = r(\rho)$, so that $\theta_{\max}$ is independent of $\epsilon$. Thus

$$r(\sigma^{(\epsilon)}) = \min_{0<\theta<\theta_{\max}} H_{\rho^{(\epsilon)}}(\theta).$$

Since $\rho^{(\epsilon)}$ converges toward $\rho$, and thus $H_{\rho^{(\epsilon)}}$ converges to $H_\rho$ uniformly on all compact subsets of $(0, \theta_{\max})$, this implies

$$\limsup_{\epsilon\to 0} r(\sigma^{(\epsilon)}) \leq r(\sigma).$$

On the other hand, we know more about this convergence: We claim that, for each $\theta \in (0, \theta_{\max})$,

$$H_\rho(\theta) \leq H_{\rho^{(\epsilon)}}(\theta) \tag{4.5}$$

and that the function $\epsilon \mapsto H_{\rho^{(\epsilon)}}(\theta)$ is actually non-decreasing for $\epsilon \in (0, r(\rho))$. (Notice this implies that $\epsilon \mapsto r(\sigma^{(\epsilon)})$ is non-decreasing.) Indeed, we can write

$$H_{\rho^{(\epsilon)}}(\theta) = \frac{1}{\theta} + \int_{-\infty}^{r(\rho)-\epsilon} \frac{\alpha u}{\alpha - \theta u} \rho(\mathrm{d}u) + \frac{\alpha r(\rho)}{\alpha - \theta r(\rho)} \rho((r(\rho)-\epsilon, r(\rho)]) = \frac{1}{\theta} + \int_{\mathbb{R}} f_{\alpha,\theta}^{(\epsilon)}(u)\rho(\mathrm{d}u)$$

where $f_{\alpha,\theta}^{(\epsilon)} : \mathrm{supp}(\rho) \to \mathbb{R}$ is defined by

$$f_{\alpha,x}^{(\epsilon)}(u) = \begin{cases} \frac{\alpha u}{\alpha - \theta u} & \text{if } u < r(\rho) - \epsilon, \\ \frac{\alpha r(\rho)}{\alpha - \theta r(\rho)} & \text{if } u \geq r(\rho) - \epsilon. \end{cases}$$

[3] Translating the notation: Their $m$ is our $\epsilon^{-1}$, and their $\epsilon$ is our $N^{-1}$. Thus their $\widetilde{\mu_\epsilon}$ is the law of $\lambda_{\max}(H_{\epsilon^{-1}})$, and their $\mu_{\epsilon,m}$ is the law of $\lambda_{\max}(H_{\epsilon^{-1}}^{(m^{-1})})$.

Since $u \mapsto \frac{\alpha u}{\alpha - \theta u}$ is strictly increasing on the support of $\rho$ and positive for $u > 0$, the map $\epsilon \mapsto f^{(\epsilon)}_{\alpha,\theta}(u)$ is, for each $u \in \operatorname{supp}(\rho)$, non-decreasing on the set $\epsilon \in (0, r(\rho))$. This shows that $\epsilon \mapsto H_{\rho^{(\epsilon)}}(\theta)$ is non-decreasing for small enough $\epsilon$ (uniformly in $\theta$), and thus that

$$\liminf_{\epsilon \downarrow 0} r(\sigma^{(\epsilon)}) \geq r(\sigma),$$

finishing the proof of (4.2).

Now we prove uniform convergence of $I^{(\epsilon)}$ to $I$ on compact sets of $(r(\sigma), +\infty)$. Recall that $\theta_{\max}$ does not depend on $\epsilon$. If $x > r(\sigma)$, then for $\epsilon$ sufficiently small we have $x > r(\sigma^{(\epsilon)})$ and thus

$$\begin{aligned} I^{(\epsilon)}(x) &= \frac{1}{2} \int_{r(\sigma^{(\epsilon)})}^{x} \left( \widetilde{G}_{\sigma^{(\epsilon)}}(t) - G_{\sigma^{(\epsilon)}}(t) \right) \mathrm{d}t = \frac{1}{2} \int_{r(\sigma^{(\epsilon)})}^{x} \int_0^{\theta_{\max}} \mathbb{1}_{G_{\sigma^{(\epsilon)}}(t) \leq u \leq \widetilde{G}_{\sigma^{(\epsilon)}}(t)} \,\mathrm{d}u \,\mathrm{d}t \\ &= \frac{1}{2} \int_0^x \int_0^{\theta_{\max}} \mathbb{1}_{H_{\rho^{(\epsilon)}}(u) \leq t} \,\mathrm{d}u \,\mathrm{d}t. \end{aligned}$$

Similarly,

$$I(x) = \frac{1}{2} \int_0^x \int_0^{\theta_{\max}} \mathbb{1}_{H_\rho(u) \leq t} \,\mathrm{d}u \,\mathrm{d}t.$$

Define

$$\begin{aligned} D^{(\epsilon)} &:= \{(t,u) \in (0,+\infty) \times (0, \theta_{\max}) : H_{\rho^{(\epsilon)}}(u) \geq t > H_\rho(u)\}, \\ R_x &:= (0,x) \times (0, \theta_{\max}), \\ D^{(\epsilon)}_x &:= D^{(\epsilon)} \cap R_x. \end{aligned}$$

From (4.5), we actually have $I^{(\epsilon)}(x) \leq I(x)$ and

$$I(x) - I^{(\epsilon)}(x) = \frac{1}{2} \int_0^x \int_0^{\theta_{\max}} \mathbb{1}_{H_\rho(u) < t \leq H_{\rho^{(\epsilon)}}(u)} \,\mathrm{d}u \,\mathrm{d}t = \frac{1}{2} \operatorname{Leb}(D^{(\epsilon)}_x).$$

Therefore, if $[a,b] \subset (r(\sigma), +\infty)$ then for all $x \in [a,b]$ and all $\epsilon < \epsilon_0(a)$ we have

$$|I^{(\epsilon)}(x) - I(x)| \leq \operatorname{Leb}(D^{(\epsilon)}_b).$$

Since $H_{\rho^{(\epsilon)}}$ decreases to $H_\rho$, the sets $D^{(\epsilon)}_b$ are nested, and their intersection over all $\epsilon > 0$ is empty. Since their Lebesgue measures are bounded above by $\operatorname{Leb}(R_b) < \infty$, we have $\lim_{\epsilon \downarrow 0} \operatorname{Leb}(D^{(\epsilon)}_b) = 0$, proving the uniform convergence of $I^{(\epsilon)}$ towards $I$ on compact sets of $(r(\sigma), +\infty)$. □

*Proof of Lemma 4.3.* Deterministically, we have

$$\|H_N - H^{(\epsilon)}_N\| = \frac{1}{M} \|Z^T(\Gamma - \Gamma^{(\epsilon)})Z\| \leq \frac{1}{M} \|Z\|^2 \|\Gamma - \Gamma^{(\epsilon)}\| \leq \frac{\epsilon \|Z\|^2}{M}.$$

Therefore it is sufficient to prove that for every $K > 0$ there exists $t_K > 0$ such that, for all $t > t_K$,

$$\limsup_{N \to \infty} \frac{1}{N} \log \mathbb{P}[\sqrt{M}^{-1} \|Z\| \geq t] \leq -K.$$

This can be deduced from the sub-Gaussian character of the entries of $Z$ and $\lim_N \frac{M}{N} = \alpha$ using for instance the arguments of [16, Section 2]. □

*Proof of Lemma 4.4.* Define

$$J_\delta(x) = \liminf_{\epsilon\downarrow 0} \inf_{y\in(x-\delta,x+\delta)} I^{(\epsilon)}(y),$$

which is non-increasing in $\delta$. If $x < r(\sigma)$, then by Lemma 4.2 there exists $\delta > 0$ such that $x + \delta < r(\sigma^{(\epsilon)})$ for all sufficiently small $\epsilon$, so $J_\delta(x) = +\infty$, and thus $J(x) = +\infty$. If $x > r(\sigma)$, then there exists $\delta > 0$ with $x - \delta > r(\sigma^{(\epsilon)})$ for all sufficiently small $\epsilon$. Since each $I^{(\epsilon)}$ is non-decreasing, this gives $J_\delta(x) = \liminf_{\epsilon\downarrow 0} I^{(\epsilon)}(x-\delta) = I(x-\delta)$, again by Lemma 4.2, and thus $J(x) = I(x)$. Finally, we let $x = r(\sigma)$. Then for every $\delta > 0$ and all $\epsilon < \epsilon_0(\delta)$, we have $r(\sigma^{(\epsilon)}) \in (x-\delta, x+\delta)$, so that $J_\delta(x) = 0 = J(x)$. This completes the proof that $I = J$, and $I$ is clearly a good rate function, since it is infinite on $(-\infty, r(\sigma))$, vanishes uniquely at $r(\sigma)$, and is strictly increasing.

Now we check (4.4), splitting into cases according to whether $\alpha(F) = \inf_{y\in F} I(y)$ is infinite or finite. If $\alpha(F) = +\infty$, then necessarily $F \subset (-\infty, r(\sigma))$, with $\sup\{y : y \in F\} < r(\sigma)$. Then Lemma 4.2 gives $\inf_{y\in F} I^{(\epsilon)}(y) = +\infty$ for all $\epsilon$ sufficiently small. If $\alpha(F) = 0$, there is nothing to prove. If $0 < \alpha(F) < \infty$, then $F \subset (r(\sigma)+\delta, +\infty)$ for some $\delta > 0$, and with $y_F = \min\{y : y \in F\}$ we have $\alpha(F) = I(y_F)$. Whenever $\epsilon$ is small enough that $y_F > r(\sigma^{(\epsilon)})$, we have $\inf_{y\in F} I^{(\epsilon)}(y) = I^{(\epsilon)}(y_F)$; as $\epsilon \downarrow 0$ this tends to $I(y_F)$, by Lemma 4.2. □

**Remark 4.5.** In the case where $x_c$ is finite, one can wonder about what happens about the previous tilting strategy. In fact, if we try to adapt this strategy, a natural candidate for the tilt for $x > x_c$ is $\mathbb{P}^{\theta_x^N}$, where $\theta_x^N = \widetilde{G_{\sigma_N}}(x)$, with $\sigma_N = \hat{\mu}_{\Gamma_M} = \frac{1}{M}\sum_{i=1}^M \delta_{d_i}$. However, here our strategy collapses, since for every such $x > x_c$, we then have that $\lim_{N\to\infty} \theta_x^N = \theta_{\max}$. In particular, the argument we have for the lower bound argument does not enable us to state Lemma 3.20, since we cannot prove that $\mathbb{P}^{\theta_x^N}[|\lambda_{\max}(H_N) - x| \le \epsilon] = \mathrm{o}(1)$ for $x > x_c$.

In fact, another clue that something qualitatively different happens for $x > x_c$ lies in a closer examination of the tilts $\mathbb{P}^{\theta_x^N}$. Indeed, considering for simplicity's sake the Gaussian case with $\lambda_{\min}(\Gamma) \ge \epsilon$ for some fixed $\epsilon > 0$, the largest eigenvalue is not exponentially tight under those tilts. Indeed, since the distribution of $Z$ is invariant by the action by right multiplication of the orthogonal group, one can then write for any $\theta$:

$$\mathbb{P}^\theta = \frac{1}{v(\mathbb{S}^{N-1})}\int_{e\in\mathbb{S}^{N-1}} \mathbb{P}^{\theta,e}\,\mathrm{d}e$$

where

$$\frac{d\mathbb{P}^{\theta,e}}{d\mathbb{P}}(Z) = \frac{e^{N\frac{\theta}{2}\left\langle e, \frac{1}{M}Z^*\Gamma Z e\right\rangle}}{\mathbb{E}_{H_N}[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}]}$$

with $v(\mathbb{S}_{N-1})$ the volume of the sphere.

In other words, the $\mathbb{P}^\theta$ are a mixture of the $\mathbb{P}^{\theta,e}$ for unit vectors $e$. In fact, we claim that, while the law of $\mathbb{P}^{\theta,e}$ over $Z \in \mathbb{R}^{M\times N}$ depends very much on $e$, the law that $\mathbb{P}^{\theta,e}$ induces on $\lambda_{\max}(M^{-1}Z^*\Gamma Z)$ does *not* depend on $e$ (and therefore that this common induced law matches the law of $\lambda_{\max}(M^{-1}Z^*\Gamma Z)$ under $\mathbb{P}^\theta$). Indeed, for any unit vector $e \in \mathbb{S}^{N-1}$ and any $N$-dimensional orthogonal matrix $O$, since $ZO \overset{d}{=} Z$ under the original Gaussian measure, for any $A \subset \mathbb{R}^{M\times N}$ we have

$$\begin{aligned}\mathbb{P}^{\theta,e}(A) &= \frac{\mathbb{E}_{H_N}[\mathbb{1}\{Z\in A\}e^{N\frac{\theta}{2}\left\langle e,\frac{1}{M}Z^*\Gamma Z e\right\rangle}]}{\mathbb{E}_{H_N}[e^{N\frac{\theta}{2}\left\langle e,\frac{1}{M}Z^*\Gamma Z e\right\rangle}]}\\ &= \frac{\mathbb{E}_{H_N}[\mathbb{1}\{ZO\in A\}e^{N\frac{\theta}{2}\left\langle Oe,\frac{1}{M}Z^*\Gamma Z(Oe)\right\rangle}]}{\mathbb{E}_{H_N}[e^{N\frac{\theta}{2}\left\langle Oe,\frac{1}{M}Z^*\Gamma Z(Oe)\right\rangle}]} = \mathbb{P}^{\theta,Oe}(AO^{-1})\end{aligned}$$

where $AO^{-1} = \{Z \in \mathbb{R}^{M\times N} : ZO \in A\}$. Of course $A \neq AO^{-1}$ in general. But if $A$ depends on $Z$ only through $\lambda_{\max}(M^{-1}Z^*\Gamma Z) = \lambda_{\max}(M^{-1}(ZO)^*\Gamma(ZO))$, then $A = AO^{-1}$, proving that this induced law is independent of $e$. In particular, to show that $\lambda_{\max}(M^{-1}Z^*\Gamma Z)$ is not exponentially tight under $\mathbb{P}^{\theta_x^N}$, we can show that it is not exponentially tight under $\mathbb{P}^{\theta_x^N,e}$ for our favorite $e$, which we will choose as the first vector $e_1$ of the canonical basis.

By writing down the scalar product at the exponent, one can easily see that, under $\mathbb{P}^{\theta_x^N,e_1}$, the entries of $Z$ remain Gaussian, centered, and independent, with variances that change only in the first row, namely to

$$\mathrm{Var}(Z_{i,1}) = \frac{1}{1-\frac{N\theta_x^N d_i}{M}}.$$

If we now assume that $d_1$ is the largest $d_i$ so that $\lim_{N\to\infty} d_1 = r(\rho)$, then $\lim_{N\to\infty} 1 - \frac{N\theta_x^N d_1}{M} = 0$, so

$$\lim_{N\to\infty} \mathrm{Var}(Z_{1,1}) = +\infty.$$

This means that $\frac{1}{\sqrt{M}}|Z_{1,1}|$ (hence the largest singular value of $\frac{1}{\sqrt{M}}Z$, hence the largest eigenvalue of $\frac{1}{M}Z^*Z$) is not exponentially tight under $\mathbb{P}^{\theta_x^N,e_1}$. Since we assumed $\lambda_{\min}(\Gamma) \geq \epsilon > 0$, we have $\lambda_{\max}(M^{-1}Z^*\Gamma Z) \geq \epsilon\lambda_{\max}(M^{-1}Z^*Z)$ deterministically, and thus $\lambda_{\max}(M^{-1}Z^*\Gamma Z)$ is not exponentially tight under $\mathbb{P}^{\theta_N^x,e_1}$ (therefore, by our discussion above, not exponentially tight under $\mathbb{P}^{\theta_N^x}$). This is a strong clue that this strategy then falls outside the purview of the asymptotic method we use in this article.

## 5 The complex case

In this section, we will review the changes needed to adapt the proof of Theorem 2.15 to Theorem 2.21.

- We keep Definition 3.1 and Definition 3.2. However we need to modify Definition 3.7 by replacing $\theta/2$ by $\theta$:

$$\frac{\mathrm{d}\mathbb{P}^\theta}{\mathrm{d}\mathbb{P}}(Z) = \frac{\mathbb{E}_e[e^{N\theta\left\langle e,\frac{1}{M}Z^*\Gamma Ze\right\rangle}]}{\mathbb{E}_{e,H_N}[e^{N\theta\langle e,H_Ne\rangle}]}.$$

- In Propositions 3.8 and 3.9 we multiply by 2 both right-hand sides:

$$\limsup_{\delta\downarrow 0}\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}_N^\theta(|\lambda_{\max}(H_N)-x|\leq\delta)\begin{cases}\leq -2(\widetilde{I_\sigma}(x)-I_\sigma(x,\theta)) & \text{if } x\in D,\\ = -\infty & \text{otherwise,}\end{cases}$$

  and

$$\liminf_{\delta\downarrow 0}\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}_N(|\lambda_{\max}(H_N)-x|<\delta)\geq -2\widetilde{I_\sigma}(x).$$

- In Lemma 3.11 the equation (3.1) is replaced by

$$\lim_{N\to\infty}\frac{1}{N}\log\mathbb{E}_{e,H_N}[e^{N\theta\langle e,H_Ne\rangle}] = 2F(\rho,\theta).$$

  In the proof of this Lemma, the Hubbard-Stratonovich transformation becomes

$$\mathbb{E}_{H_N}[e^{N\theta\langle e,H_Ne\rangle}] = \prod_{k=1}^{M}\frac{1}{\pi}\int_{w\in\mathbb{C}}\mathbb{E}\left[e^{2\Re(\overline{w}\langle z_k,e\rangle)\sqrt{\frac{N}{M}\theta d_k}}\right]e^{-|w|^2}\,\mathrm{d}w.$$

  In the Gaussian case we have:

$$\prod_{j=1}^{N}\mathbb{E}[e^{2\Re(\overline{w}\overline{(z_k)}_j e_j)\sqrt{\frac{N}{M}\theta d_k)}}] = e^{|w|^2\frac{N}{M}\theta d_k}$$

and then:

$$\frac{1}{N}\log \mathbb{E}_{e,H_N}[e^{N\theta\langle e,H_N e\rangle}] = -\frac{M}{N}\int_{\mathbb{R}} \log\left(1-\frac{N}{M}\theta t\right)\hat{\mu}_\Gamma(\mathrm{d}t).$$

In the sharp sub-Gaussian case we get for any $w \in \mathbb{C}$ and $c \in \mathbb{R}$:

$$\begin{aligned}\mathbb{E}[\exp(c\Re(\overline{w}\,\langle z_k, e\rangle))] &= \prod_{j=1}^{N}\mathbb{E}[\exp(c\Re(\overline{w(z_k)_j}e_j))] \\ &\leq \prod_{j=1}^{N}\exp\left(\frac{c^2|w|^2|e_j|^2}{4}\right) = \exp\left(\frac{c^2|w|^2}{4}\right),\end{aligned}$$

leading to:

$$\frac{1}{N}\log \mathbb{E}_{e,H_N}[e^{N\theta\langle e,H_N e\rangle}] \leq -\frac{M}{N}\int_{\mathbb{R}} \log\left(1-\frac{N}{M}\theta t\right)\hat{\mu}_\Gamma(\mathrm{d}t).$$

Similar modifications happen for the lower bound.

- In Lemma 3.17, we modify the equations to

$$\begin{aligned}&\lim_{\epsilon\downarrow 0}\limsup_{\delta\downarrow 0}\limsup_{N\to\infty}\sup_{Z\in\mathcal{A}^L_{x,\delta,\epsilon}}\left|\frac{1}{N}\log\mathbb{E}_e[e^{N\theta\langle e,H_N e\rangle}] - 2J\left(\sigma,\frac{\theta}{2},x\right)\right| \\ &= \lim_{\delta\downarrow 0}\limsup_{\epsilon\downarrow 0}\limsup_{N\to\infty}\sup_{Z\in\mathcal{A}^L_{x,\delta,\epsilon}}\left|\frac{1}{N}\log\mathbb{E}_e[e^{N\theta\langle e,H_N e\rangle}] - 2J\left(\sigma,\frac{\theta}{2},x\right)\right| = 0.\end{aligned}$$

  The proof is actually identical, we merely use the complex version (that is, $\beta = 2$ of Theorem 6.2 in [17]). One has to careful that the conventions for the function $J$ differ between this paper and [17]. If we denote $J^{GH}$ the function $J$ used in [17],

$$J^{GH}(\mu,\theta,x) = 2J\Big(\mu,\frac{\theta}{2},x\Big).$$

- In Lemma 3.18 we once again have to multiply by 2 the right hand side:

$$\limsup_{\delta\downarrow 0}\limsup_{\epsilon\downarrow 0}\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}^{\theta}(\mathcal{A}^L_{x,\delta,\epsilon}) \leq -2(\widetilde{I_\sigma}(x) - I_\sigma(x,\theta)).$$

  The modifications made to Lemmas 3.11 and 3.17 carry over to the proof which otherwise remains identical.
- In Lemma 3.3, the expression of $\theta_x$ stays the same.
- In the proof of Proposition 3.9, once again the modifications made to Lemmas 3.11, 3.17 carry over and we get

$$\begin{aligned}\liminf_{\delta\downarrow 0}\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}(|\lambda_{\max}(H_N)-x|\leq\delta) &\geq 2F(\rho,\theta_x) - 2J\left(\sigma,\frac{\theta_x}{2},x\right) \\ &= -2I_\sigma(x,\theta_x) \geq -2\widetilde{I_\sigma}(x).\end{aligned}$$

## A Concentration for empirical measures of generalized sample covariance matrices

The proof of Lemma 3.15 will follow from the following general result for concentration of the empirical spectral measure of generalized sample covariance matrices, when the underlying randomness is compactly supported. Such a result was anticipated by [19] (and our proof uses their results as input), but Remark 3 on p. 127 of that paper does not provide details, and what is written there is not quite straightforward to implement because the natural class of approximating functions does not preserve the bounded-Lipschitz property. So for completeness we provide the full details below. In our case, the size of the compact support will depend on $N$, so we do need to track the dependence on the size of the support to ensure that all our factors of $N$ work out.

We note that the proof relies fundamentally on the linearization of a generalized sample covariance matrix (see Lemma A.2), which is only possible when $\Gamma$ is positive semidefinite. This is one of the main reasons for this restriction on $\Gamma$ in the sharp sub-Gaussian case.

The first version of this paper gave a different argument for Lemma 3.15, based on adjusting the class of approximating functions mentioned above. We thank Ofer Zeitouni for suggesting the following argument, which is based on the conceptually simpler observations that the empirical measures of generalized sample covariance matrices are related to pushforwards of the empirical measures of their linearizations under the map $x \mapsto x^2$ (see Lemma A.4), and that this pushforward has nice properties with respect to $d_{\mathrm{BL}}$ (see Lemma A.3). For readers comparing to [19], we note that their $|K|$ is the diameter of the set $K$, which we write as $\operatorname{diam}(K)$.

**Proposition A.1.** *Let $A \in \mathbb{R}^{M\times N}$ have i.i.d. entries whose laws are supported in some compact set $K$; let $\Gamma \in \mathbb{R}^{M\times M}$ be deterministic, diagonal, positive semi-definite, and have all entries bounded above by $d_{\max}$; and consider the matrix $H \in \mathbb{R}^{N\times N}$ given by*

$$H = \frac{1}{M} A^T \Gamma A.$$

*Let $\delta_1(N,M) = \frac{8\operatorname{diam}(K)\sqrt{\pi}\sqrt{d_{\max}}}{\sqrt{M(N+M)}}$, $S = \max_{z\in K} |z|^2$, and $P = \sqrt{8Sd_{\max}\frac{N+M}{M}}$. Then whenever $\delta$ satisfies both of the implicit conditions*

$$\frac{\delta}{2} > (128(P+\sqrt{\delta/2})\delta_1(N,M))^{2/5}, \tag{A.1}$$

$$\left(\frac{\delta}{4(1+\frac{N}{N+M}d_{\max}S)}\right)^{3/2} > \left(128\left(P+\left(\frac{\delta}{4(1+\frac{N}{N+M}d_{\max}S)}\right)^{3/2}\right)\delta_1(N,M)\right)^{2/5}, \tag{A.2}$$

*we have*

$$\begin{aligned}
&\mathbb{P}\left(d_{\mathrm{BL}}(\hat{\mu}_H, \mathbb{E}[\hat{\mu}_H]) \geq \frac{N+M}{2N}\delta\right) \\
&\leq \frac{128(P+\sqrt{\delta/2})}{(\delta/2)^{3/2}} \exp\left(-M(N+M)\frac{1}{16\operatorname{diam}(K)^2 d_{\max}}\left[\frac{(\delta/2)^{5/2}}{128(P+\sqrt{\delta/2})} - \delta_1(N,M)\right]^2\right) \\
&+ \frac{128\left(P+\left(\frac{\delta}{4(1+\frac{N}{N+M}d_{\max}S)}\right)^{3/4}\right)}{\left(\frac{\delta}{4(1+\frac{N}{N+M}d_{\max}S)}\right)^{9/4}} \exp\Bigg(-M(N+M)\frac{1}{16\operatorname{diam}(K)^2 d_{\max}} \\
&\qquad \left[\frac{\left(\frac{\delta}{4(1+\frac{N}{N+M}d_{\max}S)}\right)^{15/4}}{128\left(P+\left(\frac{\delta}{4(1+\frac{N}{N+M}d_{\max}S)}\right)^{3/4}\right)} - \delta_1(N,M)\right]^2\Bigg)
\end{aligned} \tag{A.3}$$

Before proving Proposition A.1, we show that it implies Lemma 3.15.

*Proof of Lemma 3.15.* In our case, $d_{\max}$ is order-one, but $\operatorname{diam}(K)$ is order $N^{\gamma}$. Thus $\delta_1(N,M)$ is order $N^{\gamma-1}$, the quantity $S$ is order $N^{2\gamma}$, and $P$ is order $N^{\gamma}$. This means that, if $\delta$ is order-one, then (A.1) and (A.2) are both satisfied as long as $\gamma < 2/19$. For such $\gamma$, the right-hand side of (A.3) is order $\exp(-N^{2-2\gamma-2\gamma}) + \exp(-N^{2-2\gamma-17\gamma}) \sim \exp(-N^{2-19\gamma})$ up to polynomial errors; thus $\frac{1}{N}\log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^A}, \mathbb{E}[\hat{\mu}_{H_N^A}]) > \epsilon) \to -\infty$ as soon as $\gamma < 1/19$ as claimed. □

Our proof of Proposition A.1 relies on the following results, which we prove after. In them, we use the notation $\mu\sharp x^2$ for the pushforward of a probability measure $\mu$ by the map $x \mapsto x^2$, i.e., $\int f(y)(\mu\sharp x^2)(\mathrm{d}y) = \int f(y^2)\mu(\mathrm{d}y)$.

**Lemma A.2.** *In the setup of Proposition A.1, consider the matrix $\widetilde{H} \in \mathbb{R}^{(N+M)\times(N+M)}$ given in block form by*

$$\widetilde{H} = \frac{1}{\sqrt{M}}\begin{pmatrix} 0 & A^T\Gamma^{1/2} \\ \Gamma^{1/2}A & 0 \end{pmatrix}.$$

*Then whenever $\delta$ satisfies the implicit condition $\delta > (128(P+\sqrt{\delta})\delta_1(N,M))^{2/5}$, we have*

$$\mathbb{P}\left(d_{\mathrm{BL}}(\hat{\mu}_{\widetilde{H}}, \mathbb{E}[\hat{\mu}_{\widetilde{H}}]) \geq \delta\right) \leq \frac{128(P+\sqrt{\delta})}{\delta^{3/2}} \exp\left(-M(N+M)\frac{1}{16\operatorname{diam}(K)^2 d_{\max}}\left[\frac{\delta^{5/2}}{128(P+\sqrt{\delta})} - \delta_1(N,M)\right]^2\right)$$

**Lemma A.3.** *Suppose that $\mu$ and $\nu$ are probability measures on the right half-line. Then for any $L$ we have*

$$d_{\mathrm{BL}}(\mu\sharp x^2, \nu\sharp x^2) \leq (1+2L)d_{\mathrm{BL}}(\mu,\nu) + \mu((L,\infty)) + \nu((L,\infty)).$$

**Lemma A.4.** *We have*

$$d_{\mathrm{BL}}(\hat{\mu}_{\widetilde{H}}\sharp x^2, \mathbb{E}[\hat{\mu}_{\widetilde{H}}]\sharp x^2) = \frac{2N}{N+M} d_{\mathrm{BL}}(\hat{\mu}_H, \mathbb{E}[\hat{\mu}_H]).$$

*Proof of Proposition A.1.* By Lemma A.4, it suffices to bound $\mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{\widetilde{H}}\sharp x^2, \mathbb{E}[\hat{\mu}_{\widetilde{H}}]\sharp x^2) > \delta)$. We do this with Lemma A.3, first using Markov's inequality to upper-bound

$$\mu((L,\infty)) \leq \frac{\int_{\mathbb{R}} t^2 \mu(\mathrm{d}t)}{L^2},$$

then using the deterministic bound

$$\int_{\mathbb{R}} t^2 \hat{\mu}_{\widetilde{H}}(\mathrm{d}t) = \frac{1}{M(N+M)}\sum_{i=1}^{M}\sum_{j=1}^{N} d_i A_{ij}^2 \leq \frac{N}{N+M} d_{\max} S,$$

then choosing $L = d_{\mathrm{BL}}(\hat{\mu}_{\widetilde{H}}, \mathbb{E}[\hat{\mu}_{\widetilde{H}}])^{-1/3}$ to obtain

$$d_{\mathrm{BL}}(\hat{\mu}_{\widetilde{H}}\sharp x^2, \mathbb{E}[\hat{\mu}_{\widetilde{H}}]\sharp x^2) \leq d_{\mathrm{BL}}(\hat{\mu}_{\widetilde{H}}, \mathbb{E}[\hat{\mu}_{\widetilde{H}}]) + d_{\mathrm{BL}}(\hat{\mu}_{\widetilde{H}}, \mathbb{E}[\hat{\mu}_{\widetilde{H}}])^{2/3}(2 + 2\frac{N}{N+M} d_{\max} S).$$

Applying Lemma A.2 twice finishes the proof. □

*Proof of Lemma A.2.* This follows immediately from Theorem 1.3(b) of [19], thinking of their matrix $A$ as our matrix

$$\sqrt{\frac{N+M}{M}}\begin{pmatrix} 0 & 11^T\Gamma^{1/2} \\ \Gamma^{1/2}11^T & 0 \end{pmatrix},$$

(i.e., the top-right block has constant columns and the bottom-left block has constant rows), which has entries uniformly bounded by $\sqrt{d_{\max}\frac{N+M}{M}}$. □

*Proof of Lemma A.3.* Take $f$ such that $\|f\|_\infty + \|f\|_{\text{Lip}} \leq 1$, define $g$ by $g(x) = f(x^2)$, and for given $L$ define $g_L$ by

$$g_L(x) = \begin{cases} g(x) & x \leq L, \\ (L+1-x)g(L) & \text{if } L \leq x \leq L+1, \\ 0 & \text{if } x \geq L+1. \end{cases}$$

We have $\|g_L\|_\infty \leq 1$ and $\|g_L\|_{\text{Lip}} \leq 2L$ (technically for the restriction to the right half-line) as well as $|g(x) - g_L(x)| \leq \mathbb{1}\{x > L\}$ for $x \geq 0$. Thus

$$\begin{aligned} \left| \int f(y)(\mu \sharp x^2 - \nu \sharp x^2)(\mathrm{d}y) \right| &= \left| \int g(y)(\mu - \nu)(\mathrm{d}y) \right| \\ &\leq \left| \int g_L(y)(\mu - \nu)(\mathrm{d}y) \right| + \mu((L,\infty)) + \nu((L,\infty)) \\ &\leq (1+2L) d_{\text{BL}}(\mu,\nu) + \mu((L,\infty)) + \nu((L,\infty)). \qquad \square \end{aligned}$$

*Proof of Lemma A.4.* Since

$$\widetilde{H}^2 = \frac{1}{M} \begin{pmatrix} A^T \Gamma A & 0 \\ 0 & \Gamma^{1/2} A A^T \Gamma^{1/2} \end{pmatrix},$$

for any measurable function $f$ we have

$$\operatorname{tr}(f(\widetilde{H}^2)) = 2 \operatorname{tr}(f(H)) + (\max(M,N) - \min(M,N)) f(0),$$

and thus

$$\begin{aligned} &\int f(y)(\hat{\mu}_{\widetilde{H}} \sharp x^2 - \mathbb{E}[\hat{\mu}_{\widetilde{H}}] \sharp x^2)(\mathrm{d}y) \\ &= \int f(y^2)(\hat{\mu}_{\widetilde{H}} - \mathbb{E}[\hat{\mu}_{\widetilde{H}}])(\mathrm{d}y) = \frac{1}{N+M}(\operatorname{tr}(f(\widetilde{H}^2)) - \mathbb{E}[\operatorname{tr}(f(\widetilde{H}^2))]) \\ &= \frac{2}{N+M}(\operatorname{tr}(f(H)) - \mathbb{E}[\operatorname{tr}(f(H))]) = 2\frac{N}{N+M} \int f(y)(\hat{\mu}_H - \mathbb{E}[\hat{\mu}_H])(\mathrm{d}y). \qquad \square \end{aligned}$$

## B Concentration for multidimensional product measures

This appendix deals with a straightforward extension of classic results of Talagrand [37] and Guionnet-Zeitouni [19] on concentration for product measures, in order to consider complex-Hermitian random matrices with real and imaginary parts that are not necessarily independent of one another.

In the 1990s, Talagrand developed a theory of concentration for products of compactly-supported measures, obtaining results of the form "If $f : [-1,1]^N \to \mathbb{R}$ is Lipschitz and has convex sublevel sets, and $(X_i)_{i=1}^N$ are independent random variables each valued in $[-1,1]$, then the random variable $f(X_1, \dots, X_N)$ concentrates about its median" [37, Theorem 6.6]. Guionnet and Zeitouni translated his results into random matrices, using them to show results of the form "If the real-symmetric or complex-Hermitian Wigner matrix $W_N$ has compactly-supported entries, then $\hat{\mu}_{W_N}$ concentrates about its mean $\mathbb{E}[\hat{\mu}_{W_N}]$ in Wasserstein-1 distance" [19, Corollary 1.4(b)]. However, since Talagrand's result was written for the most digestible case of $f : [-1,1]^N \to \mathbb{R}$, the complex-Hermitian case Guionnet and Zeitouni's result required the entries of $W_N$ to have *independent* real and imaginary parts; then linear statistics of $W_N$ could indeed be nice functions of the $N^2$ independent random variables $(\operatorname{Re} W_{ij}, \operatorname{Im} W_{ij})_{1 \leq i < j \leq N} \cup (W_{ii})_{i=1}^N$.

We want to prove results about slightly more general Wigner matrices $W_N$, where the real and imaginary parts of each $W_{ij}$ are allowed to be correlated with each other, as long as the entries $W_{ij} \in \mathbb{C}$ remain independent for different upper-triangular values of $i$ and $j$. In order to do this, we need to extend Corollary 1.4(b) of [19], which in turn requires the following extension of Theorem 6.6 of [37]. We copy Talagrand's language and most of his notation, so that the reader can more easily compare, but we introduce the $d$-dimensional Euclidean unit balls

$$B_d = \{x \in \mathbb{R}^d : \|x\|_2^2 \leq 1\}.$$

**Proposition B.1.** *Consider a real-valued function $f$ defined on $(B_d)^N$. We assume that, for each real number $a$,*

$$\text{the set } \{f \leq a\} \text{ is convex.}$$

*Consider a convex set $C \subset (B_d)^N$, consider $\sigma > 0$ and assume that the restriction of $f$ to $C$ has a Lipschitz constant at most $\sigma$; that is,*

$$\forall\, x, y \in C, \qquad |f(x) - f(y)| \leq \sigma \|x - y\|,$$

*where $\|x\|$ denotes the norm $\|(x_1, \dots, x_N)\|^2 = \sum_{i=1}^N \|x_i\|_2^2$.*

*Consider independent random variables $(X_i)_{i \leq N}$ valued in $B_d$, and consider the random variable*

$$h = f(X_1, \dots, X_N).$$

*Then, if $M$ is a median of $h$, we have, for all $t > 0$, that*

$$\mathbb{P}(|h - M| \geq t) \leq 4c + \frac{4}{1 - 2c} \exp\left(-\frac{t^2}{16\sigma^2}\right) \tag{B.1}$$

*where we assume*

$$c = \mathbb{P}((X_1, \dots, X_N) \notin C) < \frac{1}{2}.$$

The proof of Proposition B.1 is almost the same as Talagrand's $d = 1$ original; one can quickly check that eq. (6.20) in [37] is still valid in our setting, and the remainder of Talagrand's proof of [37, Theorem 6.6] goes through verbatim. We remark that (B.1) has no dependence on $d$, which may be initially surprising, since we make no assumptions about the correlations between the $d$ entries of each $X_i$. However, the lack of $d$-dependence is essentially because we have chosen to extend Talagrand's $d = 1$ compact set $[-1, 1]$ to $B_d$, which has Euclidean diameter $2$ for each $d$, rather then, e.g., to replace $[-1, 1]$ with $[-1, 1]^d$, which has Euclidean diameter $2\sqrt{d}$. (If we instead considered $f : ([-1, 1]^d)^N \to \mathbb{R}$ and variables $X_i \in [-1, 1]^d$, we would obtain a variant of (B.1) with right-hand side $4c + \frac{4}{1-2c} \exp\left(-\frac{t^2}{16d\sigma^2}\right)$.)

We suspect that an extension of this form has already appeared in the literature, perhaps more than once, but we have not been able to find it.

By thinking $\mathbb{C} \cong \mathbb{R}^2$ and using the $d = 2$ case of Proposition B.1, we obtain the following extension of Corollary 1.4(b) of [19]. Again we copy their language and notation for ease of comparison. We consider inhomogeneous complex-Hermitian random matrices $X_A$ given by

$$X_A = ((X_A)_{ij})_{1 \leq i,j \leq N}, \quad X_A = X_A^*, \quad (X_A)_{ij} = \frac{1}{\sqrt{N}} A_{ij} \omega_{ij}$$

with

$$\begin{aligned} &\omega = (\omega^R + \mathrm{i}\omega^I) = (\omega_{ij})_{1 \leq i,j \leq N} = (\omega_{ij}^R + \sqrt{-1}\omega_{ij}^I)_{1 \leq i,j \leq N}, \quad \omega_{ij} = \overline{\omega_{ji}}, \\ &A = (A_{ij})_{1 \leq i,j \leq N}, \quad A_{ij} = \overline{A_{ji}}. \end{aligned}$$

Here $\{\omega_{ij}, 1 \le i \le j \le N\}$ are independent complex random variables with laws $\{P_{ij}, 1 \le i \le j \le N\}$, and the $P_{ij}$'s are probability measures on $\mathbb{C}$, but now with no assumptions on the relationship between their real and imaginary marginals, except the condition that $P_{ii}$ is supported on $\mathbb{R}$ in order to keep $X_A$ Hermitian. Here $A$ is a non-random complex matrix with entries $\{A_{ij}, 1 \le i \le j \le N\}$ uniformly bounded by, say, $a$. (By choosing all the $P_{ij}$'s to be supported on $\mathbb{R}$ and all entries of $A$ real, we can of course obtain results about real-symmetric random matrices.)

**Proposition B.2.** *Assume that the $(P_{ij}, i \le j)$ are uniformly compactly supported, that is that there exists a compact set $K \subset \mathbb{C}$ so that for any $1 \le i \le j \le N$, $P_{ij}(K^c) = 0$. Write*

$$\|K\| = \max\{\|x\| : x \in K\}.$$

*Fix $\delta_1(N) = 8\|K\|\sqrt{\pi}a/N$ and $M = a\|K\|\sqrt{8}$. For any $\delta > (128(M + \sqrt{\delta})\delta_1(N))^{2/5}$,*

$$\begin{aligned}&\mathbb{P}(\mathrm{W}_1(\hat{\mu}_{X_A}, \mathbb{E}[\hat{\mu}_{X_A}]) > \delta) \\ &\le \frac{128(M + \sqrt{\delta})}{\delta^{3/2}} \exp\left(-N^2 \frac{1}{16\|K\|^2 a^2}\left[\frac{\delta^{5/2}}{128(M+\sqrt{\delta})} - \delta_1(N)\right]^2\right).\end{aligned}$$

We again omit the proof, which just mimics that of Guionnet and Zeitouni; the only observation is that here $\|K\|$ is defined in such a way that $\frac{K}{\|K\|} \subset B_2$, so that we can use Proposition B.1 when Guionnet and Zeitouni use [37, Theorem 6.6].

## C The "compact or log-Sobolev" assumption

In this appendix we explain how to remove a certain technical assumption from previous tilting results on top-eigenvalue LDPs in the sub-Gaussian case, which we will call the "compact or log-Sobolev" assumption.

This assumption appears in various forms throughout the literature: earlier papers tend to literally require some underlying measure to have either compact support or to satisfy the log-Sobolev inequality, whereas later papers tend to require some statement about concentration of the empirical measure which is easy to verify in the compact-or-log-Sobolev case.

Our techniques to remove this assumption use a recent strengthening of the continuity properties of spherical integrals, due to Guionnet and the first author [17]. This works as follows: With $\hat{\mu}_N$ the empirical spectral measure of the matrix one is studying as defined in (1.4), $\mu_\infty$ its deterministic limit, and $d_{\mathrm{BL}}$ the bounded-Lipschitz distance from (1.3), previous results required estimates of the form

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_N, \mu_\infty) > N^{-\delta}) = -\infty \tag{C.1}$$

for some small $\delta > 0$. Under the better continuity properties, it suffices to show

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_N, \mu_\infty) > \epsilon) = -\infty \tag{C.2}$$

for every $\epsilon > 0$. We see two main benefits of (C.2) over (C.1): First, we are about to show that (C.2) is often provable without the compact-or-log-Sobolev assumption, essentially by carefully truncating the matrix entries and using concentration results of Guionnet and Zeitouni, in the style of Talagrand, for compactly supported product measures. Second, one can typically show (C.2) without relying on local laws, which had been used in previous results to verify (C.1) (see, e.g., [31, 22]). Since local laws are only available in some cases, we think it may be useful for future LDP results that their use can be bypassed.

We demonstrate the ideas in the simplest setting of sharp sub-Gaussian Wigner matrices, showing the following result, which removes the "compact-or-log-Sobolev" assumption from the main result of [16].

**Theorem C.1.** *Let $\mu$ be a centered probability measure on $\mathbb{R}$ with unit variance, and let $X_N$ be the corresponding Wigner matrix, i.e., $X_N$ is an $N \times N$ real-symmetric random matrix with i.i.d. entries up to symmetry distributed according to $\frac{\mu}{\sqrt{N}}$. If $\mu$ is sharp sub-Gaussian, then $\lambda_{\max}(X_N)$ satisfies an LDP at speed $N$ with the good rate function*

$$I(x) = \begin{cases} +\infty & \text{if } x < 2, \\ \frac{1}{2}\int_2^x \sqrt{y^2-4}\,\mathrm{d}y & \text{if } x \geq 2. \end{cases}$$

Shortly prior to the posting of this paper on the arXiv, we learned that simultaneous independent work of Cook, Ducatez, and Guionnet [10] also removes the compact-or-log-Sobolev assumption, but with different techniques. Roughly speaking, our argument works at the matrix level, while their argument works at the scalar level: in their Appendix A they prove a tail bound for convex Lipschitz functions of sub-Gaussian variables, upgrading Talagrand's classical tail bound for convex Lipschitz functions of compactly supported variables.

*Proof.* We claim that, for every $\epsilon > 0$, we have

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{X_N}, \rho_{\mathrm{sc}}) > \epsilon) = -\infty. \tag{C.3}$$

The original paper [16] of Guionnet and the first author shows something slightly stronger, under the compact-or-log-Sobolev assumption, namely that there exists some small $\kappa > 0$ with

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{X_N}, \rho_{\mathrm{sc}}) > N^{-\kappa}) = -\infty.$$

However, mimicking the arguments in our Sections 3.4 and 3.5 shows that (C.3) suffices.

To prove (C.3), we mimic the proof of Proposition 3.12, decomposing $X = A + B$ with $A_{ij} = X_{ij}\mathbb{1}\{|X_{ij}| \leq N^{\gamma-1/2}\}$ for some $\gamma = \gamma(\epsilon) > 0$ to be chosen, and then prove analogues of (3.4), (3.5), (3.6), and (3.7). The proof of (3.4) is as in the sample-covariance case, except that we use the classical result $d_{\mathrm{KS}}(\hat{\mu}_{X_N}, \hat{\mu}_{A_N}) \leq \frac{1}{N}\operatorname{rank}(X_N - A_N)$ [2, Theorem A.43], which comes from interlacing of eigenvalues, instead of [2, Theorem A.44] from interlacing of singular values as in the main text. The estimates (3.5) and (3.7) are just as in the main text, and the estimate (3.6) is actually easier in the Wigner case: Define $P = \sqrt{8}N^{\gamma}$ and $\epsilon_1(N) = 16\sqrt{\pi}N^{\gamma-1}$. Since $\sqrt{N}A_N$ has entries compactly supported in $[-N^{\gamma}, N^{\gamma}]$, [19, Theorem 1.4(a)] gives

$$\mathbb{P}(\mathrm{W}_1(\hat{\mu}_{A_N}, \mathbb{E}[\hat{\mu}_{A_N}]) > \epsilon) \leq \frac{128(P+\sqrt{\epsilon})}{\epsilon^{3/2}} \exp\left(-N^2 \frac{1}{64N^{2\gamma}}\left[\frac{\epsilon^{5/2}}{128(P+\sqrt{\epsilon})} - \epsilon_1(N)\right]^2\right)$$

as long as $\epsilon$ satisfies the implicit equation $\epsilon > (128(P+\sqrt{\epsilon})\epsilon_1(N))^{2/5}$, the right-hand side of which is order $N^{(2\gamma-1)(2/5)}$, so that the implicit equation is satisfied for all $N$ large enough. Then the argument of the exponential is order $N^{2-2\gamma+2(-\gamma)}$, and the power in the exponential is at least $1$ for $\gamma$ small enough, which suffices. □

**Remark C.2.** This proof allows the laws of the entries of $X_N$ to be sharp sub-Gaussian without having to be compactly supported or satisfying a log-Sobolev inequality. Let us provide an example of such a law. For this, let us consider $Z$ a standard Gaussian variable, $\mathcal{F} = \sigma(\{Z \in [n, n+1[\} : n \in \mathbb{Z})$ and let us define

$$\tilde{Z} = \operatorname{sign}(Z)\sqrt{\mathbb{E}[Z^2|\mathcal{F}]}.$$

The variable $\tilde{Z}$ is centered of variance $1$. It is obviously not compactly supported, and since it is supported on a discrete set of points, it cannot satisfy any log-Sobolev inequality. It remains to see that $\tilde{Z}$ is sharp sub-Gaussian. For this, let us look at its moments. Since the law of $\tilde{Z}$ is symmetric, for $k \in \mathbb{N}$ we have

$$\mathbb{E}[\tilde{Z}^{2k+1}] = 0$$

and for even moments, applying Jensen's inequality, we have

$$\mathbb{E}[\tilde{Z}^{2k}] = \mathbb{E}[\mathbb{E}[Z^2|\mathcal{F}]^k] \leq \mathbb{E}[Z^{2k}]$$

and $\mathrm{Var}(\tilde{Z}) = \mathbb{E}[\tilde{Z}^2] = \mathbb{E}[\mathbb{E}[Z^2|\mathcal{F}]] = \mathbb{E}[Z^2] = 1$. So since $Z$ is Gaussian, we have for $t \in \mathbb{R}$

$$\mathbb{E}[\exp(t\tilde{Z})] \leq \mathbb{E}[\exp(tZ)] \leq \exp\left(\frac{t^2}{2}\right),$$

meaning $\tilde{Z}$ is indeed sharp sub-Gaussian.

## D Deformed Wigner matrices

In this appendix, we use our techniques to improve previous results of the second author for so-called "deformed Wigner matrices" [31]. In this model, one considers an $N \times N$ random matrix

$$X_N = \frac{W_N}{\sqrt{N}} + D_N,$$

either real-symmetric or complex-Hermitian, where $W_N$ has i.i.d. entries up to symmetry each distributed according to some centered probability measure $\mu$ (on $\mathbb{R}$ if $\beta = 1$ or on $\mathbb{C}$ if $\beta = 2$), and where the deterministic matrix $D_N$ satisfies the following assumption, which is weaker than the corresponding assumption from [31].

**Assumption D.1.** *The matrix $D_N$ is real, diagonal, and deterministic, and its empirical measure $\hat{\mu}_{D_N}$ tends weakly as $N \to \infty$ to a compactly supported probability measure $\mu_D$, and there are asymptotically no external outliers, in the sense that*

$$\lambda_{\max}(D_N) \to r(\mu_D),$$
$$\lambda_{\min}(D_N) \to \ell(\mu_D).$$

**Remark D.2.** A word on notation: Many of the objects we used in studying the generalized sample-covariance-case have analogues here. We have chosen to overload the notation rather than cluttering it. For example, the generalized-sample-covariance case has a threshold $x_c$, defined in (2.3). We will need an analogous threshold here, and even though the definition (D.1) is different, we still call the new version $x_c$ rather than, e.g., $x_c^{(\mathrm{dw})}$. We have similarly overloaded $H$, the rate function $I$, and so on, with one notable exception: The special function $J(\mu, \theta, \lambda)$ given in Definition 3.1 is exactly the same in both models.

The typical limiting behavior of the empirical measure for this model is classical [34, 38]: We have

$$\hat{\mu}_{X_N} \to \rho_{\mathrm{sc}} \boxplus \mu_D =: \mu_D^{\mathrm{sc}},$$

where $\rho_{\mathrm{sc}}$ is the semicircle law $\rho_{\mathrm{sc}}(\mathrm{d}x) = \frac{\sqrt{(4-x^2)_+}}{2\pi}\,\mathrm{d}x$, the notation $\boxplus$ denotes the free (additive) convolution of two compactly supported probability measures, and $\mu_D^{\mathrm{sc}}$ is shorthand. We recall that $\boxplus$ is defined in terms of the Voiculescu $R$-transform, which will be important for us: For a compactly supported probability measure $\mu$, we recall the Stieltjes transform $G_\mu$, which is a decreasing bijection from $(r(\mu), +\infty)$

to $(0, G_\mu(r(\mu)))$. We write its inverse as $K_\mu : (0, G_\mu(r(\mu))) \to (r(\mu), +\infty)$, and its $R$-transform as $R_\mu(y) = K_\mu(y) - \frac{1}{y}$, which linearizes free convolution in the sense that $R_{\mu \boxplus \nu} = R_\mu + R_\nu$.

The following lemma defines functions we need to write the rate function.

**Lemma D.3.** *The function $H : (0, \infty) \to \mathbb{R}$ defined by*

$$H(y) := \begin{cases} y + K_{\mu_D}(y) & \text{if } 0 \leq y \leq G_{\mu_D}(r(\mu_D)), \\ y + r(\mu_D) & \text{if } y \geq G_{\mu_D}(r(\mu_D)), \end{cases}$$

*has the following properties:*

- *$H$ is continuous and convex.*
- *$H$ is uniquely minimized at $y_c = G_{\rho_{\mu_D^{\text{sc}}}}(r(\mu_D^{\text{sc}}))$, which is in $(0, G_{\mu_D}(r(\mu_D))]$, and $H(y_c) = r(\mu_D^{\text{sc}})$.*
- *There exists a continuous, strictly increasing function $\widetilde{\mathfrak{G}} : [r(\mu_D^{\text{sc}}), +\infty) \to \mathbb{R}$ with the following properties:*
  - *If $0 \leq y \leq G_{\mu_D}(r(\mu_D))$, then $H(\widetilde{\mathfrak{G}}(y)) = y$, and*
  $$\{w : H(w) = y\} = \{G_{\mu_D^{\text{sc}}}(y), \widetilde{\mathfrak{G}}(y)\}.$$
  - *We have $\widetilde{\mathfrak{G}}(y) > G_{\mu_D^{\text{sc}}}(y)$ for $y > r(\mu_D^{\text{sc}})$, and $\widetilde{\mathfrak{G}}(r(\mu_D^{\text{sc}})) = G_{\mu_D^{\text{sc}}}(r(\mu_D^{\text{sc}}))$.*

**Definition D.4.** *Define $I : \mathbb{R} \to [0, +\infty]$ by*

$$I(x) = \begin{cases} +\infty & \text{if } x < r(\rho_{\text{sc}} \boxplus \mu_D), \\ \frac{1}{2} \int_{r(\rho_{\text{sc}} \boxplus \mu_D)}^{x} (\widetilde{\mathfrak{G}}(y) - G_{\rho_{\text{sc}} \boxplus \mu_D}(y)) \, \mathrm{d}y & \text{if } x \geq r(\rho_{\text{sc}} \boxplus \mu_D). \end{cases}$$

**Theorem D.5.** *Suppose that Assumption D.1 holds, and that $\mu$ is sharp sub-Gaussian (in the sense of Definition 2.1 if $\beta = 1$, or in the sense of Definition 2.20 if $\beta = 2$). Then $\lambda_{\max}(X_N)$ satisfies a large deviation principle at speed $N$ with the good rate function $I^{(\beta)} = \beta I$. This function is convex, strictly increasing on $[r(\rho_{\text{sc}} \boxplus \mu_D), +\infty)$ (in particular, vanishes uniquely at $r(\rho_{\text{sc}} \boxplus \mu_D)$), and*

$$\lim_{x \to +\infty} \frac{I^{(\beta)}(x)}{\frac{\beta x^2}{4}} = 1.$$

*If $\mu$ is actually Gaussian, then by rotational invariance we do not need to assume that $D_N$ is diagonal; we can just assume it is symmetric (if $\beta = 1$) or Hermitian (if $\beta = 2$) and satisfies the rest of Assumption D.1.*

**Remark D.6.** This improves upon the result of [31], which showed a weaker version of Theorem D.5 requiring three additional assumptions: First, that either $\mu$ actually be Gaussian measure (i.e., that $\frac{W_N}{\sqrt{N}}$ be GOE/GUE), or that the important threshold

$$x_c = x_c(\mu_D) := \begin{cases} r(\mu_D) + G_{\mu_D}(r(\mu_D)) & \text{if } G_{\mu_D}(r(\mu_D)) < +\infty, \\ +\infty & \text{otherwise} \end{cases} \tag{D.1}$$

be infinite; second, that $\mu$ be either compactly supported or satisfy the log-Sobolev inequality; third, that the deformation tend to its limit at some mild polynomial speed, $d_{\text{BL}}(\hat{\mu}_{D_N}, \mu_D) \lesssim N^{-\epsilon}$ for some $\epsilon > 0$. However, the rate function was given there in a different form; below we show that the forms are equivalent.

We get rid of the second and third assumptions using the methods of Appendix C. We get rid of the first assumption as in the main text, namely by approximating $x_c < +\infty$ models with a sequence of $x_c = +\infty$ models, then using textbook results about approximating LDPs.

**Definition D.7.** *For any $x \geq r(\rho_{\text{sc}} \boxplus \mu_D)$ and any $\theta \geq 0$, define*

$$I(x,\theta) := J(\rho_{\text{sc}} \boxplus \mu_D, \theta, x) - \theta^2 - J(\mu_D, \theta, r(\mu_D)).$$

*Using this, define the rate function $\widetilde{I} : \mathbb{R} \to [0,+\infty]$ by*

$$\widetilde{I}(x) := \begin{cases} +\infty & \text{if } x < r(\rho_{\text{sc}} \boxplus \mu_D), \\ \sup_{\theta \geq 0} I(x,\theta) & \text{if } x \geq r(\rho_{\text{sc}} \boxplus \mu_D). \end{cases}$$

**Lemma D.8.** *If $x_c(\mu_D) = +\infty$, then $I = \widetilde{I}$.*

**Lemma D.9.** *If Assumption D.1 holds, and $\mu$ is sharp sub-Gaussian, and*

$$x_c(\mu_D) = +\infty,$$

*then $\lambda_{\max}(X_N)$ satisfies a large deviation principle at speed $N$ with the good rate function $I$.*

**Lemma D.10.** *Lemma D.9 implies Theorem D.5.*

*Proof of Lemma D.3.* Continuity is easy to check. Since $G_{\mu_D}$ is strictly convex on $(r(\mu_D), +\infty)$, the function $K_{\mu_D}$ is strictly convex on $(0, G_{\mu_D}(r(\mu_D)))$, so that $H$ is convex, indeed strictly on $(0, G_{\mu_D}(r(\mu_D)))$. Since the boundary behavior is $\lim_{y \to 0} H(y) = \lim_{y \to +\infty} H(y) = +\infty$, we have that $H$ is uniquely minimized at some $y_c$, which must be in $(0, G_{\mu_D}(r(\mu_D))]$. On the other hand, let $y_s := G_{\mu_D^{\text{sc}}}(r(\mu_D^{\text{sc}}))$. Lemma 6.1 of [18] gives $y_s \leq \min(G_{\rho_{\text{sc}}}(r(\rho_{\text{sc}})), G_{\mu_D}(r(\mu_D)))$, and a computation in the proof of Proposition 6.1 in [31] shows $G'_{\mu_D}(y_s) = -1$, and thus, via differentiating $y = K_{\mu_D}(G_{\mu_D}(y))$ in $y$ and evaluating at $y_s$, that $H'(y_s) = 0$. Thus $y_s = y_c$. Another computation in the proof of Proposition 6.1 in [31] shows $H(y_s) = r(\mu_D^{\text{sc}})$.

Now we study $\widetilde{\mathfrak{G}}$. Since $R_{\rho_{\text{sc}}}(y) = y$ in our normalization, we have $H(y) = R_{\rho_{\text{sc}}}(y) + R_{\mu_D}(y) + \frac{1}{y} = K_{\rho_{\text{sc}} \boxplus \mu_D}(y)$ for $y < y_c$. Since $H$ is continuous, strictly convex on $(0,b)$ for some $b$, affine increasing on $(b,+\infty)$, and minimized at $y_c \in (0,b]$, it is easy to see that $\#\{w : H(w) = y\}$ is zero for $y < H(y_c)$, one for $y = H(y_c)$, and two for $y > H(y_c)$; that the smaller of these elements is $G_{\rho_{\text{sc}} \boxplus \mu_D}(y)$; and that, if $\widetilde{\mathfrak{G}}$ denotes the inverse of $H$ on $(y_c, +\infty)$, then $\widetilde{\mathfrak{G}}$ has the claimed properties. □

*Proof of Lemma D.8.* The proof goes as in the generalized-sample-covariance case, i.e., by showing that $I$ and $\widetilde{I}$ have the same derivative on $(r(\rho_{\text{sc}} \boxplus \mu_D), +\infty)$, and both vanish at $r(\rho_{\text{sc}} \boxplus \mu_D)$. We rely on two computations already carried out in Proposition 6.1 of [31]. The first of these shows that $\widetilde{I}$ vanishes at $r(\rho_{\text{sc}} \boxplus \mu_D)$. The second shows that, for each $x > r(\rho_{\text{sc}} \boxplus \mu_D)$, we have $\widetilde{I}(x) = I(x, \theta_x)$, where $\theta_x$ is the unique solution to the constrained problem

$$H(2\theta_x) = 2\theta_x + K_{\mu_D}(2\theta_x) = x \quad \text{subject to} \quad 2\theta_x \in (G_{\rho_{\text{sc}} \boxplus \mu_D}(r(\rho_{\text{sc}} \boxplus \mu_D)), G_{\mu_D}(r(\mu_D))),$$

which in the new language of branches of Stieltjes transforms we recognize as $\theta_x = \frac{1}{2}\widetilde{\mathfrak{G}}(x)$; and that this maximizer is unique, i.e., $I(x,\theta) < \widetilde{I}(x)$ if $\theta \neq \theta_x$. Thus

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\widetilde{I}(x) = \frac{\partial}{\partial x} I(x,\theta)\bigg|_{\theta=\theta_x} &= \frac{\partial}{\partial x} J(\rho_{\text{sc}} \boxplus \mu_D, \theta, x)\bigg|_{\theta=\theta_x} \\ &= \theta_x - \frac{1}{2} G_{\rho_{\text{sc}} \boxplus \mu_D}(x) = \frac{1}{2}(\widetilde{\mathfrak{G}}(x) - G_{\rho_{\text{sc}} \boxplus \mu_D}), \end{aligned}$$

which completes the proof. □

*Proof of Lemma D.9.* As stated above, this is the main result of [31], except that (a) the rate function was written there as $\beta\widetilde{I}$ (which Lemma D.8 shows is irrelevant), and (b) that paper required $\mu$ to be either compactly supported or to satisfy log-Sobolev, and required

$$d_{\mathrm{BL}}(\hat{\mu}_{D_N}, \mu_D) \leq CN^{-\epsilon} \tag{D.2}$$

for some $C$ and $\epsilon$. Under the additional (b) assumptions, Lemma 5.3 of [31] showed

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{X_N}, \rho_{\mathrm{sc}} \boxplus \mu_D) > N^{-\kappa}) = -\infty$$

for $\kappa > 0$ sufficiently small. Obtaining this small polynomial speed (a) required (a very weak consequence) of the local law [13], and (b) was the essential reason for requiring (D.2). But Appendix C explains that it actually suffices to show

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{X_N}, \rho_{\mathrm{sc}} \boxplus \mu_D) > \epsilon) = -\infty \tag{D.3}$$

for every $\epsilon > 0$; this allows us to drop the requirement of (D.2), and to give a proof bypassing the local law. To do this, as in Appendix C, we split $W_N N^{-1/2} = A + B = A_N + B_N$ with $A_{ij} = W_{ij}N^{-1/2}\mathbb{1}\{\left|W_{ij}N^{-1/2}\right| \leq N^{\gamma-1/2}\}$ for some $\gamma = \gamma(\epsilon) > 0$ to be chosen, then decompose $X_N = (A + D_N) + (B + D_N)$ and show analogues of (3.4), (3.5), (3.6), and (3.7). The analogues of (3.4), (3.5), and (3.7) go through exactly as before. The analogue of (3.6) is the estimate

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{A+D_N}, \mathbb{E}[\hat{\mu}_{A+D_N}]) > \epsilon) = -\infty.$$

In Appendix C, when $D_N = 0$, we noted that $\sqrt{N}A$ had entries compactly supported in $[-N^\gamma, N^\gamma]$, and applied results of [19]. When $D_N = 0$, the matrix $\sqrt{N}(A + D_N)$ has entries compactly supported in boxes of size order $N^\gamma$, but whose centers have shifted away from zero. This requires a straightforward modification of the results of [19], which was already stated as Lemma 5.9 of [31]. This completes the proof of (D.3). □

*Proof of Lemma D.10.* This proof also goes as in the generalized-sample-covariance case. We use throughout results of Lemma D.3, and only write the case $\beta = 1$ for simplicity.

From its definition and Lemma D.3, we see that the rate function has the form $I(x) = \frac{1}{2}\int_{r(\rho_{\mathrm{sc}} \boxplus \mu_D)}^x g(y)\,\mathrm{d}y$, where $g$ is strictly increasing, positive for arbitrarily small arguments, and $\lim_{x\to+\infty} \frac{g(x)}{x} = 1$; this proves the claimed properties of $I$.

Fix once and for all some $\mu_D$ with $x_c(\mu_D) < +\infty$, and write $d_i = d_i^{(N)}$ for the diagonal entries of $D_N$, i.e., $D_N = \mathrm{diag}(d_1, \dots, d_N)$. For $\epsilon > 0$, define $D_N^{(\epsilon)} = \mathrm{diag}(d_1^{(\epsilon)}, \dots, d_N^{(\epsilon)})$, where $d_i^{(\epsilon)} = d_i$ if $d_i \leq r(\mu_D) - \epsilon$ and $d_i^{(\epsilon)} = r(\mu_D)$ otherwise. Then set

$$X_N^{(\epsilon)} = N^{-1/2}W_N + D_N^{(\epsilon)},$$

coupled with $X_N$ by using the same randomness $W_N$. Set $\mu_D^{(\epsilon)}$ as in (4.1), and notice that if $\epsilon \to 0$ avoids any atoms present near $r(\mu_D)$, we have that the empirical measure of $X_N^{(\epsilon)}$ tends to $\rho_{\mathrm{sc}} \boxplus \mu_D^{(\epsilon)}$, with a corresponding function $H^{(\epsilon)}$. By construction, $r(\mu_D^{(\epsilon)}) = r(\mu_D)$, but $G_{\mu_D^{(\epsilon)}}(r(\mu_D^{(\epsilon)})) = +\infty$ for each $\epsilon$, so that $\lambda_{\max}(X_N^{(\epsilon)})$ satisfies an LDP at speed $N$ with the good rate function $I^{(\epsilon)}$ defined as

$$I^{(\epsilon)}(x) = \begin{cases} +\infty & \text{if } x < r(\rho_{\mathrm{sc}} \boxplus \mu_D^{(\epsilon)}), \\ \frac{1}{2}\int_{r(\rho_{\mathrm{sc}} \boxplus \mu_D)}^x (\widetilde{\mathfrak{G}}^{(\epsilon)}(y) - G_{\rho_{\mathrm{sc}} \boxplus \mu_D^{(\epsilon)}}(y))\,\mathrm{d}y & \text{if } x \geq r(\rho_{\mathrm{sc}} \boxplus \mu_D^{(\epsilon)}), \end{cases}$$

where $\widetilde{\mathfrak{G}}^{(\epsilon)}$ is defined as in Lemma D.3 for the measure $\mu_D^{(\epsilon)}$.

To finish the proof, we just need analogues of Lemmas 4.2, 4.3, and 4.4. The analogue of Lemma 4.3 is even easier here, since $\|X_N - X_N^{(\epsilon)}\| \leq \epsilon$ deterministically. Lemma 4.4 was essentially a consequence of Lemma 4.2, and this remains true here, so we only need the analogue of Lemma 4.2.

Towards this result: Similar arguments as in the main text show that, for each $x > r(\mu_D) = r(\mu_D^{(\epsilon)})$, the function $\epsilon \mapsto G_{\mu_D^{(\epsilon)}}(x)$ is non-decreasing for small $\epsilon > 0$, so that $\epsilon \mapsto K_{\mu_D^{(\epsilon)}}(y)$ is non-decreasing, and thus $\epsilon \mapsto H^{(\epsilon)}(y)$ is non-decreasing. Thus $\epsilon \mapsto r(\rho_{\text{sc}} \boxplus \mu_D^{(\epsilon)})$ is non-decreasing, and $\liminf_{\epsilon \downarrow 0} r(\rho_{\text{sc}} \boxplus \mu_D^{(\epsilon)}) \geq r(\rho_{\text{sc}} \boxplus \mu_D)$. For the other inequality, it is clear that $H^{(\epsilon)}$ converges to $H$ uniformly on all compact subsets of $(0, G_{\mu_D}(r(\mu_D))]$. But $H^{(\epsilon)}(y) - y$ is strictly decreasing in $y$, and tends to $r(\mu_D)$ as $y \to +\infty$; thus $H^{(\epsilon)}$ tends to $H$ uniformly on all compact subsets of $(0, +\infty)$, and hence $\lim_{\epsilon \to 0} r(\rho_{\text{sc}} \boxplus \mu_D^{(\epsilon)}) = r(\rho_{\text{sc}} \boxplus \mu_D)$. Finally, for $r > r(\rho_{\text{sc}} \boxplus \mu_D)$ and $\epsilon$ small enough we have the representations

$$I^{(\epsilon)}(x) = \frac{1}{2} \int_0^x \int_0^\infty \mathbb{1}_{H^{(\epsilon)}(u) \leq t} \, \mathrm{d}u \, \mathrm{d}t \quad \text{and} \quad I(x) = \frac{1}{2} \int_0^x \int_0^\infty \mathbb{1}_{H(u) \leq t} \, \mathrm{d}u \, \mathrm{d}t,$$

so that again $I^{(\epsilon)}(x) \leq I(x)$ with $\sup_{x \in [a,b]} \left| I^{(\epsilon)}(x) - I(x) \right| \leq \mathrm{Leb}(D_b^{(\epsilon)})$, with $D_b^{(\epsilon)}$ redefined appropriately. The sets $D_b^{(\epsilon)}$ are again nested with empty intersection, so their Lebesgue measure tends to zero if it is finite. Before this finiteness was immediate, but here takes a moment's thought: Since $\lim_{y \to +\infty} H(y) = +\infty$, there exists $c$ with $H(y) \geq b$ for all $y \geq c$, and then $D_b^{(\epsilon)} \subset [0,b] \times [0,c]$, so $\mathrm{Leb}(D_b^{(\epsilon)}) < +\infty$. The rest of the proof goes as in the main text. □

## References

[1] Fanny Augeri, Alice Guionnet, and Jonathan Husson, *Large deviations for the largest eigenvalue of sub-Gaussian matrices*, Comm. Math. Phys. **383** (2021), no. 2, 997–1050. MR4239836

[2] Zhidong Bai and Jack W. Silverstein, *Spectral analysis of large dimensional random matrices*, second ed., Springer Series in Statistics, Springer, New York, 2010. MR2567175

[3] Zhigang Bao, Guangming Pan, and Wang Zhou, *Universality for the largest eigenvalue of sample covariance matrices with general population*, Ann. Statist. **43** (2015), no. 1, 382–421. MR3311864

[4] Serban Belinschi, Alice Guionnet, and Jiaoyang Huang, *Large deviation principles via spherical integrals*, Probab. Math. Phys. **3** (2022), no. 3, 543–625. MR4520314

[5] George Bennett, *Upper bounds on the moments and probability inequalities for the sum of independent, bounded random variables*, Biometrika **52** (1965), 559–569. MR205358

[6] Giulio Biroli and Alice Guionnet, *Large deviations for the largest eigenvalues and eigenvectors of spiked Gaussian random matrices*, Electron. Commun. Probab. **25** (2020), Paper No. 70, 13. MR4158230

[7] Charles Bordenave and Pietro Caputo, *A large deviation principle for Wigner matrices without Gaussian tails*, Ann. Probab. **42** (2014), no. 6, 2454–2496. MR3265172

[8] Charles Bordenave, Pietro Caputo, and Djalil Chafaï, *Spectrum of non-Hermitian heavy tailed random matrices*, Comm. Math. Phys. **307** (2011), no. 2, 513–560. MR2837123

[9] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013. MR3185193

[10] Nicholas Cook, Raphaël Ducatez, and Alice Guionnet, *Full large deviation principles for the largest eigenvalue of sub-Gaussian Wigner matrices*, 2023, arXiv:2302.14823v1. MR4239836

[11] Amir Dembo and Ofer Zeitouni, *Large deviations techniques and applications*, Stochastic Modelling and Applied Probability, vol. 38, Springer-Verlag, Berlin, 2010, Corrected reprint of the second (1998) edition. MR2571413

[12] Noureddine El Karoui, *Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices*, Ann. Probab. **35** (2007), no. 2, 663–714. MR2308592

[13] László Erdős, Torben Krüger, and Dominik Schröder, *Random matrices with slow correlation decay*, Forum Math. Sigma **7** (2019), e8, 89. MR3941370

[14] Zhou Fan and Iain M. Johnstone, *Tracy-Widom at each edge of real covariance and MANOVA estimators*, Ann. Appl. Probab. **32** (2022), no. 4, 2967–3003. MR4474524

[15] A. Guionnet and M. Maïda, *A Fourier view on the $R$-transform and related asymptotics of spherical integrals*, J. Funct. Anal. **222** (2005), no. 2, 435–490. MR2132396

[16] Alice Guionnet and Jonathan Husson, *Large deviations for the largest eigenvalue of Rademacher matrices*, Ann. Probab. **48** (2020), no. 3, 1436–1465. MR4112720

[17] Alice Guionnet and Jonathan Husson, *Asymptotics of $k$ dimensional spherical integrals and applications*, ALEA Lat. Am. J. Probab. Math. Stat. **19** (2022), no. 1, 769–797. MR4436026

[18] Alice Guionnet and Mylène Maïda, *Large deviations for the largest eigenvalue of the sum of two random matrices*, Electron. J. Probab. **25** (2020), Paper No. 14, 24. MR4073675

[19] Alice Guionnet and Ofer Zeitouni, *Concentration of the spectral measure for large matrices*, Electron. Comm. Probab. **5** (2000), 119–136. MR1781846

[20] Alice Guionnet and Ofer Zeitouni, *Large deviations asymptotics for spherical integrals*, J. Funct. Anal. **188** (2002), no. 2, 461–515. MR1883414

[21] Walid Hachem, Adrien Hardy, and Jamal Najim, *Large complex correlated Wishart matrices: fluctuations and asymptotic independence at the edges*, Ann. Probab. **44** (2016), no. 3, 2264–2348. MR3502605

[22] Jonathan Husson, *Large deviations for the largest eigenvalue of matrices with variance profiles*, Electron. J. Probab. **27** (2022), Paper No. 74, 44. MR4440063

[23] Tosio Kato, *Variation of discrete spectra*, Comm. Math. Phys. **111** (1987), no. 3, 501–504. MR900507

[24] Eytan Katzav and Isaac Pérez Castillo, *Large deviations of the smallest eigenvalue of the Wishart-Laguerre ensemble*, Phys. Rev. E (3) **82** (2010), no. 4, 040104, 4. MR2788021

[25] Antti Knowles and Jun Yin, *Anisotropic local laws for random matrices*, Probab. Theory Related Fields **169** (2017), no. 1-2, 257–352. MR3704770

[26] Ji Oon Lee and Kevin Schnelli, *Tracy-Widom distribution for the largest eigenvalue of real sample covariance matrices with general population*, Ann. Appl. Probab. **26** (2016), no. 6, 3786–3839. MR3582818

[27] Antoine Maillard, *Large deviations of extreme eigenvalues of generalized sample covariance matrices*, EPL (Europhysics Letters) **133** (2021), no. 2, 20005.

[28] Satya N. Majumdar and Massimo Vergassola, *Large deviations of the maximum eigenvalue for Wishart and Gaussian random matrices*, Phys. Rev. Lett. **102** (2009), 060601.

[29] V. A. Marčenko and L. A. Pastur, *Distribution of eigenvalues for some sets of random matrices*, Math. USSR-Sb. **1** (1967), no. 4, 457–483.

[30] A. Matytsin, *On the large-$N$ limit of the Itzykson-Zuber integral*, Nuclear Phys. B **411** (1994), no. 2-3, 805–820. MR1257846

[31] Benjamin McKenna, *Large deviations for extreme eigenvalues of deformed Wigner random matrices*, Electron. J. Probab. **26** (2021), Paper No. 34, 37. MR4235485

[32] Pierre Mergny and Marc Potters, *Right large deviation principle for the top eigenvalue of the sum or product of invariant random matrices*, J. Stat. Mech. Theory Exp. **2022** (2022), no. 6, Paper No. 063301, 65. MR4483761

[33] Alexei Onatski, *The Tracy-Widom limit for the largest eigenvalues of singular complex Wishart matrices*, Ann. Appl. Probab. **18** (2008), no. 2, 470–490. MR2398763

[34] L. A. Pastur, *The spectrum of random matrices*, Teoret. Mat. Fiz. **10** (1972), no. 1, 102–112. MR475502

[35] Kevin Schnelli and Yuanyuan Xu, *Convergence rate to the Tracy–Widom laws for the largest eigenvalue of sample covariance matrices*, Ann. Appl. Probab. **33** (2023), no. 1, 677–725. MR4551561

[36] Jack W. Silverstein and Z. D. Bai, *On the empirical distribution of eigenvalues of a class of large dimensional random matrices*, J. Multivariate Anal. **54** (1995), 175–192. MR1345534

[37] Michel Talagrand, *A new look at independence*, Ann. Probab. **24** (1996), no. 1, 1–34. MR1387624

[38] Dan Voiculescu, *Limit laws for random matrices and free products*, Invent. Math. **104** (1991), no. 1, 201–220. MR1094052

[39] Haoyu Wang, *Quantitative universality for the largest eigenvalue of sample covariance matrices*, Ann. Appl. Probab. **34** (2024), no. 3, 2539–2565. MR4756952

[40] Nikita Zhivotovskiy, *Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle*, Electron. J. Probab. **29** (2024), Paper No. 13, 28. MR4693860

**Acknowledgments.** We wish to thank Alice Guionnet for proposing the problem and for many helpful discussions, Ofer Zeitouni for suggesting a nicer proof of Lemma 3.15 than that given in the first version of this paper, and the anonymous referee for helpful suggestions. We also thank Yizhe Zhu for directing us to the reference [40].