Probab. Theory Relat. Fields (2012) 154:677–702
DOI 10.1007/s00440-011-0381-4

A concentration inequality and a local law for the sum of two random matrices

Vladislav Kargin

Received: 22 November 2010 / Revised: 9 June 2011 / Published online: 7 July 2011

Abstract Let $H_N = A_N + U_N B_N U_N^*$ where A_N and B_N are two N-by-N Hermitian matrices and U_N is a Haar-distributed random unitary matrix, and let μ_{H_N}, μ_{A_N}, μ_{B_N} be empirical measures of eigenvalues of matrices H_N, A_N, and B_N, respectively. Then, it is known (see Pastur and Vasilchuk in Commun Math Phys 214:249–286, 2000) that for large N, the measure μ_{H_N} is close to the free convolution of measures μ_{A_N} and μ_{B_N}, where the free convolution is a non-linear operation on probability measures. The large deviations of the cumulative distribution function of μ_{H_N} from its expectation have been studied by Chatterjee (J Funct Anal 245:379–389, 2007). In this paper we improve Chatterjee's concentration inequality and show that it holds with the rate which is quadratic in N. In addition, we prove a local law for eigenvalues of H_{N_N}, by showing that the normalized number of eigenvalues in an interval approaches the density of the free convolution of μ_A and μ_B provided that the interval has width $(\log N)^{-1/2}$.

Mathematics Subject Classification (2010) 60B20 · 60B10 · 46L54

1 Introduction

If A and B are two Hermitian matrices with a known spectrum, it is a classical problem to determine all possibilities for the spectrum of $A+B$. The problem goes back at least to Weyl [21]. Later, Horn [13] suggested a list of inequalities which must be satisfied by eigenvalues of $A + B$, and recently, Knutson and Tao [15] using earlier ideas by Klyachko, proved that this list is complete.

For large matrices, it is natural to consider the probabilistic analogue of this problem, when matrices A and B are "in general position". Namely, let $H_N = A_N + U_N B_N U_N^*$, where A_N and B_N are two fixed N-by-N Hermitian matrices, and U_N is

V. Kargin (✉)
Department of Mathematics, Stanford University, Stanford, CA 94305, USA
e-mail: kargin@stanford.edu

a random unitary matrix with the Haar distribution on the unitary group $\mathcal{U}(N)$. Then, the eigenvalues of H_N are random and we are interested in their joint distribution.

Let $\lambda_1^{(A)} \geq \cdots \geq \lambda_N^{(A)}$ denote eigenvalues of A_N, and define the *spectral measure* of A_N as $\mu_{A_N} := N^{-1} \sum_{k=1}^{N} \delta_{\lambda_k^{(A)}}$. Define μ_{B_N} and μ_{H_N} similarly, and note that μ_{H_N} is random even if μ_{A_N} and μ_{B_N} are non-random. What can be said about relationship of μ_{A_N}, μ_{B_N}, and μ_{H_N}?

An especially interesting case occurs when N is large. This case was investigated by Voiculescu [20] and Speicher [18] who found that as N grows μ_{H_N} approaches $\mu_{A_N} \boxplus \mu_{B_N}$, where $\boxplus$ denotes *free convolution*, a non-linear operation on probability measures introduced by Voiculescu in his studies of operator algebras. Their proofs are based on calculating traces of large powers of matrices and use ingenious combinatorics. Later, Pastur and Vasilchuk [16] applied the method of Stieltjes transforms to this problem and extended the results of Speicher and Voiculescu to measures with unbounded support.

It appears natural to ask the question about deviations of μ_{H_N} from $\mu_{A_N} \boxplus \mu_{B_N}$.

In order to illuminate the issues that arise, suppose first that we place N points independently on a fixed interval $[a, b]$, each according to a measure ν. Let the number of points in a sub-interval I be denoted $\mathcal{N}_I$. Then, $\mathcal{N}_I$ is a sum of independent Bernoulli variables and satisfies the familiar central limit law and large deviation estimates. In particular,

$$\Pr\left\{\left|\frac{\mathcal{N}_I}{N|I|} - \mathbb{E}\left(\frac{\mathcal{N}_I}{N|I|}\right)\right| > \delta\right\} \sim c_1 \exp[-c_2\delta^2 N] \tag{1}$$

for large N.

A remarkable fact is that for random points corresponding to eigenvalues of classical random matrix ensembles, the asymptotic is different and given by the formula

$$\Pr\left\{\left|\frac{\mathcal{N}_I}{N|I|} - \mathbb{E}\left(\frac{\mathcal{N}_I}{N|I|}\right)\right| > \delta\right\} \sim c_1 \exp[-c_2 f(\delta) N^2]. \tag{2}$$

Intuitively, there is a repulsion force between eigenvalues which makes large deviations of $\mathcal{N}_I$ much more unlikely for large N.

For classical ensembles this fact was rigorously shown in a more general form in [4]. Later, this result was extended to matrices of the form $A_N + sX_N$, where A_N is an Hermitian N-by-N matrix and X is an Hermitian Gaussian N-by-N matrix; see for an explanation Sects. 4.3 and 4.4 in [1].

The fluctuations of eigenvalues of matrices $H_N = A_N + U_N B_N U_N^*$ were considered by Chatterjee [8]. By an ingenious application of the Stein method he proved that for every $x \in \mathbb{R}$,

$$\Pr\left\{\left|\mathcal{F}_{H_N}(x) - \mathbb{E}\mathcal{F}_{H_N}(x)\right| > \delta\right\} \leq 2\exp\left[-c\delta^2 \frac{N}{\log N}\right],$$

where $\mathcal{F}_{H_N}(x) := N^{-1}\mathcal{N}_{(-\infty,x]}$ denotes the cumulative distribution function for eigenvalues of H_N, symbol $\mathbb{E}$ denotes the expectation with respect to the Haar measure, and c is a numeric constant. Note that the rate in this estimate is sublinear in N, hence the estimate is weaker than (2). In fact, it is even weaker than the estimate in (1) because of the logarithmic factor $(\log N)^{-1}$, and therefore it does not contain any evidence of the repulsion between eigenvalues.

The first main result of this paper is an improvement of this estimate and is as follows.

Assumption $A1$. The measure $\mu_{A_N} \boxplus \mu_{B_N}$ is absolutely continuous everywhere on $\mathbb{R}$, and its density is bounded by a constant T_N.

Theorem 1 *Suppose that Assumption* $A1$ *holds. Let* $\mathcal{F}_{H_N}$ *and* $\mathcal{F}_{\boxplus,N}$ *be cumulative distribution functions for the eigenvalues of* $H_N = A_N + U_N B_N U_N^*$ *and for* $\mu_{A_N} \boxplus \mu_{B_N}$*, respectively. Then, for all* $N \geq \exp((c_1/\delta)^{4/\varepsilon})$,

$$P\left\{\sup_x |\mathcal{F}_{H_N}(x) - \mathcal{F}_{\boxplus,N}(x)| > \delta\right\} \leq \exp[-c_2\delta^2 N^2 (\log N)^{-\varepsilon}], \tag{3}$$

where c_1, c_2 *are positive and depend only on* $K_N := \max\{\|A_N\|, \|B_N\|\}, T_N$*, and* $\varepsilon \in (0, 2]$.

Up to a logarithmic factor, the rate in this inequality is proportional to N^2, which is consistent with the possibility that the eigenvalues of matrix $H_N = A_N + U_N B_N U_N^*$ repulse each other.

With respect to Assumption $A1$, it is pertinent to note that if $\mu_{A_N}(\{x\}) < 1/2$ and $\mu_{B_N}(\{x\}) < 1/2$ for every $x \in \mathbb{R}$ (i.e., if the multiplicity of every eigenvalue of A_N and B_N is less than $N/2$), then $\mu_{A_N} \boxplus \mu_{B_N}$ has no atoms (see Theorem 7.4 in [5]). Moreover, since μ_{A_N} and μ_{B_N} are atomic, the results of [3] imply that the density of $\mu_{A_N} \boxplus \mu_{B_N}$ is analytic (i.e., in C^∞ class) everywhere on $\mathbb{R}$ where it is positive. In particular, Assumption $A1$ holds.

If Assumption $A1$ is relaxed, then it is still possible to prove a result similar to the result in Theorem 1. Namely, if $\mu_{\boxplus,N}$ is absolutely-continuous at the endpoints of interval I, then it is possible to show that for all sufficiently large N,

$$P\left\{\left|\frac{\mathcal{N}_I}{N|I|} - \mu_{\boxplus,N}(I)\right| > \delta\right\} \leq \exp[-c_2\delta^2 N^2 (\log N)^{-\varepsilon}]. \tag{4}$$

Indeed, the only place where Assumption $A1$ is used is when the distance between $\mathcal{F}_{H_N}$ and $\mathcal{F}_{\boxplus,N}$ is estimated in terms of the distance between $m_H(z)$ and $m_{\boxplus,N}(z)$ and this is done by using Bai's theorem. In order to prove (4), the original proof should be modified by using techniques from the proof of Corollary 4.2 in [10] instead of Bai's theorem. In this paper, however, we choose to concentrate on the proof of inequality (3).

In addition, if Assumption $A1$ fails and x is an atom of $\mu_{A_N} \boxplus \mu_{B_N}$ then by Thm 7.4 in [5] there exist x_A and x_B such that $x_A + x_B = x$, and

$$\mu_{A_N}(\{x_A\}) + \mu_{B_N}(\{x_B\}) - 1 = \mu_{A_N} \boxplus \mu_{B_N}(\{x\}).$$

These x_A and x_B are eigenvalues of A_N and $U_N B_N U_N^*$ with multiplicities $\mu_{A_N}(\{x_A\})N$ and $\mu_{B_N}(\{x_B\})N$, respectively. Hence, by counting dimensions and using the fact that eigenspaces of A_N and $U_N B_N U_N^*$ are in general position, we conclude that with probability 1, $x_A + x_B$ is an eigenvalue of H_N with multiplicity

$$(\mu_{A_N}(\{x_A\}) + \mu_{B_N}(\{x_B\}) - 1)N.$$

Hence, if x is an atom of $\mu_{A_N} \boxplus \mu_{B_N}$, then we have the exact equality

$$\mu_{H_N}(\{x\}) = \mu_{A_N} \boxplus \mu_{B_N}(\{x\}).$$

These considerations suggest that perhaps Assumption $A1$ can be eliminated or weakened as a condition of Theorem 1.

Our main tools in the proof of Theorem 1 are the Stieltjes transform method and standard concentration inequalities applied to functions on the unitary group.

In the first step, we establish the N^2 rate for large deviations of the Stieltjes transform of μ_{H_N}, which we denote $m_{H_N}(z)$. This follows from results in [1] and the fact that the Stieltjes transform of μ_{H_N} is Lipschitz as a function of U_N and its Lipschitz constant can be explicitly estimated.

It is not possible to prove a concentration inequality for $\mathcal{F}_{H_N}(x)$ by a similar method because for some x this function is not Lipschitz in U_N. An alternative is to use an inequality by Bai (Theorem 23 in this paper), which gives a bound on $\sup_x |\mathcal{F}_{H_N}(x) - \mathbb{E}\mathcal{F}_{H_N}(x)|$ in terms of $\sup_x |m_{H_N}(z) - \mathbb{E}m_{H_N}(z)|$, where $z = x + i\eta$. However, the second term in this inequality depends on smoothness of $\mathbb{E}\mathcal{F}_{H_N}(x)$, which is difficult to establish.

Instead, we show that $\sup_x |\mathbb{E}m_{H_N}(z) - m_{\boxplus,N}(z)|$ is small for $\eta := \mathrm{Im} z > c/\sqrt{\log N}$. (Here $m_{\boxplus,N}(z)$ denote the Stieltjes transform of $\mu_{A_N} \boxplus \mu_{B_N}$.) This estimate allows us to use Bai's inequality and estimate $\sup_x |\mathcal{F}_{H_N}(x) - \mathbb{E}\mathcal{F}_{\boxplus,N}(x)|$ in terms of the sum of $\sup_x |m_{H_N}(z) - \mathbb{E}m_{H_N}(z)|$ and $\sup_x |\mathbb{E}m_{H_N}(z) - m_{\boxplus,N}(z)|$, which are both small. The benefit of this change is that smoothness of $\mathcal{F}_{\boxplus,N}(x)$ is easier to establish than the smoothness of $\mathbb{E}\mathcal{F}_{H_N}(x)$. In our case it is guaranteed by Assumption $A1$.

For large $\mathrm{Im} z$, the difference $|\mathbb{E}m_{H_N}(z) - m_{\boxplus,N}(z)|$ can be estimated by applying Newton's iteration method (as perfected by Kantorovich [14]) to the Pastur–Vasilchuk system for $\mathbb{E}m_{H_N}(z)$. Namely, we use $m_{\boxplus,N}(z)$ as the starting point for this method and show that for sufficiently large N the difference of the solution of the system, $\mathbb{E}m_{H_N}(z)$, and the starting point is less than any fixed $\delta > 0$.

This method fails for small $\mathrm{Im} z$. We use a modification of Hadamard's three circle theorem [12] in order to estimate the difference $|\mathbb{E}m_{H_N}(z) - m_{\boxplus,N}(z)|$ in the region close to the real axis.

Theorem 1 implies the following local law result. Let $N_\eta(E)$ denote the number of eigenvalues of H_N in an interval of width 2η centered at E, and let $\varrho_{\boxplus,N}(E)$ denote the density of $\mu_{A_N} \boxplus \mu_{B_N}$ at E.

Theorem 2 *Suppose that* $\eta = \eta(N)$ *and* $1/\sqrt{\log N} \ll \eta \ll 1$. *Let assumption* $A1$ *hold with* $T_N = T$. *Assume also that* $\max\{\|A_N\|, \|B_N\|\} \le K$ *for all* N. *Then, for all* *sufficiently large* N,

$$P\left\{\sup_E\left|\frac{\mathcal{N}_\eta(E)}{2N\eta}-\varrho_{\boxplus,N}(E)\right|\geq\delta\right\}\leq\exp\left(-c\delta^2\frac{(\eta N)^2}{(\log N)^2}\right),$$

where $c>0$ *depends only on* K *and* T.

(Here the notation $(N)\ll g(N)$ means that $\lim_{N\to\infty}g(N)/f(N)=+\infty$.)

The plan of the rest of the paper is as follows. We start in Sect. 2 by establishing our notation. Section 3 provides a large deviation estimate for the Stieltjes transform of μ_{H_N} and a related function. In Sect. 4, we use this estimate to bound error terms in the Pastur–Vasilchuk system, which we re-derive for reader's convenience. Section 5 is devoted to estimating $|\mathbb{E}m_{H_N}(z)-m_{\boxplus,N}(z)|$ in the region where $\mathrm{Im}z\geq\eta_0$, and Sect. 6 is concerned with estimating it in the region $\mathrm{Im}z\gg 1/\sqrt{\log N}$. Section 7 completes the proof of our two main theorems. Several concluding remarks are made in Sect. 8.

2 Definitions and notations

We define $H_N=A_N+U_NB_NU_N^*$. The *spectral measure* of H_N is $\mu_{H_N}:=N^{-1}\sum_{k=1}^N\delta_{\lambda_k^{(H)}}$, where $\lambda_k^{(H)}$ are eigenvalues of H, counted with multiplicity. Its *cumulative distribution function* is denoted $\mathcal{F}_{H_N}(x):=\mu_{H_N}((-\infty,x])$. The number of eigenvalues of H_N in interval I is denoted $\mathcal{N}_I:=N\mu_{H_N}(I)$, and $\mathcal{N}_\eta(E):=\mathcal{N}_{(E-\eta,E+\eta]}$ denotes the number of eigenvalues in the interval of width 2η centered at E.

The *resolvent* of H_N is defined as $G_H(z):=(H_N-z)^{-1}$. Similarly, $G_A(z):=(A_N-z)^{-1}$ and $G_B(z):=(B_N-z)^{-1}$. (For brevity, we will omit the subscript N in the notation for resolvents and Stieltjes transforms.)

The *Stieltjes transform* of H_N is defined as

$$m_H(z):=N^{-1}\mathrm{Tr}\,G_H(z)=\int_{\mathbb{R}}\frac{\mu_{H_N}(d\lambda)}{\lambda-z},$$

where Tr denotes the usual matrix trace. The Stieltjes transforms of A_N and B_N are defined similarly, e.g., $m_A(z)=N^{-1}\mathrm{Tr}\,G_A(z)$. More generally, if μ is a probability measure, then its Stieltjes transform is defined as

$$m_\mu(z):=\int_{\mathbb{R}}\frac{\mu(d\lambda)}{\lambda-z}.$$

In addition, we define the following quantities:

$$f_B(z):=N^{-1}\mathrm{Tr}\left(U_NB_NU_N^*\frac{1}{H_N-z}\right)$$

and

$$f_A(z) := N^{-1}\mathrm{Tr}\left(A_N \frac{1}{H_N - z}\right)$$

Next, we define the free convolution. Consider the following system:

$$\begin{aligned} m(z) &= m_A(z - S_B(z)),\\ m(z) &= m_B(z - S_A(z)),\\ z + \frac{1}{m(z)} &= S_A(z) + S_B(z), \end{aligned} \tag{5}$$

where $m(z)$, $S_A(z)$, $S_B(z)$ are unknown functions.

Proposition 3 *There exists a unique triple of analytic functions $m(z)$, $S_A(z)$, $S_B(z)$ that are defined in $\mathbb{C}^+ = \{z : \mathrm{Im} z > 0\}$, satisfy system (5), and have the following asymptotics as $z \to \infty$:*

$$\begin{aligned} m(z) &= -z^{-1} + O(z^{-2}),\\ S_{A,B}(z) &= O(1). \end{aligned} \tag{6}$$

Moreover, the function $m(z)$ maps $\mathbb{C}^+$ to $\mathbb{C}^+$ and the functions $S_{A,B}(z)$ map $\mathbb{C}^+$ to $\mathbb{C}^- = \{z : \mathrm{Im} z < 0\}$.

Proposition 3 implies that the first function in this triple, $m_{\boxplus,N}(z)$, is the Stieltjes transform of a probability measure. This measure is called the *free convolution* of measures μ_{A_N} and μ_{B_N} and denoted $\mu_{A_N} \boxplus \mu_{B_N}$. (For shortness, we will sometimes write this measure as $\mu_{\boxplus,N}$.) The two other functions in this triple, $S_A(z)$ and $S_B(z)$, are called *subordination functions*.

Proof of Proposition 3 The uniqueness of the solution of system (5) was proved in Prop. 3.3 in [16]. However, it appears that their proof does not show that the solution exists everywhere in the upper half-plane. We prove the existence and uniqueness differently, by establishing a one-to-one correspondence between solutions of (5) and certain objects in free probability theory. After this correspondence is established, the existence, uniqueness and claimed properties of the solution follow from the corresponding properties of the free probability objects.

Recall that in the traditional definition of free convolution [19], one defines the R-transform of measure μ_A by the formula $R_A(t) = m_A^{(-1)}(-t) - 1/t$, where $m_A^{(-1)}$ is the functional inverse of m_A, chosen in such a fashion that $R_A(t)$ is analytic at $t = 0$. The function $R_B(t)$ is defined similarly. Then, one proves that $R = R_A + R_B$ is the R-transform of a probability measure, and one calls this measure the free convolution of μ_A and μ_B. In fact, this definition of free convolution is equivalent to the definition we have given above.

Indeed, let $m_{\boxplus,N}$ be the Stieltjes transform of $\mu_{A_N} \boxplus \mu_{B_N}$ as it is usually defined, that is, let it equal the functional inverse of $R + 1/t$ multiplied by -1. By definition of R_A, the first equation of (5) can be written equivalently as

$$S_B(z) = z + \frac{1}{m_{\boxplus,N}(z)} - R_A(-m_{\boxplus,N}(z)),$$

which we can use as a definition of $S_B(z)$. This definition holds only for sufficiently large z. However, by the results of Biane ([6]), $S_B(z)$ can be analytically continued to the whole of $\mathbb{C}^+$. If we write the second equation in a similar form, add them together, and use the equality $R = R_A + R_B$, then we get:

$$\begin{aligned} S_A(z) + S_B(z) &= 2z + \frac{2}{m_{\boxplus,N}(z)} - R\left(-m_{\boxplus,N}(z)\right) \\ &= z + \frac{1}{m_{\boxplus,N}(z)}, \end{aligned}$$

which is the third equation of system (5). By analytic continuation it holds everywhere in $\mathbb{C}^+$. This shows that if $m_{\boxplus,N}(z)$, $S_A(z)$, and $S_B(z)$ are defined using the traditional definition of free convolution, then they satisfy system (5). In particular this shows the existence of the solution of (5) as a triple of analytic functions defined everywhere in $\mathbb{C}^+$.

Conversely, if $m_{\boxplus,N}(z)$, $S_A(z)$, and $S_B(z)$ satisfy (5) with asymptotic conditions (6), then in a neighborhood of infinity we can write

$$\begin{aligned} R_A(-m_{\boxplus,N}(z)) &= m_A^{(-1)}(m_{\boxplus,N}(z)) + 1/m_{\boxplus,N}(z) \\ &= z - S_B(z) + 1/m_{\boxplus,N}(z), \end{aligned}$$

where the first line is the definition of R_A and the second uses the first equation of (5). If we write a similar expression for $R_B(-m_{\boxplus,N}(z))$, add them together, and use the third equation of (5), then we find that

$$R(-m_{\boxplus,N}(z)) = z + 1/m_{\boxplus,N}(z).$$

This shows that $m_{\boxplus,N}(z)$ satisfies the same functional equation as the Stieltjes transform of the free convolution measure defined in the traditional fashion. Since their power expansions at infinity are the same, these functions coincide. In particular, this shows that the solution of (5) is unique as a triple of analytic functions in $\mathbb{C}^+$ that satisfy asymptotic conditions (6).

Finally, the claimed properties of $m_{\boxplus,N}(z)$ and $S_{A,B}(z)$ follow from the properties of the Stieltjes transform of a probability measure and of the subordination functions. The latter were established by Biane in [6]. □

We denote the cumulative distribution function of $\mu_{A_N} \boxplus \mu_{B_N}$ as $\mathcal{F}_{\boxplus,N}$ and its density (when it exists) as $\varrho_{\boxplus,N}$.

The integration over U using the Haar measure will be denoted as $\mathbb{E}$. (This operation is often denoted as $\langle\cdot\rangle$ in the literature.) Correspondingly, $P(\omega)$ denotes the Haar measure of event ω.

We will usually write $z = E + i\eta$, where E and η denote the real and imaginary parts of z. We will also use the following notation:

$$\Omega_{\eta_0,c} = \{z \in \mathbb{C} : \mathrm{Im} z \geq \eta_0, \mathrm{Im} z \geq c\mathrm{Re} z\}.$$

3 Concentration for the Stieltjes transform and associated functions

The main result of this section is the following large deviation estimates for $m_H(z)$ and $f_B(z)$.

Proposition 4 *Let $z = E + i\eta$ where $\eta > 0$. Then, for a numeric $c > 0$ and every $\delta > 0$,*

$$P\{|m_H(z) - \mathbb{E}m_H(z)| > \delta\} \leq \exp\left(-\frac{c\delta^2\eta^4}{\|B\|^2}N^2\right), \tag{7}$$

and

$$P\{|f_B(z) - \mathbb{E}f_B(z)| > \delta\} \leq \exp\left[-\frac{c\delta^2\eta^4}{\|B\|^4}N^2 / \left(1 + \frac{\eta}{\|B\|}\right)^2\right]. \tag{8}$$

Proof The first claim of this proposition follows directly from Corollary 4.4.30 in [1]. The second claim can be obtained by a modification of the proof of this Corollary. For the convenience of the reader we give a short proof of both claims.

Both claims are consequences of the Gromov–Milman results about the concentration of Lipschitz functions on Riemannian manifolds [11]. In a small neighborhood of identity matrix, all unitary matrices can be written as $U = e^{iX}$, where X is Hermitian. We identify the space of Hermitian matrices X with $T\mathcal{U}$, the tangent space to $\mathcal{U}(N)$ at point I. By left translations this identification can be extended to the tangent space at any point of $\mathcal{U}(N)$. Define an inner product norm in $T\mathcal{U}$ by the formula $\|X\|_2 = (\sum_{ij} |X_{ij}|^2)^{1/2}$. This gives us a Riemannian metric ds on $\mathcal{U}(N)$. The Riemannian metric on $\mathcal{SU}(N)$ can be defined by restriction.

The (real or complex-valued) function $f(x)$ on a metric space M is called Lipschitz with constant L if for every two points $x, y \in M$, it is true that $|f(x) - f(y)| \leq Ld(x, y)$, where $d(x, y)$ is the shortest distance between x and y.

Proposition 5 *Let $g : (\mathcal{SU}(N), ds) \to \mathbb{R}$ be an L-Lipschitz function and let $\mathbb{E}g = 0$. Then*

(i) $\mathbb{E}\exp(tg) \leq \exp(ct^2L^2/N)$ *for every $t \in \mathbb{R}$ and some numeric $c > 0$, and*

(ii) $P\{|g| > \delta\} \leq \exp(-c_1N\delta^2/L^2)$ *for every $\delta > 0$ and some numeric $c_1 > 0$.*

For the proof, see Theorems 3.8.3 and 3.9.2 in [7] and Theorem 4.4.27 in [1].

In order to apply this result, we need to estimate the Lipschitz constants for $m_H(z)$ and $f_B(z)$. If M is a Riemannian manifold and f is a differentiable function on M, then it is Lipschitz with constant L provided that $|d_X f(x)| \leq L$ for every $x \in M$

and every unit vector $X \in TM_x$. Here d_X denotes the derivative in the direction of vector X. We will apply this general observation to the manifold $\mathcal{SU}(N)$.

Let $\widetilde{B}$ denote UBU^*, $B(x) = e^{iX}\widetilde{B}e^{-iX}$ and let

$$m_H(z, X) = (A + B(x) - z)^{-1}$$

We differentiate $m_H(z, X)$ with respect to X (and evaluate it at $X = 0$) by using the chain rule.

$$\begin{aligned} |d_X m_H(z, X)| &= \left| \sum_{x,y} \frac{\partial m_H(z)}{\partial \widetilde{B}_{xy}} d_X B_{xy}(x) \right| \\ &= \left| \frac{1}{N} \sum_{x,y} (G^2)_{yx} [X, \widetilde{B}]_{xy} \right| \\ &= \left| \frac{1}{N} \sum_{x,y} ([G^2, \widetilde{B}])_{yx} X_{xy} \right|. \end{aligned}$$

where we used the facts that $\partial m_H/\partial(\widetilde{B}_{xy}) = -N^{-1}(G^2)_{yx}$ and that $d_X B(x)|_{X=0} = [X, \widetilde{B}]$. These facts can be easily checked by a calculation. For the first one, see Lemma 9 below.

If $\|X\|_2 = 1$, then it follows that

$$\begin{aligned} |d_X m(z, X)| &\leq \frac{1}{N} \|[G^2, \widetilde{B}]\|_2 \\ &\leq \frac{1}{\sqrt{N}} \|[G^2, \widetilde{B}]\| \\ &\leq \frac{2\|B\|}{\sqrt{N}\eta^2} \end{aligned}$$

Together with Proposition 5, this implies the first claim of the lemma. For the second claim, let $f_B(z, X) = B(x)(A + B(x) - z)^{-1}$. Note that $\widetilde{B}(A + \widetilde{B} - z)^{-1} = I - (A - z)(A + \widetilde{B} - z)^{-1}$. This allows us to calculate:

$$\frac{\partial}{\partial \widetilde{B}_{xy}} (f_B(z, X)) = \frac{1}{N} (G(A - z)G)_{yx}.$$

Hence,

$$\begin{aligned} |d_X f(z, X)| &= \left| \sum_{x,y} \frac{\partial f_B(z)}{\partial \widetilde{B}_{xy}} d_X B_{xy}(x) \right| \\ &= \left| \frac{1}{N} \sum_{x,y} ([G(A - z)G, \widetilde{B}])_{yx} X_{xy} \right| \end{aligned}$$

$$\leq \frac{1}{N}\|[G(A-z)G, \widetilde{B}]\|_2$$
$$\leq \frac{1}{\sqrt{N}}\|[G(A-z)G, \widetilde{B}]\|.$$

Since $(A-z)G = I - \widetilde{B}G$, we can continue this as

$$|d_X f(z, X)| \leq \frac{2}{\sqrt{N}}\left(\frac{\|B\|}{\eta} + \frac{\|B\|^2}{\eta^2}\right),$$

and the rest follows from Proposition 5. □

Later, we will need the following consequence of Proposition 4.

Corollary 6 *Let $I_\eta = [-2K + i\eta, 2K + i\eta]$. Then for some positive c and c_1 which may depend on K and for all $\delta > 0$,*

$$P\{\sup_{z\in I_\eta} |m_H(z) - \mathbb{E}m_H(z)| > \delta\} \leq \exp\left(-\frac{c\delta^2\eta^4}{\|B\|^2}N^2\right),$$

provided that $N \geq c_1(\sqrt{-\log(\eta^2\delta)})/(\eta^2\delta)$.

Proof of Corollary Note that $|m_H'(z)| \leq \eta^{-2}$ and $|\mathbb{E}m_H'(z)| \leq \eta^{-2}$ and that it is enough to place $O(K/\eta^2\delta)$ points on interval I_η to create an ε-net with $\varepsilon = \eta^2\delta/4$. If $|m_H(z) - \mathbb{E}m_H(z)| \leq \delta/2$ at every point of the net, then $|m_H(z) - \mathbb{E}m_H(z)| \leq \delta$ for all $z \in I_\eta$. Hence, by Theorem 4,

$$\begin{aligned} P\left\{\sup_{z\in I_\eta} |m_H(z) - \mathbb{E}m_H(z)| > \delta\right\} &\leq \frac{c'K}{\eta^2\delta}\exp\left(-\frac{c\delta^2\eta^4}{\|B\|^2}N^2\right) \\ &= \exp\left(-\frac{c\delta^2\eta^4}{\|B\|^2}N^2 + \log\left(\frac{c'K}{\eta^2\delta}\right)\right) \\ &\leq \exp\left(-\frac{c''\delta^2\eta^4}{\|B\|^2}N^2\right), \end{aligned}$$

if $N \geq c_1(\sqrt{-\log(\eta^2\delta)})/(\eta^2\delta)$ and c_1 is sufficiently large. □

4 An estimate on error terms in the Pastur–Vasilchuk system

For the convenience of the reader, we re-derive here the Pastur–Vasilchuk system. This is a system of equations for $\mathbb{E}m_H(z)$, $\mathbb{E}f_A(z)$, and $\mathbb{E}f_B(z)$. When N is large, this system is a perturbation of system (5), and the main purpose of this section is to estimate quantitatively the size of this perturbation. Later, we will show that system (5) is stable with respect to small perturbations, and therefore for large N the function $\mathbb{E}m_H(z)$ is close to the Stieltjes transform of $\mu_{A_N} \boxplus \mu_{B_N}$.

We use notations

$$\Delta_A := (m_H - \mathbb{E}m_H)G_H - G_A(f_B - \mathbb{E}f_B)G_H$$

and

$$R_A := \frac{1}{\mathbb{E}m_H}\frac{1}{N}\mathrm{Tr}\left(\frac{1}{1+(\mathbb{E}f_B/\mathbb{E}m_H)G_A}\mathbb{E}\Delta_A\right), \tag{9}$$

with similar definitions for Δ_B and R_B.

Theorem 7 (Pastur–Vasilchuk) *The functions* $\mathbb{E}m_H(z)$, $\mathbb{E}f_A(z)$ *and* $\mathbb{E}f_B(z)$ *satisfy the following system of equations*:

$$\begin{aligned} \mathbb{E}m_H(z) &= m_A\left(z - \frac{\mathbb{E}f_B(z)}{\mathbb{E}m_H(z)}\right) + R_A(z), \\ \mathbb{E}m_H(z) &= m_B\left(z - \frac{\mathbb{E}f_A(z)}{\mathbb{E}m_H(z)}\right) + R_B(z), \\ z + \frac{1}{\mathbb{E}m_H(z)} &= \frac{\mathbb{E}f_A(z) + \mathbb{E}f_B(z)}{\mathbb{E}m_H(z)}, \end{aligned} \tag{10}$$

where R_A *and* R_B *are defined as in* (9).

The main technical tool in the proof of this theorem is the following formula due to Pastur and Vasilchuk. Recall that G_H is the resolvent of $H_N = A_N + U_N B_N U_N^*$ where U_N is the Haar distributed random unitary matrix.

Proposition 8 $\mathbb{E}(m_H G_H) = \mathbb{E}(m_H G_A - G_A f_B G_H)$.

This result immediately implies Theorem 7. Indeed, the identity in Proposition 8 can be written in the following equivalent form.

$$\begin{aligned} (\mathbb{E}m_H)\mathbb{E}G_H &= (\mathbb{E}m_H)G_A - (\mathbb{E}f_B)G_A\mathbb{E}G_H \\ &\quad + \mathbb{E}[(m_H - \mathbb{E}m_H)G_H] - G_A\mathbb{E}[(f_B - \mathbb{E}f_B)G_H] \\ &= (\mathbb{E}m_H)G_A - (\mathbb{E}f_B)G_A\mathbb{E}G_H + \mathbb{E}\Delta_A. \end{aligned}$$

This expression can be further re-written (after we multiply it by $A_N - z$ and re-arrange terms) as

$$\mathbb{E}m_H\left(A_N - \left(z - \frac{\mathbb{E}f_B}{\mathbb{E}m_H}\right)\right)\mathbb{E}G_H = \mathbb{E}m_H + (A_N - z)\mathbb{E}\Delta_A.$$

Let $z' := z - Ef_B/Em$. Then for almost all values of z,

$$\mathbb{E}m_H\mathbb{E}G_H = G_A(z')\mathbb{E}m_H + (A_N - z)G_A(z')\mathbb{E}\Delta_A.$$

Take the normalized trace and divide the resulting expression by $\mathbb{E}m_H$. Then, we obtain

$$\begin{aligned}\mathbb{E}m_H(z) &= m_A(z') + \frac{1}{\mathbb{E}m_H}\frac{1}{N}\mathrm{Tr}\left(\frac{1}{1+(\mathbb{E}f_B/\mathbb{E}m)G_A}\mathbb{E}\Delta_A\right).\\ &= m_A(z') + R_A.\end{aligned}$$

The second equation of the system is obtained similarly and the third equation is an identity.

Proof of Proposition 8 It is useful to use notation $\widetilde{B} = U_N B_N U_N^*$ and $B(x) = e^{iX}\widetilde{B}e^{-iX}$. Note that by using the resolvent identity $G_H(z) - G_A(z) = -G_A(z)\widetilde{B}\, G_H(z)$, we know that

$$\begin{aligned}\mathbb{E}(m_H G_H) &= \mathbb{E}(m_H G_A - m_H G_A \widetilde{B} G_H)\\ &= G_A \mathbb{E}(m_H - m_H \widetilde{B} G_H).\end{aligned}$$

Hence, it is enough to show that $\mathbb{E}(m_H \widetilde{B} G_H) = \mathbb{E}(f_B G_H)$,

Lemma 9 *Let A and B be two arbitrary matrices and $G(z) = (A + B - z)^{-1}$. Then, $(\partial G/\partial B_{xy})_{uv} = -G_{ux}G_{yv}$. In particular,*

$$\left(\sum_{x,y}(\partial G/\partial B_{xy})M_{xy}\right)_{uv} = -\sum_{x,y} G_{ux}M_{xy}G_{yv}.$$

Proof This is an immediate consequence of the resolvent identity $G_{X+Y}(z) - G_X(z) = -G_X(z)YG_{X+Y}(z)$ applied to $X = A + B$ and $Y = tE^{xy}$, where E^{xy} denote the matrix that have 1 in the intersection of row x and column y and zeroes elsewhere. □

Lemma 10 *For every u, v, a, b, it is true that*

$$\mathbb{E}((G_H)_{ua}(\widetilde{B}G_H)_{bv}) = \mathbb{E}((G_H\widetilde{B})_{ua}(G_H)_{bv}).$$

Proof Note that $d(\mathbb{E}[(A + B(x) - z)^{-1}])/dt = 0$ for every Hermitian matrix X, because the distribution of $B(x) = e^{-itX}\widetilde{B}e^{itX}$ is the same as the distribution of $\widetilde{B}$. We can compute

$$\begin{aligned}\frac{d}{dt}[(A + B(x) - z)^{-1}]\Big|_{t=0} &= \sum_{x,y}\frac{\partial G_H}{\partial \widetilde{B}_{xy}}\frac{dB(x)_{xy}}{dt}\Big|_{t=0}\\ &= i\sum_{x,y}\frac{\partial G_H}{\partial \widetilde{B}_{xy}}\sum_s[-X_{xs}\widetilde{B}_{sy} + \widetilde{B}_{xs}X_{sy}].\end{aligned}$$

Let E^{ab} denote an N-by-N matrix that has zeros everywhere except at the intersection of the a-th row and b-th column, where it has entry 1. If we set $X = E^{ab} + E^{ba}$

and use Lemma 9, then we obtain

$$-\mathbb{E}[(G_H)_{ua}(\widetilde{B}G_H)_{bv} + (G_H)_{ub}(\widetilde{B}G_H)_{av}]$$
$$+\mathbb{E}[(G_H\widetilde{B})_{ua}(G_H)_{bv} + (G_H\widetilde{B})_{ub}(G_H)_{av}] = 0.$$

If we set $X = i\left(E^{ab} - E^{ba}\right)$, then we obtain a similar expression and adding them together, we get:

$$\mathbb{E}[-(G_H)_{ua}(\widetilde{B}G_H)_{bv} + (G_H\widetilde{B})_{ua}(G_H)_{bv}] = 0.$$

□

If we take $u = a$ in the statement of Lemmas 10 , then we get

$$\mathbb{E}((G_H)_{aa}(\widetilde{B}G_H)_{bv}) = \mathbb{E}((G_H\widetilde{B})_{aa}(G_H)_{bv}).$$

By adding up these equalities over a and dividing by N, we obtain that $\mathbb{E}(m_H\widetilde{B}G_H) = \mathbb{E}(f_B G_H)$, and Proposition 8 is proved. □

Now, we are going to estimate the error terms R_A and R_B. Let $\eta_0 \geq 0$ and $\kappa > 0$. For our purposes it is sufficient to make the estimates in the region

$$\Omega_{\eta_0,\kappa} := \{z \in \mathbb{C} : \mathrm{Im} z \geq \eta_0, \mathrm{Im} z \geq \kappa \mathrm{Re} z\}.$$

Proposition 11 *Assume that* $\max\{\|A\|, \|B\|\} \leq K$ *and let* $\kappa > 0$. *There exists an* $\eta_0 = cK$ *such that for every* $z = E + i\eta \in \Omega_{\eta_0,\kappa}$, *it is true that*

$$|R_A| \leq \frac{C}{N\eta^2},$$

where $C > 0$ *and depends only on* K *and* κ.

In order to prove this result, we will proceed in two steps. First, we will estimate $\|\mathbb{E}\Delta_A\|$. Then we estimate the multipliers before $\mathbb{E}\Delta_A$ in the definition of R_A.

Proposition 12 *Let* $z = E + i\eta$. *Assume that* $\eta \geq \eta_0$ *and that* $\max\{\|A\|, \|B\|\} \leq K$. *Then*

$$P\{\|\Delta_A(z)\| \geq \varepsilon\} \leq \exp\left[-c\varepsilon^2\eta^6 N^2\right],$$

and $\|\mathbb{E}\Delta_A(z)\| \leq c/\left(N\eta^3\right)$ *where constants depend only on* K *and* η_0.

Proof of Proposition 12

Lemma 13 *Let* $z = E + i\eta$, *where* $\eta > 0$. *Then for a numeric* $c > 0$,
a)

$$P\{\|(m_H(z) - \mathbb{E}m_H(z))\, G_H\| \geq \varepsilon\} \leq \exp\left[-c\frac{\varepsilon^2\eta^6}{\|B\|^2}N^2\right],$$

and b)

$$P\left\{\|G_A\left(f_B(z)-\mathbb{E}f_B(z)\right)G_H\|\geq\varepsilon\right\}\leq\exp\left[-c\frac{\varepsilon^2\eta^8}{\|B\|^4}N^2/\left(1+\frac{\eta}{\|B\|}\right)^2\right].$$

Proof Note that if X is a Hermitian matrix and $\eta>0$, then $\|(X-i\eta)^{-1}\|\leq 1/\eta$. By using this fact and Proposition 4, we get

$$P\{\|(m_H(z)-\mathbb{E}m_H(z))G_H\|\geq\delta/\eta\}\leq\exp\left[-c\frac{\delta^2\eta^4}{\|B\|^2}N^2\right].$$

Claim (a) of the lemma follows if we set $\varepsilon=\delta/\eta$. Claim (b) follows from Proposition 4 in a similar fashion. □

The first claim of Proposition 12 directly follows from Lemma 13.

For the second claim, note that $\|\mathbb{E}\Delta_A\|\leq\mathbb{E}\|\Delta_A\|$ by the convexity of norm, and $\mathbb{E}\|\Delta_A\|$ can be estimated by using the first claim of Proposition 12 and the equality

$$\mathbb{E}X=\int_0^\infty(1-\mathcal{F}_X(t))dt,$$

valid for every positive random variable X and its cumulative distribution function $\mathcal{F}_X(t)$. In our case, we obtain

$$\mathbb{E}\|\Delta_A\|\leq\int_0^\infty\exp[-ct^2\eta^6N^2]dt=\frac{c'}{N\eta^3}.$$

□

Proposition 14 *Let $z=E+i\eta$ where $\eta>0$. Assume that $\{\|A\|,\|B\|\}\leq K$. Then, there exists such an $\eta_0=cK$ with numeric $c>0$, that for every $\eta\geq\eta_0$, $\|(1+(\mathbb{E}f_B/\mathbb{E}m_H)G_A(z))^{-1}\|\leq 2$.*

The proof uses the following result.

Lemma 15 *Assume that $\{\|A\|,\|B\|\}\leq K$. Then, for some numeric $c>0$, the functions $\mathbb{E}m(z)$, $\mathbb{E}f_B(z)$, and $\mathbb{E}m(z)/\mathbb{E}f_B(z)$ can be represented by uniformly convergent series in z^{-1} in the area $|z|\geq cK$,*

$$\mathbb{E}m(z)=-z^{-1}+\sum_{k=2}^\infty a_k[m]z^{-k},$$

$$\mathbb{E}f_B(z)=\sum_{k=1}^\infty a_k[f_B]z^{-k},$$

$$\frac{\mathbb{E}f_B(z)}{\mathbb{E}m(z)}=\sum_{k=0}^\infty\beta_k z^{-k}.$$

The proof of the first two equalities is by expansion of $(A+B-z)^{-1}$ and $B(A+B-z)^{-1}$ in convergent series of z^{-1} and estimating the coefficients in these series. This establishes the uniform convergence in the area $|z|>cK$ and ensures that it is possible to take expectation and trace of the series in a term-by-term fashion. The third equality follows from the first two. □

Proof of Proposition 14 By the previous lemma, $\mathbb{E}f_B/\mathbb{E}m_H$ is analytic in z^{-1} and therefore bounded if $|z|>cK$. Since $\|G_A(z)\|\leq 1/\eta$, we can choose $\eta_0=cK$ with sufficiently large c, so that $\eta>\eta_0$ ensures that

$$\left\|\frac{\mathbb{E}f_B(z)}{\mathbb{E}m_H(z)}G_A(z)\right\|<1/2,$$

and

$$\left\|\left(1+\frac{\mathbb{E}f_B(z)}{\mathbb{E}m_H(z)}G_A(z)\right)^{-1}\right\|<2.$$

□

Proof of Proposition 11 For every matrix X, it is true that $|N^{-1}\mathrm{Tr}(x)|\leq\|X\|$. Hence, by using Propositions 12 and 14,

$$\left|\frac{1}{N}\mathrm{Tr}\left(\frac{1}{1+(\mathbb{E}f_B/\mathbb{E}m_H)\,G_A}\mathbb{E}\Delta_A\right)\right|\leq\left\|\frac{1}{1+(\mathbb{E}f_B/\mathbb{E}m_H)\,G_A}\right\|\|\mathbb{E}\Delta_A\|\leq\frac{c}{N\eta^3},$$

provided that $\eta>\eta_0=cK$.

By using the power expansion for $m(z)$, we find $m(z)^{-1}\leq 2\,|z|\leq 2\eta\sqrt{1+\kappa^{-2}}$ if $|z|>cK$. It follows that for $z\in\Omega_{\eta_0,\kappa}$,

$$|R_A|\leq\frac{c}{N\eta^2}\sqrt{1+\kappa^{-2}}.$$

□

5 Stability of the Pastur–Vasilchuk system

By results of [16], the solution of system (10) exists and unique in the upper half-plane $\mathbb{C}^+$. We are going to show that the solutions of systems (10) and (5) are close to each other.

Proposition 16 *For all* $z\in\Omega_{cK,\kappa}$,

$$\max\left\{\left|\mathbb{E}m_H(z)-m_{\boxplus,N}(z)\right|\right\}\leq\frac{c'}{N\eta},$$

where c *and* c' *depends on* K *and* κ *only.*

The idea of proof is to use the solution of the system (5) as the starting point of the Newton-Kantorovich algorithm [14] that computes the solution of system (10).

It is convenient to use a more uniform notation, so we write system (10) in a more compact form:

$$
\begin{aligned}
x_1 - m_A\left(z - \frac{x_3}{x_1}\right) - R_A &= 0,\\
x_1 - m_B\left(z - \frac{x_2}{x_1}\right) - R_B &= 0,\\
zx_1 - x_2 - x_3 + 1 &= 0,
\end{aligned}
\tag{11}
$$

The starting point of the algorithm is $x^{\boxplus} = (m_{\boxplus,N}, S_A m_{\boxplus,N}, S_B m_{\boxplus,N})$, where $m_{\boxplus,N}(z)$, $S_A(z)$, and $S_B(z)$ are the solutions of (5).The variable z plays the role of a parameter.

We assume that R_A and R_B are evaluated at the solution of (10) and fixed. Hence, in (11), R_A and R_B do not depend on x. The solution of (10) remains a solution of this simplified system.

In a shorter form, system (11) can be written as

$$P(x) = 0. \tag{12}$$

Now, let us explain the Newton-Kantorovich method. Let (12) be a general non-linear functional equation where P is a non-linear operator that sends elements of a Banach space X to a Banach space Y. Let P be twice differentiable, and assume that the operator $P'(x)$ has an inverse $[P'(x)]^{-1} \in L(Y, X)$ where $L(Y, X)$ denotes the space of bounded linear operators from Y to X. Then the Newton-Kantorovich method is given by the equation

$$x_{n+1} = x_n - [P'(x_n)]^{-1} P(x_n).$$

The Kantorovich theorem (i) gives the sufficient conditions for the convergence of this process, (ii) estimates the speed of convergence, and (iii) estimates the distance of the solution x^* from the initial point x_0. We give the statement of the theorem omitting the claim about the speed of convergence, which is not important for us.

Theorem 17 (Kantorovich) *Suppose that the following conditions hold*:

1. *for an initial approximation* x_0, *the operator* $P'(x_0)$ *possesses an inverse operator* $\Gamma_0 = [P'(x_0)]^{-1}$ *whose norm has the following estimate:* $\|\Gamma_0\| \leq C_0$,
2. $\|\Gamma_0 P(x_0)\| \leq \delta_0$,
3. *the second derivative* $P''(x)$ *is bounded in the domain determined by inequality* (13) *below; namely,* $\|P''(x)\| \leq M$,

4. *the constants* C_0, δ_0, M *satisfy the relation* $h_0 = C_0\delta_0 M \leq 1/2$.
   Then equation (12) *has a solution* x^*, *which lies in a neighborhood of* x_0 *determined by the inequality*

$$\|x - x_0\| \leq \frac{1 - \sqrt{1 - 2h_0}}{h_0}\delta_0, \tag{13}$$

and the successive approximations x_n *of the Newton method converge to* x^*.

Proof of Proposition 16 In order to apply the Newton-Kantorovich method, let us calculate the derivative $P'(x)$ for our system:

$$P'(x) = \begin{pmatrix} 1 - m_A'\left(z - \frac{x_3}{x_1}\right)\frac{x_3}{x_1^2} & 0 & m_A'\left(z - \frac{x_3}{x_1}\right)\frac{1}{x_1} \\ 1 - m_B'\left(z - \frac{x_2}{x_1}\right)\frac{x_2}{x_1^2} & m_B'\left(z - \frac{x_2}{x_1}\right)\frac{1}{x_1} & 0 \\ z & -1 & -1 \end{pmatrix}.$$

Then, the determinant is

$$\det(P') = -\frac{m_A' + m_B'}{x_1} + \frac{m_A' m_B'}{x_1^3}(-zx_1 + x_2 + x_3),$$

where m_A' and m_B' are short notations for $m_A'\left(z - \frac{x_3}{x_1}\right)$ and $m_B'\left(z - \frac{x_2}{x_1}\right)$, respectively.

The power expansions from Lemma 15 and the definitions of m_A and m_B imply that $x_1 \sim -z^{-1}$, $x_2 \sim \alpha_0 z^{-1}$, $x_3 \sim \beta_0 z^{-1}$, $m_A' \sim z^{-2}$, and $m_B' \sim z^{-2}$ for $z \to \infty$. Hence

$$\det(P') = \frac{1}{z} + O(1),$$

in the area $|z| > cK$, where the constant in $O(1)$ depends only on K.

(The proof that we gave for Lemma 15 holds only for $x_1 = m_{H_N}(z)$, $x_2 = f_{A_N}(z)$, and $x_3 = f_{B_N}(z)$. However, by using results from free probability, these power expansions can be established in the case when x_1, x_2 and x_3 are defined as $m_{\boxplus,N}$, $S_A m_{\boxplus,N}$ and $S_B m_{\boxplus,N}$, respectively.)

Now, it is easy to calculate the inverse of the derivative and find that

$$\Gamma_0 = [P'(x^{\boxplus})]^{-1} = z\begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} + O(1). \tag{14}$$

Hence

$$\begin{aligned} \|\Gamma_0\| &= |z| + O(1) \\ &\leq 2|z|, \end{aligned}$$

if $z \in \Omega_{cK,\kappa}$ and c is sufficiently large.

By using formula (14), we calculate for $z \in \Omega_{cK,\kappa}$:

$$\begin{aligned}\|\Gamma_0 P(x^{\boxplus})\| &\le |z|(|R_A| + |R_B|) + O(|R_A| + |R_B|)\\ &\le c\eta(|R_A| + |R_B|) \le \frac{c'}{N\eta},\end{aligned}$$

where c' depends only on K and κ by Proposition 11.

The next step is to estimate $\|P''(x)\|$. Assume that $\|x - x^{\boxplus}\| \le \frac{1}{2}|z|^{-1}$. [Later we will show that for large N this disc contains the disc given by (13).] By direct computation of the second derivatives, it is easy to check that if c is sufficiently large and $z \in \Omega_{cK,\kappa}$, all second derivatives of $P(x)$ are bounded by a constant, which can depend on K only. Hence, $\|P''(x)\| \le M$, where M depends on K only.

Now we can apply Theorem 17 with $C_0 = 2|z|$, $\delta_0 = c'/N\eta$, M as in the previous paragraph, and $h_0 = C_0\delta_0 M$. For all sufficiently large N, $h_0 \le 1/2$ and disc (13) is inside the disc $\|x - x^{\boxplus}\| \le \frac{1}{2}|z|^{-1}$ so that the estimate for the second derivative holds.

Hence by Theorem 17, if $z \in \Omega_{cK,\kappa}$, then the Newton algorithm which starts at $x^{\boxplus}$ will converge to a solution of $P(x) = 0$ and this solution satisfies inequality $\|x - x^{\boxplus}\| \le 2\delta_0 = c/(N\eta)$. This completes the proof of Proposition 16. □

6 Hadamard's three circle theorem

So far, we established the behavior of the difference $|\mathbb{E}m_{H_N}(z) - m_{\boxplus,N}(z)|$ only for the points where $\mathrm{Im}z \ge \eta_0$. Here we prove a result about its behavior for small $\mathrm{Im}z$.

Proposition 18 *Let I_{η_N} be a straight line segment between points $-2K + i\eta_N$ and $2K + i\eta_N$, where $\eta_N \ge c_1/\sqrt{\log N}$, and c_1 is a positive constant that can depend on K. Then,*

$$\sup_{z\in I_{\eta_N}} \left|\mathbb{E}m_H(z) - m_{\boxplus,N}(z)\right| \le \exp\left(-c\sqrt{\log N}\right),$$

where c depends only on K.

Corollary 19 *Let $\eta_N = c_1 (\log N)^{-\alpha}$, where $0 < \alpha \le 1/2$ and I_{η_N} be a straight line segment between points $-2K + i\eta_N$ and $2K + i\eta_N$. Then,*

$$P\left\{\sup_{z\in I_{\eta_N}} |m_H(z) - m_{\boxplus,N}(z)| > \delta\right\} \le \exp(-c_2\delta^2 N^2(\log N)^{-4\alpha}).$$

Constants c_1 and c_2 depend only on K.

Proof of Corollary 19 This result follows from Corollary 6 and Proposition 18, which estimate $|m_H - \mathbb{E}m_H|$ and $\left|\mathbb{E}m_H - m_{\boxplus,N}\right|$, respectively, if we note that for sufficiently large N, $\left|\mathbb{E}m_H(z) - m_{\boxplus,N}(z)\right| < \delta$ for all $z \in I_{\eta_N}$. □

For the proof of Proposition 18, we use the three circle theorem by Hadamard [12,17].

Theorem 20 (Hadamard's three circle theorem) *Suppose that $f(z)$ is a function of a complex variable z, holomorphic for $|z| < 1$, and let $M(r) = \sup_\theta f(re^{i\theta})$ for $r < 1$. Then $M(r)$ possesses the following properties:*
1. *$M(r)$ is an increasing function of r;*
2. *$\log M(r)$ is a convex function of $\log r$, so that*

$$\log M(r) \leq \frac{\log(r_2/r)}{\log(r_2/r_1)} \log M(r_2) + \frac{\log(r/r_1)}{\log(r_2/r_1)} \log M(r_1)$$

if

$$0 < r_1 \leq r \leq r_2 < 1.$$

We will need the following consequence of this theorem.

Lemma 21 *Suppose $f(z)$ is holomorphic for $|z| < 1$, and let $M(r)$ be defined as in Theorem 20. Suppose that $M(r) \leq c/(1-r)$ for all $r < 1$ and that $M(e^{-1}) \leq \delta$, where $0 < \delta < \delta_0$ and δ_0 depends only on c. Let*

$$r(\delta) = \exp\left(-4\sqrt{c/\log(1/\delta)}\right),$$
$$\varepsilon(\delta) = \exp\left(-\sqrt{c\log(1/\delta)}\right).$$

Then

$$M(r) \leq \varepsilon(\delta)$$

for all $r \leq r(\delta)$ (Fig. 1).

(Note that if $\delta \to 0$, then $r(\delta) \to 1$ and $\varepsilon(\delta) \to 0$.)
Proof Let $L_\delta = \log(1/\delta)$, $r_0 = \exp(-\sqrt{c/L_\delta})$, and $s_0 = \log r_0 = -\sqrt{c/L_\delta}$. By assumption,

$$M(r_0) \leq \frac{c}{1-\exp(-\sqrt{c/L_\delta})} \leq 2\sqrt{cL_\delta}$$

for all sufficiently small δ. In the plane $(\log r, \log M)$, the equation of the straight line that goes through points $(-1, -L_\delta)$ and $(\log r_0, 2\sqrt{cL_\delta})$ is given by

$$l(s) = \frac{2\sqrt{cL_\delta} + L_\delta}{-\sqrt{c/L_\delta} + 1}(s+1) - L_\delta$$
$$= L_\delta \left[\frac{2\sqrt{c} + \sqrt{L_\delta}}{-\sqrt{c} + \sqrt{L_\delta}}(s+1) - 1\right].$$

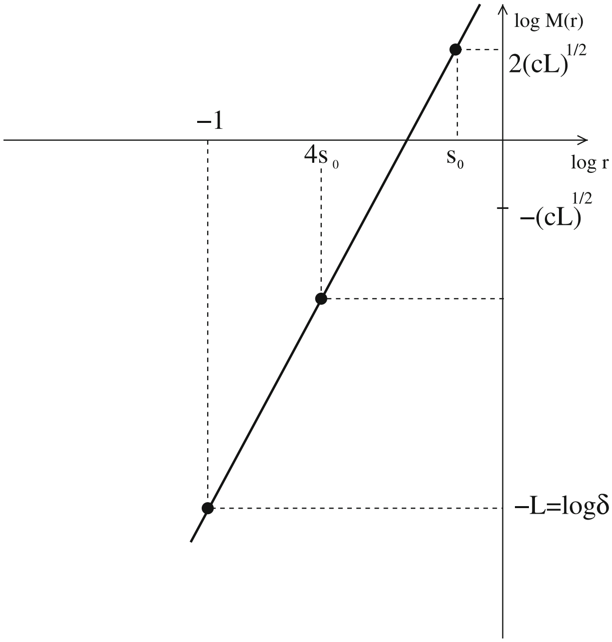

Fig. 1 Illustration to the proof of Lemma 21

By Hadamard's theorem, $\log M(e^s) \leq l(s)$ for all $s \in [-1, s_0]$. Let us set $\bar{s} = 4s_0 = -4\sqrt{c/L_\delta}$. Then,

$$l(\bar{s}) = -\sqrt{cL_\delta} - \frac{9c\sqrt{L_\delta}}{-\sqrt{c} + \sqrt{L_\delta}} \leq -\sqrt{cL_\delta}$$

if $L_\delta > c$.

Hence,

$$\log M(e^s) \leq -\sqrt{cL_\delta}$$

if $s = \log r \leq \bar{s} = -4\sqrt{c/L_\delta}$ and $\delta \leq \delta_0(c)$. □

Since we are interested in functions on the upper half-plane rather than on the unit disc, we have to make a change of variables before we are able to apply Hadamard's theorem. Consider the following map:

$$z = \frac{w - ia}{w + ia},$$

where a is a positive real number. This map sends the upper half-plane $\mathbb{C}^+ = \{w : \mathrm{Im} w \geq 0\}$ bijectively to the unit disc $D = \{z : |z| \leq 1\}$. In particular, it sends point ia

to the center of the disc. The inverse transformation is

$$w = ia\frac{1+z}{1-z}.$$

Let $x \in \mathbb{R}$ and let $\xi = (x - ia)/(x + ia) \in \partial D$. Then

$$\frac{1}{x-w} = \frac{1}{2ai}\frac{(1-\xi)(1-z)}{\xi - z}.$$

Let

$$g(w) = \int_{-\infty}^{\infty} \frac{d\mu(x)}{x-w},$$

where $\mathrm{Im} w > 0$ and μ is the difference of two probability measures. After the change of variable $w = w(z)$, this function becomes a function of variable $z \in D$. We will denote it as $f(z)$. Then,

$$f(z) = \frac{1}{2ai}\int_{\partial D} \frac{(1-\xi)(1-z)}{\xi - z} d\nu(\xi), \tag{15}$$

where $|z| < 1$ and ν is the forward image of μ, hence it is the difference of two probability measures on the unit circle ∂D.

Evidently, $f(z)$ is analytic for $|z| < 1$.

Lemma 22 *Let $f(z)$ be defined by formula (15) with ν which is the difference of two probability measures on ∂D. Then, $M(r) \leq 4a^{-1}(1-r)^{-1}$.*

Proof Clearly, $|1-\xi| \leq 2$, $|1-z| \leq 2$, and $|\xi - z| \geq 1 - |z|$. It remains to notice that the total variation of ν is bounded by 2, since it is a difference of two probability measures. These facts imply that $|f(z)| \leq 4a^{-1}(1-|z|)^{-1}$. □

Proof of 18 The map $w = ia\frac{1+z}{1-z}$ sends disc $B(0, e^{-1})$ to a disc $D_1 \in \mathbb{C}^+$ that has the diameter

$$\left[ia\frac{e-1}{e+1}, ia\frac{e+1}{e-1}\right].$$

By an appropriate choice of a, disc D_1 can be placed arbitrarily far from the real axis, hence we can apply Proposition 16 and write

$$\sup_{w \in D_1} |\mathbb{E}m_H(w) - m_{\boxplus,N}(w)| \leq \frac{c'}{aN}, \tag{16}$$

where c' depends on K.

Next, define $\delta = c'/(aN)$ and let $r(\delta) = \exp(-8a^{-1}/\sqrt{\log(1/\delta)})$ as in Lemma 21 with parameter $c = 4a^{-1}$. The map $w = ia\frac{1+z}{1-z}$ sends disc $B(0, r(\delta))$ to disc $D_2 \in \mathbb{C}^+$ with the diameter

$$ia\left[\frac{1-r(\delta)}{1+r(\delta)}, \frac{1+r(\delta)}{1-r(\delta)}\right].$$

Note that the radius of D_2 approaches infinity as $\delta \downarrow 0$, and that

$$ia\frac{1-r(\delta)}{1+r(\delta)} \sim 4i\sqrt{\frac{a}{\log(1/\delta)}} = 4i\sqrt{\frac{a}{\log(aN/c')}}.$$

It follows that there exists a $c_1 > 0$ such that for $\eta_N = c_1/\sqrt{\log N}$ all the points of the segment I_{η_N} are located inside the disc D_2.

Hence, Lemma 21 and estimate (16) imply that

$$\sup_{w \in I_{\eta_N}} |\mathbb{E}m_H(w) - m_{\boxplus,N}(w)| \le \exp(-2\sqrt{a^{-1}\log(aN/c')}) \tag{17}$$

$$\le \exp(-c_2\sqrt{\log N}). \tag{18}$$

□

7 Proof of Theorems 1 and 2

We use the following result due to Bai (see Theorems 2.1, 2.2, and Corollary 2.3 in [2]). We formulate it in the form suitable for our application

Theorem 23 (Bai) *Let* $K = \max\{\|A_N.B_N\|\}$. *Then,*

$$\sup_x |\mathcal{F}_{H_N}(x) - \mathcal{F}_{\boxplus,N}(x)| \le c_1 \left[\int_{-c_2K}^{c_2K} |m_H(E+i\eta) - m_{\boxplus,N}(E+i\eta)| dE + \frac{1}{\eta}\sup_E \int_{|x|\le 4\eta} |\mathcal{F}_{\boxplus,N}(E+x) - \mathcal{F}_{\boxplus,N}(E)| dx\right], \tag{19}$$

where c_1 *and* c_2 *are numeric.*

Proof of Theorem 1 By using Assumption $A1$, we can estimate

$$|\mathcal{F}_{\boxplus,N}(E+x) - \mathcal{F}_{\boxplus,N}(E)| \le T_N|x|,$$

and therefore the second term on the right-hand side of (19) is bounded by $16T_N\eta$.

Let us set $\eta_N = c_1(\log N)^{-\varepsilon/4}$, where $0 < \varepsilon \leq 2$. By Proposition 18, we can make

$$\sup_{z \in I_{\eta_N}} |\mathbb{E} m_H(z) - m_{\boxplus,N}(z)| \leq \delta/3,$$

provided that $N > (3/\delta)^{c \log(3/\delta)}$. We can also make $16 T_N \eta_N \leq \delta/3$ by choosing $N \geq \exp((c/\delta)^{4/\varepsilon})$.

Then, we can use Bai's theorem and Corollary 19, and find that for all sufficiently large N

$$\begin{aligned} P\left\{\sup_x |\mathcal{F}_{H_N} - \mathcal{F}_{\boxplus,N}| > \delta\right\} &\leq P\{\sup_{z \in I_{\eta_N}} |m_H(z) - m_{\boxplus,N}(z)| \geq c\delta\} \\ &\leq P\{\sup_{z \in I_{\eta_N}} |m_H(z) - \mathbb{E} m_H(z)| \geq c_1\delta\} \\ &\leq \exp(-c_2\delta^2 N^2 (\log N)^{-\varepsilon}), \end{aligned}$$

where to make sure that the last inequality holds, it is enough to take

$$N \geq c_1(\sqrt{\log(1/(\eta^2\delta))})/(\eta^2\delta).$$

For small δ, the most binding inequality on N is $N \geq \exp((c/\delta)^{4/\varepsilon})$. □

By using Theorem 1, we can derive the following corollary and prove Theorem 2. Recall that $\mathcal{N}_I$ denotes the number of eigenvalues of H in the interval I.

Corollary 24 *Suppose the assumptions of Theorem 1 hold, and assume in addition that $\eta \geq c/(\varepsilon\sqrt{\log N})$. Then the following inequality holds*:

$$P\left\{\sup_{I,|I|=\eta} \left|\frac{\mathcal{N}_I}{N|I|} - \frac{\mu_{\boxplus,N}(I)}{|I|}\right| \geq \varepsilon\right\} \leq \exp\left(-c\varepsilon^2 \frac{(\eta N)^2}{(\log N)^2}\right),$$

where $c > 0$ depends only on K and T.

Proof Let $I = (a, b]$. Then $\mathcal{N}_I/N = \mathcal{F}_{H_N}(b) - \mathcal{F}_{H_N}(a)$ and $\mu_{\boxplus,N}(I) = \mathcal{F}_{\boxplus,N}(b) - \mathcal{F}_{\boxplus,N}(a)$, and therefore

$$\begin{aligned} &P\left\{\sup_{I,|I|=\eta} \left|\frac{\mathcal{N}_I}{N|I|} - \frac{\mu_{\boxplus,N}(I)}{|I|}\right| \geq \varepsilon\right\} \\ &= P\left\{\sup_{a,b:b-a=\eta} |\mathcal{F}_{H_N}(b) - \mathcal{F}_{\boxplus,N}(b) - (\mathcal{F}_{H_N}(a) - \mathcal{F}_{\boxplus,N}(a))| \geq \varepsilon\eta\right\}, \end{aligned}$$

and the corollary is the direct consequence of Theorem 1. The assumption about η is needed to ensure that N in Theorem 1 is sufficiently large and is forced by assumptions of Proposition 18. □

Proof of Theorem 2 Assumption $A1$ with uniform T ensures that $\mu_{\boxplus,N}(I)/|I|$ approaches $\varrho_{\boxplus,N}(E)$ when $I = (E-\eta, E+\eta]$ and $\eta \to 0$. Moreover, the convergence is uniform in E. Hence the conclusion of the theorem is implied by Corollary 24. □

8 Concluding remarks

We have shown that the probability of a large deviation of the empirical c.d.f. of eigenvalues of $A_N + U_N B_N U_N^*$ from the c.d.f. of $\mu_{A_N} \boxplus \mu_{B_N}$ is bounded by $\exp(-c\delta^2 N^2/\log^{\varepsilon} N)$. The same results holds for the ensemble in which U_N denotes a Haar-distributed real orthogonal matrix. In this case Lemma 10 does not hold as stated and should be corrected. After this correction the identity in Proposition 8 becomes:

$$\mathbb{E}(m_H G_H) = \mathbb{E}(m_H G_A - G_A f_B G_H) - \frac{1}{N} G_A \mathbb{E}([(G_H)^T, B] G_H).$$

Hence, we need to re-define Δ_A by adding an additional term

$$-N^{-1} G_A [(G_H)^T, B] G_H.$$

The norm of this term is bounded by $c/(N\eta^3)$, therefore the estimate $\|E\Delta_A\| \le c/(N\eta^3)$ from Proposition 12 remains valid and further analysis can be carried through without changes.

It would be interesting to investigate whether the empirical measure of eigenvalues satisfies the large deviation principle. At the very least, it should be expected that the limit

$$\lim_{N\to\infty} -\frac{1}{N^2} \log P\{|\mathcal{F}_{H_N}(x) - \mathbb{E}\mathcal{F}_{H_N}(x)| > \delta\}$$

exists and is positive. It is also likely that the large deviation principle holds at the level of measures. For classical Gaussian ensembles the large deviation rate is closely related to the free entropy of a probability measure:

$$\Sigma(\mu) = \int \log[x-y] d\mu(x) d\mu(y).$$

For more general large matrices with Gaussian entries, the large deviation rates were obtained in the work of Guionnet. It is not clear if there are similar formulas for the large deviation rate in the case of sums of random matrices.

The second contribution of this paper is a local law for eigenvalues. It was shown that the local law holds on the scale $(\log N)^{-1/2}$. It would be interesting to extend this law to smaller scales. In the case when the eigenvalue distributions of matrices A_N and B_N converge to limiting distributions μ_A and μ_B with the free convolution $\mu_A \boxplus \mu_B$, the author expects that the local law holds on the scale $N^{-1+\varepsilon}$ at all points where the density of the free convolution exists. (A trivial cases when μ_A or μ_B are concentrated on a single point should of course be ruled out.)

Currently, the limit laws on this scale are known for the Gaussian symmetric and sample covariance matrices, where they are implied by the explicit description of the limiting eigenvalue process on the scale N^{-1}. They have also been established in [9] for the Wigner and sample covariance random matrices. In this case, the local laws have been used as the first step in the proof of the universality conjecture for this class of random matrices.

Another area of possible further research is to understand better the local structure of the eigenvalues, in particular, the point process of eigenvalues and compare it to the structure of eigenvalues in classical ensembles of random matrices. One would expect that the point process of eigenvalues converges to a universal limit.

References

1. Anderson, G.W., Guionnet, A., Zeitouni, O.: An introduction to random matrices. In: Cambridge Studies in Advanced Mathematics, vol. 118. Cambridge University Press, Cambridge (2009)
2. Bai, Z.D.: Convergence rate of expected spectral distributions of large random matrices, Part I Wigner matrices. Ann. Probab. **21**, 625–648 (1993)
3. Belinschi, S.T.: The lebesgue decomposition of the free additive convolution of two probability distributions. Probab. Theory Relat. Fields **142**, 125–150 (2008)
4. Arous, G.B., Guionnet, A.: Large deviations for Wigner's law and Voiculescu's non-commutative entropy. Probab. Theory Relat. Fields **108**, 517–542 (1997)
5. Bercovici H., Voiculescu, D.: Regularity questions for free convolutions. In: Bercovici, H., Foias, C. (eds.) Nonselfadjoint Operator Algebras, Operator Theory and Related Topics, Operator Theory Advances and Applications, vol. 104, pp. 37–47. Birkhauser, Basel (1998)
6. Biane, P.: Processes with free increments. Mathematische Zeitschrift **227**, 143–174 (1998)
7. Blower, G.: Random Matrices: High Dimensional Phenomena volume 367 of London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge (2009)
8. Chatterjee, S.: Concentration of Haar measures with an application to random matrices. J. Funct. Anal. **245**, 379–389 (2007)
9. Erdos, L., Schlein, B., Yau, H.-T., Yin, J.: The local relaxation flow approach to universality of the local statistics for random matrices. preprint arXiv:0911.3687 (2009)
10. Erdos, L., Schlein, B., Yau, H.-T.: Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices. Ann. Probab. **37**, 815–852 (2009)
11. Gromov, M., Milman, V.D.: A topological application of isoperimetric inequality. Am. J. Math. **105**(4), 843–854 (1983)
12. Hardy, G.H.: The mean value of the modulus of an analytic function. In: Proceedings of the London Mathematical Society, vol. 14, pp. 269–277 (1915)
13. Horn, A.: Eigenvalues of sums of Hermitian matrices. Pac. J. Math. **12**, 225–241 (1962)
14. Kantorovich, L.V.: Functional analysis and applied mathematics. Uspekhi Matematicheskih Nauk 3(6):89–185 (1948). English translation available in Kantorovich, L.V.: Selected Works, vol. 2, pp. 171–280. Gordon and Breach Science Publishers, New York (1996)
15. Knutson, A., Tao, T.: The honeycomb model of $GL_n(\mathbb{C})$ tensor products I: proof of the saturation conjecture. J. Am. Math. Soc. **12**, 1055–1090 (1999)
16. Pastur, L., Vasilchuk, V.: On the law of addition of random matrices. Commun. Math. Phys. **214**, 249–286 (2000)
17. Riesz, F.: Sur les valeurs moyennes du module des fonctions harmonique et des fonctions analytiques. Acta Litterarum ac Scientiarum, 1:27–32 (Available in vol. 1 of the collected papers by F. Riesz) (1922/23)
18. Speicher, R.: Free convolution and the random sum of matrices. Publications of RIMS (Kyoto University) **29**, 731–744 (1993)

19. Voiculescu, D., Dykema, K., Nica, A.: Free Random Variables. A.M.S. Providence (CRM Monograph series, No.1) (1992)
20. Voiculescu, D.: Limit laws for random matrices and free products. Invent. Math. **104**, 201–220 (1991)
21. Weyl, H.: Das asymptotische Verteilungsgesetz der Eigenwerte lineare partieller Differentialgleichungen. Math. Ann. **71**, 441–479 (1912)