Journal of Functional Analysis 282 (2022) 109394

Contents lists available at ScienceDirect

Journal of Functional Analysis

www.elsevier.com/locate/jfa

Thermalisation for Wigner matrices

Giorgio Cipolloni [a], László Erdős [a], Dominik Schröder [b,*,1]

[a] *IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria*
[b] *Institute for Theoretical Studies, ETH Zurich, Clausiusstr. 47, 8092 Zurich, Switzerland*

ARTICLE INFO

Article history:
Received 19 February 2021
Accepted 3 January 2022
Available online 21 January 2022
Communicated by Ivan Corwin

MSC:
60B20
15B52
46L54

Keywords:
Global law
Local law
Non-crossing partitions
Asymptotic freeness

ABSTRACT

We compute the deterministic approximation of products of Sobolev functions of large Wigner matrices W and provide an optimal error bound on their fluctuation with very high probability. This generalizes Voiculescu's seminal theorem [36] from polynomials to general Sobolev functions, as well as from tracial quantities to individual matrix elements. Applying the result to e^{itW} for large t, we obtain a precise decay rate for the overlaps of several deterministic matrices with temporally well separated Heisenberg time evolutions; thus we demonstrate the thermalisation effect of the unitary group generated by Wigner matrices.

1. Introduction

Since E. Wigner's pioneering idea [37], random matrices are ubiquitously used to model complex quantum Hamiltonians. Most works deal with the spectacular universality phenomenon of local eigenvalue statistics [30] but the applicability of random matrix

* Corresponding author.
E-mail addresses: giorgio.cipolloni@ist.ac.at (G. Cipolloni), lerdos@ist.ac.at (L. Erdős), dschroeder@ethz.ch (D. Schröder).

[1] Supported by Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zürich Foundation.

https://doi.org/10.1016/j.jfa.2022.109394

theory goes well beyond. The current paper has been motivated to understand the joint distribution of the unitary operator $e^{\mathrm{i}tW}$, i.e. the quantum evolution corresponding to a large $N \times N$ Wigner matrix W, at different, typically large times.

More generally, in our main result we compute the leading deterministic approximation for the random quantity

$$\langle f_1(W)A_1 f_2(W)A_2 \dots f_k(W)A_k\rangle, \tag{1.1}$$

and we provide an optimal error bound on its fluctuation. Here f_i's are Sobolev test-functions, A_i's are bounded deterministic matrices (observables) and $\langle R\rangle := \frac{1}{N}\operatorname{Tr} R$ denotes the normalized trace of any matrix $R \in \mathbf{C}^{N\times N}$. The deterministic approximation is a sum of several explicit terms, labelled by non-crossing partitions of k elements. Whenever all $f_i = p_i$ are polynomials, such formulas are routinely generated in free probability theory by evaluating $\tau(p_1(s)a_1p_2(s)a_2 \dots p_k(s)a_k)$ in a non-commutative $*$-algebra $\mathcal{A}$ with a tracial state τ, where s is a semicircular element and the set $\{a_1, a_2, \dots, a_k\}$ is *freely independent* of s. Voiculescu's classical result [36] and its extensions from Gaussian (GUE) to general Wigner matrices and to include deterministic matrices, see [3, Theorem 5.4.5] and [31, Sect. 4, Thm. 20], assert that

$$\mathbf{E}\langle p_1(W)A_1p_2(W)A_2 \dots p_k(W)A_k\rangle \to \tau(p_1(s)a_1p_2(s)a_2 \dots p_k(s)a_k), \tag{1.2}$$

where the k-tuple $(a_1, \dots, a_k) \in \mathcal{A}^k$ is the distributional limit of $(A_1, \dots, A_k) \in (\mathbf{C}^{N\times N})^k$ as $N \to \infty$. Several independent Wigner matrices can also be considered on the left hand side; they are modelled by freely independent semicircular elements in the right hand side.

Our Theorem 2.6 extends (1.2) in several important directions. First, we can handle general Sobolev functions $f_i \in H^2(\mathbf{R})$ and not only polynomials since we circumvent the moment method used in free probability theory. We can even consider certain N-dependent functions living on mesoscopic scales. Second, we control the convergence in (1.2) immediately in very high probability and not only in expectation, saving additional variance and high moment calculations typically performed separately with the moment method. This strengthening allows us to directly handle several independent random matrices instead of a single W, just by simple conditioning; the similar extension in the standard free probability approach requires considerably more sophisticated combinatorics. Third, we obtain an optimal error term of order N^{-1} involving the k-th Sobolev norms of f_i and we have a freedom to trade in weaker bounds for less smoothness assumption down to $f_i \in H^2$. Fourth, we obtain similar deterministic approximations with optimal error terms not only for the normalized traces (1.1) but for all matrix elements $\langle \boldsymbol{x}, f_1(W)A_1 \dots f_k(W)A_k\boldsymbol{y}\rangle$ with any deterministic vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^N$. Note that individual matrix elements have no counterpart in the limiting algebra $\mathcal{A}$, so they are beyond the scope of standard free probability theory. Finally, our deterministic approximations are obtained before the $N \to \infty$ limit is taken, hence the convergence of the deterministic matrices $A_1, \dots, A_k$ is not required.

In our main applications we consider (1.1) with the exponential functions $f_j(x) = e^{\mathrm{i}s_j x}$ and we are primarily interested in the decay of (1.1) for large times $s_j \gg 1$. This problem has two related motivations originating from mathematical physics and free probability theory, respectively, that we briefly explain.

The classical RAGE theorem [15, Section 5.4] for self-adjoint operators H on an infinite dimensional Hilbert space $\mathcal{H}$ shows that the Heisenberg time evolution $A(t) = e^{\mathrm{i}tH}Ae^{-\mathrm{i}tH}$ of a compact operator A asymptotically vanishes on any state $\psi \in \mathcal{H}$ in the continuous spectral subspace of H; more precisely $\langle \psi, A(t)\psi\rangle$ tends to zero in Cesaro mean for large time t. Since acting on a finite dimensional space, large $N \times N$ Wigner matrices W do not have continuous spectrum in a literal sense, but for many physical purposes they still behave as an operator with continuous spectrum; for example their eigenvectors are completely delocalized [20,21,8]. Hence the analogue of the RAGE theorem for Wigner matrices would assert that the matrix elements of $A(t) := e^{\mathrm{i}tW}Ae^{-\mathrm{i}tW}$ at any fixed deterministic vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^N$ become very close to their limiting value for large times, i.e.

$$\langle \boldsymbol{x}, A(t)\boldsymbol{y}\rangle \approx \langle \boldsymbol{x}, \boldsymbol{y}\rangle\langle A\rangle \qquad \text{for} \quad t \gg 1. \tag{1.3}$$

We call this phenomenon *thermalisation* as it corresponds to a decay to a certain equilibrium. Similarly, for two bounded deterministic matrices (observables) A and B one expects that $A(t)$ and B become thermalised, i.e.

$$\langle A(t)B\rangle \approx \langle A\rangle\langle B\rangle \qquad \text{for} \quad t \gg 1. \tag{1.4}$$

Exact equalities are not expected in (1.3) and (1.4) even after the $t \to \infty$ limit as a consequence of the finite dimensionality. Our Theorem 2.6 in this context proves the thermalisation mechanism with a precise decay rate for large times, in particular we show that

$$\begin{aligned} \langle \boldsymbol{x}, A(t)\boldsymbol{y}\rangle &= \langle \boldsymbol{x}, \boldsymbol{y}\rangle\langle A\rangle + \theta(t)^2\frac{\langle \boldsymbol{x}, \mathring{A}\boldsymbol{y}\rangle}{t^3} + \mathcal{O}\left(N^{\epsilon}\frac{t^2}{N^{1/2}}\right) \\ \langle A(t)B\rangle &= \langle A\rangle\langle B\rangle + \theta(t)^2\frac{\langle \mathring{A}\mathring{B}\rangle}{t^3} + \mathcal{O}\left(N^{\epsilon}\frac{t^2}{N}\right) \end{aligned} \tag{1.5}$$

holds with very high probability for the oscillatory order one function $\theta(t) := J_1(2t)t^{1/2}$, with J_1 a Bessel function of the first kind, and where $\mathring{A} := A - \langle A\rangle$ denotes the traceless part of A. We thus obtain an *approximate RAGE theorem* for Wigner matrices with a precise decay rate in time and with an N-dependent error bound due to the finite dimensionality of the system. The effective error terms in (1.5) allow for a simultaneous limit for large N and t in a certain range. Interestingly, the inverse cubic decay rate stems from the square root singularity of the Wigner semicircle density at the spectral edges. Since this square root behaviour is typical for the density of states in a large

class of random matrix ensembles [1,2], the cubic decay rate is expected to be fairly universal. For brevity, in this paper we focus on the simplest Wigner case, deferring the more general ensembles to future work.

We obtain similar thermalisation results for the multiple time evolutions of several observables and identify the precise rate of time decay in each case. The deterministic approximation has a hierarchical structure that allows us to identify the sector with the slowest (dominant) thermalisation rate. We find that if some observables or their products are traceless, the thermalisation is enhanced. For example, if $B(s) = e^{\mathrm{i}sW} B e^{-\mathrm{i}sW}$ is the time evolution of another deterministic B matrix with $s \gg 1$, $t - s \gg 1$ and C is a third observable, then we obtain

$$\begin{aligned}\langle A(t)B(s)C\rangle = \langle A\rangle\langle B\rangle\langle C\rangle + \theta(s)^2\frac{\langle A\rangle\langle \mathring{B}\mathring{C}\rangle}{s^3} + \theta(t)^2\frac{\langle B\rangle\langle \mathring{A}\mathring{C}\rangle}{t^3} + \theta(t-s)^2\frac{\langle C\rangle\langle \mathring{A}\mathring{B}\rangle}{(t-s)^3} \\ + \theta(s)\theta(t)\theta(t-s)\frac{\langle \mathring{A}\mathring{B}\mathring{C}\rangle}{s^{3/2}t^{3/2}(t-s)^{3/2}} + \mathcal{O}\left(N^{\epsilon}\frac{t^3}{N}\right)\end{aligned} \tag{1.6}$$

with very high probability. Note that the prevailing decay rate is strongly influenced by the possible vanishing of some of the numerators in (1.6). In particular, if all three observables are traceless, $\langle A\rangle = \langle B\rangle = \langle C\rangle = 0$, and the large times t, s and $t - s$ are comparable, the decay rate is the $\frac{9}{2}$-th power of the time.

Our second motivation comes from Voiculescu's theorem [3, Theorem 5.4.5] and [31, Sect. 4, Thm. 20] (see also [36,18,35] for previous results) which asserts that independent $N \times N$ Wigner matrices, $W_1, W_2, \dots W_k$, are asymptotically free. This means that for any collection of polynomials $p_1, p_2, \dots, p_r$ that are (asymptotically) traceless, i.e. $\langle p_j(W)\rangle \to 0$ as $N \to \infty$, we have

$$\langle p_1(W_{i_1})p_2(W_{i_2})\dots p_r(W_{i_r})\rangle \to 0, \qquad \text{as} \quad N \to \infty, \tag{1.7}$$

in expectation and almost surely, where the product is *alternating* in the sense that $i_1 \neq i_2$, $i_2 \neq i_3, \dots, i_{r-1} \neq i_r$.

The asymptotic freeness property (1.7) of *independent* Wigner matrices is a fundamental result that connects random matrices with free probability. Using the thermalisation mechanism we show that not only independent Wigner matrices are asymptotically free, but different long time evolutions by the *very same* Wigner matrix also make deterministic observables asymptotically free. More precisely, we show that the Heisenberg time evolutions of arbitrary deterministic observables, $A_1(t_1), A_2(t_2), \dots, A_k(t_k)$ are asymptotically free whenever all time differences $|t_i - t_j|$ are very large. Equivalently, we prove that with very high probability for any polynomials $p_1, \dots, p_r$

$$\langle p_1(A_{i_1}(t_{i_1}))\cdots p_r(A_{i_r}(t_{i_r}))\rangle \to 0 \quad \text{as} \quad N \to \infty \quad \text{and} \quad \min_{i\neq j}|t_i - t_j| \to \infty, \tag{1.8}$$

whenever $i_1 \neq i_2$, $i_2 \neq i_3, \dots, i_{r-1} \neq i_r$ and $\langle p_j(A_{i_j}(t_{i_j}))\rangle \to 0$ for all $j = 1, \dots, k$. The precise statement with effective error bounds is given in Corollary 2.12. We stress that the mechanism to obtain asymptotic freeness via thermalisation in (1.8) is very different from the one behind (1.7) relying on independence. A freeness mechanism similar to ours was demonstrated for different powers of the *same* Haar unitary matrix by Haagerup and Larsen in [22, Lemma 3.7].

In order to understand (1.1), we first derive a new *multi-resolvent local law* in Theorem 3.4, i.e. we identify the deterministic approximation of $G(z_1)A_1G(z_2)A_2 \dots G(z_k)$ for the resolvents, $G(z) = (W - z)^{-1}$, and then extend it to general Sobolev functions via the Helffer-Sjöstrand calculus. For a single resolvent the deterministic approximation $G(z) \approx m(z)I$ is given by the unique scalar solution $m = m(z)$ to the Dyson equation $-1/m = m + z$, both in *averaged sense*, $\langle G(z)\rangle \approx m(z)$, and in *isotropic sense*, $\langle \boldsymbol{x}, G(z)\boldsymbol{y}\rangle \approx m(z)\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ for any vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^N$. The multi-resolvent local law is proven by recursively analysing a system of self-consistent equations that is an adapted version of the deterministic Schwinger-Dyson equation obtained from second moment Gaussian calculation. The fluctuation term in this approximation has been estimated in our recent work [11]. Our approach also works in the mesoscopic regime, i.e. when the imaginary part of the spectral parameter in $G(z)$ is small as a negative power of N. In turn, this allows us to analyse the unitary time evolution $e^{\mathrm{i}tW}$ for very long, even N-dependent, times. The mesoscopic regime, however, requires to identify a multiple cancellation effect in the deterministic approximation. Amusingly, we need two very different, but eventually equivalent formulas for this approximation; one is based on non-crossing graphs (see Lemma 5.2 later) and arises naturally from the recursive structure of the Dyson equation. The other one from (5.12) is a partial resummation of the first one in terms of non-crossing partitions and the free cumulant function of divided differences of m, manifesting the cancellation.

Voiculescu's theorem (1.2) or asymptotic freeness in the form (1.7) for independent Wigner matrices has traditionally been proven with the moment method using very involved combinatorics. It efficiently handles polynomials of fixed degree as stated in (1.7) and can be extended to general functions by polynomial approximation. However, to obtain effective controls (e.g. explicit speed of convergence) or possibly N dependent test functions (like mesoscopic linear statistics) usually requires high (N-dependent) degree for the polynomials that, in turn, are increasingly difficult for the moment method as well as for the analytic subordination method [7]. Thus the extension of the moment method to more general functions has natural limitations, although there is a remarkable recent development for rational functions [24,28,38,14]. The trace of a smooth cut-off function of a polynomial in GUE and deterministic matrices has been analysed via the Master equation and linearization in [23,29] for the purpose of identifying the norm of the polynomial. Recently general smooth functions were considered in the same setup with a new interpolation method between the GUE matrices and their infinite dimensional limits, the semicircular elements [13]. A large N-expansion to arbitrary order was also obtained [33]. We follow a different route via the local laws for resolvents that works

for general Wigner matrices and also for matrix elements, it handles mesoscopic regimes very efficiently and it yields optimal control in very high probability sense offering an alternative to the customary free probability approach.

Notation and conventions. We introduce some notations we use throughout the paper. For integers $l, k \in \mathbf{N}$ we use the notations $[k] := \{1, \dots, k\}$, and

$$[k, l) := \{k, k+1, \dots, l-1\}, \qquad [k, l] := \{k, k+1, \dots, l-1, l\}$$

for $k < l$. For positive quantities f, g we write $f \lesssim g$ and $f \sim g$ if $f \leq Cg$ or $cg \leq f \leq Cg$, respectively, for some constants $c, C > 0$ which depend only on the constants appearing in (2.1). We denote vectors by bold-faced lower case Roman letters $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^N$, for some $N \in \mathbf{N}$. Vector and matrix norms, $\|\boldsymbol{x}\|$ and $\|A\|$, indicate the usual Euclidean norm and the corresponding induced matrix norm. For any $N \times N$ matrix A we use the notation $\langle A \rangle := N^{-1} \operatorname{Tr} A$ to denote the normalized trace of A. Moreover, for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^N$ we define

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \sum \overline{x}_i y_i, \qquad A_{\boldsymbol{xy}} := \langle \boldsymbol{x}, A\boldsymbol{y} \rangle,$$

with $A \in \mathbf{C}^{N \times N}$. For any $z \in \mathbf{C}$, by $\Re z$ and $\Im z$ we denote the real and imaginary part of z, respectively. We will use the concept of "with very high probability" meaning that for any fixed $D > 0$ the probability of the N-dependent event is bigger than $1 - N^{-D}$ if $N \geq N_0(D)$. Moreover, we use the convention that $\xi > 0$ denotes an arbitrary small constant which is independent of N.

Acknowledgment. The authors are very grateful to Roland Speicher for useful correspondence on the problem and pointing out additional references. The authors also thank the anonymous referees whose comments significantly improved the readability of the manuscript, and also Jana Reker for carefully reading the manuscript and spotting several typos.

2. Main results

We consider real symmetric or complex Hermitian $N \times N$ Wigner matrices W. We formulate the following assumptions on the entries of W.

Assumption 2.1. We assume that the matrix elements w_{ab} are independent up to the Hermitian symmetry $w_{ab} = \overline{w_{ba}}$ and identically distributed in the sense that $w_{ab} \stackrel{\mathrm{d}}{=} N^{-1/2}\chi_{\mathrm{od}}$, for $a < b$, $w_{aa} \stackrel{\mathrm{d}}{=} N^{-1/2}\chi_{\mathrm{d}}$, with χ_{od} being a real or complex random variable and χ_{d} being a real random variable such that $\mathbf{E}\,\chi_{\mathrm{od}} = \mathbf{E}\,\chi_{\mathrm{d}} = 0$ and $\mathbf{E}\,|\chi_{\mathrm{od}}|^2 = 1$. In the complex case we also assume that $\mathbf{E}\,\chi_{\mathrm{od}}^2 \in \mathbf{R}$. In addition, we assume the existence of the high moments of χ_{od}, χ_{d}, i.e. that there exist constants $C_p > 0$, for any $p \in \mathbf{N}$, such that

$$\mathbf{E}|\chi_{\mathrm{d}}|^p + \mathbf{E}|\chi_{\mathrm{od}}|^p \le C_p. \tag{2.1}$$

Our main result is the asymptotic evaluation of products of multiple time-evolved observables $e^{-\mathrm{i}tW}Ae^{\mathrm{i}tW}$ for general deterministic matrices A. More generally, we prove that alternating products of functions of Wigner and deterministic matrices like $\langle f(W)Ag(W)B\ldots\rangle$ with high probability concentrate around a deterministic limit which we compute explicitly. In order to state the result we first introduce *non-crossing partitions* [27] and related objects.

Definition 2.2 *(Lattice of non-crossing partitions).* Let $S \subset \mathbf{N}$ be a finite set of integers. We call a partition π of the set S *crossing* if there exist blocks $B \neq B' \in \pi$ with $a, b \in B$, $c, d \in B'$ and $a < c < b < d$, otherwise we call it *non-crossing* and we denote the set of non-crossing partitions by $\mathrm{NC}(S)$. For each non-crossing partition $\pi = \{B_1, \ldots, B_n\} \in \mathrm{NC}(S)$ we denote the number of blocks in the partition by $|\pi| := n$.

We define a partial order $\le$ on $\mathrm{NC}(S)$, the *refinement order*, such that $\pi \le \sigma$ if and only if π is a refinement of σ, i.e. if for each $B \in \pi$ there exists $B' \in \sigma$ such that $B \subset B'$. The partially ordered set $(\mathrm{NC}(S), \le)$ is in fact a lattice as any two $\pi, \sigma \in \mathrm{NC}(S)$ admit unique least upper and greatest lower bounds $\pi \vee \sigma, \pi \wedge \sigma \in \mathrm{NC}(S)$. Moreover, there exist unique maximal and minimal elements $0_S, 1_S \in \mathrm{NC}(S)$ defined by $0_S := \{\{a\} \mid a \in S\}$ and $1_S := \{S\}$.

The following definition is combinatorially identical to the definition of free cumulants of random variables in free probability in terms of the trace functional (see [34, Section 4] or [4] for connections with classical, Boolean and monotone cumulants).

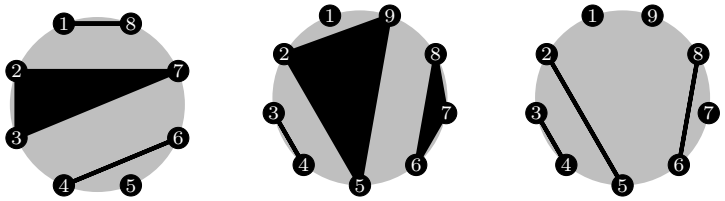
Definition 2.3 *(Free cumulant function).* Fix $k \in \mathbf{N}$ and let $f\colon 2^{[k]} \to \mathbf{C}$ be a function mapping subsets of $[k]$ to scalars. We then implicitly define the *free cumulant function of* f as the unique map $f_\circ\colon 2^{[k]} \to \mathbf{C}$ satisfying that for any $S \subset [k]$ we have

$$f[S] = \sum_{\pi \in \mathrm{NC}(S)} \prod_{B \in \pi} f_\circ[B]. \tag{2.2}$$

The implicit relation (2.2) in Definition 2.3 can be recursively turned into an explicit definition of $f_\circ$. Indeed, for $|S| = 1$ the relation (2.2) implies $f_\circ[i] = f[i]$, so that using (2.2) for $|S| = 2$ it follows that $f_\circ[i,j] = f[i,j] - f[i]f[j]$. For general $S \subset [k]$ the free cumulant function can be written explicitly as

$$f_\circ[S] = \sum_{\pi \in \mathrm{NC}(S)} \mu(\pi, 1_S) \prod_{B \in \pi} f[B], \quad \mu(\pi, \sigma) := \begin{cases} 1, & \pi = \sigma, \\ -\sum_{\pi < \nu \le \sigma} \mu(\nu, \sigma), & \pi < \sigma, \end{cases} \tag{2.3}$$

in terms of the *Möbius function* $\mu\colon \{(\pi, \sigma) \mid \pi \le \sigma \in \mathrm{NC}(S)\} \to \mathbf{Z}$, see Lemma 2.16 later for an alternative non-recursive definition. We note that for $|S| > 3$ the Möbius function depends on the elements of the blocks and not only on the block sizes of π,



(a) {18|237|46|5} (b) {1|259|34|678} (c) {1|25|34|68|7|9}

Fig. 1. Some partitions together with the corresponding disjoint convex hulls.

e.g. $\mu(\{12|3|4\},\{1234\}) = 2 \neq 1 = \mu(\{13|2|4\},\{1234\})$. This is because the concept of non-crossing partition relies on the ordering of $\mathbf{N}$; the Möbius function for the lattice of *all* partitions would be a function of the block sizes alone.

Non-crossing partitions have an alternative geometrical definition. Arrange the elements of S equidistantly in counter-clockwise order on the circle and for each $B \in \pi$ consider the P_B convex hull of the points $x \in B$. Then π is non-crossing if and only if the polygons $\{P_B \mid B \in \pi\}$ are pairwise disjoint, see Fig. 1 for some examples (note that the partition in Fig. 1c is a refinement of the partition in Fig. 1b). We note that for any $\pi \in \mathrm{NC}(S)$ the complement $\mathbf{D} \setminus \cup_{B\in\pi} P_B$ of the polygons $P_B, B \in \pi$ in the disk $\mathbf{D}$ has $|S| - |\pi| + 1$ connected components. The geometrical interpretation is particularly useful for defining the Kreweras complement of non-crossing partitions [27].

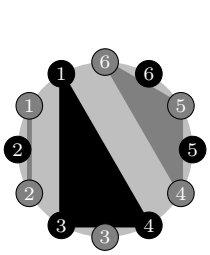
Definition 2.4 *(Kreweras complement).* Let $S \subset \mathbf{N}$ be a set of integers equidistantly arranged in counter-clockwise order on the circle and label the arcs between the points also by S in such a way that the arc x succeeds the point x in counter-clockwise order. Then for $\pi \in \mathrm{NC}(S)$ we define the *Kreweras complement* $K(\pi) \in \mathrm{NC}(S)$ such that $x, y \in S$ belong to the same block of $K(\pi)$ if and only if the arcs x, y are in the same connected component of $\mathbf{D} \setminus \cup_{B\in\pi} P_B$.

In Fig. 2 we give two examples of partitions and their Kreweras complements. We note that $|\pi| + |K(\pi)| = |S| + 1$ for any $\pi \in \mathrm{NC}(S)$. Moreover, $K^2 = K \circ K$ is simply a rotation in the sense that $K^2(\pi)$ is the partition where for $S = \{s_1, \ldots, s_n\}$ the elements in each block of π are shifted by $s_1 \mapsto s_2 \mapsto \cdots \mapsto s_n \mapsto s_1$. In particular the map K on $\mathrm{NC}(S)$ is invertible.

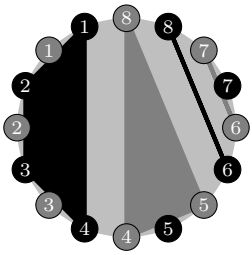
We are now ready to state our main result. We define the semicircular average of any function $f\colon [-2,2] \to \mathbf{C}$ as

$$\langle f \rangle_{\mathrm{sc}} := \int_{-2}^{2} f(x)\rho_{\mathrm{sc}}(x)\,\mathrm{d}x, \quad \rho_{\mathrm{sc}}(x) := \frac{\sqrt{4-x^2}}{2\pi}.$$

Furthermore, we define the π-*partial-trace* and the π-*trace* for partitions $\pi \in \mathrm{NC}[k]$.



(a) $K(\{134|2|5|6\}) = \{12|3|456\}$



(b) $K(\{1234|68|5|7\}) = \{1|2|3|458|67\}$

Fig. 2. Example of partitions (in black) and their Kreweras complements (in dark gray).

Definition 2.5. Let $(\mathcal{A}, \langle\cdot\rangle)$ be a complex tracial algebra. Then for $k \in \mathbf{N}$ and $\pi \in \mathrm{NC}[k]$ we define the *π-partial-trace*[2] $\mathrm{pTr}_\pi \colon \mathcal{A}^{k-1} \to \mathcal{A}$ as

$$\begin{aligned} \langle A_1, \dots, A_k \rangle_\pi &:= \prod_{B \in \pi} \left\langle \prod_{j \in B} A_j \right\rangle, \\ \mathrm{pTr}_\pi(A_1, \dots, A_{k-1}) &:= \left(\prod_{j \in B(k) \setminus \{k\}} A_j \right) \prod_{\substack{B \in \pi \\ B \not\ni k}} \left\langle \prod_{j \in B} A_j \right\rangle, \end{aligned} \tag{2.4}$$

where $B(k) \in \pi$ denotes the block containing k, and all products are ordered increasingly in the indices.

Theorem 2.6. *Let $k \geq 2$, let $A_1, \dots, A_k$ be deterministic matrices with $\|A_i\| \lesssim 1$, and let $f_1, \dots, f_k$ be Sobolev functions $f_i \in H^k([-3,3])$ normalised such that $\|f_i\|_{L^\infty} \sim 1$. Then for any $\xi > 0$ and any deterministic vectors $\boldsymbol{x}, \boldsymbol{y}$ with $\|\boldsymbol{x}\| + \|\boldsymbol{y}\| \lesssim 1$ we have*

$$\begin{aligned} \langle f_1(W) A_1 \dots f_k(W) A_k \rangle &= \sum_{\pi \in \mathrm{NC}[k]} \langle A_1, \dots, A_k \rangle_{K(\pi)} \prod_{B \in \pi} \mathrm{sc}_\circ[B] + \mathcal{O}\left(N^\xi \frac{\max_i \|f_i\|_{H^k}}{N} \right) \\ \langle \boldsymbol{x}, f_1(W) A_1 \dots f_k(W) \boldsymbol{y} \rangle &= \sum_{\pi \in \mathrm{NC}[k]} \langle \boldsymbol{x}, \mathrm{pTr}_{K(\pi)}(A_1, \dots, A_{k-1}) \boldsymbol{y} \rangle \prod_{B \in \pi} \mathrm{sc}_\circ[B] \\ &\quad + \mathcal{O}\left(N^\xi \frac{\max_i \|f_i\|_{H^k}}{N^{1/2}} \right) \end{aligned} \tag{2.5}$$

with very high probability, where $\mathrm{sc}_\circ$ is the free cumulant function from Definition 2.3 of $\mathrm{sc}[i_1, \dots, i_n] := \langle f_{i_1} f_{i_2} \cdots f_{i_n} \rangle_{\mathrm{sc}}$. For $k = 1$ the same result holds with $f \in H^k$ and $\|\cdot\|_{H^k}$ replaced by $f \in H^2$ and $\|\cdot\|_{H^2}$, respectively. A straightforward generalization of (2.5) to include several independent Wigner matrices is given in Extension 2.13.

[2] An analogous partial trace is used in free probability theory for the representation of operator valued conditional expectations, cf. for example the formula for $\tilde{\phi}_\sigma$ above [31, Theorem 19].

Note that by eigenvalue rigidity (see e.g. [19, Theorem 7.6] or [21]) the spectrum of W is contained in $[-2-\epsilon, 2+\epsilon]$, for any small $\epsilon > 0$, with very high probability, hence $f(W)$ is well defined for $f \in H^k([-3,3])$, i.e. functions defined only on $[-3,3]$. In fact, without loss of generality, we may assume that $f \in H_0^k([-3,3])$ by multiplying the original f with a smooth cutoff function without changing $f(W)$. (see Section 4.1 for more details).

Remark 2.7. The average version of (2.5) for polynomial test functions f_i has a long history. In expectation sense the first result of this type was proved for several independent Gaussian (GUE) random matrices with $A_i = I$ in Voiculescu's seminal paper [36]; later upgraded to almost sure convergence in [35]. The extension to Wigner matrices with general entry distribution as well as the inclusion of special block diagonal deterministic matrices A_i was achieved in [18]. The case with arbitrary deterministic matrices can be found in [3, Theorem 5.4.5] and [31, Sect. 4, Thm. 20] (see also [6] under relaxed moment conditions on the entry distribution). The only results beyond polynomials are in [38,14] for certain class of rational functions; general Sobolev functions have not been considered before the current work. Furthermore, the isotropic version of (2.5) is new even for polynomials.

We note that in the language of free probability theory the r.h.s. of (2.5) can be interpreted as follows. If the deterministic matrices A_i converge in the sense of moments to some elements $a_i \in \mathcal{A}$,

$$A_i \xrightarrow{\text{distr}} a_i, \quad \text{as} \quad N \to \infty,$$

of some non-commutative probability space $(\mathcal{A}, \phi)$, then we asymptotically have

$$\langle f_1(W) A_1 \cdots f_k(W) A_k \rangle \to \phi(f_1(s) a_1 \cdots f_k(s) a_k)$$

for a semicircular element $s \in \mathcal{A}$ freely independent of $a_1, \ldots, a_k$.

Example 2.8. For $k = 1, 2, 3$ the deterministic approximation F_k of $f_1(W)A_1 \cdots f_k(W)$ in (2.5) is given as follows. The deterministic approximations of $\langle f_1(W)A_1 \cdots f_k(W)A_k \rangle$ follow by multiplying the expressions below by A_k and taking the trace.

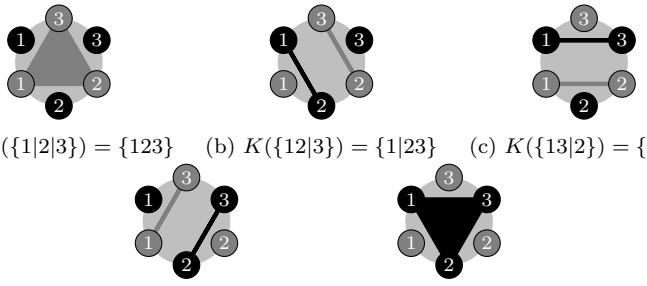
($k = 1$) Here we simply have $F_1 = \langle f_1 \rangle_{\mathrm{sc}}$ since $\mathrm{sc}_\circ[1] = \langle f_1 \rangle_{\mathrm{sc}}$.

($k = 2$) We have

$$\begin{aligned} F_2 &= A_1 \, \mathrm{sc}_\circ[1] \, \mathrm{sc}_\circ[2] + \langle A_1 \rangle \, \mathrm{sc}_\circ[1,2] \\ &= A_1 \langle f_1 \rangle_{\mathrm{sc}} \langle f_2 \rangle_{\mathrm{sc}} + \langle A_1 \rangle (\langle f_1 f_2 \rangle_{\mathrm{sc}} - \langle f_1 \rangle_{\mathrm{sc}} \langle f_2 \rangle_{\mathrm{sc}}) \end{aligned}$$

using $\mathrm{sc}_\circ[1,2] = \langle f_1 f_2 \rangle_{\mathrm{sc}} - \langle f_1 \rangle_{\mathrm{sc}} \langle f_2 \rangle_{\mathrm{sc}}$.

($k = 3$) For $k = 3$ there are five terms (corresponding to the non-crossing partitions visualised in Fig. 3)



(a) $K(\{1|2|3\}) = \{123\}$ (b) $K(\{12|3\}) = \{1|23\}$ (c) $K(\{13|2\}) = \{12|3\}$

(d) $K(\{1|23\}) = \{13|2\}$ (e) $K(\{123\}) = \{1|2|3\}$

Fig. 3. List of all non-crossing partitions NC([3]) on three vertices (in black) together with their Kreweras complement (in dark gray).

$$F_3 = A_1A_2\,\mathrm{sc}_\circ[1]\,\mathrm{sc}_\circ[2]\,\mathrm{sc}_\circ[3] + \langle A_1A_2\rangle\,\mathrm{sc}_\circ[1,3]\,\mathrm{sc}_\circ[2] + A_1\langle A_2\rangle\,\mathrm{sc}_\circ[1]\,\mathrm{sc}_\circ[2,3] + A_2\langle A_1\rangle\,\mathrm{sc}_\circ[1,2]\,\mathrm{sc}_\circ[3] + \langle A_1\rangle\langle A_2\rangle\,\mathrm{sc}_\circ[1,2,3]$$

with

$$\mathrm{sc}_\circ[1,2,3] = \langle f_1f_2f_3\rangle_{\mathrm{sc}} - \langle f_1f_2\rangle_{\mathrm{sc}}\langle f_3\rangle_{\mathrm{sc}} - \langle f_1f_3\rangle_{\mathrm{sc}}\langle f_2\rangle_{\mathrm{sc}} - \langle f_2f_3\rangle_{\mathrm{sc}}\langle f_1\rangle_{\mathrm{sc}} + 2\langle f_1\rangle_{\mathrm{sc}}\langle f_2\rangle_{\mathrm{sc}}\langle f_3\rangle_{\mathrm{sc}}$$

and $\mathrm{sc}_\circ[i], \mathrm{sc}_\circ[i,j]$ as before.

2.1. Thermalisation and asymptotic freeness

We now specialise Theorem 2.6 to the functions $f(x) := e^{\mathrm{i}sx}$, with $s \in \mathbf{R}$, and define

$$\varphi(s) := \langle e^{\mathrm{i}s\cdot}\rangle_{\mathrm{sc}} = \int_{-2}^{2} \rho_{\mathrm{sc}}(x)e^{\mathrm{i}sx}\,\mathrm{d}x = \frac{J_1(2s)}{s}, \tag{2.6}$$

where J_1 is a Bessel function of the first kind. We note that by standard asymptotic of the Bessel function we have

$$J_1(x) = -\cos\Big(x + \frac{\pi}{4}\Big)\sqrt{\frac{2}{\pi x}} + \mathcal{O}\Big(\frac{1}{x^{3/2}}\Big), \qquad \text{for} \quad x \gg 1. \tag{2.7}$$

Corollary 2.9. *Let $k \geq 2$, $s_1, \dots, s_k \in \mathbf{R}$ and let $A_1, \dots, A_k, \boldsymbol{x}, \boldsymbol{y}$ be deterministic matrices and vectors with $\|A_i\| \lesssim 1$ and $\|\boldsymbol{x}\| + \|\boldsymbol{y}\| \lesssim 1$. Then*

$$\langle e^{\mathrm{i}s_1W}A_1\cdots A_{k-1}e^{\mathrm{i}s_kW}A_k\rangle = \sum_{\pi\in\mathrm{NC}[k]} \langle A_1,\dots,A_k\rangle_{K(\pi)} \prod_{B\in\pi}\varphi_\circ[B] + \mathcal{O}\Big(N^\xi \frac{\max_i|s_i|^k}{N}\Big),$$

$$\langle \boldsymbol{x}, e^{\mathrm{i}s_1W}A_1\cdots A_{k-1}e^{\mathrm{i}s_kW}\boldsymbol{y}\rangle = \sum_{\pi\in\mathrm{NC}[k]} \langle \boldsymbol{x}, \mathrm{pTr}_{K(\pi)}(A_1,\dots,A_{k-1})\boldsymbol{y}\rangle \prod_{B\in\pi}\varphi_\circ[B] \tag{2.8}$$

$$+ \mathcal{O}\Big(N^{\xi} \frac{\max_i |s_i|^k}{N^{1/2}} \Big),$$

with very high probability for any $\xi > 0$, where $\varphi_\circ$ is the free cumulant function from Definition 2.3 of $\varphi[i_1, \dots, i_n] := \varphi(s_{i_1} + \cdots + s_{i_n})$, with φ being defined in (2.6).

Corollary 2.9 ensures a time decay if some or all of the involved matrices are traceless. More precisely, we obtain:

Corollary 2.10. *Let $k \geq 2$, and let $s_1, \dots, s_k \in \mathbf{R}$. Then*

(i) for deterministic traceless matrices $A_1, \dots, A_k$ we have the averaged estimate

$$\big| \langle e^{\mathrm{i}s_1 W} A_1 \cdots e^{\mathrm{i}s_k W} A_k \rangle \big| \lesssim \max_{i \neq j} \Big(\frac{1}{(1+|s_i|)(1+|s_j|)} \Big)^{3/2} + N^{\xi} \frac{\max_i |s_i|^k}{N}, \tag{2.9}$$

(ii) and for deterministic traceless matrices $A_1, \dots, A_{k-1}$ we have the isotropic estimate

$$\big| \langle \boldsymbol{x}, e^{\mathrm{i}s_1 W} A_1 \cdots A_{k-1} e^{\mathrm{i}s_k W} \boldsymbol{y} \rangle \big| \lesssim \max_{i} \Big(\frac{1}{1+|s_i|} \Big)^{3/2} + N^{\xi} \frac{\max_i |s_i|^k}{N^{1/2}}, \tag{2.10}$$

both with very high probability.

In particular, for the unitary time evolution

$$A(t) := e^{\mathrm{i}tW} A e^{-\mathrm{i}tW} \tag{2.11}$$

and times $t_1, \dots, t_k \in \mathbf{R}$ with consecutive differences $s_i := t_i - t_{i-1}$, $t_0 := t_k$ we have the very high probability bound

$$|\langle A_1(t_1) \dots A_k(t_k) \rangle| \lesssim \max_{i \neq j} \Big(\frac{1}{(1+|s_i|)(1+|s_j|)} \Big)^{3/2} + N^{\xi} \frac{\max_i |s_i|^k}{N}, \tag{2.12}$$

as a consequence of (2.9) and similarly for the isotropic case.

In the first term of (2.12) we observe a thermalisation decay for k traceless observables A_i, each of them evolved some time t_i, as long as at least two consecutive time differences grow. If only some of the observables are traceless but the times t_i are *ordered* (equivalently, all but one s_i in Corollary 2.10 have the same sign), then we have a decay factor for each traceless observable, hence the decay rate is typically much faster. This is the content of the following corollary.

Corollary 2.11 *(Thermalisation decay). Let $k \geq 2$, and let $t_1 < \cdots < t_k$ be ordered times. Then*

(i) if $\mathfrak{a}$ *of the deterministic matrices* $A_1, \dots, A_k$ *are traceless, then we have the averaged bound*

$$|\langle A_1(t_1) \dots A_k(t_k) \rangle| \lesssim \max_i \Big(\frac{1}{1+|s_i|} \Big)^{3(1+\lceil \mathfrak{a}/2 \rceil)/2} + N^{\xi} \frac{\max_i |s_i|^k}{N}, \tag{2.13}$$

for $s_i = t_i - t_{i-1}$, $t_0 = t_k$,

(ii) and if $\mathfrak{a}$ *of the deterministic matrices* $A_1, \dots, A_{k-1}$ *are traceless, then we have the isotropic bound*

$$|\langle \boldsymbol{x}, A_1(t_1) \dots A_k(t_k) \boldsymbol{y} \rangle| \lesssim \max_i \Big(\frac{1}{1+|s_i|} \Big)^{3\lceil \mathfrak{a}/2 \rceil/2} + N^{\xi} \frac{\max_i |s_i|^k}{N^{1/2}}, \tag{2.14}$$

for $s_1 = t_1, s_k = t_k$ *and* $s_i = t_i - t_{i-1}$ *otherwise.*

The estimate (2.12) in particular implies that time evolutions $A(t_1), A(t_2)$ become *asymptotically free* as $|t_1 - t_2| \to \infty$. We recall that elements $a_1, \dots, a_p$ of some non-commutative probability space $(\mathcal{A}, \langle \cdot \rangle)$ are called *free* if for any $k \in \mathbf{N}$, any $i_1 \neq i_2 \neq \dots \neq i_k \in [p]$ and polynomials $p_1, \dots, p_k$ satisfying $\langle p_j(a_{i_j}) \rangle = 0$ it holds that

$$\langle p_1(a_{i_1}) \cdots p_k(a_{i_k}) \rangle = 0. \tag{2.15}$$

Corollary 2.12. *Let* $p \in \mathbf{N}$ *and let* $A_1, \dots, A_p$ *be sequences of deterministic matrices with* $\|A_i\| \lesssim 1$. *Consider* p *sequences of times* $\{t_1^N\}_{N \in \mathbf{N}}$, $\{t_2^N\}_{N \in \mathbf{N}} \dots$, $\{t_p^N\}_{N \in \mathbf{N}}$ *such that* $\min_{i \neq j} |t_i^N - t_j^N| \to \infty$ *as* $N \to \infty$, *then the unitary time evolutions* $A_1(t_1), \dots, A_p(t_p)$ *are asymptotically free. More precisely, for any* $k \in \mathbf{N}$ *and* $i_1 \neq \dots \neq i_k \in [p]$, *and any polynomials* $p_1, \dots, p_k$ *with* $\langle p_j(A_{i_j}) \rangle = 0$ *(which also implies* $\langle p_j(A_{i_j}(t_{i_j})) \rangle = 0$*) we have (with* $t_i = t_i^N$ *for brevity)*

$$\lim_{\substack{N \to \infty \\ \min_{i \neq j} |t_i - t_j| \to \infty}} \langle p_1(A_{i_1}(t_{i_1})) \cdots p_k(A_{i_k}(t_{i_k})) \rangle \tag{2.16}$$

with very high probability, as long as $\max_i |t_i| \lesssim N^{1/k - \epsilon}$ *for some* $\epsilon > 0$.

Corollary 2.12 can be easily extended to the asymptotic freeness of $A_1(t_1), \dots, A_p(t_p)$ and the algebra generated by arbitrary deterministic matrices $\{D_1, \dots, D_q\}$.

We stress that conjugation by the unitary time evolutions with respect to the very same Wigner matrix yields asymptotically free observables in (2.16). In the language of free-probability theory, a statement analogous to Corollary 2.12 is[3] that for self-adjoint $a_1, \dots, a_p$ and unitary elements $u_1, \dots, u_p$ in some non-commutative probability space

[3] We thank Roland Speicher for bringing this to our attention in private communication, and pointing out the reference [25, Proposición 3.21] for the simple proof of the assertion.

$(\mathcal{A}, \phi)$ the conjugated elements $u_1 a_1 u_1^*, \dots, u_p a_p u_p^*$ are free whenever $\{a_1, \dots, a_p\}$ is $*$-free from $\{u_1, \dots, u_p\}$ and $\phi(u_i u_j^*) = 0$ for $i \neq j$. Note that here freeness among the unitaries is not required, exactly as in Corollary 2.12 where the unitaries $e^{\mathrm{i}t_i W}, e^{\mathrm{i}t_j W}$ are not independent. In this context we also mention that for a set S and a Haar unitary u in a free probability space the conjugations of S with respect to different powers of the same unitary u, i.e. $S, uSu^*, u^2 S (u^*)^2$, are free, see [22, Lemma 3.7]. The estimate (2.16) shows that for large t the unitary random matrix $e^{\mathrm{i}tW}$ is close to a Haar unitary in this sense.

2.2. Extensions of Theorem 2.6

We close this section with a few extensions of Theorem 2.6. The first one is straightforward and we illustrate its proof immediately by an example, whilst the other two will be proven in Section 4.2.

Extension 2.13 *(Multiple independent Wigner matrices). Due to the high-probability sense of Theorem 2.6 we immediately obtain generalisations to multiple independent Wigner matrices both in averaged and isotropic sense by applying Theorem 2.6, to resolve each Wigner matrix iteratively while conditioning on all others. For example, let W, W' denote two independent Wigner matrices satisfying Assumption 2.1. Then, as an example, we obtain*

$$\begin{aligned}\langle f_1(W) A_1 f_2(W') A_2 f_3(W) A_3 \rangle &= \langle f_2 \rangle_{\mathrm{sc}} \langle f_1(W) A_1 A_2 f_3(W) A_3 \rangle + \mathcal{O}\Big(\frac{N^\xi}{N} \|f_2\|_{H^2} \Big) \\ &= \langle A_1 A_2 A_3 \rangle \langle f_1 \rangle_{\mathrm{sc}} \langle f_2 \rangle_{\mathrm{sc}} \langle f_3 \rangle_{\mathrm{sc}} + \mathcal{O}\Big(\frac{N^\xi}{N} \max_i \|f_i\|_{H^2} \Big) \\ &\quad + \langle A_1 A_2 \rangle \langle A_3 \rangle \langle f_2 \rangle_{\mathrm{sc}} \big(\langle f_1 f_3 \rangle_{\mathrm{sc}} - \langle f_1 \rangle_{\mathrm{sc}} \langle f_3 \rangle_{\mathrm{sc}} \big)\end{aligned}$$

with very high probability, where in the first step we used Theorem 2.6 for the random matrix W' after conditioning on W, while in the second step we used Theorem 2.6 again for W. This result should be compared with corresponding expression for $W = W'$ in Example 2.8 which implies

$$\begin{aligned}&\langle f_1(W) A_1 f_2(W) A_2 f_3(W) A_3 \rangle \\ &= \langle A_1 A_2 A_3 \rangle \langle f_1 \rangle_{\mathrm{sc}} \langle f_2 \rangle_{\mathrm{sc}} \langle f_2 \rangle_{\mathrm{sc}} + \langle A_1 A_2 \rangle \langle A_3 \rangle \langle f_2 \rangle_{\mathrm{sc}} \big(\langle f_1 f_3 \rangle_{\mathrm{sc}} - \langle f_1 \rangle_{\mathrm{sc}} \langle f_3 \rangle_{\mathrm{sc}} \big) \\ &\quad + \langle A_1 A_3 \rangle \langle A_2 \rangle \langle f_1 \rangle_{\mathrm{sc}} \big(\langle f_2 f_3 \rangle_{\mathrm{sc}} - \langle f_2 \rangle_{\mathrm{sc}} \langle f_3 \rangle_{\mathrm{sc}} \big) \\ &\quad + \langle A_2 A_3 \rangle \langle A_1 \rangle \langle f_3 \rangle_{\mathrm{sc}} \big(\langle f_1 f_2 \rangle_{\mathrm{sc}} - \langle f_1 \rangle_{\mathrm{sc}} \langle f_2 \rangle_{\mathrm{sc}} \big) \\ &\quad + \langle A_1 \rangle \langle A_2 \rangle \langle A_3 \rangle \big(\langle f_1 f_2 f_3 \rangle_{\mathrm{sc}} - \langle f_1 f_2 \rangle_{\mathrm{sc}} \langle f_3 \rangle_{\mathrm{sc}} - \langle f_1 f_3 \rangle_{\mathrm{sc}} \langle f_2 \rangle_{\mathrm{sc}} \\ &\qquad\qquad - \langle f_2 f_3 \rangle_{\mathrm{sc}} \langle f_1 \rangle_{\mathrm{sc}} + 2 \langle f_1 \rangle_{\mathrm{sc}} \langle f_2 \rangle_{\mathrm{sc}} \langle f_3 \rangle_{\mathrm{sc}} \big) + \mathcal{O}\Big(N^{\xi - 1} \max_i \|f_i\|_{H^2} \Big).\end{aligned}$$

Similar statements hold for arbitrary number of independent Wigner matrices $W_1, W_2, W_3, \dots$ *with possible repetitions.*

Extension 2.14 *(Mesoscopic version). For* $0 < a < 1$ *and mesoscopically rescaled* $f_i(x) = g_i(N^a(x-E))$ *with* $|E| \le 3$ *and* $g_i \in H_0^k$ *compactly supported, the result in Theorem 2.6 holds with the bound* $N^{\xi-1}\max_i \|g_i\|_{H^k}$ *replacing the rhs. in (2.5).*

Extension 2.15 *(Regularity of test functions). Theorem 2.6 gives an error bound practically of order* N^{-1} *under a relatively high regularity assumption on the functions* f_i*. Similar result holds for less regular functions with a weaker error bound. For example, if all* $f_i \in H^p([-3,3])$*, with some* $p \in [2,k]$*, then we obtain the averaged bound in (2.5) with an error term* $N^{\xi-(p-1)/(k-1)}\prod_i \|f_i\|_{H^p}$ *in the rhs., and the isotropic bound in (2.5) with an error term* $N^{\xi}[N^{-(p-1)/(2k-3)} \vee N^{-1/2}]\prod_i \|f_i\|_{H^p}$*.*

2.3. Möbius for non-crossing partitions

Finally, in terms of the Kreweras complement, we provide an explicit expression for the Möbius function $\mu(\pi, 1_S)$ defined recursively in (2.3). Lemma 2.16 is a standard result in the free probability literature but we present the proof for convenience.

Lemma 2.16. *For any finite* $S \subset \mathbf{N}$ *and any* $\pi \in \mathrm{NC}(S)$ *we have*

$$\mu(\pi, 1_S) = (-1)^{|\pi|-1} \prod_{B \in K(\pi)} C_{|B|-1},$$

where C_n *denotes the* n*-th Catalan number, i.e.* $(C_0, C_1, C_2, C_3, \dots) = (1, 1, 2, 5, 14, \dots)$.

Proof. As noticed in [34, Proposition 1] the interval $[\pi, 1_S]$ (with respect to the refinement partial order) is isomorphic to products of elementary intervals of the form $[0_B, 1_B]$ and therefore the general formula for the Möbius function follows directly from the special case $\mu(0_B, 1_B) = (-1)^{|B|-1}C_{|B|-1}$, as computed in [34, Corollary 5]. Following e.g. [32, Lemma 2.14], this idea can conveniently be presented by using the Kreweras complement, noting that K is an anti-automorphism in the sense that $\pi \le \sigma$ if and only if $K(\pi) \ge K(\sigma)$, and thus we have the isomorphism

$$[\pi, 1_S] \cong [K(1_S), K(\pi)] = [0_S, K(\pi)] \cong \prod_{B \in K(\pi)} [0_B, 1_B].$$

Consequently, since the Möbius function as defined in (2.3) is multiplicative, it follows that

$$\mu(\pi, 1_S) = \prod_{B \in K(\pi)} \mu(0_B, 1_B) = \prod_{B \in K(\pi)} (-1)^{|B|-1} C_{|B|-1} = (-1)^{|S|-|K(\pi)|} \prod_{B \in K(\pi)} C_{|B|-1}$$

and the claim follows from $|\pi| + |K(\pi)| = |S| + 1$. □

3. Multi resolvent local laws

Before stating the local laws for $G_1A_1G_2\cdots A_{k-1}G_k$, we introduce the commonly used definition of stochastic domination (see, e.g. [19]):

Definition 3.1 *(Stochastic Domination).* If

$$X = \left(X^{(N)}(u) \,\middle|\, N \in \mathbf{N}, u \in U^{(N)}\right) \quad \text{and} \quad Y = \left(Y^{(N)}(u) \,\middle|\, N \in \mathbf{N}, u \in U^{(N)}\right)$$

are families of non-negative random variables indexed by N, and possibly some parameter u, then we say that X is stochastically dominated by Y, if for all $\epsilon, D > 0$ we have

$$\sup_{u\in U^{(N)}} \mathbf{P}\left[X^{(N)}(u) > N^{\epsilon}Y^{(N)}(u)\right] \leq N^{-D}$$

for large enough $N \geq N_0(\epsilon, D)$. In this case we use the notation $X \prec Y$ or $X = \mathcal{O}_{\prec}(Y)$.

For $k = 1$ it is well known that as $N \to +\infty$ the resolvent G is well approximated by the unique solution $m = m(z)$ of the Dyson equation

$$-\frac{1}{m} = m + z, \quad \Im m \Im z > 0. \tag{3.1}$$

The optimal local law for a single G is well known [21,26,9] (see e.g. [10, Appendix A] to extend the local law to $\eta \geq N^{-100}$):

Theorem 3.2 *(Single G local laws). Let $z \in \mathbf{C} \setminus \mathbf{R}$ with $\eta := |\Im z| \geq N^{-100}$. Then for deterministic matrices and vectors $A, \boldsymbol{x}, \boldsymbol{y}$ with bounded norms $\|A\| + \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \lesssim 1$ we have*

$$\langle GA\rangle = m\langle A\rangle + \mathcal{O}_{\prec}\left(\frac{1}{N\eta}\right), \quad \langle \boldsymbol{x}, G\boldsymbol{y}\rangle = m\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \mathcal{O}_{\prec}\left(\sqrt{\frac{\rho}{N\eta}}\right)$$

with $\rho := \pi^{-1}|\Im m|$.

For $k \geq 2$ the resolvent identity $G(z_1)G(z_2) = [G(z_1) - G(z_2)]/[z_1 - z_2]$ suggests that *divided differences* of m provide the deterministic approximation to $G_1G_2\cdots G_k$, i.e. in the case when the deterministic matrices are $A_1 = A_2 = \cdots = A_{k-1} = I$.

Definition 3.3 *(Divided differences).* For finite multi-sets $\{z_1, \ldots, z_n\} \subset \mathbf{C} \setminus \mathbf{R}$ we recursively define

$$m[z_1, \ldots, z_n] := \frac{m[z_2, \ldots, z_n] - m[z_1, \ldots, z_{n-1}]}{z_n - z_1} \tag{3.2}$$

in case there are two distinct $z_1 \neq z_n$ among $z_1, \ldots, z_n$, and otherwise we set

$$m[\underbrace{z,\dots,z}_{n \text{ times}}] := \frac{m^{(n-1)}(z)}{(n-1)!}. \tag{3.3}$$

We note that this is well defined in the sense that $m[z_1,\dots,z_n]$ is independent of the ordering of the multi-set $\{z_1,\dots,z_n\}$.

The main technical result to prove Theorem 2.6 is the local law for alternating products of resolvents and deterministic matrices, which will be proven later in Section 5.2.

Theorem 3.4 *(Multi-resolvent local law). Let W be a Wigner matrix satisfying Assumption 2.1 with resolvent $G(z) = (W-z)^{-1}$. For $k \in \mathbf{N}$ let $z_1,\dots,z_k \in \mathbf{C}\setminus\mathbf{R}$ be such that $|\Re z_i| \le 3$ and $|\Im z_i| \ge N^{-1}$, and let $A_1,\dots,A_k, \boldsymbol{x}, \boldsymbol{y}$ be arbitrary deterministic matrices and vectors with $\|\boldsymbol{x}\|, \|\boldsymbol{y}\|, \|A_i\| \lesssim 1$, and define*

$$M_{[k]} := \sum_{\pi\in \mathrm{NC}([k])} \mathrm{pTr}_{K(\pi)}(A_1,\dots,A_{k-1}) \prod_{B\in\pi} m_\circ[B], \tag{3.4}$$

where $m_\circ$ is the free cumulant function from Definition 2.3 of the divided difference $m[i_1,\dots,i_n] := m[z_{i_1},\dots,z_{i_n}]$ from Definition 3.3. Then with

$$\eta_* := \min_i |\Im z_i|, \qquad \rho := \max_i \rho(z_i), \qquad \rho(z) := \frac{|\Im m(z)|}{\pi} \tag{3.5}$$

we have

$$\langle \boldsymbol{x}, G_1 A_1 G_2 \dots A_{k-1} G_k \boldsymbol{y}\rangle = \langle \boldsymbol{x}, M_{[k]}\boldsymbol{y}\rangle + \mathcal{O}_\prec\Big(\frac{1}{\eta_*^{k-1}}\sqrt{\frac{\rho}{N\eta_*}}\Big), \tag{3.6a}$$

$$\langle G_1 A_1 \dots A_{k-1} G_k A_k\rangle = \langle M_{[k]} A_k\rangle + \mathcal{O}_\prec\Big(\frac{1}{N\eta_*^k}\Big). \tag{3.6b}$$

The error estimates in (3.6a)–(3.6b) are optimal. In fact, elementary calculations show that e.g. for GUE W we have

$$\begin{aligned} \sqrt{\mathbf{E}|\langle G-m\rangle|^2} &= \frac{1}{N}\frac{\Im m}{\eta}\frac{1}{|1-m^2|}\Big(1+\mathcal{O}\Big(\frac{1}{N\eta}\Big)\Big) \sim \frac{1}{N\eta}, \\ \sqrt{\mathbf{E}|\langle \boldsymbol{x}, (G-m)\boldsymbol{y}\rangle|^2} &= \|\boldsymbol{x}\|\|\boldsymbol{y}\|\sqrt{\frac{\Im m|m|}{N\eta}}\Big(1+\mathcal{O}\Big(\frac{1}{N\eta}\Big)\Big) \sim \sqrt{\frac{\rho}{N\eta}}, \end{aligned} \tag{3.7}$$

demonstrating the optimality for $k=1$. For several G's we have by the resolvent identity $GG^* = \Im G/\Im z$, and thus the optimality for $k \ge 2$ follows from (3.7).

Remark 3.5. If the imaginary parts $\Im z_i$ vary in size, and the intermediate matrices are simply identity matrices, $A_i = I$, then the error bounds in (3.6) can be improved to

$$\begin{aligned}\langle \boldsymbol{x}, G_1 G_2 \cdots G_k \boldsymbol{y}\rangle &= \langle \boldsymbol{x}, \boldsymbol{y}\rangle m[z_1, \dots, z_k] + \mathcal{O}\left(\frac{1}{\prod_i |\Im z_i|}\sqrt{\frac{\min_i |\Im z_i|}{N}}\right), \\ \langle G_1 G_2 \cdots G_k\rangle &= m[z_1, \dots, z_k] + \mathcal{O}\left(\frac{1}{N\prod_i |\Im z_i|}\right),\end{aligned} \tag{3.8}$$

by using the improved bound from [12, Theorem 3.5] instead of (5.17)–(5.18) later. However, in the absence of deterministic matrices the local law (3.8) can alternatively also be derived from integrating the single-G local law, see [5, Lemma 3.9].

Example 3.6. We consider some examples.

(i) For $k = 2$ there are only two non-crossing partitions $\{12\}, \{1|2\}$ and thus

$$M_{[2]} = A_1 m_\circ[1] m_\circ[2] + \langle A_1\rangle m_\circ[1,2] = A_1 m[z_1] m[z_2] + \langle A_1\rangle (m[z_1, z_2] - m[z_1] m[z_2]).$$

(ii) For $k = 3$ all non-crossing partitions are given in Fig. 3 and thus we obtain

$$\begin{aligned}M_{[3]} = {} & A_1 A_2 m_\circ[1] m_\circ[2] m_\circ[3] + A_2 \langle A_1\rangle m_\circ[3] m_\circ[1,2] + \langle A_1 A_2\rangle m_\circ[2] m_\circ[1,3] \\ & + A_1 \langle A_2\rangle m_\circ[1] m_\circ[2,3] + \langle A_1\rangle\langle A_2\rangle m_\circ[1,2,3]\end{aligned}$$

with $m_\circ[i] = m[z_i]$, $m_\circ[i,j] = m[z_i, z_j] - m[z_i] m[z_j]$ and

$$\begin{aligned}m_\circ[i,j,k] = {} & m[z_i, z_j, z_k] - m[z_i] m[z_j, z_k] - m[z_j] m[z_i, z_k] - m[z_k] m[z_i, z_j] \\ & + 2 m[z_i] m[z_j] m[z_k]\end{aligned}$$

due to (2.3).

(iii) For $A_1 = \cdots = A_{k-1} = I$ we have $\mathrm{pTr}_\pi(A_1, \dots, A_{k-1}) = I$ for any π and thus

$$M_{[k]} = \sum_{\pi \in \mathrm{NC}([k])} \prod_{B \in \pi} m_\circ[B] = m[z_1, \dots, z_k]$$

due to (2.2).

The deterministic approximations $M_{[k]}$ satisfy the following bounds (which will be proven in Section 4.3).

Lemma 3.7. *For any $k \geq 1$ with η_*, ρ as in (3.5) we have the bound*

$$\|M_{[k]}\| \lesssim \frac{\rho}{\eta_*^{k-1}} \prod_{j=1}^{k-1} \|A_j\|, \tag{3.9}$$

and if $\mathfrak{a}$ out of the matrices $A_1, \dots, A_k$ are traceless, then we also have

$$|\langle M_{[k]} A_k \rangle| \lesssim \frac{\rho}{\eta_*^{k-1-\lceil \mathfrak{a}/2 \rceil}} \prod_{j=1}^{k} \|A_j\|. \tag{3.10}$$

In fact, for generic matrices A_i and $z_1 = \cdots = z_k$, with $\eta_* = \Im z_i$, these bounds are optimal, which shows that error terms in (3.6a)–(3.6b) are typically smaller than the deterministic leading terms whenever $N\eta_*\rho \gg 1$, i.e. in the regime where η is larger than the local eigenvalue spacing.

Remark 3.8. The bound (3.9) is consistent with the corresponding bound on resolvent chains

$$|\langle \boldsymbol{x}, G_1 A_1 \dots A_{k-1} G_k \boldsymbol{y} \rangle| \lesssim \frac{\sqrt{\langle \boldsymbol{x}, \Im G_1 \boldsymbol{x} \rangle \langle \boldsymbol{y}, \Im G_k \boldsymbol{y} \rangle}}{\eta_*^{k-1}} \leq \frac{\rho}{\eta_*^{k-1}}$$

obtained via Cauchy-Schwarz, the Ward identity

$$GG^* = \frac{\Im G}{\Im z}, \tag{3.11}$$

and the norm bound $\|G\| \leq 1/|\Im z|$. Similarly, (3.10) in the critical $\eta_* \sim N^{-1+\epsilon}$ regime is consistent with the recently established asymptotic orthogonality [11, Theorem 2.2] of the eigenvectors $\{\boldsymbol{u}_i\}$ of W with respect to traceless observables $\langle A \rangle = 0$ in the sense $|\langle \boldsymbol{u}_i, A \boldsymbol{u}_j \rangle| \prec N^{-1/2}$. Indeed, by spectral decomposition it follows that

$$|\langle G_1 A_1 \dots G_k A_k \rangle| \leq \frac{1}{N} \sum_{a_1 \cdots a_k} \frac{|\langle \boldsymbol{u}_{a_1}, A_1 \boldsymbol{u}_{a_2} \rangle| \cdots |\langle \boldsymbol{u}_{a_k}, A_k \boldsymbol{u}_{a_1} \rangle|}{|\lambda_{a_1} - z_1| \cdots |\lambda_{a_k} - z_k|} \prec N^{k-1-\mathfrak{a}/2}$$

in the case when $\mathfrak{a}$ of the matrices $A_1, \dots, A_k$ are traceless.

4. Proof of the main results

Using the Helffer-Sjöstrand representation, we express $f_1(W) A_1 \cdots f_k(W)$ as an integral of products of resolvents at different spectral parameters. Without loss of generality may assume (see Section 4.1 for details) that $f_i \in H_0^k([-3,3])$ and we consider such functions naturally extended to the entire real line by setting it zero outside of $[-3,3]$.

We now recall the Helffer-Sjöstrand representation. For any $f \in H_0^k([-3,3])$ we define its almost analytic extension by

$$f_{\mathbf{C}}(z) = f_{\mathbf{C},k}(z) = f_{\mathbf{C},k}(x + \mathrm{i}\eta) := \left[\sum_{j=0}^{k-1} \frac{(\mathrm{i}\eta)^j}{j!} f^{(j)}(x) \right] \chi(\eta), \tag{4.1}$$

with $\chi(\eta)$ a smooth cut-off equal to one on $[-5,5]$ and equal to zero on $[-10,10]^c$. Then by Helffer-Sjöstrand representation [16], we have

$$f(\lambda) = \frac{1}{\pi} \int_{\mathbf{C}} \frac{\partial_{\bar{z}} f_{\mathbf{C}}}{\lambda - z} \, \mathrm{d}^2 z, \tag{4.2}$$

where $\mathrm{d}^2 z = \mathrm{d}x \, \mathrm{d}\eta$ denotes the Lebesgue measure on $\mathbf{C} \equiv \mathbf{R}^2$ with $z = x + \mathrm{i}\eta$, and $\partial_{\bar{z}} := (\partial_x + \mathrm{i}\partial_\eta)/2$.

By (4.2) we get

$$f_1(W)A_1 \cdots f_k(W) = \frac{1}{\pi^k} \int_{\mathbf{C}^k} \prod_{i=1}^{k} \mathrm{d}^2 z_i \left[\prod_{i=1}^{k} (\partial_{\bar{z}} (f_i)_{\mathbf{C}})(z_i) \right] G(z_1) A_1 \cdots A_{k-1} G(z_k), \tag{4.3}$$

where $G(z_i) := (W - z_i)^{-1}$. In particular, using (4.3), we reduce the analysis of (2.5) to proving a *local law* for alternating chains of resolvents $G_i = G(z_i)$ and bounded deterministic matrices A_i, i.e. for

$$G_1 A_1 G_2 \cdots A_{k-1} G_k. \tag{4.4}$$

See Theorem 3.4 for the precise statement.

In Section 4.1 we prove our main result Theorem 2.6, then in Section 4.2 we will prove some of its corollaries and extensions. Finally, in Section 4.3 we prove some additional technical results used within the proof of Theorem 2.6.

4.1. Proof of Theorem 2.6

Set $H^k := H^k([-3,3])$. If $\max_i \|f_i\|_{H^k} \geq N^{1-\xi}$, with $\xi > 0$ from the statement of Theorem 2.6, then there is nothing to prove since the lhs. of (2.5) is clearly bounded by $\prod_i \|f_i\|_{L^\infty} \lesssim 1$. In the remainder of the proof we can thus assume that $\|f_i\|_{H^k} \leq N^{1-\xi}$, which implies $\|f^{(p)}\|_{L^2} \leq N^{1-\xi}$ for any $1 \leq p \leq k$. Additionally, without loss of generality we can assume that $f \in H_0^k([-3,3])$, indeed if this is not the case is then it is enough to consider $f_\chi(x) := f(x)\chi(x)$, with χ a smooth cut-off function which is equal to one on $[-5/2, 5/2]$ and equal to zero on $[-3,3]^c$. Then, by eigenvalue rigidity (see e.g. [19, Theorem 7.6] or [21]), $f_\chi(W) = f(W)$ with very high probability. Furthermore, we consider any function $f \in H_0^k([-3,3])$ to be extended to $\mathbf{R}$ by zero outside of $[-3,3]$. We first prove the average case in (2.5), and then we explain the very minor changes required in the isotropic case.

In the following computations we will often use the bound

$$\int_{\mathbf{R}} \mathrm{d}x |\partial_{\bar{z}} f_{\mathbf{C},k}(x + \mathrm{i}\eta)| \lesssim \eta^{k-1} \|f\|_{H^k} \tag{4.5}$$

from (4.1). We now prove that the regime $|\eta_i| \geq \eta_0$, for some $i \in [k]$ and with $\eta_0 := N^{-1+\xi/2}$, in the integral representation of $\langle f_1(W)A_1 \dots f_k(W)A_k \rangle$ from (4.3) is negligible. From now on we use the notation $f_{\mathbf{C}}(z) = f_{\mathbf{C},k}(z)$. We first consider the case

when only a single $|\eta_i| \le \eta_0$ and then we explain the minor changes in the case more then one η_i's are small. Without loss of generality we assume that $|\eta_1| \le \eta_0$; in this regime we will prove that

$$\begin{aligned}\left| \int \mathrm{d}x_1 \cdots \mathrm{d}x_k \int\limits_{\substack{|\eta_i|\ge\eta_0,\\ i\in[2,k]}} \mathrm{d}\eta_2 \cdots \mathrm{d}\eta_k \int\limits_{-\eta_0}^{\eta_0} \mathrm{d}\eta_1 \left(\prod_{i=1}^{k} (\partial_{\bar z} (f_i)_{\mathbf{C}})(z_i) \right) \langle G(z_1) A_1 \cdots G(z_k) A_k \rangle \right| \\ \prec \eta_0 \max_i \|f_i\|_{H^k}.\end{aligned} \tag{4.6}$$

To prove the bound (4.6) we will use Stokes theorem in the following form:

$$\int\limits_{-10}^{10} \int\limits_{\widetilde\eta}^{10} \partial_{\bar z} \psi(x + \mathrm{i}\eta) h(x + \mathrm{i}\eta) \, \mathrm{d}\eta \, \mathrm{d}x = \frac{1}{2\mathrm{i}} \int\limits_{-10}^{10} \psi(x + \mathrm{i}\widetilde\eta) h(x + \mathrm{i}\widetilde\eta) \, \mathrm{d}x, \tag{4.7}$$

for any $\widetilde\eta \in [0, 10]$, and for any $\psi, h \in H^1(\mathbf{C}) \equiv H^1(\mathbf{R}^2)$ such that $\partial_{\bar z} h = 0$ on the domain of integration and for ψ vanishing at the left, right and top boundary of the domain of integration. Note that by (4.7) we readily conclude that

$$\begin{aligned}&\int\limits_{\mathbf{R}} \mathrm{d}x_i \int\limits_{\eta_0}^{10} \mathrm{d}\eta_i (\partial_{\bar z} (f_i)_{\mathbf{C}})(z_i) \langle G(z_1) A_1 \dots A_{i-1} G(z_i) A_i \dots G(z_k) A_k \rangle \\ &\quad = \int\limits_{\mathbf{R}} \mathrm{d}x_i (f_i)_{\mathbf{C}}(x_i + \mathrm{i}\eta_0) \langle G(z_1) A_1 \dots A_{i-1} G(x_i + \mathrm{i}\eta_0) A_i \dots G(z_k) A_k \rangle,\end{aligned} \tag{4.8}$$

for any fixed $z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_k$. Then, using (4.8) repeatedly for the $z_2, \dots, z_k$-variables, and defining $\eta_r := N^{-100}$, we conclude

$$\begin{aligned}&|\text{lhs. of (4.6)}| \\ &= \left| \int \prod_{i=1}^{k} \mathrm{d}x_i \int\limits_{-\eta_0}^{\eta_0} \mathrm{d}\eta_1 (\partial_{\bar z} (f_1)_{\mathbf{C}})(x_1 + \mathrm{i}\eta_1) \prod_{i=2}^{k} (f_i)_{\mathbf{C}}(x_i + \mathrm{i}\eta_0) \right. \\ &\qquad\qquad \left. \times \langle G(z_1) A_1 G(x_2 + \mathrm{i}\eta_0) \cdots G(x_k + \mathrm{i}\eta_0) A_k \rangle | \right| \\ &\prec \|f_1\|_{H^k} \left(\int\limits_{\eta_r \le |\eta_1| \le \eta_0} \left(\eta_1^{k-3/2} + \frac{\eta_1^{k-5/2}}{N} \right) \frac{1}{\eta_0^{k-3/2}} \, \mathrm{d}\eta_1 + \int\limits_{|\eta_1| \le \eta_r} \eta_1^{k-2} \eta_0^{-k+1} \, \mathrm{d}\eta_1 \right) \\ &\lesssim \eta_0 \|f_1\|_{H^k}.\end{aligned} \tag{4.9}$$

In the first inequality in the regime $\eta_r = N^{-100} \leq |\eta_1| \leq \eta_0$ we used (4.5) for $\int \mathrm{d}x_1 |\partial_{\overline{z}}(f_1)_{\mathbf{C}}|$, and the bound

$$\begin{aligned}
&\langle G(z_1)A_1G(x_2+\mathrm{i}\eta_0)\cdots G(x_k+\mathrm{i}\eta_0)A_k\rangle \\
&\leq \langle G(z_1)A_1A_1^*G(z_1)^*\rangle^{1/2} \\
&\qquad \times \langle G(x_2+\mathrm{i}\eta_0)A_2\cdots G(x_k+\mathrm{i}\eta_0)A_kA_k^*G(x_k+\mathrm{i}\eta_0)^*\cdots A_2^*G(x_2+\mathrm{i}\eta_0)^*\rangle^{1/2} \\
&\prec \frac{1}{\sqrt{\eta_1\eta_0}\eta_0^{k-2}}\langle \Im G(z_1)\rangle^{1/2}\langle \Im G(x_2+\mathrm{i}\eta_0)\rangle^{1/2} \\
&\prec \frac{1}{\sqrt{\eta_1\eta_0}\eta_0^{k-2}}\left(1+\frac{1}{N\eta_1}\right).
\end{aligned} \tag{4.10}$$

Here, to go from first to the second line we used a Schwarz inequality, to go from the second to the third line we used the norm bounds $\|A_1A_1^*\| \lesssim 1$ and

$$\|A_2\cdots G(x_k+\mathrm{i}\eta_0)A_kA_k^*G(x_k+\mathrm{i}\eta_0)^*\cdots A_2^*\| \lesssim \eta_0^{-2k+4}$$

and Ward identity (3.11). Finally, to go to the last line we used the averaged local law from Theorem 3.2 to show the boundedness of $\langle \Im G\rangle$. In the complementary regime $|\eta_1| \leq N^{-100}$ we used the norm bound

$$\left|\langle G(z_1)A_1G(x_2+\mathrm{i}\eta_0)A_2\cdots G(x_k+\mathrm{i}\eta_0)A_k\rangle\right| \leq \|G(z_1)A_1\|\prod_{i=2}^{k}\|G(x_i+\mathrm{i}\eta_0)A_i\| \lesssim \frac{1}{\eta_1\eta_0^{k-1}},$$

together with the estimate (4.5) for $\partial_{\overline{z}}(f_1)_{\mathbf{C}}$. Note that in the penultimate inequality of (4.9) we also used that

$$\|(f_i)_{\mathbf{C}}(\cdot+\mathrm{i}\eta_0)\|_{L^1} \lesssim \|(f_i)_{\mathbf{C}}\|_{L^2} \lesssim \sum_{j=0}^{k-1}\frac{\eta_0^j}{j!}\|f_i^{(j)}\|_{L_2} \lesssim 1,$$

by (4.1) and $\|f_i\|_{H^k} \leq N^{1-\xi}$.

The bound for the regime when more than one η_i are smaller than η_0 (in absolute value) is completely analogous giving an even smaller bound. In particular, if $|\eta_1|,\ldots,|\eta_l| \leq \eta_0$, with $l \in [k]$, then we perform an integration by parts in the $z_{l+1},\ldots,z_k$-variables and use the bound (4.5) for $\partial_{\overline{z}}(f_i)_{\mathbf{C}}$ for $i \in [l]$ (see (4.9) below for the case $l=1$), which gives a bound

$$\eta_0^{(l-1)k+1}\left(\prod_{i=1}^{l}\|f_i\|_{H^k}\right) \leq \eta_0^{(l-1)k+1}N^{l-1}\max_i\|f_i\|_{H^k}$$

by $\|f_i\|_{H^k} \leq N^{1-\xi}$. This concludes the proof of the fact that the small η_i regime is negligible. We now estimate the regime when $|\eta_i| \geq \eta_0$ for any $i \in [k]$.

By (4.3) and the local law (3.6b), we conclude that

$$\begin{aligned}\langle f_1(W)A_1\cdots f_k(W)A_k\rangle &= \frac{1}{\pi^k}\int_{\mathbf{R}^k}\int_{\eta_0\le|\eta_i|\le 10} \mathrm{d}^2 z_1\cdots \mathrm{d}^2 z_k(\partial_{\bar z}(f_1)_{\mathbf{C}})(z_1)\cdots(\partial_{\bar z}(f_k)_{\mathbf{C}})(z_k)\langle M_{[k]}A_k\rangle \\ &\quad + \mathcal{O}_{\prec}\left(\eta_0 \max_i \|f_i\|_{H^k}\right).\end{aligned} \tag{4.11}$$

Note that in (4.11), proceeding as in (4.9), we estimated the error term coming from the local law (3.6b) by

$$\begin{aligned}\frac{1}{\pi^k}\int_{\mathbf{R}^k}\int_{\eta_0\le|\eta_i|\le 10} \mathrm{d}^2\boldsymbol{z}\prod_{i=1}^k(\partial_{\bar z}(f_i)_{\mathbf{C}})(z_i)\langle(G(z_1)A_1\ldots G(z_k)-M_{[k]})A_k\rangle \\ = \mathcal{O}_{\prec}\left(N^{-1}\max_i\|f_i\|_{H^k}\right),\end{aligned} \tag{4.12}$$

with $\mathrm{d}^2\boldsymbol{z} := \mathrm{d}^2 z_1\ldots\mathrm{d}^2 z_k$. More precisely, in (4.12) we considered the regime $\eta_1\le\eta_2\le\cdots\le\eta_k$ (all the other regimes give the same contribution by symmetry) and performed $k-1$ integration by parts in the z_i-variables, $i\in[2,k]$, as in (4.8), and then estimated the remaining $\partial_{\bar z}(f_1)_{\mathbf{C}}(z_1)$ by (4.5). Note that the factors $N^{\xi/2}$ from $\eta_0 = N^{-1+\xi/2}$, and $|\log\eta_0|$ from the integration of the error term in the local law were all included in the $\mathcal{O}_{\prec}(\cdot)$ notation.

By the formula for $M_{[k]}$ in (3.4) and the definition of $m_\circ$ from Definition 2.3, to compute the rhs. of (4.11) it is enough to compute the integral of $m[z_1,\ldots,z_p]$ for any $p\in\mathbf{N}$. This technical lemma will be proven at the end of this section.

Lemma 4.1. *For any $p\in\mathbf{N}$ denote $\boldsymbol{z} := (z_1,\ldots,z_p)\in\mathbf{C}^p$, then it holds*

$$\frac{1}{\pi^p}\int_{\mathbf{R}^p}\int_{\eta_r\le|\eta_i|\le 10}\mathrm{d}^2\boldsymbol{z}\prod_{i=1}^p(\partial_{\bar z}(f_i)_{\mathbf{C}})(z_i)m[z_1,\ldots,z_p] = \left\langle\prod_{i=1}^p f_i\right\rangle_{\mathrm{sc}} + \mathcal{O}(\eta_r), \tag{4.13}$$

where $\eta_r := N^{-100}$.

Finally, using that by (4.6) and Lemma 3.7 the regime $\eta_i\in[\eta_r,\eta_0]$ can be added back to (4.11) at the price of an error $\eta_0\max_i\|f_i\|_{H^k}$, and using (4.13) repeatedly together with the definition of $\mathrm{sc}_\circ$ given in the statement of Theorem 2.6, we conclude the proof of the average case in (2.5), modulo the proof of Lemma 4.1.

The proof of the isotropic case in (2.5) is very similar. The only differences are the following: (i) to bound the small η_i-regime we have to replace (4.10) by

$$\left|\langle\boldsymbol{x}, G(z_1)A_1G(x_2+\mathrm{i}\eta_0)\cdots G(x_k+\mathrm{i}\eta_0)\boldsymbol{y}\rangle\right| \prec \frac{1}{\sqrt{\eta_1\eta_0}\eta_0^{k-2}}\left(1+\frac{1}{\sqrt{N\eta_1}}\right),$$

which still gives exactly the same bound (4.6); (ii) to estimate the error term coming from the isotropic local law (3.6a) (used in the regime when $|\eta_i| \geq \eta_0$ for all $i \in [k]$) we have to replace (4.12) by

$$\frac{1}{\pi^k} \int_{\mathbf{R}^k} \int_{\eta_0 \leq |\eta_i| \leq 10} \mathrm{d}^2 \boldsymbol{z} \prod_{i=1}^{k} (\partial_{\bar{z}} (f_i)_{\mathbf{C}})(z_i) \langle \boldsymbol{x}, (G(z_1) A_1 \dots G(z_k) - M_{[k]}) \boldsymbol{y} \rangle = \mathcal{O}_{\prec} \left(N^{-1/2} \max_i \|f_i\|_{H^k} \right). \tag{4.14}$$

The proof of (4.14) is exactly the same as the proof of (4.12). □

4.2. Proof of the corollaries and extensions of Theorem 2.6

Proof of Corollary 2.9. The leading term (2.8) is given exactly by the leading term in (2.5) choosing $f_i(x) = e^{\mathrm{i} s_i x}$, and using the definition of $\varphi(x)$ in (2.6). The bound for the error term readily follows from $\|f_i\|_{H^k} \lesssim |s_i|^k$ and $\|f_i\|_{L^\infty} \lesssim 1$. □

Proof of Corollary 2.10. For the proof of (2.9) we note that in (2.8) only $\pi \in \mathrm{NC}[k]$ contribute for which all blocks of $K(\pi)$ contain at least two elements. However, this can only be the case if π contains at least two singleton blocks and the claim follows since $\varphi_\circ[\{i\}] \lesssim (1 + |s_i|)^{-3/2}$. □

Proof of Corollary 2.11. For the proof of (2.13) we note that due to the ordering of times it follows that

$$\varphi_\circ[B] = \begin{cases} \mathcal{O}((1 + \min_i |s_i|)^{-3/2}), & B \subsetneq [k], \\ 1, & B = [k]. \end{cases} \tag{4.15}$$

For a traceless A_i's only partitions π for which $K(\pi)$ has at most $\lfloor \mathfrak{a}/2 \rfloor$ blocks contribute, i.e. only partitions π with at least $\lceil \mathfrak{a}/2 \rceil + 1$ blocks and also (2.13) follows. For the proof of (2.14) we similarly note that only partitions contribute for which $K(\pi)$ as at most $\lfloor \mathfrak{a}/2 \rfloor + 1$ blocks (since the block containing k has no trace restriction). □

Proof of Corollary 2.12. By linearity it is clearly sufficient to check that for $k \geq 2$, $i_1 \neq i_2 \neq \ldots \neq i_k \in [p]$ and any $a_i \in \mathbf{N}$ we have

$$\left\langle \left(A_{i_1} (t_{i_1})^{a_1} - \langle A_{i_1} (t_{i_1})^{a_1} \rangle \right) \cdots \left(A_{i_k} (t_{i_k})^{a_k} - \langle A_{i_k} (t_{i_k})^{a_k} \rangle \right) \right\rangle = o(1). \tag{4.16}$$

This is indeed the case for $1 \ll |t_1 - t_2| \ll N^{1/k - \epsilon}$, since the lhs. of (4.16) simplifies to

$$\left\langle (A_{i_1}^{a_1})^\circ (t_{i_1}) \cdots (A_{i_k}^{a_k})^\circ (t_{i_k}) \right\rangle = \mathcal{O} \left(N^\xi \frac{|t_1 - t_2|^k}{N} + (1 + |t_1 - t_2|)^{-3} \right),$$

where $A^\circ := A - \langle A \rangle$ denotes the traceless part of A, and we used (2.12). □

Proof of the bound in Extension 2.14. The proof of this bound is completely analogous to the proof of (2.5), the only difference is that instead of (4.1) we consider the almost analytic extension

$$f_{\mathbf{C},k}(z) = f_{\mathbf{C},k}(x + \mathrm{i}\eta) := \left[\sum_{j=0}^{k-1} \frac{(\mathrm{i}\eta)^j}{j!} f^{(j)}(x) \right] \chi(N^a \eta),$$

for each $f = f_i$, where $f_i = g_i(N^a(x - E))$ and $a \in (0,1)$, $|E| < 2$. □

Proof of the bound in Extension 2.15. The proof of this bound is similar to the proof of Theorem 2.6, the only difference is that the regime $|\eta_i| \le \eta_0$ and the error term in the local law are estimated differently. Similarly to the proof of Theorem 2.6, we prove the bound in Extension 2.15 in the average case and then we explain the very minor changes in the isotropic case. Without loss of generality we assume that $f_i \in H_0^p([-3,3])$. We first show how to bound the small η_i-regimes; here we again only consider the case when only $|\eta_1| \le \eta_0$. Recall that for H^p functions we have

$$\int_{\mathbf{R}} \mathrm{d}x_i |\partial_{\bar{z}} (f_i)_{\mathbf{C}}| \lesssim \eta_i^{p-1} \|f_i\|_{H^p} \tag{4.17}$$

by (4.5), where we used the short-hand notation $f_{\mathbf{C}}(z) = f_{\mathbf{C},p}(z)$. Then, using this bound, we get

$$\left| \prod_{i=1}^k \int \mathrm{d}x_i \prod_{i=2}^k \int_{|\eta_i| \ge \eta_0} \mathrm{d}\eta_i \int_{-\eta_0}^{\eta_0} \mathrm{d}\eta_1 \prod_{i=1}^k (\partial_{\bar{z}} (f_i)_{\mathbf{C}})(z_i) \langle G(z_1) A_1 \cdots G(z_k) A_k \rangle \right| \prec \eta_0^{p-1} \prod_i \|f_i\|_{H^p}, \tag{4.18}$$

for some $N^{-1} \ll \eta_0 \ll 1$ that we will choose shortly. In the estimate (4.18) we also used the norm bound $|\langle G(z_1) A_1 \cdots G(z_k) A_k \rangle| \lesssim \prod_i |\eta_i|^{-1}$. Similarly to (4.12), using the bound (4.17) and the average local law (3.6b), we conclude that

$$\begin{aligned} &\left| \prod_{i=1}^k \int \mathrm{d}x_i \prod_{i=1}^k \int_{|\eta_i| \ge \eta_0} \mathrm{d}\eta_i \prod_{i=1}^k (\partial_{\bar{z}} (f_i)_{\mathbf{C}})(z_i) \langle (G(z_1) A_1 \dots G(z_k) - M_{[k]}) A_k \rangle \right| \\ &\qquad \prec \frac{1}{N \eta_0^{k-p}} \prod_i \|f_i\|_{H^p} \end{aligned} \tag{4.19}$$

Optimising the bounds in (4.18) and (4.19) we find that $\eta_0 = N^{-1/(k-1)}$, concluding the proof of the bound stated in Extension 2.15 in the average case. The proof in the

isotropic case is exactly the same, the only difference is that (4.19) has to be replaced by

$$\left|\prod_{i=1}^{k}\int \mathrm{d}x_i \prod_{i=1}^{k}\int_{|\eta_i|\geq\eta_0} \mathrm{d}\eta_i \prod_{i=1}^{k}(\partial_{\bar{z}}(f_i)_{\mathbf{C}})(z_i)\langle \boldsymbol{x}, (G(z_1)A_1\ldots G(z_k)-M_{[k]})\boldsymbol{y}\rangle\right| \tag{4.20}$$
$$\prec \frac{1}{\sqrt{N}}\left(\frac{1}{\eta_0^{k-p-1/2}}\vee 1\right)\prod_i \|f_i\|_{H^p},$$

where we used the isotropic local law (3.6a). Optimising the error terms in (4.18) and (4.20) we conclude that in the isotropic case $\eta_0 = N^{-1/(2k-3)}$. □

4.3. Proof of additional results used within the proof of Theorem 2.6

Proof Lemma 4.1. We claim that

$$m[z_1,\ldots,z_k] = \int_{\mathbf{R}} \rho(x)\prod_{i=1}^{p}\frac{1}{(x-z_i)}\,\mathrm{d}x. \tag{4.21}$$

The proof of (4.21) follows by induction. For $p=1$ (4.21) is trivial, for $p=2$, we have

$$m[z_1,z_2] = \frac{m[z_1]-m[z_2]}{z_1-z_2} = \int \frac{\rho(x)}{z_1-z_2}\left[\frac{1}{x-z_1}-\frac{1}{x-z_2}\right]\mathrm{d}x = \int \frac{\rho(x)}{(x-z_1)(x-z_2)}\,\mathrm{d}x.$$

Now assume that (4.21) holds for p, then it holds for $p+1$ as well:

$$\begin{aligned} m[z_1,z_2,z_3,\ldots,z_{p+1}] &= \frac{m[z_1,z_3,\ldots,z_{p+1}]-m[z_2,z_3,\ldots,z_{p+1}]}{z_1-z_2} \\ &= \int \frac{\rho(x)}{z_1-z_2}\left[\frac{1}{x-z_1}-\frac{1}{x-z_2}\right]\prod_{i=3}^{p}\frac{1}{(x-z_i)}\,\mathrm{d}x \\ &= \int \rho(x)\prod_{i=1}^{p+1}\frac{1}{(x-z_i)}\,\mathrm{d}x, \end{aligned}$$

concluding the proof of (4.21). Finally, using (4.21) we readily conclude (4.13), where we used that for any fixed $x\in\mathbf{R}$ we have

$$\begin{aligned} \int_{\mathbf{R}}\mathrm{d}x_i \int_{\eta_r\leq|\eta_i|\leq 10}\mathrm{d}\eta_i(\partial_{\bar{z}}(f_i)_{\mathbf{C}}(z_i))\frac{1}{x-z} &= \int_{\mathbf{R}}\mathrm{d}x_i(f_i)_{\mathbf{C}}(x_i+\mathrm{i}\eta_r)\frac{\eta_r}{(x-x_i)^2+\eta_r^2} \\ &= \pi f(x)+\mathcal{O}(\eta_r). \end{aligned}$$

We remark that in the first equality we used (4.7). □

Proof of Lemma 3.7. We first claim that for $k \geq 2$

$$|m[z_1, \dots, z_k]| \leq \eta_*^{1-k} \max_i |\Im m(z_i)| \tag{4.22}$$

for any $z_1, \dots, z_k$. The bound (4.22) follows immediately from (4.21) and estimating

$$\begin{aligned} |m[z_1, \dots, z_k]| &\leq \int \rho_{\mathrm{sc}}(x) \prod_{i=1}^{k} \frac{1}{|x - z_i|} \, \mathrm{d}x \\ &\leq \frac{\eta_*^{2-k}}{2} \int \rho_{\mathrm{sc}}(x) \Big(\frac{1}{|x - z_1|^2} + \frac{1}{|x - z_2|^2} \Big) \, \mathrm{d}x \\ &= \frac{\eta_*^{2-k}}{2} \Big(\frac{|\Im m(z_1)|}{|\Im z_1|} + \frac{|\Im m(z_2)|}{|\Im z_2|} \Big) \leq \eta_*^{1-k} \max_i |\Im m(z_i)|. \end{aligned}$$

From (4.22) it follows that for any $B \subset [k]$ with $|B| \geq 2$ we have $|m_\circ[B]| \lesssim \rho \eta_*^{1-|B|}$ due to (2.3) where single-block partition of B yields the worst bound, and thus

$$\Big| \prod_{B \in \pi} m_\circ[B] \Big| \lesssim 1 + \rho \eta_*^{|\pi|-k} \tag{4.23}$$

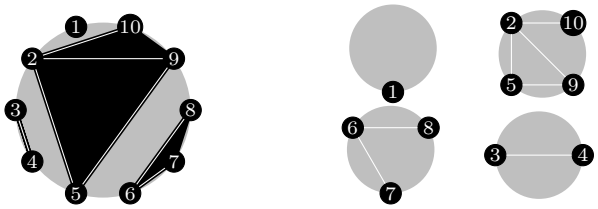
for any $\pi \in \mathrm{NC}[k]$, concluding the proof of (3.9) using definition (3.4). Finally, for the proof of (3.10) note that after taking the trace $\langle M_{[k]} A_k \rangle$ with $M_{[k]}$ as in (3.4) for $\mathfrak{a}$ traceless A_i only those $\pi \in \mathrm{NC}[k]$ give a non-zero contribution for which $K(\pi)$ has at most $|K(\pi)| \leq k - \lceil \mathfrak{a}/2 \rceil$ blocks. Equivalently, π necessarily has at least $|\pi| \geq \lceil \mathfrak{a}/2 \rceil + 1$ blocks, concluding also the proof of (3.10) using (4.23). □

5. Proof of the local law

5.1. Alternative representations of $m, m_\circ, M$

To prepare the proof of Theorem 3.4 we first provide explicit alternative representations of the divided differences $m[\cdot]$ and their free-cumulant version $m_\circ[\cdot]$ based upon non-crossing graphs instead of non-crossing partitions. We begin with the definition of non-crossing graphs.

Definition 5.1. Let $S \subset \mathbf{N}$ be a finite set of integers arranged on a circle as in Definition 2.4. We call an undirected graph (S, E) *crossing* if there exist two edges $(ab), (cd) \in E$ with $a < c < b < d$, otherwise we call it *non-crossing* and we denote the set of *non-crossing graphs* by $\mathrm{NCG}(S)$. We call a graph (S, E) a *dissection graph* if for $S = \{s_1 < \cdots < s_k\}$ we have $(s_1 s_2), (s_2 s_3), \dots, (s_k s_1) \notin E$, i.e. if all edges dissect the polygon spanned by $s_1, \dots, s_k$, and denote the set of dissection graphs by $\mathrm{NCG_d}(S)$. Finally, we denote the set of connected graphs by $\mathrm{NCG_c}(S)$.



(a) Full graph with induced partition (b) Connected subgraphs

Fig. 4. Decomposition of the graph $\{(25), (59), (29), (2\ 10), (34), (68), (67)\}$ according to partition $\{1|2, 5, 9, 10|3, 4|6, 7, 8\}$ of its connected components.

Each non-crossing graph $(S, E) \in \mathrm{NCG}(S)$ trivially induces a non-crossing partition $\pi \in \mathrm{NC}(S)$ with blocks representing the vertices in the connected components of (S, E) and thus we can represent

$$\mathrm{NCG}(S) = \bigsqcup_{\pi \in \mathrm{NC}(S)} \prod_{B \in \pi} \mathrm{NCG}_{\mathrm{c}}(B), \tag{5.1}$$

see Fig. 4 for an example.

Lemma 5.2. *Let $k \in \mathbf{N}$ and let $z_1, \dots, z_k \in \mathbf{C}$. For $S \subset [k]$ and the divided difference $m[S] := m[\{z_s \mid s \in S\}]$ we have*

$$m[S] = \left(\prod_{s \in S} m_s\right) \sum_{E \in \mathrm{NCG}(S)} q_E, \qquad q_E := \prod_{ab \in E} q_{ab}, \quad q_{ab} := \frac{m_a m_b}{1 - m_a m_b} \tag{5.2}$$

with $m_a := m(z_a)$. Moreover, for the free-cumulant function of m we have

$$m_\circ[S] = \left(\prod_{s \in S} m_s\right) \sum_{E \in \mathrm{NCG}_{\mathrm{c}}(S)} q_E. \tag{5.3}$$

We stress that in contrast to the formulas (2.2)–(2.3) for $f[S]$ and $f_\circ[S]$ valid for any function f, the representations (5.2)–(5.3) of $m[S]$ and $m_\circ[S]$ in terms of q hold only for the specific function m, the solution to equation (3.1).

Proof of Lemma 5.2. We first prove (5.2) by induction on $|S|$ with $|S| = 1$ being trivial. The $|S| = 2$ case

$$m[z_i, z_j] = \frac{m_i - m_j}{z_i - z_j} = \frac{m_i - m_j}{m_j + 1/m_j - m_i - 1/m_i} = \frac{m_i m_j}{1 - m_i m_j} = q_{ij} = m_i m_j (1 + q_{ij}) \tag{5.4}$$

follows directly from (3.1).

For the induction step we may consider, without loss of generality, the set $S = [n]$ and spectral parameters $z_1 \neq z_n$. The general case follows by relabelling the spectral parameters and continuity in case of equal spectral parameter (for a direct argument

in the case of equal spectral parameters see Remark 5.3 below). From the induction hypothesis we may assume that (5.2) has been established for $S' = (1,n],[1,n)$. Using the non-crossing property we partition $\mathrm{NCG}[1,n]$ into

$$\mathrm{NCG}[1,n] = \mathrm{NCG}^{(1n)}[1,n] \sqcup \mathrm{NCG}^{\neg(1n)}[1,n], \tag{5.5}$$

i.e. the subsets of graphs containing or not containing the edge $(1n)$, with

$$\mathrm{NCG}^{\neg(ab)}[a,b] := \{E \in \mathrm{NCG}[a,b] | (ab) \notin E\}, \quad \mathrm{NCG}^{(ab)}[a,b] := \mathrm{NCG}^{\neg(ab)}[a,b] \times \{\{(ab)\}\}$$

and then further partition $\mathrm{NCG}^{\neg(1n)}[n]$ according to the maximal vertex connected to 1, i.e.

$$\mathrm{NCG}^{\neg(1n)}[1,n] = \mathrm{NCG}(1,n] \sqcup \bigsqcup_{j=2}^{n-1} (\mathrm{NCG}^{(1j)}[j] \times \mathrm{NCG}[j,n]). \tag{5.6}$$

We then obtain

$$\begin{aligned} \sum_{E \in \mathrm{NCG}[1,n]} \frac{q_E}{1+q_{1n}} &= \sum_{E \in \mathrm{NCG}^{\neg(1n)}[1,n]} q_E \\ &= \sum_{E \in \mathrm{NCG}(1,n]} q_E + \sum_{j=2}^{n-1} m_1 m_j \Big(\sum_{E \in \mathrm{NCG}[1,j]} q_E \Big) \Big(\sum_{E \in \mathrm{NCG}[j,n]} q_E \Big) \\ &= \sum_{E \in \mathrm{NCG}[1,n)} q_E + \sum_{j=2}^{n-1} m_j m_n \Big(\sum_{E \in \mathrm{NCG}[1,j]} q_E \Big) \Big(\sum_{E \in \mathrm{NCG}[j,n]} q_E \Big), \end{aligned} \tag{5.7}$$

where the first equality in (5.7) follows from (5.5), and the second equality then follows from (5.5)–(5.6) together with $(1+q_{1j})m_1 m_j = q_{1j}$. The argument for the third equality is completely symmetric. Then from (5.4) and the induction hypothesis we have

$$\begin{aligned} \frac{m(1,n] - m[1,n)}{z_n - z_1} &= \frac{q_{1n}}{m_n - m_1} \Big(m(1,n] - m[1,n) \Big) \\ &= \Big(\prod_{i=1}^{n} m_i \Big) \frac{q_{1n}}{m_n - m_1} \Big(\frac{1}{m_1} \sum_{E \in \mathrm{NCG}(1,n]} q_E - \frac{1}{m_n} \sum_{E \in \mathrm{NCG}[1,n)} q_E \Big). \end{aligned} \tag{5.8}$$

By solving (5.7) for $m_1^{-1} \sum_{E \in \mathrm{NCG}(1,n]} q_E$ and $m_n^{-1} \sum_{E \in \mathrm{NCG}[1,n)} q_E$ and noting that for the difference of the two used in (5.8) the $\sum_j$ terms cancel, we obtain

$$\begin{aligned} &\frac{m(1,n] - m[1,n)}{z_n - z_1} \\ &= \Big(\prod_{i=1}^{n} m_i \Big) \frac{q_{1n}}{m_n - m_1} \frac{1}{1+q_{1n}} \Big(\frac{1}{m_1} - \frac{1}{m_n} \Big) \sum_{E \in \mathrm{NCG}[1,n]} q_E = \Big(\prod_{i=1}^{n} m_i \Big) \sum_{E \in \mathrm{NCG}[1,n]} q_E, \end{aligned}$$

completing the induction step.

We now turn to (5.3) and reorganise (5.2) in terms of the vertex sets of the connected components as in (5.1) and obtain

$$m[S] = \sum_{\pi \in \mathrm{NC}(S)} \prod_{B \in \pi} \Big(\Big(\prod_{s \in B} m_s \Big) \sum_{E \in \mathrm{NCG_c}(B)} q_E \Big),$$

so that (5.3) follows immediately from Definition 2.3 and the uniqueness of $m_\circ$, cf. (2.3). □

Remark 5.3 *(Alternative proof of Lemma 5.2 for equal spectral parameters).* In the case where all spectral parameters are equal we consider $S = [n]$ with $z_1 = \cdots = z_n = z$. It is well known that the generating functions $a_n(w), b_n(w)$ of the graphs in $\mathrm{NCG}([n]), \mathrm{NCG_d}([n])$ with j edges satisfy [17, Eq. (9)]

$$a_n(w) = \begin{cases} 1, & n = 1, \\ 1 + w, & n = 2, \\ (1+w)^n b_n(w), & n > 2. \end{cases} \qquad (5.9)$$

$$b_{n+1}(w) = \begin{cases} 1, & n = 2, \\ (1+2w) b_n(w) + \frac{2w(1+w)}{n} b_n'(w), & n \geq 3, \end{cases}$$

and therefore (5.2) is equivalent to

$$\frac{m^{(n-1)}}{(n-1)!} = m^n a_n(q), \qquad (5.10)$$

which is obvious for $n = 1$. For $n = 2$ the identity (5.10) follows from differentiating (3.1) yielding $(1 - m^2)m' = m^2$ and therefore $m' = q = m^2(1+q)$, and similarly for $n = 3$ we obtain

$$\frac{m''}{2} = \frac{q'}{2} = \frac{m^3}{(1-m^2)^3} = m^3(1+q)^3 = m^3 a_3(q). \qquad (5.11)$$

Assuming (5.10) for some $n \geq 3$ and differentiating yields

$$\begin{aligned} \frac{m^{(n)}}{n!} &= \Big(m^{n-1} q (1+q)^n + m^n q' (1+q)^{n-1} \Big) b_n(q) + \frac{m^n q' (1+q)^n b_n'(q)}{n} \\ &= m^{n+1}(1+q)^{n+1} \Big((1+2q) b_n(q) + \frac{2q(1+q) b_n'(q)}{n} \Big) = m^{n+1} a_{n+1}(q) \end{aligned}$$

from (5.11), completing the induction step.

Using Lemma 5.2 and (5.1) we immediately obtain an alternative representation of $M_{[k]}$ in the form

$$\begin{aligned} M_{[k]} &= \Big(\prod_{i=1}^{k} m_i\Big) \sum_{\pi\in \mathrm{NC}[k]} \mathrm{pTr}_{K(\pi)}(A_1,\dots,A_{k-1}) \prod_{B\in\pi}\Big(\sum_{E\in \mathrm{NCG}_{\mathrm{c}}(B)} q_E\Big) \\ &= \Big(\prod_{i=1}^{k} m_i\Big) \sum_{E\in \mathrm{NCG}[k]} \mathrm{pTr}_{K(\pi(E))}(A_1,\dots,A_{k-1}) q_E, \end{aligned} \tag{5.12}$$

where $\pi(E)\in \mathrm{NC}[k]$ denotes the non-crossing partition induced by the connected components of $E\in \mathrm{NCG}[k]$. Using (5.12) we obtain a third equivalent (this time recursive) definition of $M_{[k]}$ and also a simple tracial recursive relationship expressing the divided difference structure. For $1\le i<j\le k$ we define $M_{[i,j]}$ exactly as in (3.4) in terms of the non-crossing partitions $\pi\in \mathrm{NC}([i,j])$ and partial traces of the matrices $A_i,\dots,A_{j-1}$. For brevity of notations we furthermore set $M_{(i,j]} := M_{[i+1,j]}$ and $M_{[i,j)} := M_{[i,j-1]}$.

Lemma 5.4. *For any $k\ge 2$ we have the recursive relations*

$$\begin{aligned} M_{[k]} &= m_1\Big(A_1M_{(1,k]} + q_{1k}\langle A_1M_{(1,k]}\rangle + \sum_{j=2}^{k-1}\langle M_{[1,j]}\rangle\Big(M_{[j,k]} + q_{1k}\langle M_{[j,k]}\rangle\Big)\Big) \\ &= m_k\Big(M_{[1,k)}A_{k-1} + q_{1k}\langle M_{[1,k)}A_{k-1}\rangle + \sum_{j=2}^{k-1}\langle M_{[1,j]}\rangle\Big(M_{[j,k]} + q_{1k}\langle M_{[j,k]}\rangle\Big)\Big) \end{aligned} \tag{5.13}$$

and if $z_1\neq z_k$ then also

$$\langle M_{[k]}\rangle = \frac{\langle M_{[1,k)}A_{k-1} - A_1M_{(1,k]}\rangle}{z_1-z_k}. \tag{5.14}$$

Proof. From the partition (5.5)–(5.6) and (5.12) we obtain

$$\begin{aligned} &\frac{M_{[k]}}{m_1\cdots m_k} \\ &= \sum_{E\in \mathrm{NCG}(1,k]} q_E\Big(\mathrm{pTr}_{K(\pi(E))}\,A_{[1,k)} + q_{1k}\,\mathrm{pTr}_{K(\pi(E\cup\{(1k)\}))}\,A_{[1,k)}\Big) \\ &+ \sum_{j=2}^{k-1}\sum_{E_j\in \mathrm{NCG}^{(1j)}[j]\times \mathrm{NCG}[j,k]} q_{E_j}\Big(\mathrm{pTr}_{K(\pi(E_j))}\,A_{[1,k)} + q_{1k}\,\mathrm{pTr}_{K(\pi(E_j\cup\{(1k)\}))}\,A_{[1,k)}\Big). \end{aligned} \tag{5.15}$$

Since $E\in \mathrm{NCG}(1,k]$ implies $\{1\}\in\pi(E)$ it follows that in $K(\pi(E))$ both $1,k$ are in the same block and therefore

$$\mathrm{pTr}_{K(\pi(E))}\, A_{[1,k)} = A_1\, \mathrm{pTr}_{K(\pi(E))|_{(1,k]}}\, A_{(1,k)}.$$

Similarly, for $E_j = E_j^1 \sqcup E_j^2 \in \mathrm{NCG}^{(1j)}[j] \times \mathrm{NCG}[j,k]$ we note that

$$\mathrm{pTr}_{K(\pi(E_j))}\, A_{[1,k)} = \mathrm{pTr}_{K(\pi(E_j^1))|_{[j]}}\, A_{[1,j)}\, \mathrm{pTr}_{K(\pi(E_j^2))|_{[j,k]}}\, A_{[j,k)}$$

since $1, j$ are necessarily in the same block from $\pi(E)$ and therefore $[1,j)$ and $[j,k)$ are in different blocks of $K(\pi(E))$. Finally, we note that

$$\mathrm{pTr}_{K(\pi(E\cup\{(1k)\}))}\, A_{[1,k)} = \langle \mathrm{pTr}_{K(\pi(E\cup\{(1k)\}))}\, A_{[1,k)} \rangle$$

and

$$\sum_{E_j^1 \in \mathrm{NCG}^{(1j)}[j]} q_{E_j^1} \langle \mathrm{pTr}_{K(\pi(E_j^1))}\, A_{[1,j)} \rangle = \frac{q_{1j}}{1+q_{1j}} \sum_{E_j^1 \in \mathrm{NCG}[j]} q_{E_j^1} \langle \mathrm{pTr}_{K(\pi(E_j^1))}\, A_{[1,j)} \rangle,$$

so that (5.15) yields the first equality in (5.13). The proof of the second equality in (5.13) is completely analogous by partitioning the non-crossing graphs according to the edges connected to k rather than 1, hence details are omitted.

Using both equalities of (5.13) we obtain

$$\begin{aligned} \langle M_{[k]} \rangle &= \frac{m_1}{m_1 - m_k} \langle M_{[k]} \rangle - \frac{m_k}{m_1 - m_k} \langle M_{[k]} \rangle \\ &= (1+q_{1k}) \frac{m_1 m_k}{m_1 - m_k} \langle M_{[1,k)} A_{k-1} - A_1 M_{(1,k]} \rangle \end{aligned} \tag{5.16}$$

and (5.14) follows from

$$(1+q_{1k}) \frac{m_1 m_k}{m_1 - m_k} = \frac{q_{1k}}{m_1 - m_k} = \frac{1}{z_1 - z_k}$$

with the last equality due to

$$\frac{z_1 - z_k}{m_1 - m_k} = \frac{m_k - m_1}{m_1 - m_k} + \frac{1/m_k - 1/m_1}{m_1 - m_k} = \frac{1}{m_1 m_k} - 1 = \frac{1 - m_1 m_k}{m_1 m_k} = \frac{1}{q_{1k}}. \quad \square$$

5.2. Proof of Theorem 3.4

The proof of Theorem 3.4 is inductive over $K \geq 1$. For $K = k = 1$ both (3.6a)–(3.6b) follow directly from Theorem 3.2. For the induction we proceed in several steps:

(S1) Proof of (3.6b) for $k = K$ assuming (3.6b) for $k < K$.

- (S1a) Proof of (3.6b) for $k = K$, $A_k = I$ and $\rho_k = \max_j \rho_j$.
- (S1b) Proof of (3.6b) for $k = K$ traceless $\langle A_k \rangle = 0$.

(S2) Proof of (3.6a) for $k = K$ assuming (3.6b) for $k \leq K$ and (3.6a) for $k < K$.

We first note that (S1a)–(S1b) imply (S1). Indeed, by cyclicity we may rearrange $\langle G_1 A_1 \cdots G_k A_k \rangle$ in such a way that $\rho_k = \max_j \rho_j$ and by decomposing $A_k = \langle A_k \rangle I + (A_k - \langle A_k \rangle I)$ we conclude (S1) from linearity, (S1a) for $\langle A_k \rangle I$ and (S1b) for $A_k - \langle A_k \rangle I$. It thus remains to establish (S1a), (S1b) and (S2).

The main input for arguments below is the bound on renormalized chains of resolvents established in [11, Theorem 4.1, Remark 4.3] in the form of

$$\left| \langle \underline{W G_1 A_1 \cdots G_k A_k} \rangle \right| \prec \frac{\rho}{N \eta_*^k} \tag{5.17}$$

$$\left| \langle \boldsymbol{x}, \underline{W G_1 A_1 \cdots G_k} \boldsymbol{y} \rangle \right| \prec \sqrt{\frac{\rho}{N \eta_*}} \frac{1}{\eta_*^{k-1}} \tag{5.18}$$

Here the renormalization, denoted by underlining, is defined as

$$\underline{W f(W)} := W f(W) - \widetilde{\mathbf{E}} \widetilde{W} (\partial_{\widetilde{W}} f)(W),$$

with $\partial_{\widetilde{W}}$ denoting the directional derivative in direction $\widetilde{W}$ and $\widetilde{W}$ denotes an independent GUE-matrix with expectation $\widetilde{\mathbf{E}}$. Using the resolvent identity $WG - zG = I$, equation (3.1) and

$$\widetilde{\mathbf{E}} \widetilde{W} \partial_{\widetilde{W}} G = -\widetilde{\mathbf{E}} \widetilde{W} G \widetilde{W} G = -\langle G \rangle G$$

we obtain

$$G = m - m \underline{WG} + m \langle G - m \rangle G. \tag{5.19}$$

Proof of (S1a). In case $\Im z_1 \Im z_k < 0$ we use the resolvent identity with the abbreviation $G_{[a,b]} := G_a A_a G_{a+1} \cdots A_{b-1} G_b$ to obtain

$$\begin{aligned} \langle G_{[k]} \rangle &= \left\langle \frac{G_1 - G_k}{z_1 - z_k} A_1 G_{(1,k)} A_{k-1} \right\rangle \\ &= \frac{\langle M_{[1,k)} A_{k-1} - A_1 M_{(1,k]} \rangle}{z_1 - z_k} + \mathcal{O}_\prec \left(\frac{1}{N \eta_*^k} \right) = \langle M_{[k]} \rangle + \mathcal{O}_\prec \left(\frac{1}{N \eta_*^k} \right) \end{aligned} \tag{5.20}$$

from (5.14), the averaged local law for $k-1$ resolvents and $|z_1 - z_k| \geq \eta_1 \vee \eta_k \geq \eta_*$.

In case $\Im z_1 \Im z_k > 0$ we instead use (5.19) and

$$\underline{W G_{[k]}} = \underline{W G_1} A_1 G_{(1,k]} + \sum_{j=2}^{k} \langle G_{[j]} \rangle G_{[j,k]}$$

to obtain

$$G_{[k]} = m_1 A_1 G_{(1,k]} - m_1 \underline{WG_{[k]}} + \sum_{j=2}^{k-1} m_1 \langle G_{[j]} \rangle G_{[j,k]} + m_1 m_k \langle G_{[k]} \rangle + m_1 \langle G_1 - m_1 \rangle G_{[k]} + m_1 \langle G_{[k]} \rangle (G_k - m_k). \tag{5.21}$$

From the averaged local law (3.6b) for up to $k-1$ resolvents we have

$$\langle G_{[j]} \rangle \langle G_{[j,k]} \rangle = \langle M_{[j]} \rangle \langle M_{[j,k]} \rangle + \mathcal{O}_{\prec}\left(\frac{\rho}{N\eta_*^k} \right) \tag{5.22}$$

from

$$|\langle G_{[j]} \rangle| \leq \sqrt{\langle (G_1 A_1)(G_1 A_1)^* \rangle} \sqrt{\langle G^*_{2\cdots j} G_{2\cdots j} \rangle} \lesssim \frac{\sqrt{\langle \Im G_1 \rangle \langle \Im G_j \rangle}}{(\eta_1 \eta_j)^{1/2} \eta_2 \cdots \eta_{j-1}} \leq \frac{\rho}{\eta_*^{j-1}},$$

where in the second inequality we used Ward identity (3.11). By taking the averaged trace of (5.21) we obtain

$$\begin{aligned} &\Big(1 - m_1 m_k - m_1 \langle G_1 - m_1 \rangle - m_1 \langle G_k - m_k \rangle\Big) \langle G_{[k]} \rangle \\ &\quad = m_1 \langle A_1 G_{(1,k]} \rangle - m_1 \langle \underline{WG_{[k]}} \rangle + \sum_{j=2}^{k-1} m_1 \langle G_{[j]} \rangle \langle G_{[j,k]} \rangle \\ &\quad = m_1 \langle A_1 M_{(1,k]} \rangle + \sum_{j=2}^{k-1} m_1 \langle M_{[j]} \rangle \langle M_{[j,k]} \rangle + \mathcal{O}_{\prec}\left(\frac{\rho}{N\eta_*^k} \right), \end{aligned} \tag{5.23}$$

where in the second step we used (5.17), (5.22) and

$$\langle A_1 G_{(1,k]} \rangle = \langle A_1 M_{(1,k]} \rangle + \mathcal{O}_{\prec}\left(\frac{1}{N\eta_*^{k-1}} \right) = \langle A_1 M_{(1,k]} \rangle + \mathcal{O}_{\prec}\left(\frac{\rho}{N\eta_*^k} \right)$$

from the induction hypothesis. By (5.13) the first two terms on the r.h.s. of (5.23) can be written as

$$m_1 \Big(\langle A_1 M_{(1,k]} \rangle + \sum_{j=2}^{k-1} \langle M_{[1,j]} \rangle \langle M_{[j,k]} \rangle \Big) = \frac{\langle M_{[k]} \rangle}{1 + q_{1k}} = (1 - m_1 m_k) \langle M_{[k]} \rangle \tag{5.24}$$

so that due to

$$|\langle G_i - m_i \rangle \langle M_{[k]} \rangle| \prec \frac{\rho}{N\eta_*^k}$$

from (3.9) and Theorem 3.2 we finally conclude

$$\Big(1 - m_1 m_k - m_1 \langle G_1 - m_1 \rangle - m_1 \langle G_k - m_k \rangle\Big) \langle G_{[k]} - M_{[k]} \rangle = \mathcal{O}_{\prec}\left(\frac{\rho}{N\eta_*^k} \right). \tag{5.25}$$

By elementary estimates from the definition of m in (3.1) and the fact that $\Im z_1 \Im z_k > 0$ we easily conclude $|1 - m_1 m_k| \gtrsim \rho_k = \rho$ and therefore together with $|\langle G_i - m_i \rangle| \prec 1/(N\eta_*) \ll \rho$ we immediately conclude (5.20). □

Proof of (S1b). For such A_k with $\langle A_k \rangle = 0$, multiplying (5.21) from the rhs. and taking the trace, we obtain

$$\begin{aligned} &\left(1 + \mathcal{O}_\prec\left(\frac{1}{N\eta_*}\right)\right)\langle G_{[k]} A_k \rangle \\ &= m_1 \langle A_1 G_{(1,k]} A_k \rangle + \sum_{j=2}^{k-1} m_1 \langle G_{[j]} \rangle \langle G_{[j,k]} A_k \rangle + \mathcal{O}_\prec\left(\frac{1}{N\eta_*^k}\right) \\ &= \langle M_{[k]} A_k \rangle + \mathcal{O}_\prec\left(\frac{1}{N\eta_*^k}\right) \end{aligned} \tag{5.26}$$

from (5.13) and the local laws for up to $k-1$ resolvents. □

Proof of (S2). We take the inner product of (5.21) with $\boldsymbol{x}, \boldsymbol{y}$ to obtain

$$\begin{aligned} &\langle \boldsymbol{x}, G_{[k]} \boldsymbol{y} \rangle \\ &= m_1 \langle \boldsymbol{x}, A_1 G_{(1,k]} \boldsymbol{y} \rangle + \sum_{j=2}^{k-1} m_1 \langle G_{[j]} \rangle \langle \boldsymbol{x}, G_{[j,k]} \boldsymbol{y} \rangle + m_1 m_k \langle \boldsymbol{x}, \boldsymbol{y} \rangle \langle G_{[k]} \rangle \\ &\quad + \mathcal{O}_\prec\left(\sqrt{\frac{\rho}{N\eta_*}} \frac{1}{\eta_*^{k-1}}\right) \\ &= m_1 \langle \boldsymbol{x}, A_1 M_{(1,k]} \boldsymbol{y} \rangle + \sum_{j=2}^{k-1} m_1 \langle M_{[j]} \rangle \langle \boldsymbol{x}, M_{[j,k]} \boldsymbol{y} \rangle + m_1 m_k \langle \boldsymbol{x}, \boldsymbol{y} \rangle \langle M_{[k]} \rangle \\ &\quad + \mathcal{O}_\prec\left(\sqrt{\frac{\rho}{N\eta_*}} \frac{1}{\eta_*^{k-1}}\right) \end{aligned} \tag{5.27}$$

from the renormalization bound (5.18), the averaged local laws (3.6b) for up to k resolvents and isotropic local laws (3.6a) up to $k-1$ resolvents. Now from the recursive relation (5.13) and $(1 + q_{1k}) m_1 m_k = q_{1k}$ we have

$$m_1 A_1 M_{(1,k]} + \sum_{j=2}^{k-1} m_1 \langle M_{[j]} \rangle M_{[j,k]} + m_1 m_k \langle M_{[k]} \rangle = M_{[k]} \tag{5.28}$$

and the claim follows together with (5.27). □

By combining (S1)–(S2) we conclude the induction step and thereby the proof of Theorem 3.4.

References

[1] O. Ajanki, L. Erdős, T. Krüger, Singularities of solutions to quadratic vector equations on the complex upper half-plane, Commun. Pure Appl. Math. 70 (2017) 1672–1705, MR3684307.
[2] J. Alt, L. Erdős, T. Krüger, The Dyson equation with linear self-energy: spectral bands, edges and cusps, Doc. Math. 25 (2020) 1421–1539, MR4164728.
[3] G.W. Anderson, A. Guionnet, O. Zeitouni, An Introduction to Random Matrices, Cambridge Studies in Advanced Mathematics, vol. 118, Cambridge University Press, Cambridge, 2010, xiv+492, MR2760897.
[4] O. Arizmendi, T. Hasebe, F. Lehner, C. Vargas, Relations between cumulants in noncommutative probability, Adv. Math. 282 (2015) 56–92, MR3374523.
[5] Z. Bao, Y. He, Quantitative CLT for linear eigenvalue statistics of Wigner matrices, preprint, arXiv: 2103.05402, 2021.
[6] S.T. Belinschi, M. Capitaine, Spectral properties of polynomials in independent Wigner and deterministic matrices, J. Funct. Anal. 273 (2017) 3901–3963, MR3711884.
[7] S.T. Belinschi, T. Mai, R. Speicher, Analytic subordination theory of operator-valued free additive convolution and the solution of a general random matrix problem, J. Reine Angew. Math. 732 (2017) 21–53, MR3717087.
[8] L. Benigni, P. Lopatto, Optimal delocalization for generalized Wigner matrices, preprint, arXiv: 2007.09585, 2020.
[9] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, J. Yin, Isotropic local laws for sample covariance and generalized Wigner matrices, Electron. J. Probab. 19 (33) (2014) 53, MR3183577.
[10] G. Cipolloni, L. Erdős, D. Schröder, Edge universality for non-Hermitian random matrices, Probab. Theory Relat. Fields 179 (2021) 1–28, MR4221653.
[11] G. Cipolloni, L. Erdős, D. Schröder, Eigenstate thermalization hypothesis for Wigner matrices, preprint, arXiv:2012.13215, 2020.
[12] G. Cipolloni, L. Erdős, D. Schröder, Functional central limit theorems for Wigner matrices, preprint, arXiv:2012.13218, 2020.
[13] B. Collins, A. Guionnet, F. Parraud, On the operator norm of non-commutative polynomials in deterministic matrices and iid GUE matrices, preprint, arXiv:1912.04588, 2019.
[14] B. Collins, T. Mai, A. Miyagawa, F. Parraud, S. Yin, Convergence for noncommutative rational functions evaluated in random matrices, preprint, arXiv:2103.05962, 2021.
[15] H.L. Cycon, R.G. Froese, W. Kirsch, B. Simon, Schrödinger Operators with Application to Quantum Mechanics and Global Geometry, Study, Texts and Monographs in Physics, Springer-Verlag, Berlin, 1987, x+319, MR883643.
[16] E.B. Davies, The functional calculus, J. Lond. Math. Soc. (2) 52 (1995) 166–176, MR1345723.
[17] C. Domb, A.J. Barrett, Enumeration of ladder graphs, Discrete Math. 9 (1974) 341–358, MR345872.
[18] K. Dykema, On certain free product factors via an extended matrix model, J. Funct. Anal. 112 (1993) 31–60, MR1207936.
[19] L. Erdős, A. Knowles, H.-T. Yau, J. Yin, The local semicircle law for a general class of random matrices, Electron. J. Probab. 18 (59) (2013) 58, MR3068390.
[20] L. Erdős, B. Schlein, H.-T. Yau, Local semicircle law and complete delocalization for Wigner random matrices, Commun. Math. Phys. 287 (2009) 641–655, MR2481753.
[21] L. Erdős, H.-T. Yau, J. Yin, Rigidity of eigenvalues of generalized Wigner matrices, Adv. Math. 229 (2012) 1435–1515, MR2871147.
[22] U. Haagerup, F. Larsen, Brown's spectral distribution measure for R-diagonal elements in finite von Neumann algebras, J. Funct. Anal. 176 (2000) 331–367, MR1784419.
[23] U. Haagerup, S. Thorbjørnsen, A new application of random matrices: $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ is not a group, Ann. Math. (2) 162 (2005) 711–775, MR2183281.
[24] J.W. Helton, T. Mai, R. Speicher, Applications of realizations (aka linearizations) to free probability, J. Funct. Anal. 274 (2018) 1–79, MR3718048.
[25] S.R.M. Jacobo, Aplicaciones de probabilidad libre a la teoría de información cuántica, MA thesis Centro de Investigación en Mathemáticas, A.C, 2019.
[26] A. Knowles, J. Yin, The isotropic semicircle law and deformation of Wigner matrices, Commun. Pure Appl. Math. 66 (2013) 1663–1750, MR3103909.
[27] G. Kreweras, Sur les partitions non croisées d'un cycle, Discrete Math. 1 (1972) 333–350, MR309747.
[28] T. Mai, R. Speicher, Free probability, random matrices, and representations of non-commutative rational functions, in: Computation and Combinatorics in Dynamics, Stochastics and Control, in: Abel Symp, vol. 13, Springer, Cham, 2018, pp. 551–577, MR3967396.

[29] C. Male, The norm of polynomials in large random and deterministic matrices, Probab. Theory Relat. Fields 154 (2012) 477–532, MR3000553.
[30] M.L. Mehta, Random Matrices and the Statistical Theory of Energy Levels, Academic Press, New York-London, 1967, x+259, MR0220494.
[31] J.A. Mingo, R. Speicher, Free probability and random matrices, in: Fields Institute Monographs, vol. 35, Springer, New York, Fields Institute for Research in Mathematical Sciences, Toronto, ON, 2017, xiv+336, MR3585560.
[32] H. Mühle, Noncrossing Set Partitions, Lecture notes (TU Dresden) 2017.
[33] F. Parraud, Asymptotic expansion of smooth functions in polynomials in deterministic matrices and iid GUE matrices, preprint, arXiv:2011.04146, 2020.
[34] R. Speicher, Multiplicative functions on the lattice of noncrossing partitions and free convolution, Math. Ann. 298 (1994) 611–628, MR1268597.
[35] D. Voiculescu, A strengthened asymptotic freeness result for random matrices with applications to free entropy, Int. Math. Res. Not. 41 (63) (1998) MR1601878.
[36] D. Voiculescu, Limit laws for random matrices and free products, Invent. Math. 104 (1991) 201–220, MR1094052.
[37] E.P. Wigner, Characteristic vectors of bordered matrices with infinite dimensions, Ann. Math. (2) 62 (1955) 548–564, MR77805.
[38] S. Yin, Non-commutative rational functions in strongly convergent random variables, Adv. Oper. Theory 3 (2018) 178–192, MR3730345.