Electronic Journal of Probability

Electron. J. Probab. **27** (2022), article no. 117, 1–38.
ISSN: 1083-6489 https://doi.org/10.1214/22-EJP838

Optimal multi-resolvent local laws for Wigner matrices*

Giorgio Cipolloni[†] László Erdős[‡] Dominik Schröder[§]

Abstract

We prove local laws, i.e. optimal concentration estimates for arbitrary products of resolvents of a Wigner random matrix with deterministic matrices in between. We find that the size of such products heavily depends on whether some of the deterministic matrices are traceless. Our estimates correctly account for this dependence and they hold optimally down to the smallest possible spectral scale.

Keywords: global law; local law; random matrices.
MSC2020 subject classifications: 60B20; 15B52.
Submitted to EJP on January 14, 2022, final version accepted on August 11, 2022.

1 Introduction

A remarkable feature of large Hermitian random matrices H is that their resolvents $G(z) = (H - z)^{-1}$ tend to concentrate around a deterministic matrix $M = M(z)$ for spectral parameters $z \in \mathbf{C}$ even just slightly away from the real axis. If the correlation among the matrix entries of H is sufficiently weak, in particular for *Wigner matrices* with independent (up to Hermitian symmetry) and identically distributed matrix elements, this phenomenon holds as long as $|\Im z|$ is just slightly above the typical eigenvalue spacing around $\Re z$. While the random matrix H strongly fluctuates around its mean $\mathbf{E}\, H$, it is surprising that the resolvent has such a strong concentration property even on small spectral scales. Rigorous results of this type are generally called *local laws* and they play a fundamental role in random matrix theory since they are able to resolve spectral properties of H almost down to individual eigenvalues. We remark that for Wigner matrices $M(z) = m(z)I$ is the multiple of the identity matrix, where m is the Stieltjes transform of Wigner's semicircle distribution. For more general ensembles M is given as the solution of the *(matrix) Dyson equation*, a non-linear deterministic equation [3].

[*] L. Erdős was supported by ERC Advanced Grant "RMTBeyond" No. 101020331. D. Schröder was supported by Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zürich Foundation.

[†] Princeton Center for Theoretical Science, Princeton University, Princeton, NJ 08544, USA. E-mail: gc4233@princeton.edu

[‡] IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria. E-mail: lerdos@ist.ac.at

[§] Institute for Theoretical Studies, ETH Zurich, Clausiusstr. 47, 8092 Zurich, Switzerland. E-mail: dschroeder@ethz.ch

Historically, the primary motivation for local laws was to provide the necessary a priori estimates in the *three step strategy* to prove the Wigner-Dyson-Mehta spectral universality for random matrices via the Dyson Brownian Motion (DBM), see [32] for a comprehensive summary. The first local law was proved for Wigner matrices in the tracial sense [30]; extended later to more general *entry-wise* [33] and *isotropic* [40] senses, as well as to much more general classes of random matrices, including nonzero expectation [38, 42, 44], nontrivial variance profile [4], and even correlations [3, 29]. Numerous related works focused on local laws for band matrices [13, 26, 50–52], sparse matrices [8–10, 27, 43, 43], heavy tails [2, 12], accurate error terms [18, 35], general invariant β-ensembles [1, 14–17, 24, 39, 45, 49] and many more.

With a very few recent exceptions, listed at the end of Section 1.1, all local laws so far concerned a single resolvent. Their *averaged* and *isotropic* versions assert that for any fixed $\epsilon > 0$, deterministic test matrix B and test vectors $\boldsymbol{x}, \boldsymbol{y}$, the bounds

$$|\langle (G(z) - M(z))B\rangle| \le \frac{N^\epsilon \|B\|}{N\eta}, \qquad |\langle \boldsymbol{x}, (G(z) - M(z))\boldsymbol{y}\rangle| \le \frac{N^\epsilon \|\boldsymbol{x}\|\|\boldsymbol{y}\|}{\sqrt{N\eta}}, \qquad \eta := |\Im z| \tag{1.1}$$

hold with very high probability, where N is a dimension of H, $\langle R\rangle := \frac{1}{N}\operatorname{Tr} R$ denotes the normalized trace and $\langle\cdot,\cdot\rangle$ denotes the scalar product in $\mathbf{C}^N$. The estimates (1.1) are optimal in the critical small η regime (up to the factor N^ϵ).

This paper is concerned with the multi-resolvent generalizations of (1.1). If G is approximated by M, what approximates the square of the resolvent? The naive answer $G^2 \approx M^2$ is wrong, even for the simplest Wigner case since the approximation $G \approx M$ in (1.1) holds true only in weak sense; it cannot be "squared". Nevertheless G^2 still concentrates and the hint given by the identity $G(z)^2 = \partial_z G(z)$ leads to the correct answer. Indeed $G(z)^2 \approx \partial_z M(z)$ in the sense

$$|\langle (G(z)^2 - \partial_z M(z))B\rangle| \le \frac{N^\epsilon \|B\|}{N\eta^2}, \qquad |\langle \boldsymbol{x}, (G(z)^2 - \partial_z M(z))\boldsymbol{y}\rangle| \le \frac{N^\epsilon \|\boldsymbol{x}\|\|\boldsymbol{y}\|}{\sqrt{N}\eta^{3/2}}, \tag{1.2}$$

and again the error terms are optimal. Note that these error terms match the differentiation procedure; indeed (1.2) can formally be obtained by "differentiating" (1.1).

Such algebraic ideas, however, do not help much further if we ask for concentration of the alternating product

$$G(z_1)B_1G(z_2)B_2G(z_3)\dots B_{k-1}G(z_k) \tag{1.3}$$

of resolvents and deterministic matrices $B_1, B_2, \dots$, and more generally for

$$f_1(H)B_1f_2(H)B_2\dots B_{k-1}f_k(H), \tag{1.4}$$

where f_i's are arbitrary functions on $\mathbf{R}$. The product (1.3) still concentrates but its deterministic approximation, denoted by $M(z_1, B_1, z_2, \dots, B_{k-1}, z_k)$, is non-trivial even for the Wigner case and it was identified only recently in [23, Theorem 3.4] (however, formulas for traces of (1.4) when f_i's are polynomials have already been obtained within free probability theory, see e.g. [6, Theorem 5.4.5] or [48, Sect 4. Thm 20.]). The main result of the current work is to prove the optimal error term for this approximation and thus to establish the optimal local law for any product of the type (1.3) when H is from the Wigner ensemble (Theorem 2.5). These optimal multi-resolvent local laws will then be used to establish the universality of the Gaussian fluctuations of (1.4) in subsequent works. To keep the current paper focused, we present here only one simple application of our new local law to improve our control on the thermalisation effect of the Wigner matrices (see Remark 2.8 below).

In connection with CLT for linear eigenvalue statistics, special cases of tracial local laws for (1.3) for $k = 2, 3$ have been proven in [7, 19, 31, 36, 37, 46, 47]. These results, however, considered the special $B_i = I$ case, where resolvent identities can directly reduce the number of G's. More importantly, the accurate analysis of the case with general B's must handle traceless B's separately as we explain in the next subsection.

1.1 The role of the traceless matrices

The major complication for the multi-resolvent local law is that the size of $M(z_1, B_1, z_2, \ldots, B_{k-1}, z_k)$ heavily depends on whether some of the matrices B_i are traceless or not, and the error term must match the size of M to be considered optimal. For example, if $B_1 = B_2 = \ldots = B_{k-1} = I$, then $\langle M(z_1, I, z_2, \ldots, I, z_k)\rangle \sim (1/\eta)^{k-1}$ with $\eta := \min |\Im z_i|$ in the interesting regime where $\eta \lesssim 1$, and the corresponding local law

$$|\langle G(z_1)G(z_2)G(z_3)\ldots G(z_k) - M(z_1, I, z_2, \ldots, I, z_k)\rangle| \leq \frac{N^\epsilon}{N\eta^k} = \frac{1}{\eta^{k-1}}\frac{N^\epsilon}{N\eta} \tag{1.5}$$

is optimal (up to N^ϵ) for $\eta \lesssim 1$. Note that the error term is by a factor $N^\epsilon/N\eta$ smaller than the deterministic approximation, hence (1.5) proves concentration for any $\eta \gg 1/N$.

Exactly the same estimate holds for (1.3) with general deterministic matrices B_i with $\|B_i\| = 1$ instead of $B_i = I$, see [23, Theorem 3.4]. However, if all $B_1, B_2, \ldots, B_k$ are traceless, $\langle B_i\rangle = 0$, then in the $\eta \lesssim 1$ regime typically

$$\langle M(z_1, B_1, z_2, \ldots, B_{k-1}, z_k)B_k\rangle \sim \frac{1}{\eta^{\lfloor k/2\rfloor - 1}}, \tag{1.6}$$

therefore $N^\epsilon/(N\eta^k)$ in (1.5) is much bigger than the deterministic approximation. This indicates that the robust error term proven in [23, Theorem 3.4] for general matrices is far from being optimal when traceless matrices are involved, but it does not give a hint what the optimal error term should be.

The correct answer, in a heuristic form, can be formulated by the following rule of thumb that we coin the $\sqrt{\eta}$-rule (in the $\eta \lesssim 1$ regime):

$\sqrt{\eta}$-rule: Each traceless matrix B_i reduces both the size of M and the error term by a factor $\sqrt{\eta}$.

Establishing the $\sqrt{\eta}$-rule for M is relatively straightforward given its explicit form, but for the error term it is much harder – this is the main content of the current paper.

The special role of a traceless deterministic matrix even for the single resolvent local law was observed only recently in [22], where it was shown that

$$|\langle (G(z) - m(z))B\rangle| = |\langle G(z)B\rangle| \leq \frac{N^\epsilon}{N\sqrt{\eta}}$$

if $\langle B\rangle = 0$ in contrast to the much bigger error of order $1/(N\eta)$ for general B in (1.1). In fact, $G - m$ has two different fluctuation modes, a tracial and a traceless one, expressed somewhat informally in the following two-scale central limit theorem

$$\langle (G(z) - m(z))B\rangle \approx \langle B\rangle \frac{\xi_1}{N\eta} + \langle \mathring{B}\mathring{B}^*\rangle^{1/2}\frac{\xi_2}{N\sqrt{\eta}} \tag{1.7}$$

where ξ_1 and ξ_2 are independent Gaussian variables and $\mathring{B} := B - \langle B\rangle$ is the traceless part of B. The asymptotics $\approx$ in (1.7) is understood in the sense of all moments and in the limit as $N\eta \gg 1$; see [23, Theorem 4.1] for the precise statement.

Tracking the influence of the traceless deterministic matrices in multi-resolvent local laws for Wigner matrices played an essential role in our proof of the *Eigenstate* *thermalisation hypothesis* [21], and in the *functional central limit theorems* to understand the fluctuation modes of $f(W)$ as a matrix [22]. However, in these papers only two- and three-resolvent local laws were necessary and suboptimal error was sufficient. For example, a key technical ingredient in [21] was the local law

$$\langle G(z)BG^*(z)B\rangle = |m(z)|^2\langle BB^*\rangle + O\Big(\frac{N^\epsilon}{\sqrt{N\eta}}\Big) \tag{1.8}$$

for any $\langle B\rangle = 0$ with $\|B\| \lesssim 1$, which in particular implied the upper bound

$$\langle G(z)BG(z)B\rangle = \mathcal{O}(1), \qquad \text{for } N\eta \ge N^{2\epsilon}$$

in agreement with (1.6) applied to $k = 2$. In the relevant small η regime the error in (1.8) is better than the robust error of order $1/(N\eta^2)$ from (1.5) valid irrespective whether B is traceless or not, but (1.8) is still far from optimal. The $\sqrt{\eta}$-rule predicts an error term of order $1/(N\eta)$ in (1.8), a factor of $(\sqrt{\eta})^2$ better than the robust error (1.5), while (1.8) does not even get the optimal N-power that is naturally expected in the $\eta \sim 1$ regime. Similarly, specific three-resolvent local laws that were proven in [22, Proposition 3.4], also came with suboptimal errors. Finally, we mention a related two-resolvent local law for the Hermitization of an i.i.d. matrix in [20, Theorem 5.2] where the mechanism for the reduced error term is different from the $\sqrt{\eta}$-rule.

1.2 Strategy of the proof

We developed a very concise new method to prove multi-resolvent local laws. The basic idea for all local law proofs is to show that G, or in the multi-resolvent case $GBGB\dots G$ from (1.3), approximately satisfies the Dyson equation, the defining equation of the corresponding M. In the previous approaches the fluctuating error term in this approximation was treated separately and it was shown to be negligible with the help of a high moment cumulant expansion. The expansion generated many terms and a fairly involved Feynman diagrammatic representation was needed to bookkeep and estimate them. This becomes especially cumbersome where some additional smallness effect needs to be consistently tracked along the whole expansion. For example, in the main technical Theorem 4.1 in [21], we meticulously counted the number of “effectively” traceless B factors, struggling with the complication that some B factor becomes B^2 along the cumulant expansion, losing its smallness effect. Even suboptimal error terms for small k as in (1.8) required major efforts and the general case was out of reach.

Our new method drastically simplifies this procedure using two unrelated ideas. First, the large Feynman diagrammatic representation is actually due to an overexpansion of the fluctuating error term which can be considerably reduced if one expands “minimalistically”, so to say. In the context of single resolvent averaged local laws this idea appeared first in [43], coined as *recursive moment estimates*, we will use this philosophy for the multi-resolvent situation and also for the isotropic case.

Second, the fundamental concern in the proofs of multi-resolvent local laws is how to truncate the resulting hierarchy involving longer and longer chains of the form $GBGB\dots G$. The cumulant expansion for a chain of length k as in (1.3) will contain chains of length up to $2k$. For the single resolvent local law, $k = 1$, this problem is usually solved by the Ward identity $GG^* = \Im G/\eta$, immediately reducing longer chains to a single resolvent. If traceless matrices are in between G’s such identity is not directly applicable. In [21] we solved this problem by considering the positive quantity $\Lambda^2 := \langle \Im GB\Im GB\rangle$ for traceless B and estimated all longer chains in terms of Λ, to arrive, finally, at a simple Gromwall-type inequality for Λ, roughly of the type

$$\Lambda^2 \lesssim 1 + \frac{\Lambda^2}{N\eta}, \tag{1.9}$$

from which $\Lambda \lesssim 1$ immediately follows. The reduction of longer chains to Λ's involved a careful Schwarz inequality within the spectral decomposition of H, for example for an averaged chain involving $2k$ resolvents (using $\Im G$'s instead of G for illustrational simplicity) we used

$$\begin{aligned} |\langle (\Im G\, B)^{2k}\rangle| &= \frac{1}{N}\Big| \sum_{i_1\ldots i_{2k}} \langle \boldsymbol{u}_{i_1}, B\boldsymbol{u}_{i_2}\rangle \langle \boldsymbol{u}_{i_2}, B\boldsymbol{u}_{i_3}\rangle \ldots \langle \boldsymbol{u}_{i_{2k}}, B\boldsymbol{u}_{i_1}\rangle \prod_{j=1}^{2k} \Im \frac{1}{\lambda_{i_j} - z}\Big| \\ &\le \frac{1}{N}\Big(\sum_{ij} |\langle \boldsymbol{u}_i, B\boldsymbol{u}_j\rangle|^2 \Im \frac{1}{\lambda_i - z} \Im \frac{1}{\lambda_j - z}\Big)^k \\ &= N^{k-1} \langle \Im G B \Im G B\rangle^k. \end{aligned} \tag{1.10}$$

Here λ_i and $\boldsymbol{u}_i$ are the eigenvalues and the orthonormal eigenvectors of H, respectively. The size of the l.h.s., based upon its deterministic approximation (1.6), is η^{-k+1}, while the r.h.s. is of order N^{k-1} hence this inequality lost a factor $(N\eta)^{k-1}$. Very roughly, each summation in (1.10) effectively runs over $N\eta$ different i indices and if each summand were independent, then an effective central limit theorem would reduce the size by a factor $1/\sqrt{(N\eta)^{2k}} = (N\eta)^{-k}$, in reality this effect is weaker by a factor $N\eta$. Nevertheless, for larger k's this loss in the Schwarz inequality in (1.10) cannot be recovered from the smallness of higher order cumulants, which eventually results in suboptimal error terms in the local law in [21]. Another complication is that the bound (1.10) is also needed for $(GB)^k$. Since spectrally G is much less localized than $\Im G$, technically we could not do the analysis locally in the spectrum and Λ was actually defined after taking a supremum over the real parts of the spectral parameters z_i in G's.

The basic objects in the current paper are the appropriately rescaled versions of the *differences* $(GB)^k - M_k B$ between alternating chains of length k and their deterministic counterparts M_k. More precisely, we set

$$\Psi_k^{\mathrm{av}} := N\eta^{k/2} |\langle (GB)^k - M_k B\rangle|, \tag{1.11}$$

and its isotropic version Ψ_k^{iso} is defined similarly. The general definition allows for different spectral parameters and different B matrices in the $GBGBGB$.... chain but we ignore this technicality here. The rescaling is chosen such that $\Psi_k^{\mathrm{av,iso}} \lesssim 1$ corresponds to the optimal local laws to be proven.

The "minimalistic" cumulant expansion applied directly to the moments of Ψ's generates further chains of alternating products of resolvents and B's. Each of them is expressed as their deterministic "main term" M plus the error term involving Ψ's, i.e. for this purpose we write (1.11) as

$$\langle (GB)^k\rangle = \langle M_k B\rangle + \mathcal{O}\Big(\frac{\Psi_k^{\mathrm{av}}}{N\eta^{k/2}}\Big),$$

and similarly for matrix elements $[(GB)^k]_{ab}$. The explicit M_k terms can be directly estimated, leaving us with a nonlinear infinite hierarchy of coupled *master inequalities* for Ψ_k^{av} and Ψ_k^{iso} for each k (Proposition 3.5). The estimate for Ψ_k still contains terms involving Ψ_{2k} since the cumulant expansion generates longer chains. This time, however, we truncate the hierarchy in the most economical way; roughly speaking a chain of length $2k$ is split into two chains of length k instead of k chains of length two as in (1.10). Hence many fewer $N\eta$ factors are lost in the analogue of (1.10); the loss is only $(N\eta)^2$ for the averaged bounds and $N\eta$ in the isotropic bound, independently of k (see Lemma 3.6 below).

Even after the reduction of longer chains to shorter ones, the new truncated system of master inequalities cannot be closed by a simple algebra, in contrast to the single inequality (1.9) derived for Λ. We first prove a non-optimal *a priori* bound $\Psi_k^{\mathrm{av,iso}} \lesssim \sqrt{N\eta}$ *for all* k with a step-two induction argument and successively improving the power of $N\eta$ in each step. Then we start the procedure all over again, but now we will not use the reduction of Ψ_{2k}'s back to Ψ_k's that would cost us $(N\eta)$ or $(N\eta)^2$ factors; we rather use the already proven a priori bound $\Psi_{2k} \lesssim \sqrt{N\eta}$ that loses only $\sqrt{N\eta}$. It turns out that such a loss can finally be compensated by the smaller size of the higher cumulants.

Summarizing, the key conceptual novelty in the current approach compared with [21] is twofold. First, in [21] we operated with upper bounds on size of the chains, like (1.10), while now we operate on the level of the much more precise Ψ's measuring the fluctuations of the chains, i.e. their deviations from their deterministic counterpart. This enables us to determine the leading order term for resolvent chains of any length, and perform a more accurate analysis purely on the level of sub-leading deviations. Second, longer chains are split only into two smaller chains, yielding much less $(N\eta)$-factors lost. However, the price for this higher accuracy is that we need to handle a new infinite system of inequalities for the Ψ's. Finally, two important technical differences are that (i) we can work locally in the spectrum and (ii) now we use the minimalistic cumulant expansion that considerably shortens the argument.

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 $k < l$. By $\lceil \cdot \rceil$, $\lfloor \cdot \rfloor$ we denote the upper and lower integer part, respectively, i.e. for $x \in \mathbf{R}$ we define $\lceil x \rceil := \min\{m \in \mathbf{N} : \ m \geq x\}$ and $\lfloor x \rfloor := \max\{m \in \mathbf{N} : \ m \leq x\}$. 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 the moment condition, see (2.1) later. 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 $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^N$ and matrices $A \in \mathbf{C}^{N \times N}$ we define

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \sum_{i=1}^{N} \overline{x}_i y_i, \qquad A_{\boldsymbol{x}\boldsymbol{y}} := \langle \boldsymbol{x}, A\boldsymbol{y} \rangle.$$

We will use the concept of "with very high probability" meaning that for any fixed $D > 0$ the probability of an 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. We introduce the notion of *stochastic domination* (see e.g. [28]): given two families of non-negative random variables

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

indexed by N (and possibly some parameter u) 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 $X = \mathcal{O}_\prec(Y)$.

2 Main results

We start with the definition of the matrix model we consider.

Definition 2.1. *We call W a Wigner matrix if it is an $N \times N$ random Hermitian matrix which satisfies the following properties. The off-diagonal matrix elements below the diagonal are centred independent, identically distributed (i.i.d) real ($\beta = 1$) or complex ($\beta = 2$) random variables with $\mathbf{E}|w_{ij}|^2 = 1/N$. Additionally, in the complex case we assume that $\mathbf{E}\, w_{ij}^2 = 0$. The diagonal elements are centred i.i.d. real random variables with $\mathbf{E}\, w_{ii}^2 = 2/(N\beta)$. Furthermore, we assume that for every $q \in N$ there is a constant C_q such that*

$$\mathbf{E}|\sqrt{N} w_{ij}|^q \le C_q. \tag{2.1}$$

Remark 2.2. The assumptions $\mathbf{E}\, w_{ij}^2 = 0$ in the complex case, and $\mathbf{E}\, w_{ii}^2 = 2/(\beta N)$ are made to make the presentation clearer. All our results can be easily extended to this case as well, but we refrain from doing it for notational simplicity.

We set $G(z) := (W - z)^{-1}$ to be resolvent of the Wigner matrix W with spectral parameter $z \in \mathbf{C} \setminus \mathbf{R}$. The *optimal local law* asserts that $G(z)$ is approximately equal to $m(z)I$ down to the microscopic scale $|\Im z| \gg 1/N$, where

$$m(z) = m_{\mathrm{sc}}(z) := \int_{-2}^{2} \frac{1}{x - z} \rho_{\mathrm{sc}}(x)\, \mathrm{d}x, \qquad \rho_{\mathrm{sc}}(x) := \frac{\sqrt{4 - x^2}}{2\pi} \tag{2.2}$$

is the Stieltjes transform of the semicircular distribution.

Theorem 2.3. *For any $z \in \mathbf{C} \setminus \mathbf{R}$ with $|z| \le N^{100}$, $d := \mathrm{dist}(z, [-2, 2])$, $\eta := |\Im z|$ and any deterministic vectors $\boldsymbol{x}, \boldsymbol{y}$ it holds that*

$$|\langle G - m \rangle| \prec \begin{cases} \frac{1}{N\eta}, & d < 1 \\ \frac{1}{Nd^2}, & d \ge 1, \end{cases} \qquad |\langle \boldsymbol{x}, (G - m)\boldsymbol{y} \rangle| \prec \|\boldsymbol{x}\| \|\boldsymbol{y}\| \begin{cases} \frac{\sqrt{|\Im m(z)|}}{\sqrt{N\eta}} + \frac{1}{N\eta}, & d < 1 \\ \frac{1}{\sqrt{N} d^2}, & d \ge 1. \end{cases} \tag{2.3}$$

theorem 2.3 in this form, including both the $d < 1$ and $d \ge 1$ regimes, can be found in [29, Theorem 2.1] even for much more general random matrix ensembles allowing for correlations. Its tracial version and its special entry-wise version (where $\boldsymbol{x}, \boldsymbol{y}$ are coordinate vectors) have already been established in [5, Lemma B.1]. However, the really interesting $d < 1$ regime has been proven much earlier: tracial version in [30], entry-wise version in [33] and isotropic version [40]; with many other refinements and generalisations mentioned in the introduction. The $d \ge 1$ regime, sometimes called the *global law*, is much easier and most papers on the local law naturally excluded it for convenience albeit they could have handled this regime, too, with some minor extra effort.

In case of several spectral parameters $z_1, z_2, \dots$ we use the abbreviation $G_i := G(z_i)$. For our main result we recall from [23] that the deterministic approximation to $G_1 B_1 G_2 \cdots G_{k-1} B_{k-1} G_k$ for arbitrary deterministic matrices $B_1, \dots, B_{k-1}$ is given by

$$M(z_1, B_1, \dots, B_{k-1}, z_k) := \sum_{\pi \in \mathrm{NC}[k]} \mathrm{pTr}_{K(\pi)}(B_1, \dots, B_{k-1}) \prod_{B \in \pi} m_\circ[B], \tag{2.4}$$

where $\mathrm{NC}[k]$ denotes the non-crossing partitions of the set $[k] = \{1, \dots, k\}$ arranged in increasing order, and $K(\pi)$ denotes the Kreweras complement of π [41], e.g. $K(\{134|2|5|6\}) = \{12|3|456\}$. Moreover, the partial trace pTr_π with respect to a partition π is given by

$$\mathrm{pTr}_\pi(B_1, \dots, B_{k-1}) = \prod_{B \in \pi \setminus B(k)} \left\langle \prod_{j \in B} B_j \right\rangle \prod_{j \in B(k) \setminus \{k\}} B_j, \tag{2.5}$$

with $B(k) \in \pi$ denoting the unique block containing k. Finally for any subset $B \subset [k]$ we define $m[B] := m_{\mathrm{sc}}[z_B]$ as the iterated divided difference of m_{sc} evaluated in $z_B := \{z_i \mid i \in B\}$, and by $m_\circ[\cdot]$ denote the free-cumulant transform of $m[\cdot]$ which is defined implicitly by the relation

$$m[B] = \sum_{\pi \in \mathrm{NC}(B)} \prod_{B' \in \pi} m_\circ[B'], \qquad \forall B \subset [k], \tag{2.6}$$

e.g. $m_\circ[i,j] = m[\{i,j\}] - m[\{i\}]m[\{j\}]$. We note that the iterated divided difference admits the representation

$$m_{\mathrm{sc}}[\{z_i \mid i \in B\}] = \int_{-2}^{2} \rho_{\mathrm{sc}}(x) \prod_{i \in B} \frac{1}{(x - z_i)} \, \mathrm{d}x. \tag{2.7}$$

For more details on these notations, see [23, Section 2]. As an example we have

$$\begin{aligned} M(z_1, B_1, z_2) &= \langle B_1 \rangle (m_{\mathrm{sc}}[z_1, z_2] - m_{\mathrm{sc}}(z_1) m_{\mathrm{sc}}(z_2)) + B_1 m_{\mathrm{sc}}(z_1) m_{\mathrm{sc}}(z_2) \\ &= \frac{\langle B_1 \rangle}{2\pi} \int_{-2}^{2} \frac{\sqrt{4 - x^2}}{(x - z_1)(x - z_2)} \, \mathrm{d}x + (B_1 - \langle B_1 \rangle) m_{\mathrm{sc}}(z_1) m_{\mathrm{sc}}(z_2) \end{aligned} \tag{2.8}$$

for any matrix B_1 and

$$\begin{aligned} &M(z_1, A_1, z_2, A_2, z_3) \\ &\quad = \langle A_1 A_2 \rangle (m_{\mathrm{sc}}[z_1, z_3] - m_{\mathrm{sc}}(z_1) m_{\mathrm{sc}}(z_3)) m_{\mathrm{sc}}(z_2) + A_1 A_2 m_{\mathrm{sc}}(z_1) m_{\mathrm{sc}}(z_2) m_{\mathrm{sc}}(z_3) \\ &M(z_1, A_1, z_2, A_2, z_3, A_3, z_4) \\ &\quad = \langle A_1 A_2 A_3 \rangle (m_{\mathrm{sc}}[z_1, z_4] - m_{\mathrm{sc}}(z_1) m_{\mathrm{sc}}(z_4)) m_{\mathrm{sc}}(z_2) m_{\mathrm{sc}}(z_3) \\ &\qquad + A_1 A_2 A_3 m_{\mathrm{sc}}(z_1) m_{\mathrm{sc}}(z_2) m_{\mathrm{sc}}(z_3) m_{\mathrm{sc}}(z_4) \\ &\qquad + A_1 \langle A_2 A_3 \rangle (m_{\mathrm{sc}}[z_2, z_4] - m_{\mathrm{sc}}(z_2) m_{\mathrm{sc}}(z_4)) m_{\mathrm{sc}}(z_1) m_{\mathrm{sc}}(z_3) \\ &\qquad + A_3 \langle A_1 A_2 \rangle (m_{\mathrm{sc}}[z_1, z_3] - m_{\mathrm{sc}}(z_1) m_{\mathrm{sc}}(z_3)) m_{\mathrm{sc}}(z_2) m_{\mathrm{sc}}(z_4) \end{aligned} \tag{2.9}$$

for traceless matrices A_1, A_2, A_3. In the sequel we follow the notational convention that general deterministic matrices are denoted by B, while the letter A is used to denote explicitly traceless matrices.

We now give bounds on the size of the deterministic term $M(z_1, B_1 \ldots, z_k, B_k, z_k)$. The proof of this lemma is presented in Appendix A.

Lemma 2.4. *If a out of the k matrices $B_1, \ldots, B_k$ with $\|B_i\| \lesssim 1$ are traceless, i.e. $\langle B_j \rangle = 0$ holds for a different indices (for some $0 \le a \le k$), then it holds that*

$$\begin{aligned} \|M(z_1, B_1 \ldots, z_k, B_k, z_{k+1})\| &\lesssim \begin{cases} \frac{1}{\eta^{k - \lceil a/2 \rceil}} & d \le 1 \\ \frac{1}{d^{k+1}} & d \ge 1, \end{cases} \\ |\langle M(z_1, B_1, \ldots, z_{k-1}, B_{k-1}, z_k) B_k \rangle| &\lesssim \begin{cases} \frac{1}{\eta^{k - 1 - \lceil a/2 \rceil}} & d \le 1 \\ \frac{1}{d^{k}} & d \ge 1, \end{cases} \end{aligned} \tag{2.10}$$

with $\eta := \min_j |\Im z_j|$ and $d := \min_j \mathrm{dist}(z_j, [-2, 2])$. Generically, both bounds are sharp when not all $\Im z_i$ have the same sign.

Theorem 2.5 (Multi-resolvent local law). *Fix $\epsilon > 0$, let $k \ge 1$ and consider $z_1, \ldots, z_{k+1} \in \mathbf{C}$ with $\max_j |z_j| \le N^{100}$, $\min_j |\Im z_j| \ge N^{-1+\epsilon}$, and let $B_1, \ldots, B_k$ be deterministic matrices of norm $\|B_j\| \lesssim 1$, such that a of them are traceless for some $0 \le a \le k$. Let $\eta := \min_j |\Im z_j|$ and $d := \min_j$ dist$(z_j, [-2, 2])$. Then for arbitrary deterministic vectors $\boldsymbol{x}, \boldsymbol{y}$ of norm $\|\boldsymbol{x}\| + \|\boldsymbol{y}\| \lesssim 1$ we have the* optimal averaged local law

$$|\langle G_1 B_1 \cdots G_k B_k - M(z_1, B_1, \ldots, B_{k-1}, z_k) B_k \rangle| \prec \begin{cases} \frac{1}{N \eta^{k - a/2}} & d \le 1 \\ \frac{1}{N d^{k+1}} & d \ge 1, \end{cases} \tag{2.11a}$$

and the optimal isotropic local law

$$\left|\Big\langle \boldsymbol{x}, \Big(G_1B_1\cdots G_kB_kG_{k+1} - M(z_1,B_1,\dots,B_k,z_{k+1})\Big)\boldsymbol{y}\Big\rangle\right| \prec \begin{cases} \frac{1}{\sqrt{N}\eta^{k-a/2+1/2}} & d\le 1 \\ \frac{1}{\sqrt{N}d^{k+2}} & d\ge 1, \end{cases} \tag{2.11b}$$

where $G_j := G(z_j)$*.*

Remark 2.6. (a) In the regime $d\le 1$ the error terms in (2.11) are generically smaller, by a factor of $1/(N\eta)$ and $1/\sqrt{N\eta}$, respectively, than the leading terms $\langle \boldsymbol{x}, M_{k+1}\boldsymbol{y}\rangle$ and $\langle M_kB_k\rangle$, with the shorthand notation $M_j := M(z_1,B_1,\dots,B_{j-1},z_j)$, c.f. lemma 2.4. For $d\ge 1$ the error terms are smaller by a factor $1/(Nd)$ and $1/(\sqrt{N}d)$, respectively.

(b) The estimates (2.11a) and (2.11b) are optimal. This can be easily seen from the proof since in the Gaussian case the leading term of the variance (4.17) and (4.34) is estimated sharply due to the optimality of lemma 2.4.

(c) The really interesting part of theorem 2.5 is the $d\le 1$ regime, since the effect of traceless matrices is only relevant when at least some of the spectral parameters is close to the limiting spectrum $[-2,2]$. In fact, for $d\ge 1$ very similar bounds were already given in [23, Theorem 3.4]. However, the proof in [23] relied on the fairly involved diagrammatic expansion used in [21, Theorem 4.1]. With our new method, we can give a much shorter alternative proof for this regime as well; this will be explained separately in Appendix B.

(d) With our new method we could also present a simplified proof of the single resolvent local law as stated in theorem 2.3. In this way we could circumvent citing the quite involved [29, Theorem 2.1] that was designed to handle much more general ensembles than Wigner. The proof of the easier $d\ge 1$ regime is especially simple in this new way, which would eliminate the main reason for citing [29] instead of earlier and simpler single resolvent local law proofs for $d\le 1$. For the sake of brevity we refrain from reproving theorem 2.3, and instead we assume it as an input within the proof of theorem 2.5.

By Theorem 2.5 we will also conclude the following corollary.

Corollary 2.7. *Let* $k\ge 3$*, let* $B_1,\dots,B_k$ *be deterministic matrices with* $\|B_i\|\lesssim 1$*, such that* a *of them are traceless for some* $0\le a\le k$*. Let* $f_1,\dots,f_k$ *be Sobolev functions* $f_i\in H^{\lceil k-a/2\rceil}(\mathbf{R})$ *such that* $\|f_i\|_{L^\infty}\lesssim 1$*. Then for any deterministic vectors* $\boldsymbol{x},\boldsymbol{y}$ *with* $\|\boldsymbol{x}\|+\|\boldsymbol{y}\|\lesssim 1$ *we have*

$$\begin{aligned} \langle f_1(W)B_1\dots f_k(W)B_k\rangle &= \sum_{\pi\in\mathrm{NC}[k]}\langle B_1,\dots,B_k\rangle_{K(\pi)}\prod_{B\in\pi}\mathrm{sc}_\circ[B] + \mathcal{O}_\prec\left(\frac{\max_i\|f_i\|_{H^{\lceil k-a/2\rceil}}}{N}\right) \\ \langle \boldsymbol{x}, f_1(W)B_1\dots f_k(W)\boldsymbol{y}\rangle &= \sum_{\pi\in\mathrm{NC}[k]}\langle \boldsymbol{x},\mathrm{pTr}_{K(\pi)}(B_1,\dots,B_{k-1})\boldsymbol{y}\rangle\prod_{B\in\pi}\mathrm{sc}_\circ[B] \\ &\quad + \mathcal{O}_\prec\left(\frac{\max_i\|f_i\|_{H^{\lceil k-a/2\rceil}}}{N^{1/2}}\right), \end{aligned} \tag{2.12}$$

where $\mathrm{sc}_\circ$ *is the free cumulant function from* (2.6) *of* $\mathrm{sc}[i_1,\dots,i_n] := \langle f_{i_1}f_{i_2}\cdots f_{i_n}\rangle_{\mathrm{sc}}$*, with* $\langle f\rangle_{\mathrm{sc}} := \int f(x)\rho_{\mathrm{sc}}(x)\,\mathrm{d}x$*. For* $k=2$ *and* $a=0,1$ *exactly the same result holds. In the remaining case* $k=2$*,* $a=2$ (2.12) *also holds with* $f_i\in H^{\lceil k-a/2\rceil}$ *and* $\|\cdot\|_{H^{\lceil k-a/2\rceil}}$ *replaced by* $f_i\in H^2$ *and* $\|\cdot\|_{H^2}$*, respectively. The results in* (2.12) *can be extended straightforwardly to include several independent Wigner matrices (see [23, Remark 2.13]).*

Exactly the same result (2.12) for $k=1$ and $f\in H^2$ was proven in [22], where we actually even proved a CLT for $\langle f(W)A\rangle$.

We remark that in Corollary 2.7 there is a significant improvement in the error term compared to [23, Theorem 2.6] where the matrices B_j do not necessarily have trace zero. Namely, the Sobolev norm $\|\cdot\|_{H^k}$ in the error term of [23, Theorem 2.6] is here replaced by $\|\cdot\|_{H^{\lceil k-a/2\rceil}}$, with a denoting the number of traceless matrices. For $a=0$ the error terms in Corollary 2.7 coincide with the ones in [23, Theorem 2.6].

Remark 2.8 (Thermalisation)**.** We now specialise Corollary 2.7 to $f(x)=e^{\mathrm{i}sx}$, with $s>0$, and define

$$\varphi(s):=\int_{-2}^{2}e^{\mathrm{i}sx}\rho_{\mathrm{sc}}(x)\,\mathrm{d}x=\frac{J_1(2s)}{s},\tag{2.13}$$

where J_1 is the Bessel function of the first kind. The thermalisation result from [23, Corollaries 2.9-2.10] asserts that the unitary Heisenberg evolution generated by the Wigner matrix renders deterministic observables (matrices) asymptotically independent for large times. More precisely,

$$\langle e^{\mathrm{i}sW}B_1e^{-\mathrm{i}sW}B_2\rangle=\langle B_1\rangle\langle B_2\rangle+\varphi(s)^2\langle B_1B_2\rangle+\mathcal{O}_\prec\left(\frac{s^2}{N}\right),\tag{2.14}$$

for any deterministic matrices B_1,B_2 (for simplicity we only stated the case $k=2$).

Using the optimal local law for two resolvents in (2.11a), by a very similar proof to the one of Corollary 2.7, we conclude

$$\langle e^{\mathrm{i}sW}A_1e^{-\mathrm{i}sW}A_2\rangle=\varphi(s)^2\langle A_1A_2\rangle+\mathcal{O}_\prec\left(\frac{s}{N}\right),\tag{2.15}$$

with $\langle A_1\rangle=\langle A_2\rangle=0$. Note the improved error term in (2.15) compared to s^2N^{-1} from (2.14), which allow us to prove that

$$\langle e^{\mathrm{i}sW}A_1e^{-\mathrm{i}sW}A_2\rangle\approx\varphi(s)^2\langle A_1A_2\rangle$$

for any $s\ll N^{1/4}$ (instead of $s\ll N^{1/5}$ from (2.14)), where we used that $\varphi(s)^2\sim s^{-3}$ for $s\gg 1$. We remark that by Corollary 2.7 we obtain a similar improvement for any $k\geq 3$, but we refrain from stating it for notational simplicity.

3 Proof of the multi-resolvent local law in the $d\leq 1$ regime

We give a detailed proof of Theorem 2.5 for the much more involved $d\leq 1$ regime, in particular in this case $\eta\leq 1$. In Appendix B we explain the necessary modifications for the $d\geq 1$ case. At a certain technical point (within the proof of Lemma 5.1), the proof for the $d\leq 1$ uses (2.11a) for the $d\geq 1$ regime, but this lemma is not needed for the proof in the $d\geq 1$ regime, so our argument is not circular. With the exception of Appendix B, throughout the rest of the paper we assume that $d\leq 1$, hence $\eta\leq 1$.

For traceless deterministic matrices A_j, $\|A_j\|\leq 1$, $\langle A_j\rangle=0$, deterministic bounded vectors $\boldsymbol{x},\boldsymbol{y}$, $\|\boldsymbol{x}\|+\|\boldsymbol{y}\|\leq 1$ and for $k\geq 1$ we introduce the normalized differences

$$\begin{aligned}\Psi_k^{\mathrm{av}}(\boldsymbol{z}_k,\boldsymbol{A}_k)&:=N\eta^{k/2}|\langle G_1A_1\cdots G_kA_k-M(z_1,A_1,\dots,A_{k-1},z_k)A_k\rangle|,\\ \Psi_k^{\mathrm{iso}}(\boldsymbol{z}_{k+1},\boldsymbol{A}_k,\boldsymbol{x},\boldsymbol{y})&:=\sqrt{N\eta^{k+1}}\left|\Big(G_1A_1\cdots A_kG_{k+1}-M(z_1,A_1,\dots,A_k,z_{k+1})\Big)_{\boldsymbol{x}\boldsymbol{y}}\right|,\end{aligned}\tag{3.1}$$

where

$$G_k:=G(z_k),\quad \eta:=\min_i|\Im z_i|,\quad \boldsymbol{z}_k:=(z_1,\dots,z_k),\quad \boldsymbol{A}_k:=(A_1,\dots,A_k).\tag{3.2}$$

For convenience we extend these definitions to $k = 0$ by

$$\Psi_0^{\mathrm{av}}(z) := N\eta|\langle G(z) - m_{\mathrm{sc}}(z)\rangle|, \quad \Psi_0^{\mathrm{iso}}(z, \boldsymbol{x}, \boldsymbol{y}) := \sqrt{N\eta}|\langle \boldsymbol{x}, (G(z) - m_{\mathrm{sc}}(z))\boldsymbol{y}\rangle|, \quad \eta := |\Im z|, \tag{3.3}$$

and note that

$$\Psi_0^{\mathrm{av}} + \Psi_0^{\mathrm{iso}} \prec 1 \tag{3.4}$$

by the well known single-resolvent local law [11,34,40]. Note that the index k counts the number of traceless matrices.

For notational convenience we also introduce the concept of *ϵ-uniform bounds*.

Definition 3.1. *Fix any $\epsilon > 0$. Let $k \in \mathbf{N}$, then we say that the bounds*

$$\begin{aligned} |\langle G(z_1)B_1 \cdots G(z_k)B_k - M(z_1, B_1, \dots, B_{k-1}, z_k)B_k\rangle| &\prec \mathcal{E}^{\mathrm{av}}, \\ \left|\big(G(z_1)B_1 \cdots B_k G(z_{k+1}) - M(z_1, B_1, \dots, z_k, B_k, z_{k+1})\big)_{\boldsymbol{x}\boldsymbol{y}}\right| &\prec \mathcal{E}^{\mathrm{iso}} \end{aligned} \tag{3.5}$$

hold ϵ-uniformly *for some control parameters $\mathcal{E}^{\mathrm{av/iso}} = \mathcal{E}^{\mathrm{av/iso}}(N, \eta)$, depending only on N, η, if the implicit constants in* (3.5) *are uniform in bounded deterministic matrices $\|B_j\| \le 1$, deterministic vectors $\|\boldsymbol{x}\|, \|\boldsymbol{y}\| \le 1$, and spectral parameters z_j with $1 \ge \eta := \min_j |\Im z_j| \ge N^{-1+\epsilon}$, $|z_j| \le N^{100}$. Moreover, we may allow for additional restrictions on the deterministic matrices, and talk about uniformity under the additional assumption that some of the matrices are traceless, or some of them is a multiple of the identity matrix, etc.*

Note that (3.5) is stated for each fixed choice of the spectral parameters z_j in the left hand side, but in fact it is equivalent to an apparently stronger statement, when the same bounds hold with suprema over the spectral parameters z_j. More precisely, if $\mathcal{E}^{\mathrm{av}} \ge N^{-C}$ for some constant C, then (3.5) implies

$$\sup_{z_1, z_2, \dots, z_k} |\langle G(z_1)B_1 \cdots G(z_k)B_k - M(z_1, B_1, \dots, B_{k-1}, z_k)B_k\rangle| \prec \mathcal{E}^{\mathrm{av}} \tag{3.6}$$

(and similarly for the isotropic bound), where the supremum is taken over all choices of z_j's in the admissible spectral domain, i.e. with $|z_j| \le N^{100}$ and $1 \ge \min_j |\Im z_j| \ge N^{-1+\epsilon}$. This bound follows from (3.5) by the usual *grid argument*. Indeed, we may apply (3.5) for a dense N^{-10k}-grid of k-tuples of complex numbers within the spectral domain. The number of such tuples is at most polynomial in N and we use the standard property of stochastic domination to conclude $\max_i X_i \prec C$ from $X_i \prec C$ as long as the number of i's is at most polynomial in N. Finally, we can use the Lipschitz continuity (with Lipschitz constant at most $\eta^{-k-1} \le N^{k+1}$) of the left hand side of (3.5) to extend the bound for all spectral parameters in the spectral domain. In the sequel we will frequently use this equivalence between (3.5) and (3.6), e.g. when we integrate such bounds over some spectral parameter.

We first establish the following key lemma which allows us to conclude multi-resolvent local laws for general deterministic matrices from the special case where each deterministic matrix is traceless.

Lemma 3.2. *Fix $\epsilon > 0$ and $k > 0$ and assume that for all $1 \le j \le k$ and some control parameters $\psi_j^{\mathrm{av/iso}}$ the a priori bounds*

$$\Psi_j^{\mathrm{av}}(\boldsymbol{z}_j, \boldsymbol{A}_j) \prec \psi_j^{\mathrm{av}}, \qquad \Psi_j^{\mathrm{iso}}(\boldsymbol{z}_j, \boldsymbol{A}_j, \boldsymbol{x}, \boldsymbol{y}) \prec \psi_j^{\mathrm{iso}} \tag{3.7}$$

have been established ϵ-uniformly in traceless matrices*. Then it holds that*

$$\begin{aligned}\langle G(z_1)B_1\cdots G(z_k)B_k\rangle &= \langle M(z_1,B_1,\dots,B_{k-1},z_k)B_k\rangle + \mathcal{O}_\prec\left(\frac{\sum_{j=a}^{k-b}\psi_j^{\mathrm{av}}}{N\eta^{k-a/2}}\right)\\ \Big(G(z_1)B_1G(z_2)\cdots B_kG(z_{k+1})\Big)_{\boldsymbol{xy}} &= M(z_1,B_1,\dots,B_k,z_{k+1})_{\boldsymbol{xy}} + \mathcal{O}_\prec\left(\frac{\sum_{j=a}^{k-b}\psi_j^{\mathrm{iso}}}{\sqrt{N}\eta^{k-a/2+1/2}}\right),\end{aligned} \tag{3.8}$$

ϵ*-uniformly in vectors* $\boldsymbol{x},\boldsymbol{y}$ *and deterministic matrices* $B_1,\dots,B_k$*, out of which* $0\le a\le k$ *are traceless and* $0\le b\le k$ *are a multiple of the identity.*

Using Lemma 3.2 we reduce Theorem 2.5 to the following Lemma.

Lemma 3.3 (Final estimate on $\Psi_k^{\mathrm{av/iso}}$)**.** *For any* $\epsilon>0$ *and* $k\ge 1$ *we have*

$$\Psi_k^{\mathrm{av}} + \Psi_k^{\mathrm{iso}} \prec 1 \tag{3.9}$$

ϵ*-uniformly in traceless matrices.*

Proof of Theorem 2.5. Theorem 2.5 is equivalent to Lemma 3.3 in case when all matrices are traceless. The general case follows from Lemma 3.2 and setting $\psi_k^{\mathrm{av/iso}}=1$ due to Lemma 3.3. □

We prove Lemma 3.3 in two steps and first establish a weaker bound as stated in the following lemma.

Lemma 3.4 (A priori estimate on $\Psi_k^{\mathrm{av/iso}}$)**.** *For any* $\epsilon>0$ *and* $k\ge 1$ *we have*

$$\Psi_k^{\mathrm{av}} + \Psi_k^{\mathrm{iso}} \prec \sqrt{N\eta} \tag{3.10}$$

ϵ*-uniformly in traceless matrices.*

The rest of the proof is organised as follows: First, we prove Lemma 3.2, then in Section 3.1 we state the *master inequalities* on the $\Psi_k^{\mathrm{av/iso}}$ parameters, which we then use to prove Lemmas 3.3 and 3.4 in Section 3.2. Finally, the proof of the master inequalities will be presented in Section 4.

Proof of Lemma 3.2. We start the proof by splitting all those $k-a-b$ matrices B_i that are neither traceless nor multiples of the identity as $B_i=\langle B_i\rangle+\mathring{B}_i$. Since (2.4) is multi-linear in the B-matrices and the error terms in (3.8) are monotonically decreasing as a or b are increased, it is sufficient to prove Lemma 3.2 for the special case when $a+b=k$, i.e. all matrices are either traceless or multiple of the identity.

Moreover, if $\Im z_i\Im z_j<0$ then we use the resolvent identity $G(z_i)G(z_j)=[G(z_i)-G(z_j)]/(z_i-z_j)$ and $|z_i-z_j|\ge\eta$ repeatedly to further reduce the lemma to the special case

$$\left(\prod_{j=1}^{k_1}G(z_{1,j})\right)A_1\left(\prod_{j=1}^{k_2}G(z_{2,j})\right)A_2\cdots \tag{3.11}$$

where $\langle A_i\rangle=0$ and $\mathrm{sgn}(\Im z_{i,1})=\cdots=\mathrm{sgn}(\Im z_{i,k_i})$ for all i. We note that (2.4) satisfies the same relation since

$$M(\dots,z_i,I,z_{i+1},\dots)=\frac{M(\dots,z_i,\dots)-M(\dots,z_{i+1},\dots)}{z_i-z_{i+1}} \tag{3.12}$$

due to

$$m[z_i,z_{i+1}]=\frac{m_{\mathrm{sc}}(z_i)-m_{\mathrm{sc}}(z_{i+1})}{z_i-z_{i+1}} \tag{3.13}$$

by definition. Finally, from the residue theorem we have that

$$\prod_{j=1}^{k} G(z_j) = \frac{1}{\pi} \int_{\mathbf{R}} \Im G(x+\mathrm{i}\eta) \prod_{j=1}^{k} \frac{1}{x - z_j + \operatorname{sgn}(\Im z_j)\mathrm{i}\eta} \, \mathrm{d}x \tag{3.14}$$

whenever $0 < \eta < \min_j \Im z_j$ or $\max_j \Im z_j < -\eta < 0$. We note that M from (2.4) satisfies the same relation since

$$M(\dots, z_i, I, z_{i+1}, I, \dots, I, z_{i+n}, \dots) = \frac{1}{2\pi\mathrm{i}} \int_{\mathbf{R}} \frac{M(\dots, x+\mathrm{i}\eta, \dots) - M(\dots, x-\mathrm{i}\eta, \dots)}{(x+\sigma\mathrm{i}\eta - z_i)\cdots(x+\sigma\mathrm{i}\eta - z_{i+n})} \, \mathrm{d}x \tag{3.15}$$

for $\sigma = \operatorname{sgn}(\Im z_i) = \dots = \operatorname{sgn}(\Im z_{i+n})$ due to multi-linearity and

$$m[z_i, \dots, z_{i+n}] = \frac{1}{2\pi\mathrm{i}} \int_{\mathbf{R}} \frac{m(x+\mathrm{i}\eta) - m(x-\mathrm{i}\eta)}{(x+\sigma\mathrm{i}\eta - z_1)\cdots(x+\sigma\mathrm{i}\eta - z_n)} \, \mathrm{d}x \tag{3.16}$$

from the residue theorem. By using (3.14) for each product in (3.11) obtain an alternating chain of traceless matrices and resolvents, so that the bound follows by the assumptions in (3.7). □

3.1 Master inequalities and reduction lemma

From now on every deterministic matrix A_i is assumed to be traceless and uniformity is understood as uniformity in traceless matrices.

Proposition 3.5 (A priori estimates on $\Psi^{\mathrm{av/iso}}$). *(i) Assume that*

$$\Psi_j^{\mathrm{av/iso}} \prec \psi_j^{\mathrm{av/iso}}, \qquad 1 \le j \le 4 \tag{3.17}$$

uniformly. Then it holds that

$$\Psi_1^{\mathrm{av}} \prec 1 + \frac{\psi_1^{\mathrm{iso}} + (\psi_1^{\mathrm{av}})^{1/2} + (\psi_2^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}} \tag{3.18a}$$

$$\Psi_2^{\mathrm{av}} \prec 1 + \psi_1^{\mathrm{av}} + \frac{\psi_2^{\mathrm{iso}} + (\psi_2^{\mathrm{av}})^{1/2} + (\psi_4^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}} + \frac{(\psi_1^{\mathrm{iso}})^2 + (\psi_1^{\mathrm{av}})^2 + \psi_1^{\mathrm{iso}}(\psi_2^{\mathrm{av}})^{1/2}}{N\eta} \tag{3.18b}$$

$$\Psi_1^{\mathrm{iso}} \prec 1 + \frac{\psi_1^{\mathrm{iso}} + \psi_1^{\mathrm{av}}}{\sqrt{N\eta}} + \frac{(\psi_2^{\mathrm{iso}})^{1/2}}{(N\eta)^{1/4}} \tag{3.18c}$$

$$\Psi_2^{\mathrm{iso}} \prec 1 + \psi_1^{\mathrm{iso}} + \frac{\psi_2^{\mathrm{iso}} + (\psi_1^{\mathrm{iso}}\psi_3^{\mathrm{iso}})^{1/2} + \psi_1^{\mathrm{av}}}{\sqrt{N\eta}} + \frac{\psi_1^{\mathrm{iso}}\psi_1^{\mathrm{av}}}{N\eta} + \frac{(\psi_3^{\mathrm{iso}})^{1/2} + (\psi_4^{\mathrm{iso}})^{1/2}}{(N\eta)^{1/4}}, \tag{3.18d}$$

again uniformly.

(ii) Now, let $k > 2$ and assume that a priori bounds

$$\Psi_j^{\mathrm{av}} \prec \begin{cases} \psi_j^{\mathrm{av}} := \sqrt{N\eta}, & j \le k-2, \\ \psi_j^{\mathrm{av}}, & k-1 \le j \le 2k, \end{cases}$$
$$\Psi_j^{\mathrm{iso}} \prec \begin{cases} \psi_j^{\mathrm{iso}} := \sqrt{N\eta}, & j \le k-2, \\ \psi_j^{\mathrm{iso}}, & k-1 \le j \le 2k, \end{cases} \tag{3.19}$$

have been established uniformly. Then it holds that

$$\Psi_k^{\mathrm{av}} \prec 1 + \sum_{j=1}^{k-1} \psi_j^{\mathrm{av}} + \frac{\psi_{k-1}^{\mathrm{iso}} + \psi_k^{\mathrm{iso}} + \sum_{j=\lceil k/2\rceil}^{k} (\psi_{2j}^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}} \tag{3.20a}$$

$$\Psi_k^{\mathrm{iso}} \prec 1 + \sum_{j=1}^{k-1} \psi_j^{\mathrm{iso}} + \frac{\psi_k^{\mathrm{iso}} + (\psi_{k+1}^{\mathrm{iso}}\psi_{k-1}^{\mathrm{iso}})^{1/2} + \psi_{k-1}^{\mathrm{av}}}{\sqrt{N\eta}} + \frac{\sum_{j=k+1}^{2k} (\psi_j^{\mathrm{iso}})^{1/2}}{(N\eta)^{1/4}} \tag{3.20b}$$

again uniformly.

Since in Proposition 3.5 resolvent chains of length k are estimated by resolvent chains of length up to $2k$ we will need the following *reduction lemma* in order avoid an infinite hierarchy of inequalities with higher and higher k-indices.

Lemma 3.6 (Reduction inequality)**.** *Fix $k \geq 1$ and assume that $\Psi_n^{\mathrm{av/iso}} \prec \psi_n^{\mathrm{av/iso}}$ holds for $0 \leq n \leq 2k$ uniformly (in the sense explained in Proposition 3.5). Then it holds that*

$$\Psi_{2k}^{\mathrm{av}} \prec \begin{cases} (N\eta)^2 + (\psi_k^{\mathrm{av}})^2, & k \textit{ even} \\ (N\eta)^2 + N\eta(\psi_{k-1}^{\mathrm{av}} + \psi_{k+1}^{\mathrm{av}}) + \psi_{k-1}^{\mathrm{av}}\psi_{k+1}^{\mathrm{av}}, & k \textit{ odd.} \end{cases} \tag{3.21}$$

Moreover, for $j \leq k$ and for k even, we have

$$\Psi_{k+j}^{\mathrm{iso}} \prec N\eta\Big(1 + \frac{\psi_k^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\Big(1 + \frac{(\psi_{2j}^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}}\Big), \tag{3.22}$$

again uniformly.

The proofs of Proposition 3.5 and Lemma 3.6 will be given in Section 4 and Section 5, respectively.

3.2 Proof of the bounds on $\Psi^{\mathrm{av/iso}}$ in Lemmas 3.3 and 3.4

Proof of Lemma 3.4. Within the proof we repeatedly appeal to a simple argument we call *iteration*. By this we mean that whenever $X \prec x$ implies

$$X \prec A + \frac{x}{B} + x^{1-\alpha}C^{\alpha}, \tag{3.23}$$

for some constants $B \geq N^{\delta}$, $A, C > 0$, and exponent $0 < \alpha < 1$, and we know that $X \prec N^D$ initially (here δ, α and D are N-independent positive constants, other quantities may depend on N) then we can iterate (3.23) finitely many times (depending only on δ, α and D) until we arrive at

$$X \prec A + C. \tag{3.24}$$

In other words, (3.23) implies (3.24).

The proof of Lemma 3.4 is a two-step induction on k. Our first step is to establish the induction hypothesis

$$\Psi_1^{\mathrm{av/iso}} \prec 1 \leq \sqrt{N\eta}, \quad \Psi_2^{\mathrm{av/iso}} \prec \sqrt{N\eta}, \tag{3.25}$$

in fact for $\Psi_1^{\mathrm{av/iso}}$ we will establish the stronger $\prec 1$ bound immediately. We start with (3.18b) which together with (3.21) implies

$$\begin{aligned} \Psi_2^{\mathrm{av}} &\prec 1 + \psi_1^{\mathrm{av}} + \frac{\psi_2^{\mathrm{iso}} + (\psi_2^{\mathrm{av}})^{1/2} + (\psi_4^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}} + \frac{(\psi_1^{\mathrm{iso}})^2 + (\psi_1^{\mathrm{av}})^2 + \psi_1^{\mathrm{iso}}(\psi_2^{\mathrm{av}})^{1/2}}{N\eta} \\ &\prec \sqrt{N\eta} + \psi_1^{\mathrm{av}} + \frac{\psi_2^{\mathrm{iso}} + \psi_2^{\mathrm{av}}}{\sqrt{N\eta}} + \frac{(\psi_1^{\mathrm{iso}})^2 + (\psi_1^{\mathrm{av}})^2 + \psi_1^{\mathrm{iso}}(\psi_2^{\mathrm{av}})^{1/2}}{N\eta} \end{aligned} \tag{3.26}$$

and hence, using iteration and a Schwarz inequality $\psi_1^{\mathrm{iso}}(\psi_2^{\mathrm{av}})^{1/2} \leq (\psi_1^{\mathrm{iso}})^2 + \psi_2^{\mathrm{av}}$ for the last term, we get

$$\Psi_2^{\mathrm{av}} \prec \sqrt{N\eta} + \psi_1^{\mathrm{av}} + \frac{\psi_2^{\mathrm{iso}}}{\sqrt{N\eta}} + \frac{(\psi_1^{\mathrm{iso}})^2 + (\psi_1^{\mathrm{av}})^2}{N\eta}. \tag{3.27}$$

Next, we consider (3.18d) and eliminate $\psi_3^{\mathrm{iso}}, \psi_4^{\mathrm{iso}}$ from it by first using (3.21) and (3.22) in the form

$$\begin{aligned}
\Psi_3^{\mathrm{iso}} &\prec N\eta\Big(1+\frac{\psi_2^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\Big(1+\frac{\psi_2^{\mathrm{av}}}{N\eta}\Big)^{1/2}\\
&\prec N\eta\Big(1+\frac{\psi_2^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\Big(1+\frac{\psi_2^{\mathrm{iso}}}{(N\eta)^{3/2}}+\frac{(\psi_1^{\mathrm{av}})^2+(\psi_1^{\mathrm{iso}})^2}{(N\eta)^2}\Big)^{1/2},\\
\Psi_4^{\mathrm{iso}} &\prec N\eta\Big(1+\frac{\psi_2^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\Big(1+\frac{\psi_4^{\mathrm{av}}}{N\eta}\Big)^{1/2} \prec (N\eta)^{3/2}\Big(1+\frac{\psi_2^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\Big(1+\frac{\psi_2^{\mathrm{av}}}{N\eta}\Big)\\
&\prec (N\eta)^{3/2}\Big(1+\frac{\psi_2^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\Big(1+\frac{\psi_2^{\mathrm{iso}}}{(N\eta)^{3/2}}+\frac{(\psi_1^{\mathrm{av}})^2+(\psi_1^{\mathrm{iso}})^2}{(N\eta)^2}\Big),
\end{aligned} \tag{3.28}$$

where in the second step we also eliminated ψ_2^{av} using (3.27). Plugging these bounds into (3.18d) yields

$$\begin{aligned}
\Psi_2^{\mathrm{iso}} &\prec 1+\psi_1^{\mathrm{iso}}+\frac{\psi_2^{\mathrm{iso}}+(\psi_1^{\mathrm{iso}}\psi_3^{\mathrm{iso}})^{1/2}+\psi_1^{\mathrm{av}}}{\sqrt{N\eta}}+\frac{\psi_1^{\mathrm{iso}}\psi_1^{\mathrm{av}}}{N\eta}+\frac{(\psi_3^{\mathrm{iso}})^{1/2}+(\psi_4^{\mathrm{iso}})^{1/2}}{(N\eta)^{1/4}}\\
&\prec \sqrt{N\eta}+\psi_1^{\mathrm{iso}}+\frac{\psi_2^{\mathrm{iso}}}{\sqrt{N\eta}}+\frac{(\psi_1^{\mathrm{iso}})^2+(\psi_1^{\mathrm{av}})^2}{N\eta}+\sqrt{\psi_2^{\mathrm{iso}}}(N\eta)^{1/4}\Big(1+\frac{\psi_1^{\mathrm{av}}+\psi_1^{\mathrm{iso}}}{N\eta}\Big).
\end{aligned} \tag{3.29}$$

By iteration we thus obtain

$$\Psi_2^{\mathrm{iso}} \prec \sqrt{N\eta}+\psi_1^{\mathrm{iso}}+\frac{(\psi_1^{\mathrm{iso}})^2+(\psi_1^{\mathrm{av}})^2}{N\eta}, \tag{3.30}$$

and by feeding (3.30) back into (3.27) we conclude

$$\Psi_2^{\mathrm{av}} \prec \sqrt{N\eta}+\psi_1^{\mathrm{av}}+\frac{(\psi_1^{\mathrm{iso}})^2+(\psi_1^{\mathrm{av}})^2}{N\eta}. \tag{3.31}$$

By using (3.30) in (3.18c) we immediately obtain

$$\Psi_1^{\mathrm{iso}} \prec 1+\frac{\psi_1^{\mathrm{iso}}+\psi_1^{\mathrm{av}}}{\sqrt{N\eta}}+\frac{\psi_1^{\mathrm{iso}}+\psi_1^{\mathrm{av}}}{(N\eta)^{3/4}} \prec 1+\frac{\psi_1^{\mathrm{av}}}{(N\eta)^{1/2}} \tag{3.32}$$

and together with (3.18a) we also have that

$$\Psi_1^{\mathrm{av}} \prec 1+\frac{(\psi_2^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}}. \tag{3.33}$$

Finally, by combining (3.31) to (3.33) we obtain

$$\Psi_2^{\mathrm{av}} \prec \sqrt{N\eta}+\frac{(\psi_2^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}}+\frac{\psi_2^{\mathrm{av}}}{(N\eta)^2} \prec \sqrt{N\eta} \tag{3.34}$$

and therefore $\Psi_1^{\mathrm{av/iso}} \prec 1$ and finally, by (3.30), all statements in the claim (3.25) hold. This completes the initial step of the induction.

Now we turn to the induction step: we assume that $k \geq 4$ is even and that the bounds

$$\Psi_n^{\mathrm{av/iso}} \prec \sqrt{N\eta}, \qquad n \leq k-2 \tag{3.35}$$

have already been proved. We will prove the same bounds for $n = k-1, k$.

For any $j \le k$ and under the assumption (3.35) the reduction inequalities (3.21) and (3.22) simplify (recall that k is even) to

$$\begin{aligned}\Psi_{k+j}^{\mathrm{iso}} &\prec N\eta\Big(1+\frac{\psi_k^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\Big(1+\frac{\psi_{2j}^{\mathrm{av}}}{N\eta}\Big)^{1/2}\\ &\prec N\eta\Big(1+\frac{\psi_k^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\begin{cases}\sqrt{N\eta}+\frac{\psi_j^{\mathrm{av}}}{\sqrt{N\eta}}, & j \text{ even},\\ \sqrt{N\eta}+\sqrt{\psi_{j-1}^{\mathrm{av}}+\psi_{j+1}^{\mathrm{av}}+\psi_{j-1}^{\mathrm{iso}}\psi_{j+1}^{\mathrm{av}}/N\eta}, & j \text{ odd},\end{cases}\\ &\prec (N\eta)^{3/2}\Big(1+\frac{\psi_k^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\Big(1+\frac{\psi_k^{\mathrm{av}}}{N\eta}\Big)^{\mathbf{1}(j=k)+\mathbf{1}(j=k-1)/2} \prec (N\eta)^{3/2}\Big(1+\frac{\psi_k^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\Big(1+\frac{\psi_k^{\mathrm{av}}}{N\eta}\Big)\end{aligned} \tag{3.36}$$

and

$$\Psi_{2j}^{\mathrm{av}} \prec \left.\begin{cases}(N\eta)^2+(\psi_k^{\mathrm{av}})^2, & j=k,\\ (N\eta)^2+N\eta\psi_k^{\mathrm{av}}, & j=k-1,\\ (N\eta)^2, & \text{else},\end{cases}\right\} \prec (N\eta)^2+(\psi_k^{\mathrm{av}})^2. \tag{3.37}$$

Then together with (3.20a) and (3.20b) it follows that

$$\begin{aligned}\Psi_{k-1}^{\mathrm{av}} &\prec \sqrt{N\eta}+\frac{\psi_{k-1}^{\mathrm{iso}}+\sum_{j=k/2}^{k-1}(\psi_{2j}^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}} \prec \sqrt{N\eta}+\frac{\psi_{k-1}^{\mathrm{iso}}+\psi_k^{\mathrm{av}}}{\sqrt{N\eta}}\\ \Psi_{k-1}^{\mathrm{iso}} &\prec \sqrt{N\eta}+\frac{\sum_{j=k}^{2k-2}(\psi_j^{\mathrm{iso}})^{1/2}}{(N\eta)^{1/4}} \prec \sqrt{N\eta}\Big(1+\frac{\psi_k^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)^{1/2}\Big(1+\frac{\psi_k^{\mathrm{av}}}{N\eta}\Big)^{1/2}\end{aligned} \tag{3.38}$$

and

$$\begin{aligned}\Psi_k^{\mathrm{av}} &\prec \sqrt{N\eta}+\psi_{k-1}^{\mathrm{av}}+\frac{\psi_{k-1}^{\mathrm{iso}}+\psi_k^{\mathrm{iso}}+\sum_{j=k/2}^{k}(\psi_{2j}^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}}\\ &\prec \sqrt{N\eta}+\psi_{k-1}^{\mathrm{av}}+\frac{\psi_{k-1}^{\mathrm{iso}}+\psi_k^{\mathrm{iso}}+\psi_k^{\mathrm{av}}}{\sqrt{N\eta}}\\ \Psi_k^{\mathrm{iso}} &\prec \sqrt{N\eta}+\psi_{k-1}^{\mathrm{iso}}+\frac{\psi_{k-1}^{\mathrm{av}}+(\psi_{k-1}^{\mathrm{iso}}\psi_{k+1}^{\mathrm{iso}})^{1/2}}{\sqrt{N\eta}}+\frac{\sum_{j=k+1}^{2k}(\psi_j^{\mathrm{iso}})^{1/2}}{(N\eta)^{1/4}}\\ &\prec \sqrt{N\eta}+\psi_{k-1}^{\mathrm{iso}}+\frac{\psi_{k-1}^{\mathrm{av}}+\psi_k^{\mathrm{iso}}}{\sqrt{N\eta}}+\sqrt{N\eta}\Big(1+\frac{\psi_k^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)^{1/2}\Big(1+\frac{\psi_k^{\mathrm{av}}}{N\eta}\Big)^{1/2}\end{aligned} \tag{3.39}$$

where we used the first inequality of (3.36) to estimate $\psi_{k+1}^{\mathrm{iso}}$ in the $\sqrt{\psi_{k-1}^{\mathrm{iso}}\psi_{k+1}^{\mathrm{iso}}}$-term with $\psi_2^{\mathrm{av}} = \sqrt{N\eta}$. Iterating (3.39) yields

$$\begin{aligned}\Psi_k^{\mathrm{av}} &\prec \sqrt{N\eta}+\psi_{k-1}^{\mathrm{av}}+\frac{\psi_{k-1}^{\mathrm{iso}}+\psi_k^{\mathrm{iso}}}{\sqrt{N\eta}}\\ \Psi_k^{\mathrm{iso}} &\prec \sqrt{N\eta}+\psi_{k-1}^{\mathrm{iso}}+\frac{\psi_{k-1}^{\mathrm{av}}+\psi_k^{\mathrm{av}}}{\sqrt{N\eta}},\end{aligned} \tag{3.40}$$

and by using (3.38) in (3.40) it follows that

$$\begin{aligned}\Psi_k^{\mathrm{av}} &\prec \sqrt{N\eta}+\frac{\psi_k^{\mathrm{iso}}+\psi_k^{\mathrm{av}}}{\sqrt{N\eta}}+\Big(1+\frac{\psi_k^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)^{1/2}\Big(1+\frac{\psi_k^{\mathrm{av}}}{N\eta}\Big)^{1/2}\\ &\prec \sqrt{N\eta}+\frac{\psi_k^{\mathrm{iso}}}{\sqrt{N\eta}}\\ \Psi_k^{\mathrm{iso}} &\prec \sqrt{N\eta}\Big(1+\frac{\psi_k^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)^{1/2}\Big(1+\frac{\psi_k^{\mathrm{av}}}{N\eta}\Big)^{1/2}+\frac{\psi_k^{\mathrm{av}}}{\sqrt{N\eta}}\\ &\prec \sqrt{N\eta}+\sqrt{\psi_k^{\mathrm{av}}}+\frac{\psi_k^{\mathrm{av}}}{\sqrt{N\eta}} \prec \sqrt{N\eta}+\frac{\psi_k^{\mathrm{av}}}{\sqrt{N\eta}}.\end{aligned} \tag{3.41}$$

From (3.41) we immediately conclude $\Psi_k^{\mathrm{av/iso}} \prec \sqrt{N\eta}$ and by feeding this back into (3.39) finally that

$$\Psi_{k-1}^{\mathrm{av/iso}} + \Psi_k^{\mathrm{av/iso}} \prec \sqrt{N\eta}, \tag{3.42}$$

concluding the induction step. □

Proof of Lemma 3.3. This follows directly from Lemma 3.4 and proposition 3.5 and induction on k. □

4 Proof of the master inequalities, Proposition 3.5

We recall the definition of the *second order renormalisation*, denoted by underlining, from [21]. For matrix-valued functions $f(W), g(W)$ of the random matrix W we define

$$\underline{f(W)Wg(W)} := f(W)Wg(W) - \mathbf{E}_{\widetilde{W}}\Big[(\partial_{\widetilde{W}} f)(W)\widetilde{W}g(W) + f(W)\widetilde{W}(\partial_{\widetilde{W}} g)(W)\Big], \tag{4.1}$$

where $\partial_{\widetilde{W}}$ denotes the directional derivative in the direction of a GUE matrix $\widetilde{W}$ that is independent of W. The expectation is w.r.t. this GUE matrix. Note that if W itself is a GUE matrix, then $\mathbf{E}\,\underline{f(W)Wg(W)} = 0$, while for W with a general distribution this expectation is independent of the first two moments of W; in other words the underline renormalises $f(W)Wg(W)$ up to second order. We note that underline in (4.1) is a well-defined notation only when the position of the "middle" W to which the renormalisation refers is unambiguous. This is the case in all of our proof since f, g will be products of resolvents not explicitly involving monomials of W.

We also note that the directional derivative of the resolvent is given by

$$\partial_{\widetilde{W}} G = -G\widetilde{W}G, \tag{4.2}$$

furthermore, we have

$$\mathbf{E}_{\widetilde{W}}\, \widetilde{W}A\widetilde{W} = \langle A\rangle \cdot I. \tag{4.3}$$

For example, in case of $f = I$ and $g(W) = (W-z)^{-1} = G$ we have

$$\underline{WG} = WG + \langle G\rangle G.$$

Similarly, for $G_i = G(z_i)$ we also have

$$\underline{WG_1G_2} = WG_1G_2 + \langle G_1\rangle G_1G_2 + \langle G_1G_2\rangle G_2, \quad \underline{G_1WG_2} = G_1WG_2 + \langle G_1\rangle G_1G_2 + \langle G_2\rangle G_1G_2$$

indicating that the definition of the underline in (4.1) depends on the "left" and "right" functions f and g, and even though $f(W)Wg(W) = Wf(W)g(W) = f(W)g(W)W$, their second order renormalisations are not the same.

Using this underline notation and the defining equation for $m = m_{\mathrm{sc}}$, we have

$$G = m - m\underline{WG} + m\langle G - m\rangle G = m - m\underline{GW} + m\langle G - m\rangle G. \tag{4.4}$$

The key idea of the proof of Proposition 3.5 is using (4.4) for some G_j in $G_1A_1 \ldots A_{k-1}G_k$ and extending the renormalisation to the whole product at the expense adding resolvent products of lower order. For example,

$$\begin{aligned} &G_1A_1G_2A_2G_3\Big(1 + \mathcal{O}\Big(\frac{1}{N\eta}\Big)\Big) \\ &= m_2G_1A_1A_2G_3 - m_2G_1A_1\underline{WG_2}A_2G_3 \\ &= m_2\Big(G_1A_1A_2G_3 + \langle G_1A_1\rangle G_1G_2A_2G_3 + G_1A_1G_3\langle G_2A_2G_3\rangle - \underline{G_1A_1WG_2A_2G_3}\Big), \end{aligned} \tag{4.5}$$

where on the rhs. only products of resolvent with one deterministic matrix need to be understood. The renormalisation of the whole product will be handled by cumulant expansion exploiting that its expectation vanishes up to second order. We note that while $\underline{WG} = \underline{GW}$, replacing G_2 by $m_2 - m_2\underline{G_2W}$ instead of $m_2 - m_2\underline{WG_2}$ in (4.5) still gives a slightly different expression:

$$\begin{aligned}&G_1A_1G_2A_2G_3\Big(1+\mathcal{O}\Big(\frac{1}{N\eta}\Big)\Big)\\&= m_2G_1A_1A_2G_3 - m_2G_1A_1\underline{G_2W}A_2G_3\\&= m_2\Big(G_1A_1A_2G_3 + \langle G_1A_1G_2\rangle G_1A_2G_3 + G_1A_1G_2G_3\langle A_2G_3\rangle - \underline{G_1A_1G_2WA_2G_3}\Big).\end{aligned} \tag{4.6}$$

A key ingredient for the proof is the following lemma which shows that the deterministic approximation M defined in (2.4) satisfies the same recursive relations as suggested by (4.5) and (4.6) after ignoring the full underline term and the $1/(N\eta)$ error terms.

Lemma 4.1. *Let $z_1,\dots,z_k$ by spectral parameters, and $A_1,\dots,A_{k-1}$ be deterministic matrices. Then for any $1\le j\le k$ we have the relations*

$$\begin{aligned}M(z_1,\dots,z_k) &= m_jM(z_1,\dots,z_{j-1},A_{j-1}A_j,z_{j+1},\dots,z_k)\\&\quad + m_j\sum_{l=1}^{j-1}M(z_1,\dots,A_{l-1},z_l,I,z_j,A_j,\dots,z_k)\langle M(z_l,A_l,\dots,z_{j-1})A_{j-1}\rangle\\&\quad + m_j\sum_{l=j+1}^{k}M(z_1,\dots,A_{j-1},z_l,A_l,\dots,z_k)\langle M(z_j,A_j,\dots,z_l)\rangle,\end{aligned} \tag{4.7}$$

and

$$\begin{aligned}M(z_1,\dots,z_k) &= m_jM(z_1,\dots,z_{j-1},A_{j-1}A_j,z_{j+1},\dots,z_k)\\&\quad + m_j\sum_{l=1}^{j-1}M(z_1,\dots,A_{l-1},z_l,A_j,\dots,z_k)\langle M(z_l,A_l,\dots,z_j)\rangle\\&\quad + m_j\sum_{l=j+1}^{k}M(z_1,\dots,A_{j-1},z_j,I,z_l,A_l\dots,z_k)\langle M(z_l,A_j,\dots,z_{l-1})A_{l-1}\rangle.\end{aligned} \tag{4.8}$$

We remark that the special $j=1$ case of this lemma was already proven in [23, Lemma 5.4]. We will present a direct combinatorial proof for the general case in appendix A. Alternatively, lemma 4.1 can also be deduced from the original expansions for resolvent products with the full underline term. For example, taking the expectation of (4.5) for W being a GUE matrix and letting $N\to\infty$ removes the full underline term and the error terms. Since the local law [23, Theorem 3.4] asserts that $G_1A_1G_2A_2G_3$ asymptotically equals $M(z_1,A_1,z_2,A_2,z_3)$ in the $N\to\infty$ limit for any fixed spectral parameters, we obtain the corresponding identity (4.7) for $k=3$. The argument for general k is identical.

4.1 Proof of Proposition 3.5

The proofs of the averaged and isotropic bounds are done separately below. For simplicity we do not carry the dependence on the spectral parameters z_j and traceless matrices A_j but instead simply write G and A.

4.1.1 Averaged bounds (3.18a), (3.18b) and (3.20a)

Within the proof we repeatedly make use of the a priori bounds (3.17) and (3.19) for $j \le 2k$. It is important to stress that after possibly applying Lemma 3.2 no chains of length more than $2k$ arise along our expansion hence the a priori bounds are needed up to index $2k$ only.

By (4.4) for the first G and using the local law $|\langle G - m\rangle| \prec 1/(N\eta)$ we obtain

$$\begin{aligned}&\langle (GA)^k\rangle\Big(1+\mathcal{O}_\prec\big((N\eta)^{-1}\big)\Big)\\ &\quad= m\langle A(GA)^{k-1}\rangle - m\langle \underline{WGA}(GA)^{k-1}\rangle\\ &\quad= m\langle A(GA)^{k-1}\rangle + m\sum_{j=1}^{k-1}\langle (GA)^j G\rangle\langle (GA)^{k-j}\rangle - m\langle \underline{W(GA)^k}\rangle.\end{aligned} \tag{4.9}$$

By assumption (3.17) and (3.19) and Lemma 3.2 we have

$$\begin{aligned}\big|\langle A(GA)^{k-1}\rangle - \langle AM_{k-1}A\rangle\big| &\prec \frac{\psi^{\mathrm{av}}_{k-2}}{N\eta^{k/2}} + \frac{\psi^{\mathrm{av}}_{k-1}}{N\eta^{k/2-1/2}} \lesssim \frac{\psi^{\mathrm{av}}_{k-1}+\psi^{\mathrm{av}}_{k-2}}{N\eta^{k/2}}\\ |\langle (GA)^j G - M_{j+1}\rangle| &\prec \frac{\psi^{\mathrm{av}}_j}{N\eta^{j/2+1}},\\ \big|\langle (GA)^{k-j}\rangle - \langle M_{k-j}A\rangle\big| &\prec \frac{\psi^{\mathrm{av}}_{k-j}}{N\eta^{(k-j)/2}},\end{aligned} \tag{4.10}$$

so we can replace each resolvent chain by its deterministic M-value plus the error term. In particular, for the middle term in the third line of (4.9) by a telescopic summation we have

$$\begin{aligned}&\left|\sum_{j=1}^{k}\Big(\langle (GA)^j G\rangle\langle (GA)^{k-j}\rangle - \sum_{j=1}^{k}\langle M_{j+1}\rangle\langle M_{k-j}A\rangle\Big)\right|\\ &\prec \sum_{j=2}^{k-1}\frac{1}{\eta^{\lfloor j/2\rfloor}}\frac{\psi^{\mathrm{av}}_{k-j}}{N\eta^{(k-j)/2}} + \sum_{j=1}^{k-4}\frac{1}{\eta^{\lfloor (k-j)/2\rfloor}}\frac{\psi^{\mathrm{av}}_j}{N\eta^{j/2}} + \frac{\psi^{\mathrm{av}}_{k-2}}{N\eta^{k/2}} + \frac{\psi^{\mathrm{av}}_{k-3}}{N\eta^{k/2-1/2}} + \sum_{j=1}^{k-1}\frac{\psi^{\mathrm{av}}_j\psi^{\mathrm{av}}_{k-j}}{N^2\eta^{k/2+1}}\\ &\lesssim \frac{1}{N\eta^{k/2}}\left(\psi^{\mathrm{av}}_{k-2} + \sum_{j=1}^{k-1}\psi^{\mathrm{av}}_j\Big(1+\frac{\psi^{\mathrm{av}}_{k-j}}{N\eta}\Big)\right),\end{aligned} \tag{4.11}$$

where we used that by assumption $\eta \lesssim 1$, the bounds (2.10) and $\langle M_2\rangle = \langle M_1 A\rangle = 0$. Together with the deterministic identity (4.7) we conclude

$$\big(\langle (GA)^k\rangle - \langle M_k A\rangle\big)\Big(1+\mathcal{O}_\prec\big((N\eta)^{-1}\big)\Big) = -m\langle \underline{W(GA)^k}\rangle + \mathcal{O}_\prec(\mathcal{E}^{\mathrm{av}}_k) \tag{4.12}$$

with

$$\mathcal{E}^{\mathrm{av}}_k := \frac{1}{N\eta^{k/2}}\left(1+\sum_{j=1}^{k-1}\psi^{\mathrm{av}}_j\Big(1+\frac{\psi^{\mathrm{av}}_{k-j}}{N\eta}\Big)\right), \tag{4.13}$$

where we used $|\langle M_2 A\rangle| \lesssim 1$ and $|\langle M_k A\rangle| \lesssim \eta^{1-k/2}$ for $k \ge 3$.

We recall the cumulant expansion

$$\mathbf{E}\, w_{ab} f(W) = \mathbf{E}\,\frac{\partial_{ba} f(W) + \sigma\partial_{ab} f(W)}{N} + \sum_{k=2}^{R}\sum_{q+q'=k}\frac{\kappa^{q+1,q'}_{ab}}{N^{(k+1)/2}}\,\mathbf{E}\,\partial^q_{ab}\partial^{q'}_{ba} f(W) + \Omega_R, \tag{4.14}$$

from [21, Eq. (79)] with an error term Ω_R which for the application in (4.15) below can be easily seen to be of size $\Omega_R = \mathcal{O}(N^{-2p})$ for $R = 12p$. Here the first fraction represents the Gaussian contribution and $\sigma = N\,\mathbf{E}\,w_{12}^2 \in \{0,1\}$ is determined by the complex/real symmetry class of W due to Definition 2.1. The sum in (4.14) represents the non-Gaussian contribution and $\kappa_{ab}^{p,q}$ denotes the joint cumulant of p copies of $\sqrt{N}w_{ab}$ and q copies of $\sqrt{N}\overline{w_{ab}}$. Using (4.12) and (4.14) and distributing the derivatives we obtain

$$\begin{aligned}
&\mathbf{E}|\langle (GA)^k - M_kA\rangle|^{2p}\\
&\lesssim \Big|-m\,\mathbf{E}\langle \underline{W(GA)^k}\rangle\langle (GA)^k - M_kA\rangle^{p-1}\langle (G^*A)^k - M_k^*A\rangle^p\Big| + \mathcal{O}_\prec\big((\mathcal{E}_k^{\mathrm{av}})^{2p}\big)\\
&\lesssim \mathbf{E}\,|m|\frac{|\langle (GA)^{2k}G\rangle| + |\langle (GA)^k(G^*A)^kG^*\rangle|}{N^2}\big|\langle (GA)^k - M_kA\rangle\big|^{2p-2}\\
&\quad + \sum_{|\boldsymbol{l}|+\sum(J\cup J_*)\geq 2} \mathbf{E}\,\Xi_k^{\mathrm{av}}(\boldsymbol{l},J,J^*)\big|\langle (GA)^k - M_kA\rangle\big|^{2p-1-|J\cup J_*|} + \mathcal{O}_\prec\big((\mathcal{E}_k^{\mathrm{av}})^{2p}\big),
\end{aligned} \tag{4.15}$$

where $\Xi_k^{\mathrm{av}}(\boldsymbol{l},J,J_*)$ is defined as

$$\Xi_k^{\mathrm{av}} := |m| N^{-(|\boldsymbol{l}|+\sum(J\cup J_*)+3)/2}\sum_{ab}|\partial^{\boldsymbol{l}}((GA)^k)_{ba}|\prod_{\boldsymbol{j}\in J}|\partial^{\boldsymbol{j}}\langle (GA)^k\rangle|\prod_{\boldsymbol{j}\in J_*}|\partial^{\boldsymbol{j}}\langle (G^*A)^k\rangle|, \tag{4.16}$$

and the summation in (4.15) is taken over tuples $\boldsymbol{l}\in\mathbf{Z}_{\geq 0}^2$ and multisets of tuples $J, J_*\subset \mathbf{Z}_{\geq 0}^2\setminus\{0,0\}$. Moreover, we set $\partial^{(l_1,l_2)} := \partial_{ab}^{l_1}\partial_{ba}^{l_2}$, $|(l_1,l_2)| = l_1+l_2$ and $\sum J := \sum_{\boldsymbol{j}\in J}|\boldsymbol{j}|$. For the first term in the third line of (4.15) we have

$$|m|\frac{|\langle (GA)^{2k}G\rangle| + |\langle (GA)^k(G^*A)^kG^*\rangle|}{N^2} \prec \frac{1}{N^2\eta^k}\Big(1+\frac{\psi_{2k}^{\mathrm{av}}}{N\eta}\Big) \tag{4.17}$$

due to Lemma 3.2 and $|\langle M_{2k+1}\rangle|\lesssim \eta^{-k}$ from (2.10).

We now turn to the estimate on Ξ^{av} from (4.16). Due to the Leibniz rule the derivatives can be written as a sum of products of (aa, bb, ab, ba)-entries of resolvent chains of the form $GAG\cdots AG(A)$, e.g.

$$\begin{aligned}
\partial_{ab}\partial_{ba}(GAGA)_{ba} &= G_{ba}G_{bb}(GAGA)_{aa} + G_{bb}G_{aa}(GAGA)_{ba} + G_{bb}(GAG)_{aa}(GA)_{ba}\\
&\quad + G_{ba}(GAG)_{bb}(GA)_{aa} + (GAG)_{ba}G_{bb}(GA)_{aa} + (GAG)_{bb}G_{aa}(GA)_{ba}\\
\partial_{ba}\partial_{ab}\langle GA\rangle &= \frac{G_{bb}(GAG)_{aa} + (GAG)_{bb}G_{aa}}{N}.
\end{aligned} \tag{4.18}$$

Thus we have the *naive bounds*

$$\begin{aligned}
|\partial^{\boldsymbol{l}}((GA)^k)_{ba}| &\prec \frac{1}{\eta^{(k-1)/2}}\sum_{k_0+\cdots+k_{|\boldsymbol{l}|}=k-1}\prod_i\Big(1+\frac{\psi_{k_i}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big) \prec \frac{1}{\eta^{(k-1)/2}}\Big(1+\frac{\psi_{k-1}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big),\\
|\partial^{\boldsymbol{j}}\langle (G^{(*)}A)^k\rangle| &\prec \frac{1}{N\eta^{k/2}}\sum_{k_1+\cdots+k_{|\boldsymbol{j}|}=k}\prod_i\Big(1+\frac{\psi_{k_i}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big) \prec \frac{1}{N\eta^{k/2}}\Big(1+\frac{\psi_k^{\mathrm{iso}}+\psi_{k-1}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big),
\end{aligned} \tag{4.19}$$

where we used that $\psi_{k_i}^{\mathrm{iso}} = \sqrt{N\eta}$ for $k_i\leq k-2$ by (3.19) by assumption. In the proof of the bounds (4.19) we used that

$$\big|((GA)^{k_i})_{ab}\big| = \left|(M_{k_i}A)_{ab} + \mathcal{O}_\prec\left(\frac{\psi_{k_i}^{\mathrm{iso}}}{\sqrt{N\eta^{k_i+1}}}\right)\right| \prec \frac{1}{\eta^{k_i/2}}\left(1+\frac{\psi_{k_i}^{\mathrm{iso}}}{\sqrt{N\eta}}\right), \tag{4.20}$$

by (3.1) and the norm bound in (2.10) for the deterministic term. We will use (4.19) for any $k\neq 2$, the $k=2$ case will be done slightly differently later.

For $k \neq 2$, by (4.19) we obtain

$$|\Xi_k^{\mathrm{av}}| \prec N^{(2-|\boldsymbol{l}|-\sum(J\cup J_*))/2}\frac{\sqrt{N\eta}}{N\eta^{k/2}}\Big(\frac{1}{N\eta^{k/2}}\Big)^{|J\cup J_*|}\Big(1+\frac{\psi_{k-1}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\Big(1+\frac{\psi_k^{\mathrm{iso}}+\psi_{k-1}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)^{|J\cup J_*|}. \tag{4.21}$$

Note that estimating Ξ_k^{av} is necessary only if $|\boldsymbol{l}| + \sum(J\cup J_*) \geq 2$ by (4.15), so the N-prefactor in (4.21) comes with a non-positive power. In fact, if $|\boldsymbol{l}| + \sum(J\cup J_*) \geq 3$, then this factor removes the $\sqrt{N\eta}$ factor from the numerator, which will be sufficient for our purpose.

In case $|\boldsymbol{l}| + \sum(J\cup J_*) = 2$ we still wish to remove the $\sqrt{N\eta}$ factor, so we need to improve (4.21). We use a standard procedure, called the *Ward improvement*, which relies on the fact that sums of the form $\sum_{ab}((GA)^nG)_{ab}$ can be estimated more efficiently then just estimating each term one by one. Note that in (4.16), after distributing the derivatives according to the Leibniz rule, necessarily some resolvent chain[1] $(GAG\cdots AG[A])$ appears with off-diagonal indices (a,b) or (b,a). Indeed, an off-diagonal term comes from one of the products in (4.16) when $|\boldsymbol{j}| = 1$ for some $\boldsymbol{j} \in J\cup J_*$, and it comes from the $\partial^{\boldsymbol{l}}((GA)^k)_{ba}$ factor when $|\boldsymbol{l}| = 0$ or $|\boldsymbol{l}| = 2$, by parity considerations. For such off-diagonal resolvent chains we use

$$\begin{aligned}\Big|\sum_{ab}(G[A])_{ab}\Big| &\leq N\Big(\sum_{ab}|(G[A])_{ab}|^2\Big)^{1/2} = N^{3/2}\sqrt{\langle G[A^2]G^*\rangle}\\ &\leq N^{3/2}[\|A\|]\sqrt{\langle GG^*\rangle} \prec \frac{N^{3/2}}{\eta^{1/2}}\Big(1+\frac{(\psi_0^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}}\Big)\\ \sum_{ab}|((GA)^nG[A])_{ab}| &\leq N^{3/2}\sqrt{\langle (GA)^nG[A^2]G^*(AG^*)^n\rangle}\\ &\leq [\|A\|]N^{3/2}\sqrt{\langle (GA)^nGG^*(AG^*)^n\rangle} \prec \frac{N^{3/2}}{\eta^{(n+1)/2}}\Big(1+\sqrt{\frac{\psi_{2n}^{\mathrm{av}}}{N\eta}}\Big)\end{aligned} \tag{4.22}$$

for $n \geq 1$. This allows us to gain a factor of $(N\eta)^{-1/2}$ compared with the naive bounds

$$\begin{aligned}\Big|\sum_{ab}(G[A])_{ab}\Big| &\prec N^2\Big(1+\frac{\psi_0^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\\ \sum_{ab}|((GA)^nG[A])_{ab}| &\prec \frac{N^2}{\eta^{n/2}}\Big(1+\frac{\psi_n^{\mathrm{iso}}}{\sqrt{N\eta}}\Big).\end{aligned} \tag{4.23}$$

that were used in (4.21), at the expense at the expense of replacing $1+\psi_n^{\mathrm{iso}}/\sqrt{N\eta}$ by $1+\sqrt{\psi_{2n}^{\mathrm{av}}/N\eta}$. Thus, in case $|\boldsymbol{l}| + \sum(J\cup J_*) = 2$ we can also improve upon (4.21) by a factor of $(N\eta)^{-1/2}$ and obtain

$$|\Xi_k^{\mathrm{av}}| \prec \Big(\frac{1}{N\eta^{k/2}}\Big)^{1+|J\cup J^*|}\Big(1+\frac{\psi_{k-1}^{\mathrm{iso}}+\psi_k^{\mathrm{iso}}+\sum_{j=\lfloor k/2\rfloor}^{k}(\psi_{2j}^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}}\Big)^{|J\cup J^*|+1}, \tag{4.24}$$

where we used that $\psi_{2j}^{\mathrm{av}} = \sqrt{N\eta}$ for $j < \lfloor k/2\rfloor$ from (3.19). Combining this with the earlier discussed $|\boldsymbol{l}| + \sum(J\cup J_*) \geq 3$ case, we obtain (4.24) for all cases. By plugging (4.17) and (4.24) into (4.15) we conclude

$$\begin{aligned}&\mathbf{E}\big|\langle (GA)^k - M_kA\rangle\big|^{2p} \prec (\mathcal{E}_k^{\mathrm{av}})^{2p}\\ &+\sum_{m=1}^{2p}\Big[\frac{1}{N\eta^{k/2}}\Big(1+\frac{\psi_{k-1}^{\mathrm{iso}}+\psi_k^{\mathrm{iso}}+\sum_{j=\lfloor k/2\rfloor}^{k}(\psi_{2j}^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}}\Big)\Big]^m\Big(\mathbf{E}\big|\langle (GA)^k - M_kA\rangle\big|^{2p}\Big)^{1-m/2p}\end{aligned} \tag{4.25}$$

[1] Here the $[A]$ in square brackets indicates an optional matrix A which may or may not be present.

and get the appropriate estimate $\mathbf{E}|\cdots|^{2p}$ using Young inequalities. Since p is arbitrary, it follows that

$$\begin{aligned} |\langle (GA)^k - M_k A\rangle| &\prec \mathcal{E}_k^{\mathrm{av}} + \frac{1}{N\eta^{k/2}}\Big(1 + \frac{\psi_{k-1}^{\mathrm{iso}} + \psi_k^{\mathrm{iso}} + \sum_{j=\lfloor k/2\rfloor}^{k} (\psi_{2j}^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}}\Big) \\ &\prec \frac{1}{N\eta^{k/2}}\Bigg(1 + \sum_{j=1}^{k-1} \psi_j^{\mathrm{av}} + \frac{\psi_{k-1}^{\mathrm{iso}} + \psi_k^{\mathrm{iso}} + \sum_{j=\lceil k/2\rceil}^{k} (\psi_{2j}^{\mathrm{av}})^{1/2}}{\sqrt{N\eta}}\Bigg), \end{aligned} \tag{4.26}$$

concluding the proof of (3.18a) and (3.20a). Here we used that at least one factor in the $\psi_j^{\mathrm{av}}\psi_{k-j}^{\mathrm{av}}$ product from $\mathcal{E}_k^{\mathrm{av}}$ is equal to $\sqrt{N\eta}$ by using (3.19), since either j or $k-j$ is smaller or equal than $k-2$ for $k \neq 2$.

The proof of (3.18b), i.e. the $k=2$ case, is identical except that in the second line of (4.19)

$$|\partial^{\boldsymbol{j}} \langle (G^{(*)}A)^2\rangle| \prec \frac{1}{N\eta} \sum_{k_1+\cdots+k_{|\boldsymbol{j}|}=2} \prod_i \Big(1 + \frac{\psi_{k_i}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big) \prec \frac{1}{N\eta^{k/2}}\Big(1 + \frac{\psi_2^{\mathrm{iso}}}{\sqrt{N\eta}} + \frac{(\psi_1^{\mathrm{iso}})^2}{N\eta}\Big) \tag{4.27}$$

and in $\mathcal{E}_2^{\mathrm{av}}$ there are quadratic terms resulting in $(\psi_1^{\mathrm{iso}})^2, (\psi_1^{\mathrm{av}})^2$ in (3.18b). This completes the estimates for the averaged quantities.

4.1.2 Isotropic bounds (3.18c), (3.18d) and (3.20b)

Similarly to (4.9), for the isotropic local law we start by comparing $((GA)^k G - M_{k+1})_{\boldsymbol{xy}}$ and $(\underline{GAW(GA)^{k-1}G})_{\boldsymbol{xy}}$

$$\begin{aligned} &((GA)^k G)_{\boldsymbol{xy}}\Big(1 + \mathcal{O}_\prec\big((N\eta)^{-1}\big)\Big) \\ &\quad = m(GA(AG)^{k-1})_{\boldsymbol{xy}} - m(GA\underline{WG}(AG)^{k-1})_{\boldsymbol{xy}} \\ &\quad = m(GA(AG)^{k-1})_{\boldsymbol{xy}} - m(\underline{GAWG(AG)^{k-1}})_{\boldsymbol{xy}} + m\langle GA\rangle (G^2(AG)^{k-1})_{\boldsymbol{xy}} \\ &\quad\quad + m\sum_{j=1}^{k-1} \langle (GA)^j G\rangle ((GA)^{k-j}G)_{\boldsymbol{xy}}. \end{aligned} \tag{4.28}$$

We again replace the G-chains with their deterministic counterparts using

$$\begin{aligned} (GA(AG)^{k-1})_{\boldsymbol{xy}} &= (M(z_1, A^2, z_3, \ldots))_{\boldsymbol{xy}} + \mathcal{O}_\prec\left(\frac{\psi_{k-1}^{\mathrm{iso}}}{\sqrt{N\eta^k}} + \frac{\psi_{k-2}^{\mathrm{iso}}}{\sqrt{N\eta^{k+1}}}\right) \\ &= (M(z_1, A^2, z_3, \ldots))_{\boldsymbol{xy}} + \mathcal{O}_\prec\left(\frac{\psi_{k-1}^{\mathrm{iso}} + \psi_{k-2}^{\mathrm{iso}}}{\sqrt{N\eta^{k+1}}}\right) \\ \big|\langle GA\rangle (G^2(AG)^{k-1})_{\boldsymbol{xy}}\big| &\prec \frac{\psi_1^{\mathrm{av}}}{N\eta^{1/2}} \frac{1}{\eta^{(k+1)/2}}\Big(1 + \frac{\psi_{k-1}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big), \end{aligned} \tag{4.29}$$

where we used the upper bound on $(G^2(AG)^{k-1})_{\boldsymbol{xy}}$ from (3.8). By a telescopic replacement we have

$$\begin{aligned} &\left|\sum_{j=1}^{k-1} \Big(\langle (GA)^j G\rangle ((GA)^{k-j}G)_{\boldsymbol{xy}} - \langle M_{j+1}\rangle (M_{k-j+1})_{\boldsymbol{xy}}\Big)\right| \\ &\prec \sum_{j=2}^{k-1} |\langle M_{j+1}\rangle| \frac{\psi_{k-j}^{\mathrm{iso}}}{\sqrt{N\eta^{k-j+1}}} + \sum_{j=1}^{k-1} \frac{\psi_j^{\mathrm{av}}}{N\eta^{j/2+1}} |(M_{k-j+1})_{\boldsymbol{xy}}| + \sum_{j=1}^{k-1} \frac{\psi_j^{\mathrm{av}}}{N\eta^{j/2+1}} \frac{\psi_{k-j}^{\mathrm{iso}}}{\sqrt{N\eta^{k-j+1}}} \\ &\lesssim \sum_{j=1}^{k-1} \frac{\psi_{k-j}^{\mathrm{iso}}}{\sqrt{N\eta^{k+1}}}\Big(1 + \frac{\psi_j^{\mathrm{av}}}{N\eta}\Big) + \sum_{j=1}^{k-1} \frac{\psi_j^{\mathrm{av}}}{N\eta^{k/2+1}}. \end{aligned} \tag{4.30}$$

and together with (4.7) we conclude from (4.28) that

$$((GA)^k G - M_{k+1})_{\boldsymbol{x}\boldsymbol{y}}\Big(1+\mathcal{O}_\prec\big((N\eta)^{-1}\big)\Big) \quad = -m(\underline{GAWG(AG)^{k-1}})_{\boldsymbol{x}\boldsymbol{y}} + \mathcal{O}_\prec\big(\mathcal{E}_k^{\mathrm{iso}}\big), \tag{4.31}$$

where

$$\mathcal{E}_k^{\mathrm{iso}} := \frac{1}{\sqrt{N}\eta^{(k+1)/2}}\left(1+\sum_{j=1}^{k-1}\Big[\psi_{k-j}^{\mathrm{iso}}\Big(1+\frac{\psi_j^{\mathrm{av}}}{N\eta}\Big)+\frac{\psi_j^{\mathrm{av}}}{\sqrt{N\eta}}\Big]+\mathbf{1}(k=1)\frac{\psi_1^{\mathrm{av}}}{\sqrt{N\eta}}\right). \tag{4.32}$$

Thus,

$$\begin{aligned}
&\mathbf{E}|((GA)^k G - M_{k+1})_{\boldsymbol{x}\boldsymbol{y}}|^{2p}\\
&\lesssim \Big|-m\,\mathbf{E}(\underline{GAWG(AG)^{k-1}})_{\boldsymbol{x}\boldsymbol{y}}((GA)^k G - M_{k+1})_{\boldsymbol{x}\boldsymbol{y}}^{p-1}((G^*A)^k G^* - M_{k+1}^*)_{\boldsymbol{y}\boldsymbol{x}}^{p}\Big|\\
&\quad + \mathcal{O}_\prec\big((\mathcal{E}_k^{\mathrm{iso}})^{2p}\big)\\
&\lesssim \mathbf{E}\,\widetilde{\Xi}_k^{\mathrm{iso}}\big|((GA)^k G - M_{k+1})_{\boldsymbol{x}\boldsymbol{y}}\big|^{2p-2} + \mathcal{O}_\prec\big((\mathcal{E}_k^{\mathrm{iso}})^{2p}\big)\\
&\quad + \sum_{|\boldsymbol{l}|+\sum(J\cup J_*)\geq 2}\mathbf{E}\,\Xi_k^{\mathrm{iso}}(\boldsymbol{l},J,J_*)\big|((GA)^k G - M_{k+1})_{\boldsymbol{x}\boldsymbol{y}}\big|^{2p-1-|J\cup J_*|},
\end{aligned} \tag{4.33}$$

where

$$\begin{aligned}
\widetilde{\Xi}_k^{\mathrm{iso}} := |m|\sum_{j=0}^{k}\Big(&\frac{|(G(AG)^j G(AG)^{k-1})_{\boldsymbol{x}\boldsymbol{y}}(G(AG)^{k-j+1})_{\boldsymbol{x}\boldsymbol{y}}|}{N}\\
&+\frac{|(G^*(AG^*)^j G(AG)^{k-1})_{\boldsymbol{y}\boldsymbol{y}}(GA(G^*A)^{k-j}G^*)_{\boldsymbol{x}\boldsymbol{x}}|}{N}\Big)
\end{aligned} \tag{4.34}$$

and $\Xi_k^{\mathrm{iso}}(\boldsymbol{l},J,J_*)$ is defined as

$$\begin{aligned}
\Xi_k^{\mathrm{iso}} := |m|N^{-(|\boldsymbol{l}|+\sum(J\cup J_*)+1)/2}&\sum_{ab}|\partial^{\boldsymbol{l}}[(GA)_{\boldsymbol{x}a}(G(AG)^{k-1})_{b\boldsymbol{y}}]|\\
&\times\prod_{\boldsymbol{j}\in J}|\partial^{\boldsymbol{j}}((GA)^k G)_{\boldsymbol{x}\boldsymbol{y}}|\prod_{\boldsymbol{j}\in J_*}|\partial^{\boldsymbol{j}}((G^*A)^k G^*)_{\boldsymbol{y}\boldsymbol{x}}|.
\end{aligned} \tag{4.35}$$

For (4.34) we estimate

$$\begin{aligned}
\widetilde{\Xi}_k^{\mathrm{iso}} &\prec \sum_{j=0}^{k}\Big(\|M_{k+j}\|+\frac{\psi_{k+j-1}^{\mathrm{iso}}}{\sqrt{N}\eta^{(k+j)/2}}\Big)\Big(\|M_{k-j+2}\|+\frac{\psi_{k-j+1}^{\mathrm{iso}}}{\sqrt{N}\eta^{(k-j+2)/2}}\Big)\\
&\prec \frac{1}{N\eta^{k+1}}\sum_{j=0}^{k}\Big(1+\frac{\psi_{k+j-1}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\Big(1+\frac{\psi_{k-j+1}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big).
\end{aligned} \tag{4.36}$$

In order to estimate Ξ_k^{iso} we use the entrywise bounds

$$\begin{aligned}
|\partial^{\boldsymbol{j}}((G^{(*)}A)^k G^{(*)})_{\boldsymbol{x}\boldsymbol{y}}| &\prec \frac{1}{\eta^{k/2}}\sum_{k_0+\cdots+k_{|\boldsymbol{j}|}=k}\prod_i\Big(1+\frac{\psi_{k_i}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\\
&\prec \frac{1}{\eta^{k/2}}\Big(1+\frac{\psi_k^{\mathrm{iso}}+\psi_{k-1}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\\
|\partial^{\boldsymbol{l}}(G(AG)^{k-1})_{b\boldsymbol{y}}| &\prec \frac{1}{\eta^{(k-1)/2}}\sum_{k_0+\cdots+k_{|\boldsymbol{l}|+1}=k-1}\prod_i\Big(1+\frac{\psi_{k_i}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\\
&\prec \frac{1}{\eta^{(k-1)/2}}\Big(1+\frac{\psi_{k-1}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big).
\end{aligned} \tag{4.37}$$

Note that in the second step of the first inequality we tacitly assumed that $k\neq 2$; the special case $k=2$ will be discussed at the end of the proof. From (4.37) we directly obtain the naive bound

$$|\Xi_k^{\mathrm{iso}}|\prec\frac{\sqrt{N\eta}}{N^{(|\boldsymbol{l}|+\sum(J\cup J_*)-2)/2}}\Big(\frac{\sqrt{N\eta}}{\sqrt{N\eta^{k+1}}}\Big)^{1+|J\cup J_*|}\Big(1+\frac{\psi_{k-1}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)\Big(1+\frac{\psi_k^{\mathrm{iso}}+\psi_{k-1}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)^{|J\cup J_*|}. \tag{4.38}$$

Recalling the definition (4.35) and that we need to estimate Ξ_k^{iso} only when $|\boldsymbol{l}|+\sum(J\cup J_*)\geq 2$ by (4.33), we claim that we can improve upon (4.38) by

(a) 4 factors of $(N\eta)^{-1/2}$ in case $|\boldsymbol{l}|=0$ and $|\boldsymbol{j}|=1$ for some $\boldsymbol{j}\in J\cup J_*$ (implying $|J\cup J_*|\geq 2$),

(b) 3 factors of $(N\eta)^{-1/2}$ in case $|\boldsymbol{l}|=0$ and $|J\cup J_*|\geq 1$,

(c) 3 factors of $(N\eta)^{-1/2}$ in case $|\boldsymbol{j}|=1$ for some $\boldsymbol{j}\in J\cup J_*$,

(d) 2 factor of $(N\eta)^{-1/2}$ otherwise,

at the expense of replacing of a multiplicative factor of $1+\psi_{k_i}^{\mathrm{iso}}/\sqrt{N\eta}$ by $1+(\psi_{2k_i}^{\mathrm{iso}})^{1/2}/(N\eta)^{1/4}$ for each such improvement. Indeed, estimating

$$\begin{aligned}\sum_a|((GA)^nG)_{\boldsymbol{x}a}|&\leq\sqrt{N}\sqrt{\sum_a|((GA)^nG)_{\boldsymbol{x}a}|^2}\leq N^{1/2}\sqrt{((GA)^nGG^*(AG^*)^n)_{\boldsymbol{x}\boldsymbol{x}}}\\&\prec\sqrt{\frac{N}{\eta^{n+1}}}\Big(1+\frac{\psi_{2n}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)^{1/2}\end{aligned} \tag{4.39a}$$

$$\begin{aligned}\sum_a|((GA)^nG)_{\boldsymbol{x}a}||((GA)^mG)_{\boldsymbol{y}a}|&\leq\sqrt{((GA)^nGG^*(AG^*)^n)_{\boldsymbol{x}\boldsymbol{x}}}\sqrt{((GA)^mGG^*(AG^*)^m)_{\boldsymbol{x}\boldsymbol{x}}}\\&\prec\frac{1}{\sqrt{\eta^{n+1}}}\Big(1+\frac{\psi_{2n}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)^{1/2}\frac{1}{\sqrt{\eta^{m+1}}}\Big(1+\frac{\psi_{2m}^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)^{1/2}\end{aligned} \tag{4.39b}$$

gains factors of $(N\eta)^{-1/2}$ and $(N\eta)^{-1}$ respectively, compared to the naive bounds

$$\begin{aligned}\sum_a|((GA)^nG)_{\boldsymbol{x}a}|&\prec\frac{N}{\eta^{n/2}}\Big(1+\frac{\psi_n^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)^{1/2}\\\sum_a|((GA)^nG)_{\boldsymbol{x}a}||((GA)^mG)_{\boldsymbol{y}a}|&\prec\frac{N}{\eta^{n/2+m/2}}\Big(1+\frac{\psi_n^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)^{1/2}\Big(1+\frac{\psi_m^{\mathrm{iso}}}{\sqrt{N\eta}}\Big)^{1/2},\end{aligned} \tag{4.40}$$

for one and two off-diagonal chains per summation index. Similar gains are possible for the summation over the b-index. We call a chain evaluated in $\boldsymbol{x},a$ or $\boldsymbol{y},a$ an a-chain (as in (4.39a)-(4.39b)), and a chain evaluated in $\boldsymbol{x},b$ or $\boldsymbol{y},b$ a b-chain.

We now check that, when performing the a and b summations, in each of the cases (a) to (d) the gains (4.39a) and (4.39b) can be used sufficiently often to obtain the claimed number of $(N\eta)^{-1/2}$ factors. Note that even if there were many a-chains, a gain is possible from at most two of them.

(a) Here both the $\boldsymbol{l}$-factor $[(GA)_{\boldsymbol{x}a}(G(AG)^{k-1})_{b\boldsymbol{y}}]$ (see (4.35)) and the $\boldsymbol{j}$-factor $\partial^{\boldsymbol{j}}((GA)^kG)_{\boldsymbol{x}\boldsymbol{y}}$, after performing the derivative, contain exactly one a- and one b-chain each. Hence (4.39b) can be used for both summations, and we gain four factors.

(b) Here the $\boldsymbol{l}$-factor contains one a-chain and one b-chain, while the $\boldsymbol{j}$-factor contains either an a- or b-chain, and thus both (4.39a) and (4.39b) can be used once for the a and once for the b-summation, gaining three factors.

(c) Due to $|\boldsymbol{j}| = 1$, the $\boldsymbol{j}$-factor contains one a- and one b-chain, while the $\boldsymbol{l}$-factor contains either an a- or b-chain, and thus both (4.39a) and (4.39b) can be used once, gaining three factors.

(d) The $\boldsymbol{l}$-factor contains either one a- and one b-chain, or two a-chains, or two b-chains. In the first case we use (4.39a) twice, and in the latter two cases we use (4.39b) once in order to gain two factors in total.

Now we collect these improvements for (4.38). If $|\boldsymbol{l}| + \sum(J \cup J_*) - |J \cup J_*| = 0$, then we are in case (a) and can gain 4 factors. If $|\boldsymbol{l}| + \sum(J \cup J_*) - |J \cup J_*| = 1$, then either $|\boldsymbol{l}| = 0$ and we are in case (b), or $|\boldsymbol{j}| = 1$ for all $\boldsymbol{j} \in J \cup J_*$ and we are in case (c), yielding three gained factors in both cases. Finally, if $|\boldsymbol{l}| + \sum(J \cup J_*) - |J \cup J_*| \geq 2$, then case (d) applies with a two factor gain. Note that the fewer gains are compensated by the higher power of $1/N$ in the prefactor in (4.38). Altogether we can conclude that

$$|\Xi_k^{\mathrm{iso}}| \prec \Big(\frac{1}{N\eta^{k+1}}\Big)^{(1+|J\cup J_*|)/2}\Bigg(1 + \frac{\psi_{k-1}^{\mathrm{iso}} + \psi_k^{\mathrm{iso}}}{\sqrt{N\eta}} + \frac{(\psi_{k+1}^{\mathrm{iso}})^{1/2} + \cdots + (\psi_{2k}^{\mathrm{iso}})^{1/2}}{(N\eta)^{1/4}}\Bigg)^{|J\cup J_*|+1}. \tag{4.41}$$

By plugging (4.36) and (4.41) into (4.33) we conclude (3.18c) and (3.20b). This proves (3.18c) and (3.20b).

For the special $k = 2$ case, i.e. for the proof of (3.18d) we note that in the first equality of (4.37) and in the estimate on $\mathcal{E}_k^{\mathrm{iso}}$ there are additional quadratic terms $(\psi_1^{\mathrm{iso}})^2$ and $\psi_1^{\mathrm{iso}}\psi_1^{\mathrm{av}}$ but otherwise the proof remains unchanged. □

5 Proof of the reduction inequalities, lemma 3.6

In order to prove lemma 3.6 we first infer local laws for resolvent chains including some absolute value $|G|$ from resolvent chains without absolute value. To formulate the precise statement, for any choices of $g_i(x) \in \{1/(x-z_i), 1/|x-z_i|\}$ we first generalise (2.4) to

$$M(g_1, A_1, g_2, \dots, A_{k-1}, g_k) := \sum_{\pi\in\mathrm{NC}[k]} \mathrm{pTr}_{K(\pi)}(A_1, \dots, A_{k-1}) \prod_{B\in\pi} \mathrm{sc}_\circ[B], \tag{5.1}$$

where $\mathrm{sc}_\circ$ is the free cumulant function of $\mathrm{sc}[i_1, \dots, i_n] := \langle g_{i_1} \cdots g_{i_k}\rangle_{\mathrm{sc}}$. We note that the bounds (2.10) and their proofs verbatim also apply to this more generalised M. The following lemma generalises lemma 3.2 to absolute values.

Lemma 5.1. *Fix $\epsilon > 0$ and $k > 0$ and assume that for $1 \leq j \leq k$ a priori bounds*

$$\Psi_j^{\mathrm{av}}(\boldsymbol{z}_j, \boldsymbol{A}_j) \prec \psi_j^{\mathrm{av}}, \qquad \Psi_j^{\mathrm{iso}}(\boldsymbol{z}_j, \boldsymbol{A}_j, \boldsymbol{x}, \boldsymbol{y}) \prec \psi_j^{\mathrm{iso}} \tag{5.2}$$

have been established ϵ-uniformly in traceless matrices. Then $z_1, \dots, z_{k+1} \in \mathbf{C}$ with $\eta = \min_i|\Im z_i|$ and $G_i \in \{G(z_i), |G(z_i)|\}$ and corresponding $g_i(x) \in \{1/(x - z_i), 1/|x - z_i|\}$ it holds that

$$\begin{aligned}\langle G_1B_1\cdots G_kB_k\rangle &= \langle M(g_1, B_1, \dots, B_{k-1}, g_k)B_k\rangle + \mathcal{O}_\prec\Bigg(\frac{\sum_{j=a}^k \psi_j^{\mathrm{av}} \wedge 1}{N\eta^{k-a/2}}\Bigg)\\ \Big(G_1B_1G_2\cdots B_kG_{k+1}\Big)_{\boldsymbol{x}\boldsymbol{y}} &= M(g_1, B_1, \dots, B_k, g_{k+1})_{\boldsymbol{x}\boldsymbol{y}} + \mathcal{O}_\prec\Bigg(\frac{\sum_{j=a}^k \psi_j^{\mathrm{iso}} \wedge 1}{\sqrt{N}\eta^{k-a/2+1/2}}\Bigg),\end{aligned} \tag{5.3}$$

ϵ-uniformly in vectors $\boldsymbol{x}, \boldsymbol{y}$ and deterministic matrices $B_1, \dots, B_k$, out of which a are traceless.

Proof. The proof is analogous to the special case given in lemma 3.2, with the additional step first of representing any $|G|$ via

$$|G(E+\mathrm{i}\eta)| = \frac{1}{\mathrm{i}\pi}\int_0^\infty \frac{G(E+\mathrm{i}(\eta^2+s^2)^{1/2}) - G(E-\mathrm{i}(\eta^2+s^2)^{1/2})}{(\eta^2+s^2)^{1/2}}\,\mathrm{d}s \tag{5.4}$$

as an integral over resolvents. Here we used the identity

$$\frac{1}{|x-\mathrm{i}\eta|} = \frac{1}{\mathrm{i}\pi}\int_0^\infty \left(\frac{1}{x-\mathrm{i}(\eta^2+s^2)^{1/2}} - \frac{1}{x-\mathrm{i}(\eta^2+s^2)^{1/2}}\right)\frac{1}{(\eta^2+s^2)^{1/2}}\,\mathrm{d}s. \tag{5.5}$$

We note that M for $g(x) = |x-E-\mathrm{i}\eta|^{-1}$ satisfies the analogous identity

$$M(\dots,g,\dots) = \frac{1}{\mathrm{i}\pi}\int_0^\infty \frac{M(\dots,E+\mathrm{i}(\eta^2+s^2)^{1/2},\dots) - M(\dots,E-\mathrm{i}(\eta^2+s^2)^{1/2},\dots)}{(\eta^2+s^2)^{1/2}}\,\mathrm{d}s \tag{5.6}$$

by multi-linearity. In (5.6) the lhs. is understood in the sense of (5.1), and the rhs. in the sense of (2.4).

It remains to estimate the integral of the error term obtained from using (5.4) for each $|G|$ and replacing the resulting resolvent chains by their deterministic equivalents. From now on we only consider the case $a=k$ in the averaged version (the isotropic one is analogous). Proceeding as in lemma 3.2, the general case $0\le a\le k-1$ is completely analogous and so omitted. For notational simplicity in the following we denote all the deterministic matrices by A and resolvents by G (even if they are evaluated at different spectral parameters). For concreteness we assume that only two $g_i(x)$'s are equal to $|x-z_i|^{-1}$, the rest is $(x-z_i)^{-1}$, i.e. $k_1+k_2+2=k$. Introducing the shorthand notations $z_{i,s} := E_i + \mathrm{i}\sqrt{\eta_i^2+t^2}$, $M(z_{1,s}, z_{k_1+2,s}) := M(z_{1,s}, A, z_2, \dots, z_{k_1+1}, A, z_{k_1+2,s}, A, z_{k_1+3}, \dots, z_k)$, we have

$$\begin{aligned}
&\big|\langle |G(E_1+\mathrm{i}\eta_1)|A(GA)^{k_1}|G(E_{k_1+2}+\mathrm{i}\eta_{k_1+2})|A(GA)^{k_2}\rangle - \langle M(g_1,A,\dots,A,g_k)A\rangle\big| \\
&\lesssim \left|\iint_0^\infty \langle G(z_{1,s})A(GA)^{k_1}G(z_{k_1+2,s})A(GA)^{k_2} - M(z_{1,s},z_{k_1+2,t})A\rangle \frac{\mathrm{d}s\,\mathrm{d}t}{\sqrt{\eta_1^2+s^2}\sqrt{\eta_{k_1+2}^2+t^2}}\right| \\
&\lesssim \left|\iint_0^{N^{5k}} \langle G(z_{1,s})A(GA)^{k_1}G(z_{k_1+2,s})A(GA)^{k_2} - M(z_{1,s},z_{k_1+2,t})A\rangle \frac{\mathrm{d}s\,\mathrm{d}t}{\sqrt{\eta_1^2+s^2}\sqrt{\eta_{k_1+2}^2+t^2}}\right| \\
&\quad + \mathcal{O}\left(N^{-2}\right) \\
&\prec \frac{\psi_k^{\mathrm{av}}}{N\eta^{k/2}}\left(\int_0^1\int_0^{N^{5k}} + \int_0^{N^{5k}}\int_0^1\right)\frac{\mathrm{d}s\,\mathrm{d}t}{\sqrt{\eta_1^2+s^2}\sqrt{\eta_{k_1+2}^2+t^2}} \\
&\quad + \frac{1}{N\eta^k}\int_1^{N^{5k}}\int_1^{N^{5k}} \frac{\mathrm{d}s\,\mathrm{d}t}{\sqrt{\eta_1^2+s^2}\sqrt{\eta_{k_1+2}^2+t^2}} + \mathcal{O}\left(N^{-2}\right) \\
&\prec \frac{\psi_k^{\mathrm{av}}\wedge 1}{N\eta^{k/2}}.
\end{aligned} \tag{5.7}$$

Note that to go from the second to the third line we used the trivial norm bound $\|G(E+\mathrm{i}\eta)\|\lesssim \eta^{-1}$ to remove the very large s and t regime (and a similar bound for the deterministic term). Additionally, in the penultimate inequality we used (5.2) to bound the regime $\eta\le 1$, with $\eta := \min_i |\Im z_i|$, and the averaged local law (2.11a) in the regime $\eta \ge 1$. Alternatively, we could have used [23, Theorem 3.4] in this latter regime. □

Proof of Lemma 3.6. Similarly to Section 4, to make the presentation simpler we do not carry the dependence on the spectral parameters z_j and traceless matrices A_j but instead simply write G and A.

We first start with the bound in the average case and we distinguish two cases depending on whether k is even or odd. Let $\{\lambda_i\}_{i\in[N]}$ be the eigenvalues of W, and let $\boldsymbol{u}_i$ be the corresponding eigenvectors. For even k, using the shorthand notation $T := A(GA)^{k/2-1}$, we have

$$\begin{aligned}
&\Psi_{2k}^{\mathrm{av}} = N\eta^k |\langle (GA)^{2k} - M_{2k}A\rangle| \\
&\lesssim N\eta + \frac{N\eta^k}{N}\left|\sum_{ijml}\frac{\langle \boldsymbol{u}_i, T\boldsymbol{u}_j\rangle\langle \boldsymbol{u}_j, T\boldsymbol{u}_m\rangle\langle \boldsymbol{u}_m, T\boldsymbol{u}_l\rangle\langle \boldsymbol{u}_l, T\boldsymbol{u}_i\rangle}{(\lambda_i - z_1)(\lambda_j - z_{k/2+1})(\lambda_m - z_{k+1})(\lambda_l - z_{(3k)/2+1})}\right| \\
&\lesssim N\eta + \frac{N\eta^k}{N}\sum_{ijml}\frac{|\langle \boldsymbol{u}_i, A(GA)^{k/2-1}\boldsymbol{u}_j\rangle|^2|\langle \boldsymbol{u}_m, A(GA)^{k/2-1}\boldsymbol{u}_l\rangle|^2}{|(\lambda_i - z_1)(\lambda_j - z_{k/2+1})(\lambda_m - z_{k+1})(\lambda_l - z_{(3k)/2+1})|} \\
&= N\eta + N^2\eta^k\langle |G|A(GA)^{k/2-1}|G|A(G^*A)^{k/2-1}\rangle\langle |G|A(GA)^{k/2-1}|G|A(G^*A)^{k/2-1}\rangle \\
&\lesssim N\eta + N^2\eta^k\left(\frac{1}{\eta^{k/2-1}} + \frac{\psi_k^{\mathrm{av}}}{N\eta^{k/2}}\right)^2 \le (N\eta + \psi_k^{\mathrm{av}})^2\,.
\end{aligned} \tag{5.8}$$

In the last line we used Lemma 5.1. This concludes the bound for even k.

Similarly, for odd k we have

$$\begin{aligned}
\Psi_{2k}^{\mathrm{av}} &= N\eta^k |\langle (GA)^{2k} - M_{2k}A\rangle| \\
&\lesssim N\eta + N^2\eta^k\langle |G|A(GA)^{(k+1)/2-1}|G|A(GA)^{(k+1)/2-1}\rangle \\
&\quad\times\langle |G|A(GA)^{(k-1)/2-1}|G|A(GA)^{(k-1)/2-1}\rangle \\
&\lesssim (N\eta)^2 + N\eta(\psi_{k+1}^{\mathrm{av}} + \psi_{k-1}^{\mathrm{av}}) + \psi_{k+1}^{\mathrm{av}}\psi_{k-1}^{\mathrm{av}},
\end{aligned} \tag{5.9}$$

where to go to the last line we again used Lemma 5.1. Additionally, to go from the first to the second line of (5.9) we used (with the shorthand notation $T := A(GA)^{(k+1)/2-1}$, $S := A(GA)^{(k-1)/2-1}$)

$$\begin{aligned}
&\langle (GA)^{2k}\rangle \\
&= \frac{1}{N}\sum_{ijml}\frac{\langle \boldsymbol{u}_i, T\boldsymbol{u}_j\rangle\langle \boldsymbol{u}_j, T\boldsymbol{u}_m\rangle\langle \boldsymbol{u}_m, S\boldsymbol{u}_l\rangle\langle \boldsymbol{u}_l, S\boldsymbol{u}_i\rangle}{(\lambda_i - z_1)(\lambda_j - z_{(k+1)/2+1})(\lambda_m - w_{k+2})(\lambda_l - w_{(3k+1)/2+1})} \\
&\lesssim \frac{1}{N}\sum_{ijml}\frac{|\langle \boldsymbol{u}_i, A(GA)^{(k+1)/2-1}\boldsymbol{u}_j\rangle|^2|\langle \boldsymbol{u}_m, A(GA)^{(k-1)/2-1}\boldsymbol{u}_l\rangle|^2}{|(\lambda_i - z_1)(\lambda_j - z_{(k+1)/2+1})(\lambda_m - w_{k+2})(\lambda_l - w_{(3k+1)/2+1})|} \\
&= N\langle |G|A(GA)^{(k+1)/2-1}|G|A(G^*A)^{(k+1)/2-1}\rangle\langle |G|A(GA)^{(k-1)/2-1}|G|A(G^*A)^{(k-1)/2-1}\rangle.
\end{aligned} \tag{5.10}$$

We now consider the isotropic case when k is even and $j \ge 1$:

$$\begin{aligned}
\Psi_{k+j}^{\mathrm{iso}} &\lesssim \sqrt{N\eta} + \sqrt{N}\eta^{(k+j+1)/2}\langle \boldsymbol{x}, (GA)^{k+j}G\boldsymbol{y}\rangle \\
&= \sqrt{N\eta} + \sqrt{N}\eta^{(k+j+1)/2}\langle \boldsymbol{x}, (GA)^{k/2}GA(GA)^{j-1}G(AG)^{k/2}\boldsymbol{y}\rangle \\
&\lesssim \sqrt{N\eta} + N\eta^{(k+j+1)/2}\langle \boldsymbol{x}, (GA)^{k/2}|G|(AG^*)^{k/2}\boldsymbol{x}\rangle^{1/2}\langle \boldsymbol{y}, (GA)^{k/2}|G|(AG^*)^{k/2}\boldsymbol{y}\rangle^{1/2} \\
&\qquad\quad\times\langle |G|A(GA)^{j-1}|G|(AG^*)^{j-1}A\rangle^{1/2} \\
&\lesssim \sqrt{N\eta} + N\eta^{(k+j+1)/2}\left(\frac{1}{\eta^{k/2}} + \frac{\psi_k^{\mathrm{iso}}}{\sqrt{N\eta^{2k+1}}}\right)\left(\frac{1}{\eta^{j-1}} + \frac{\psi_{2j}^{\mathrm{av}}}{N\eta^j}\right)^{1/2} \\
&\lesssim \big(N\eta + (N\eta)^{1/2}\psi_k^{\mathrm{iso}}\big)\big(1 + (N\eta)^{-1/2}(\psi_{2j}^{\mathrm{av}})^{1/2}\big),
\end{aligned} \tag{5.11}$$

Additionally, to go from the second to the third line we used that

$$\begin{aligned}
&\langle \boldsymbol{x}, (GA)^{k/2} GA (GA)^{j-1} G (AG)^{k/2} \boldsymbol{y}\rangle \\
&\quad = \sum_{ij} \frac{\langle \boldsymbol{x}, (GA)^{k/2}\boldsymbol{u}_i\rangle \langle \boldsymbol{u}_i, A(GA)^{j-1}\boldsymbol{u}_j\rangle \langle \boldsymbol{u}_j, (AG)^{k/2}\boldsymbol{y}\rangle}{(\lambda_i - z_{k/2+1})(\lambda_j - z_{k/2+j+1})} \\
&\quad \le \left(\sum_{ij} \frac{\langle |\boldsymbol{x}, (GA)^{k/2}\boldsymbol{u}_i\rangle|^2 |\langle \boldsymbol{u}_j, (AG)^{k/2}\boldsymbol{y}\rangle|^2}{|(\lambda_i - z_{k/2+1})(\lambda_j - z_{k/2+j+1})|}\right)^{1/2} \left(\sum_{ij} \frac{|\langle \boldsymbol{u}_i, A(GA)^{j-1}\boldsymbol{u}_j\rangle|^2}{|(\lambda_i - z_{k/2+1})(\lambda_j - z_{k/2+j+1})|}\right)^{1/2} \\
&\quad = N^{1/2} \langle \boldsymbol{x}, (GA)^{k/2}|G|(AG^*)^{k/2}\boldsymbol{x}\rangle^{1/2} \langle \boldsymbol{y}, (GA)^{k/2}|G|(AG^*)^{k/2}\boldsymbol{y}\rangle^{1/2} \\
&\qquad \times \langle |G|A(GA)^{j-1}|G|(AG^*)^{j-1}A\rangle^{1/2}.
\end{aligned} \tag{5.12}$$

□

6 Proof of Corollary 2.7

The proof of this corollary relies on the Helffer-Sjöstrand representation [25], i.e. we express each $f_i(W)$ in $f_1(W)A_1 \cdots f_k(W)$ as an integral of resolvents at different spectral parameters. Note that by eigenvalue rigidity (see e.g. [28, Theorem 7.6] or [34]) the spectrum of W is contained in $[-2-\epsilon, 2+\epsilon]$, for any small $\epsilon > 0$, with very high probability. In particular this implies that it is enough to consider test functions $f_i \in H_0^{\lceil k-a/2\rceil}([-3,3])$, i.e. Sobolev functions on $\mathbf{R}$ which are non-zero only on $[-3,3]$. In fact, this can be always achieved by multiplying the original f with a smooth cut-off function without changing $f(W)$ up to an event of very small probability.

We present the proof only when all the matrices are traceless, i.e. when $a = k$. The proof in the general case is completely analogous and so omitted.

Let $f \in H_0^{\lceil k/2\rceil}([-3,3])$ then we define its almost analytic extension by

$$f_{\mathbf{C}}(z) = f_{\mathbf{C},k}(z) = f_{\mathbf{C},k}(x + \mathrm{i}\eta) := \left[\sum_{j=0}^{\lceil k/2\rceil - 1} \frac{(\mathrm{i}\eta)^j}{j!} f^{(j)}(x)\right] \chi(\eta), \tag{6.1}$$

where $\chi(\eta)$ is a smooth cut-off equal to one on $[-5,5]$ and equal to zero on $[-10,10]^c$ and $f^{(j)}$ denotes the j-th derivative. Then we have

$$f(\lambda) = \frac{1}{\pi} \int_{\mathbf{C}} \frac{\partial_{\bar z} f_{\mathbf{C}}(z)}{\lambda - z}\, \mathrm{d}^2 z, \tag{6.2}$$

where $\mathrm{d}^2 z = \mathrm{d}x\, \mathrm{d}\eta$ denotes the Lebesgue measure on $\mathbf{C} \equiv \mathbf{R}^2$ with $z = x + \mathrm{i}\eta$.

Consider $f_1, \dots, f_k \in H_0^{\lceil k/2\rceil}([-3,3])$, then by (6.2) we get

$$f_1(W)A_1 \cdots f_k(W) = \frac{1}{\pi^k} \int_{\mathbf{C}^k} \prod_{i=1}^k \mathrm{d}^2 z_i \left[\prod_{i=1}^k (\partial_{\bar z}(f_i)_{\mathbf{C}})(z_i)\right] G(z_1)A_1 \cdots A_{k-1}G(z_k), \tag{6.3}$$

where $G(z_i) := (W - z_i)^{-1}$.

Proof of Corollary 2.7. This argument is very similar to the proof of [23, Theorem 2.6], hence here we only explain the main differences.

Pick any $\xi > 0$ as a tolerance exponent in the definition of $\mathcal{O}_\prec()$. Without loss of generality we can assume that $\max_i \|f_i\|_{H^{\lceil k/2\rceil}} \lesssim N^{1-\xi}$ (otherwise there is nothing to prove). We first prove the averaged case in (2.12), and then we explain the very minor changes required in the isotropic case.

We start with the bound

$$\int_{\mathbf{R}} \mathrm{d}x |\partial_{\bar{z}} f_{\mathbf{C},k}(x+\mathrm{i}\eta)| \lesssim \eta^{\lceil k/2\rceil -1} \|f\|_{H^{\lceil k/2\rceil}} \tag{6.4}$$

which easily follows from (6.1). Set $\eta_0 := N^{-1+\xi/2}$; first we prove that the regime $|\eta_i| \le \eta_0$, for some $i \in [k]$ in the integral representation of $\langle f_1(W)A_1 \dots f_k(W)A_k\rangle$ from (6.3) is negligible. Here we only present the proof in the case when $|\eta_i| \le \eta_0$ happens only for a single index i; the changes when more than one η_i's are small are exactly the same as explained above [23, Eq. (3.21)], giving an even smaller bound.

Without loss of generality we assume that $|\eta_1| \le \eta_0$. In this regime we claim that (with $z_i = x_i + \mathrm{i}\eta_i$)

$$\begin{aligned} &\left| \int \mathrm{d}x_1 \cdots \mathrm{d}x_k \int_{\substack{|\eta_i|\ge \eta_0,\\ i\in[2,k]}} \mathrm{d}\eta_2 \cdots \mathrm{d}\eta_k \int_{-\eta_0}^{\eta_0} \mathrm{d}\eta_1 \left(\prod_{i=1}^k (\partial_{\bar{z}} (f_i)_{\mathbf{C}})(z_i) \right) \langle G(z_1)A_1 \cdots G(z_k)A_k\rangle \right| \\ &\qquad \prec \eta_0 (N\eta_0)^{k/2-1} \max_i \|f_i\|_{H^{\lceil k/2\rceil}}. \end{aligned} \tag{6.5}$$

To prove (6.5) we will use Stokes theorem in the following form:

$$\int_{-10}^{10} \int_{\widetilde{\eta}}^{10} \partial_{\bar{z}} \psi(x+\mathrm{i}\eta) h(x+\mathrm{i}\eta)\, \mathrm{d}x\, \mathrm{d}\eta = \frac{1}{2\mathrm{i}} \int_{-10}^{10} \psi(x+\mathrm{i}\widetilde{\eta}) h(x+\mathrm{i}\widetilde{\eta})\, \mathrm{d}x, \tag{6.6}$$

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. We will use (6.6) and the compact support of $(f_i)_{\mathbf{C}}$ to conclude that

$$\begin{aligned} &\int_{\mathbf{R}} \mathrm{d}x_i \int_{\eta_0}^{10} \mathrm{d}\eta_i (\partial_{\bar{z}} (f_i)_{\mathbf{C}})(z_i) \langle G(z_1)A_1 \dots A_{i-1} G(z_i) A_i \dots G(z_k)A_k\rangle \\ &\qquad = \frac{1}{2\mathrm{i}} \int_{\mathbf{R}} \mathrm{d}x_i (f_i)_{\mathbf{C}}(x_i + \mathrm{i}\eta_0) \langle G(z_1)A_1 \dots A_{i-1} G(x_i+\mathrm{i}\eta_0) A_i \dots G(z_k)A_k\rangle, \end{aligned} \tag{6.7}$$

for any fixed $z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_k$. Using (6.7) repeatedly for the $z_2, \dots, z_k$-variables, we conclude

$$\begin{aligned} |\text{lhs. of (6.5)}| = \frac{1}{2^{k-1}} \Bigg| \int \prod_{i=1}^k \mathrm{d}x_i \int_{-\eta_0}^{\eta_0} \mathrm{d}\eta_1 (\partial_{\bar{z}} (f_1)_{\mathbf{C}})(x_1 + \mathrm{i}\eta_1) \prod_{i=2}^k (f_i)_{\mathbf{C}}(x_i + \mathrm{i}\eta_0) \\ \times \langle G(z_1)A_1 G(x_2+\mathrm{i}\eta_0) \cdots G(x_k+\mathrm{i}\eta_0)A_k\rangle \Bigg|. \end{aligned} \tag{6.8}$$

Additionally, we will use the following bound on products of k resolvents which holds uniformly in $|\eta| \ge N^{-10k}$. For this bound we introduce $\rho(z) := \pi^{-1} |\Im m_{\mathrm{sc}}(z)|$, for any $z \in \mathbf{C} \setminus \mathbf{R}$, as the harmonic extension of the semicircle density noting that $\rho(x + \mathrm{i}0) = \rho_{\mathrm{sc}}(x)$.

Lemma 6.1. *For any $k \in \mathbf{N}$, $z_i := x_i + \mathrm{i}\eta_i$, with $|x_i| \le 2$ and $|\eta_i| \ge N^{-10k}$, with $i \in [k]$, it holds*

$$|\langle G(z_1)AG(z_2) \dots AG(z_k)A\rangle| \prec N^{k/2-1} \prod_{i\in[k]} \frac{1}{\rho(x_i + \mathrm{i}N^{-2/3})} \left(1 + \frac{1}{N|\eta_i|}\right), \tag{6.9}$$

$$|\langle \boldsymbol{x}, G(z_1)AG(z_2) \dots AG(z_k)\boldsymbol{y}\rangle| \prec N^{(k-1)/2} \prod_{i\in[k]} \frac{1}{\rho(x_i + \mathrm{i}N^{-2/3})} \left(1 + \frac{1}{N|\eta_i|}\right), \tag{6.10}$$

uniformly for deterministic traceless matrices $\|A\| \lesssim 1$, vectors $\|\boldsymbol{x}\| + \|\boldsymbol{y}\| \lesssim 1$, and z_i as above.

Armed with all these ingredients, we have the following chain of inequalities in order to prove (6.5):

$$\begin{aligned}
&|\text{rhs. of (6.5)}|\\
&\quad\prec \int \mathrm{d}x_1 \frac{|f_1^{(\lceil k/2\rceil)}(x_1)|}{\rho(x_1+\mathrm{i}N^{-2/3})}\int_{\eta_r\le|\eta_1|\le\eta_0}\eta_1^{\lceil k/2\rceil-1}\left(1+\frac{1}{N|\eta_1|}\right)\mathrm{d}\eta_1\left(\prod_{i=2}^k\int\mathrm{d}x_i\frac{1}{\rho(x_i+\mathrm{i}N^{-2/3})}\right)\\
&\qquad+\eta_0\|f_1\|_{H^{\lceil k/2\rceil}}\int_{|\eta_1|\le\eta_r}\eta_1^{\lceil k/2\rceil-2}\eta_0^{-k+1}\,\mathrm{d}\eta_1\\
&\quad\lesssim \eta_0(N\eta_0)^{k/2-1}\left(\int\mathrm{d}x_1\left|f_1^{(\lceil k/2\rceil)}(x_1)\right|^2\right)^{1/2}\left(\int\frac{\mathrm{d}x_1}{\rho(x_1+\mathrm{i}N^{-2/3})^2}\right)^{1/2}+\eta_0\|f_1\|_{H^{\lceil k/2\rceil}}\\
&\quad\lesssim \eta_0(N\eta_0)^{k/2-1}\|f_1\|_{H^{\lceil k/2\rceil}},
\end{aligned}\tag{6.11}$$

where in the first step we first used $\|f_i\|_\infty\lesssim 1$ for $i\in[2,k]$ and after splitting the η_1 integration, in the regime $\eta_r\le|\eta_1|\le\eta_0$ we used (6.9) together with

$$|\partial_{\bar z}(f_1)(x_1+\mathrm{i}\eta_1)|\lesssim\eta_1^{\lceil k/2\rceil-1}|f_1^{(\lceil k/2\rceil)}(x_1)|$$

for any $|x_1|\le 2$, $|\eta_1|\le\eta_0$ from (6.1). In the complementary regime $|\eta_1|<\eta_r$ we used the trivial norm bound $|\langle G(z_1)A_1\cdots G(z_k)A_k\rangle|\le\prod_i\|G(z_i)A_i\|\le\prod_i|\eta_i|^{-1}$ together with (6.4). In the penultimate inequality of (6.11) we also used that $\int 1/\rho$ is finite due to the square root singularity of ρ, and that $\int 1/\rho^2\lesssim\log N$ thanks to the tiny $N^{2/3}$-regularisation. This concludes the proof of (6.5).

We now estimate the integration regime in (6.8) where $|\eta_i|\ge\eta_0$ for all $i\in[k]$. By (6.3) and the local law (2.11a), we conclude that

$$\begin{aligned}
&\langle f_1(W)A_1\cdots f_k(W)A_k\rangle\\
&\quad=\frac{1}{\pi^k}\int_{\mathbf{R}^k}\int_{\eta_0\le|\eta_i|\le 10}\mathrm{d}^2z_1\cdots\mathrm{d}^2z_k(\partial_{\bar z}(f_1)_{\mathbf{C}})(z_1)\cdots(\partial_{\bar z}(f_k)_{\mathbf{C}})(z_k)\langle M_{[k]}A_k\rangle\\
&\qquad+\mathcal{O}_\prec\left(\eta_0(N\eta)^{k/2-1}\max_i\|f_i\|_{H^{\lceil k/2\rceil}}\right).
\end{aligned}\tag{6.12}$$

where we abbreviated $M_{[k]}=M(z_1,A_1,\dots,z_{k-1},A_{k-1},z_k)$. Note that in (6.12) we estimated the error term $N^{-1}(\min|\eta_i|)^{-k/2}$ coming from the local law (2.11a) by

$$\begin{aligned}
&\frac{1}{\pi^k}\int_{\mathbf{R}^k}\int_{\eta_0\le|\eta_i|\le 10}\mathrm{d}^2\boldsymbol{z}\prod_{i=1}^k(\partial_{\bar z}(f_i)_{\mathbf{C}})(z_i)\langle(G(z_1)A_1\dots G(z_k)-M_{[k]})A_k\rangle\\
&\qquad\qquad=\mathcal{O}_\prec\left(N^{-1}\max_i\|f_i\|_{H^{\lceil k/2\rceil}}\right),
\end{aligned}\tag{6.13}$$

with $\mathrm{d}^2\boldsymbol{z}:=\mathrm{d}^2z_1\dots\mathrm{d}^2z_k$. More precisely, in (6.13) 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 (6.7), and then estimated the remaining $\partial_{\bar z}(f_1)_{\mathbf{C}}(z_1)$ by (6.4). The error term $N^{-1}|\eta_1|^{-k/2}$ from the local law together with the $|\eta_1|^{\lceil k/2\rceil-1}$ bound from (6.4) and the integration in η_1 yields (6.13).

Finally, using that by (6.5) the regime $\eta_i\in[\eta_r,\eta_0]$ can be added back to (6.12) at the price of an error $\eta_0(N\eta_0)^{k/2-1}\max_i\|f_i\|_{H^{\lceil k/2\rceil}}$ we conclude the proof of the averaged case in (2.12) modulo the computation of the leading deterministic term which is done exactly as in [23, Proof of Theorem 2.6] and so the details are omitted.

The proof of the isotropic case in (2.12) is very similar. The only differences are the following: (i) to bound the small η_i-regime we have to use (6.10) instead of (6.9), which still gives exactly the same bound (6.5); (ii) to estimate the error term coming from the isotropic local law (2.11b) (used in the regime when $|\eta_i| \geq \eta_0$ for all $i \in [k]$) we have to replace (6.13) by

$$\begin{aligned}&\frac{1}{\pi^k}\int_{\mathbf{R}^k}\int_{\eta_0 \leq |\eta_i| \leq 10} \mathrm{d}^2 \boldsymbol{z} \prod_{i=1}^{k} (\partial_{\bar{z}}(f_i)_{\mathbf{C}})(z_i) \langle \boldsymbol{x}, (G(z_1)A_1 \dots G(z_k) - M_{[k]})\boldsymbol{y}\rangle \\ &\qquad\qquad = \mathcal{O}_{\prec}\left(N^{-1/2} \max_i \|f_i\|_{H^{\lceil k/2 \rceil}}\right).\end{aligned} \tag{6.14}$$

The proof of (6.14) is exactly the same as the proof of (6.13). □

A Additional proofs

Proof of Lemma 2.4. We first note that the inequality

$$|m_{\mathrm{sc}}[z_1, \dots, z_j]| \lesssim \frac{1}{\eta^{j-1}} \tag{A.1}$$

is a direct consequence of the integral representation (2.7). The bound (A.1) is sharp only when not all $\Im z_i$ have the same sign. If all signs agree, then the iterated divided difference remains bounded by the smoothness of m_{sc} in the bulk. By Möbius inversion [23, Eq. (2.3), Lemma 2.16] we have

$$\begin{aligned} m_\circ[B] &= \sum_{\pi \in \mathrm{NC}(B)} (-1)^{|\pi|-1} \Bigg(\prod_{S \in K(\pi)} C_{|S|-1} \Bigg) \prod_{T \in \pi} m[T] \\ &= m[B] + \sum_{\substack{\pi \in \mathrm{NC}(B) \\ |\pi| \geq 2}} (-1)^{|\pi|-1} \Bigg(\prod_{S \in K(\pi)} C_{|S|-1} \Bigg) \prod_{T \in \pi} m[T] \\ &= m[B] + \sum_{\substack{\pi \in \mathrm{NC}(B) \\ |\pi| \geq 2}} \mathcal{O}\bigg(\frac{1}{\eta^{|B|-|\pi|}} \bigg) \lesssim \frac{1}{\eta^{|B|-1}}, \end{aligned} \tag{A.2}$$

where C_n is the n-th Catalan number. Here we used (A.1) in the third and fourth step recalling that

$$m[T] = m\big[\{z_i \mid i \in T\}\big]. \tag{A.3}$$

We note that (A.2) is sharp since (A.1) is sharp and leading order cancellations are impossible in the ultimate line.

From the definition (2.5) it follows that $\mathrm{pTr}_{K(\pi)}$ is non-zero only when no block of $K(\pi)$ is a singleton $\{i\}$ with $\langle B_i \rangle = 0$, and therefore $|K(\pi)| \leq k - \lceil a/2 \rceil$ or equivalently $|\pi| \geq 1 + \lceil a/2 \rceil$. Thus (2.10) follows directly from (2.4). □

Proof of lemma 4.1. We only prove (4.7) as the proof of (4.8) is completely analogous. We recall the alternative definition of M from [23, Eq. (5.12)]

$$\begin{aligned} \frac{M(z_1, \dots, z_k)}{m_1 \cdots m_k} &= \sum_{E \in \mathrm{NCG}[1,k]} \mathrm{pTr}_{K(\pi(E))}(\boldsymbol{A}_{[1,k)}) q_E, \\ q_E &:= \prod_{e \in E} q_e, \quad q_{ij} := \frac{m_i m_j}{1 + m_i m_j}, \quad \boldsymbol{A}_S := \big(A_i \mid i \in S\big), \end{aligned} \tag{A.4}$$

where $\mathrm{NCG}[1,k]$ denotes the set of *non-crossing graphs* on the vertex set $[1,k] = \{1, \dots, k\}$, i.e. graphs without crossing edges $(ab), (cd)$ with $a < c < b < d$. The graphs are identified with their edge sets E. Note that the connected components of any non-crossing graph E form a non-crossing partition of the set $[1,k]$ that we denoted by $\pi(E)$ in (A.4).

For any fixed $j\in[1,k]$, we now partition the set of non-crossing graphs as

$$\mathrm{NCG}[1,k]=\mathcal{G}_j\sqcup\bigsqcup_{l=1}^{j-1}\big(\mathcal{G}_{lj}^{\mathrm{i}}\times\mathcal{G}_{lj}^{\mathrm{o}}\big)\sqcup\bigsqcup_{l=j+1}^{k}\big(\mathcal{G}_{jl}^{\mathrm{i}}\times\mathcal{G}_{jl}^{\mathrm{o}}\big),\tag{A.5}$$

according to the idea that each non-crossing graph either

(i) has j as an isolated vertex, or

(ii) has a maximal $l<j$ with $(lj)\in E$, and the graph can be written as the product of a graph *inside* and a graph *outside* the interval $[l,j]$, or

(iii) has no $l<j$ with $(lj)\in E$ but there is a maximal $l>j$ with $(jl)\in E$, and the graph can be written as the product of a graph *inside* and a graph *outside* the interval $[j,l]$.

The corresponding formal definitions used in (A.5) are given

$$\begin{aligned}&\mathcal{G}_j:=\mathrm{NCG}([1,k]\setminus\{j\})\\&\mathcal{G}_{lj}^{\mathrm{i}}:=\mathrm{NCG}[l,j),\qquad\mathcal{G}_{lj}^{\mathrm{o}}:=\{E\in\mathrm{NCG}([1,l]\cup[j,k])\mid(lj)\in E\}\\&\mathcal{G}_{jl}^{\mathrm{i}}:=\{E\in\mathrm{NCG}[j,l]\mid(jl)\in E\},\qquad\mathcal{G}_{jl}^{\mathrm{o}}:=\mathrm{NCG}([1,j)\cup[l,k]).\end{aligned}\tag{A.6}$$

We note that for graphs $E\in\mathrm{NCG}[1,k]$ with an isolated vertex j whose edge-set is given by the edge-set $E=E_1\in\mathcal{G}_j$ of its restriction to $[1,k]\setminus\{j\}$ we have

$$\mathrm{pTr}_{K(\pi(E))}(\boldsymbol{A}_{[1,k)})=\mathrm{pTr}_{K(\pi(E_1))}(\boldsymbol{A}_{[1,j-2]},A_{j-1}A_j,\boldsymbol{A}_{[j+1,k)}).\tag{A.7}$$

Similarly for $E=E_1\cup E_2$ with $E_1\in\mathcal{G}_{lj}^{\mathrm{i}},E_2\in\mathcal{G}_{lj}^{\mathrm{o}}$ for some $l<j$ we have

$$\mathrm{pTr}_{\pi(E)}(\boldsymbol{A}_{[1,k)})=\langle\mathrm{pTr}_{K(\pi(E_1))}(\boldsymbol{A}_{[l,j-2]})A_{j-1}\rangle\,\mathrm{pTr}_{K(\pi(E_2))}(\boldsymbol{A}_{[1,l)},I,\boldsymbol{A}_{[j,k)})\tag{A.8}$$

since the vertices $l+1,\dots,j-1$ are necessarily in distinct connected components than the vertices $1,\dots,l-1,j+1,\dots,k$ due to the non-crossing property. Finally, for $E=E_1\cup E_2$ with $E_1\in\mathcal{G}_{jl}^{\mathrm{i}},E_2\in\mathcal{G}_{lj}^{\mathrm{o}}$ for some $l>j$ we have

$$\mathrm{pTr}_{\pi(E)}(\boldsymbol{A}_{[1,k)})=\langle\mathrm{pTr}_{K(\pi_1)}(\boldsymbol{A}_{[j,l)})\rangle\,\mathrm{pTr}_{K(\pi(E_2))}(\boldsymbol{A}_{[1,j)},\boldsymbol{A}_{[l,k)})\tag{A.9}$$

by the same reasoning.

Using this decomposition in (A.4), we thus obtain

$$\begin{aligned}&\frac{M(z_1,\dots,z_k)}{m_1\cdots m_k}\\&=\sum_{E\in\mathcal{G}_j}q_E\,\mathrm{pTr}_{K(\pi(E))}(\boldsymbol{A}_{[1,j-2]},A_{j-1}A_j,\boldsymbol{A}_{[j+1,k)})\\&\quad+\sum_{l=1}^{j-1}\sum_{E_1\in\mathcal{G}_{lj}^{\mathrm{i}}}q_{E_1}\langle\mathrm{pTr}_{K(\pi(E_1))}(\boldsymbol{A}_{[l,j-2]})A_{j-1}\rangle\sum_{E_2\in\mathcal{G}_{lj}^{\mathrm{o}}}q_{E_2}\,\mathrm{pTr}_{K(\pi(E_2))}(\boldsymbol{A}_{[1,l)},I,\boldsymbol{A}_{[j,k)})\\&\quad+\sum_{l=j+1}^{k}\sum_{E_1\in\mathcal{G}_{jl}^{\mathrm{i}}}q_{E_1}\langle\mathrm{pTr}_{K(\pi_1)}(\boldsymbol{A}_{[j,l)})\rangle\sum_{E_2\in\mathcal{G}_{jl}^{\mathrm{o}}}q_{E_2}\,\mathrm{pTr}_{K(\pi(E_2))}(\boldsymbol{A}_{[1,j)},\boldsymbol{A}_{[l,k)}).\end{aligned}\tag{A.10}$$

By (A.4) it follows directly that

$$\begin{aligned}\sum_{E_1\in\mathcal{G}^{\mathrm{i}}_{lj}} q_{E_1}\,\mathrm{pTr}_{K(\pi(E_1))}(\boldsymbol{A}_{[l,j-2]}) &= \frac{M(z_l,A_l,\dots,A_{j-2},z_{j-1})}{m_l\cdots m_{j-1}}\\ \sum_{E_2\in\mathcal{G}^{\mathrm{o}}_{jl}} q_{E_2}\,\mathrm{pTr}_{K(\pi(E_2))}(\boldsymbol{A}_{[1,j)},\boldsymbol{A}_{[l,k)}) &= \frac{M(z_1,\dots,A_{j-1},z_l,\dots,z_k)}{m_1\cdots m_{j-1}m_l\cdots m_k},\end{aligned} \tag{A.11}$$

while for $\mathcal{G}^{\mathrm{o}}_{lj}$ and $\mathcal{G}^{\mathrm{i}}_{jl}$ we note that the graphs with or without the edges (lj) or (jl), respectively, give exactly the same tracial expression, and therefore

$$\begin{aligned}\sum_{E_2\in\mathcal{G}^{\mathrm{o}}_{lj}} q_{E_2}\,\mathrm{pTr}_{K(\pi(E_2))}(\boldsymbol{A}_{[1,l)},I,\boldsymbol{A}_{[j,k)}) &= \frac{q_{lj}}{1+q_{lj}}\frac{M(z_1,\dots,z_l,I,z_j,\dots,z_k)}{m_1\cdots m_l m_j\cdots m_k}\\ \sum_{E_1\in\mathcal{G}^{\mathrm{i}}_{jl}} q_{E_1}\langle\mathrm{pTr}_{K(\pi_1)}(\boldsymbol{A}_{[j,l)})\rangle &= \frac{q_{jl}}{1+q_{jl}}\frac{M(z_j,\dots,z_l)}{m_1\cdots m_l}.\end{aligned} \tag{A.12}$$

The claim now follows from using (A.11) and (A.12) within (A.10) and using $q_{lj}/(1+q_{lj}) = m_l m_j$. □

Proof of Lemma 6.1. Let $\epsilon > 0$ be arbitrary small and set $J := N^\epsilon$. For any $|x| \le 2$, define $z(x,J) = x + \mathrm{i}\eta(x,J)$ where $\eta(x,J)$ is uniquely defined implicitly via the equation $N\eta(x,J)\rho(z(x,J)) = J$. Note that $\eta(x,J) \gtrsim N^{-1+\epsilon}$. Denote by λ_i the eigenvalues of W and by $\boldsymbol{u}_i$ the corresponding orthonormal eigenvectors. Additionally, we define the quantiles γ_i implicitly by

$$\int_{-\infty}^{\gamma_i}\rho_{\mathrm{sc}}(x)\,\mathrm{d}x = \frac{i}{N},\qquad i\in[N] \tag{A.13}$$

and we recall the *rigidity* bound (see e.g. [28, Theorem 7.6] or [34])

$$\big|\lambda_i-\gamma_i\big| \prec \frac{1}{N^{2/3}(N+1-i)^{1/3}},\qquad 1\le i\le N.$$

Using this eigenvalue rigidity and the spectral decomposition of W, it is easy to see the following bound on the *overlaps* of the eigenvectors with a test matrix A

$$\begin{aligned}|\langle\boldsymbol{u}_i,A\boldsymbol{u}_j\rangle|^2 &\prec \frac{1}{N\rho(z(\gamma_i,J))\rho(z(\gamma_j,J))}\langle\Im G(z(\gamma_i,J))A\Im G(z(\gamma_j,J))A\rangle\\ &\prec \frac{1}{N\rho(z(\gamma_i,J))\rho(z(\gamma_j,J))}\end{aligned} \tag{A.14}$$

for any $i,j\in[N]$. Here we neglected N^ϵ-factors since $\epsilon>0$ is arbitrary small and eventually it can be incorporated in the $\prec$-notation. Note that in the last inequality of (A.14) we used (2.11a) with $k=a=2$ and that the corresponding deterministic term, a linear combination of $\langle M(z_i,A_1,z_j)A_2\rangle$ is bounded, see (2.10), where $z_i = z(\gamma_i,J)$ or $z_i = \bar{z}(\gamma_i,J)$.

Given the overlap bound (A.14), we now present the proof of (6.9); the proof of (6.10) is completely analogous and so omitted. By spectral decomposition for each resolvent together with (A.14), using that $\rho(z(x,J)) \sim \rho(x+\mathrm{i}N^{-2/3})$ for any $|x|\le 2$ (modulo N^ϵ-factor), we find that

$$\begin{aligned}|\langle G(z_1)AG(z_2)\dots AG(z_k)\rangle| &\prec N^{k/2-1}\prod_{j=1}^k\frac{1}{N}\sum_{i=1}^N\frac{1}{|\lambda_i-z_j|\rho(\gamma_i+\mathrm{i}N^{-2/3})}\\ &\prec N^{k/2-1}\prod_{j=1}^k\frac{1}{\rho(x_j+\mathrm{i}N^{-2/3})}\left(1+\frac{1}{N|\eta_j|}\right),\end{aligned} \tag{A.15}$$

where we used that

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\lambda_i-z_j|\rho(\gamma_i+\mathrm{i}N^{-2/3})}\prec\frac{1}{N}&\sum_{|i-i_0|\le N^\delta}\frac{1}{|\lambda_i-z_j|\rho(\gamma_i+\mathrm{i}N^{-2/3})}\\&+\frac{1}{N}\sum_{|i-i_0|>N^\delta}\frac{1}{|\gamma_i-\gamma_{i_0}|\rho(\gamma_i+\mathrm{i}N^{-2/3})}\\&\prec\frac{1}{\rho(x_j+\mathrm{i}N^{-2/3})}\left(1+\frac{1}{N|\eta_j|}\right).\end{aligned}\tag{A.16}$$

Here $\delta>0$ is an arbitrary small constant (and we neglected N^δ-factors since eventually it can be incorporated in the $\prec$-notation), and $i_0=i_0(j)$ is the index such that $\gamma_{i_0(j)}$ is the closest quantile to the fixed $x_j=\Re z_j$. In the first inequality in (A.16) we used rigidity to replace λ_i and z_j with the closest quantiles. In the last step in (A.16) we first used that $\rho(\gamma_i+\mathrm{i}N^{-2/3})$ and $\rho(x_j+\mathrm{i}N^{-2/3})$ are comparable up to an N^δ factor, again by rigidity, and then we used the trivial bound $1/|\lambda_i-z_j|\le 1/|\eta_j|$ in the first sum and performed the second sum using the regular spacing of the quantiles. □

B Proof of the multi-resolvent local law in the $d\ge 1$ regime

The $d\ge 1$ regime is conceptually much simpler than $d\le 1$ for several reasons. First, there is no need to keep track of the traceless matrices separately. Second, the trivial norm estimate $\|G(z)\|\le 1/d$ is affordable without much loss. These two facts mean that long chains of the form $GAGA\dots G$ can affordably be reduced to much shorter chains by estimating intermediate A and G factors simply by norm. This trivially takes care of the reduction problem, the key difficulty in the proof when $d\le 1$; in particular no analogue of Lemma 3.6 is needed. Furthermore, we will not need to introduce the quantities $\Psi^{\mathrm{iso/av}}$ and $\psi^{\mathrm{iso/av}}$ and gradually improve the estimate on them; the system of master inequalities reduces to a simple induction on the length k of the resolvent chain.

We will present the proof of the averaged law (2.11a) for $d\ge 1$, the corresponding isotropic law (2.11b) is completely analogous and will be omitted. The backbone of the argument is a very simplified form of Section 4. For notational simplicity, we again do not carry the precise dependence of the resolvents on the spectral parameters and we denote every deterministic matrix A_i generically by A. Note that A's are not necessarily traceless.

We prove (2.11a) by induction on k, the initial $k=1$ case will be proven along the way. We now fix some $k\ge 1$ and in the case $k\ge 2$, we assume that (2.11a) has been proven for all resolvent chains of length at most $k-1$. The starting point of the proof of (2.11a) for k is formula (4.9) that we repeat here

$$\begin{aligned}&\langle (GA)^k\rangle\Big(1+\mathcal{O}_\prec\Big(\frac{1}{Nd^2}\Big)\Big)\\&\qquad=m\langle A(GA)^{k-1}\rangle+m\sum_{j=1}^{k-1}\langle (GA)^jG\rangle\langle (GA)^{k-j}\rangle-m\langle\underline{W(GA)^k}\rangle.\end{aligned}\tag{B.1}$$

Note that the $1/(N\eta)$ in the error term in the lhs. is replaced with $1/(Nd^2)$ since it came from the standard single resolvent local law from theorem 2.3. Notice that all but one chains in the rhs. of (B.1) have less than k resolvents, these can be approximated by their deterministic counterparts using the induction hypothesis of the form

$$\begin{aligned} \left|\langle A(GA)^{k-1}\rangle - \langle AM_{k-1}A\rangle\right| &\prec \frac{1}{Nd^k}, && k\ge 2 \\ |\langle (GA)^jG - M_{j+1}\rangle| &\prec \frac{1}{Nd^{j+2}}, && 1\le j\le k-2 \\ \left|\langle (GA)^{k-j}\rangle - \langle M_{k-j}A\rangle\right| &\prec \frac{1}{Nd^{k-j+1}}, && j\le k-1. \end{aligned} \tag{B.2}$$

The $k=1$ case is particularly simple, since the first term in the rhs. of (B.1) is simply $m\langle A\rangle$ and the sum is absent. In the $k\ge 2$ case, for the remaining $\langle (GA)^{k-1}G\rangle$ term we instead use the integral representation (3.14) and (3.15) in order to also estimate this term using the induction hypothesis as

$$|\langle (GA)^{k-1}G - M_k\rangle| \prec \frac{1}{Nd^{k+1}}. \tag{B.3}$$

Thus, similarly to the telescopic summation (4.11) and using the deterministic identity (4.7), we obtain the following analogue of (4.12):

$$\langle (GA)^k - M_kA\rangle = -m\langle \underline{W(GA)^k}\rangle + \mathcal{O}_\prec\left(\widetilde{\mathcal{E}}_k^{\mathrm{av}}\right), \quad \text{with} \quad \widetilde{\mathcal{E}}_k^{\mathrm{av}} := \frac{1}{Nd^{k+1}}, \tag{B.4}$$

where the error term $\widetilde{\mathcal{E}}_k^{\mathrm{av}}$ has been appropriately redefined compared with (4.12).

Now we fix any integer p and compute the $2p$-th moment of the lhs. of (B.4) exactly as in (4.15) with the definition of Ξ_k^{av} given in (4.16). We follow the calculation from (4.15) through (4.27) but the estimates are greatly simplified as follows. Instead of (4.17) we now have

$$|m|\frac{|\langle (GA)^{2k}G\rangle| + |\langle (GA)^k(G^*A)^kG^*\rangle|}{N^2} \prec \frac{1}{N^2d^{2k+2}} = \left(\widetilde{\mathcal{E}}_k^{\mathrm{av}}\right)^2 \tag{B.5}$$

by a trivial norm bound and $d\ge 1$. Note that we exploited the additional decay $|m|\lesssim 1/d$ unlike in (4.17) where $|m|\lesssim 1$ was used.

Now we turn to the estimate of Ξ_k^{av}. The naive bounds (4.19) become

$$|\partial^{\boldsymbol{l}}((GA)^k)_{ba}| \prec \frac{1}{d^{k+|\boldsymbol{l}|}}, \qquad |\partial^{\boldsymbol{j}}\langle (G^{(*)}A)^k\rangle| \prec \frac{1}{Nd^{k+|\boldsymbol{j}|}} \tag{B.6}$$

as long as $\boldsymbol{j}\ne 0$, and they again follow from the trivial norm estimates. Using these bounds in (4.16), we have

$$\Xi_k^{\mathrm{av}} \prec N^{-(|\boldsymbol{l}|+\sum(J\cup J_*)+3)/2}N^2\frac{1}{d^{k+|\boldsymbol{l}|}}\left(\frac{1}{Nd^{k+1}}\right)^{\sum(J\cup J_*)} \le N^{(1-|\boldsymbol{l}|)/2}\left(\widetilde{\mathcal{E}}_k^{\mathrm{av}}\right)^{1+\sum(J\cup J_*)}. \tag{B.7}$$

If $|\boldsymbol{l}|\ge 1$, then this naive bound is already sufficient. When $|\boldsymbol{l}|=0$, then we perform the $\sum_{ab}$ summation a bit more carefully, similarly to the second line of (4.22):

$$\sum_{ab}|((GA)^k)_{ba}| \le N^{3/2}\sqrt{\langle (GA)^{k-1}GG^*(AG^*)^{k-1}\rangle} \prec N^{3/2}d^{-k}.$$

Note that this bound gains a factor $1/\sqrt{N}$ compared to the trivial bound in (B.7) since the double sum now contributes only by a factor $N^{3/2}$ instead of N^2. This gain is sufficient to improve (B.7) to

$$\Xi_k^{\mathrm{av}} \prec \left(\widetilde{\mathcal{E}}_k^{\mathrm{av}}\right)^{1+\sum(J\cup J_*)}. \tag{B.8}$$

Plugging this estimate together with (B.5) into (4.15), using a Young inequality as we did when going from (4.25) to (4.26) and recalling that p was arbitrary, we obtain

$$|\langle (GA)^k - M_kA\rangle| \prec \widetilde{\mathcal{E}}_k^{\mathrm{av}}$$

i.e. we proved (2.11a) in the $d \geq 1$ regime.

We omit the proof of (2.11b) in the same regime since can be obtained analogously, following a substantial simplification of the argument in Section 4.1.2 along the same lines as the average bound was simplified following Section 4.1.1.

References

[1] A. Adhikari and J. Huang, *Dyson Brownian motion for general ß and potential at the edge*, Probab. Theory Related Fields **178**, 893–950 (2020), MR4168391.

[2] A. Aggarwal, *Bulk universality for generalized Wigner matrices with few moments*, Probab. Theory Related Fields **173**, 375–432 (2019), MR3916110.

[3] O. H. Ajanki, L. Erdős, and T. Krüger, *Stability of the matrix Dyson equation and random matrices with correlations*, Probab. Theory Related Fields **173**, 293–373 (2019), MR3916109.

[4] O. H. Ajanki, L. Erdős, and T. Krüger, *Universality for general Wigner-type matrices*, Probab. Theory Related Fields **169**, 667–727 (2017), MR3719056.

[5] J. Alt, L. Erdős, T. Krüger, and Y. Nemish, *Location of the spectrum of Kronecker random matrices*, Ann. Inst. Henri Poincaré Probab. Stat. **55**, 661–696 (2019), MR3949949.

[6] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, Vol. 118, Cambridge Studies in Advanced Mathematics (Cambridge University Press, Cambridge, 2010), pp. xiv+492, MR2760897.

[7] Z. Bao and Y. He, *Quantitative CLT for linear eigenvalue statistics of Wigner matrices*, preprint (2021), arXiv:2103.05402.

[8] R. Bauerschmidt, J. Huang, A. Knowles, and H.-T. Yau, *Edge rigidity and universality of random regular graphs of intermediate degree*, Geom. Funct. Anal. **30**, 693–769 (2020), MR4135670.

[9] R. Bauerschmidt, J. Huang, and H.-T. Yau, *Local Kesten-McKay law for random regular graphs*, Comm. Math. Phys. **369**, 523–636 (2019), MR3962004.

[10] R. Bauerschmidt, A. Knowles, and H.-T. Yau, *Local semicircle law for random regular graphs*, Comm. Pure Appl. Math. **70**, 1898–1960 (2017), MR3688032.

[11] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Isotropic local laws for sample covariance and generalized Wigner matrices*, Electron. J. Probab. **19**, no. 33, 53 (2014), MR3183577.

[12] C. Bordenave and A. Guionnet, *Localization and delocalization of eigenvectors for heavy-tailed random matrices*, Probab. Theory Related Fields **157**, 885–953 (2013), MR3129806.

[13] P. Bourgade, F. Yang, H.-T. Yau, and J. Yin, *Random band matrices in the delocalized phase, II: generalized resolvent estimates*, J. Stat. Phys. **174**, 1189–1221 (2019), MR3934695.

[14] P. Bourgade, L. Erdős, and H.-T. Yau, *Bulk universality of general ß-ensembles with non-convex potential*, J. Math. Phys. **53**, 095221, 19 (2012), MR2905803.

[15] P. Bourgade, L. Erdős, and H.-T. Yau, *Edge universality of beta ensembles*, Comm. Math. Phys. **332**, 261–353 (2014), MR3253704.

[16] P. Bourgade, L. Erdős, and H.-T. Yau, *Universality of general ß-ensembles*, Duke Math. J. **163**, 1127–1190 (2014), MR3192527.

[17] P. Bourgade, K. Mody, and M. Pain, *Optimal local law and central limit theorem for ß-ensembles*, Comm. Math. Phys. **390**, 1017–1079 (2022), MR4389077.

[18] C. Cacciapuoti, A. Maltsev, and B. Schlein, *Bounds for the Stieltjes transform and the density of states of Wigner matrices*, Probab. Theory Related Fields **163**, 1–59 (2015), MR3405612.

[19] G. Cipolloni and L. Erdős, *Fluctuations for differences of linear eigenvalue statistics for sample covariance matrices*, Random Matrices Theory Appl. **9**, 2050006, 32 (2020), MR4119592.

[20] G. Cipolloni, L. Erdős, and D. Schröder, *Central Limit Theorem for Linear Eigenvalue Statistics of non-Hermitian Random Matrices*, Comm. Pure Appl. Math. (2019), arXiv:1912.04100.

[21] G. Cipolloni, L. Erdős, and D. Schröder, *Eigenstate thermalization hypothesis for Wigner matrices*, Comm. Math. Phys. **388**, 1005–1048 (2021), MR4334253.

[22] G. Cipolloni, L. Erdős, and D. Schröder, *Functional Central Limit Theorems for Wigner Matrices*, Accepted for publication in Ann. Appl. Probab (2020), arXiv:2012.13218.

[23] G. Cipolloni, L. Erdős, and D. Schröder, *Thermalisation for Wigner matrices*, J. Funct. Anal. **282**, Paper No. 109394, 37 (2022), MR4372147.

[24] T. Claeys, B. Fahs, G. Lambert, and C. Webb, *How much can the eigenvalues of a random Hermitian matrix fluctuate?*, Duke Math. J. **170**, 2085–2235 (2021), MR4278668.

[25] E. B. Davies, *The functional calculus*, J. London Math. Soc. (2) **52**, 166–176 (1995), MR1345723.

[26] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Delocalization and diffusion profile for random band matrices*, Comm. Math. Phys. **323**, 367–416 (2013), MR3085669.

[27] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Spectral statistics of Erdős-Rényi graphs I: Local semicircle law*, Ann. Probab. **41**, 2279–2375 (2013), MR3098073.

[28] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *The local semicircle law for a general class of random matrices*, Electron. J. Probab. **18**, no. 59, 58 (2013), MR3068390.

[29] L. Erdős, T. Krüger, and D. Schröder, *Random matrices with slow correlation decay*, Forum Math. Sigma **7**, e8, 89 (2019), MR3941370.

[30] L. Erdős, B. Schlein, and H.-T. Yau, *Local semicircle law and complete delocalization for Wigner random matrices*, Comm. Math. Phys. **287**, 641–655 (2009), MR2481753.

[31] L. Erdős and D. Schröder, *Fluctuations of rectangular Young diagrams of interlacing Wigner eigenvalues*, Int. Math. Res. Not. IMRN, 3255–3298 (2018), MR3805203.

[32] L. Erdős and H.-T. Yau, *A dynamical approach to random matrix theory*, Vol. 28, Courant Lecture Notes in Mathematics (Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2017), pp. ix+226, MR3699468.

[33] L. Erdős, H.-T. Yau, and J. Yin, *Bulk universality for generalized Wigner matrices*, Probab. Theory Related Fields **154**, 341–407 (2012), MR2981427.

[34] L. Erdős, H.-T. Yau, and J. Yin, *Rigidity of eigenvalues of generalized Wigner matrices*, Adv. Math. **229**, 1435–1515 (2012), MR2871147.

[35] F. Götze, A. Naumov, and A. Tikhomirov, *Local semicircle law under fourth moment condition*, J. Theoret. Probab. **33**, 1327–1362 (2020), MR4125959.

[36] Y. He and A. Knowles, *Mesoscopic eigenvalue density correlations of Wigner matrices*, Probab. Theory Related Fields **177**, 147–216 (2020), MR4095015.

[37] Y. He and A. Knowles, *Mesoscopic eigenvalue statistics of Wigner matrices*, Ann. Appl. Probab. **27**, 1510–1550 (2017), MR3678478.

[38] Y. He, A. Knowles, and R. Rosenthal, *Isotropic self-consistent equations for mean-field random matrices*, Probab. Theory Related Fields **171**, 203–249 (2018), MR3800833.

[39] J. Huang and B. Landon, *Rigidity and a mesoscopic central limit theorem for Dyson Brownian motion for general ß and potentials*, Probab. Theory Related Fields **175**, 209–253 (2019), MR4009708.

[40] A. Knowles and J. Yin, *The isotropic semicircle law and deformation of Wigner matrices*, Comm. Pure Appl. Math. **66**, 1663–1750 (2013), MR3103909.

[41] G. Kreweras, *Sur les partitions non croisees d'un cycle*, Discrete Math. **1**, 333–350 (1972), MR309747.

[42] J. O. Lee and K. Schnelli, *Local deformed semicircle law and complete delocalization for Wigner matrices with random potential*, J. Math. Phys. **54**, 103504, 62 (2013), MR3134604.

[43] J. O. Lee and K. Schnelli, *Local law and Tracy-Widom limit for sparse random matrices*, Probab. Theory Related Fields **171**, 543–616 (2018), MR3800840.

[44] J. O. Lee, K. Schnelli, B. Stetler, and H.-T. Yau, *Bulk universality for deformed Wigner matrices*, Ann. Probab. **44**, 2349–2425 (2016), MR3502606.

[45] Y. Li, *Rigidity of Eigenvalues for beta Ensemble in Multi-Cut Regime*, Thesis (Ph.D.)–Brandeis University (ProQuest LLC, Ann Arbor, MI, 2017), p. 282, MR3755113.

[46] Y. Li, K. Schnelli, and Y. Xu, *Central limit theorem for mesoscopic eigenvalue statistics of deformed Wigner matrices and sample covariance matrices*, Ann. Inst. Henri Poincaré Probab. Stat. **57**, 506–546 (2021), MR4255183.

[47] Y. Li and Y. Xu, *On fluctuations of global and mesoscopic linear statistics of generalized Wigner matrices*, Bernoulli **27**, 1057–1076 (2021), MR4255226.

[48] J. A. Mingo and R. Speicher, *Free probability and random matrices*, Vol. 35, Fields Institute Monographs (Springer, New York; Fields Institute for Research in Mathematical Sciences, Toronto, ON, 2017), pp. xiv+336, MR3585560.

[49] P. Sosoe and P. Wong, *Local semicircle law in the bulk for Gaussian ß-ensemble*, J. Stat. Phys. **148**, 204–232 (2012), MR2966359.

[50] F. Yang, H.-T. Yau, and J. Yin, *Delocalization and quantum diffusion of random band matrices in high dimensions I: Self-energy renormalization*, preprint (2021), arXiv:2104.12048.

[51] F. Yang, H.-T. Yau, and J. Yin, *Delocalization and quantum diffusion of random band matrices in high dimensions II: T-expansion*, preprint (2021), arXiv:2107.05795.

[52] F. Yang and J. Yin, *Random band matrices in the delocalized phase, III: averaging fluctuations*, Probab. Theory Related Fields **179**, 451–540 (2021), MR4221663.