CENTRAL LIMIT THEOREM FOR LINEAR EIGENVALUE STATISTICS OF NON-HERMITIAN RANDOM MATRICES

GIORGIO CIPOLLONI[†‡] AND LÁSZLÓ ERDŐS[†]

IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria

DOMINIK SCHRÖDER[*]

Institute for Theoretical Studies, ETH Zurich, Clausiusstr. 47, 8092 Zurich, Switzerland

ABSTRACT. We consider large non-Hermitian random matrices X with complex, independent, identically distributed centred entries and show that the linear statistics of their eigenvalues are asymptotically Gaussian for test functions having $2 + \epsilon$ derivatives. Previously this result was known only for a few special cases; either the test functions were required to be analytic [73], or the distribution of the matrix elements needed to be Gaussian [74], or at least match the Gaussian up to the first four moments [82], [56]. We find the exact dependence of the limiting variance on the fourth cumulant that was not known before. The proof relies on two novel ingredients: (i) a local law for a product of *two* resolvents of the Hermitisation of X with different spectral parameters and (ii) a coupling of several weakly dependent Dyson Brownian Motions. These methods are also the key inputs for our analogous results on the linear eigenvalue statistics of *real* matrices X that are presented in the companion paper [31].

1. INTRODUCTION

Eigenvalues of random matrices form a strongly correlated point process. One manifestation of this fact is the unusually small fluctuation of their linear statistics making the eigenvalue process distinctly different from a Poisson point process. Suppose that the $n \times n$ random matrix X has i.i.d. entries of zero mean and variance $1/n$. The empirical density of the eigenvalues $\{\sigma_i\}_{i=1}^n$ converges to a limit distribution; it is the uniform distribution on the unit disk in the non-Hermitian case *(circular law)* and the semicircular density in the Hermitian case *(Wigner semicircle law)*. For test functions f defined on the spectrum one may consider the fluctuation of the linear statistics and one expects that

$$L_n(f) := \sum_{i=1}^n f(\sigma_i) - \mathbf{E} \sum_{i=1}^n f(\sigma_i) \sim \mathcal{N}(0, V_f) \tag{1.1}$$

converges to a centred normal distribution as $n \to \infty$. The variance V_f is expected to depend only on the second and fourth moments of the single entry distribution. Note that, unlike in the usual central limit theorem, there is no $1/\sqrt{n}$ rescaling in (1.1) which is a quantitative indication of a strong correlation. The main result of the current paper is the proof of (1.1) for non-Hermitian random matrices with complex i.i.d. entries and for general test functions f. We give an explicit formula for V_f that involves the fourth cumulant of X as well, disproving a conjecture by Chafaï [24]. By polarisation, from (1.1) it

also follows that the limiting joint distribution of $(L_n(f_1), L_n(f_2), \ldots, L_n(f_k))$ for a fixed number of test functions is jointly Gaussian.

We remark that another manifestation of the strong eigenvalue correlation is the repulsion between neighbouring eigenvalues. For Gaussian ensembles the local repulsion is directly seen from the well-known determinantal structure of the joint distribution of all eigenvalues; both in the non-Hermitian *Ginibre* case and in the Hermitian *GUE/GOE* case. In the spirit of *Wigner-Dyson-Mehta universality* of the local correlation functions [67] level repulsion should also hold for random matrices with general distributions. While for the Hermitian case the universality has been rigorously established for a large class of random matrices (see e.g. [41] for a recent monograph), the analogous result for the non-Hermitian case is still open in the bulk spectrum (see, however, [30] for the edge regime and [82] for entry distributions whose first four moments match the Gaussian).

These two manifestations of the eigenvalue correlations cannot be deduced from each other, however the proofs often share common tools. For n-independent test functions f, (1.1) apparently involves understanding the eigenvalues only on the macroscopic scales, while the level repulsion is expressly a property on the microscopic scale of individual eigenvalues. However the suppression of the usual $\sqrt{n}$ fluctuation is due to delicate correlations on all scales, so (1.1) also requires understanding local scales.

Hermitian random matrices are much easier to handle, hence fluctuation results of the type (1.1) have been gradually obtained for more and more general matrix ensembles as well as for broader classes of test functions, see, e.g. [53, 9, 55, 66, 75] and [80] for the weakest regularity conditions on f. Considering n-dependent test functions, Gaussian fluctuations have been detected even on mesoscopic scales [22, 23, 49, 51, 52, 35, 58, 63, 11].

Non-Hermitian random matrices pose serious challenges, mainly because their eigenvalues are potentially very unstable. When X has i.i.d. centred Gaussian entries with variance $1/n$ (this is called the *Ginibre ensemble*), the explicit determinantal formulas for the correlation functions may be used to compute the distribution of the linear statistics $L_n(f)$. Forrester in [45] proved (1.1) for complex Ginibre ensemble and radially symmetric f and obtained the variance $V_f = (4\pi)^{-1} \int_{\mathbf{D}} |\nabla f|^2 \, \mathrm{d}^2 z$ where $\mathbf{D}$ is the unit disk. He also gave a heuristic argument based on Coulomb gas theory for general f and his calculations predicted an additional boundary term $\frac{1}{2} \|f\|^2_{\dot{H}^{1/2}(\partial \mathbf{D})}$ in the variance V_f. Rider considered test functions f depending only on the angle [72] when $f \notin H^1(\mathbf{D})$ and accordingly V_f grows with $\log n$ (similar growth is proved for $f = \log$ in [68]). Finally, Rider and Virág in [74] have rigorously verified Forrester's prediction for general $f \in C^1(\mathbf{D})$ using a cumulant formula for determinantal processes found first by Costin and Lebowitz [33] and extended by Soshnikov [79]. They also presented a *Gaussian free field (GFF)* interpretation of the result that we extend in Section 2.1.

The first result beyond the explicitly computable Gaussian case is due to Rider and Silverstein [73, Theorem 1.1] who proved (1.1) for X with i.i.d. complex matrix elements and for test functions f that are analytic on a large disk. Analyticity allowed them to use contour integration and thus deduce the result from analysing the resolvent at spectral parameters far away from the actual spectrum. The domain of analyticity was optimized in [70], where extensions to elliptic ensembles were also proven. Polynomial test functions via the alternative moment method were considered by Nourdin and Peccati in [69]. The analytic method of [73] was recently extended by Coston and O'Rourke [34] to fluctuations of linear statistics for *products* of i.i.d. matrices. However, these method fail for a larger class of test functions.

Since the first four moments of the matrix elements fully determine the limiting eigenvalue statistics, Tao and Vu were able to compare the fluctuation of the local eigenvalue density for a general non-Gaussian X with that of a Ginibre matrix [82, Corollary 10] assuming the first four moments of X match those of the complex Ginibre ensemble. This method was extended by Kopel [56, Corollary 1] to general smooth test functions with an additional study on the real eigenvalues when X is real (see also the work of Simm for polynomial statistics of the real eigenvalues [78]).

Our result removes the limitations of both previous approaches: we allow general test functions and general distribution for the matrix elements without constraints on matching moments. We remark that the dependence of the variance V_f on the fourth cumulant of the single matrix entry escaped all previous works. The Ginibre ensemble with its vanishing fourth cumulant clearly cannot catch this dependence. Interestingly, even though the fourth cumulant in general is not zero in the work Rider

and Silverstein [73], it is multiplied by a functional of f that happens to vanish for analytic functions (see (2.6), (2.8) and Remark 2.5 later). Hence this result did not detect the precise role of the fourth cumulant either. This may have motivated the conjecture [24] that the variance does not depend on the fourth cumulant at all.

In order to focus on the main new ideas, in this paper we consider the problem only for X with genuinely complex entries. Our method also works for real matrices where the real axis in the spectrum plays a special role that modifies the exact formula for the expectation and the variance V_f in (1.1). This leads to some additional technical complications that we have resolved in a separate work [31] which contains the real version of our main Theorem 2.2.

Finally, we remark that the problem of fluctuations of linear statistics has been considered for β-log-gases in one and two dimensions; these are closely related to the eigenvalues of the Hermitian, resp. non-Hermitian Gaussian matrices for classical values $\beta = 1, 2, 4$ and for quadratic potential. In fact, in two dimensions the logarithmic interaction also corresponds to the Coulomb gas from statistical physics. Results analogous to (1.1) in one dimension were obtained e.g. in [53, 76, 17, 14, 57, 13, 52, 1]. In two dimensions similar results have been established both in the macroscopic [61] and in the mesoscopic [12] regimes.

We now outline the main ideas in our approach. We use Girko's formula [47] in the form given in [82] to express linear eigenvalue statistics of X in terms of resolvents of a family of $2n \times 2n$ Hermitian matrices

$$H^z := \begin{pmatrix} 0 & X - z \\ X^* - \overline{z} & 0 \end{pmatrix} \tag{1.2}$$

parametrized by $z \in \mathbf{C}$. This formula asserts that

$$\sum_{\sigma \in \mathrm{Spec}(X)} f(\sigma) = -\frac{1}{4\pi} \int_{\mathbf{C}} \Delta f(z) \int_0^\infty \Im \mathrm{Tr}\, G^z(i\eta) \, \mathrm{d}\eta \, \mathrm{d}^2 z \tag{1.3}$$

for any smooth, compactly supported test function f (the apparent divergence of the η-integral at infinity can easily be removed, see (3.10)). Here we set $G^z(w) := (H^z - w)^{-1}$ to be the resolvent of H^z. We have thus transformed our problem to a Hermitian one and all tools and results developed for Hermitian ensembles in the recent years are available.

Utilizing Girko's formula requires a good understanding of the resolvent of H^z along the imaginary axis for all $\eta > 0$. On very small scales $\eta \ll n^{-1}$, there are no eigenvalues thus $\Im \mathrm{Tr}\, G^z(i\eta)$ is negligible. All other scales $\eta \gtrsim n^{-1}$ need to be controlled carefully since *a priori* they could all contribute to the fluctuation of $L_n(f)$, even though *a posteriori* we find that the entire variance comes from scales $\eta \sim 1$.

In the mesoscopic regime $\eta \gg n^{-1}$, *local laws* from [4, 5] accurately describe the leading order deterministic behaviour of $\frac{1}{n} \mathrm{Tr}\, G^z(i\eta)$ and even the matrix elements $G_{ab}^z(i\eta)$; now we need to identify the next order fluctuating term in the local law. In other words we need to prove a central limit theorem for the traces of resolvents G^z. In fact, based upon (1.3), for the higher k-th moments of $L_n(f)$ we need the joint distribution of $\mathrm{Tr}\, G^{z_l}(i\eta)$ for different spectral parameters $z_1, z_2, \ldots, z_k$. This is one of our main technical achievements. Note that the asymptotic joint Gaussianity of traces of Wigner resolvents $\mathrm{Tr}(H - w_1)^{-1}, \mathrm{Tr}(H - w_2)^{-1}, \ldots$ at different spectral parameters has been obtained in [50, 51]. However, the method of this result is not applicable since the role of the spectral parameter z in (1.2) is very different from w; it is in an off-diagonal position thus these resolvents do not commute and they are not in the spectral resolution of a single matrix.

The microscopic regime, $\eta \sim n^{-1}$, is much more involved than the mesoscopic one. Local laws and their fluctuations are not sufficient, we need to trace the effect of the individual eigenvalues $0 \le \lambda_1^z \le \lambda_2^z, \ldots$ of H^z near zero (the spectrum of H^z is symmetric, we may focus on the positive eigenvalues). Moreover, we need their *joint* distribution for different z parameters which, for arbitrary z's, is not known even in the Ginibre case. We prove, however, that λ_1^z and $\lambda_1^{z'}$ are asymptotically independent if z and z' are far away, say $|z - z'| \ge n^{-1/100}$. A similar result holds simultaneously for several small eigenvalues. Notice that due to the z-integration in (1.3), when the k-th moment of $L_n(f)$ is computed, the integration variables $z_1, z_2, \ldots, z_k$ are typically far away from each other. The resulting

independence of the spectra of $H^{z_1}, H^{z_2}, \ldots$ near zero ensures that the microscopic regime eventually does not contribute to the fluctuation of $L_n(f)$.

The proof of the independence of λ_1^z and $\lambda_1^{z'}$ relies on the analysis of the *Dyson Brownian motion* (DBM) developed in the recent years [41] for the proof of the Wigner-Dyson-Mehta universality conjecture for Wigner matrices. The key mechanism is the fast local equilibration of the eigenvalues $\boldsymbol{\lambda}^z(t) := \{\lambda_i^z(t)\}$ along the stochastic flow generated by adding a small time-dependent Gaussian component to the original matrix. This Gaussian component can then be removed by the *Green function comparison theorem* (GFT). One of the main technical results of [26] (motivated by the analogous analysis in [59] for Wigner matrices that relied on coupling and homogenisation ideas introduced first in [20]) asserts that for any fixed z the DBM process $\boldsymbol{\lambda}^z(t)$ can be pathwise approximated by a similar DBM with a different initial condition by *exactly* coupling the driving Brownian motions in their DBMs. We extend this idea to simultaneously trailing $\boldsymbol{\lambda}^z(t)$ and $\boldsymbol{\lambda}^{z'}(t)$ by their independent Ginibre counterparts. The evolutions of $\boldsymbol{\lambda}^z(t)$ and $\boldsymbol{\lambda}^{z'}(t)$ are not independent since their driving Brownian motions are correlated; the correlation is given by the eigenfunction overlap $\langle u_i^z, u_j^{z'} \rangle \langle v_j^{z'}, v_i^z \rangle$ where $w_i^z = (u_i^z, v_i^z) \in \mathbf{C}^n \times \mathbf{C}^n$ denotes the eigenvector of H^z belonging to λ_i^z. However, this overlap turns out to be small if z and z' are far away and i is not too big. Thus the analysis of the microscopic regime has two ingredients: (i) extending the coupling idea to driving Brownian motions whose distributions are not identical but close to each other; and (ii) proving the smallness of the overlap.

While (i) can be achieved by relatively minor modifications to the proofs in [26], (ii) requires to develop a new type of local law. Indeed, the overlap can be estimated in terms of traces of products of resolvents, $\mathrm{Tr}\, G^z(\mathrm{i}\eta)G^{z'}(\mathrm{i}\eta')$ with $\eta, \eta' \sim n^{-1+\epsilon}$ in the mesoscopic regime. Customary local laws, however, do not apply to a quantity involving *products* of resolvents. In fact, even the leading deterministic term needs to be identified by solving a new type of deterministic Dyson equation. We first show the stability of this new equation using the lower bound on $|z - z'|$. Then we prove the necessary high probability bound for the error term in the Dyson equation by a diagrammatic cumulant expansion adapted to the new situation of product of resolvents. The key novelty is to extract the effect that G^z and $G^{z'}$ are weakly correlated when z and z' are far away from each other.

We close this section with an important remark concerning the proofs for Hermitian versus non-Hermitian matrices. Similarly to Girko's formula (1.3), the linear eigenvalue statistics for Hermitian matrices are also expressed by an integral of the resolvents over all spectral parameters. However, in the corresponding Helffer-Sjöstrand formula, sufficient regularity of f directly neutralizes the potentially singular behaviour of the resolvent near the real axis, giving rise to CLT results even with suboptimal control on the resolvent in the mesoscopic regime. A similar trade-off in (1.3) is not apparent; it is unclear if and how the integration in z could help regularize the η integral. This is a fundamental difference between CLTs for Hermitian and non-Hermitian ensembles that explains the abundance of Hermitian results in contrast to the scarcity of available non-Hermitian CLTs.

Notations and conventions. We introduce some notations we use throughout the paper. For integers $k \in \mathbf{N}$ we use the notation $[k] := \{1, \ldots, k\}$. We write $\mathbf{H}$ for the upper half-plane $\mathbf{H} := \{z \in \mathbf{C} \mid \Im z > 0\}$, $\mathbf{D} \subset \mathbf{C}$ for the open unit disk, and for any $z \in \mathbf{C}$ we use the notation $\mathrm{d}^2 z := 2^{-1}\mathrm{i}(\mathrm{d}z \wedge \mathrm{d}\bar{z})$ for the two dimensional volume form on $\mathbf{C}$. For positive quantities f, g we write $f \lesssim g$ and $f \sim g$ if $f \le Cg$ or $cg \le f \le Cg$, respectively, for some constants $c, C > 0$ which depend only on the constants appearing in (2.1). For any two positive real numbers $\omega_*, \omega^* \in \mathbf{R}_+$ by $\omega_* \ll \omega^*$ we denote that $\omega_* \le c\omega^*$ for some small constant $0 < c \le 1/100$. We denote vectors by bold-faced lower case Roman letters $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^k$, for some $k \in \mathbf{N}$. Vector and matrix norms, $\|\boldsymbol{x}\|$ and $\|A\|$, indicate the usual

Euclidean norm and the corresponding induced matrix norm. For any $2n \times 2n$ matrix A we use the notation $\langle A \rangle := (2n)^{-1} \operatorname{Tr} A$ to denote the normalized trace of A. Moreover, for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{C}^n$ and matrices $A, B \in \mathbf{C}^{2n \times 2n}$ we define

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \sum \overline{x}_i y_i, \qquad \langle A, B \rangle := \langle A^* B \rangle.$$

We will use the concept of "with very high probability" meaning that for any fixed $D > 0$ the probability of the 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.

2. Main results

We consider *complex i.i.d. matrices* X, i.e. $n \times n$ matrices whose entries are independent and identically distributed as $x_{ab} \overset{d}{=} n^{-1/2} \chi$ for some complex random variable χ, satisfying the following:

Assumption 2.1. *We assume that* $\mathbf{E}\, \chi = \mathbf{E}\, \chi^2 = 0$ *and* $\mathbf{E}|\chi|^2 = 1$. *In addition we assume the existence of high moments, i.e. that there exist constants* $C_p > 0$, *for any* $p \in \mathbf{N}$, *such that*

$$\mathbf{E}|\chi|^p \leq C_p. \tag{2.1}$$

The *circular law* [8, 10, 48, 15, 46, 47, 71, 81] asserts that the empirical distribution of eigenvalues $\{\sigma_i\}_{i=1}^n$ of a complex i.i.d. matrix X converges to the uniform distribution on the unit disk $\mathbf{D}$, i.e.

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n f(\sigma_i) = \frac{1}{\pi} \int_{\mathbf{D}} f(z)\, \mathrm{d}^2 z, \tag{2.2}$$

with very high probability for any continuous bounded function f. Our main result is a central limit theorem for the centred *linear statistics*

$$L_n(f) := \sum_{i=1}^n f(\sigma_i) - \mathbf{E} \sum_{i=1}^n f(\sigma_i) \tag{2.3}$$

for general complex i.i.d. matrices and generic test functions f.

In order to state the result we introduce some notations and certain Sobolev spaces. We fix some open bounded $\Omega \subset \mathbf{C}$ containing the closed unit disk $\overline{\mathbf{D}} \subset \Omega$ and having a piecewise C^1-boundary, or, more generally, any boundary satisfying the *cone property* (see e.g. [65, Section 8.7]). We consider test functions $f \in H_0^{2+\delta}(\Omega)$ in the Sobolev space $H_0^{2+\delta}(\Omega)$ which is defined as the completion of the smooth compactly supported functions $C_c^\infty(\Omega)$ under the norm

$$\|f\|_{H^{2+\delta}(\Omega)} := \|(1 + |\xi|)^{2+\delta} \widehat{f}(\xi)\|_{L^2(\Omega)}$$

and we note that by Sobolev embedding such functions are continuously differentiable, and vanish at the boundary of Ω. For notational convenience we identify $f \in H_0^{2+\delta}(\Omega)$ with its extension to all of $\mathbf{C}$ obtained from setting $f \equiv 0$ in $\mathbf{C} \setminus \Omega$. We note that our results can trivially be extended to bounded test functions with non-compact support since due to [5, Theorem 2.1], with high probability, all eigenvalues satisfy $|\sigma_i| \leq 1 + \epsilon$ and therefore non-compactly supported test functions can simply be smoothly cut-off. For h defined on the boundary of the unit disk $\partial \mathbf{D}$ we define its Fourier transform

$$\widehat{h}(k) = \frac{1}{2\pi} \int_0^{2\pi} h(e^{\mathrm{i}\theta}) e^{-\mathrm{i}\theta k}\, \mathrm{d}\theta, \qquad k \in \mathbf{Z}. \tag{2.4}$$

For $f, g \in H_0^{2+\delta}(\Omega)$ we define the homogeneous semi-inner products

$$\langle g, f \rangle_{\dot{H}^{1/2}(\partial \mathbf{D})} := \sum_{k \in \mathbf{Z}} |k| \widehat{f}(k) \overline{\widehat{g}(k)}, \qquad \|f\|_{\dot{H}^{1/2}(\partial \mathbf{D})}^2 := \langle f, f \rangle_{\dot{H}^{1/2}(\partial \mathbf{D})}, \tag{2.5}$$

where, with a slight abuse of notation, we identified f and g with their restrictions to $\partial \mathbf{D}$.

Theorem 2.2 (Central Limit Theorem for linear statistics)**.** *Let* X *be a complex* $n \times n$ *i.i.d. matrix satisfying Assumption 2.1 with eigenvalues* $\{\sigma_i\}_{i=1}^n$, *and denote the fourth cumulant of* χ *by* $\kappa_4 := \mathbf{E}|\chi|^4 -$

2. *Fix $\delta > 0$, an open complex domain Ω with $\overline{\mathbf{D}} \subset \Omega \subset \mathbf{C}$ and a complex valued test function $f \in H_0^{2+\delta}(\Omega)$. Then the centred linear statistics $L_n(f)$, defined in (2.3), converges*

$$L_n(f) \Longrightarrow L(f),$$

to a complex Gaussian random variable $L(f)$ with expectation $\mathbf{E}\, L(f) = 0$ and variance $\mathbf{E}|L(f)|^2 = C(f,f) =: V_f$ and $\mathbf{E}\, L(f)^2 = C(\overline{f}, f)$, where

$$
\begin{aligned}
C\,(g,f) := {} & \frac{1}{4\pi} \langle \nabla g, \nabla f \rangle_{L^2(\mathbf{D})} + \frac{1}{2} \langle g, f \rangle_{\dot{H}^{1/2}(\partial \mathbf{D})} \\
& + \kappa_4 \left(\frac{1}{\pi} \int_{\mathbf{D}} \overline{g(z)}\, \mathrm{d}^2 z - \frac{1}{2\pi} \int_0^{2\pi} \overline{g(e^{i\theta})}\, \mathrm{d}\theta \right) \\
& \times \left(\frac{1}{\pi} \int_{\mathbf{D}} f(z)\, \mathrm{d}^2 z - \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\theta})\, \mathrm{d}\theta \right).
\end{aligned}
\tag{2.6}
$$

More precisely, any finite moment of $L_n(f)$ converges at a rate $n^{-c(k)}$, for some small $c(k) > 0$, i.e.

$$\mathbf{E}\, L_n(f)^k \overline{L_n(f)}^l = \mathbf{E}\, L(f)^k \overline{L(f)}^l + \mathcal{O}\left(n^{-c(k+l)}\right). \tag{2.7}$$

Moreover, the expectation in (2.3) is given by

$$\mathbf{E}\sum_{i=1}^n f(\sigma_i) = \frac{n}{\pi} \int_{\mathbf{D}} f(z)\, \mathrm{d}^2 z + \frac{1}{8\pi} \int_{\mathbf{D}} \Delta f(z)\, \mathrm{d}^2 z - \frac{\kappa_4}{\pi} \int_{\mathbf{D}} f(z)(2|z|^2 - 1)\, \mathrm{d}^2 z + \mathcal{O}\left(n^{-c}\right) \tag{2.8}$$

for some small constant $c > 0$. The implicit constants in the error terms in (2.7)–(2.8) depend on the $H^{2+\delta}$-norm of f and C_p from (2.1).

Remark 2.3 (V_f is strictly positive)**.** *The variance $V_f = \mathbf{E}|L(f)|^2$ in Theorem 2.2 is strictly positive. Indeed, by the Cauchy-Schwarz inequality it follows that*

$$\left| \frac{1}{\pi} \int_{\mathbf{D}} f(z)\, \mathrm{d}^2 z - \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\theta})\, d\theta \right|^2 \leq \frac{1}{8\pi} \int_{\mathbf{D}} |\nabla f|^2\, \mathrm{d}^2 z.$$

Hence, since $\kappa_4 \geq -1$ in (2.6), this shows that

$$V_f \geq \frac{1}{8\pi} \int_{\mathbf{D}} |\nabla f|^2\, \mathrm{d}^2 z + \frac{1}{2} \|f\|_{\dot{H}^{1/2}(\partial \mathbf{D})}^2 > 0.$$

By polarisation, a multivariate Central Limit Theorem readily follows from Theorem 2.2:

Corollary 2.4. *Let X be an $n \times n$ i.i.d. complex matrix satisfying Assumption 2.1, and let $L_n(f)$ be defined in (2.3). For a fixed open bounded complex domain Ω with $\overline{\mathbf{D}} \subset \Omega \subset \mathbf{C}$, $\delta > 0$, $p \in \mathbf{N}$ and for any finite collection of test functions $f^{(1)}, \dots, f^{(p)} \in H_0^{2+\delta}(\Omega)$ the vector*

$$(L_n(f^{(1)}), \dots, L_n(f^{(p)})) \Longrightarrow (L(f^{(1)}), \dots, L(f^{(p)})), \tag{2.9}$$

converges to a multivariate complex Gaussian of zero expectation $\mathbf{E}\, L(f) = 0$ and covariance $\mathbf{E}\, L(f)\overline{L(g)} = \mathbf{E}\, L(f)L(\overline{g}) = C(f,g)$ with C as in (2.6). Moreover, for any mixed k-moments we have an effective convergence rate of order $n^{-c(k)}$, as in (2.7)

Remark 2.5. *We may compare Theorem 2.2 with the previous results in [84, Eq. (5.18)], [74, Theorem 1] and [73, Theorem 1.1]:*

   (i) *For $\kappa_4 = 0$ the expansion (2.8) agrees with the subleading order correction to the circular law from [84, Eq. (5.18)] (see also [86, Eq. (5.16)] and [62, Eq. (1.14)]).*

   (ii) *Note that for a single $f \colon \mathbf{C} \to \mathbf{R}$ in the Ginibre case, i.e. $\kappa_4 = 0$, Theorem 2.2 implies [74, Theorem 1] with $\sigma_f^2 + \widetilde{\sigma}_f^2 = C(f,f)$, using the notation therein and with $C(f,f)$ defined in (2.6).*

   (iii) *If additionally f is complex analytic in a neighbourhood of $\overline{\mathbf{D}}$, using the notation $\partial := \partial_z$, the expressions in (2.6),(2.8) of Theorem 2.2 simplify to*

$$\mathbf{E}\sum_{i=1}^n f(\sigma_i) = n f(0) + \mathcal{O}\left(n^{-\delta'}\right), \quad C\,(f,g) = \frac{1}{\pi} \int_{\mathbf{D}} \partial f(z) \overline{\partial g(z)}\, \mathrm{d}^2 z, \tag{2.10}$$

where we used that for any f, g complex analytic in a neighbourhood of $\overline{\mathbf{D}}$ we have

$$\frac{1}{2\pi}\int_{\mathbf{D}}\langle \nabla g, \nabla f\rangle \,\mathrm{d}^2 z = \frac{1}{\pi}\int_{\mathbf{D}}\partial f(z)\overline{\partial g(z)}\,\mathrm{d}^2 z = \sum_{k\in\mathbf{Z}}|k|\widehat{f\restriction_{\partial\mathbf{D}}}(k)\overline{\widehat{g\restriction_{\partial\mathbf{D}}}(k)}, \qquad (2.11)$$

and that

$$\frac{1}{\pi}\int_{\mathbf{D}}f(z)\,\mathrm{d}^2 z = \frac{1}{2\pi}\int_0^{2\pi}f(e^{i\theta})\,\mathrm{d}\theta = f(0).$$

The second equality in (2.11) *follows by writing f and g in Fourier series. The result in* (2.10) *exactly agrees with* [73, *Theorem 1.1*].

Remark 2.6 (Mesoscopic regime). *We formulated our result for* macroscopic *linear statistics, i.e. for test functions f that are independent of n. One may also consider* mesoscopic *linear statistics as well when $f(\sigma)$ is replaced with $\varphi(n^a(\sigma - z_0))$ for some fixed scale $a > 0$, reference point $z_0 \in \mathbf{D}$ and function $\varphi \in H^{2+\delta}(\mathbf{C})$. Our proof can directly handle this situation as well for any small $a \leq 1/500$[1], say, since all our error terms are effective as a small power of $1/n$. For $a > 0$ the leading term to the variance V_f comes solely from the $\|\nabla f\|^2$ term in* (2.6)*, in particular the effect of the fourth cumulant is negligible.*

2.1. **Connection to the Gaussian free field.** It has been observed in [74] that for the special case $\kappa_4 = 0$ the limiting random field $L(f)$ can be viewed as a variant of the *Gaussian free field* [77]. The Gaussian free field on some bounded domain $\Omega \subset \mathbf{C}$ can formally be defined as a *Gaussian Hilbert space* of random variables $h(f)$ indexed by functions in the homogeneous Sobolev space $f \in \dot{H}_0^1(\Omega)$ such that the map $f \mapsto h(f)$ is linear and

$$\mathbf{E}\, h(f) = 0, \quad \mathbf{E}\, \overline{h(f)}h(g) = \langle f, g\rangle_{\dot{H}^1(\Omega)}. \qquad (2.12)$$

Here for $\Omega \subset \mathbf{C}$ we defined the homogeneous Sobolev space $\dot{H}_0^1(\Omega)$ as the completion of smooth compactly supported function $C_c^\infty(\Omega)$ with respect to the semi-inner product

$$\langle g, f\rangle_{\dot{H}^1(\Omega)} := \langle \nabla g, \nabla f\rangle_{L^2(\Omega)}, \qquad \|f\|_{\dot{H}^1(\Omega)}^2 := \langle f, f\rangle_{\dot{H}^1(\Omega)}.$$

By the Poincaré inequality the space $\dot{H}_0^1(\Omega)$ is in fact a Hilbert space and as a vector space coincides with the usual Sobolev space $H_0^1(\Omega)$ with an equivalent norm but a different scalar product.

Since $\overline{\mathbf{D}} \subset \Omega$, the Sobolev space $\dot{H}_0^1(\Omega)$ can be orthogonally decomposed as

$$\dot{H}_0^1(\Omega) = \dot{H}_0^1(\mathbf{D}) \oplus \dot{H}_0^1(\overline{\mathbf{D}}^c) \oplus \dot{H}_0^1((\partial\mathbf{D})^c)^\perp,$$

where the complements are understood as the complements within Ω. The orthogonal complement $\dot{H}_0^1((\partial\mathbf{D})^c)^\perp$ is (see e.g. [77, Thm. 2.17]) given by the closed subspace of functions which are harmonic in $\mathbf{D} \cup \overline{\mathbf{D}}^c = (\partial\mathbf{D})^c$, i.e. away from the unit circle. For closed subspaces $S \subset \dot{H}_0^1(\Omega)$ we denote the orthogonal projection onto S by P_S. Then by orthogonality and conformal symmetry it follows [74, Lemma 3.1][2] that

$$\left\|P_{\dot{H}_0^1(\mathbf{D})}f + P_{\dot{H}_0^1((\partial\mathbf{D})^c)^\perp}f\right\|_{\dot{H}^1(\Omega)}^2 = \|f\|_{\dot{H}^1(\mathbf{D})}^2 + \|P_{\dot{H}_0^1((\partial\mathbf{D})^c)^\perp}f\|_{\dot{H}^1(\mathbf{D})}^2$$

$$= \|f\|_{\dot{H}^1(\mathbf{D})}^2 + 2\pi\|f\|_{\dot{H}^{1/2}(\partial\mathbf{D})}^2, \qquad (2.13)$$

where we canonically identify $f \in \dot{H}_0^1(\Omega)$ with its restriction to $\mathbf{D}$. If $\kappa_4 = 0$, then the rhs. of (2.13) is precisely $4\pi C(f, f)$ and therefore $L(f)$ can be interpreted [74, Corollary 1.2] as the projection

$$L = (4\pi)^{-1/2}Ph, \qquad P := \left(P_{\dot{H}_0^1(\mathbf{D})} + P_{\dot{H}_0^1((\partial\mathbf{D})^c)^\perp}\right) \qquad (2.14)$$

---

[1]The upper bound $1/500$ for a is a crude overestimate, we did not optimise it along the proof. The actual value of a comes from the fact that it has to be smaller than ω_d (see of Proposition 3.5) and from Lemma 7.9 (which is the main input of Proposition 3.5) it follows that $\omega_d \leq 1/100$.

[2]In Eq. (3.1), and in the last displayed equation of the proof of Lemma 3.1 factors of 2 are missing. In the notation of [74] the correct equations read

$$\frac{1}{2}\|P_H f\|_{\dot{H}^1(\mathbf{C})}^2 = \|P_H f\|_{\dot{H}^1(\mathbf{U})}^2 = 2\pi\|f\|_{\dot{H}^{1/2}(\partial\mathbf{U})}^2 \quad \text{and} \quad \langle g_1, g_2\rangle_{\dot{H}^1(\mathbf{U})} = 2\pi\langle g_1, g_2\rangle_{\dot{H}^{1/2}(\partial\mathbf{U})}.$$

of the Gaussian free field h onto $\dot{H}_0^1(\mathbf{D}) \oplus \dot{H}_0^1((\partial\mathbf{D})^c)^\perp$, i.e. the Gaussian free field conditioned to be harmonic in $\mathbf{D}^c$. The projection (2.14) is defined via duality, i.e. $(Ph)(f) := h(Pf)$ so that indeed

$$\mathbf{E}\left|\left[\frac{1}{\sqrt{4\pi}}Ph\right](f)\right|^2 = \frac{1}{4\pi}\left(\|f\|_{\dot{H}^1(\mathbf{D})}^2 + 2\pi\|f\|_{\dot{H}^{1/2}(\partial\mathbf{D})}^2\right) = C(f,f) = \mathbf{E}|L(f)|^2.$$

If $\kappa_4 > 0$, then L can be interpreted as the sum

$$L = \frac{1}{\sqrt{4\pi}}Ph + \sqrt{\kappa_4}\Big(\langle\cdot\rangle_{\mathbf{D}} - \langle\cdot\rangle_{\partial\mathbf{D}}\Big)\Xi \tag{2.15}$$

of the Gaussian free field Ph conditioned to be harmonic in $\mathbf{D}^c$, and an independent standard real Gaussian Ξ multiplied by difference of the averaging functionals $\langle\cdot\rangle_{\mathbf{D}}, \langle\cdot\rangle_{\partial\mathbf{D}}$ on $\mathbf{D}$ and $\partial\mathbf{D}$. For $\kappa_4 < 0$ there seems to be no direct interpretation of L similar to (2.15).

3. Proof strategy

For the proof of Theorem 2.2 we study the $2n \times 2n$ matrix H^z defined in (1.2), that is the Hermitisation of $X - z$. Denote by $\{\lambda_{\pm i}^z\}_{i=1}^n$ the eigenvalues of H^z labelled in an increasing order (we omit the index $i = 0$ for notational convenience). As a consequence of the block structure of H^z its spectrum is symmetric with respect to zero, i.e. $\lambda_{-i}^z = -\lambda_i^z$ for any $i \in [n]$.

Let $G(w) = G^z(w) := (H^z - w)^{-1}$ denote the resolvent of H^z with $\eta = \Im w \neq 0$. It is well known (e.g. see [4, 5]) that G^z becomes approximately deterministic, as $n \to \infty$, and its limit is expressed via the unique solution of the scalar equation

$$-\frac{1}{m^z} = w + m^z - \frac{|z|^2}{w + m^z}, \quad \eta\Im m^z(w) > 0, \quad \eta = \Im w \neq 0, \tag{3.1}$$

which is a special case of the *matrix Dyson equation* (MDE), see e.g. [2]. We note that on the imaginary axis $m^z(\mathrm{i}\eta) = \mathrm{i}\Im m^z(\mathrm{i}\eta)$. To find the limit of G^z we define a $2n \times 2n$ block-matrix

$$M^z(w) := \begin{pmatrix} m^z(w) & -zu^z(w) \\ -\overline{z}u^z(w) & m^z(w) \end{pmatrix}, \quad u^z(w) := \frac{m^z(w)}{w + m^z(w)}, \tag{3.2}$$

where each block is understood to be a scalar multiple of the $n \times n$ identity matrix. We note that m, u, M are uniformly bounded in z, w, i.e.

$$\|M^z(w)\| + |m^z(w)| + |u^z(w)| \lesssim 1. \tag{3.3}$$

Indeed, taking the imaginary part of (3.1) we have (dropping z, w)

$$\beta_*\Im m = (1 - \beta_*)\Im w, \quad \beta_* := 1 - |m|^2 - |u|^2|z|^2, \tag{3.4}$$

which implies

$$|m|^2 + |u|^2|z|^2 < 1, \tag{3.5}$$

as $\Im m$ and $\Im w$ have the same sign. Note that (3.5) saturates if $\Im w \to 0$ and $\Re w$ is in the support of the *self-consistent density of states*, $\rho^z(E) := \pi^{-1}\Im m^z(E + \mathrm{i}0)$. Moreover, (3.1) is equivalent to $u = -m^2 + u^2|z|^2$, thus $|u| < 1$ and (3.3) follows.

For our analysis the derivative $m'(w)$ in the w-variable plays a central role and we note that by taking the derivative of (3.1) we obtain

$$m' = \frac{1 - \beta}{\beta}, \quad \beta := 1 - m^2 - u^2|z|^2. \tag{3.6}$$

On the imaginary axis, $w = \mathrm{i}\eta$, where by taking the real part of (3.1) it follows that $\Re m(\mathrm{i}\eta) = 0$, we can use [5, Eq. (3.13)]

$$\Im m(\mathrm{i}\eta) \sim \begin{cases} \eta^{1/3} + |1 - |z|^2|^{1/2} & \text{if } |z| \leq 1, \\ \frac{\eta}{|z|^2 - 1 + \eta^{2/3}} & \text{if } |z| > 1, \end{cases} \quad \eta \lesssim 1 \tag{3.7}$$

to obtain asymptotics for

$$\beta_* \sim \frac{\eta}{\Im m}, \quad \beta = \beta_* + 2(\Im m)^2, \quad \eta \lesssim 1. \tag{3.8}$$

The optimal local law from Theorem [4, Theorem 5.2] and [5, Theorem 5.2][3], which for the application in Girko's formula (1.3) is only needed on the imaginary axis, asserts that $G^z \approx M^z$ in the following sense:

Theorem 3.1 (Optimal local law for G). *The resolvent G^z is very well approximated by the deterministic matrix M^z in the sense*

$$|\langle (G^z(\mathrm{i}\eta) - M^z(\mathrm{i}\eta))A\rangle| \leq \frac{\|A\|n^\xi}{n\eta}, \qquad |\langle \boldsymbol{x}, (G^z(\mathrm{i}\eta) - M^z(\mathrm{i}\eta))\boldsymbol{y}\rangle| \leq \frac{\|\boldsymbol{x}\|\|\boldsymbol{y}\|n^\xi}{\sqrt{n\eta}}, \qquad (3.9)$$

with very high probability, uniformly for $\eta > 0$ and for any deterministic matrices and vectors $A, \boldsymbol{x}, \boldsymbol{y}$.

The matrix H^z can be related to the linear statistics of eigenvalues σ_i of X via the precise (regularised) version of Girko's Hermitisation formula (1.3)

$$
\begin{aligned}
L_n(f) &= \frac{1}{4\pi}\int_{\mathbf{C}} \Delta f(z)\Big[\log|\det(H^z - \mathrm{i}T)| - \mathbf{E}\log|\det(H^z - \mathrm{i}T)|\Big]\,\mathrm{d}^2 z \\
&\quad - \frac{n}{2\pi\mathrm{i}}\int_{\mathbf{C}} \Delta f\left[\left(\int_0^{\eta_0} + \int_{\eta_0}^{\eta_c} + \int_{\eta_c}^{T}\right)\big[\langle G^z(\mathrm{i}\eta) - \mathbf{E}\,G^z(\mathrm{i}\eta)\rangle\big]\,\mathrm{d}\eta\right]\mathrm{d}^2 z \\
&=: J_T + I_0^{\eta_0} + I_{\eta_0}^{\eta_c} + I_{\eta_c}^{T},
\end{aligned}
\qquad (3.10)
$$

for

$$\eta_0 := n^{-1-\delta_0}, \quad \eta_c := n^{-1+\delta_1}, \qquad (3.11)$$

and some very large $T > 0$, say $T = n^{100}$. Note that in (3.10) we used that $\langle G^z(\mathrm{i}\eta)\rangle = \mathrm{i}\langle\Im G^z(\mathrm{i}\eta)\rangle$ by spectral symmetry. The test function $f\colon \mathbf{C} \to \mathbf{C}$ is in $H^{2+\delta}$ and it is compactly supported. J_T in (3.10) consists of the first line in the rhs., whilst $I_0^{\eta_0}, I_{\eta_0}^{\eta_c}, I_{\eta_c}^{T}$ corresponds to the three different η-regimes in the second line of the rhs. of (3.10).

Remark 3.2. *We remark that in (3.10) we split the η-regimes in a different way compared to [30, Eq. (32)]. We also use a different notation to identify the η-scales: here we use the notation $J_T, I_0^{\eta_0}, I_{\eta_0}^{\eta_c}, I_{\eta_c}^{T}$, whilst in [30, Eq. (32)] we used the notation I_1, I_2, I_3, I_4.*

The different regimes in (3.10) will be treated using different techniques. More precisely, the integral J_T is easily estimated as in [5, Proof of Theorem 2.3], which uses similar computations to [4, Proof of Theorem 2.5]. The term $I_0^{\eta_0}$ is estimated using the fact that with high probability there are no eigenvalues in the regime $[0, \eta_0]$; this follows by [83, Theorem 3.2]. Alternatively (see Remark 4.2 and Remark 4.5 later), the contribution of the regime $I_0^{\eta_0}$ can be estimated without resorting to the quite sophisticated proof of [83, Theorem 3.2] if the entries of X satisfy the additional assumption (4.3). More precisely, this can be achieved using [4, Proposition 5.7] (which follows adapting the proof of [16, Lemma 4.12]) to bound the very small regime $[0, n^{-l}]$, for some large $l \in \mathbf{N}$, and then using [32, Corollary 4] to bound the regime $[n^{-l}, \eta_0]$.

The main novel work is done for the integrals $I_{\eta_0}^{\eta_c}$ and $I_{\eta_c}^{T}$. The main contribution to $L_n(f)$ comes from the mesoscopic regime in $I_{\eta_c}^{T}$, which is analysed using the following Central Limit Theorem for resolvents.

Proposition 3.3 (CLT for resolvents). *Let $\epsilon, \xi > 0$ be arbitrary. Then for $z_1, \ldots, z_p \in \mathbf{C}$ and $\eta_1, \ldots, \eta_p \geq n^{\xi-1}\max_{i\neq j}|z_i - z_j|^{-2}$, denoting the pairings on $[p]$ by Π_p, we have*

$$
\begin{aligned}
\mathbf{E}\prod_{i\in[p]}\langle G_i - \mathbf{E}\,G_i\rangle &= \sum_{P\in\Pi_p}\prod_{\{i,j\}\in P}\mathbf{E}\langle G_i - \mathbf{E}\,G_i\rangle\langle G_j - \mathbf{E}\,G_j\rangle + \mathcal{O}\left(\Psi\right) \\
&= \frac{1}{n^p}\sum_{P\in\Pi_p}\prod_{\{i,j\}\in P}\frac{V_{i,j} + \kappa_4 U_i U_j}{2} + \mathcal{O}\left(\Psi\right),
\end{aligned}
\qquad (3.12)
$$

---

[3]The local laws in [4, Theorem 5.2] and [5, Theorem 5.2] have been proven for $\eta \geq \eta_f(z)$, with $\eta_f(z)$ being the fluctuation scale defined in [5, Eq. (5.2)], but they can be easily extend to any $\eta > 0$ by a standard argument, see [30, Appendix A].

where $G_i = G^{z_i}(i\eta_i)$,

$$\Psi := \frac{n^\epsilon}{(n\eta_*)^{1/2}} \frac{1}{\min_{i\neq j}|z_i - z_j|^4} \prod_{i\in[p]} \frac{1}{|1 - |z_i||n\eta_i|}, \tag{3.13}$$

$\eta_* := \min_i \eta_i$, *and* $V_{i,j} = V_{i,j}(z_i, z_j, \eta_i, \eta_j)$ *and* $U_i = U_i(z_i, \eta_i)$ *are defined as*

$$V_{i,j} := \frac{1}{2}\partial_{\eta_i}\partial_{\eta_j} \log\big[1 + (u_i u_j |z_i||z_j|)^2 - m_i^2 m_j^2 - 2u_i u_j \Re z_i \overline{z_j}\big],$$

$$U_i := \frac{i}{\sqrt{2}}\partial_{\eta_i} m_i^2, \tag{3.14}$$

with $m_i = m^{z_i}(i\eta_i)$ *and* $u_i = u^{z_i}(i\eta_i)$.

Moreover, the expectation of G is given by

$$\langle \mathbf{E}\, G\rangle = \langle M\rangle - \frac{i\kappa_4}{4n}\partial_\eta(m^4) + \mathcal{O}\Big(\frac{1}{|1 - |z||n^{3/2}(1+\eta)} + \frac{1}{|1 - |z||(n\eta)^2}\Big). \tag{3.15}$$

Remark 3.4. *In Section 4 we will apply this proposition in the regime where $\min_{i\neq j}|z_i - z_j|$ is quite large, i.e. it is at least $n^{-\delta}$, for some small $\delta > 0$, hence we did not optimise the estimates for the opposite regime. However, using the more precise [31, Lemma 6.1] instead of Lemma 6.1 within the proof, one can immediately strengthen Proposition 3.3 on two accounts. First, the condition on $\eta_* = \min \eta_i$ can be relaxed to*

$$\eta_* \gtrsim n^{\xi-1}\Big(\min_{i\neq j}|z_i - z_j|^2 + \eta_*\Big)^{-1}.$$

Second, the denominator $\min_{i\neq j}|z_i - z_j|^4$ in (3.13) can be improved to

$$\Big(\min_{i\neq j}|z_i - z_j|^2 + \eta_*\Big)^2.$$

In order to show that the contribution of $I^{\eta_c}_{\eta_0}$ to $L_n(f)$ is negligible, in Proposition 3.5 we prove that $\langle G^{z_1}(i\eta_1)\rangle$ and $\langle G^{z_2}(i\eta_2)\rangle$ are asymptotically independent if z_1, z_2 are far enough from each other, they are well inside $\mathbf{D}$, and $\eta_0 \leq \eta_1, \eta_2 \leq \eta_c$.

Proposition 3.5 (Independence of resolvents with small imaginary part)**.** *Fix $p \in \mathbf{N}$. For any sufficiently small $\omega_d, \omega_h, \omega_f > 0$ such that $\omega_h \ll \omega_f$, there exist $\omega, \widehat{\omega}, \delta_0, \delta_1 > 0$ such that $\omega_h \ll \delta_m \ll \widehat{\omega} \ll \omega \ll \omega_f$, for $m = 0, 1$, such that for any $|z_l| \leq 1 - n^{-\omega_h}$, $|z_l - z_m| \geq n^{-\omega_d}$, with $l, m \in [p]$, $l \neq m$, it holds*

$$\mathbf{E}\prod_{l=1}^p \langle G^{z_l}(i\eta_l)\rangle = \prod_{l=1}^p \mathbf{E}\langle G^{z_l}(i\eta_l)\rangle + \mathcal{O}\left(\frac{n^{p(\omega_h+\delta_0)+\delta_1}}{n^\omega} + \frac{n^{\omega_f+3\delta_0}}{\sqrt{n}}\right), \tag{3.16}$$

for any $\eta_1, \ldots, \eta_p \in [n^{-1-\delta_0}, n^{-1+\delta_1}]$.

The paper is organised as follows: In Section 4 we conclude Theorem 2.2 by combining Propositions 3.3 and 3.5. In Section 5 we prove a local law for $G_1 A G_2$, for a deterministic matrix A. In Section 6, using the result in Section 5 as an input, we prove Proposition 3.3, the Central Limit Theorem for resolvents. In Section 7 we prove Proposition 3.5 using the fact that the correlation among small eigenvalues of H^{z_1}, H^{z_2} is "small", if z_1, z_2 are far from each other, as a consequence of the local law in Section 5.

4. Central limit theorem for linear statistics

In this section, using Proposition 3.3–3.5 as inputs, we prove our main result Theorem 2.2.

4.1. **Preliminary reductions in Girko's formula.** In this section we prove that the main contribution to $L_n(f)$ in (3.10) comes from the regime $I_{\eta_c}^T$. This is made rigorous in the following lemma.

Lemma 4.1. *Fix $p \in \mathbf{N}$ and some bounded open $\overline{\mathbf{D}} \subset \Omega \subset \mathbf{C}$, and for any $l \in [p]$ let $f^{(l)} \in H_0^{2+\delta}(\Omega)$. Then*

$$\mathbf{E} \prod_{l=1}^{p} L_n\big(f^{(l)}\big) = \mathbf{E} \prod_{l=1}^{p} I_{\eta_c}^T\big(f^{(l)}\big) + \mathcal{O}\left(n^{-c(p)}\right), \tag{4.1}$$

for some small $c(p) > 0$, with $L_n(f^{(l)})$ and $I_{\eta_c}^T(f^{(l)})$ defined in (3.10). The constant in $\mathcal{O}(\cdot)$ may depend on p and on the L^2-norm of $\Delta f^{(1)}, \ldots, \Delta f^{(p)}$.

Remark 4.2. *In the remainder of this section we need to ensure that with high probability the matrix H^z, defined in (1.2), does not have eigenvalues very close to zero, i.e. that*

$$\mathbf{P}\left(\mathrm{Spec}(H^z) \cap \left[-n^{-l}, n^{-l}\right] \neq \emptyset\right) \leq C_l n^{-l/2}, \tag{4.2}$$

for any $l \geq 2$ uniformly in $|z| \leq 1$. The bound (4.2) directly follows from [83, Theorem 3.2]. Alternatively, (4.2) follows by [4, Proposition 5.7] (which follows adapting the proof of [16, Lemma 4.12]), without recurring to the quite sophisticated proof of [83, Theorem 3.2], under the additional assumption that there exist $\alpha, \beta > 0$ such that the random variable χ has a density $g \colon \mathbf{C} \to [0, \infty)$ which satisfies

$$g \in L^{1+\alpha}(\mathbf{C}), \qquad \|g\|_{L^{1+\alpha}(\mathbf{C})} \leq n^{\beta}. \tag{4.3}$$

We start proving *a priori* bounds for the integrals defined in (3.10).

Lemma 4.3. *Fix some bounded open $\overline{\mathbf{D}} \subset \Omega \subset \mathbf{C}$ and let $f \in H_0^{2+\delta}(\Omega)$. Then for any $\xi > 0$ the bounds*

$$|J_T| \leq \frac{n^{1+\xi}\|\Delta f\|_{L^1(\Omega)}}{T^2}, \qquad \big|I_0^{\eta_0}\big| + \big|I_{\eta_0}^{\eta_c}\big| + |I_{\eta_c}^T| \leq n^{\xi}\|\Delta f\|_{L^2(\Omega)}|\Omega|^{1/2}, \tag{4.4}$$

hold with very high probability, where $|\Omega|$ denotes the Lebesgue measure of the set Ω.

Proof. The proof of the bound for J_T is identical to [5, Proof of Theorem 2.3] and so omitted.

The bound for $I_0^{\eta_0}, I_{\eta_0}^{\eta_c}, I_{\eta_c}^T$ relies on the local law of Theorem 3.1. More precisely, by Theorem 3.1 and (3.15) of Proposition 3.3 it follows that

$$|\langle G^z - \mathbf{E}\, G^z \rangle| \leq \frac{n^{\xi}}{n\eta}, \tag{4.5}$$

with very high probability uniformly in $\eta > 0$ and $|z| \leq C$ for some large $C > 0$. First of all we remove the regime $[0, n^{-l}]$ by [83, Theorem 3.2], i.e. its contribution is smaller than n^{-l}, for some large $l \in \mathbf{N}$, with very high probability. Alternatively, this can be achieved by [4, Proposition 5.7] under the additional assumption (4.3) in Remark 4.2. Then for any $a, b \geq n^{-l}$, by (4.5), we have

$$n\left|\int_{\Omega} \mathrm{d}^2 z \Delta f(z) \int_a^b \mathrm{d}\eta \big[\langle G(i\eta) - \mathbf{E}\, G(i\eta) \rangle\big]\right| \lesssim n^{\xi}|\Omega|^{1/2}\|\Delta f\|_{L^2(\Omega)}, \tag{4.6}$$

with very high probability. This concludes the proof of the second bound in (4.4). □

We have a better bound for $I_0^{\eta_0}, I_{\eta_0}^{\eta_c}$ which holds true in expectation.

Lemma 4.4. *Fix some bounded open $\overline{\mathbf{D}} \subset \Omega \subset \mathbf{C}$ and let $f \in H_0^{2+\delta}(\Omega)$. Then there exists $\delta' > 0$ such that*

$$\mathbf{E}|I_0^{\eta_0}| + \mathbf{E}\big|I_{\eta_0}^{\eta_c}\big| \leq n^{-\delta'}\|\Delta f\|_{L^2(\Omega)}. \tag{4.7}$$

Proof of Lemma 4.1. Lemma 4.1 readily follows (see e.g. [30, Lemma 4.2]) combining Lemma 4.3 and Lemma 4.4. □

We conclude this section with the proof of Lemma 4.4.

Proof of Lemma 4.4. The bound for $\mathbf{E}|I_0^{\eta_0}|$ immediately follows by [83, Theorem 3.2] (see also Remark 4.5 for an alternative proof).

By the local law outside the spectrum, given in the second part of [5, Theorem 5.2], it follows that for $0 < \gamma < 1/2$ we have

$$|\langle G^z(i\eta) - M^z(i\eta)\rangle| \leq \frac{n^\xi}{n^{1+\gamma/3}\eta}, \tag{4.8}$$

uniformly for all $|z|^2 \geq 1 + (n^\gamma\eta)^{2/3} + n^{(\gamma-1)/2}, \eta > 0$, and $|z| \leq 1 + \tau^*$, for some $\tau^* \sim 1$. We remark that the local law (4.8) was initially proven only for η above the fluctuation scale $\eta_f(z)$, which is defined in [5, Eq. (5.2)], but it can be easily extend to any $\eta > 0$ using the monotonicity of the function $\eta \mapsto \eta\langle\Im G(i\eta)\rangle$ and the fact that

$$\left|n^\xi\eta_f(z)\langle M^z(in^\xi\eta_f(z))\rangle\right| + |\eta\langle M^z(i\eta)\rangle| \lesssim n^{2\xi}\frac{\eta_f(z)^2}{|z|^2-1}, \tag{4.9}$$

uniformly in $\eta > 0$, since $\Im M^z(i\eta) = \Im m^z(i\eta)I$ by (3.2), with I the $2n \times 2n$ identity matrix, and $\Im m^z(i\eta) \leq \eta(|z|^2 - 1)^{-1}$ by [5, Eq. (3.13)]. Note that we assumed the additional term $n^{(\gamma-1)/2}$ in the lower bound for $|z|^2$ compared with [5, Theorem 5.2] in order to ensure that the rhs. in (4.9), divided by η, is smaller than the error term in (4.8).

Next, in order to bound $\mathbf{E}|I_{\eta_0}^{\eta_c}|$, we consider

$$\mathbf{E}|I_{\eta_0}^{\eta_c}|^2 = -\frac{n^2}{4\pi^2}\int_{\mathbf{C}}\mathrm{d}^2 z_1(\Delta f)(z_1)\int_{\mathbf{C}}\mathrm{d}^2 z_2(\Delta\overline{f})(z_2)\int_{\eta_0}^{\eta_c}\mathrm{d}\eta_1\int_{\eta_0}^{\eta_c}\mathrm{d}\eta_2\, F \tag{4.10}$$

$$F = F(z_1, z_2, \eta_1, \eta_2) := \mathbf{E}\left[\langle G^{z_1}(i\eta_1) - \mathbf{E}\, G^{z_1}(i\eta_1)\rangle\langle G^{z_2}(i\eta_2) - \mathbf{E}\, G^{z_2}(i\eta_2)\rangle\right]. \tag{4.11}$$

By (4.6) it follows that the regimes $1 - n^{-2\omega_h} \leq |z_l|^2 \leq 1 + n^{-2\omega_h}$, with $l = 1, 2$, and $|z_1 - z_2| \leq n^{-\omega_d}$ in (4.10), with ω_h, ω_d defined in Proposition 3.5, are bounded by $n^{-2\omega_h+\xi}$ and $n^{-\omega_d/2+\xi}$, respectively. Moreover, the contribution from the regime $|z_l| \geq 1 + n^{-2\omega_h}$ is also bounded by $n^{-2\omega_h+\xi}$ using (4.8) with $\gamma \leq 1 - 3\omega_h - 2\delta_1$, say $\gamma = 1/4$. After collecting these error terms we conclude that

$$\mathbf{E}|I_{\eta_0}^{\eta_c}|^2 = \frac{n^2}{4\pi^2}\int_{|z_1|\leq 1-n^{-\omega_h}}\mathrm{d}^2 z_1\Delta f(z_1)\int_{\substack{|z_2|\leq 1-n^{-\omega_h}, \\ |z_2-z_1|\geq n^{-\omega_d}}}\mathrm{d}^2 z_2\Delta\overline{f(z_2)}$$

$$\times \int_{\eta_0}^{\eta_c}\mathrm{d}\eta_1\int_{\eta_0}^{\eta_c}\mathrm{d}\eta_2 F + \mathcal{O}\left(\frac{n^\xi}{n^{\omega_h}} + \frac{n^\xi}{n^{\omega_d/2}}\right). \tag{4.12}$$

We remark that the implicit constant in $\mathcal{O}(\cdot)$ in (4.12) and in the remainder of the proof may depend on $\|\Delta f\|_{L^2(\Omega)}$.

Then by Proposition 3.5 it follows that

$$\mathbf{E}\left[\langle G^{z_1}(i\eta_1) - \mathbf{E}\langle G^{z_1}(i\eta_1)\rangle\langle G^{z_2}(i\eta_2) - \mathbf{E}\, G^{z_2}(i\eta_2)\rangle\right] = \mathcal{O}\left(\frac{n^{c(\omega_h+\delta_0)+\delta_1}}{n^\omega}\right), \tag{4.13}$$

with $\omega_h \ll \delta_0 \ll \omega$. Hence, plugging (4.13) into (4.12) it follows that

$$\mathbf{E}|I_{\eta_0}^{\eta_c}|^2 = \mathcal{O}\left(\frac{n^{c(\omega_h+\delta_0)+2\delta_1}}{n^\omega}\right). \tag{4.14}$$

This concludes the proof under the assumption $\omega_h \ll \delta_m \ll \omega$, with $m = 0, 1$, of Proposition 3.5 (see Section 7.2.3 later for a summary on all the scales involved in the proof of Proposition 3.5). □

Remark 4.5 (Alternative proof of the bound for $\mathbf{E}|I_0^{\eta_0}|$). *Under the additional assumption (4.3) in Remark 4.2, we can prove the same bound for $\mathbf{E}|I_0^{\eta_0}|$ in (4.7) without relying on the fairly sophisticated proof of [83, Theorem 3.2].*

In order to bound $\mathbf{E}|I_0^{\eta_0}|$ we first remove the regime $\eta \in [0, n^{-l}]$ as in the proof of Lemma 4.3. Then, using (4.6) to bound the integral over the regime $|1 - |z|^2| \leq 1 + n^{-2\omega_h}$, with ω_h defined in Proposition 3.5, and (4.8) for the regime $|z|^2 \geq 1 - n^{-2\omega_h}$, we conclude that

$$\mathbf{E}|I_0^{\eta_0}| = \mathbf{E}\frac{n}{2\pi}\int_{|z|\leq 1-n^{-2\omega_h}}|\Delta f|\left|\int_0^{\eta_0}\langle G^z - \mathbf{E}\, G^z\rangle\,\mathrm{d}\eta\right|\mathrm{d}^2 z + \mathcal{O}\left(\frac{n^\xi}{n^{\omega_h}}\right). \tag{4.15}$$

By universality of the smallest eigenvalue of H^z (which directly follows by Proposition 7.14 for any fixed $|z|^2 \leq 1 - n^{-2\omega_h}$; see also [26]), and the bound in [32, Corollary 2.4] we have that

$$\mathbf{P}\left(\lambda_1^z \leq \eta_0\right) \leq n^{-\delta_0/4},$$

with $\eta_0 = n^{-1-\delta_0}$ and $\omega_h \ll \delta_0$. This concludes the bound in (4.7) for $I_0^{\eta_0}$ following exactly the same proof of [30, Lemma 4.6], by (4.15). We warn the reader that in [32, Corollary 2.4] λ_1 denotes the smallest eigenvalue of $(X-z)(X-z)^*$, whilst here λ_1^z denotes the smallest (positive) eigenvalue of H^z.

4.2. **Computation of the expectation in Theorem 2.2.** In this section we compute the expectation $\mathbf{E}\sum_i f(\sigma_i)$ in (2.8) using the computation of $\mathbf{E}\langle G\rangle$ in (3.15) of Proposition 3.3 as an input. More precisely, we prove the following lemma. Note that (4.16) proves (2.8) in Theorem 2.2.

Lemma 4.6. *Fix some bounded open $\overline{\mathbf{D}} \subset \Omega \subset \mathbf{C}$ and let $f \in H_0^{2+\delta}(\Omega)$, and let $\kappa_4 := n^2[\mathbf{E}|x_{11}|^4 - 2(\mathbf{E}|x_{11}|^2)]$, then*

$$\mathbf{E}\sum_{i=1}^n f(\sigma_i) = \frac{n}{\pi}\int_{\mathbf{D}} f(z)\,\mathrm{d}^2 z + \frac{1}{8\pi}\int_{\mathbf{D}}\Delta f(z)\,\mathrm{d}^2 z - \frac{\kappa_4}{\pi}\int_{\mathbf{D}} f(z)(2|z|^2 - 1)\,\mathrm{d}^2 z + \mathcal{O}\left(n^{-\delta'}\right),$$
(4.16)

for some small $\delta' > 0$.

Proof. By estimating the regimes $\eta < \eta_0$ and $\eta > T$ using [83, Thm. 3.2] and [5, Proof of Thm. 2.3], respectively, we have that

$$\mathbf{E}\sum_i f(\sigma_i) = -\frac{n}{2\pi\mathrm{i}}\int_{\mathbf{C}}\Delta f(z)\left(\int_{\eta_0}^{\eta_c} + \int_{\eta_c}^{T}\right)\mathbf{E}\langle G^z(\mathrm{i}\eta)\rangle\,\mathrm{d}\eta\,\mathrm{d}^2 z + \mathcal{O}(n^{-c}) \qquad (4.17)$$

for some small $c > 0$.

We now consider the regime $\eta \in [\eta_c, T]$. Since the error term in (3.15) is not affordable for $1 - |z|$ very close to zero, we remove the regime $|1 - |z|^2| \leq n^{-2\nu}$ in the z-integral by the optimal local law at the expense of an error term $n^{-\nu+\xi}$, for some very small $\nu > 0$ we will choose shortly. The regime $|1 - |z|^2| \geq n^{-2\nu}$, instead, is computed using (3.15). Hence, collecting these error terms we conclude that there exists $\delta' > 0$ such that

$$-\frac{n}{2\pi\mathrm{i}}\int \mathrm{d}^2 z\Delta f\int_{\eta_c}^{T}\mathrm{d}\eta\,\mathbf{E}\langle G\rangle \qquad (4.18)$$

$$= -\frac{n}{2\pi\mathrm{i}}\int \mathrm{d}^2 z\Delta f\int_{\eta_c}^{T}\mathrm{d}\eta\left(\langle M\rangle - \frac{\mathrm{i}\kappa_4}{4n}\partial_\eta(m^4)\right) + \mathcal{O}\left(n^{-\delta'} + n^{-\nu+\xi} + \frac{n^{2\nu}}{n\eta_c}\right)$$

$$= \frac{n}{\pi}\int_{\mathbf{D}} f(z)\,\mathrm{d}^2 z - \frac{\kappa_4}{\pi}\int_{\mathbf{D}} f(z)(2|z|^2 - 1)\,\mathrm{d}^2 z + \mathcal{O}\left(n^{-\delta'} + \frac{n^{2\nu}}{n\eta_c} + n^{2\nu}\eta_c + n^{-\nu+\xi}\right),$$

with $\eta_c = n^{-1+\delta_1}$ defined in (3.11). To add back the regimes $\eta \in [0, \eta_c]$, $\eta \geq T$, and the regime $|1 - |z|^2| \leq n^{-2\nu}$ we used that $|\partial_\eta(m^4)| \lesssim n^{2\nu}$ and that using $|m| \leq \eta^{-1}$ we have $|\partial_\eta(m^4)| \lesssim \eta^{-5}$ by (3.6). Choosing $\nu, \delta' > 0$ so that $\nu \ll \delta_1 \ll \delta'$ we conclude

$$\mathbf{E}\sum_i f(\sigma_i) = \frac{n}{\pi}\int_{\mathbf{D}} f(z)\,\mathrm{d}^2 z - \frac{\kappa_4}{\pi}\int_{\mathbf{D}} f(z)(2|z|^2 - 1)\,\mathrm{d}^2 z$$
$$- \frac{n}{2\pi\mathrm{i}}\int_{\mathbf{C}}\Delta f(z)\int_{\eta_0}^{\eta_c}\mathbf{E}\langle G^z(\mathrm{i}\eta)\rangle\,\mathrm{d}\eta\,\mathrm{d}^2 z + \mathcal{O}\left(n^{-c}\right) \qquad (4.19)$$

from (4.17)–(4.18).

Finally, we consider the regime $\eta \in [\eta_0, \eta_c]$. First we note that in the Ginibre case we have the expansion (see [84, Eq. (5.18)] and [62, Eq. (1.14)])

$$\mathbf{E}\sum_i f(\widetilde{\sigma}_i) = \frac{n}{\pi}\int_{\mathbf{D}} f(z)\,\mathrm{d}^2 z + \frac{1}{8\pi}\int_{\mathbf{D}}\Delta f(z)\,\mathrm{d}^2 z,$$

which, using (4.19) with $\kappa_4 = 0$, implies

$$-\frac{n}{2\pi i} \int_{\mathbf{C}} \Delta f(z) \int_{\eta_0}^{\eta_c} \mathbf{E}\langle \widetilde{G}^z(i\eta)\rangle \, d\eta \, d^2 z = \frac{1}{8\pi} \int_{\mathbf{D}} \Delta f(z) \, d^2 z + \mathcal{O}\left(n^{-c}\right), \qquad (4.20)$$

where $\widetilde{G}$ denote the resolvent of the Hermitization of a complex Ginibre matrix. In order to compare $\mathbf{E}\langle G\rangle$ and $\mathbf{E}\langle \widetilde{G}\rangle$ we use that for $|1 - |z|^2| > n^{-2\nu}$ we have

$$\int_{\eta_0}^{\eta_c} d\eta \, \mathbf{E}\langle G^z\rangle = \int_{\eta_0}^{\eta_c} d\eta \, \mathbf{E}\langle \widetilde{G}^z\rangle + \mathcal{O}\left(n^{-1-c}\right), \qquad (4.21)$$

for some small $c > 0$, the complementary regime is negligible by its small volume. The relation (4.21) follows using Lemma 7.6 and computations analogous (actually easier) to (7.40). By combining (4.19), (4.20) and (4.21) the proof of the lemma is concluded. $\qquad \square$

4.3. Computation of the second and higher moments in Theorem 2.2.

In this section we conclude the proof of Theorem 2.2, i.e. we compute

$$\mathbf{E} \prod_{i \in [p]} L_n(f^{(i)}) = \mathbf{E} \prod_{i \in [p]} I_{\eta_c}^T(f^{(i)}) + \mathcal{O}(n^{-c(p)})$$

$$= \mathbf{E} \prod_{i \in [p]} \left[-\frac{n}{2\pi i} \int_{\mathbf{C}} \Delta f^{(i)}(z) \int_{\eta_c}^T \langle G^z(i\eta) - \mathbf{E}\, G^z(i\eta)\rangle \, d\eta \, d^2 z \right] \qquad (4.22)$$

$$+ \mathcal{O}(n^{-c(p)})$$

to leading order using (3.12).

Lemma 4.7. *Let $f^{(i)}$ be as in Theorem 2.2 and set $f^{(i)} = f$ or $f^{(i)} = \overline{f}$ for any $i \in [p]$, and recall that Π_p denotes the set of pairings on $[p]$. Then*

$$\mathbf{E} \prod_{i \in [p]} \left[-\frac{n}{2\pi i} \int_{\mathbf{C}} \Delta f^{(i)}(z) \int_{\eta_c}^T \langle G^z(i\eta) - \mathbf{E}\, G^z(i\eta)\rangle \, d\eta \, d^2 z \right]$$

$$= \sum_{P \in \Pi_p} \prod_{\{i,j\} \in P} \left[-\int_{\mathbf{C}} d^2 z_i \Delta f^{(i)} \int_{\mathbf{C}} d^2 z_j \Delta f^{(j)} \int_0^\infty d\eta_i \int_0^\infty d\eta_j \frac{V_{i,j} + \kappa_4 U_i U_j}{8\pi^2} \right] \qquad (4.23)$$

$$+ \mathcal{O}(n^{-c(p)}),$$

for some small $c(p) > 0$, where $V_{i,j}$ and U_i are as in (3.14). The implicit constant in $\mathcal{O}(\cdot)$ may depend on p.

Proof. In order to prove the lemma we have to check that the integral of the error term in (3.12) is at most of size $n^{-c(p)}$, and that the integral of $V_{i,j} + \kappa_4 U_i U_j$ for $\eta_i \leq \eta_c$ or $\eta_i \geq T$ is similarly negligible. In the remainder of the proof we assume that p is even, since the terms with p odd are of lower order by (3.12).

Note that by the explicit form of m_i, u_i in (3.1)–(3.2), by the definition of $V_{i,j}, U_i, U_j$ in (3.14), the fact that by $-m_i^2 + |z_i|^2 u_i^2 = u_i$ we have

$$V_{i,j} = \frac{1}{2} \partial_{\eta_i} \partial_{\eta_j} \log\left(1 - u_i u_j \left[1 - |z_i - z_j|^2 + (1 - u_i)|z_i|^2 + (1 - u_j)|z_j|^2\right]\right),$$

and using $|\partial_{\eta_i} m_i| \leq [\Im m^{z_i}(i\eta_i) + \eta_i]^{-2}$ by (3.6)–(3.8), we conclude (see also (6.6)–(6.7) later)

$$|V_{i,j}| \lesssim \frac{[(\Im m^{z_i}(i\eta_i) + \eta_i)(\Im m^{z_j}(i\eta_j) + \eta_j)]^{-2}}{[|z_i - z_j|^2 + (\eta_i + \eta_j)(\min\{\Im m^{z_i}, \Im m^{z_j}\}^2)]^2}, \quad |U_i| \lesssim \frac{1}{\Im m^{z_i}(i\eta_i)^2 + \eta_i^3}. \qquad (4.24)$$

Using the bound (4.6) to remove the regime $Z_i := \{|1 - |z_i|^2| \leq n^{-2\nu}\}$ for any $i \in [p]$, for some small $\nu > 0$, we conclude that the lhs. of (4.23) is equal to

$$\frac{(-n)^p}{(2\pi i)^p} \prod_{i \in [p]} \int_{Z_i^c} d^2 z_i \Delta f^{(i)}(z_i) \, \mathbf{E} \prod_{i \in [p]} \int_{\eta_c}^T \langle G^{z_i}(i\eta_i) - \mathbf{E}\, G^{z_i}(i\eta_i)\rangle \, d\eta_i + \mathcal{O}\left(\frac{n^{p\xi}}{n^\nu}\right), \qquad (4.25)$$

for any very small $\xi > 0$. Additionally, since the error term Ψ defined in (3.13) behaves badly for small $|z_i - z_j|$, we remove the regime

$$\widehat{Z}_i := \bigcup_{j<i} \{z_j : |z_i - z_j| \le n^{-2\nu}\}$$

in each z_i-integral in (4.25) using (4.6), and, denoting $f^{(i)} = f^{(i)}(z_i)$, get

$$\frac{(-n)^p}{(2\pi\mathrm{i})^p} \prod_{i\in[p]} \int_{Z_i^c \cap \widehat{Z}_i^c} \mathrm{d}^2 z_i \Delta f^{(i)} \, \mathbf{E} \prod_{i\in[p]} \int_{\eta_c}^T \langle G^{z_i}(\mathrm{i}\eta_i) - \mathbf{E}\, G^{z_i}(\mathrm{i}\eta_i) \rangle \, \mathrm{d}\eta_i + \mathcal{O}\left(\frac{n^{p\xi}}{n^\nu}\right). \quad (4.26)$$

Plugging (3.12) into (4.26), and using the first bound in (4.4) to remove the regime $\eta_i \ge T$ for the lhs. of (4.23) we get

$$\frac{1}{(2\pi\mathrm{i})^p} \prod_{i\in[p]} \int_{Z_i^c \cap \widehat{Z}_i^c} \mathrm{d}^2 z_i \Delta f^{(i)} \sum_{P\in\Pi_p} \prod_{\{i,j\}\in P} \int_0^\infty \int_0^\infty -\frac{V_{i,j} + \kappa_4 U_i U_j}{8\pi^2} \, \mathrm{d}\eta_j \, \mathrm{d}\eta_i$$
$$+ \mathcal{O}\left(\frac{n^{p\xi}}{n^\nu} + \frac{n^{20\nu p + \delta_1}}{n} + \frac{n^{\xi p + 2p\nu}}{n^{\delta_1/2}}\right), \quad (4.27)$$

where $\eta_c = n^{-1+\delta_1}$, the second last error term comes from adding back the regimes $\eta_i \in [0, \eta_c]$ using that

$$|V_{i,j}| \le \frac{n^{20\nu}}{(1+\eta_i^2)(1+\eta_j^2)}, \qquad |U_i| \le \frac{n^{4\nu}}{1+\eta_i^3},$$

for $z_i \in Z_i^c \cap \widehat{Z}_i^c$ and $z_j \in Z_j^c \cap \widehat{Z}_j^c$ by (4.24). The last error term in (4.27) comes from the integral of Ψ, with Ψ defined in (3.13). Finally, we perform the η-integrations using the explicit formulas (4.28) and (4.29) below. After that, we add back the domains Z_i and $\widehat{Z}_i$ for $i \in [p]$ at a negligible error, since these domains have volume of order $n^{-2\nu}$, $\Delta f^{(i)} \in L^2$, and the logarithmic singularities from (4.28) are integrable. This concludes (4.23) choosing ν so that $\nu \ll \delta_1 \ll 1$. $\qquad\square$

In the next three sub-sections we compute the integrals in (4.23) for any i, j's. To make our notation simpler we use only the indices $1, 2$, i.e. we compute the integral of $V_{1,2}$ and $U_1 U_2$.

4.3.1. *Computation of the (η_1, η_2)-integrals.* Using the relations in (3.14) we explicitly compute the (η_1, η_2)-integral of $V_{1,2}$:

$$-\int_0^\infty \int_0^\infty V_{1,2} \, \mathrm{d}\eta_1 \, \mathrm{d}\eta_2 = -\frac{1}{2} \log A \upharpoonright_{\substack{\eta_1=0,\\ \eta_2=0}}$$

$$= \Theta(z_1, z_2) := \frac{1}{2} \begin{cases} -\log|z_1 - z_2|^2, & |z_1|, |z_2| \le 1, \\ \log|z_l|^2 - \log|z_1 - z_2|^2, & |z_m| \le 1, |z_l| > 1, \\ \log|z_1 z_2|^2 - \log|1 - z_1 \overline{z}_2|^2, & |z_1|, |z_2| > 1, \end{cases} \quad (4.28)$$

with $A(\eta_1, \eta_2, z_1, z_2)$ defined by

$$A(\eta_1, \eta_2, z_1, z_2) := 1 + (u_1 u_2 |z_1||z_2|)^2 - m_1^2 m_2^2 - 2u_1 u_2 \Re z_1 \overline{z}_2.$$

Then the η_i-integral of U_i, for $i \in \{1, 2\}$, is given by

$$\int_0^\infty U_i \, \mathrm{d}\eta_i = \frac{\mathrm{i}}{\sqrt{2}}(1 - |z_i|^2). \quad (4.29)$$

Before proceeding we rewrite $\Theta(z_1, z_2)$ as

$$2\Theta(z_1, z_2) = -\log|z_1 - z_2|^2 + \log|z_1|^2 \mathbf{1}(|z_1| > 1) + \log|z_2|^2 \mathbf{1}(|z_2| > 1)$$
$$+ \left[\log|z_1 - z_2|^2 - \log|1 - z_1 \overline{z}_2|^2\right] \mathbf{1}(|z_1|, |z_2| > 1).$$

In the remainder of this section we use the notations

$$\mathrm{d}z := \mathrm{d}z + \mathrm{i}\,\mathrm{d}y, \quad \mathrm{d}\overline{z} := \mathrm{d}x - \mathrm{i}\,\mathrm{d}y, \qquad \partial_z := \frac{\partial_x - \mathrm{i}\partial_y}{2}, \quad \partial_{\overline{z}} := \frac{\partial_x + \mathrm{i}\partial_y}{2},$$

and $\partial_l := \partial_{z_l}$, $\overline{\partial}_l := \partial_{\overline{z}_l}$. With this notation $\Delta_{z_l} = 4\partial_{z_l}\partial_{\overline{z}_l}$.

We split the computation of the leading term in the rhs. of (4.23) into two parts: the integral of $V_{1,2}$, and the integral of $U_1 U_2$.

4.3.2. *Computation of the (z_1, z_2)-integral of $V_{1,2}$.* In this section we compute the integral of $V_{1,2}$ in (4.23). To make our notation easier in the remainder of this section we use the notation f and g, instead of $f^{(1)}$, $f^{(2)}$, with f in Theorem 2.2 and $g = f$ or $g = \overline{f}$.

Lemma 4.8. *Let $V_{1,2}$ be defined in (3.14), then*

$$
-\frac{1}{8\pi^2}\int_{\mathbf{C}}\mathrm{d}^2 z_1\int_{\mathbf{C}}\mathrm{d}^2 z_2 \Delta f(z_1)\Delta\overline{g(z_2)}\int_0^\infty \mathrm{d}\eta_1\int_0^\infty \mathrm{d}\eta_2 V_{1,2}
$$
$$
= \frac{1}{4\pi}\int_{\mathbf{D}}\langle\nabla g, \nabla f\rangle\,\mathrm{d}^2 z + \frac{1}{2}\sum_{m\in\mathbf{Z}}|m|\widehat{f\restriction_{\partial\mathbf{D}}}(m)\overline{\widehat{g\restriction_{\partial\mathbf{D}}}(m)}.
$$

(4.30)

Note that the rhs. of (4.30) gives exactly the first two terms in (2.6).

Using the expression of $V_{1,2}$ in (3.14) and the computation of its (η_1, η_2)-integral in (4.28), we have that

$$
-\frac{1}{8\pi^2}\int_{\mathbf{C}}\mathrm{d}^2 z_1\int_{\mathbf{C}}\mathrm{d}^2 z_2 \Delta f(z_1)\Delta\overline{g(z_2)}\int_0^\infty \mathrm{d}\eta_1\int_0^\infty \mathrm{d}\eta_2 V_{1,2}
$$
$$
= \frac{2}{\pi^2}\int_{\mathbf{C}}\mathrm{d}^2 z_1\int_{\mathbf{C}}\mathrm{d}^2 z_2 \partial_1\overline{\partial}_1 f(z_1)\partial_2\overline{\partial}_2\overline{g(z_2)}\Theta(z_1, z_2),
$$

(4.31)

with $\Theta(z_1, z_2)$ is defined in the rhs. of (4.28).

We compute the r.h.s. of (4.31) as stated in Lemma 4.9. The proof of this lemma is postponed to Appendix A.

Lemma 4.9. *Let $\Theta(z_1, z_2)$ be defined in (4.28), then we have that*

$$
\frac{2}{\pi^2}\int_{\mathbf{C}}\mathrm{d}^2 z_1\int_{\mathbf{C}}\mathrm{d}^2 z_2 \partial_1\overline{\partial}_1 f(z_1)\partial_2\overline{\partial}_2\overline{g(z_2)}\Theta(z_1, z_2) = \frac{1}{4\pi}\int_{\mathbf{D}}\langle\nabla g, \nabla f\rangle\,\mathrm{d}^2 z
$$
$$
+ \lim_{\epsilon\to 0}\left[\frac{1}{2\pi^2}\int_{|z_1|\geq 1}\mathrm{d}^2 z_1\int_{\substack{|1-z_1\overline{z}_2|\geq\epsilon,\\|z_2|\geq 1}}\mathrm{d}^2 z_2\,\partial_1 f(z_1)\overline{\partial}_2\overline{g(z_2)}\frac{1}{(1-\overline{z}_1 z_2)^2}\right.
$$
$$
\left.+ \frac{1}{2\pi^2}\int_{|z_1|\geq 1}\mathrm{d}^2 z_1\int_{\substack{|1-z_1\overline{z}_2|\geq\epsilon,\\|z_2|\geq 1}}\mathrm{d}^2 z_2\,\overline{\partial}_1 f(z_1)\partial_2\overline{g(z_2)}\frac{1}{(1-z_1\overline{z}_2)^2}\right].
$$

(4.32)

Proof of Lemma 4.8. By Lemma 4.9 it follows that to prove Lemma 4.8 it is enough to compute the last two lines in the rhs. of (4.32).

Note that using the change of variables $\overline{z}_1 \to 1/\overline{z}_1$, $z_2 \to 1/z_2$ the integral in the rhs. of (4.32) is equal to the same integral on the domain $|z_1|, |z_2| \leq 1$, $|1 - z_1\overline{z}_2| \geq \epsilon$. By a standard density argument, using that $f, g \in H_0^{2+\delta}$, it is enough to compute the limit in (4.32) only for polynomials, hence, from now on, we consider polynomials f, g of the form

$$
f(z_1) = \sum_{k,l\geq 0} z_1^k\overline{z}_1^l a_{kl}, \qquad g(z_2) = \sum_{k,l\geq 0} z_2^k\overline{z}_2^l b_{kl},
$$

(4.33)

for some coefficients $a_{kl}, b_{kl} \in \mathbf{C}$. We remark that the summations in (4.33) are finite since f and g are polynomials. Then, using that

$$
\lim_{\epsilon\to 0}\int_{|z_1|\leq 1}\int_{\substack{|1-z_1\overline{z}_2|\geq\epsilon,\\|z_2|\leq 1}} z_1^\alpha\overline{z}_1^\beta z_2^{\alpha'}\overline{z}_2^{\beta'}\,\mathrm{d}^2 z_1\,\mathrm{d}^2 z_2 = \frac{\pi^2}{(\alpha+1)(\alpha'+1)}\delta_{\alpha,\beta}\delta_{\alpha',\beta'},
$$

we compute the limit in the rhs. of (4.32) as follows

$$
\lim_{\epsilon \to 0} \sum_{k,l,k',l',m \geq 0} \frac{1}{2\pi^2} \int_{|z_1| \leq 1} \int_{\substack{|1-z_1\overline{z}_2| \geq \epsilon, \\ |z_2| \leq 1}} \mathrm{d}^2 z_1 \, \mathrm{d}^2 z_2 \, m a_{kl} \overline{b_{k'l'}}
$$
$$
\times \left[kk' z_1^{k-1} \overline{z}_1^{l+m-1} z_2^{l'+m-1} \overline{z}_2^{k'-1} + ll' z_1^{k+m-1} \overline{z}_1^{l-1} z_2^{k'+m-1} \overline{z}_2^{l'-1} \right]
$$
$$
= \frac{1}{2} \sum_{\substack{k,l,k',l', \\ m \geq 0}} m a_{kl} \overline{b_{k'l'}} \left[\delta_{k,l+m} \delta_{k',l'+m} + \delta_{k,l-m} \delta_{k',l'-m} \right] \tag{4.34}
$$
$$
= \frac{1}{2} \sum_{\substack{k,l,k',l' \geq 0, \\ m \in \mathbf{Z}}} |m| a_{kl} \overline{b_{k'l'}} \delta_{k,l+m} \delta_{k',l'+m}.
$$

On the other hand

$$
\sum_{m \in \mathbf{Z}} |m| \widehat{f \restriction_{\partial \mathbf{D}}}(m) \overline{\widehat{g \restriction_{\partial \mathbf{D}}}(m)} = \sum_{m \in \mathbf{Z}} |m| \sum_{k,l,k',l' \geq 0} a_{kl} \overline{b_{k'l'}} \delta_{m,k-l} \delta_{m,k'-l'}, \tag{4.35}
$$

where

$$
\widehat{f \restriction_{\partial \mathbf{D}}}(k) := \frac{1}{2\pi} \int_0^{2\pi} f \restriction_{\partial \mathbf{D}} (e^{i\theta}) e^{-ik\theta} \, \mathrm{d}\theta, \quad f \restriction_{\partial \mathbf{D}} (e^{i\theta_j}) = \sum_{k \in \mathbf{Z}} \widehat{f \restriction_{\partial \mathbf{D}}}(k) e^{i\theta_j k}.
$$

Finally, combining (4.32) and (4.34)–(4.35), we conclude the proof of (4.30).

$\square$

4.3.3. *Computation of the (z_1, z_2)-integral of $U_1 U_2$.* In order to conclude the proof of Theorem 2.2, in this section we compute the integral of $U_1 U_2$ in (4.23). Similarly to the previous section, we use the notation f and g, instead of $f^{(1)}$, $f^{(2)}$, with f in Theorem 2.2 and $g = f$ or $g = \overline{f}$.

Lemma 4.10. *Let $\kappa_4 = n^2[\mathbf{E}|x_{11}|^2 - 2(\mathbf{E}|x_{11}|^2)]$, and let U_1, U_2 be defined in (3.14), then*

$$
- \frac{\kappa_4}{8\pi^2} \int_{\mathbf{C}} \mathrm{d}^2 z_1 \int_{\mathbf{C}} \mathrm{d}^2 z_2 \Delta f(z_1) \Delta \overline{g(z_2)} \int_0^\infty \mathrm{d}\eta_1 \int_0^\infty \mathrm{d}\eta_2 U_1 U_2
$$
$$
= \kappa_4 \left(\frac{1}{\pi} \int_{\mathbf{D}} f(z) \, \mathrm{d}^2 z - \widehat{f \restriction_{\partial \mathbf{D}}}(0) \right) \left(\frac{1}{\pi} \int_{\mathbf{D}} \overline{g(z)} \, \mathrm{d}^2 z - \overline{\widehat{g \restriction_{\partial \mathbf{D}}}(0)} \right). \tag{4.36}
$$

Proof of Theorem 2.2. Theorem 2.2 readily follows combining Lemma 4.7, Lemma 4.8 and Lemma 4.10.

$\square$

Proof of Lemma 4.10. First of all, we recall the following formulas of integration by parts

$$
\int_{\mathbf{D}} \partial_z f(z, \overline{z}) \, \mathrm{d}^2 z = \frac{i}{2} \int_{\partial \mathbf{D}} f(z, \overline{z}) \, \mathrm{d}\overline{z}, \quad \int_{\mathbf{D}} \partial_{\overline{z}} f(z, \overline{z}) \, \mathrm{d}^2 z = -\frac{i}{2} \int_{\partial \mathbf{D}} f(z, \overline{z}) \, \mathrm{d}z. \tag{4.37}
$$

Then, using the computation of the η-integral of U in (4.29), and integration by parts (4.37) twice, we conclude that

$$
\int_{\mathbf{C}} \Delta f \int_0^\infty U \, \mathrm{d}\eta \, \mathrm{d}^2 z = i2\sqrt{2} \int_{\mathbf{D}} \partial \overline{\partial} f(z)(1 - |z|^2) \, \mathrm{d}^2 z = i2\sqrt{2} \int_{\mathbf{D}} \overline{\partial} f(z) \overline{z} \, \mathrm{d}^2 z
$$
$$
= -i2\sqrt{2} \left(\int_{\mathbf{D}} f(z) \, \mathrm{d}^2 z + \frac{i}{2} \int_{\partial(\mathbf{D})} f(z) \overline{z} \, \mathrm{d}z \right)
$$
$$
= -i2\sqrt{2} \left(\int_{\mathbf{D}} f(z) \, \mathrm{d}^2 z - \pi \widehat{f \restriction_{\partial \mathbf{D}}}(0) \right).
$$

This concludes the proof of this lemma.

$\square$

5. LOCAL LAW FOR PRODUCTS OF RESOLVENTS

The main technical result of this section is a local law for *products* of resolvents with different spectral parameters $z_1 \neq z_2$. Our goal is to find a deterministic approximation to $\langle A G^{z_1} B G^{z_2} \rangle$ for generic bounded deterministic matrices A, B. Due to the correlation between the two resolvents the deterministic approximation to $\langle A G^{z_1} B G^{z_2} \rangle$ is not simply $\langle A M^{z_1} B M^{z_2} \rangle$. In the context of linear statistics such local laws for products of resolvents have previously been obtained e.g. for Wigner matrices in [40] and for sample-covariance matrices in [28] albeit with weaker error bounds. In the current non-Hermitian setting we need such local law twice; for the resolvent CLT in Proposition 3.3, and for the asymptotic independence of resolvents in Proposition 3.5. The key point for the latter is to obtain an improvement in the error term for mesoscopic separation $|z_1 - z_2| \sim n^{-\epsilon}$, a fine effect that has not been captured before.

Our proof applies verbatim to both real and complex i.i.d. matrices, as well as to resolvents $G^z(w)$ evaluated at an arbitrary spectral parameter $w \in \mathbf{H}$. We therefore work with this more general setup in this section, even though for the application in the proofs of Propositions 3.3–3.5 this generality is not necessary.

We recall from [5] that with the shorthand notations

$$G_i := G^{z_i}(w_i), \quad M_i := M^{z_i}(w_i), \tag{5.1}$$

the deviation of G_i from M_i is computed from the identity

$$G_i = M_i - M_i \underline{W G_i} + M_i \mathcal{S}[G_i - M_i] G_i, \quad W := \begin{pmatrix} 0 & X \\ X^* & 0 \end{pmatrix}. \tag{5.2}$$

The relation (5.2) requires some definitions. First, the linear *covariance* or *self-energy operator* $\mathcal{S} \colon \mathbf{C}^{2n \times 2n} \to \mathbf{C}^{2n \times 2n}$ is given by

$$\mathcal{S}\left[\begin{pmatrix} A & B \\ C & D \end{pmatrix} \right] := \widetilde{\mathbf{E}} \widetilde{W} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \widetilde{W} = \begin{pmatrix} \langle D \rangle & 0 \\ 0 & \langle A \rangle \end{pmatrix}, \quad \widetilde{W} = \begin{pmatrix} 0 & \widetilde{X} \\ \widetilde{X}^* & 0 \end{pmatrix}, \tag{5.3}$$

where $\widetilde{X} \sim \mathrm{Gin}_{\mathbf{C}}$, i.e. it averages the diagonal blocks and swaps them. Here $\mathrm{Gin}_{\mathbf{C}}$ stands for the standard complex Ginibre ensemble. The ultimate equality in (5.3) follows directly from $\mathbf{E} \, \widetilde{x}_{ab}^2 = 0$, $\mathbf{E} |\widetilde{x}_{ab}|^2 = n^{-1}$. Second, underlining denotes, for any given function $f \colon \mathbf{C}^{2n \times 2n} \to \mathbf{C}^{2n \times 2n}$, the *self-renormalisation* $\underline{W f(W)}$ defined by

$$\underline{W f(W)} := W f(W) - \widetilde{\mathbf{E}} \widetilde{W} (\partial_{\widetilde{W}} f)(W), \tag{5.4}$$

where ∂ indicates a directional derivative in the direction $\widetilde{W}$ and $\widetilde{W}$ denotes an independent random matrix as in (5.3) with $\widetilde{X}$ a complex Ginibre matrix with expectation $\widetilde{\mathbf{E}}$. Note that we use complex Ginibre $\widetilde{X}$ irrespective of the symmetry class of X. Therefore, using the resolvent identity, it follows that

$$\underline{W G} = W G + \widetilde{\mathbf{E}} \widetilde{W} G \widetilde{W} G = W G + \mathcal{S}[G] G.$$

We now use (5.2) and (5.4) to compute

$$\begin{aligned}
G_1 B G_2 &= M_1 B G_2 - M_1 \underline{W G_1} B G_2 + M_1 \mathcal{S}[G_1 - M_1] G_1 B G_2 \\
&= M_1 B M_2 + M_1 B (G_2 - M_2) - M_1 \underline{W G_1 B G_2} + M_1 \mathcal{S}[G_1 B G_2] M_2 \\
&\quad + M_1 \mathcal{S}[G_1 B G_2](G_2 - M_2) + M_1 \mathcal{S}[G_1 - M_1] G_1 B G_2,
\end{aligned} \tag{5.5}$$

where, in the second equality, we used

$$\begin{aligned}
\underline{W G_1 B G_2} &= \underline{W G_1} B G_2 + \mathcal{S}[G_1] G_1 B G_2 + \mathcal{S}[G_1 B G_2] G_2 \\
&= \underline{W G_1} B G_2 + \mathcal{S}[G_1 B G_2] G_2.
\end{aligned}$$

Assuming that the self-renormalised terms and the ones involving $G_i - M_i$ in (5.5) are small, (5.5) implies

$$G_1 B G_2 \approx M_B^{z_1, z_2}, \tag{5.6}$$

where

$$M_B^{z_1, z_2}(w_1, w_2) := (1 - M^{z_1}(w_1) \mathcal{S}[\cdot] M^{z_2}(w_2))^{-1}[M^{z_1}(w_1) B M^{z_2}(w_2)]. \tag{5.7}$$

We define the corresponding 2-*body stability operator*

$$\widehat{\mathcal{B}} = \widehat{\mathcal{B}}_{12} = \widehat{\mathcal{B}}_{12}(z_1, z_2, w_1, w_2) := 1 - M_1 \mathcal{S}[\cdot] M_2, \tag{5.8}$$

acting on the space of $2n \times 2n$ matrices equipped with the usual Euclidean matrix norm which induces a natural norm for $\widehat{\mathcal{B}}$.

Our main technical result of this section is making (5.6) rigorous in the sense of Theorem 5.2 below. To keep notations compact, we first introduce a commonly used (see, e.g. [36]) notion of high-probability bound.

Definition 5.1 (Stochastic Domination)**.** *If*

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

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

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

for large enough $n \geq n_0(\epsilon, D)$. In this case we use the notation $X \prec Y$.

Theorem 5.2. *Fix $z_1, z_2 \in \mathbf{C}$ and $w_1, w_2 \in \mathbf{C}$ with $|\eta_i| := |\Im w_i| \geq n^{-1}$ such that*

$$\eta_* := \min\{|\eta_1|, |\eta_2|\} \geq n^{-1+\epsilon} \|\widehat{\mathcal{B}}_{12}^{-1}\|$$

for some $\epsilon > 0$. Assume that $G^{z_1}(w_1), G^{z_2}(w_2)$ satisfy the local laws in the form

$$|\langle A(G^{z_i} - M^{z_i}) \rangle| \prec \frac{\|A\|}{n|\eta_i|}, \quad |\langle \boldsymbol{x}, (G^{z_i} - M^{z_i}) \boldsymbol{y} \rangle| \prec \frac{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}{\sqrt{n|\eta_i|}}$$

for any bounded deterministic matrix and vectors $A, \boldsymbol{x}, \boldsymbol{y}$. Then, for any bounded deterministic matrix B, with $\|B\| \lesssim 1$, the product of resolvents $G^{z_1} B G^{z_2} = G^{z_1}(w_1) B G^{z_2}(w_2)$ is approximated by $M_B^{z_1, z_2} = M_B^{z_1, z_2}(w_1, w_2)$ defined in (5.7) in the sense that

$$|\langle A(G^{z_1} B G^{z_2} - M_B^{z_1, z_2}) \rangle| \prec \frac{\|A\| \|\widehat{\mathcal{B}}_{12}^{-1}\|}{n\eta_* |\eta_1 \eta_2|^{1/2}}$$

$$\times \left(\eta_*^{1/12} + \eta_*^{1/4} \|\widehat{\mathcal{B}}_{12}^{-1}\| + \frac{1}{\sqrt{n\eta_*}} + \frac{\|\widehat{\mathcal{B}}_{12}^{-1}\|^{1/4}}{(n\eta_*)^{1/4}} \right), \tag{5.9}$$

$$|\langle \boldsymbol{x}, (G^{z_1} B G^{z_2} - M_B^{z_1, z_2}) \boldsymbol{y} \rangle| \prec \frac{\|\boldsymbol{x}\| \|\boldsymbol{y}\| \|\widehat{\mathcal{B}}_{12}^{-1}\|}{(n\eta_*)^{1/2} |\eta_1 \eta_2|^{1/2}}$$

for any deterministic $A, \boldsymbol{x}, \boldsymbol{y}$.

The estimates in (5.9) will be complemented by a upper bound on $\|\widehat{\mathcal{B}}^{-1}\|$ in Lemma 6.1, where we will prove in particular that $\|\widehat{\mathcal{B}}^{-1}\| \lesssim n^{2\delta}$ whenever $|z_1 - z_2| \gtrsim n^{-\delta}$, for some small fixed $\delta > 0$.

The proof of Theorem 5.2 will follow from a bootstrap argument once the main input, the following high-probability bound on $\underline{W G_1 B G_2}$ has been established.

Proposition 5.3. *Under the assumptions of Theorem 5.2, the following estimates hold uniformly in $n^{-1} \lesssim |\eta_1|, |\eta_2| \lesssim 1$.*

(i) *We have the isotropic bound*

$$|\langle \boldsymbol{x}, \underline{W G_1 B G_2} \boldsymbol{y} \rangle| \prec \frac{1}{(n\eta_*)^{1/2} |\eta_1 \eta_2|^{1/2}} \tag{5.10a}$$

uniformly for deterministic vectors and matrix $\|\boldsymbol{x}\| + \|\boldsymbol{y}\| + \|B\| \leq 1$.

(ii) *Assume that for some positive deterministic $\theta = \theta(z_1, z_2, \eta_*)$ an a priori bound*

$$|\langle A G_1 B G_2 \rangle| \prec \theta \tag{5.10b}$$

has already been established uniformly in deterministic matrices $\|A\| + \|B\| \leq 1$. *Then we have the improved averaged bound*

$$|\langle WG_1BG_2A\rangle| \prec \frac{1}{n\eta_*|\eta_1\eta_2|^{1/2}}\Big((\theta\eta_*)^{1/4} + \frac{1}{\sqrt{n\eta_*}} + \eta_*^{1/12}\Big), \qquad (5.10c)$$

again uniformly in deterministic matrices $\|A\| + \|B\| \leq 1$.

Proof of Theorem 5.2. We note that from (5.7) and (3.3) we have

$$\|M_B^{z_1,z_2}\| \lesssim \|\widehat{\mathcal{B}}^{-1}\| \qquad (5.11)$$

and abbreviate $G_{12} := G_1 B G_2$, $M_{12} := M_B^{z_1,z_2}$. We now assume an a priori bound $|\langle G_{12}A\rangle| \prec \theta_1$, i.e. that (5.10b) holds with $\theta = \theta_1$. In the first step we may take $\theta_1 = |\eta_1\eta_2|^{-1/2}$ due to the local law for G_i from which it follows that

$$|\langle AG_1BG_2\rangle| \leq \sqrt{\langle AG_1G_1^*A^*\rangle}\sqrt{\langle BG_2G_2^*B^*\rangle}$$

$$= \frac{1}{\sqrt{|\eta_1||\eta_2|}}\sqrt{\langle A\Im G_1 A^*\rangle}\sqrt{\langle B\Im G_2 B^*\rangle} \prec \theta_1.$$

By (5.5) and (5.7) we have

$$\widehat{\mathcal{B}}[G_{12} - M_{12}] = M_1 B(G_2 - M_2) - M_1 \underline{WG_{12}} + M_1\mathcal{S}[G_{12}](G_2 - M_2)$$
$$+ M_1\mathcal{S}[G_1 - M_1]G_{12}, \qquad (5.12)$$

and from (3.9) and (5.10c) we obtain

$$|\langle A(G_{12} - M_{12})\rangle| = |\langle A^*, \widehat{\mathcal{B}}^{-1}\widehat{\mathcal{B}}[G_{12} - M_{12}]\rangle| = |\langle (\widehat{\mathcal{B}}^*)^{-1}[A^*]^*\widehat{\mathcal{B}}[G_{12} - M_{12}]\rangle|$$

$$\prec \|\widehat{\mathcal{B}}^{-1}\|\Big[\frac{1}{n\eta_*} + \frac{(\theta_1\eta_*)^{1/4} + (\sqrt{n\eta_*})^{-1} + \eta_*^{1/12}}{n\eta_*|\eta_1\eta_2|^{1/2}} + \frac{\theta_1}{n\eta_*}\Big].$$

For the terms involving $G_i - M_i$ we used that $\mathcal{S}[R] = \langle RE_2\rangle E_1 + \langle RE_1\rangle E_2$ with the $2n \times 2n$ block matrices

$$E_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad E_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad (5.13)$$

i.e. that $\mathcal{S}$ effectively acts as a trace, so that the averaged bounds are applicable. Therefore with (5.11) it follows that

$$|\langle G_{12}A\rangle| \prec \theta_2 := \|\widehat{\mathcal{B}}^{-1}\|\Big[1 + \frac{1}{n\eta_*} + \frac{(\theta_1\eta_*)^{1/4} + (\sqrt{n\eta_*})^{-1} + \eta_*^{1/12}}{n\eta_*|\eta_1\eta_2|^{1/2}} + \frac{\theta_1}{n\eta_*}\Big]. \qquad (5.14)$$

By iterating (5.14) we can use $|\langle G_{12}A\rangle| \prec \theta_2 \ll \theta_1$ as new input in (5.10b) to obtain $|\langle G_{12}A\rangle| \prec \theta_3 \ll \theta_2$ since $n\eta_* \gg \|\widehat{\mathcal{B}}^{-1}\|$. Here θ_j, for $j = 3, 4, \ldots$, is defined iteratively by replacing θ_1 with θ_{j-1} in the rhs. of the defining equation for θ_2 in (5.14). This improvement continues until the fixed point of this iteration, i.e. until $\theta_N^{3/4}$ approaches $\|\widehat{\mathcal{B}}^{-1}\|n^{-1}\eta_*^{-7/4}$. For any given $\xi > 0$, after finitely many steps $N = N(\xi)$ the iteration stabilizes to

$$\theta_* \lesssim n^\xi\Big[\|\widehat{\mathcal{B}}^{-1}\| + \frac{\|\widehat{\mathcal{B}}^{-1}\|}{n\eta_*}\frac{\eta_*^{1/12}}{|\eta_1\eta_2|^{1/2}} + \frac{1}{\eta_*}\Big(\frac{\|\widehat{\mathcal{B}}^{-1}\|}{n\eta_*}\Big)^{4/3}\Big],$$

from which

$$|\langle A(G_{12} - M_{12})\rangle| \prec \frac{\|\widehat{\mathcal{B}}^{-1}\|}{n\eta_*|\eta_1\eta_2|^{1/2}}\Big(\eta_*^{1/12} + \eta_*^{1/4}\|\widehat{\mathcal{B}}^{-1}\| + \frac{1}{\sqrt{n\eta_*}} + \Big(\frac{\|\widehat{\mathcal{B}}^{-1}\|}{n\eta_*}\Big)^{1/4}\Big),$$

and therefore the averaged bound in (5.9) follows.

For the isotropic bound in (5.9) note that

$$\langle \boldsymbol{x}, (G_{12} - M_{12})\boldsymbol{y}\rangle = \text{Tr}\big[(\widehat{\mathcal{B}}^*)^{-1}[\boldsymbol{xy}^*]\big]^*\widehat{\mathcal{B}}[G_{12} - M_{12}]$$

and that due to the block-structure of $\widehat{\mathcal{B}}$ we have

$$(\widehat{\mathcal{B}}^*)^{-1}[\boldsymbol{xy}^*] = \sum_{i=1}^4 \boldsymbol{x}_i\boldsymbol{y}_i^*, \qquad \|\boldsymbol{x}_i\|\|\boldsymbol{y}_i\| \lesssim \|\widehat{\mathcal{B}}^{-1}\|,$$

for some vectors $\boldsymbol{x}_i, \boldsymbol{y}_i$. The isotropic bound in (5.9) thus follows in combination with the isotropic bound in (3.9), (5.12) and (5.10a) applied to the pairs of vectors $\boldsymbol{x}_i, \boldsymbol{y}_i$. This completes the proof of the theorem modulo the proof of Proposition 5.3. $\qquad\square$

5.1. **Probabilistic bound and the proof of Proposition 5.3.** We follow the graphical expansion outlined in [38, 37] adapted to the current setting. We focus on the case when X has complex entries and additionally mention the few changes required when X is a real matrix. We abbreviate $G_{12} = G_1 B G_2$ and use iterated cumulant expansions to expand $\mathbf{E}|\langle \boldsymbol{x}, \underline{WG_{12}\boldsymbol{y}}\rangle|^{2p}$ and $\mathbf{E}|\langle \underline{WG_{12}A}\rangle|^{2p}$ in terms of polynomials in entries of G. For the expansion of the first W we have in the complex case

$$
\mathbf{E}\,\mathrm{Tr}(\underline{WG_{12}A})\,\mathrm{Tr}(\underline{WG_{12}A})^{p-1}\,\mathrm{Tr}(A^*\underline{G_{12}^*W})^p
$$
$$
= \frac{1}{n}\,\mathbf{E}\sum_{ab} R_{ab}\,\mathrm{Tr}(\Delta^{ab}G_{12}A)\partial_{ba}\Big[\mathrm{Tr}(\underline{WG_{12}A})^{p-1}\,\mathrm{Tr}(A^*\underline{G_{12}^*W})^p\Big]
$$
$$
+ \sum_{k\geq 2}\sum_{ab}\sum_{\boldsymbol{\alpha}\in\{ab,ba\}^k}\frac{\kappa(ab,\boldsymbol{\alpha})}{k!}
$$
$$
\times \mathbf{E}\,\partial_{\boldsymbol{\alpha}}\Big[\mathrm{Tr}(\Delta^{ab}G_{12}A)\,\mathrm{Tr}(\underline{WG_{12}A})^{p-1}\,\mathrm{Tr}(A^*\underline{G_{12}^*W})^p\Big]
$$

(5.15)

and similarly for $\langle \boldsymbol{x}, \underline{WG_{12}\boldsymbol{y}}\rangle$, where unspecified summations $\sum_a$ are understood to be over $\sum_{a\in[2n]}$, and $(\Delta^{ab})_{cd} := \delta_{ac}\delta_{bd}$. Here we introduced the matrix $R_{ab} := \mathbf{1}(a\leq n, b>n) + \mathbf{1}(a>n, b\leq n)$ which is the rescaled second order cumulant (variance), i.e. $R_{ab} = n\kappa(ab,ba)$. For $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_k)$ we denote the joint cumulant of $w_{ab}, w_{\alpha_1},\ldots,w_{\alpha_k}$ by $\kappa(ab,\boldsymbol{\alpha})$ which is non-zero only for $\boldsymbol{\alpha}\in \{ab,ba\}^k$. The derivative $\partial_{\boldsymbol{\alpha}}$ denotes the derivative with respect to $w_{\alpha_1},\ldots,w_{\alpha_k}$. Note that in (5.15) the $k=1$ term differs from the $k\geq 2$ terms in two aspects. First, we only consider the ∂_{ba} derivative since in the complex case we have $\kappa(ab,ab) = 0$. Second, the action of the derivative on the first trace is not present since it is cancelled by the *self-renormalisation* of $\underline{WG_{12}}$.

In the real case (5.15) differs slightly. First, for the $k=1$ terms both ∂_{ab} and ∂_{ba} have to be taken into account with the same weight R since $\kappa(ab,ab) = \kappa(ab,ba)$. Second, we chose only to renormalise the effect of the ∂_{ba}-derivative and hence the ∂_{ab}-derivative acts on all traces. Thus in the real case, compared to (5.15) there is an additional term given by

$$
\frac{1}{n}\,\mathbf{E}\sum_{ab} R_{ab}\partial_{ab}\Big[\mathrm{Tr}(\Delta^{ab}G_{12}A)\,\mathrm{Tr}(\underline{WG_{12}A})^{p-1}\,\mathrm{Tr}(A^*\underline{G_{12}^*W})^p\Big].
$$

The main difference to [38, Section 4] and [37, Section 4] is that therein instead of $\underline{WG_{12}}$ the single-G renormalisation $\underline{WG}$ was considered. With respect to the action of the derivatives there is, however, little difference between the two since we have

$$
\partial_{ab}G = -G\Delta^{ab}G, \quad \partial_{ab}G_{12} = -G_1\Delta^{ab}G_{12} - G_{12}\Delta^{ab}G_2.
$$

Therefore after iterating the expansion (5.15) we structurally obtain the same polynomials as in [38, 37], except of the slightly different combinatorics and the fact that exactly $2p$ of the G's are G_{12}'s and the remaining G's are either G_1 or G_2. Thus, using the local law for G_i in the form

$$
|\langle \boldsymbol{x}, G_i\boldsymbol{y}\rangle| \prec 1,
$$
$$
|\langle \boldsymbol{x}, G_{12}\boldsymbol{y}\rangle| \leq \sqrt{\langle \boldsymbol{x}, G_1 G_1^* \boldsymbol{x}\rangle}\sqrt{\langle \boldsymbol{y}, BG_2 G_2^* B^*\boldsymbol{y}\rangle}
$$
$$
= \frac{1}{\sqrt{|\eta_1||\eta_2|}}\sqrt{\langle \boldsymbol{x}, (\Im G_1)\boldsymbol{x}\rangle}\sqrt{\langle \boldsymbol{y}, B(\Im G_2)B^*\boldsymbol{y}\rangle} \prec \frac{1}{\sqrt{|\eta_1||\eta_2|}}
$$

(5.16)

for $\|\boldsymbol{x}\| + \|\boldsymbol{y}\| \lesssim 1$, we obtain exactly the same bound as in [38, Eq. (23a)] times a factor of $(|\eta_1||\eta_2|)^{-p}$ accounting for the $2p$ exceptional G_{12} edges, i.e.

$$
\mathbf{E}|\langle \boldsymbol{x}, \underline{WG_{12}\boldsymbol{y}}\rangle|^{2p} \lesssim \frac{n^\epsilon}{(n\eta_*)^p|\eta_1|^p|\eta_2|^p}, \quad \mathbf{E}|\langle \underline{WG_{12}A}\rangle|^{2p} \lesssim \frac{n^\epsilon}{(n\eta_*)^{2p}|\eta_1|^p|\eta_2|^p}.
$$

(5.17)

The isotropic bound from (5.17) completes the proof of (5.10a).

It remains to improve the averaged bound in (5.17) in order to obtain (5.10c). We first have to identify where the bound (5.17) is suboptimal. By iterating the expansion (5.15) we obtain a complicated polynomial expression in terms of entries of G_{12}, G_1, G_2 which is most conveniently represented graphically as

$$\mathbf{E}|\langle \underline{W G_{12} A}\rangle|^{2p} = \sum_{\Gamma \in \mathrm{Graphs}(p)} c(\Gamma)\, \mathbf{E}\, \mathrm{Val}(\Gamma) + \mathcal{O}\left(n^{-2p}\right) \tag{5.18}$$

for some finite collection of $\mathrm{Graphs}(p)$. Before we precisely define the *value of* Γ, $\mathrm{Val}(\Gamma)$, we first give two examples. Continuing (5.15) in case $p = 1$ we have

$$
\begin{aligned}
&\mathbf{E}\, \mathrm{Tr}(\underline{W G_{12} A})\, \mathrm{Tr}(A^* \underline{G_{12}^* W}) \\
&= \sum_{ab} \frac{R_{ab}}{n}\, \mathbf{E}\, \mathrm{Tr}(\Delta^{ab} G_{12} A)\, \mathrm{Tr}(A^* G_{12}^* \Delta^{ba}) \\
&\quad - \sum_{ab} \frac{R_{ab}}{n}\, \mathbf{E}\, \mathrm{Tr}(\Delta^{ab} G_{12} A)\, \mathrm{Tr}(A^* \underline{G_2^* \Delta^{ba} G_{12}^* W}) \\
&\quad - \sum_{ab} \frac{R'_{ab}}{2n^{3/2}}\, \mathbf{E}\, \mathrm{Tr}(\Delta^{ab} G_1 \Delta^{ba} G_{12} A)\, \mathrm{Tr}(A^* G_{12}^* \Delta^{ba}) + \cdots
\end{aligned}
\tag{5.19a}
$$

where, for illustration, we only kept two of the three Gaussian terms (the last being when W acts on G_1^*) and one non-Gaussian term. For the non-Guassian term we set $R'_{ab} := n^{3/2} \kappa(ab, ba, ba)$, $|R'_{ab}| \lesssim 1$. Note that in the case of i.i.d. matrices with $\sqrt{n} x_{ab} \overset{\mathrm{d}}{=} x$, we have $R'_{ab} = \kappa(x, \overline{x}, \overline{x})$ for $a \leq n, b > n$ and $R'_{ab} = \kappa(x, x, \overline{x}) = \overline{\kappa(x, \overline{x}, \overline{x})}$ for $a > n, b \leq n$. For our argument it is of no importance whether matrices representing cumulants of degree at least three like R' are block-constant. It is important, however, that the variance $\kappa(ab, ba)$ represented by R is block-constant since later we will perform certain resummations. For the second term on the rhs. of (5.19a) we then obtain by another cumulant expansion that

$$
\begin{aligned}
&\sum_{ab} \frac{R_{ab}}{n}\, \mathbf{E}\, \mathrm{Tr}(\Delta^{ab} G_{12} A)\, \mathrm{Tr}(A^* \underline{G_2^* \Delta^{ba} G_{12}^* W}) \\
&= -\sum_{ab}\sum_{cd} \frac{R_{ab} R_{cd}}{n^2}\, \mathbf{E}(G_{12} \Delta^{dc} G_2 A)_{ba}\, \mathrm{Tr}(A^* G_2^* \Delta^{ba} G_{12}^* \Delta^{cd}) + \cdots \\
&\quad - \sum_{ab}\sum_{cd} \frac{R_{ab} R'_{cd}}{2! n^{5/2}}\, \mathbf{E}(G_{12} \Delta^{dc} G_2 A)_{ba}\, \mathrm{Tr}(A^* G_2^* \Delta^{ba} G_{12}^* \Delta^{dc} G_1^* \Delta^{cd}),
\end{aligned}
\tag{5.19b}
$$

where we kept one of the two Gaussian terms and one third order term. After writing out the traces, (5.19a)–(5.19b) become

$$
\begin{aligned}
&\sum_{ab} \frac{R_{ab}}{n}\, \mathbf{E}(G_{12} A)_{ba} (A^* G_{12}^*)_{ab} + \cdots \\
&- \sum_{ab} \frac{R'_{ab}}{n^{3/2}}\, \mathbf{E}(G_1)_{bb} (G_{12} A)_{aa} (A^* G_{12}^*)_{ab} \\
&+ \sum_{ab}\sum_{cd} \frac{R_{ab} R_{cd}}{n^2}\, \mathbf{E}(G_{12})_{bd} (G_2 A)_{ca} (A^* G_2^*)_{db} (G_{12}^*)_{ac} \\
&+ \sum_{ab}\sum_{cd} \frac{R_{ab} R'_{cd}}{2! n^{5/2}}\, \mathbf{E}(G_{12})_{bd} (G_2 A)_{ca} (G_1^*)_{cc} (A^* G_2^*)_{db} (G_{12}^*)_{ad}.
\end{aligned}
\tag{5.19c}
$$

If X is real, then in (5.19) some additional terms appear since $\kappa(ab, ab) = \kappa(ab, ba)$ in the real case, while $\kappa(ab, ab) = 0$ in the complex case. In the first equality of (5.19) this results in additional terms

like

$$\sum_{ab} \frac{R_{ab}}{n} \mathbf{E}\Big(-\operatorname{Tr}(\Delta^{ab}G_1\Delta^{ab}G_{12}A)\operatorname{Tr}(A^*\underline{G_{12}^*W})$$

$$+ \operatorname{Tr}(\Delta^{ab}G_{12}A)\operatorname{Tr}(A^*G_{12}^*\Delta^{ab}) \tag{5.20}$$

$$- \operatorname{Tr}(\Delta^{ab}G_{12}A)\operatorname{Tr}(A^*\underline{G_2^*\Delta^{ab}G_{12}^*W}) + \dots \Big).$$

Out of the three terms in (5.20), however, only the first one is qualitatively different from the terms already considered in (5.19) since the other two are simply transpositions of already existing terms. After another expansion of the first term in (5.20) we obtain terms like

$$\sum_{ab} \frac{R_{ab}}{n}(G_{12}A)_{ba}(A^*G_{12}^*)_{ba} + \cdots$$

$$+ \sum_{ab}\sum_{cd} \frac{R_{ab}R_{cd}}{n^2}(G_1)_{ba}(G_{12}A)_{ba}(A^*G_2^*)_{dc}(G_{12}^*)_{dc} \tag{5.21}$$

$$+ \sum_{ab}\sum_{cd} \frac{R_{ab}R'_{cd}}{2!n^{5/2}}(G_{12})_{bc}(G_2A)_{da}(A^*G_2^*)_{da}(G_{12}^*)_{bd}(G_2^*)_{cc}$$

specific to the real case.

Now we explain how to encode (5.19) in the graphical formalism (5.18). The summation labels a_i, b_i correspond to vertices, while matrix entries correspond to edges between respective labelled vertices. We distinguish between the cumulant- or κ-edges E_κ, like R, R' and G-edges E_G, like $(A^*G_2^*)_{db}$ or $(G_{12}^*)_{ab}$, but do not graphically distinguish between $G_1, G_{12}, A^*G_2^*$, etc. The four terms from the rhs. of (5.19) would thus be represented as

$$\tag{5.22}$$
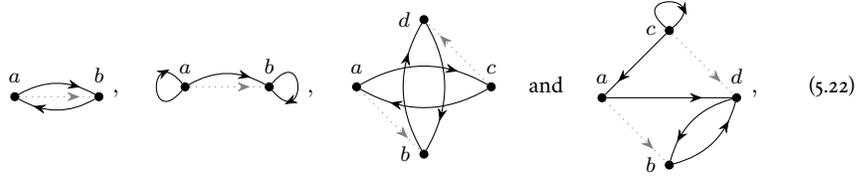
where the edges from E_G are solid and those from E_κ dotted. Similarly, the three examples from (5.21) would be represented as

$$\tag{5.23}$$
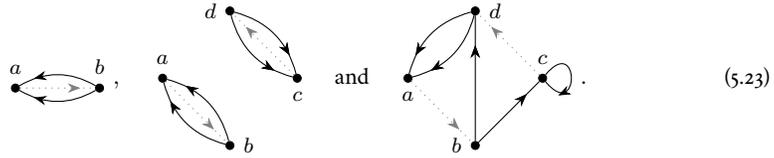
It is not hard to see that after iteratively performing cumulant expansions up to order $4p$ for each remaining W we obtain a finite collection of polynomial expressions in R and G which correspond to graphs Γ from a certain set $\operatorname{Graphs}(p)$ with the following properties. We consider a directed graph $\Gamma = (V, E_\kappa \cup E_G)$ with an even number $|V| = 2k$ of vertices, where k is the number of cumulant expansions along the iteration. The edge set is partitioned into two types of disjoint edges, the elements of E_κ are called *cumulant edges* and the elements of E_G are called *G-edges*. For $u \in V$ we define the G-degree of u as

$$d_G(u) := d_G^{\text{out}}(u) + d_G^{\text{in}}(u),$$

$$d_G^{\text{out}}(u) := |\{v \in V \mid (uv) \in E_G\}|, \quad d_G^{\text{in}}(u) := |\{v \in V \mid (vu) \in E_G\}|.$$

We now record some structural attributes.

(A1) The graph (V, E_κ) is a perfect matching and in particular $|V| = 2|E_\kappa|$. For convenience we label the vertices by $u_1, \ldots, u_k, v_1, \ldots, v_k$ with cumulant edges $(u_1 v_1), \ldots, (u_k v_k)$. The ordering of the elements of E_κ indicated by $1, \ldots, k$ is arbitrary and irrelevant.

(A2) The number of κ-edges is bounded by $|E_\kappa| \le 2p$ and therefore $|V| \le 4p$

(A3) For each $(u_i v_i) \in E_\kappa$, the G-degree of both vertices agrees, i.e. $d_G(u_i) = d_G(v_i) =: d_G(i)$. Furthermore the G-degree satisfies $2 \le d_G(i) \le 4p$. Note that loops (uu) contribute a value of 2 to the degree.

(A4) If $d_G(i) = 2$, then no loops are adjacent to either u_i or v_i.

(A5) We distinguish two types of G-edges $E_G = E_G^1 \cup E_G^2$ whose numbers are given by

$$|E_G^2| = 2p, \quad |E_G^1| = \sum_i d_G(i) - 2p, \quad |E_G| = |E_G^1| + |E_G^2|.$$

Note that in the examples (5.23) above we had $|E_\kappa| = 1$ in the first and $|E_\kappa| = 2$ in the other two cases. For the degrees we had $d_G(1) = 2$ in the first case, $d_G(1) = d_G(2) = 2$ in the second case, and $d_G(1) = 2, d_G(2) = 3$ in the third case. The number of G-edges involving G_{12} is 2 in all cases, while the number of remaining G-edges is 0, 2 and 3, respectively, in agreement with (A5). We now explain how we relate the graphs to the polynomial expressions they represent.

(I1) Each vertex $u \in V$ corresponds to a summation $\sum_{a \in [2n]}$ with a label a assigned to the vertex u.

(I2) Each G-edge $(uv) \in E_G^1$ represents a matrix $\mathcal{G}^{(uv)} = A_1 G_i A_2$ or $\mathcal{G}^{(uv)} = A_1 G_i^* A_2$ for some norm-bounded deterministic matrices A_1, A_2. Each G-edge $(uv) \in E_G^2$ represents a matrix $\mathcal{G}^{(uv)} = A_1 G_{12} A_2$ or $\mathcal{G}^{(uv)} = A_1 G_{12}^* A_2$ for norm bounded matrices A_1, A_2. We denote the matrices $\mathcal{G}^{(uv)}$ with a calligraphic "G" to avoid confusion with the ordinary resolvent matrix G.

(I3) Each κ-edge (uv) represents the matrix

$$R_{ab}^{(uv)} = \kappa(\underbrace{\sqrt{n} w_{ab}, \ldots, \sqrt{n} w_{ab}}_{d_G^{\text{in}}(u)}, \underbrace{\sqrt{n} \overline{w_{ab}}, \ldots, \sqrt{n} \overline{w_{ab}}}_{d_G^{\text{out}}(u)}),$$

where $d_G^{\text{in}}(u) = d_G^{\text{out}}(v)$ and $d_G^{\text{out}}(u) = d_G^{\text{in}}(v)$ are the in- and out degrees of u, v.

(I4) Given a graph Γ we define its value[4] as

$$\text{Val}(\Gamma) := n^{-2p} \prod_{(u_i v_i) \in E_\kappa} \left(\sum_{a_i, b_i \in [2n]} n^{-d_G(i)/2} R_{a_i b_i}^{(u_i v_i)} \right) \prod_{(u_i v_i) \in E_G} \mathcal{G}_{a_i b_i}^{(u_i v_i)}, \qquad (5.24)$$

where $R^{(u_i v_i)}$ is as in (I3) and a_i, b_i are the summation indices associated with u_i, v_i.

Proof of (5.18). In order to prove (5.18) we have to check that the graphs representing the polynomial expressions of the cumulant expansion up to order $4p$ indeed have the attributes (A1)–(A5). Here (A1)–(A3) follow directly from the construction, with the lower bound $d_G(i) \ge 2$ being a consequence of $\mathbf{E} w_{ab} = 0$ and the upper bound $d_G(i) \le 4p$ being a consequence of the fact that we trivially truncate the expansion after the $4p$-th cumulant. The error terms from the truncation are estimated trivially using (5.16). The fact (A4) that no G-loops may be adjacent to degree two κ-edges follows since due to the self-renormalisation $\underline{W G_{12}}$ the second cumulant of W can only act on some W or G in another trace, or if it acts on some G in its own trace then it generates a $\kappa(ab, ab)$ factor (only possible when X is real). In the latter case one of the two vertices has two outgoing, and the other one two incoming G-edges, and in particular no loops are adjacent to either of them. The counting of G_{12}-edges in E_G^2 in (A5) is trivial since along the procedure no G_{12}-edges can be created or removed. For the counting of G_i edges in E_G^1 note that the action of the k-th order cumulant in the expansion of $\underline{W G_{12}}$ may remove k_1 W's and may create additional k_2 G_i's with $k = k_1 + k_2$, $k_1 \ge 1$. Therefore, since the number of G_i edges is 0 in the beginning, and the number of W's is reduced from $2p$ to 0 the second equality in (A5) follows.

---

[4] In [37] we defined the value with an expectation so that (5.18) holds without expectation. In the present paper we follow the convention of [38] and consider the value as a random variable.

It now remains to check that with the interpretations (I₁)–(I₄) the values of the constructed graphs are consistent in the sense of (5.18). The constant $c(\Gamma) \sim 1$ accounts for combinatorial factors in the iterated cumulant expansions and the multiplicity of identical graphs. The factor n^{-2p} in (I₄) comes from the $2p$ normalised traces. The relation (I₃) follows from the fact that the k-th order cumulant of k_1 copies of w_{ab} and k_2 copies of $\overline{w_{ab}} = w_{ba}$ comes together with k_1 copies of Δ^{ab} and k_2 copies of Δ^{ba}. Thus a is the first index of some G a total of k_2 times, while the remaining k_1 times the first index is b, and for the second indices the roles are reversed. $\qquad\square$

Having established the properties of the graphs and the formula (5.18), we now estimate the value of any individual graph.

Naive estimate. We first introduce the so called *naive estimate*, N-Est(Γ), of a graph Γ as the bound on its value obtained by estimating the factors in (5.24) as $|\mathcal{G}^e_{ab}| \prec 1$ for $e \in E^1_G$ and $|\mathcal{G}^e_{ab}| \prec (|\eta_1||\eta_2|)^{-1/2}$ for $e \in E^2_G$, $|R^e_{ab}| \lesssim 1$ and estimating summations by their size. Thus, we obtain

$$\mathrm{Val}(\Gamma) \prec \mathrm{N\text{-}Est}(\Gamma) := \frac{1}{n^{2p}|\eta_1|^p|\eta_2|^p} \prod_i \Big(n^{2-d_G(i)/2}\Big)$$

$$\le \frac{n^{|E^2_\kappa|}n^{|E^3_\kappa|/2}}{n^{2p}|\eta_1|^p|\eta_2|^p} \le \frac{1}{|\eta_1|^p|\eta_2|^p}, \tag{5.25}$$

where

$$E^j_\kappa := \{(u_i, v_i) \mid d_G(i) = j\}$$

is the set of degree j κ-edges, and in the last inequality we used $|E^2_\kappa| + |E^3_\kappa| \le |E_\kappa| \le 2p$.

Ward estimate. The first improvement over the naive estimate comes from the effect that sums of resolvent entries are typically smaller than the individual entries times the summation size. This effect can easily be seen from the *Ward* or resolvent identity $G^*G = \Im G/\eta = (G - G^*)/(2i\eta)$. Indeed, the naive estimate of $\sum_a G_{ab}$ is n using $|G_{ab}| \prec 1$. However, using the Ward identity we can improve this to

$$\left|\sum_a G_{ab}\right| \le \sqrt{2n}\sqrt{\sum_a |G_{ab}|^2} = \sqrt{2n}\sqrt{(G^*G)_{bb}} = \sqrt{\frac{2n}{\eta}}\sqrt{(\Im G)_{bb}} \prec n\frac{1}{\sqrt{n\eta}},$$

i.e. by a factor of $(n\eta)^{-1/2}$. Similarly, we can gain two such factors if the summation index a appears in two G-factors off-diagonally, i.e.

$$\left|\sum_a (G_1)_{ab}(G_2)_{ca}\right| \le \sqrt{(G^*_1 G_1)_{bb}}\sqrt{(G_2 G^*_2)_{cc}} \prec n\frac{1}{n\eta}.$$

However, it is impossible to gain more than two such factors per summation. We note that we have the same gain also for summations of G_{12}. For example, the naive estimate on $\sum_a (G_{12})_{ab}$ is $n|\eta_1\eta_2|^{-1/2}$ since $|(G_{12})_{ab}| \prec |\eta_1\eta_2|^{-1/2}$. Using the Ward identity, we obtain an improved bound of

$$\left|\sum_a (G_{12})_{ab}\right| \le \sqrt{2n}\sqrt{(G^*_{12}G_{12})_{bb}} = \sqrt{\frac{2n}{|\eta_1|}}\sqrt{(G^*_2 B^*(\Im G_1)BG_2)_{bb}}$$

$$\lesssim \sqrt{\frac{n}{|\eta_1|^2}}\sqrt{(G^*_2 G_2)_{bb}} \prec \frac{\sqrt{n}}{|\eta_1||\eta_2|^{1/2}} \le \frac{n}{|\eta_1\eta_2|^{1/2}}\frac{1}{\sqrt{n\eta_*}},$$

where we recall $\eta_* = \min\{|\eta_1|, |\eta_2|\}$. Each of these improvements is associated with a specific G-edge with the restriction that one cannot gain simultaneously from more than two edges adjacent to any given vertex $u \in V$ while summing up the index a associated with u. Note, however, that globally it is nevertheless possible to gain from arbitrarily many G-edges adjacent to any given vertex, as long as the summation order is chosen correctly. In order to count the number edges giving rise to such improvements we recall a basic definition [64] from graph theory.

Definition 5.4. *For $k \ge 1$ a graph $\Gamma = (V, E)$ is called k-degenerate if any induced subgraph has minimal degree at most k.*

The relevance of this definition in the context of counting the number of gains of $(n\eta_*)^{-1/2}$ lies in the following equivalent characterisation [44].

Lemma 5.5. *A graph* $\Gamma = (V, E)$ *is k-degenerate if and only if there exists an ordering of vertices* $\{v_1, \ldots, v_n\} = V$ *such that for each $m \in [n]$ it holds that*

$$\deg_{\Gamma[\{v_1,\ldots,v_m\}]}(v_m) \leq k \tag{5.26}$$

where for $V' \subset V$, $\Gamma[V']$ denotes the induced subgraph on the vertex set V'.

We consider a subset of non-loop edges $E_{\mathrm{Ward}} \subset E_G \setminus \{(vv) \mid v \in V\}$ for which Ward improvements will be obtained. We claim that if $\Gamma_{\mathrm{Ward}} = (V, E_{\mathrm{Ward}})$ is 2-degenerate, then we may gain a factor of $(n\eta_*)^{-1/2}$ from each edge in E_{Ward}. Indeed, take the ordering $\{v_1, \ldots, v_{2|E_\kappa|}\}$ guaranteed to exist in Lemma 5.5 and first sum up the index a_1 associated with v_1. Since Γ_{Ward} is 2-degenerate there are at most two edges from E_{Ward} adjacent to v_1 and we can gain a factor of $(n\eta_*)^{-1/2}$ for each of them. Next, we can sum up the index associated with vertex v_2 and again gain the same factor for each edge in E_{Ward} adjacent to v_2. Continuing this way we see that in total we can gain a factor of $(n\eta_*)^{-|E_{\mathrm{Ward}}|/2}$ over the naive bound (5.25).

Definition 5.6 (Ward estimate). *For a graph Γ with fixed subset $E_{\mathrm{Ward}} \subset E_G$ of edges we define*

$$\mathrm{W\text{-}Est}(\Gamma) := \frac{\mathrm{N\text{-}Est}(\Gamma)}{(n\eta_*)^{|E_{\mathrm{Ward}}|/2}}.$$

By considering only G-edges adjacent to κ-edges of degrees 2 and 3 it is possible to find such a 2-degenerate set with

$$|E_{\mathrm{Ward}}| = \sum_i (4 - d_G(i))_+$$

elements, cf. [37, Lemma 4.7]. As a consequence, as compared with the first inequality in (5.25), we obtain an improved bound

$$\mathrm{Val}(\Gamma) \prec \mathrm{W\text{-}Est}(\Gamma) = \frac{1}{n^{2p}|\eta_1\eta_2|^p} (n\eta_*)^{-|E_{\mathrm{Ward}}|/2} \prod_i \left(n^{2-d_G(i)/2}\right)$$

$$= \frac{1}{n^{2p}|\eta_1\eta_2|^p} \prod_{d_G(i)=2} \left(\frac{n}{n\eta_*}\right) \prod_{d_G(i)=3} \left(\frac{\sqrt{n}}{\sqrt{n\eta_*}}\right) \prod_{d_G(i)\geq 4} \left(n^{2-d_G(i)/2}\right) \tag{5.27}$$

$$\lesssim \frac{1}{(n\eta_*)^{2p}|\eta_1\eta_2|^p} \eta_*^{2p+\sum_i(d_G(i)/2-2)} \lesssim \frac{1}{(n\eta_*)^{2p}|\eta_1\eta_2|^p},$$

where in the penultimate inequality we used $n^{-1} \leq \eta_*$, and in the ultimate inequality that $d_G(i) \geq 2$ and $|E_\kappa| \leq 2p$ which implies that the exponent of η_* is non-negative and $\eta_* \lesssim 1$. Thus we gained a factor of $(n\eta_*)^{-2p}$ over the naive estimate (5.25).

Resummation improvements. The bound (5.27) is optimal if $z_1 = z_2$ and if η_1, η_2 have opposite signs. In the general case $z_1 \neq z_2$ we have to use two additional improvements which both rely on the fact that the summations $\sum_{a_i,b_i}$ corresponding to $(u_i, v_i) \in E_\kappa^2$ can be written as matrix products since $d_G(u_i) = d_G(v_i) = 2$. Therefore we can sum up the G-edges adjacent to (u_iv_i) as

$$\sum_{a_ib_i} G_{xa_i}G_{a_iy}G_{zb_i}G_{b_iw}R_{a_ib_i}$$

$$= \sum_{a_ib_i} G_{xa_i}G_{a_iy}G_{zb_i}G_{b_iw}\left[\mathbf{1}(a_i > n, b_i \leq n) + \mathbf{1}(a_i \leq n, b_i > n)\right] \tag{5.28a}$$

$$= (GE_1G)_{xy}(GE_2G)_{zw} + (GE_2G)_{xy}(GE_1G)_{zw},$$

where E_1, E_2 are defined in (5.13), in the case of four involved G's and $d_G^{\mathrm{in}} = d_G^{\mathrm{out}} = 1$. If one vertex has two incoming, and the other two outgoing edges (which is only possible if X is real), then we similarly can sum up

$$\sum_{ab} G_{xa}G_{ya}G_{bz}G_{bw}R_{ab} = (GE_1G^t)_{xy}(G^tE_2G)_{zw} + (GE_2G^t)_{xy}(G^tE_1G)_{zw}, \tag{5.28b}$$

so merely some G is replaced by its transpose G^t compared to (5.28a) which will not change any estimate. In the remaining cases with two and three involved G's we similarly have

$$\sum_{ab} G_{ba}G_{ab}R_{ab} = \operatorname{Tr} GE_1GE_2 + \operatorname{Tr} GE_2GE_1$$

$$\sum_{ab} G_{xa}G_{ab}G_{by}R_{ab} = (GE_1GE_2G)_{xy} + (GE_2GE_1G)_{xy}.$$

(5.28c)

By carrying out all available *partial summations* at degree-2 vertices as in (5.28) for the value $\operatorname{Val}(\Gamma)$ of some graph Γ we obtain a collection of *reduced graphs*, in which cycles of G's are contracted to the trace of their matrix product, and chains of G's are contracted to single edges, also representing the matrix products with two *external* indices. We denote generic cycle-subgraphs of k edges from E_G with vertices of degree two by Γ_k°, and generic chain-subgraphs of k edges from E_G with *internal* vertices of degree two and external vertices of degree at least three by Γ_k^-. With a slight abuse of notation we denote the *value* of Γ_k° by $\operatorname{Tr}\Gamma_k^\circ$, and the *value* of Γ_k^- with external indices (a,b) by $(\Gamma_k^-)_{ab}$, where for a fixed choice of E_1, E_2 in (5.28) the internal indices are summed up. The actual choice of E_1, E_2 is irrelevant for our analysis, hence we will omit it from the notation. The concept of the *naive* and *Ward* estimates of any graph Γ carry over naturally to these chain and cycle-subgraphs by setting

$$\text{N-Est}(\Gamma_k^\circ) := \frac{n^k}{|\eta_1\eta_2|^{|E_G^2(\Gamma_k^\circ)|/2}}, \quad \text{N-Est}(\Gamma_k^-) := \frac{n^{k-1}}{|\eta_1\eta_2|^{|E_G^2(\Gamma_k^-)|/2}},$$

$$\text{W-Est}(\Gamma_k^{\circ/-}) = \frac{\text{N-Est}(\Gamma_k^{\circ/-})}{(n\eta_*)^{|E_{\text{Ward}}(\Gamma_k^{\circ/-})|/2}}, \quad E_{\text{Ward}}(\Gamma_k^{\circ/-}) = E_G(\Gamma_k^{\circ/-}) \cap E_{\text{Ward}}(\Gamma).$$

(5.29)

After contracting the chain- and cycle-subgraphs we obtain $2^{|E_\kappa^2|}$ reduced graphs Γ_{red} on the vertex set

$$V(\Gamma_{\text{red}}) := \{v \in V(\Gamma) \mid d_G(v) \geq 3\}$$

with κ-edges

$$E_\kappa(\Gamma_{\text{red}}) := E_\kappa^{\geq 3}(\Gamma)$$

and G-edges

$$E_G(\Gamma_{\text{red}}) := \{(uv) \in E_G(\Gamma) \mid \min\{d_G(u), d_G(v)\} \geq 3\} \cup E_G^{\text{chain}}(\Gamma_{\text{red}}),$$

with additional *chain-edges*

$$E_G^{\text{chain}}(\Gamma_{\text{red}}) := \left\{(u_1u_{k+1}) \,\middle|\, \begin{matrix} k \geq 2, \, u_1, u_{k+1} \in V(\Gamma_{\text{red}}), \, \exists \Gamma_k^- \subset \Gamma, \\ V(\Gamma_k^-) = (u_1, \ldots, u_{k+1}) \end{matrix}\right\}.$$

The additional chain edges $(u_1u_{k+1}) \in E_G^{\text{chain}}$ naturally represent the matrices

$$\mathcal{G}^{(u_1u_{k+1})} := \left((\Gamma_k^-)_{ab}\right)_{a,b\in[2n]}$$

whose entries are the values of the chain-subgraphs. Note that due to the presence of E_1, E_2 in (5.28) the matrices associated with some G-edges can be multiplied by E_1, E_2. However, since in the definition (I2) of G-edges the multiplication with generic bounded deterministic matrices is implicitly allowed, this additional multiplication will not be visible in the notation. Note that the reduced graphs contain only vertices of at least degree three, and only κ-edges from $E_\kappa^{\geq 3}$. The definition of value, naive estimate and Ward estimate naturally extend to the reduced graphs and we have

$$\operatorname{Val}(\Gamma) = \sum \operatorname{Val}(\Gamma_{\text{red}}) \prod_{\Gamma_k^\circ \subset \Gamma} \operatorname{Tr}\Gamma_k^\circ$$

(5.30)

and

$$\text{N-Est}(\Gamma) = \text{N-Est}(\Gamma_{\text{red}}) \prod_{\Gamma_k^\circ \subset \Gamma} \text{N-Est}(\Gamma_k^\circ),$$

$$\text{W-Est}(\Gamma) = \text{W-Est}(\Gamma_{\text{red}}) \prod_{\Gamma_k^\circ \subset \Gamma} \text{W-Est}(\Gamma_k^\circ).$$

(5.31)

The irrelevant summation in (5.30) of size $2^{|E_\kappa^2|}$ is due to the sums in (5.28).

Let us revisit the examples (5.22) to illustrate the summation procedure. The first two graphs in (5.22) only have degree-2 vertices, so that the reduced graphs are empty with value $n^{-2p} = n^{-2}$, hence

$$\mathrm{Val}(\Gamma) = \frac{1}{n^2} \sum \mathrm{Tr}\,\Gamma_2^\circ \qquad \mathrm{Val}(\Gamma) = \frac{1}{n^2} \sum (\mathrm{Tr}\,\Gamma_2^\circ)(\mathrm{Tr}\,\Gamma_2^\circ),$$

where the summation is over two and, respectively, four terms. The third graph in (5.22) results in no traces but in four reduced graphs

$$\mathrm{Val}(\Gamma) = \sum \mathrm{Val}(\ \begin{array}{c}\includegraphics{}\end{array}\ ),$$

where for convenience we highlighted the chain-edges E_G^{chain} representing Γ_k^- by double lines (note that the two endpoints of a chain edge may coincide, but it is not interpreted as a cycle graph since this common vertex has degree more than two, so it is not summed up into a trace along the reduction process). Finally, to illustrate the reduction for a more complicated graph, we have

where we labelled the vertices for convenience, and the summation on the rhs. is over four assignments of E_1, E_2.

Since we have already established a bound on $\mathrm{Val}(\Gamma) \prec \mathrm{W\text{-}Est}(\Gamma)$ we only have to identify the additional gain from the resummation compared to the *Ward-estimate* (5.27).

We will need to exploit two additional effects:

(i) The Ward-estimate is sub-optimal whenever, after resummation, we have some contracted cycle $\mathrm{Tr}\,\Gamma_k^\circ$ or a reduced graph with a chain-edge Γ_k^- with $k \geq 3$.

(ii) When estimating $\mathrm{Tr}\,\Gamma_k^\circ$, $k \geq 2$ with Γ_k° containing some G_{12}, then also the improved bound from (i) is sub-optimal and there is an additional gain from using the a priori bound $|\langle G_{12}A\rangle| \prec \theta$.

We now make the additional gains (i)–(ii) precise.

Lemma 5.7. *For $k \geq 2$ let Γ_k° and Γ_k^- be some cycle and chain subgraphs.*

(i) *We have*
$$|\mathrm{Tr}\,\Gamma_k^\circ| \prec (n\eta_*)^{-(k-2)/2}\,\mathrm{W\text{-}Est}(\Gamma_k^\circ) \tag{5.32a}$$
and for all a, b
$$|(\Gamma_k^-)_{ab}| \prec (n\eta_*)^{-(k-2)/2}\,\mathrm{W\text{-}Est}(\Gamma_k^-). \tag{5.32b}$$

(ii) *If Γ_k° contains at least one G_{12} then we have a further improvement of $(\eta_*\theta)^{1/2}$, i.e.*
$$|\mathrm{Tr}\,\Gamma_k^\circ| \prec \sqrt{\eta_*\theta}\,(n\eta_*)^{-(k-2)/2}\,\mathrm{W\text{-}Est}(\Gamma_k^\circ), \tag{5.32c}$$
where θ is as in (5.10b).

The proof of Lemma 5.7 follows from the following optimal bound on general products $G_{j_1\ldots j_k}$ of resolvents and generic deterministic matrices.

Lemma 5.8. *Let $w_1, w_2, \ldots, z_1, z_2, \ldots$ denote arbitrary spectral parameters with $\eta_i = \Im w_i > 0$. With $G_j = G^{z_j}(w_j)$ we then denote generic products of resolvents $G_{j_1}, \ldots G_{j_k}$ or their adjoints/-transpositions (in that order) with arbitrary bounded deterministic matrices in between by $G_{j_1\ldots j_k}$, e.g. $G_{1i1} = A_1 G_1 A_2 G_i A_3 G_1 A_4$.*

(i) *For $j_1, \ldots j_k$ we have the isotropic bound*

$$|\langle \boldsymbol{x}, G_{j_1 \ldots j_k} \boldsymbol{y}\rangle| \prec \|\boldsymbol{x}\|\|\boldsymbol{y}\|\sqrt{\eta_{j_1}\eta_{j_k}}\Big(\prod_{n=1}^{k} \eta_{j_n}\Big)^{-1}. \tag{5.33a}$$

(ii) *For $j_1, \ldots, j_k$ and any $1 \leq s < t \leq k$ we have the averaged bound*

$$|\langle G_{j_1 \ldots j_k}\rangle| \prec \sqrt{\eta_{j_s}\eta_{j_t}}\Big(\prod_{n=1}^{k} \eta_{j_n}\Big)^{-1}. \tag{5.33b}$$

Lemma 5.8 for example implies $|(G_{1i})_{ab}| \prec (\eta_1\eta_i)^{-1/2}$ or $|(G_{i1i})_{ab}| \prec (\eta_1\eta_i)^{-1}$. Note that the averaged bound (5.33b) can be applied more flexibly by choosing s, t freely, e.g.

$$|\langle G_{1i1i}\rangle| \prec \min\{\eta_1^{-1}\eta_i^{-2}, \eta_1^{-2}\eta_i^{-1}\},$$

while $|\langle \boldsymbol{x}, G_{1i1i}\boldsymbol{y}\rangle| \prec \|\boldsymbol{x}\|\|\boldsymbol{y}\|(\eta_1\eta_i)^{-3/2}$.

Proof of Lemma 5.8. We begin with

$$|\langle \boldsymbol{x}, G_{j_1 \ldots j_k}\boldsymbol{y}\rangle|$$
$$\leq \sqrt{\langle \boldsymbol{x}, G_{j_1}G_{j_1}^{*}\boldsymbol{x}\rangle}\sqrt{\langle \boldsymbol{y}, G_{j_2 \ldots j_k}^{*}G_{j_2 \ldots j_k}\boldsymbol{y}\rangle} \prec \frac{\|\boldsymbol{x}\|}{\sqrt{\eta_{j_1}}}\sqrt{\langle \boldsymbol{y}, G_{j_2 \ldots j_k}^{*}G_{j_2 \ldots j_k}\boldsymbol{y}\rangle}$$
$$\lesssim \frac{\|\boldsymbol{x}\|}{\sqrt{\eta_{j_1}}}\frac{1}{\eta_{j_2}}\sqrt{\langle \boldsymbol{y}, G_{j_3 \ldots j_k}^{*}G_{j_3 \ldots j_k}\boldsymbol{y}\rangle} \lesssim \cdots$$
$$\lesssim \frac{\|\boldsymbol{x}\|}{\sqrt{\eta_{j_1}}}\frac{1}{\eta_{j_2}\ldots\eta_{j_{k-1}}}\sqrt{\langle \boldsymbol{y}, G_{j_k}^{*}G_{j_k}\boldsymbol{y}\rangle} \prec \frac{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}{\sqrt{\eta_{j_1}\eta_{j_k}}}\frac{1}{\eta_{j_2}\ldots\eta_{j_{k-1}}},$$

where in each step we estimated the middle $G_{j_2}^{*}G_{j_2}, G_{j_3}^{*}G_{j_3}, \ldots$ terms trivially by $1/\eta_{j_2}^2, 1/\eta_{j_3}^2, \ldots$, and in the last step we used Ward estimate. This proves (5.33a). We now turn to (5.33b) where by cyclicity without loss of generality we may assume $s = 1$. Thus

$$|\langle G_{j_1 \ldots j_k}\rangle| \leq \sqrt{\langle G_{j_1 \ldots j_{t-1}}G_{j_1 \ldots j_{t-1}}^{*}\rangle}\sqrt{\langle G_{j_t \ldots j_k}^{*}G_{j_t \ldots j_k}\rangle}$$
$$= \sqrt{\langle G_{j_1 \ldots j_{t-1}}G_{j_1 \ldots j_{t-1}}^{*}\rangle}\sqrt{\langle G_{j_t \ldots j_k}G_{j_t \ldots j_k}^{*}\rangle}$$
$$\lesssim \Big(\prod_{n\neq 1,t}\frac{1}{\eta_{j_n}}\Big)\sqrt{\langle G_{j_1}G_{j_1}^{*}\rangle}\sqrt{\langle G_{j_t}G_{j_t}^{*}\rangle} \prec \frac{1}{\sqrt{\eta_{j_1}\eta_{j_t}}}\Big(\prod_{n\neq 1,t}\frac{1}{\eta_{j_n}}\Big),$$

where in the second step we used cyclicity of the trace, the norm-estimate in the third step und the Ward-estimate in the last step. $\qquad\square$

Proof of Lemma 5.7. For the proof of (5.32a) we recall from the definition of the Ward-estimate in (5.29) that for a cycle Γ_k° we have

$$\text{W-Est}(\Gamma_k^{\circ}) \geq \frac{\text{N-Est}(\Gamma_k^{\circ})}{(n\eta_*)^{k/2}} = \frac{n^{k/2}}{|\eta_1\eta_2|^{|E_G^2(\Gamma_k^{\circ})|/2}}\frac{1}{\eta_*^{k/2}}$$

since $|E_{\text{Ward}}(\Gamma_k^{\circ})| \leq |E_G(\Gamma_k^{\circ})| \leq k$. Thus, together with (5.33b) and interpreting $\text{Tr}\,\Gamma_k^{\circ}$ as a trace of a product of $k + |E_G^2(\Gamma_k^{\circ})|$ factors of G's we conclude

$$|\text{Tr}\,\Gamma_k^{\circ}| \prec \frac{n}{|\eta_1\eta_2|^{|E_G^2(\Gamma_k^{\circ})|}\eta_*^{k-|E_G^2(\Gamma_k^{\circ})|-1}} \leq \frac{n}{|\eta_1\eta_2|^{|E_G^2(\Gamma_k^{\circ})|/2}\eta_*^{k-1}} \leq \frac{\text{W-Est}(\Gamma_k^{\circ})}{(n\eta_*)^{k/2-1}}. \tag{5.34}$$

Note that Lemma 5.8 is applicable here even though therein (for convenience) it was assumed that all spectral parameters w_i have positive imaginary parts. However, the lemma also applies to spectral parameters with negative imaginary parts since it allows for adjoints and $G^z(\overline{w}) = (G^z(w))^*$. The first inequality in (5.34) elementarily follows from (5.33b) by distinguishing the cases $|E_G^2| = k, k-1$

or $\leq k - 2$, and always choosing s and t such that the $\sqrt{\eta_{j_s}\eta_{j_t}}$ factor contains the highest possible η_* power. Similarly to (5.34), for (5.32b) we have, using (5.33a),

$$|(\Gamma_k^-)_{ab}| \prec \frac{n^{k-1}}{|\eta_1\eta_2|^{|E_G^2(\Gamma_k^-)|/2}}\frac{1}{(n\eta_*)^{k/2}} \leq \frac{\text{W-Est}(\Gamma_k^-)}{(n\eta_*)^{k/2-1}}. \tag{5.35}$$

For the proof of (5.32c) we use a Cauchy-Schwarz estimate to isolate a single G_{12} factor from the remaining G's in Γ_l°. We may represent the "square" of all the remaining factors by an appropriate cycle graph $\Gamma_{2(k-1)}^\circ$ of length $2(k-1)$ with $|E_G^2(\Gamma_{2(k-1)}^\circ)| = 2(|E_G^2(\Gamma_k^\circ)| - 1)$. We obtain

$$|\text{Tr}\,\Gamma_k^\circ| \leq \sqrt{\text{Tr}(G_{12}G_{12}^*)}\sqrt{|\text{Tr}\,\Gamma_{2(k-1)}^\circ|} = \sqrt{\text{Tr}\,G_1^*G_1BG_2G_2^*B^*}\sqrt{|\text{Tr}\,\Gamma_{2(k-1)}^\circ|}$$

$$= \frac{\sqrt{\text{Tr}(\Im G_1)B(\Im G_2)B^*}\sqrt{|\text{Tr}\,\Gamma_{2(k-1)}^\circ|}}{\sqrt{|\eta_1\eta_2|}}$$

$$\prec \frac{\sqrt{\theta n}}{\sqrt{|\eta_1\eta_2|}}\frac{\sqrt{n}}{|\eta_1\eta_2|^{|E_2^G(\Gamma_k^\circ)|/2-1/2}\eta_*^{k-3/2}}$$

$$\leq \sqrt{\eta_*\theta}(n\eta_*)^{-(k-2)/2}\,\text{W-Est}(\Gamma_k^\circ)$$

where in the penultimate step we wrote out $\Im G = (G - G^*)/(2i)$ in order to use (5.10b), and used (5.34) for $\Gamma_{2(k-1)}^\circ$. □

Now it remains to count the gains from applying Lemma 5.7 for each cycle- and chain subgraph of Γ. We claim that

$$\text{W-Est}(\Gamma) \leq \left(\eta_*^{1/6}\right)^{d_{\geq 3}}\frac{1}{(n\eta_*)^{2p}|\eta_1\eta_2|^p}, \qquad d_{\geq 3} := \sum_{d_G(i)\geq 3}d_G(i). \tag{5.36a}$$

Furthermore, suppose that Γ has c degree-2 cycles Γ_k° which according to (A3) has to satisfy $0 \leq c' := |E_\kappa^2| - c \leq |E_\kappa^2|$. Then we claim that

$$|\text{Val}(\Gamma)| \prec \left(\frac{1}{n\eta_*}\right)^{(c'-d_{\geq 3}/2)_+}\left(\sqrt{\eta_*\theta}\right)^{(p-c'-d_{\geq 3}/2)_+}\text{W-Est}(\Gamma). \tag{5.36b}$$

Assuming (5.36a)–(5.36b) it follows immediately that

$$|\text{Val}(\Gamma)| \prec \frac{1}{(n\eta_*)^{2p}|\eta_1\eta_2|^p}\left(\sqrt{\eta_*\theta} + \frac{1}{n\eta_*} + \eta_*^{1/6}\right)^p,$$

implying (5.10c). In order to complete the proof of the Proposition 5.3 it remains to verify (5.36a) and (5.36b).

Proof of (5.36a). This follows immediately from the penultimate inequality in (5.27) and

$$\eta_*^{2p+\sum_i(d_G(i)/2-2)} \leq \eta_*^{\sum_i(d_G(i)/2-1)} = \eta_*^{\frac{1}{2}\sum_{d_G(i)\geq 3}(d_G(i)-2)} \leq \eta_*^{\frac{1}{6}\sum_{d_G(i)\geq 3}d_G(i)},$$

where we used (A2) in the first inequality. □

Proof of (5.36b). For cycles Γ_k° or chain-edges Γ_k^- in the reduced graph we say that $\Gamma_k^{\circ/-}$ has $(k-2)_+$ *excess G-edges*. Note that for cycles Γ_k° every additional G beyond the minimal number $k \geq 2$ is counted as an excess G-edge, while for chain-edges Γ_k^- the first additional G beyond the minimal number $k \geq 1$ is not counted as an excess G-edge. We claim that:

(C1) The total number of excess G-edges is at least $2c' - d_{\geq 3}$.
(C2) There are at least $p - c' - d_{\geq 3}/2$ cycles in Γ containing G_{12}.

Since the vertices of the reduced graph are u_i, v_i for $d_G(i) \geq 3$, it follows that the reduced graph has $\sum_{d_G(i)\geq 3}(d_G(u_i) + d_G(v_i))/2 = d_{\geq 3}$ edges while the total number of G's beyond the minimally required G's (i.e. two for cycles and one for edges) is $2c'$. Thus in the worst case there are at least $2c' - d_{\geq 3}$ excess G-edges, confirming (C1).

The total number of G_{12}'s is $2p$, while the total number of G_i's is $2|E_\kappa^2| + d_{\geq 3} - 2p$, according to (A5). For fixed c the number of cycles with G_{12}'s is minimised in the case when all G_i's are in cycles of length 2 which results in $|E_\kappa^2| - p + \lfloor d_{\geq 3}/2 \rfloor$ cycles without G_{12}'s. Thus, there are at least

$$c - \left(|E_\kappa^2| - p + \lfloor d_{\geq 3}/2 \rfloor \right) = p - c' - \lfloor d_{\geq 3}/2 \rfloor \geq p - c' - d_{\geq 3}/2$$

cycles with some G_{12}, confirming also (C2).

The claim (5.36b) follows from (C1)–(C2) in combination with Lemma 5.7. $\qquad\square$

6. CENTRAL LIMIT THEOREM FOR RESOLVENTS

The goal of this section is to prove the CLT for resolvents, as stated in Proposition 3.3. We begin by analysing the 2-body stability operator $\widehat{\mathcal{B}}$ from (5.8), as well as its special case, the 1-body stability operator

$$\mathcal{B} := \widehat{\mathcal{B}}(z, z, w, w) = 1 - M\mathcal{S}[\cdot]M. \tag{6.1}$$

Note that other than in the previous Section 5, all spectral parameters $\eta, \eta_1, \ldots, \eta_p$ considered in the present section are positive, or even, $\eta, \eta_i \geq 1/n$.

Lemma 6.1. *For $w_1 = i\eta_1, w_2 = i\eta_2 \in i\mathbf{R} \setminus \{0\}$ and $z_1, z_2 \in \mathbf{C}$ we have*

$$\|\widehat{\mathcal{B}}^{-1}\|^{-1} \gtrsim (|\eta_1| + |\eta_2|) \min\{(\Im m_1)^2, (\Im m_2)^2\} + |z_1 - z_2|^2. \tag{6.2}$$

Moreover, for $z_1 = z_2 = z$ and $w_1 = w_2 = i\eta$ the operator $\mathcal{B} = \widehat{\mathcal{B}}$ has two non-trivial eigenvalues β, β_* with β, β_* as in (3.4), (3.6), and the remaining eigenvalues being 1.

Proof. Throughout the proof we assume that $\eta_1, \eta_2 > 0$, all the other cases are completely analogous. With the shorthand notations $m_i := m^{z_i}(w_i), u_i := u^{z_i}(w_i)$ and the partial trace $\mathrm{Tr}_2 \colon \mathbf{C}^{2n \times 2n} \to \mathbf{C}^4$ rearranged into a 4-dimensional vector, the stability operator $\widehat{\mathcal{B}}$, written as a 4×4 matrix is given by

$$\widehat{\mathcal{B}} = 1 - \mathrm{Tr}_2^{-1} \circ \begin{pmatrix} T_1 & 0 \\ T_2 & 0 \end{pmatrix} \circ \mathrm{Tr}_2, \quad \mathrm{Tr}_2 \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} := \begin{pmatrix} \langle R_{11} \rangle \\ \langle R_{22} \rangle \\ \langle R_{12} \rangle \\ \langle R_{21} \rangle \end{pmatrix}. \tag{6.3}$$

Here we defined

$$T_1 := \begin{pmatrix} z_1 \overline{z_2} u_1 u_2 & m_1 m_2 \\ m_1 m_2 & \overline{z_1} z_2 u_1 u_2 \end{pmatrix}, \quad T_2 := \begin{pmatrix} -z_1 u_1 m_2 & -z_2 u_2 m_1 \\ -\overline{z_2} u_2 m_1 & -\overline{z_1} u_1 m_2 \end{pmatrix},$$

and Tr_2^{-1} is understood to map $\mathbf{C}^4$ into $\mathbf{C}^{2n \times 2n}$ in such a way that each $n \times n$ block is a constant multiple of the identity matrix. From (6.3) it follows that $\widehat{\mathcal{B}}$ has eigenvalue 1 in the $4(n^2-1)$-dimensional kernel of Tr_2, and that the remaining four eigenvalues are 1, 1 and the eigenvalues $\widehat{\beta}, \widehat{\beta}_*$ of $B_1 := 1 - T_1$, i.e.

$$\widehat{\beta}, \widehat{\beta}_* := 1 - u_1 u_2 \Re z_1 \overline{z_2} \pm \sqrt{m_1^2 m_2^2 - u_1^2 u_2^2 (\Im z_1 \overline{z_2})^2}. \tag{6.4}$$

Thus the claim about the $w_1 = w_2, z_1 = z_2$ special case follows. The bound (6.2) follows directly from

$$\left| \widehat{\beta} \widehat{\beta}_* \right| \gtrsim (\eta_1 + \eta_2) \min\{(\Im m_1)^2, (\Im m_2)^2\} + |z_1 - z_2|^2, \tag{6.5}$$

since $|\widehat{\beta}|, |\widehat{\beta}_*| \lesssim 1$ and $\|\widehat{\mathcal{B}}^{-1}\| \lesssim \|B_1^{-1}\| = (\min\{|\widehat{\beta}|, |\widehat{\beta}_*|\})^{-1}$ due to B_1 being normal.

We now prove (6.5). By (6.4), using that $u_i = -m_i^2 + u_i^2 |z_i|^2$ repeatedly, it follows that

$$\begin{aligned} \widehat{\beta} \widehat{\beta}_* &= 1 - u_1 u_2 \left[1 - |z_1 - z_2|^2 + (1 - u_1)|z_1|^2 + (1 - u_2)|z_2|^2 \right] \\ &= u_1 u_2 |z_1 - z_2|^2 + (1 - u_1)(1 - u_2) - m_1^2 u_2 \left(\frac{1}{u_1} - 1 \right) \\ &\quad - m_2^2 u_1 \left(\frac{1}{u_2} - 1 \right). \end{aligned} \tag{6.6}$$

Then, using $1 - u_i = \eta_i/(\eta_i + \Im m_i) \gtrsim \eta_i/(\Im m_i)$, that $m_i = \mathrm{i}\Im m_i$, and assuming $u_1, u_2 \in [\delta, 1]$, for some small fixed $\delta > 0$, we get that

$$
\begin{aligned}
\left|\widehat{\beta}\widehat{\beta}_*\right| &\gtrsim |z_1 - z_2|^2 + (\Im m_1)^2(1 - u_1) + (\Im m_2)^2(1 - u_2) \\
&\gtrsim |z_1 - z_2|^2 + \min\{(\Im m_1)^2, (\Im m_2)^2\}(2 - u_1 - u_2) \\
&\gtrsim |z_1 - z_2|^2 + \min\{(\Im m_1)^2, (\Im m_2)^2\}\left(\frac{\eta_1}{\Im m_1} + \frac{\eta_2}{\Im m_2}\right).
\end{aligned}
\tag{6.7}
$$

If instead at least one $u_i \in [0, \delta]$ then, by the second equality in the display above, the bound (6.5) is trivial. $\square$

We now turn to the computation of the expectation $\mathbf{E}\langle G^z(\mathrm{i}\eta)\rangle$ to higher precision beyond the approximation $\langle G\rangle \approx \langle M\rangle$. Recall the definition of the 1-body stability operator from (6.1) with non-trivial eigenvalues β, β_* as in (3.4), (3.6).

Lemma 6.2. *For $\kappa_4 \neq 0$ we have a correction of order n^{-1} to $\mathbf{E}\langle G\rangle$ of the form*

$$
\mathbf{E}\langle G\rangle = \langle M\rangle + \mathcal{E} + \mathcal{O}\left(\frac{1}{|\beta|}\left(\frac{1}{n^{3/2}(1+\eta)} + \frac{1}{(n\eta)^2}\right)\right),
\tag{6.8a}
$$

where

$$
\frac{1}{|\beta|} = \|(\mathcal{B}^*)^{-1}[1]\| \lesssim \frac{1}{|1 - |z|^2| + \eta^{2/3}}
\tag{6.8b}
$$

and

$$
\mathcal{E} := \frac{\kappa_4}{n}m^3\left(\frac{1}{1 - m^2 - |z|^2} - 1\right) = -\frac{\mathrm{i}\kappa_4}{4n}\partial_\eta(m^4).
\tag{6.8c}
$$

Proof. Using (5.2) we find

$$
\begin{aligned}
\langle G - M\rangle &= \langle 1, \mathcal{B}^{-1}\mathcal{B}[G - M]\rangle = \langle(\mathcal{B}^*)^{-1}[1], \mathcal{B}[G - M]\rangle \\
&= -\langle M^*(\mathcal{B}^*)^{-1}[1], \underline{WG}\rangle + \langle M^*(\mathcal{B}^*)^{-1}[1], \mathcal{S}[G - M](G - M)\rangle \\
&= -\langle M^*(\mathcal{B}^*)^{-1}[1], \underline{WG}\rangle + \mathcal{O}_\prec\left(\frac{\|(\mathcal{B}^*)^{-1}[1]\|}{(n\eta)^2}\right).
\end{aligned}
\tag{6.9}
$$

With

$$
A := \left((\mathcal{B}^*)^{-1}[1]\right)^* M
$$

we find from the explicit formula for $\mathcal{B}$ given in (6.3) and (3.6) that

$$
\langle MA\rangle = \frac{1 - \beta}{\beta} = \frac{1}{1 - m^2 - |z|^2u^2} - 1 = -\mathrm{i}\partial_\eta m,
\tag{6.10}
$$

and, using a cumulant expansion we find

$$
\mathbf{E}\langle \underline{W}GA\rangle = \sum_{k\geq 2}\sum_{ab}\sum_{\boldsymbol{\alpha}\in\{ab,ba\}^k}\frac{\kappa(ba, \boldsymbol{\alpha})}{k!}\,\mathbf{E}\,\partial_{\boldsymbol{\alpha}}\langle\Delta^{ba}GA\rangle.
\tag{6.11}
$$

We first consider $k = 2$ where by parity at least one G factor is off-diagonal, e.g.

$$
\frac{1}{n^{5/2}}\sum_{a\leq n}\sum_{b>n}\mathbf{E}\,G_{ab}G_{aa}(GA)_{bb}
$$

and similarly for $a > n, b \leq n$. By writing $G = M + G - M$ and using the isotropic structure of the local law (3.9) we obtain

$$
\begin{aligned}
&\frac{1}{n^{5/2}}\sum_{a\leq n}\sum_{b>n}\mathbf{E}\,G_{ab}G_{aa}(GA)_{bb} \\
&= \frac{1}{n^{5/2}}\,\mathbf{E}\,m(MA)_{n+1,n+1}\langle E_1\mathbf{1}, GE_2\mathbf{1}\rangle + \mathcal{O}_\prec\left(n^2 n^{-5/2}(n\eta)^{-3/2}|\beta|^{-1}\right) \\
&= \mathcal{O}_\prec\left(\frac{1}{|\beta|n^{3/2}(1+\eta)} + \frac{1}{|\beta|n^2\eta^{3/2}}\right),
\end{aligned}
$$

where $\mathbf{1} = (1, \ldots, 1)$ denotes the constant vector of norm $\|\mathbf{1}\| = \sqrt{2n}$. Thus we can bound all $k = 2$ terms by $|\beta|^{-1}\left(n^{-3/2}(1+\eta)^{-1} + n^{-2}\eta^{-3/2}\right)$.

For $k \geq 4$ we can afford bounding each G entrywise and obtain bounds of $|\beta|^{-1}n^{-3/2}$. Finally, for the $k = 3$ term there is an assignment $(\boldsymbol{\alpha}) = (ab, ba, ab)$ for which all G's are diagonal and which contributes a leading order term given by

$$-\frac{\kappa_4}{2n^3}\sum_{ab}{}' M_{aa}M_{bb}M_{aa}(MA)_{bb} = -\frac{\kappa_4}{n}\langle M\rangle^3\langle MA\rangle, \tag{6.12}$$

where

$$\sum_{ab}{}' := \sum_{a\leq n}\sum_{b>n} + \sum_{a>n}\sum_{b\leq n},$$

and thus

$$\sum_{k\geq 2}\sum_{ab}\sum_{\boldsymbol{\alpha}\in\{ab,ba\}^k}\frac{\kappa(ba,\boldsymbol{\alpha})}{k!}\partial_{\boldsymbol{\alpha}}\langle\Delta^{ba}GA\rangle = -\frac{\kappa_4}{n}\langle M\rangle^3\langle MA\rangle$$
$$+ \mathcal{O}\Big(\frac{1}{|\beta|n^{3/2}(1+\eta)} + \frac{1}{|\beta|n^2\eta^{3/2}}\Big), \tag{6.13}$$

concluding the proof. $\qquad\square$

We now turn to the computation of higher moments which to leading order due to Lemma 6.2 is equivalent to computing

$$\mathbf{E}\prod_{i\in[p]}\langle G_i - M_i - \mathcal{E}_i\rangle, \quad \mathcal{E}_i := \frac{\kappa_4}{n}\langle M_i\rangle^3\langle M_iA_i\rangle, \quad A_i := \left((\mathcal{B}_i^*)^{-1}[1]\right)^* M_i,$$

with G_i, M_i as in (5.1) for $z_1, \ldots, z_k \in \mathbf{C}$, $\eta_1, \ldots, \eta_k > 1/n$. Using Lemma 6.2, Eq. (6.9), $|\mathcal{E}_i| \lesssim 1/n$ and the high-probability bound

$$|\langle \underline{WG_iA_i}\rangle| \prec \frac{1}{|\beta_i|n\eta_i} \tag{6.14}$$

we have

$$\prod_{i\in[p]}\langle G_i - \mathbf{E}\,G_i\rangle = \prod_{i\in[p]}\langle -\underline{WG_iA_i} - \mathcal{E}_i\rangle + \mathcal{O}_{\prec}\Big(\frac{\psi}{n\eta}\Big), \quad \psi := \prod_{i\in[p]}\frac{1}{|\beta_i|n|\eta_i|}. \tag{6.15}$$

In order to prove Proposition 3.3 we need to compute the leading order term in the local law bound

$$\left|\prod_{i\in[p]}\langle -\underline{WG_iA_i} - \mathcal{E}_i\rangle\right| \prec \psi. \tag{6.16}$$

Proof of Proposition 3.3. To simplify notations we will not carry the β_i-dependence within the proof because each A_i is of size $\|A_i\| \lesssim |\beta_i|^{-1}$ and the whole estimate is linear in each $|\beta_i|^{-1}$. We first perform a cumulant expansion in $\underline{WG_1}$ to compute

$$\mathbf{E}\prod_{i\in[p]}\langle -\underline{WG_iA_i} - \mathcal{E}_i\rangle$$
$$= -\langle\mathcal{E}_1\rangle\,\mathbf{E}\prod_{i\neq 1}\langle -\underline{WG_iA_i} - \mathcal{E}_i\rangle$$
$$+ \sum_{i\neq 1}\mathbf{E}\,\widetilde{\mathbf{E}}\langle -\widetilde{W}G_1A_1\rangle\langle -\widetilde{W}G_iA_i + \underline{WG_i\widetilde{W}G_iA_i}\rangle\prod_{j\neq 1,i}\langle -\underline{WG_jA_j} - \mathcal{E}_j\rangle \tag{6.17}$$
$$+ \sum_{k\geq 2}\sum_{ab}\sum_{\boldsymbol{\alpha}\in\{ab,ba\}^k}\frac{\kappa(ba,\boldsymbol{\alpha})}{k!}\,\mathbf{E}\,\partial_{\boldsymbol{\alpha}}\Big[\langle -\Delta^{ba}G_1A_1\rangle\prod_{i\neq 1}\langle -\underline{WG_iA_i} - \mathcal{E}_i\rangle\Big],$$

where $\widetilde{W}$ denotes an independent copy of W with expectation $\widetilde{\mathbf{E}}$, and the underline is understood with respect to W and not $\widetilde{W}$. We now consider the terms of (6.17) one by one. For the second term on the rhs. we use the identity

$$\mathbf{E}\langle WA\rangle\langle WB\rangle = \frac{1}{2n^2}\langle AE_1BE_2 + AE_2BE_1\rangle = \frac{\langle AEBE'\rangle}{2n^2}, \tag{6.18}$$

where we recall the block matrix definition from (5.13) and follow the convention that E, E' are summed over both choices $(E, E') = (E_1, E_2), (E_2, E_1)$. Thus we obtain

$$\begin{aligned}
&\widetilde{\mathbf{E}}\langle -\widetilde{W}G_1A_1\rangle\langle -\widetilde{W}G_iA_i + \underline{WG_i\widetilde{W}G_iA_i}\rangle \\
&= \frac{1}{2n^2}\langle G_1A_1EG_iA_iE' - G_1A_1E\underline{G_iA_iWG_i}E'\rangle \\
&= \frac{1}{2n^2}\langle G_1A_1EG_iA_iE' + G_1\mathcal{S}[G_1A_1EG_iA_i]G_iE' - \underline{G_1A_1EG_iA_iWG_i}E'\rangle.
\end{aligned} \tag{6.19}$$

Here the self-renormalisation in the last term is defined analogously to (5.4), i.e.

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

which is only well-defined if it is clear to which W the action is associated, i.e. $\underline{WWf(W)}$ would be ambiguous. However, we only use the self-renormalisation notation for $f(W), \underline{g(W)}$ being (products of) resolvents and deterministic matrices, so no ambiguities should arise. For the first two terms in (6.19) we use $\|M_{AE_1}^{z_1,z_i}\| \lesssim \|\widehat{\mathcal{B}}_{1i}^{-1}\| \lesssim |z_1 - z_i|^{-2}$ due to (6.2) and the first bound in (5.9) from Theorem 5.2 (estimating the big bracket by 1) to obtain

$$\begin{aligned}
&\langle G_1A_1EG_iA_iE' + G_1\mathcal{S}[G_1A_1EG_iA_i]G_iE'\rangle \\
&= \langle M_{A_1E}^{z_1,z_i}A_iE' + M_{E'}^{z_i,z_1}\mathcal{S}[M_{A_1E}^{z_1,z_i}A_i]\rangle \\
&\quad + \mathcal{O}_{\prec}\left(\frac{1}{n|z_1 - z_i|^4\eta_*^{1i}|\eta_1\eta_i|^{1/2}} + \frac{1}{n^2|z_1 - z_i|^4(\eta_*^{1i})^2|\eta_1\eta_i|}\right),
\end{aligned} \tag{6.20}$$

where $\eta_*^{1i} := \min\{\eta_1, \eta_i\}$. For the last term in (6.19) we claim that

$$\mathbf{E}|\langle \underline{G_1A_1EG_iA_iWG_i}E'\rangle|^2 \lesssim \left(\frac{1}{n\eta_1\eta_i\eta_*^{1i}}\right)^2, \tag{6.21}$$

the proof of which we present after concluding the proof of the proposition. Thus, using (6.21) together with (6.14),

$$\begin{aligned}
&\left|n^{-2}\,\mathbf{E}\langle \underline{G_1A_1EG_iA_iWG_i}E'\rangle \prod_{j\neq 1,i}\langle -\underline{WG_jA_j} - \mathcal{E}_i\rangle\right| \\
&\lesssim \frac{n^\epsilon}{n^2}\left[\prod_{j\neq 1,i}\frac{1}{n\eta_j}\right]\left(\mathbf{E}|\langle \underline{G_1A_1EG_iA_iWG_i}E'\rangle|^2\right)^{1/2} \\
&\lesssim \frac{n^\epsilon}{n\eta_*^{1i}}\prod_j\frac{1}{n\eta_j} \leq \frac{n^\epsilon\psi}{n\eta_*}.
\end{aligned}$$

Together with (6.16) and (6.19)–(6.20) we obtain

$$\begin{aligned}
&\mathbf{E}\,\widetilde{\mathbf{E}}\langle -\widetilde{W}G_1A_1\rangle\langle -\widetilde{W}G_iA_i + \underline{WG_i\widetilde{W}G_iA_i}\rangle \prod_{j\neq 1,i}\langle -\underline{WG_jA_j} - \mathcal{E}_j\rangle \\
&= \frac{V_{1,i}}{2n^2}\,\mathbf{E}\prod_{j\neq 1,i}\langle -\underline{WG_jA_j} - \mathcal{E}_j\rangle \\
&\quad + \mathcal{O}\left(\psi n^\epsilon\left(\frac{1}{n\eta_*} + \frac{|\eta_1\eta_i|^{1/2}}{n\eta_*^{1i}|z_1 - z_i|^4} + \frac{1}{(n\eta_*^{1i})^2|z_1 - z_i|^4}\right)\right)
\end{aligned} \tag{6.22}$$

since, by an explicit computation the rhs. of (6.20) is given by $V_{1,i}$ as defined in (3.14). Indeed, from the explicit formula for $\mathcal{B}$ it follows that main term on the rhs. of (6.20) can be written as $\widetilde{V}_{1,i}$, where

$$
\begin{aligned}
\widetilde{V}_{i,j} := {} & \frac{2m_i m_j \left[2u_i u_j \Re z_i \overline{z_j} + (u_i u_j |z_i||z_j|)^2 \left[s_i s_j - 4 \right] \right]}{t_i t_j \left[1 + (u_i u_j |z_i||z_j|)^2 - m_i^2 m_j^2 - 2u_i u_j \Re z_i \overline{z_j} \right]^2} \\
& + \frac{2m_i m_j (m_i^2 + u_i^2 |z_i|^2)(m_j^2 + u_j^2 |z_j|^2)}{t_i t_j \left[1 + (u_i u_j |z_i||z_j|)^2 - m_i^2 m_j^2 - 2u_i u_j \Re z_i \overline{z_j} \right]^2},
\end{aligned}
\tag{6.23}
$$

using the notations $t_i := 1 - m_i^2 - u_i^2 |z_1|^2$, $s_i := m_i^2 - u_i^2 |z_i|^2$. By an explicit computation using the equation (3.1) for m_i, m_j it can be checked that $\widetilde{V}_{i,j}$ can be written as a derivative and is given by $\widetilde{V}_{i,j} = V_{i,j}$ with $V_{i,j}$ from (3.14).

Next, we consider the third term on the rhs. of (6.17) for $k = 2$ and $k \geq 3$ separately. We first claim the auxiliary bound

$$
|\langle \boldsymbol{x}, \underline{GBWG}\boldsymbol{y} \rangle| \prec \frac{\|\boldsymbol{x}\|\|\boldsymbol{y}\|\|B\|}{n^{1/2} \eta^{3/2}}.
\tag{6.24}
$$

Note that (6.24) is very similar to (5.10a) except that in (6.24) both G's have the same spectral parameters z, η and the order of W and G is interchanged. The proof of (6.24) is, however, very similar and we leave details to the reader.

After performing the $\boldsymbol{\alpha}$-derivative in (6.17) via the Leibniz rule, we obtain a product of $t \geq 1$ traces of the types $\langle (\Delta G_i)^{k_i} A_i \rangle$ and $\langle \underline{W(G_i \Delta)^{k_i} G_i} A_i \rangle$ with $k_i \geq 0$, $\sum k_i = k + 1$, and $p - t$ traces of the type $\langle \underline{WG_i A_i} + \mathcal{E}_i \rangle$. For the term with multiple self-renormalised G's, i.e. $\langle \underline{W(G_i\Delta)^{k_i}G_i} A_i \rangle$ with $k_i \geq 1$ we rewrite

$$
\begin{aligned}
\langle \underline{W(G\Delta)^k G} A \rangle &= \langle \underline{GAW(G\Delta)^k} \rangle \\
&= \langle \underline{GAWG}\Delta(G\Delta)^{k-1} \rangle + \sum_{j=1}^{k-1} \langle \underline{GAS}[(G\Delta)^j G](G\Delta)^{k-j} \rangle \\
&= \langle \underline{GAWG}\Delta(G\Delta)^{k-1} \rangle + \sum_{j=1}^{k-1} \langle \underline{GAE(G\Delta)^{k-j}} \rangle \langle \underline{GE'(G\Delta)^j} \rangle.
\end{aligned}
\tag{6.25}
$$

Case $k = 2$, $t = 1$. In this case the only possible term is given by $\langle \Delta G_1 \Delta G_1 \Delta G_1 A_1 \rangle$ where by parity at least one $G = G_1$ is off-diagonal and in the worst case (only one off-diagonal factor) we estimate

$$
\begin{aligned}
n^{-1-3/2} \sum_{a \leq n} \sum_{b > n} G_{aa} G_{bb} (GA)_{ab} &= \frac{m^2}{n^{5/2}} \langle E_1 \mathbf{1}, \underline{GAE_2}\mathbf{1} \rangle + \mathcal{O}_\prec \left(\frac{1}{n^{1/2}} \frac{1}{(n\eta_1)^{3/2}} \right) \\
&= \mathcal{O}_\prec \left(\frac{1}{n^{3/2}} + \frac{1}{n^2 \eta_1^{3/2}} \right),
\end{aligned}
$$

after replacing $G_{aa} = m + (G - M)_{aa}$ and using the isotropic structure of the local law in (3.9), and similarly for $\sum_{a > n} \sum_{b \leq n}$.

Case $k = 2$, $t = 2$. In this case there are $2 + 2$ possible terms

$$
\begin{aligned}
\langle \Delta G_1 \Delta G_1 A_1 \rangle \langle \Delta G_i A_i + \underline{WG_i \Delta G_i A_i} \rangle \\
+ \langle \Delta G_1 A_1 \rangle \langle \Delta G_i \Delta G_i A_i + \underline{WG_i \Delta G_i \Delta G_i A_i} \rangle.
\end{aligned}
$$

For the first two, in the worst case, we have the estimate

$$
\begin{aligned}
\frac{1}{n^{7/2}} {\sum_{ab}}' (G_1)_{aa} (G_1 A_1)_{bb} \Big((G_i A_i)_{ab} + (\underline{G_i A_i W G_i})_{ab} \Big) \\
= \mathcal{O}_\prec \left(\frac{1}{n^{5/2}} + \frac{1}{n^3 \eta_1 \eta_i^{3/2}} \right)
\end{aligned}
$$

using (6.24), where we recall the definition of $\sum'$ from (6.12). Similarly, using (6.25) and (6.24) for the ultimate two terms, we have the bound

$$\frac{1}{n^{7/2}} \mathbf{E} \sum_{ab}' (G_1 A_1)_{ab} \Big(\underline{(G_i A_i W G_i)_{aa}} (G_i)_{bb} + \frac{(G_i A_i E G_i)_{ab}(G_i E' G_i)_{ab}}{n} \Big)$$
$$= \mathcal{O}_\prec \Big(\frac{1}{n^3 \eta_1^{1/2} \eta_i^2} \Big).$$

Case $k = 2$, $t = 3$. In this final $k = 2$ case we have to consider four terms

$$\langle \Delta G_1 A_1 \rangle \langle \Delta G_i A_i + \underline{W G_i \Delta G_i A_i} \rangle \langle \Delta G_j A_j + \underline{W G_j \Delta G_j A_j} \rangle,$$

which, using (6.24), we estimate by

$$\frac{1}{n^{9/2}} \sum_{ab}' (G_1 A_1)_{ab} \Big((G_i A_i)_{ab} + \underline{(G_i A_i W G_i)_{ab}} \Big) \Big((G_j A_j)_{ab} + \underline{(G_j A_j W G_j)_{ab}} \Big)$$
$$= \mathcal{O}_\prec \Big(\frac{1}{n^4 \eta_1^{1/2} \eta_i^{3/2} \eta_j^{3/2}} \Big).$$

By inserting the above estimates back into (6.17), after estimating all untouched traces by $n^\epsilon/(n\eta_i)$ in high probability using (6.14), we obtain

$$\sum_{k=2} \sum_{ab} \sum_{\boldsymbol{\alpha} \in \{ab, ba\}^k} \frac{\kappa(ba, \boldsymbol{\alpha})}{k!} \mathbf{E} \, \partial_{\boldsymbol{\alpha}} \Big[\langle -\Delta^{ba} G_1 A_1 \rangle \prod_{i \neq 1} \langle -\underline{W G_i A_i} - \mathcal{E}_i \rangle \Big]$$
$$= \mathcal{O}\Big(\frac{\psi n^\epsilon}{\sqrt{n\eta_*}} \Big). \tag{6.26}$$

Case $k \geq 3$. In case $k \geq 3$ after the action of the derivative in (6.17) there are $1 \leq t \leq k+1$ traces involving some Δ. By writing the normalised traces involving Δ as matrix entries we obtain a prefactor of $n^{-t-(k+1)/2}$ and a $\sum_{ab}$-summation over entries of $k+1$ matrices of the type G, GA, $\underline{GAWG}$ such that each summation index appears exactly $k+1$ times. There are some additional terms from the last sum in (6.25) which are smaller by a factor $(n\eta)^{-1}$ and which can be bounded exactly as in the $k = 2$ case. If there are only diagonal G or GA-terms, then we have a naive bound of $n^{-t-(k-3)/2}$ and therefore potentially some leading-order contribution in case $k = 3$. If, however, $k > 3$, or there are some off-diagonal G, GA or some $\underline{GAWG}$ terms, then, using (6.24) we obtain an improvement of at least $(n\eta)^{-1/2}$ over the naive bound (6.16). For $k = 3$, by parity, the only possibility of having four diagonal G, GA factors, is distributing the four Δ's either into a single trace or two traces with two Δ's each. Thus the relevant terms are

$$\langle \Delta G_1 \Delta G_1 \Delta G_1 \Delta G_1 A_1 \rangle, \quad \langle \Delta G_1 \Delta G_1 A_1 \rangle \langle \Delta G_i \Delta G_i A_i \rangle.$$

For the first one we recall from (6.13) for $k = 3$ that

$$\sum_{ab} \sum_{\boldsymbol{\alpha}} \kappa(ba, \boldsymbol{\alpha}) \langle \Delta^{ba} G_1 \Delta^{\alpha_1} G_1 \Delta^{\alpha_2} G_1 \Delta^{\alpha_3} G_1 A_1 \rangle = \mathcal{E}_1 + \mathcal{O}_\prec \Big(\frac{1}{n^{3/2}} + \frac{1}{n^2 \eta_1^{3/2}} \Big). \tag{6.27}$$

For the second one we note that only choosing $\boldsymbol{\alpha} = (ab, ab, ba), (ab, ba, ab)$ gives four diagonal factors, while any other choice gives at least two off-diagonal factors. Thus

$$
\begin{aligned}
\sum_{ab} \sum_{\boldsymbol{\alpha}} & \kappa(ba, \boldsymbol{\alpha}) \langle \Delta^{ba} G_1 \Delta^{\alpha_1} G_1 \rangle \langle \Delta^{\alpha_2} G_i \Delta^{\alpha_3} G_i A_i \rangle \\
&= \frac{\kappa_4}{n^2} {\sum_{ab}}' \langle \Delta^{ba} G_1 \Delta^{ab} G_1 A_1 \rangle \big[\langle \Delta^{ab} G_i \Delta^{ba} G_i A_i \rangle + \langle \Delta^{ba} G_i \Delta^{ab} G_i A_i \rangle \big] + \mathcal{O}_\prec(\mathcal{E}) \\
&= \frac{\kappa_4}{4n^4} {\sum_{ab}}' (G_1)_{aa} (G_1 A_1)_{bb} \big[(G_i)_{bb} (G_i A_i)_{aa} + (G_i)_{aa} (G_i A_i)_{bb} \big] + \mathcal{O}_\prec(\mathcal{E}) \\
&= \frac{\kappa_4}{4n^4} {\sum_{ab}}' m_1 m_i (M_1 A_1)_{bb} \big[(M_i A_i)_{aa} + (M_i A_i)_{bb} \big] + \mathcal{O}_\prec \left(\sqrt{n \eta_*} \mathcal{E} \right) \\
&= \frac{\kappa_4}{n^2} \langle M_1 \rangle \langle M_i \rangle \langle M_1 A_1 \rangle \langle M_i A_i \rangle + \mathcal{O}_\prec \left(\frac{1}{n^{5/2} \eta_*^{1/2}} \right),
\end{aligned}
\tag{6.28}
$$

where $\mathcal{E} := (n^3 \eta_*)^{-1}$. We recall from (6.10) that

$$
\langle M_1 \rangle \langle M_i \rangle \langle M_1 A_1 \rangle \langle M_i A_i \rangle = \frac{1}{2} U_1 U_i
$$

with U_i defined in (3.14). Thus, we can conclude for the $k \geq 3$ terms in (6.17) that

$$
\begin{aligned}
\sum_{k \geq 3} \sum_{ab} \sum_{\boldsymbol{\alpha} \in \{ab, ba\}^k} & \frac{\kappa(ba, \boldsymbol{\alpha})}{k!} \, \mathbf{E} \, \partial_{\boldsymbol{\alpha}} \Big[\langle -\Delta^{ba} G_1 A_1 \rangle \prod_{i \neq 1} \langle -\underline{W G_i A_i} - \mathcal{E}_i \rangle \Big] \\
&= \langle \mathcal{E}_1 \rangle \, \mathbf{E} \prod_{i \neq 1} \langle -\underline{W G_i A_i} - \mathcal{E}_i \rangle + \sum_{i \neq 1} \frac{\kappa_4 U_1 U_i}{2n^2} \, \mathbf{E} \prod_{j \neq 1, i} \langle -\underline{W G_j A_j} - \mathcal{E}_j \rangle \\
&\quad + \mathcal{O} \Big(\frac{\psi n^\epsilon}{(n \eta_*)^{1/2}} \Big).
\end{aligned}
\tag{6.29}
$$

By combining (6.17) with (6.22), (6.26) and (6.29) we obtain

$$
\begin{aligned}
\mathbf{E} \prod_i \langle -\underline{W G_i A_i} - \mathcal{E}_i \rangle &= \sum_{i \neq 1} \frac{V_{1,i} + \kappa_4 U_1 U_i}{2n^2} \, \mathbf{E} \prod_{j \neq 1, i} \langle -\underline{W G_j A_j} - \mathcal{E}_j \rangle \\
&\quad + \mathcal{O} \left(\frac{\psi n^\epsilon}{\sqrt{n \eta_*}} + \frac{\psi n^\epsilon}{n \eta_*^{1/2} |z_1 - z_i|^4} + \frac{\psi n^\epsilon}{(n \eta_*)^2 |z_1 - z_i|^4} \right),
\end{aligned}
\tag{6.30}
$$

and thus by induction

$$
\begin{aligned}
\mathbf{E} \prod_i \langle -\underline{W G_i A_i} - \mathcal{E}_i \rangle &= \frac{1}{n^p} \sum_{P \in \Pi_p} \prod_{\{i, j\} \in P} \frac{V_{i,j} + \kappa_4 U_i U_j}{2} \\
&\quad + \mathcal{O} \left(\frac{\psi n^\epsilon}{\sqrt{n \eta_*}} + \frac{\psi n^\epsilon}{n \eta_*^{1/2} |z_1 - z_i|^4} + \frac{\psi n^\epsilon}{(n \eta_*)^2 |z_1 - z_i|^4} \right),
\end{aligned}
\tag{6.31}
$$

from which the equality $\mathbf{E} \prod_i \langle G_i - \mathbf{E} \, G_i \rangle$ and the second line of (3.12) follows, modulo the proof of (6.21). The remaining equality then follows from applying the very same equality for each element of the pairing. Finally, (3.15) follows directly from Lemma 6.2. $\qquad \square$

Proof of (6.21). Using the notation of Lemma 5.8, our goal is to prove that

$$
\mathbf{E} |\langle \underline{W G_{i1i}} \rangle|^2 \lesssim \left(\frac{1}{n \eta_1 \eta_i \eta_*^{1i}} \right)^2.
\tag{6.32}
$$

Since only η_1, η_i play a role within the proof of (6.21), we drop the indices from η_*^{1i} and simply write $\eta_* = \eta_*^{1i}$. Using a cumulant expansion we compute

$$
\begin{aligned}
&\mathbf{E}|\underline{\langle WG_{i1i}\rangle}|^2 \\
&= \mathbf{E}\,\widetilde{\mathbf{E}}\langle\widetilde{W}G_{i1i}\rangle\Big(\langle\widetilde{W}G_{i1i}\rangle + \underline{\langle WG_i\widetilde{W}G_{i1i}} + \underline{WG_{i1}\widetilde{W}G_{1i}} + \underline{WG_{i1i}\widetilde{W}G_i}\rangle\Big) \\
&\quad + \sum_{k\geq 2}\mathcal{O}\Big(\frac{1}{n^{(k+1)/2}}\Big)\sideset{}{'}\sum_{ab}\sum_{k_1+k_2=k-1}\sum_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}\mathbf{E}\langle\Delta^{ab}\partial_{\boldsymbol{\alpha}_1}G_{i1i}\rangle\langle\Delta^{ab}\partial_{\boldsymbol{\alpha}_2}G_{i1i}\rangle \\
&\quad + \sum_{k\geq 2}\mathcal{O}\Big(\frac{1}{n^{(k+1)/2}}\Big)\sideset{}{'}\sum_{ab}\sum_{k_1+k_2=k}\sum_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}\mathbf{E}\langle\Delta^{ab}\partial_{\boldsymbol{\alpha}_1}G_{i1i}\rangle\underline{\langle W\partial_{\boldsymbol{\alpha}_2}G_{i1i}\rangle},
\end{aligned}
\tag{6.33}
$$

where $\boldsymbol{\alpha}_i$ is understood to be summed over $\boldsymbol{\alpha}_i \in \{ab, ba\}^{k_i}$. In (6.33) we only kept the scaling $|\kappa(ab,\boldsymbol{\alpha})| \lesssim n^{-(k+1)/2}$ of the cumulants, and also absorb combinatorial factors as $k!$ in $\mathcal{O}(\cdot)$. We first consider those terms in (6.33) which contain no self-renormalisations $\underline{Wf(W)}$ anymore since those do not have to be expanded further. For the very first term we obtain

$$
\widetilde{\mathbf{E}}\langle\widetilde{W}G_{i1i}\rangle\langle\widetilde{W}G_{i1i}\rangle = \frac{\langle G_{i1ii1i}\rangle}{n^2} = \mathcal{O}_{\prec}\Big(\frac{1}{n^2\eta_1^2\eta_i^3}\Big).
\tag{6.34}
$$

To bound products of G_1 and G_i we use Lemma 5.8. For the second line on the rhs. of (6.33) we have to estimate

$$
\mathcal{O}\Big(\frac{1}{n^{(k+1)/2+2}}\Big)\sum_{k\geq 2}\sideset{}{'}\sum_{ab}\sum_{k_1+k_2=k-1}\sum_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}\mathbf{E}(\partial_{\boldsymbol{\alpha}_1}(G_{i1i})_{ba})(\partial_{\boldsymbol{\alpha}_2}(G_{i1i})_{ba})
$$

and we note that without derivatives we have the estimate $|(G_{i1i})| \prec (\eta_1\eta_i)^{-1}$. Additional derivatives do not affect this bound since if e.g. G_i is derived we obtain one additional G_i but also one additional product of G's with G_i in the end, and one additional product with G_i in the beginning. Due to the structure of the estimate (5.33a) the bound thus remains invariant. For example $|(\partial_{ab}G_{i1i})_{ba}| = |(G_i)_{bb}(G_{i1i})_{aa} + \ldots| \prec (\eta_1\eta_i)^{-1}$. Thus, by estimating the sum trivially we obtain

$$
\frac{1}{n^{(k+1)/2}}\sum_{\substack{k_1+k_2=k-1\\k\geq 2}}\sideset{}{'}\sum_{ab}\sum_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}\mathbf{E}\langle\Delta^{ab}\partial_{\boldsymbol{\alpha}_1}G_{i1i}\rangle\langle\Delta^{ab}\partial_{\boldsymbol{\alpha}_2}G_{i1i}\rangle = \mathcal{O}_{\prec}\Big(\frac{1}{n^{3/2}\eta_1^2\eta_i^2}\Big)
\tag{6.35}
$$

since $k \geq 2$.

It remains to consider the third line on the rhs. of (6.33) and the remaining terms from the first line. In both cases we perform a second cumulant expansion and again differentiate the Gaussian (i.e. the second order cumulant) term, and the terms from higher order cumulants. Since the two consecutive cumulant expansions commute it is clearly sufficient to consider the Gaussian term for the first line, and the full expansion for the third line. We begin with the latter and compute

$$
\begin{aligned}
&\mathbf{E}\langle\Delta^{ab}\partial_{\boldsymbol{\alpha}_1}G_{i1i}\rangle\underline{\langle W\partial_{\boldsymbol{\alpha}_2}G_{i1i}\rangle} \\
&= \widetilde{\mathbf{E}}\,\mathbf{E}\langle\Delta^{ab}\partial_{\boldsymbol{\alpha}_1}(G_i\widetilde{W}G_{i1i} + G_{i1}\widetilde{W}G_{1i} + G_{i1i}\widetilde{W}G_i)\rangle\langle\widetilde{W}\partial_{\boldsymbol{\alpha}_2}G_{i1i}\rangle \\
&\quad + \sum_{l\geq 2}\sideset{}{'}\sum_{cd}\sum_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\mathbf{E}\langle\Delta^{ab}\partial_{\boldsymbol{\alpha}_1}\partial_{\boldsymbol{\beta}_1}G_{i1i}\rangle\langle\Delta^{cd}\partial_{\boldsymbol{\alpha}_2}\partial_{\boldsymbol{\beta}_2}G_{i1i}\rangle \\
&= \frac{1}{n^2}\mathbf{E}\langle\partial_{\boldsymbol{\alpha}_1}(G_{i1i}\Delta^{ab}G_i + G_{1i}\Delta^{ab}G_{i1} + G_i\Delta^{ab}G_{i1i})\partial_{\boldsymbol{\alpha}_2}(G_{i1i})\rangle \\
&\quad + \sum_{l\geq 2}\sideset{}{'}\sum_{cd}\sum_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}\mathbf{E}\langle\Delta^{ab}\partial_{\boldsymbol{\alpha}_1}\partial_{\boldsymbol{\beta}_1}G_{i1i}\rangle\langle\Delta^{cd}\partial_{\boldsymbol{\alpha}_2}\partial_{\boldsymbol{\beta}_2}G_{i1i}\rangle,
\end{aligned}
\tag{6.36}
$$

where $\boldsymbol{\beta}_i$ are understood to be summed over $\boldsymbol{\beta}_i \in \{cd, dc\}^{l_i}$ with $l_1 + l_2 = l$. After inserting the first line of (6.36) back into (6.33) we obtain an overall factor of $n^{-3-(k+1)/2}$ as well as the $\sum_{ab}$-summation over some $\partial_{\boldsymbol{\alpha}}(\mathcal{G})_{ab}$, where $\mathcal{G}$ is a product of either $2 + 5$ or $3 + 4$ G_1's and G_i's respectively with G_i in beginning and end. We can bound $|\partial_{\boldsymbol{\alpha}}(\mathcal{G})_{ab}| \prec \eta_1^{-2}\eta_i^{-4} + \eta_1^{-3}\eta_i^{-3} \leq \eta_1^{-2}\eta_i^{-2}\eta_*^{-2}$ and thus can estimate the sum by $n^{-5/2}\eta_1^{-2}\eta_i^{-2}\eta_*^{-2}$ since $k \geq 2$. Here we used (5.33a) to estimate all matrix elements of the form $\mathcal{G}'_{ab}, \mathcal{G}'_{aa}, \ldots$ emerging after performing the derivative $\partial_{\boldsymbol{\alpha}}(\mathcal{G})_{ab}$.

Now we turn to the second line of (6.36) when inserted back into (6.33), where we obtain a total prefactor of $n^{-(k+l)/2-3}$, a summation $\sum_{abcd}$ over $(\partial_{\boldsymbol{\alpha}_1}\partial_{\boldsymbol{\beta}_1}G_{i1i})_{ab}(\partial_{\boldsymbol{\alpha}_2}\partial_{\boldsymbol{\beta}_2}G_{i1i})_{cd}$. In case $k = l = 2$, by parity, after performing the derivatives at least two factors are off-diagonal, while in case $k+l = 5$ at least one factor is off-diagonal. Thus we obtain a bound of $n^{1-(k+l)/2}\eta_1^{-2}\eta_i^{-2}$ multiplied by a Ward-improvement of $(n\eta_*)^{-1}$ in the first, and $(n\eta_*)^{-1/2}$ in the second case. Thus we conclude

$$\frac{1}{n^{(k+1)/2}}\sum_{\substack{k_1+k_2=k\\k\geq 2}}\sideset{}{'}\sum_{ab}\sum_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}\mathbf{E}\langle\Delta^{ab}\partial_{\boldsymbol{\alpha}_1}G_{i1i}\rangle\langle\underline{W\partial_{\boldsymbol{\alpha}_2}G_{i1i}}\rangle = \mathcal{O}\Big(\frac{1}{n^2\eta_1^2\eta_i^2\eta_*^2}\Big). \tag{6.37}$$

Finally, we consider the Gaussian part of the cumulant expansion of the remaining terms in the first line of (6.33), for which we obtain

$$\frac{1}{n^2}\widetilde{\mathbf{E}}\langle(G_{i1i}\widetilde{W}G_i + G_{1i}\widetilde{W}G_{i1} + G_i\widetilde{W}G_{i1i})^2\rangle = O_\prec\Big(\frac{1}{n^2\eta_1^2\eta_i^2\eta_*^2}\Big) \tag{6.38}$$

since

$$|\langle G_iG_i\rangle| \prec \frac{1}{\eta_i}, \quad |\langle G_iG_{i1}\rangle| \prec \frac{1}{\eta_1\eta_i}, \quad |\langle G_iG_{i1i}\rangle| \prec \frac{1}{\eta_1\eta_i^2},$$

$$|\langle G_{1i}G_{1i}\rangle| \prec \frac{1}{\eta_1^2\eta_i}, \quad |\langle G_{1i}G_{i1i}\rangle| \prec \frac{1}{\eta_1^2\eta_i^2}, \quad |\langle G_{i1i}G_{i1i}\rangle| \prec \frac{1}{\eta_1^2\eta_i^3}$$

due to (5.33b). By combining (6.34)–(6.38) we conclude the proof of (6.21) using (6.33). $\qquad\square$

7. Independence of the small eigenvalues of H^{z_1} and H^{z_2}

Given an $n \times n$ i.i.d. complex matrix X, for any $z \in \mathbf{C}$ we recall that the Hermitisation of $X - z$ is given by

$$H^z := \begin{pmatrix} 0 & X-z \\ X^*-\overline{z} & 0 \end{pmatrix}. \tag{7.1}$$

The block structure of H^z induces a symmetric spectrum with respect to zero, i.e. denoting by $\{\lambda_{\pm i}^z\}_{i=1}^n$ the eigenvalues of H^z, we have that $\lambda_{-i}^z = -\lambda_i^z$ for any $i \in [n]$. Denote the resolvent of H^z by G^z, i.e. on the imaginary axis G^z is defined by $G^z(\mathrm{i}\eta) := (H^z - \mathrm{i}\eta)^{-1}$, with $\eta > 0$.

Convention 7.1. *We omitted the index $i = 0$ in the definition of the eigenvalues of H^z. In the remainder of this section we always assume that all the indices are not zero, e.g we use the notation*

$$\sum_{j=-n}^n := \sum_{j=-n}^{-1} + \sum_{j=1}^n.$$

Similarly, by $|i| \leq A$, for some $A > 0$, we mean $0 < |i| \leq A$, etc.

The main result of this section is the proof of Proposition 3.5 which follows by Proposition 7.2 and rigidity estimates in Section 7.1.

Proposition 7.2. *Fix $p \in \mathbf{N}$. For any $\omega_d, \omega_f, \omega_h > 0$ sufficiently small constants such that $\omega_h \ll \omega_f$, there exits $\omega, \widehat{\omega}, \delta_0, \delta_1 > 0$ with $\omega_h \ll \delta_m \ll \widehat{\omega} \ll \omega \ll \omega_f$, for $m = 0, 1$, such that for any fixed $z_1, \ldots, z_p \in \mathbf{C}$ such that $|z_l| \leq 1 - n^{-\omega_h}, |z_l - z_m| \geq n^{-\omega_d}$, with $l, m \in [p], l \neq m$, it holds*

$$\mathbf{E}\prod_{l=1}^p\frac{1}{n}\sum_{|i_l|\leq n^{\widehat{\omega}}}\frac{\eta_l}{(\lambda_{i_l}^{z_l})^2+\eta_l^2} = \prod_{l=1}^p\mathbf{E}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat{\omega}}}\frac{\eta_l}{(\lambda_{i_l}^{z_l})^2+\eta_l^2}$$
$$+ \mathcal{O}\left(\frac{n^{\widehat{\omega}}}{n^{1+\omega}}\sum_{l=1}^p\frac{1}{\eta_l}\times\prod_{m=1}^p\left(1+\frac{n^\xi}{n\eta_m}\right) + \frac{n^{p\xi+2\delta_0}n^{\omega_f}}{n^{3/2}}\sum_{l=1}^p\frac{1}{\eta_l} + \frac{n^{p\delta_0+\delta_1}}{n^{\widehat{\omega}}}\right), \tag{7.2}$$

for any $\xi > 0$, where $\eta_1, \ldots, \eta_p \in [n^{-1-\delta_0}, n^{-1+\delta_1}]$ and the implicit constant in $\mathcal{O}(\cdot)$ may depend on p.

We recall that the eigenvalues of H^z are labelled by $\lambda_{-n} \le \cdots \le \lambda_{-1} \le \lambda_1 \le \cdots \le \lambda_n$, hence the summation over $|i_l| \le n^{\widehat{\omega}}$ in (7.2) is over the smallest (in absolute value) eigenvalues of H^z.

The remainder of Section 7 is divided as follows: in Section 7.1 we state rigidity of the eigenvalues of the matrices H^{z_l} and a local law for $\operatorname{Tr} G^{z_l}$, then using these results and Proposition 7.2 we conclude the proof of Proposition 3.5. In Section 7.2 we state the main technical results needed to prove Proposition 7.2 and conclude its proof. In Section 7.3 we estimate the overlaps of eigenvectors, corresponding to small indices, of H^{z_l}, H^{z_m} for $l \ne m$, this is the main input to prove the asymptotic independence in Proposition 7.2. In Section 7.4 we present Proposition 7.14 which is a modification of the pathwise coupling of DBMs from [59, 20] (adapted to the 2×2 matrix model (7.1) in [26]) which is needed to deal with the (small) correlation of $\boldsymbol{\lambda}^{z_l}$, the eigenvalues of H^{z_l}, for different l's. In Section 7.5 we prove some technical lemmata used in Section 7.2. Finally, in Section 7.6 we prove Proposition 7.14.

7.1. **Rigidity of eigenvalues and proof of Proposition 3.5.**

In this section, before proceeding with the actual proof of Proposition 7.2, we state the local law away from the imaginary axis, proven in [31], that will be used in the following sections. We remark that the averaged and entry-wise version of this local law for $|z| \le 1 - \epsilon$, for some small fixed $\epsilon > 0$, has already been established in [21, Theorem 3.4].

Proposition 7.3 (Theorem 3.1 of [31]). *Let $\omega_h > 0$ be sufficiently small, and define $\delta_l := 1 - |z_l|^2$. Then with very high probability it holds*

$$\left| \frac{1}{2n} \sum_{1 \le |i| \le n} \frac{1}{\lambda_i^{z_l} - w} - m^{z_l}(w) \right| \le \frac{\delta_l^{-100} n^\xi}{n \Im w}, \qquad (7.3)$$

uniformly in $|z_l|^2 \le 1 - n^{-\omega_h}$ and $0 < \Im w \le 10$. Here m^{z_l} denotes the solution of (3.1).

Note that $\delta_l := 1 - |z_l|^2$ introduced in Proposition 7.3 are not to be confused with the exponents δ_0, δ_1 introduced in Proposition 7.2.

Let $\{\lambda_{\pm i}^z\}_{i=1}^n$ denote the eigenvalues of H^z, and recall that $\rho^z(E) = \pi^{-1} \Im m^z(E + \mathrm{i}0)$ is the limiting (self-consistent) density of states. Then by Proposition 7.3 the rigidity of λ_i^z follows by a standard application of Helffer-Sjöstrand formula (see e.g. [36, Lemma 7.1, Theorem 7.6] or [43, Section 5] for a detailed derivation):

$$|\lambda_i^z - \gamma_i^z| \le \frac{\delta^{-100} n^\xi}{n}, \qquad |i| \le cn, \qquad (7.4)$$

with $c > 0$ a small constant and $\delta := 1 - |z|^2$, with very high probability, uniformly in $|z| \le 1 - n^{-\omega_h}$. The quantiles γ_i^z are defined by

$$\frac{i}{n} = \int_0^{\gamma_i^z} \rho^z(E) \, \mathrm{d}E, \qquad 1 \le i \le n, \qquad (7.5)$$

and $\gamma_{-i}^z := -\gamma_i^z$ for $-n \le i \le -1$. Note that by (7.5) it follows that $\gamma_i^z \sim i/(n\rho^z(0))$ for $|i| \le n^{1-10\omega_h}$, where $\rho^z(0) = \Im m^z(0) = (1 - |z|^2)^{1/2}$ for $|z| < 1$ by (3.7).

Using the rigidity bound in (7.4), by Proposition 7.2 we conclude the proof of Proposition 3.5.

Proof of Proposition 3.5. Let $z_1, \ldots, z_p$ such that $|z_l| \le 1 - n^{-\omega_h}$ and $|z_l - z_m| \ge n^{-\omega_d}$, for any $l, m \in [p]$, with ω_d, ω_h defined in Proposition 3.5. Let $\omega, \widehat{\omega}, \delta_0, \delta_1$ be as in Proposition 7.2, i.e.

$$\omega_h \ll \delta_m \ll \widehat{\omega} \ll \omega \ll \omega_f,$$

for $m = 0, 1$. For a detailed summary about all the different scales in the proof of Proposition 7.2 and so of Proposition 3.5 see Section 7.2.3 later. Write

$$\langle G^{z_l}(\mathrm{i}\eta_l) \rangle = \frac{\mathrm{i}}{2n} \left[\sum_{|i| \le \widehat{\omega}} + \sum_{\widehat{\omega} < |i| \le n} \right] \frac{\eta_l}{(\lambda_i^{z_l})^2 + \eta_l^2}, \qquad (7.6)$$

for $\eta_l \in [n^{-1-\delta_0}, n^{-1+\delta_1}]$. As a consequence of Proposition 7.2, the summations over $|i| \le n^{\widehat{\omega}}$ are asymptotically independent for different l's. We now prove that the sum over $n^{\widehat{\omega}} < |i| \le n$ in (7.6) is much smaller n^{-c} for some small constant $c > 0$.

Since $\omega_h \ll \widehat{\omega}$ the rigidity of the eigenvalues in (7.4) holds for $n^{\widehat{\omega}} \le |i| \le n^{1-10\omega_h}$, hence we conclude the following bound with very high probability:

$$\frac{1}{n} \sum_{n^{\widehat{\omega}} \le |i| \le n} \frac{\eta_l}{(\lambda_i^{z_l})^2 + \eta_l^2} \lesssim n^{40\omega_h} \sum_{n^{\widehat{\omega}} \le |i| \le n} \frac{n\eta_l}{i^2(\rho^{z_l}(0))^2} \lesssim \frac{n^{\delta_1+40\omega_h}}{n^{\widehat{\omega}}}, \tag{7.7}$$

where we used that $(\lambda_i^z)^2 + \eta^2 \gtrsim n^{-40\omega_h}$ for $n^{1-10\omega_h} \le |i| \le n$, and that $\eta_l \in [n^{-1-\delta_0}, n^{-1+\delta_1}]$. In particular, in (7.7) we used that by (7.5) it follows $\gamma_i^{z_l} \sim i/(n\rho^{z_l}(0))$ for $|i| \le n^{1-10\omega_h}$, where $\rho^{z_l}(0) = \Im m^{z_l}(0) = (1 - |z_l|^2)^{1/2}$ for $|z_l|^2 \le 1$ by (3.7).

Combining (7.6)–(7.7) with Proposition 7.2 we immediately conclude that

$$\begin{aligned}
\mathbf{E} \prod_{l=1}^p \langle G^{z_l}(\mathrm{i}\eta_l) \rangle &= \mathbf{E} \prod_{l=1}^p \frac{\mathrm{i}}{2n} \sum_{|i| \le n^{\widehat{\omega}}} \frac{\eta_l}{(\lambda_i^{z_l})^2 + \eta_l^2} + \mathcal{O}\left(\frac{n^{\delta_1+40\omega_h}}{n^{\widehat{\omega}}} \right) \\
&= \prod_{l=1}^p \mathbf{E} \frac{\mathrm{i}}{2n} \sum_{|i| \le n^{\widehat{\omega}}} \frac{\eta_l}{(\lambda_i^{z_l})^2 + \eta_l^2} + \mathcal{O}\left(\frac{n^{p\delta_0+\widehat{\omega}}}{n^{\omega}} + \frac{n^{\delta_1+40\omega_h}}{n^{\widehat{\omega}}} \right) \\
&= \prod_{l=1}^p \mathbf{E} \langle G^{z_l}(\mathrm{i}\eta_l) \rangle + \mathcal{O}\left(\frac{n^{\delta_1+40\omega_h}}{n^{\widehat{\omega}}} + \frac{n^{p\delta_0+\widehat{\omega}}}{n^{\omega}} \right).
\end{aligned}$$

This concludes the proof of Proposition 3.5 since $\omega_h \ll \delta_m \ll \widehat{\omega} \ll \omega$, with $m = 0, 1$. $\qquad\square$

We conclude Section 7.1 with some properties of m^z, the unique solution of (3.1). Fix $z \in \mathbf{C}$, and consider the $2n \times 2n$ matrix $A+F$, with F a Wigner matrix, whose entries are centred random variables of variance $(2n)^{-1}$, and A is a deterministic diagonal matrix $A := \mathrm{diag}(|z|, \ldots, |z|, -|z|, \ldots, -|z|)$. Then by [29, Eq. (2.1)], [37, Eq. (2.2)] it follows that the corresponding *Dyson equation* is given by

$$\begin{cases} -\frac{1}{m_1} = w - |z| + \frac{m_1+m_2}{2} \\ -\frac{1}{m_2} = w + |z| + \frac{m_1+m_2}{2}, \end{cases} \tag{7.8}$$

which has a unique solution under the assumption $\Im m_1, \Im m_2 > 0$. By (7.8) it readily follows that m^z, the solution of (3.1), satisfies

$$m^z(w) = \frac{m_1(w) + m_2(w)}{2}. \tag{7.9}$$

In particular, this implies that all the regularity properties of $m_1 + m_2$ (see e.g. [3, Theorem 2.4, Lemma A.7], [6, Proposition 2.3, Lemma A.1]) hold for m^z as well, e.g. m^z is $1/3$-Hölder continuous for any $z \in \mathbf{C}$.

7.2. Overview of the proof of Proposition 7.2.

The main result of this section is the proof of Proposition 7.2, which is divided into two further sub-sections. In Lemma 7.5, we prove that we can add a common small Ginibre component to the matrices H^{z_l}, with $l \in [p]$, $p \in \mathbf{N}$, without changing their joint eigenvalue distribution much. In Section 7.2.1, we introduce comparison processes for the process defined in (7.15) below, with initial data $\boldsymbol{\lambda}^{z_l} = \{\lambda_{\pm i}^{z_l}\}_{i=1}^n$, where we recall that $\{\lambda_i^{z_l}\}_{i=1}^n$ are the singular values of $\check{X}_{t_f} - z_l$, and $\lambda_{-i}^{z_l} = -\lambda_i^{z_l}$ (the matrix $\check{X}_{t_f}$ is defined in (7.12) below). Finally, in Section 7.2.2 we conclude the proof of Proposition 7.2. Additionally, in Section 7.2.3 we summarize the different scales used in the proof of Proposition 7.2.

Let X be an i.i.d. complex $n \times n$ matrix, and run the Ornstein-Uhlenbeck (OU) flow

$$d\widehat{X}_t = -\frac{1}{2}\widehat{X}_t \, dt + \frac{d\widehat{B}_t}{\sqrt{n}}, \qquad \widehat{X}_0 = X, \tag{7.10}$$

for a time

$$t_f := \frac{n^{\omega_f}}{n}, \tag{7.11}$$

with some small exponent $\omega_f > 0$ given in Proposition 7.2, in order to add a small Gaussian component to X. $\widehat{B}_t$ in (7.10) is a standard matrix valued complex Brownian motion independent of $\widehat{X}_0$,

i.e. $\sqrt{2}\Re\widehat{B}_{ab}$, $\sqrt{2}\Im\widehat{B}_{ab}$ are independent standard real Brownian motions for any $a, b \in [n]$. Then we construct an i.i.d. matrix $\check{X}_{t_f}$ such that

$$\widehat{X}_{t_f} \stackrel{d}{=} \check{X}_{t_f} + \sqrt{ct_f}U, \tag{7.12}$$

for some constant $c > 0$ very close to 1, and U is a complex Ginibre matrix independent of $\check{X}_{t_f}$.

Next, we define the matrix flow

$$dX_t = \frac{dB_t}{\sqrt{n}}, \quad X_0 = \check{X}_{t_f}, \tag{7.13}$$

where B_t is a standard matrix valued complex Brownian motion independent of X_0 and $\widehat{B}_t$. Note that by construction X_{ct_f} is such that

$$X_{ct_f} \stackrel{d}{=} \widehat{X}_{t_f}. \tag{7.14}$$

Define the matrix $H_t^{z_l}$ as in (7.1) replacing $X - z$ by $X_t - z_l$, for any $l \in [p]$, then the flow in (7.13) induces the following DBM flow on the eigenvalues of $H_t^{z_l}$ (cf. [39, Eq. (5.8)]):

$$d\lambda_i^{z_l}(t) = \sqrt{\frac{1}{2n}}\, db_i^{z_l} + \frac{1}{2n} \sum_{j \neq i} \frac{1}{\lambda_i^{z_l}(t) - \lambda_j^{z_l}(t)}\, dt, \qquad 1 \leq |i| \leq n, \tag{7.15}$$

with initial data $\{\lambda_{\pm i}^{z_l}(0)\}_{i=1}^n$, where $\lambda_i^{z_l}(0)$, with $i \in [n]$ and $l \in [p]$, are the singular values of $\check{X}_{t_f} - z_l$, and $\lambda_{-i}^{z_l} = -\lambda_i^{z_l}$. The well-posedness of (7.15) follows by [26, Appendix A]. It follows from this derivation that the Brownian motions $\{b_i^{z_l}\}_{i=1}^n$, omitting the t-dependence, are defined as

$$db_i^{z_l} := \sqrt{2}\left(dB_{ii}^{z_l} + \overline{dB_{ii}^{z_l}}\right), \qquad dB_{ij}^{z_l} := \sum_{a,b=1}^n \overline{u_i^{z_l}(a)}\, dB_{ab} v_j^{z_l}(b), \tag{7.16}$$

where $(\boldsymbol{u}_i^{z_l}, \pm\boldsymbol{v}_i^{z_l})$ are the orthonormal eigenvectors of $H_t^{z_l}$ with corresponding eigenvalues $\lambda_{\pm i}^{z_l}$, and B_{ab} are the entries of the Brownian motion defined in (7.13). For negative indices we define $b_{-i}^{z_l} := -b_i^{z_l}$. It follows from (7.16) that for each fixed l the collection of Brownian motions $\boldsymbol{b}^{z_l} = \{b_i^{z_l}\}_{i=1}^n$ consists of i.i.d. Brownian motions, however the families $\boldsymbol{b}^{z_l}$ are not independent for different l's, in fact their joint distribution is not necessarily Gaussian. The derivation of (7.15) follows standard steps, see e.g. [41, Section 12.2]. For the convenience of the reader we included this derivation in Appendix B.

Remark 7.4. *We point out that in the formula [26, Eq. (3.9)] analogous to (7.15) the term $j = -i$ in (7.15) is apparently missing. This additional term does not influence the results in [26, Section 3] (that are proven for the real DBM for which the term $j = -i$ is actually not present).*

As a consequence of (7.14) we conclude the following lemma.

Lemma 7.5. *Let $\boldsymbol{\lambda}^{z_l} = \{\lambda_{\pm i}^{z_l}\}_{i=1}^n$ be the eigenvalues of H^{z_l} and let $\boldsymbol{\lambda}^{z_l}(t)$ be the solution of (7.15) with initial data $\boldsymbol{\lambda}^{z_l}$, then*

$$\begin{aligned}
\mathbf{E}\prod_{l=1}^p \frac{1}{n} \sum_{|i_l| \leq n^{\widehat{\omega}}} \frac{\eta_l}{(\lambda_{i_l}^{z_l})^2 + \eta_l^2} &= \mathbf{E}\prod_{l=1}^p \frac{1}{n} \sum_{|i_l| \leq n^{\widehat{\omega}}} \frac{\eta_l}{(\lambda_{i_l}^{z_l}(ct_f))^2 + \eta_l^2} \\
&+ \mathcal{O}\left(\frac{n^{p\xi+2\delta_0}t_f}{n^{1/2}} \sum_{l=1}^p \frac{1}{\eta_l} + \frac{n^{k\delta_0+\delta_1}}{n^{\widehat{\omega}}}\right),
\end{aligned} \tag{7.17}$$

for any sufficiently small $\widehat{\omega}, \delta_0, \delta_1 > 0$ such that $\delta_m \ll \widehat{\omega}$, where $\eta_l \in [n^{-1-\delta_0}, n^{-1+\delta_1}]$ and t_f defined in (7.11).

Proof. The equality in (7.17) follows by a standard Green's function comparison (GFT) argument (e.g. see [30, Proposition 3.1]) for the $\langle G^{z_l}(i\eta_l)\rangle$, combined with the same argument as in the proof of Proposition 3.5, using the local law [4, Theorem 5.1] and (7.14), to show that the summation over $n^{\widehat{\omega}} < |i| \leq n$ is negligible. We remark that the GFT used in this lemma is much easier than the one in [30, Proposition 3.1] since here we used GFT only for a very short time $t_f \sim n^{-1+\omega_f}$, for a very small $\omega_f > 0$, whilst in [30, Proposition 3.1] the GFT is considered up to a time $t = +\infty$. The scaling in the error term in [30,

Proposition 3.1] is different compared to the error term in (7.17) since the scaling therein refers to the cusp-scaling. □

7.2.1. *Definition of the comparison processes for* $\boldsymbol{\lambda}^{z_l}(t)$. The philosophy behind the proof of Proposition 7.2 is to compare the distribution of $\boldsymbol{\lambda}^{z_l}(t) = \{\lambda_{\pm i}^{z_l}(t)\}$, the strong solutions of (7.15) for $l \in [p]$, which are correlated for different l's and realized on a probability space Ω_b, with carefully constructed independent processes $\boldsymbol{\mu}^{(l)}(t) = \{\mu_{\pm i}^{(l)}(t)\}_{i=1}^n$ on a different probability space Ω_β. We choose $\boldsymbol{\mu}^{(l)}(t)$ to be the solution of

$$\mathrm{d}\mu_i^{(l)}(t) = \frac{\mathrm{d}\beta_i^{(l)}}{\sqrt{2n}} + \frac{1}{2n}\sum_{j\neq i}\frac{1}{\mu_i^{(l)}(t) - \mu_j^{(l)}(t)}\,\mathrm{d}t, \quad \mu_i^{(l)}(0) = \mu_i^{(l)}, \tag{7.18}$$

for $|i| \leq n$, with $\mu_i^{(l)}$ the eigenvalues of the matrix

$$H^{(l)} := \begin{pmatrix} 0 & X^{(l)} \\ (X^{(l)})^* & 0 \end{pmatrix}$$

where $X^{(l)}$ are independent Ginibre matrices, $\boldsymbol{\beta}^{(l)} = \{\beta_i^{(l)}\}_{i=1}^n$ are independent vectors of i.i.d. standard real Brownian motions, and $\beta_{-i}^{(l)} = -\beta_i^{(l)}$. We let $\mathcal{F}_{\beta,t}$ denote the common filtration of the Brownian motions $\boldsymbol{\beta}^{(l)}$ on Ω_β.

In the remainder of this section we define two processes $\widetilde{\boldsymbol{\lambda}}^{(l)}, \widetilde{\boldsymbol{\mu}}^{(l)}$ so that for a time $t \geq 0$ large enough $\widetilde{\lambda}_i^{(l)}(t), \widetilde{\mu}_i^{(l)}(t)$ for small indices i will be close to $\lambda_i^{z_l}(t)$ and $\mu_i^{(l)}(t)$, respectively, with very high probability. Additionally, the processes $\widetilde{\boldsymbol{\lambda}}^{(l)}, \widetilde{\boldsymbol{\mu}}^{(l)}$ will be such that they have the same joint distribution:

$$\left(\widetilde{\boldsymbol{\lambda}}^{(1)}(t), \ldots, \widetilde{\boldsymbol{\lambda}}^{(p)}(t)\right)_{t\geq 0} \stackrel{d}{=} \left(\widetilde{\boldsymbol{\mu}}^{(1)}(t), \ldots, \widetilde{\boldsymbol{\mu}}^{(p)}(t)\right)_{t\geq 0}. \tag{7.19}$$

Fix $\omega_A > 0$ and define the process $\widetilde{\boldsymbol{\lambda}}(t)$ to be the solution of

$$\mathrm{d}\widetilde{\lambda}_i^{(l)}(t) = \frac{1}{2n}\sum_{j\neq i}\frac{1}{\widetilde{\lambda}_i^{(l)}(t) - \widetilde{\lambda}_j^{(l)}(t)}\,\mathrm{d}t + \begin{cases} \sqrt{\frac{1}{2n}}\,\mathrm{d}b_i^{z_l} & \text{if} \quad |i| \leq n^{\omega_A} \\ \sqrt{\frac{1}{2n}}\,\mathrm{d}\widetilde{b}_i^{(l)} & \text{if} \quad n^{\omega_A} < |i| \leq n, \end{cases} \tag{7.20}$$

with initial data $\widetilde{\boldsymbol{\lambda}}^{(l)}(0)$ being the singular values, taken with positive and negative sign, of independent Ginibre matrices $\widetilde{Y}^{(l)}$ independent of $\boldsymbol{\lambda}^{z_l}(0)$. Here $\mathrm{d}b_i^{z_l}$ is from (7.15); this is used for small indices. For large indices we define the driving Brownian motions to be an independent collection $\{\{\widetilde{b}_i^{(l)}\}_{i=n^{\omega_A}+1}^n \mid l \in [p]\}$ of p vector-valued i.i.d. standard real Brownian motions which are also independent of $\{\{b_{\pm i}^{z_l}\}_{i=1}^n \mid l \in [p]\}$, and that $\widetilde{b}_{-i}^{(l)} = -\widetilde{b}_i^{(l)}$. The Brownian motions $\boldsymbol{b}^{z_l}$, with $l \in [p]$, and $\{\{\widetilde{b}_i^{(l)}\}_{i=n^{\omega_A}+1}^n \mid l \in [p]\}$ are defined on a common probability space that we continue to denote by Ω_b with the common filtration $\mathcal{F}_{b,t}$.

We conclude this section by defining $\widetilde{\boldsymbol{\mu}}^{(l)}(t)$, the comparison process of $\boldsymbol{\mu}^{(l)}(t)$. It is given as the solution of the following DBM:

$$\mathrm{d}\widetilde{\mu}_i^{(l)}(t) = \frac{1}{2n}\sum_{j\neq i}\frac{1}{\widetilde{\mu}_i^{(l)}(t) - \widetilde{\mu}_j^{(l)}(t)}\,\mathrm{d}t + \begin{cases} \sqrt{\frac{1}{2n}}\,\mathrm{d}\zeta_i^{z_l} & \text{if} \quad |i| \leq n^{\omega_A} \\ \sqrt{\frac{1}{2n}}\,\mathrm{d}\widetilde{\zeta}_i^{(l)} & \text{if} \quad n^{\omega_A} < |i| \leq n, \end{cases} \tag{7.21}$$

with initial data $\widetilde{\boldsymbol{\mu}}^{(l)}(0)$ so that they are the singular values of independent Ginibre matrices $Y^{(l)}$, which are also independent of $\widetilde{Y}^{(l)}$. We now explain how to construct the driving Brownian motions in (7.21) so that (7.19) is satisfied. We only consider positive indices, since the negative indices are defined by symmetry. For indices $n^{\omega_A} < i \leq n$ we choose $\{\widetilde{\zeta}_{\pm i}^{(l)}\}_{n^{\omega_A}+1}^n$ to be independent families (for different l's) of i.i.d. Brownian motions, defined on the same probability space of $\{\boldsymbol{\beta}^{(l)} : l \in [p]\}$, that are independent of the Brownian motions $\{\beta_{\pm i}^{(l)}\}_{i=1}^n$ used in (7.18). For indices $1 \leq i \leq n^{\omega_A}$ the families $\{\{\zeta_i^{z_l}\}_{i=1}^{n^{\omega_A}} \mid l \in [p]\}$ will be constructed from the independent families $\{\{\beta_i^{(l)}\}_{i=1}^{n^{\omega_A}} \mid l \in [p]\}$ as follows.

Arranging $\{\{\beta_i^{(l)}\}_{i=1}^{n^{\omega_A}} \mid l \in [p]\}$ into a single vector, we define the pn^{ω_A}-dimensional vector

$$\underline{\beta} := (\beta_1^{(1)}, \ldots, \beta_{n^{\omega_A}}^{(1)}, \ldots, \beta_1^{(p)}, \ldots, \beta_{n^{\omega_A}}^{(p)}). \tag{7.22}$$

Similarly we define the pn^{ω_A}-dimensional vector

$$\underline{b} := (b_1^{z_1}, \ldots, b_{n^{\omega_A}}^{z_1}, \ldots, b_1^{z_p}, \ldots, b_{n^{\omega_A}}^{z_p}) \tag{7.23}$$

which is a continuous martingale. To make our notation easier, in the following we assume that $n^{\omega_A} \in \mathbf{N}$. For any $i, j \in [pn^{\omega_A}]$, we use the notation

$$i = (l-1)n^{\omega_A} + \mathsf{i}, \qquad j = (m-1)n^{\omega_A} + \mathsf{j}, \tag{7.24}$$

with $l, m \in [p]$ and $\mathsf{i}, \mathsf{j} \in [n^{\omega_A}]$. Note that in the definitions in (7.24) we used $(l-1), (m-1)$ instead of l, m so that l and m exactly indicate in which block of the matrix $C(t)$ in (7.25) the indices i, j are. With this notation, the covariance matrix of the increments of $\underline{b}$ is the matrix $C(t)$ consisting of p^2 blocks of size n^{ω_A} is defined as

$$C_{ij}(t)\,\mathrm{d}t := \mathbf{E}[\mathrm{d}b_{\mathsf{i}}^{z_l}\,\mathrm{d}b_{\mathsf{j}}^{z_m} \mid \mathcal{F}_{b,t}] = \begin{cases} \Theta_{\mathsf{ij}}^{z_l, z_m}(t)\,\mathrm{d}t & \text{if} \quad l \neq m, \\ \delta_{\mathsf{ij}}\,\mathrm{d}t & \text{if} \quad l = m. \end{cases} \tag{7.25}$$

Here

$$\Theta_{\mathsf{ij}}^{z_l, z_m}(t) := 4\Re\big[\langle \boldsymbol{u}_{\mathsf{i}}^{z_l}(t), \boldsymbol{u}_{\mathsf{j}}^{z_m}(t)\rangle\langle \boldsymbol{v}_{\mathsf{i}}^{z_m}(t), \boldsymbol{v}_{\mathsf{j}}^{z_l}(t)\rangle\big], \tag{7.26}$$

with $\{\boldsymbol{w}_{\pm i}\}_{i \in [n]} = \{(\boldsymbol{u}_i^{z_l}(t), \pm\boldsymbol{v}_i^{z_l}(t))\}_{i \in [n]}$ the orthonormal eigenvectors of $H_t^{z_l}$. Note that $\{\boldsymbol{w}_i\}_{|i| \leq n}$ are not well-defined if $H_t^{z_l}$ has multiple eigenvalues. However, without loss of generality, we can assume that almost surely $H_t^{z_l}$ does not have multiple eigenvalues for any $l \in [p]$, as a consequence of [27, Lemma 6.2] (which is the adaptation of [25, Proposition 2.3] to the 2×2 block structure of $H_t^{z_l}$).

By Doob's martingale representation theorem [54, Theorem 18.12] there exists a standard Brownian motion $\boldsymbol{\theta}_t \in \mathbf{R}^{pN^{\omega_A}}$ realized on an extension $(\widetilde{\Omega}_b, \widetilde{\mathcal{F}}_{b,t})$ of the original filtrated probability space $(\Omega_b, \mathcal{F}_{b,t})$ such that $\mathrm{d}\underline{b} = \sqrt{C}\,\mathrm{d}\boldsymbol{\theta}$. Here $\boldsymbol{\theta}_t$ and $C(t)$ are adapted to the filtration $\widetilde{\mathcal{F}}_{b,t}$ and note that $C = C(t)$ is a positive semi-definite matrix and $\sqrt{C}$ denotes its positive semi-definite matrix square root.

For the clarity of the presentation the original processes $\boldsymbol{\lambda}^{z_l}$ and the comparison processes $\boldsymbol{\mu}^{(l)}$ will be realized on completely different probability spaces. We thus construct another copy $(\Omega_\beta, \mathcal{F}_{\beta,t})$ of the filtrated probability space $(\widetilde{\Omega}_b, \widetilde{\mathcal{F}}_{b,t})$ and we construct a matrix valued process $C^\#(t)$ and a Brownian motion $\underline{\beta}$ on $(\Omega_\beta, \mathcal{F}_{\beta,t})$ such that $(C^\#(t), \underline{\beta}(t))$ are adapted to the filtration $\mathcal{F}_{\beta,t}$ and they have the same joint distribution as $(C(t), \boldsymbol{\theta}(t))$. The Brownian motion $\underline{\beta}$ is used in (7.18) for small indices.

Define the process

$$\underline{\zeta}(t) := \int_0^t \sqrt{C^\#(s)}\,\mathrm{d}\underline{\beta}(s), \quad \underline{\zeta} = (\zeta_1^{z_1}, \ldots, \zeta_{n^{\omega_A}}^{z_1}, \ldots, \zeta_1^{z_p}, \ldots, \zeta_{n^{\omega_A}}^{z_p}), \tag{7.27}$$

on the probability space Ω_β and define $\zeta_{-i}^{z_l} := -\zeta_i^{z_l}$ for any $1 \leq i \leq n^{\omega_A}, l \in [p]$. Since $\underline{\beta}$ are i.i.d. Brownian motions, we clearly have

$$\mathbf{E}[\mathrm{d}\zeta_{\mathsf{i}}^{z_l}(t)\,\mathrm{d}\zeta_{\mathsf{j}}^{z_m}(t) \mid \mathcal{F}_{\beta,t}] = C^\#(t)_{ij}\,\mathrm{d}t, \qquad |\mathsf{i}|, |\mathsf{j}| \leq n^{\omega_A}. \tag{7.28}$$

By construction we see that the processes $(\{b_{\pm i}^{z_l}\}_{i=1}^{n^{\omega_A}})_{l=1}^k$ and $(\{\zeta_{\pm i}^{z_l}\}_{i=1}^{n^{\omega_A}})_{l=1}^k$ have the same distribution. Furthermore, since by definition the two collections

$$\left\{ \{\widetilde{b}_{\pm i}^{(l)}\}_{i=n^{\omega_A}+1}^n, \{\widetilde{\zeta}_{\pm i}^{(l)}\}_{i=n^{\omega_A}+1}^n \,\Big|\, l \in [k] \right\}$$

are independent of

$$\left\{ \{b_{\pm i}^{z_l}\}_{i=1}^{n^{\omega_A}}, \{\beta_{\pm i}^{(l)}\}_{i=1}^{n^{\omega_A}} \,\Big|\, l \in [k] \right\}$$

and among each other, we have

$$\left(\{b_{\pm i}^{z_l}\}_{i=1}^{n^{\omega_A}}, \{\widetilde{b}_{\pm i}^{(l)}\}_{i=n^{\omega_A}+1}^n \right)_{l=1}^p \stackrel{d}{=} \left(\{\zeta_{\pm i}^{z_l}\}_{i=1}^{n^{\omega_A}}, \{\widetilde{\zeta}_{\pm i}^{(l)}\}_{i=n^{\omega_A}+1}^n \right)_{l=1}^p. \tag{7.29}$$

Finally, by the definitions in (7.20), (7.21), and (7.29), it follows that the Dyson Brownian motions $\widetilde{\boldsymbol{\lambda}}^{(l)}$ and $\widetilde{\boldsymbol{\mu}}^{(l)}$ have the same distribution, i.e.

$$\left(\widetilde{\boldsymbol{\lambda}}^{(1)}(t), \dots, \widetilde{\boldsymbol{\lambda}}^{(p)}(t)\right) \stackrel{d}{=} \left(\widetilde{\boldsymbol{\mu}}^{(1)}(t), \dots, \widetilde{\boldsymbol{\mu}}^{(p)}(t)\right) \tag{7.30}$$

since their initial conditions, as well as their driving processes (7.29), agree in distribution. Note that these processes are Brownian motions for each fixed l since $C_{ij}(t) = \delta_{ij}$ if $l = m$, but jointly they are not necessarily Gaussian due to the non-trivial correlation $\Theta_{ij}^{z_l, z_m}$ in (7.25).

7.2.2. *Proof of Proposition 7.2.* In this section we conclude the proof of Proposition 7.2 using the comparison processes defined in Section 7.2.1. More precisely, we use that the processes $\boldsymbol{\lambda}^{z_l}(t)$, $\widetilde{\boldsymbol{\lambda}}^{(l)}(t)$ and $\boldsymbol{\mu}^{(l)}(t)$, $\widetilde{\boldsymbol{\mu}}^{(l)}(t)$ are close pathwise at time t_f, as stated below in Lemma 7.6 and Lemma 7.7, respectively. The proofs of these lemmas are postponed to Section 7.5. They will be a consequence of Proposition 7.14, which is an adaptation to our case of the main technical estimate of [59]. The main input is the bound on the eigenvector overlap in Lemma 7.9, since it gives an upper bound on the correlation structure in (7.28). Let $\rho_{sc}(E) = \frac{1}{2\pi}\sqrt{4 - E^2}$ denote the semicircle density.

Lemma 7.6. *Fix $p \in \mathbf{N}$, and let $\boldsymbol{\lambda}^{z_l}(t)$, $\widetilde{\boldsymbol{\lambda}}^{(l)}(t)$, with $l \in [p]$, be the processes defined in (7.15) and (7.20), respectively. For any small $\omega_h, \omega_f > 0$ such that $\omega_h \ll \omega_f$ there exist $\omega, \widehat{\omega} > 0$ with $\omega_h \ll \widehat{\omega} \ll \omega \ll \omega_f$, such that for any $|z_l| \leq 1 - n^{-\omega_h}$ it holds*

$$\left|\rho^{z_l}(0)\lambda_i^{z_l}(ct_f) - \rho_{sc}(0)\widetilde{\lambda}_i^{(l)}(ct_f)\right| \leq n^{-1-\omega}, \qquad |i| \leq n^{\widehat{\omega}}, \tag{7.31}$$

with very high probability, where $t_f := n^{-1+\omega_f}$ and $c > 0$ is defined in (7.14).

Lemma 7.7. *Fix $p \in \mathbf{N}$, and let $\boldsymbol{\mu}^{(l)}(t)$, $\widetilde{\boldsymbol{\mu}}^{(l)}(t)$, with $l \in [p]$, be the processes defined in (7.18) and (7.21), respectively. For any small $\omega_h, \omega_f, \omega_d > 0$ such that $\omega_h \ll \omega_f$ there exist $\omega, \widehat{\omega} > 0$ with $\omega_h \ll \widehat{\omega} \ll \omega \ll \omega_f$, such that for any $|z_l| \leq 1 - n^{-\omega_h}$, $|z_l - z_m| \geq n^{-\omega_d}$, with $l \neq m$, it holds*

$$\left|\mu_i^{(l)}(ct_f) - \widetilde{\mu}_i^{(l)}(ct_f)\right| \leq n^{-1-\omega}, \qquad |i| \leq n^{\widehat{\omega}}, \tag{7.32}$$

with very high probability, where $t_f := n^{-1+\omega_f}$ and $c > 0$ is defined in (7.14).

Proof of Proposition 7.2. In the following we omit the trivial scaling factors $\rho^{z_l}(0)$, $\rho_{sc}(0)$ in the second term in the lhs. of (7.31) to make our notation easier. We recall that by Lemma 7.5 we have

$$\mathbf{E}\prod_{l=1}^p \frac{1}{n}\sum_{|i_l|\leq n^{\widehat{\omega}}} \frac{\eta_l}{(\lambda_{i_l}^{z_l})^2 + \eta_l^2} = \mathbf{E}\prod_{l=1}^p \frac{1}{n}\sum_{|i_l|\leq n^{\widehat{\omega}}} \frac{\eta_l}{(\lambda_{i_l}^{z_l}(ct_f))^2 + \eta_l^2}$$
$$+ \mathcal{O}\left(\frac{n^{p\xi+2\delta_0}t_f}{n^{1/2}}\sum_{l=1}^p \frac{1}{\eta_l} + \frac{n^{p\delta_0+\delta_1}}{n^{\widehat{\omega}}}\right), \tag{7.33}$$

where $\lambda_i^{z_l}(t)$ is the solution of (7.15) with initial data $\lambda_i^{z_l}$. Next we replace $\lambda_i^{z_l}(t)$ with $\widetilde{\lambda}_i^{z_l}(t)$ for small indices by using Lemma 7.6; this is formulated in the following lemma whose detailed proof is postponed to the end of this section.

Lemma 7.8. *Fix $p \in \mathbf{N}$, and let $\lambda_i^{z_l}(t)$, $\widetilde{\lambda}_i^{(l)}(t)$, with $l \in [p]$, be the solution of (7.15) and (7.20), respectively. Then*

$$\mathbf{E}\prod_{l=1}^p \frac{1}{n}\sum_{|i_l|\leq n^{\widehat{\omega}}} \frac{\eta_l}{(\lambda_{i_l}^{z_l})^2 + \eta_l^2} = \mathbf{E}\prod_{l=1}^p \frac{1}{n}\sum_{|i_l|\leq n^{\widehat{\omega}}} \frac{\eta_l}{(\widetilde{\lambda}_{i_l}^{(l)}(ct_f))^2 + \eta_l^2} + \mathcal{O}(\Psi), \tag{7.34}$$

where $\lambda_{i_l}^{z_l} = \lambda_{i_l}^{z_l}(0)$, $t_f = n^{-1+\omega_f}$, and the error term is given by

$$\Psi := \frac{n^{\widehat{\omega}}}{n^{1+\omega}}\left(\sum_{l=1}^p \frac{1}{\eta_l}\right)\cdot\prod_{l=1}^p\left(1 + \frac{n^\xi}{n\eta_l}\right) + \frac{n^{p\xi+2\delta_0}t_f}{n^{1/2}}\sum_{l=1}^p\frac{1}{\eta_l} + \frac{n^{p\delta_0+\delta_1}}{n^{\widehat{\omega}}}.$$

By (7.30) it readily follows that

$$\mathbf{E}\prod_{l=1}^{p}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat\omega}}\frac{\eta_l}{(\widetilde\lambda_{i_l}^{(l)}(ct_f))^2+\eta_l^2}=\mathbf{E}\prod_{l=1}^{p}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat\omega}}\frac{\eta_l}{(\widetilde\mu_{i_l}^{(l)}(ct_f))^2+\eta_l^2}. \qquad (7.35)$$

Moreover, by (7.32), similarly to Lemma 7.8, we conclude

$$\mathbf{E}\prod_{l=1}^{p}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat\omega}}\frac{\eta_l}{(\widetilde\mu_{i_l}^{(l)}(ct_f))^2+\eta_l^2}=\mathbf{E}\prod_{l=1}^{p}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat\omega}}\frac{\eta_l}{(\mu_{i_l}^{(l)}(ct_f))^2+\eta_l^2}+\mathcal{O}(\Psi). \qquad (7.36)$$

Additionally, by the definition of the processes $\boldsymbol\mu^{(l)}(t)$ in (7.18) it follows that $\boldsymbol\mu^{(l)}(t)$, $\boldsymbol\mu^{(m)}(t)$ are independent for $l\neq m$ and so that

$$\mathbf{E}\prod_{l=1}^{p}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat\omega}}\frac{\eta_l}{(\mu_{i_l}^{(l)}(ct_f))^2+\eta_l^2}=\prod_{l=1}^{p}\mathbf{E}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat\omega}}\frac{\eta_l}{(\mu_{i_l}^{(l)}(ct_f))^2+\eta_l^2}. \qquad (7.37)$$

Combining (7.34)–(7.37), we get

$$\mathbf{E}\prod_{l=1}^{p}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat\omega}}\frac{\eta_l}{(\lambda_{i_l}^{z_l})^2+\eta_l^2}=\prod_{l=1}^{p}\mathbf{E}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat\omega}}\frac{\eta_l}{(\mu_{i_l}^{(l)}(ct_f))^2+\eta_l^2}+\mathcal{O}(\Psi). \qquad (7.38)$$

Then, by similar computation to the ones in (7.33)–(7.38) we conclude that

$$\prod_{l=1}^{p}\mathbf{E}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat\omega}}\frac{\eta_l}{(\lambda_{i_l}^{z_l})^2+\eta_l^2}=\prod_{l=1}^{p}\mathbf{E}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat\omega}}\frac{\eta_l}{(\mu_{i_l}^{(l)}(ct_f))^2+\eta_l^2}+\mathcal{O}(\Psi). \qquad (7.39)$$

We remark that in order to prove (7.39) it would not be necessary to introduce the additional comparison processes $\widetilde{\boldsymbol\lambda}^{(l)}$ and $\widetilde{\boldsymbol\mu}^{(l)}$ of Section 7.2.1, since in (7.39) the product is outside the expectation, so one can compare the expectations one by one; the correlation between these processes for different l's plays no role. Hence, already the usual coupling (see e.g. [20, 26, 59]) between the processes $\boldsymbol\lambda^{z_l}(t)$, $\boldsymbol\mu^{(l)}(t)$ defined in (7.15) and (7.18), respectively, would be sufficient to prove (7.39).

Finally, combining (7.38)–(7.39) we conclude the proof of Proposition 7.2. □

Proof of Lemma 7.8. We show the proof for $p=2$ in order to make our presentation easier. The case $p\geq 3$ proceeds exactly in the same way. In order to make our notation shorter, for $l\in\{1,2\}$, we define

$$T_{i_l}^{(l)}:=\frac{\eta_l}{(\lambda_{i_l}^{z_l}(ct_f))^2+\eta_l^2}.$$

Similarly, replacing $\lambda_{i_l}^{z_l}(ct_f)$ with $\widetilde\lambda_{i_l}^{(l)}(ct_f)$, we define $\widetilde T_l$. Then, by telescopic sum, we have

$$\left|\mathbf{E}\prod_{l=1}^{2}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat\omega}}T_{i_l}^{(l)}-\mathbf{E}\prod_{l=1}^{2}\frac{1}{n}\sum_{|i_l|\leq n^{\widehat\omega}}\widetilde T_{i_l}^{(l)}\right|$$

$$=\frac{1}{n^2}\left|\mathbf{E}\sum_{|i_1|,|i_2|\leq n^{\widehat\omega}}\left[T_{i_1}^{(1)}-\widetilde T_{i_1}^{(1)}\right]T_{i_2}^{(2)}-\mathbf{E}\sum_{|i_1|,|i_2|\leq n^{\widehat\omega}}\left[T_{i_2}^{(2)}-\widetilde T_{i_2}^{(2)}\right]\widetilde T_{i_1}^{(1)}\right|$$

$$\lesssim\sum_{\substack{l,m=1\\l\neq m}}^{2}\left(1+\frac{n^{\xi}}{n\eta_l}\right)\mathbf{E}\frac{1}{n}\sum_{|i_m|\leq n^{\widehat\omega}}\frac{T_{i_m}^{(m)}\widetilde T_{i_m}^{(m)}}{\eta_m}\left|(\widetilde\lambda_{i_m}^{(m)}(ct_f))^2-(\lambda_{i_m}^{z_m}(ct_f))^2\right| \qquad (7.40)$$

$$\lesssim\frac{n^{\widehat\omega}}{n^{1+\omega}}\left(\frac{1}{\eta_1}+\frac{1}{\eta_2}\right)\cdot\prod_{l=1}^{2}\left(1+\frac{n^{\xi}}{n\eta_l}\right),$$

where we used the local law (3.1) in the first inequality and (7.31) in the last step. Combining (7.40) with (7.33) we conclude the proof of Lemma 7.8. □

Before we continue, we summarize the scales used in the entire Section 7.

7.2.3. *Relations among the scales in the proof of Proposition 7.2.* Scales in the proof of Proposition 7.2 are characterized by various exponents ω's of n that we will also refer to scales, for simplicity. The basic input scales in the proof of Proposition 7.2 are $0 < \omega_d, \omega_h, \omega_f \ll 1$, the others will depend on them. The exponents ω_h, ω_d are chosen within the assumptions of Lemma 7.9 to control the location of z's as $|z_l| \leq 1 - n^{-\omega_h}, |z_l - z_m| \geq n^{-\omega_p}$, with $l \neq m$. The exponent ω_f defines the time $t_f = n^{-1+\omega_f}$ so that the local equilibrium of the DBM is reached after t_f. This will provide the asymptotic independence of $\lambda_i^{z_l}, \lambda_j^{z_m}$ for small indices and for $l \neq m$.

The primary scales created along the proof of Proposition 7.2 are $\omega, \widehat{\omega}, \delta_0, \delta_1, \omega_E, \omega_B$. The scales ω_E, ω_B are given in Lemma 7.9: $n^{-\omega_E}$ measures the size of the eigenvector overlaps from (7.26) while the exponent ω_B describes the range of indices for which these overlap estimates hold. Recall that the overlaps determine the correlations among the driving Brownian motions. The scale ω quantifies the $n^{-1-\omega}$ precision of the coupling between various processes. These couplings are effective only for small indices i, their range is given by $\widehat{\omega}$ as $|i| \leq n^{\widehat{\omega}}$. Both these scales are much bigger than ω_h but much smaller than ω_f. They are determined in Lemma 7.6, Lemma 7.7, in fact both lemmas give only a necessary upper bound on the scales $\omega, \widehat{\omega}$, so we can pick the smaller of them. The exponents δ_0, δ_1 determine the range of $\eta \in [n^{-1-\delta_0}, n^{-1+\delta_1}]$ for which Proposition 7.2 holds; these are determined in Lemma 7.5 after $\omega, \widehat{\omega}$ have already been fixed. These steps yield the scales $\omega, \widehat{\omega}, \delta_0, \delta_1$ claimed in Proposition 7.2 and hence also in Proposition 3.5. We summarize order relation among all these scales as

$$\omega_h \ll \delta_m \ll \widehat{\omega} \ll \omega \ll \omega_B \ll \omega_f \ll \omega_E \ll 1, \qquad m = 0, 1. \tag{7.41}$$

We mention that three further auxiliary scales emerge along the proof but they play only a local, secondary role. For completeness we also list them here; they are $\omega_1, \omega_A, \omega_l$. Their meanings are the following: $t_1 := n^{-1+\omega_1}$, with $\omega_1 \ll \omega_f$, is the time needed for the DBM process $x_i(t, \alpha)$, defined in (7.55), to reach local equilibrium, hence to prove its universality; $t_0 := t_f - t_1$ is the initial time we run the DBM before starting with the actual proof of universality so that the solution $\boldsymbol{\lambda}^{z_l}(t_0)$ of (7.15) at time t_0 and the density $d\rho(E, t, \alpha)$ (which we will define in Section 7.6.2) satisfy certain technical regularity conditions [26, Lemma 3.3-3.5], [59, Lemma 3.3-3.5]. Note that $t_0 \sim t_f$, in fact they are almost the same. The other two scales are technical: ω_l is the scale of the short range interaction, and ω_A is a cut-off scale such that $x_i(t, \alpha)$ is basically independent of α for $|i| \leq n^{\omega_A}$. These scales are inserted in the above chain of inequalities (7.41) between ω, ω_B as follows

$$\omega_h \ll \delta_m \ll \widehat{\omega} \ll \omega \ll \omega_1 \ll \omega_l \ll \omega_A \leq \omega_B \ll \omega_f \ll \omega_E \ll 1, \quad m = 0, 1.$$

In particular, the relation $\omega_A \ll \omega_E$ ensures that the effect of the correlation is small, see the bound in (7.54) later.

We remark that introducing the additional initial time layer t_0 is not really necessary for our proof of Proposition 7.2 since the initial data $\boldsymbol{\lambda}^z(0)$ of the DBM in (7.15) and their deterministic density ρ^z already satisfy [26, Lemma 3.3-3.5], [59, Lemma 3.3-3.5] as a consequence of (7.3) (see Remark 7.10 and Remark 7.16 for more details). We keep it only to facilitate the comparison with [26, 59].

7.3. **Bound on the eigenvector overlap for large $|z_1 - z_2|$.** For any $z \in \mathbf{C}$, let $\{\boldsymbol{w}_{\pm i}^z\}_{i=1}^n$ be the eigenvectors of the matrix H^z. They are of the form $\boldsymbol{w}_{\pm i}^z = (\boldsymbol{u}_i^z, \pm \boldsymbol{v}_i^z)$, with $\boldsymbol{u}_i^z, \boldsymbol{v}_i^z \in \mathbf{C}^n$, as a consequence of the symmetry of the spectrum of H^z induced by its block structure. The main input to prove Lemma 7.6–7.7 is the following high probability bound on the almost orthogonality of the eigenvectors belonging to distant z_l, z_m parameters and eigenvalues close to zero. With the help of the Dyson Brownian motion (DBM), this information will then be used to establish almost independence of these eigenvalues.

Lemma 7.9. *Let $\{\boldsymbol{w}_{\pm i}^{z_l}\}_{i=1}^n = \{(\boldsymbol{u}_i^{z_l}, \pm \boldsymbol{v}_i^{z_l})\}_{i=1}^n$, for $l = 1, 2$, be the eigenvectors of matrices H^{z_l} of the form (7.1) with i.i.d. entries. Then for any sufficiently small $\omega_d, \omega_h > 0$ there exist $\omega_B, \omega_E > 0$ such that if $|z_1 - z_2| \geq n^{-\omega_d}, |z_l| \leq 1 - n^{-\omega_h}$ then*

$$\left| \langle \boldsymbol{u}_i^{z_1}, \boldsymbol{u}_j^{z_2} \rangle \right| + \left| \langle \boldsymbol{v}_i^{z_1}, \boldsymbol{v}_j^{z_2} \rangle \right| \leq n^{-\omega_E}, \quad 1 \leq i, j \leq n^{\omega_B}, \tag{7.42}$$

with very high probability.

Proof. Using the spectral symmetry of H^z, for any $z \in \mathbf{C}$ we write G^z in spectral decomposition as

$$G^z(\mathrm{i}\eta) = \sum_{j>0} \frac{2}{(\lambda_j^z)^2 + \eta^2} \begin{pmatrix} \mathrm{i}\eta \boldsymbol{u}_j^z(\boldsymbol{u}_j^z)^* & \lambda_j^z \boldsymbol{u}_j^z(\boldsymbol{v}_j^z)^* \\ \lambda_j^z \boldsymbol{v}_j^z(\boldsymbol{u}_j^z)^* & \mathrm{i}\eta \boldsymbol{v}_j^z(\boldsymbol{v}_j^z)^* \end{pmatrix}.$$

Let $\eta \geq n^{-1}$, then by rigidity of the eigenvalues in (7.4), for any $i_0, j_0 \geq 1$ such that $\lambda_{i_0}^{z_l}, \lambda_{j_0}^{z_l} \lesssim \eta$, with $l = 1, 2$, and any z_1, z_2 such that $n^{-\omega_d} \lesssim |z_1 - z_2| \lesssim 1$, for some $\omega_d > 0$ we will choose shortly, it follows that

$$\begin{aligned}
&\left|\langle \boldsymbol{u}_{i_0}^{z_1}, \boldsymbol{u}_{j_0}^{z_2}\rangle\right|^2 + \left|\langle \boldsymbol{v}_{i_0}^{z_1}, \boldsymbol{v}_{j_0}^{z_2}\rangle\right|^2 \\
&\lesssim \sum_{i,j=1}^n \frac{4\eta^4}{((\lambda_i^{z_1})^2 + \eta^2)((\lambda_j^{z_2})^2 + \eta^2)} \left(\left|\langle \boldsymbol{u}_i^{z_1}, \boldsymbol{u}_j^{z_2}\rangle\right|^2 + \left|\langle \boldsymbol{v}_i^{z_1}, \boldsymbol{v}_j^{z_2}\rangle\right|^2\right) \\
&= \eta^2 \mathrm{Tr}(\Im G^{z_1})(\Im G^{z_2}) \lesssim \frac{n^{8\omega_d/3}}{(n\eta)^{1/4}} + (\eta^{1/12} + n\eta^2)n^{2\omega_d} \\
&\lesssim \frac{n^{2\omega_d + 100\omega_h}}{n^{1/23}}.
\end{aligned} \tag{7.43}$$

The first inequality in the second line of (7.43) is from Theorem 5.2 and the lower bound on $|\widehat{\beta}_*|$ from (6.2). In the last inequality we choose $\eta = n^{-12/23}$, under the assumption that $\omega_d \leq 1/100$ and that $i_0, j_0 \leq n^{1/5}$ (in order to make sure that the first inequality in (7.43) hold). We also used that the first term in the lhs. of the last inequality is always smaller than the other two for $\eta \geq n^{-5/9}$, and in the second line of (7.43) we used that M_{12}, the deterministic approximation of $\mathrm{Tr}\, \Im G^{z_1} \Im G^{z_2}$ in Theorem 5.2, is bounded by $\|M_{12}\| \lesssim |z_1 - z_2|^{-2}$.

This concludes the proof by choosing $\omega_B \leq 1/5$ and $\omega_d = 1/100$, which implies a choice of $\omega_E = -(2\omega_d + 100\omega_h - 1/23)$. $\qquad\square$

7.4. Pathwise coupling of DBM close to zero. This section is the main technical result used in the proof of Lemma 7.6 and Lemma 7.7. We compare the evolution of two DBMs whose driving Brownian motions are nearly the same for small indices and are independent for large indices. In Proposition 7.14 we will show that the points with small indices in the two processes become very close to each other on a certain time scale t_f. This time scale is chosen to be larger than the local equilibration time, but not too large so that the independence of the driving Brownian motions for large indices do not yet have an effect on particles with small indices.

Remark 7.10. *The main result of this section (Proposition 7.14) is stated for general deterministic initial data $\boldsymbol{s}(0)$ satisfying Definition 7.12 even if for its applications in the proof of Proposition 7.2 we only consider initial data which are eigenvalues of i.i.d. random matrices.*

The proof of Proposition 7.14 follows the proof of fixed energy universality in [20, 26, 59], adapted to the block structure (7.1) in [26] (see also [25, 27] for further adaptations of [20, 59] to different matrix models). The main novelty in our DBM analysis compared to [20, 26, 59] is that we analyse a process for which we allow not (fully) coupled driving Brownian motions (see Assumption 7.11).

Define the processes $s_i(t), r_i(t)$ to be the solution of

$$\mathrm{d}s_i(t) = \sqrt{\frac{1}{2n}}\, \mathrm{d}\mathfrak{b}_i^s(t) + \frac{1}{2n}\sum_{j \neq i} \frac{1}{s_i(t) - s_j(t)}\, \mathrm{d}t, \qquad 1 \leq |i| \leq n, \tag{7.44}$$

and

$$\mathrm{d}r_i(t) = \sqrt{\frac{1}{2n}}\, \mathrm{d}\mathfrak{b}_i^r(t) + \frac{1}{2n}\sum_{j \neq i} \frac{1}{r_i(t) - r_j(t)}\, \mathrm{d}t, \qquad 1 \leq |i| \leq n, \tag{7.45}$$

with initial data $s_i(0) = s_i, r_i(0) = r_i$, where $\boldsymbol{s} = \{s_{\pm i}\}_{i=1}^n$ and $\boldsymbol{r} = \{r_{\pm i}\}_{i=1}^n$ are two independent sets of particles such that $s_{-i} = -s_i$ and $r_{-i} = -r_i$ for $i \in [n]$. The driving standard real Brownian motions $\{\mathfrak{b}_i^s\}_{i=1}^n$, $\{\mathfrak{b}_i^r\}_{i=1}^n$ in (7.44)–(7.45) are two i.i.d. families and they are such that $\mathfrak{b}_{-i}^s = -\mathfrak{b}_i^s$, $\mathfrak{b}_{-i}^r = -\mathfrak{b}_i^r$ for $i \in [n]$. For convenience we also assume that $\{r_{\pm i}\}_{i=1}^n$ are the singular values of $\widetilde{X}$, with $\widetilde{X}$ a Ginibre matrix. This is not a restriction; indeed, once a process with general initial data $\boldsymbol{s}$ is

shown to be close to the reference process with Ginibre initial data, then processes with any two initial data will be close.

Fix an n-dependent parameter $K = K_n = n^{\omega_K}$, for some $\omega_K > 0$. On the correlation structure between the two families of i.i.d. Brownian motions $\{\mathfrak{b}_i^s\}_{i=1}^n$, $\{\mathfrak{b}_i^r\}_{i=1}^n$ we make the following assumptions:

Assumption 7.11. *Suppose that the families $\{\mathfrak{b}_{\pm i}^s\}_{i=1}^n$, $\{\mathfrak{b}_{\pm i}^r\}_{i=1}^n$ in (7.44) and (7.45) are realised on a common probability space with a common filtration $\mathcal{F}_t$. Let*

$$L_{ij}(t)\,\mathrm{d}t := \mathbf{E}\big[\big(\mathrm{d}\mathfrak{b}_i^s(t) - \mathrm{d}\mathfrak{b}_i^r(t)\big)\big(\mathrm{d}\mathfrak{b}_j^s(t) - \mathrm{d}\mathfrak{b}_j^r(t)\big) \,\big|\, \mathcal{F}_t\big] \tag{7.46}$$

denote the covariance of the increments conditioned on $\mathcal{F}_t$. The processes satisfy the following assumptions:

(a) *$\{\mathfrak{b}_i^s\}_{i=1}^n$, $\{\mathfrak{b}_i^r\}_{i=1}^n$ are two families of i.i.d. standard real Brownian motions.*

(b) *$\{\mathfrak{b}_{\pm i}^r\}_{i=K+1}^n$ is independent of $\{\mathfrak{b}_{\pm i}^s\}_{i=1}^n$, and $\{\mathfrak{b}_{\pm i}^s\}_{i=K+1}^n$ is independent of $\{\mathfrak{b}_{\pm i}^r\}_{i=1}^n$.*

(c) *Fix $\omega_Q > 0$ so that $\omega_K \ll \omega_Q$. We assume that the subfamilies $\{\mathfrak{b}_{\pm i}^s\}_{i=1}^K$, $\{\mathfrak{b}_{\pm i}^r\}_{i=1}^K$ are very strongly dependent in the sense that for any $|i|, |j| \leq K$ it holds*

$$|L_{ij}(t)| \leq n^{-\omega_Q} \tag{7.47}$$

with very high probability for any fixed $t \geq 0$.

Furthermore we assume that the initial data $\{s_{\pm i}\}_{i=1}^n$ is regular in the following sense (cf. [26, Definition 3.1], [59, Definition 2.1], motivated by [60, Definition 2.1]).

Definition 7.12 ((g, G)-regular points). *Fix a very small $\nu > 0$, and choose g and G such that*

$$n^{-1+\nu} \leq g \leq n^{-2\nu}, \qquad G \leq n^{-\nu}.$$

A set of $2n$-points $\boldsymbol{s} = \{s_i\}_{i=1}^{2n}$ on $\mathbf{R}$ is called (g, G)-regular if there exist constants $c_\nu, C_\nu > 0$ such that

$$c_\nu \leq \frac{1}{2n}\Im \sum_{i=-n}^n \frac{1}{s_i - (E + \mathrm{i}\eta)} \leq C_\nu, \tag{7.48}$$

for any $|E| \leq G$, $\eta \in [g, 10]$, and if there is a constant C_s large enough such that $\|\boldsymbol{s}\|_\infty \leq n^{C_s}$. Moreover, $c_\nu, C_\nu \sim 1$ if $\eta \in [g, n^{-2\nu}]$ and $c_\nu \geq n^{-100\nu}$, $C_\nu \leq n^{100\nu}$ if $\eta \in [n^{-2\nu}, 10]$.

Remark 7.13. *We point out that in [26, Definition 3.1] and [59, Definition 2.1] the constants c_ν, C_ν do not depend on $\nu > 0$, but this change does not play any role since ν will always be the smallest exponent of scale involved in the analysis of the DBMs (7.44)–(7.45), hence negligible.*

Let $\rho_{\mathrm{fc},t}(E)$ be the deterministic approximation of the density of the particles $\{s_{\pm i}(t)\}_{i=1}^n$ that is obtained from the semicircular flow acting on the empirical density of the initial data $\{s_{\pm i}(0)\}_{i=1}^n$, see [59, Eq. (2.5)–(2.6)]. Recall that $\rho_{sc}(E)$ denotes the semicircular density.

Proposition 7.14. *Let the processes $\boldsymbol{s}(t) = \{s_{\pm i}(t)\}_{i=1}^n$, $\boldsymbol{r}(t) = \{r_{\pm i}(t)\}_{i=1}^n$ be the solutions of (7.44) and (7.45), respectively, and assume that the driving Brownian motions in (7.44)–(7.45) satisfy Assumption 7.11. Additionally, assume that $\boldsymbol{s}(0)$ is (g, G)-regular in the sense of Definition 7.12 and that $\boldsymbol{r}(0)$ are the singular values of a Ginibre matrix. Then for any small $\nu, \omega_f > 0$ such that $\nu \ll \omega_K \ll \omega_f \ll \omega_Q$ and that $gn^\nu \leq t_f \leq n^{-\nu}G^2$, there exist $\omega, \widehat{\omega} > 0$ with $\nu \ll \widehat{\omega} \ll \omega \ll \omega_f$, and such that it holds*

$$\big|\rho_{\mathrm{fc},t_f}(0)s_i(t_f) - \rho_{\mathrm{sc}}(0)r_i(t_f)\big| \leq n^{-1-\omega}, \qquad |i| \leq n^{\widehat{\omega}}, \tag{7.49}$$

with very high probability, where $t_f := n^{-1+\omega_f}$.

The proof of Proposition 7.14 is postponed to Section 7.6.

Remark 7.15. *Note that, without loss of generality, it is enough to prove Proposition 7.14 only for the case $\rho_{\mathrm{fc},t_f}(0) = \rho_{sc}(0)$, since we can always rescale the time: we may define $\widetilde{s}_i := (\rho_{\mathrm{fc},t_f}(0)s_i/\rho_{sc}(0))$ and notice that $\widetilde{s}_i(t)$ is a solution of the DBM (7.44) after rescaling as $t' = (\rho_{\mathrm{fc},t_f}(0)/\rho_{sc}(0))^2 t$.*

7.5. Proof of Lemma 7.6 and Lemma 7.7. In this section we prove that by Lemma 7.9 and Proposition 7.14 Lemmas 7.6–7.7 follow.

7.5.1. *Application of Proposition 7.14 to $\boldsymbol{\lambda}^{z_l}(t)$ and $\tilde{\boldsymbol{\lambda}}^{(l)}(t)$.* In this section we prove that for any fixed l the processes $\boldsymbol{\lambda}^{z_l}(t)$ and $\widetilde{\boldsymbol{\lambda}}^{(l)}(t)$ satisfy Assumption 7.11, Definition 7.12 and so that by Proposition 7.14 we conclude the lemma.

Proof of Lemma 7.6. For any fix $l \in [p]$, by the definition of the driving Brownian motions of the processes (7.15) and (7.20) it is clear that they satisfy Assumption 7.11 choosing $\boldsymbol{s}(t) = \boldsymbol{\lambda}^{z_l}(t)$, $\boldsymbol{r}(t) = \widetilde{\boldsymbol{\lambda}}^{(l)}(t)$, and $K = n^{\omega_A}$, since $L_{ij}(t) \equiv 0$ for $|i|, |j| \leq K$.

We now show that the set of points $\{\lambda_{\pm i}^{z_l}\}_{i=1}^n$, rescaled by $\rho^{z_l}(0)/\rho_{sc}(0)$, is (g, G)-*regular* for

$$g = n^{-1+\omega_h}\delta_l^{-100}, \qquad G = n^{-\omega_h}\delta_l^{10}, \qquad \nu = \omega_h. \tag{7.50}$$

with $\delta_l := 1 - |z_l|^2$, for any $l \in [p]$. By the local law (7.3), together with the regularity properties of m^{z_l} which follow by (7.9), namely that m^{z_l} is $1/3$-Hölder continuous, we conclude that there exist constants $c_{\omega_h}, C_{\omega_h} > 0$ such that

$$c_{\omega_h} \leq \Im \frac{1}{2n} \sum_{i=-n}^n \frac{1}{[\rho^{z_l}(0)\lambda_i^{z_l}/\rho_{sc}(0)] - (E + \mathrm{i}\eta)} \leq C_{\omega_h}, \tag{7.51}$$

for any $|E| \leq n^{-\omega_h}\delta_l^{10}$, $n^{-1}\delta_l^{-100} \leq \eta \leq 10$. In particular, $c_{\omega_h}, C_{\omega_h} \sim 1$ for $\eta \in [g, n^{-2\omega_h}]$, and $c_{\omega_h} \gtrsim n^{-100\omega_h}$, $C_{\omega_h} \lesssim n^{100\omega_h}$ for $\eta \in [n^{-2\omega_h}, 10]$. This implies that the set $\boldsymbol{\lambda}^{z_l} = \{\lambda_{\pm i}^{z_l}\}_{i=1}^n$ satisfies Definition 7.12 and it concludes the proof of this lemma. □

7.5.2. *Application of Proposition 7.14 to $\boldsymbol{\mu}^{(l)}(t)$ and $\tilde{\boldsymbol{\mu}}^{(l)}(t)$.* We now prove that for any fixed l the processes $\boldsymbol{\mu}^{(l)}(t)$ and $\widetilde{\boldsymbol{\mu}}^{(l)}(t)$ satisfy Assumption 7.11, Definition 7.12 and so that by Proposition 7.14 we conclude the lemma.

Proof of Lemma 7.7. For any fixed $l \in [p]$, we will apply Proposition 7.14 with the choice $\boldsymbol{s}(t) = \boldsymbol{\mu}^{(l)}(t)$, $\boldsymbol{r}(t) = \widetilde{\boldsymbol{\mu}}^{(l)}(t)$ and $K = n^{\omega_A}$. Since the initial data $s_i(0) = \mu_i^{(l)}(0)$ are the singular values of a Ginibre matrix $X^{(l)}$, it is clear that the assumption in Definition 7.12 holds choosing $g = n^{-1+\delta}$ and $G = n^{-\delta}$, and $\nu = 0$, for any small $\delta > 0$ (see e.g. the local law in (7.3)).

We now check Assumption 7.11. By the definition of the families of i.i.d. Brownian motions

$$\left(\{\zeta_{\pm i}^{z_l}\}_{i=1}^{n^{\omega_A}}, \{\widetilde{\zeta}_{\pm i}^{(l)}\}_{i=n^{\omega_A}+1}^n\right)_{l=1}^p, \qquad \left(\{\beta_{\pm i}^{(l)}\}_{i=1}^n\right)_{l=1}^p, \tag{7.52}$$

defined in (7.21) and (7.18), respectively, it immediately follows that they satisfy (a) and (b) of Assumption 7.11, since $\{\widetilde{\zeta}_{\pm i}^{(l)}\}_{i=n^{\omega_A}+1}^n$ are independent of $\{\beta_{\pm i}^{(l)}\}_{i=1}^n$ as well as $\{\beta_{\pm i}^{(l)}\}_{i=n^{\omega_A}+1}^n$ are independent of $\{\widetilde{\zeta}_{\pm i}^{(l)}\}_{i=1}^n$ by construction. Recall that $\mathcal{F}_{\beta,t}$ denotes the common filtration of all the Brownian motions $\boldsymbol{\beta}^{(m)} = \{\beta_i^{(m)}\}_{i=1}^n, m \in [p]$.

Finally, we prove that also (c) of Assumption 7.11 is satisfied. We recall the relations $i = \mathrm{i} + (l-1)n^{\omega_A}$ and $j = \mathrm{j} + (l-1)n^{\omega_A}$ from (7.24) which, for any fixed l, establish a one to one relation between a pair $\mathrm{i}, \mathrm{j} \in [n^{\omega_B}]$ and a pair i, j with $(l-1)n^{\omega_A} + 1 \leq i, j \leq ln^{\omega_A}$. By the definition of $\{\zeta_{\pm i}^{z_l}\}_{i=1}^{n^{\omega_A}}$ it follows that

$$\mathrm{d}\zeta_{\mathrm{i}}^{z_l} - \mathrm{d}\beta_{\mathrm{i}}^{(l)} = \sum_{m=1}^{pn^{\omega_A}} \left(\sqrt{C^\#(t)} - I\right)_{im} \mathrm{d}(\underline{\beta})_m, \qquad 1 \leq \mathrm{i} \leq n^{\omega_A}, \tag{7.53}$$

with $\underline{\beta}$ defined in (7.22), and so that for any $1 \leq \mathrm{i}, \mathrm{j} \leq n^{\omega_A}$ and fixed l we have

$$\mathbf{E}\left[\left(\mathrm{d}\zeta_{\mathrm{i}}^{z_l} - \mathrm{d}\beta_{\mathrm{i}}^{(l)}\right)\left(\mathrm{d}\zeta_{\mathrm{j}}^{z_l} - \mathrm{d}\beta_{\mathrm{j}}^{(l)}\right) \,\Big|\, \mathcal{F}_{\beta,t}\right]$$

$$= \sum_{m_1, m_2=1}^{pn^{\omega_A}} \left(\sqrt{C^\#(t)} - I\right)_{im_1}\left(\sqrt{C^\#(t)} - I\right)_{jm_2} \mathbf{E}\left[\mathrm{d}(\underline{\beta})_{m_1} \mathrm{d}(\underline{\beta})_{m_2} \,\big|\, \mathcal{F}_{\beta,t}\right]$$

$$= \left[\left(\sqrt{C^\#(t)} - I\right)^2\right]_{ij} \mathrm{d}t,$$

since $\sqrt{C^\#(t)}$ is real symmetric. Hence, $L_{ij}(t)$ defined in (7.46) in this case is given by

$$L_{ij}(t) = \left[\left(\sqrt{C^\#(t)} - I \right)^2 \right]_{ij}.$$

Then, by Cauchy-Schwarz inequality, we have that

$$\begin{aligned}
|L_{ij}(t)| &\leq \left[\left(\sqrt{C^\#(t)} - I \right)^2 \right]_{ii}^{1/2} \left[\left(\sqrt{C^\#(t)} - I \right)^2 \right]_{jj}^{1/2} \\
&\leq \mathrm{Tr} \left[(\sqrt{C^\#(t)} - I)^2 \right] \leq \mathrm{Tr} \left[(C^\#(t) - I)^2 \right] \lesssim \frac{p^2 n^{2\omega_A}}{n^{4\omega_E}},
\end{aligned} \tag{7.54}$$

with very high probability, where in the last inequality we used that $C^\#(t)$ and $C(t)$ have the same distribution and the bound (7.42) of Lemma 7.9 holds for $C(t)$ hence for $C^\#(t)$ as well. This implies that for any fixed $l \in [p]$ the two families of Brownian motions $\{\beta_{\pm i}^{(l)}\}_{i=1}^n$ and $(\{\zeta_{\pm i}^{z_l}\}_{i=1}^{n^{\omega_A}}, \{\widetilde{\zeta}_{\pm i}^{(l)}\}_{i=n^{\omega_A}+1}^n)$ satisfy Assumption 7.11 with $K = n^{\omega_A}$ and $\omega_Q = 4\omega_E - 2\omega_A$. Applying Proposition 7.14 this concludes the proof of Lemma 7.7. □

7.6. Proof of Proposition 7.14.

We divide the proof of Proposition 7.14 into four sub-sections. In Section 7.6.1 we introduce an interpolating process $\boldsymbol{x}(t, \alpha)$ between the processes $\boldsymbol{s}(t)$ and $\boldsymbol{r}(t)$ defined in (7.44)–(7.45), and in Section 7.6.2 we introduce a measure which approximates the particles $\boldsymbol{x}(t, \alpha)$ and prove their rigidity. In Section 7.6.3 we introduce a cut-off near zero (this scale will be denoted by ω_A later) such that we only couple the dynamics of the particles $|i| \leq n^{\omega_A}$, as defined in (c) of Assumption 7.11, i.e. we will choose $\omega_A = \omega_K$. Additionally, we also localise the dynamics on a scale ω_l (see Section 7.2.3) since the main contribution to the dynamics comes from the nearby particles. We will refer to the new process $\widehat{\boldsymbol{x}}(t, \alpha)$ (see (7.68) later) as the *short range approximation* of the process $\boldsymbol{x}(t, \alpha)$. Finally, in Section 7.6.4 we conclude the proof of Proposition 7.14.

Large parts of our proof closely follow [26, 59] and for brevity we will focus on the differences. We use [26, 59] as our main references since the 2×2 block matrix setup of [26] is very close to the current one and [26] itself closely follows [59]. However, we point out that many key ideas of this technique have been introduced in earlier papers on universality; e.g. short range cut-off and finite speed of propagation in [42, 18], coupling and homogenisation in [20]; for more historical references, see [59]. The main novelty of [59] itself is a mesoscopic analysis of the fundamental solution $p_t(x, y)$ of (7.79) which enables the authors to prove short time universality for general deterministic initial data. They also proved the result with very high probability unlike [20] that relied on level repulsion estimates. We also mention a related but different more recent technique to prove universality [19], which has been recently adapted to the singular values setup, or equivalently to the 2×2 block matrix structure, in [85].

7.6.1. Definition of the interpolated process.

For $\alpha \in [0, 1]$ we introduce the continuous interpolation process $\boldsymbol{x}(t, \alpha)$, between the processes $\boldsymbol{s}(t)$ and $\boldsymbol{r}(t)$ in (7.44)–(7.45), defined as the solution of the flow

$$\mathrm{d}x_i(t, \alpha) = \alpha \frac{\mathrm{d}\mathfrak{b}_i^s}{\sqrt{2n}} + (1 - \alpha) \frac{\mathrm{d}\mathfrak{b}_i^r}{\sqrt{2n}} + \frac{1}{2n} \sum_{j \neq i} \frac{1}{x_i(t, \alpha) - x_j(t, \alpha)} \, \mathrm{d}t, \tag{7.55}$$

with initial data

$$\boldsymbol{x}(0, \alpha) = \alpha \boldsymbol{s}(t_0) + (1 - \alpha) \boldsymbol{r}(t_0), \tag{7.56}$$

with some t_0 that is a slightly smaller than t_f. In fact we will write $t_0 + t_1 = t_f$ with $t_1 \ll t_f$, where t_1 is the time scale for the equilibration of the DBM with initial condition (7.56) (see (7.64)). To make our notation consistent with [26, 59] in the remainder of this section we assume that $t_0 = n^{-1+\omega_0}$, for some small $\omega_0 > 0$, such that $\omega_K \ll \omega_0 \ll \omega_Q$. The reader can think of $\omega_0 = \omega_f$. Note that the strong solution of (7.55) is well defined since the variance of its driving Brownian motion is smaller than $\frac{1}{2n}(1 - 2\alpha(1 - \alpha)n^{-\omega_Q})$ by (7.47), which is below the critical variance for well-posedness of the DBM since we are in the complex symmetry class (see e.g. [7, Lemma 4.3.3]).

By (7.55) it clearly follows that $\boldsymbol{x}(t, 0) = \boldsymbol{r}(t + t_0)$ and $\boldsymbol{x}(t, 1) = \boldsymbol{s}(t + t_0)$, for any $t \geq 0$. Note that the process (7.55) is almost the same as [59, Eq. (3.13)], [26, Eq. (3.13)], except for the stochastic term, which in our case depends on α. Also, to make the notation clearer, we remark that in [26, 59] the

interpolating process is denoted by $\boldsymbol{z}(t, \alpha)$. We changed this notation to $\boldsymbol{x}(t, \alpha)$ to avoid confusions with the z_l-parameters introduced in the previous sections where we apply Proposition 7.14 to the processes defined in Section 7.2.1.

Remark 7.16. *Even if all processes $\boldsymbol{\lambda}(t)$, $\widetilde{\boldsymbol{\lambda}}(t)$, $\widetilde{\boldsymbol{\mu}}(t)$, $\boldsymbol{\mu}(t)$ introduced in Section 7.2.1 already satisfy [26, Lemma 3.3-3.5], [59, Lemma 3.3-3.5] as a consequence of the local law (7.3) and the rigidity estimates (7.4), we decided to present the proof of Proposition 7.14 for general deterministic initial data $\mathbf{s}(0)$ satisfying Definition 7.12 (see Remark 7.10). Hence, an additional time t_0 is needed to ensure the validity of [26, Lemma 3.3-3.5], [59, Lemma 3.3-3.5]. More precisely, we first let the DBMs (7.44)–(7.45) evolve for a time $t_0 := n^{-1+\omega_0}$, and then we consider the process (7.55) whose initial data in (7.56) is given by a linear interpolation of the solutions of (7.44)–(7.45) at time t_0.*

Before proceeding with the analysis of (7.55) we give some definitions and state some preliminary results necessary for its analysis.

7.6.2. *Interpolating measures and particle rigidity.* Using the convention of [26, Eq. (3.10)–(3.11)], given a probability measure $\mathrm{d}\rho(E)$, we define the $2n$-quantiles γ_i by

$$
\begin{aligned}
\gamma_i &:= \inf\left\{ x \;\middle|\; \int_{-\infty}^{x} \mathrm{d}\rho(E) \geq \frac{n+i-1}{2n} \right\}, \quad 1 \leq i \leq n, \\
\gamma_i &:= \inf\left\{ x \;\middle|\; \int_{-\infty}^{x} \mathrm{d}\rho(E) \geq \frac{n+i}{2n} \right\}, \qquad -n \leq i \leq -1,
\end{aligned}
\tag{7.57}
$$

Note that $\gamma_1 = 0$ if $\mathrm{d}\rho(E)$ is symmetric with respect to 0.

Let $\rho_{fc,t}(E)$ be defined above Proposition 7.14 (see e.g. [59, Eq. (2.5)–(2.6)] for more details), and let $\rho_{sc}(E)$ denote the semicircular density, then by $\gamma_i(t)$, γ_i^{sc} we denote the $2n$-quantiles, defined as in (7.57), of $\rho_{fc,t}$ and ρ_{sc}, respectively.

Following the construction of [59, Lemma 3.3-3.4, Appendix A], [26, Section 3.2.1], we define the interpolating (random) measure $\mathrm{d}\rho(E, t, \alpha)$ for any $\alpha \in [0, 1]$. More precisely, the measure $\mathrm{d}\rho(E, t, \alpha)$ is deterministic close to zero, and it consists of delta functions of the position of the particles $x_i(t, \alpha)$ away from zero.

Denote by $\gamma_i(t, \alpha)$ the quantiles of $\mathrm{d}\rho(E, \alpha, t)$, and by $m(w, t, \alpha)$, with $w \in \mathbf{H}$, its Stieltjes transform. Fix $q_* \in (0, 1)$ throughout this section, and let $k_0 = k_0(q_*) \in \mathbf{N}$ be the largest index such that

$$
|\gamma_{\pm k_0}(t_0)|, |\gamma_{\pm k_0}^{sc}| \leq q_* G,
\tag{7.58}
$$

with G defined in (7.50), then the measure $\mathrm{d}\rho(E, t, \alpha)$ has a deterministic density (denoted by $\rho(E, \alpha, t)$ with a slight abuse of notation) on the interval

$$
\mathcal{G}_\alpha := [\alpha \gamma_{-k_0}(t_0) + (1-\alpha)\gamma_{-k_0}^{sc}, \alpha \gamma_{k_0}(t_0) + (1-\alpha)\gamma_{k_0}^{sc}].
\tag{7.59}
$$

Outside $\mathcal{G}_\alpha$ the measure $\mathrm{d}\rho(E, t, \alpha)$ consists of $1/(2n)$ times delta functions of the particle locations $\delta_{x_i(t,\alpha)}$.

Remark 7.17. *By the construction $\mathrm{d}\rho(E, t, \alpha)$ as in [59, Lemma 3.3-3.4, Appendix A], [26, Section 3.2.1] all the regularity properties of $\mathrm{d}\rho(E, \alpha, t)$, its quantiles $\gamma_i(t, \alpha)$, and its Stieltjes transform $m(E + \mathrm{i}\eta, t, \alpha)$ in [59, Lemma 3.3-3.4], [26, Lemma 3.3-3.4] hold without any change. In particular, it follows that*

$$
|\gamma_i(t, \alpha) - \gamma_j(t, \alpha)| \sim \frac{|i-j|}{n}, \qquad |i|, |j| \leq q_* G,
\tag{7.60}
$$

with q_* defined above (7.58), and G in (7.50).

Define the Stieltjes transform of the empirical measure of the particle configuration $\{x_{\pm i}(t, \alpha)\}_{i=1}^{n}$ by

$$
m_n(w, t, \alpha) := \frac{1}{2n} \sum_{i=-n}^{n} \frac{1}{x_i(t, \alpha) - w}, \quad w \in \mathbf{H}.
\tag{7.61}
$$

We recall that the summation does not include the term $i = 0$ (see Remark 7.1). The local law for $m_n(w, t, \alpha)$, hence rigidity for the interpolated particles (7.55), easily follows similarly to [52, Section

3.2]; the only minor difference is that the driving martingales in (7.55) are not independent for different indices (see also [31, Lemma 7.12] when a similar proof has been presented with more details).

Lemma 7.18. *Fix* $q \in (0,1)$ *and* $\tilde{\epsilon} > 0$. *Define* $\widehat{C}_q := \{j : |j| \leq qk_0\}$, *with* k_0 *defined in* (7.58). *Then for any* $\xi > 0$, *with very high probability we have the optimal rigidity*

$$\sup_{0 \leq t \leq t_0 n^{-\tilde{\epsilon}}} \sup_{i \in \widehat{C}_q} \sup_{0 \leq \alpha \leq 1} |x_i(t,\alpha) - \gamma_i(t,\alpha)| \leq \frac{n^{\xi+100\nu}}{n}, \tag{7.62}$$

and the local law

$$\sup_{n^{-1+\tilde{\epsilon}} \leq \eta \leq 10} \sup_{0 \leq t \leq t_0 n^{-\tilde{\epsilon}}} \sup_{0 \leq \alpha \leq 1} \sup_{E \in q\mathcal{G}_\alpha} |m_n(E + \mathrm{i}\eta, t, \alpha) - m(E + \mathrm{i}\eta, t, \alpha)| \leq \frac{n^{\xi+100\nu}}{n\eta}, \tag{7.63}$$

for sufficiently large n, *with* $\nu > 0$ *from Definition* 7.12.

Without loss of generality in Lemma 7.18 we assumed $k_1 = k_0$ in [26, Eq. (3.25)–(3.26)].

7.6.3. *Short range analysis.* In the following of this section we perform a local analysis of (7.55) adapting the analysis of [26, 59] and explaining the minor changes needed for the analysis of the flow (7.55), for which the driving Brownian motions $\mathfrak{b}^s$, $\mathfrak{b}^r$ satisfy Assumption 7.11, compared to the analysis of [26, Eq. (3.13)], [59, Eq. (3.13)]. More precisely, we run the DBM (7.55) for a time

$$t_1 := \frac{n^{\omega_1}}{n}, \tag{7.64}$$

for any $\omega_1 > 0$ such that $\nu \ll \omega_1 \ll \omega_K$, with ν, ω_K defined in Definition 7.12 and above Assumption 7.11, respectively, so that (7.55) reaches its local equilibrium (see Section 7.2.3 for a summary on the different scales). Moreover, since the dynamics of $x_i(t,\alpha)$ is mostly influenced by the particles close to it, in the following we define a short range approximation of the process $\boldsymbol{x}(t,\alpha)$ (see (7.68) later), denoted by $\widehat{\boldsymbol{x}}(t,\alpha)$, and use the homogenisation theory developed in [59], adapted in [26] for the singular values flow, for the short range kernel.

Remark 7.19. *We do not need to define the shifted process* $\widetilde{\boldsymbol{x}}(t,\alpha)$ *as in* [26, Eq. (3.29)–(3.32)] *and* [59, Eq. (3.36)–(3.40)], *since in our case the measure* $\mathrm{d}\rho(E,t,\alpha)$ *is symmetric with respect to* 0 *by assumption, hence, using the notation in* [26, Eq. (3.29)–(3.32)], *we have* $\widetilde{\boldsymbol{x}}(t,\alpha) = \boldsymbol{x}(t,\alpha) - \gamma_1(t,\alpha) = \boldsymbol{x}(t,\alpha)$. *Hence, from now on we only use* $\boldsymbol{x}(t,\alpha)$ *and the reader can think* $\widetilde{\boldsymbol{x}}(t,\alpha) \equiv \boldsymbol{x}(t,\alpha)$ *for a direct analogy with* [26, 59].

Our analysis will be completely local, hence we introduce a short range cut-off. Fix $\omega_l, \omega_A > 0$ so that

$$0 < \omega_1 \ll \omega_l \ll \omega_A \ll \omega_0 \ll \omega_Q, \tag{7.65}$$

with ω_1 defined in (7.64), ω_0 defined below (7.56), and ω_Q in (c) of Assumption 7.11. Moreover, we assume that ω_A is such that

$$K_n = n^{\omega_A}, \tag{7.66}$$

with $K_n = n^{\omega_K}$ in Assumption 7.11, i.e. $\omega_A = \omega_K$. We remark that it is enough to choose $\omega_A \ll \omega_K$, but to avoid further splitting in (7.68) we assumed $\omega_K = \omega_A$.

For any $q \in (0,1)$, define the set

$$A_q := \left\{ (i,j) \,\middle|\, |i-j| \leq n^{\omega_l} \text{ or } ij > 0, i \notin \widehat{C}_q, j \notin \widehat{C}_q \right\}, \tag{7.67}$$

and denote $A_{q,(i)} := \{j \,|\, (i,j) \in A_q\}$. In the remainder of this section we will often use the notations

$$\sum_{j}^{A_{q,(i)}} := \sum_{j \in A_{q,(i)}}, \qquad \sum_{j}^{A_{q,(i)}^c} := \sum_{j \notin A_{q,(i)}}.$$

Let $q_* \in (0,1)$ be defined above (7.58), then we define the short range process $\widehat{\boldsymbol{x}}(t,\alpha)$ (cf. [26, Eq. (3.35)–(3.36)], [59, Eq. (3.45)–(3.46)]) as follows

$$
\mathrm{d}\widehat{x}_i(t,\alpha) = \frac{1}{2n} \sum_j^{A_{q_*,(i)}} \frac{1}{\widehat{x}_i(t,\alpha) - \widehat{x}_j(t,\alpha)} \, \mathrm{d}t
$$
$$
+ \begin{cases} \alpha\frac{\mathrm{d}\mathfrak{b}^s}{\sqrt{2n}} + (1-\alpha)\frac{\mathrm{d}\mathfrak{b}^r}{\sqrt{2n}} & \text{if } |i| \le n^{\omega_A}, \\ \alpha\frac{\mathrm{d}\mathfrak{b}^s}{\sqrt{2n}} + (1-\alpha)\frac{\mathrm{d}\mathfrak{b}^r}{\sqrt{2n}} + J_i(\alpha,t)\,\mathrm{d}t & \text{if } n^{\omega_A} < |i| \le n, \end{cases}
\tag{7.68}
$$

where

$$
J_i(\alpha,t) := \frac{1}{2n} \sum_j^{A^c_{q_*,(i)}} \frac{1}{x_i(t,\alpha) - x_j(t,\alpha)},
\tag{7.69}
$$

and initial data $\widehat{\boldsymbol{x}}(0,\alpha) = \boldsymbol{x}(0,\alpha)$. Note that

$$
\sup_{0 \le t \le t_1} \sup_{0 \le \alpha \le 1} |J_1(\alpha,t)| \le \log n,
\tag{7.70}
$$

with very high probability.

Remark 7.20. *Note that the SDE defined in* (7.68) *has the same form as in* [59, Eq. (3.70)], *with* $F_i = 0$ *in our case, except for the stochastic term in* (7.68) *that looks slightly different, in particular it depends on* α. *Nevertheless, by Assumption* 7.11, *the quadratic variation of the driving Brownian motions in* (7.68) *is also bounded by one uniformly in* $\alpha \in [0,1]$. *Moreover, the process defined in* (7.68) *and the measure* $\mathrm{d}\rho(E,t,\alpha)$ *satisfy* [59, Eq. (3.71)–(3.77)].

Since when we consider the difference process $\widehat{\boldsymbol{x}}(t,\alpha) - \boldsymbol{x}(t,\alpha)$ the stochastic differential disappears, by [59, Lemma 3.8], without any modification, it follows that

$$
\sup_{0 \le t \le t_1} \sup_{0 \le \alpha \le 1} \sup_{|i| \le n} |\widehat{x}_i(t,\alpha) - x_i(t,\alpha)| \le n^{\xi+100\nu} t_1 \left(\frac{1}{n^{\omega_l}} + \frac{n^{\omega_A}}{n^{\omega_0}} + \frac{1}{\sqrt{nG}} \right),
\tag{7.71}
$$

for any $\xi > 0$ with very high probability, with G defined in (7.50). In particular, (7.71) implies that the short range process $\widehat{\boldsymbol{x}}(t,\alpha)$, defined in (7.68), approximates very well (i.e. they are closer than the fluctuation scale) the process $\boldsymbol{x}(t,\alpha)$ defined in (7.55).

Next, in order to use the smallness of (7.46)–(7.47) in Assumption 7.11 for $|i| \le n^{\omega_A}$, we define $\boldsymbol{u}(t,\alpha) := \partial_\alpha \widehat{\boldsymbol{x}}(t,\alpha)$, which is the solution of the following discrete SPDE (cf. [26, Eq. (3.38)], [59, Eq. (3.63)]):

$$
\mathrm{d}\boldsymbol{u} = \sum_j^{A_{q_*,(i)}} B_{ij}(u_j - u_i)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\xi}_1 + \boldsymbol{\xi}_2\,\mathrm{d}t = -\mathcal{B}\boldsymbol{u}\,\mathrm{d}t + \mathrm{d}\boldsymbol{\xi}_1 + \boldsymbol{\xi}_2\,\mathrm{d}t,
\tag{7.72}
$$

where

$$
B_{ij} := \frac{\mathbf{1}_{j \ne \pm i}}{2n(\widehat{x}_i - \widehat{x}_j)^2}, \quad \mathrm{d}\xi_{1,i} := \frac{\mathrm{d}\mathfrak{b}^s_i}{\sqrt{2n}} - \frac{\mathrm{d}\mathfrak{b}^r_i}{\sqrt{2n}}
$$
$$
\xi_{2,i} := \begin{cases} 0 & \text{if } |i| \le n^{\omega_A}, \\ \partial_\alpha J_i(\alpha,t) & \text{if } n^{\omega_A} < |i| \le n, \end{cases}
\tag{7.73}
$$

with $J_i(\alpha,t)$ defined in (7.69). We remark that the operator[5] $\mathcal{B}$ defined via the kernel in (7.73) depends on α and t. It is not hard to see (e.g. see [59, Eq. (3.65), Eq. (3.68)–(3.69)]) that the forcing term $\boldsymbol{\xi}_2$ is bounded with very high probability by n^C, for some $C > 0$, for $n^{\omega_A} < |i| \le n$. Note that the only difference in (7.72) compared to [26, Eq. (3.38)], [59, Eq. (3.63)] is the additional term $\mathrm{d}\boldsymbol{\xi}_1$ which will be negligible for our analysis.

Let $\mathcal{U}$ be the semigroup associated to $\mathcal{B}$, i.e. if $\partial_t \boldsymbol{v} = -\mathcal{B}\boldsymbol{v}$, then for any $0 \le s \le t$ we have that

$$
v_i(t) = \sum_{j=-n}^n \mathcal{U}_{ij}(s,t,\alpha) v_j(s), \qquad |i| \le n.
$$

---

[5]The operator $\mathcal{B}$ defined here is not to be confused with the completely unrelated one in (6.1).

The first step to analyse the equation in (7.72) is the following finite speed of propagation estimate (cf. [26, Lemma 3.9], [59, Lemma 3.7]).

Lemma 7.21. *Let $0 \leq s \leq t \leq t_1$. Fix $0 < q_1 < q_2 < q_*$, with $q_* \in (0,1)$ defined in (7.58), and $\epsilon_1 > 0$ such that $\epsilon_1 \ll \omega_A$. Then for any $\alpha \in [0,1]$ we have*

$$|U_{ji}(s,t,\alpha)| + |U_{ij}(s,t,\alpha)| \leq n^{-D}, \qquad (7.74)$$

for any $D > 0$ with very high probability, if either $i \in \widehat{C}_{q_2}$ and $|i - j| > n^{\omega_l + \epsilon_1}$, or if $i \notin \widehat{C}_{q_2}$ and $j \in \widehat{C}_{q_1}$.

Proof. The proof of this lemma follows the same lines as [59, Lemma 3.7]. There are only two differences that we point out. The first one is that [59, Eq. (4.15)], using the notation therein, has to be replaced by

$$\sum_k v_k^2 (\nu^2(\psi_k')^2 + \nu\psi_k'')\, \mathbf{E}[\mathrm{d}C_k(\alpha,t)\,\mathrm{d}C_k(\alpha,t) \mid \mathcal{F}_t], \qquad (7.75)$$

where $\mathcal{F}_t$ is the filtration defined in Assumption 7.11, and $C_k(\alpha,t)$ is defined as

$$C_k(\alpha,t) := \alpha \frac{\mathfrak{b}_k^s(t)}{\sqrt{2n}} + (1-\alpha)\frac{\mathfrak{b}_k^r(t)}{\sqrt{2n}}. \qquad (7.76)$$

We remark that ν in (7.75) should not to be confused with ν in Definition 7.12. Then, by Kunita-Watanabe inequality, it is clear that

$$\mathbf{E}[\mathrm{d}C_k(\alpha,t)\,\mathrm{d}C_k(\alpha,t) \mid \mathcal{F}_t] \lesssim \frac{\mathrm{d}t}{n}, \qquad (7.77)$$

uniformly in $|k| \leq n$, $t \geq 0$, and $\alpha \in [0,1]$. The fact that (7.77) holds is the only input needed to bound [59, Eq. (4.21)].

The second difference is that the stochastic differential $(\sqrt{2}\,\mathrm{d}B_k)/\sqrt{n}$ in [59, Eq. (4.21)] has to be replaced by $\mathrm{d}C_k(\alpha,t)$ defined in (7.76). This change is inconsequential in the bound [59, Eq. (4.26)], since $\mathbf{E}\,\mathrm{d}C_k(\alpha,t) = 0$. $\square$

Moreover, the result in [26, Lemma 3.8], [59, Lemma 3.10] hold without any change, since its proof is completely deterministic and the stochastic differential in the definition of the process $\widehat{\boldsymbol{x}}(t,\alpha)$ does not play any role.

In the remainder of this section, before completing the proof of Proposition 7.14, we describe the homogenisation argument to approximate the t-dependent kernel of $\mathcal{B}$ with a continuous kernel (denoted by $p_t(x,y)$ below). We follow verbatim [59, Section 3-4] and its adaptation to the singular value flow of [26, Section 3.4], except for the bound of the rhs. of (**??**), where we handle the additional term $\mathrm{d}\boldsymbol{\xi}_1$ in (7.73).

Fix a constant $\epsilon_B > 0$ such that $\omega_A - \epsilon_B > \omega_l$, and let $a \in \mathbf{Z}$ be such that $0 < |a| \leq n^{\omega_A - \epsilon_B}$. Define also the equidistant points $\gamma_j^f := j(2n\rho_{sc}(0))^{-1}$, which approximate the quantiles $\gamma_j(t,\alpha)$ very well for small j, i.e. $|\gamma_j^f - \gamma_j(t,\alpha)| \lesssim n^{-1}$ for $|j| \leq n^{\omega_0/2}$ (see [59, Eq. (3.91)]). Consider the solution of

$$\partial_t w_i = -(\mathcal{B}w)_i, \quad w_i(0) = 2n\delta_{ia}, \qquad (7.78)$$

and define the cut-off $\eta_l := n^{\omega_l}(2n\rho_{sc}(0))^{-1}$. Let $p_t(x,y)$ be the fundamental solution of the equation

$$\partial_t f(x) = \int_{|x-y| \leq \eta_l} \frac{f(y) - f(x)}{(x-y)^2} \rho_{sc}(0)\,\mathrm{d}y. \qquad (7.79)$$

The idea of the homogenisation argument is that the deterministic solution f of (7.79) approximates very well the random solution of (7.78). This is formulated in terms of the solution kernels of the two equations in Proposition 7.22. Following [26, Lemma 3.9-3.13, Corollary 3.14, Theorem 3.15-3.17], which are obtained adapting the proof of [59, Section 3.6], we will conclude the following proposition.

Proposition 7.22. *Let $a, i \in \mathbf{Z}$ such that $|a| \leq n^{\omega_A - \epsilon_B}$ and $|i - a| \leq n^{\omega_l}/10$. Fix $\epsilon_c > 0$ such that $\omega_1 - \epsilon_c > 0$, let $t_1 := n^{-1+\omega_1}$ and $t_2 := n^{-\epsilon_c}t_1$, then for any $\alpha \in [0,1]$ and for any $|u| \leq t_2$ we have*

$$\left| \mathcal{U}_{ia}(0, t_1 + u, \alpha) - \frac{p_{t_1}(\gamma_i^f, \gamma_a^f)}{n} \right| \leq \frac{n^{100\nu + \epsilon_c}}{nt_1} \left(\frac{(nt_1)^2}{n^{\omega_l}} + \frac{1}{(nt_1)^{1/10}} + \frac{1}{n^{3\epsilon_c/2}} \right), \qquad (7.80)$$

with very high probability.

Proof. The proof of this proposition relies on [59, Section 3.6], which has been adapted to the 2×2 block structure in [26, Lemma 3.9–3.13, Corollary 3.14, Theorem 3.15–3.17]. We thus present only the differences compared to[26, 59]; for a complete proof we defer the reader to these works.

The only difference in the proof of this proposition compared to the proof of [26, Theorem 3.17], [59, Theorem 3.11] is in [59, Eq. (3.121) of Lemma 3.14] and [59, Eq. (3.148) of Lemma 3.14]. The main goal of [59, Lemma 3.14] and [59, Lemma 3.14] is to prove that

$$\mathrm{d}\frac{1}{2n}\sum_{1\leq|i|\leq n}(w_i - f_i)^2 = -\langle \boldsymbol{w}(t) - \boldsymbol{f}(t), \mathcal{B}(\boldsymbol{w}(t) - \boldsymbol{f}(t))\rangle + \text{Lower order}, \qquad (7.81)$$

where $f_i := f(\widehat{x}_i(t,\alpha), t)$, with $\widehat{x}_i(t,\alpha)$ being the solution of (7.68), and $\boldsymbol{w}(t)$, $\boldsymbol{f}(t)$ being the solutions of (7.78) and (7.79) with $x = \widehat{x}_i(t,\alpha)$, respectively. In order to prove (7.81), following [59, Eq. (3.121)] and using the notation therein (with $N = 2n$ and replacing $\widehat{z}_i$ by $\widehat{x}_i$), we compute

$$\mathrm{d}\frac{1}{2n}\sum_{1\leq|i|\leq n}(w_i - f_i)^2$$
$$= \frac{1}{n}\sum_{1\leq|i|\leq n}(w_i - f_i)\left[\partial_t w_i\,\mathrm{d}t - (\partial_t f)(t,\widehat{x}_i)\,\mathrm{d}t - f'(t,\widehat{x}_i)\,\mathrm{d}\widehat{x}_i\right] \qquad (7.82)$$
$$+ \frac{1}{n}\sum_{1\leq|i|\leq n}\left(-(w_i - f_i)f''(t,\widehat{x}_i) + (f'(t,\widehat{x}_i))^2\right)\mathbf{E}[\mathrm{d}C_i(\alpha,t)\,\mathrm{d}C_i(\alpha,t)\,|\,\mathcal{F}_t],$$

where

$$C_i(\alpha,t) := \alpha\frac{\mathfrak{b}_i^s(t)}{\sqrt{2n}} + (1-\alpha)\frac{\mathfrak{b}_i^r(t)}{\sqrt{2n}}.$$

As a consequence of the slight difference in definition of $\mathrm{d}\widehat{x}_i$ in (7.68), compared to the definition of $\mathrm{d}\widehat{z}_i$ in [59, Eq. (3.70)], the martingale term in (7.82) is given by (cf. [59, Eq. (3.148)])

$$\mathrm{d}M_t = \frac{1}{2n}\sum_{1\leq|i|\leq n}(w_i - f_i)f_i'\,\mathrm{d}C_i(\alpha,t). \qquad (7.83)$$

The terms in the first line of the rhs. of (7.82) are bounded exactly as in [59, Eq. (3.124)–(3.146), (3.149)–(3.154)]. It remains to estimate the second line in the rhs. of (7.82).

The expectation of the second line of (7.82) is bounded by a constant times $n^{-1}\,\mathrm{d}t$, exactly as in (7.77). This is the only input needed to bound the terms (7.82) in [59, Eq. (3.122)-(3.123)]. Hence, in order to conclude the proof of this proposition we are left with the term in (7.83).

The quadratic variation of the term in (7.83) is given by

$$\mathrm{d}\langle M\rangle_t = \frac{1}{2n}\sum_{1\leq|i|,|j|\leq n}(w_i - f_i)(w_j - f_j)f_i'f_j'\,\mathbf{E}[\mathrm{d}C_i(\alpha,t)\,\mathrm{d}C_j(\alpha,t)\,|\,\mathcal{F}_t],$$

using the notation in [59, Eq. (3.155)–(3.157)] is used. By (b) of Assumption 7.11 it follows that

$$\mathrm{d}\langle M\rangle_t = \frac{1}{4n^2}\sum_{1\leq|i|,|j|\leq n^{\omega_A}}(w_i - f_i)(w_j - f_j)f_i'f_j'\,\mathbf{E}[\mathrm{d}C_i(\alpha,t)\,\mathrm{d}C_j(\alpha,t)\,|\,\mathcal{F}_t]$$
$$+ \frac{\alpha^2 + (1-\alpha)^2}{8n^3}\sum_{n^{\omega_A}<|i|\leq n}(w_i - f_i)^2(f_i')^2\,\mathrm{d}t. \qquad (7.84)$$

Then, by (c) of Assumption 7.11, for $|i|,|j|\leq n^{\omega_A}$ we have

$$\mathbf{E}[\mathrm{d}C_i(\alpha,t)\,\mathrm{d}C_j(\alpha,t)\,|\,\mathcal{F}_t] = \left[\alpha^2 + (1-\alpha)^2\right]\frac{\delta_{ij}}{2n}\,\mathrm{d}t$$
$$+ \frac{\alpha(1-\alpha)}{2n}\mathbf{E}\left[\left(\mathrm{d}\mathfrak{b}_i^s\,\mathrm{d}\mathfrak{b}_j^r + \mathrm{d}\mathfrak{b}_i^r\,\mathrm{d}\mathfrak{b}_j^s\right)\,|\,\mathcal{F}_t\right], \qquad (7.85)$$

and that

$$\mathbf{E}\left[\mathrm{d}\mathfrak{b}_i^s\,\mathrm{d}\mathfrak{b}_j^r\,|\,\mathcal{F}_t\right] = \mathbf{E}\left[(\mathrm{d}\mathfrak{b}_i^s - \mathrm{d}\mathfrak{b}_i^r)\,\mathrm{d}\mathfrak{b}_j^r\,|\,\mathcal{F}_t\right] + \delta_{ij}\,\mathrm{d}t \lesssim \left(|L_{ii}(t)|^{1/2} + \delta_{ij}\right)\mathrm{d}t, \qquad (7.86)$$

where in the last inequality we used Kunita-Watanabe inequality.

Combining (7.84)–(7.86) we finally conclude that

$$
\begin{aligned}
\mathrm{d}\langle M\rangle_t \leq {} & \frac{1}{8n^3} \sum_{1 \leq |i| \leq n} (w_i - f_i)^2 (f_i')^2 \, \mathrm{d}t \\
& + \frac{\alpha(1-\alpha)}{4n^3} \sum_{1 \leq |i|,|j| \leq n^{\omega_A}} |L_{ii}(t)|^{1/2} \big|(w_i - f_i)(w_j - f_j) f_i' f_j'\big| \, \mathrm{d}t.
\end{aligned}
\tag{7.87}
$$

Since $\alpha \in [0,1]$, $|L_{ii}(t)| \leq n^{-\omega_Q}$ and $\omega_A \ll \omega_Q$ by (7.47) and (7.65)–(7.66), using Cauchy-Schwarz in (7.87), we conclude that

$$
\mathrm{d}\langle M\rangle_t \lesssim \frac{1}{n^3} \sum_{1 \leq |i| \leq n} (w_i - f_i)^2 (f_i')^2 \, \mathrm{d}t,
\tag{7.88}
$$

which is exactly the lhs. in [59, Eq. (3.155)], hence the high probability bound in [59, Eq. (3.155)] follows. Then the remainder of the proof of [59, Lemma 3.14] proceeds exactly in the same way.

Given (7.82) as an input, the proof of (7.80) is concluded following the proof of [59, Theorems 3.16-3.17] line by line.

$\square$

7.6.4. *Proof of Proposition 7.14.* We conclude this section with the proof of Proposition 7.14 following [26, Section 3.6]. We remark that all the estimates above hold uniformly in $\alpha \in [0,1]$ when bounding an integrand by [59, Appendix E].

Proof of Proposition 7.14. For any $|i| \leq n$, by (7.71), it follows that

$$
s_i(t_0 + t_1) - r_i(t_0 + t_1) = x_i(t_1, 1) - x_i(t_1, 0) = \widehat{x}_i(t_1, 1) - \widehat{x}_i(t_1, 0) + \mathcal{O}\left(\frac{n^\xi t_1}{n^{\omega_l}}\right).
\tag{7.89}
$$

We remark that in (7.89) we ignored the scaling (7.49) since it can be removed by a simple time-rescaling (see Remark 7.15 for more details). Then, using that $u_i = \partial_\alpha \widehat{x}_i$ we have that

$$
\widehat{x}_i(t_1, 1) - \widehat{x}_i(t_1, 0) = \int_0^1 u_i(t_1, \alpha) \, \mathrm{d}\alpha.
\tag{7.90}
$$

We recall that u is a solution of

$$
\mathrm{d}\boldsymbol{u} = \mathcal{B}\boldsymbol{u} \, \mathrm{d}t + \mathrm{d}\boldsymbol{\xi}_1 + \boldsymbol{\xi}_2 \, \mathrm{d}t,
\tag{7.91}
$$

as defined in (7.72)–(7.73), with

$$
|\xi_{2,i}(t)| \leq \mathbf{1}_{\{|i|>n^{\omega_A}\}} n^C,
\tag{7.92}
$$

with very high probability for some constant $C > 0$ and any $0 \leq t \leq t_1$. The estimates to remove the forcing term $\boldsymbol{\xi}_2$ in (7.90) are completely analogous to [26, Section 3.7], and so omitted; we thus focus on the estimate of the effect of $\mathrm{d}\boldsymbol{\xi}_1$ as if $\boldsymbol{\xi}_2$ were not present in (7.91). We now split the estimate of this term into short and long range part. We start with the short range part, and then explain the relatively minor differences in the estimates for the long range part. Define $\boldsymbol{v} = \boldsymbol{v}(t)$ as the solution of

$$
\mathrm{d}\boldsymbol{v} = \mathcal{B}\boldsymbol{v} \, \mathrm{d}t + \mathbf{1}(|\cdot| > n^{\omega_A}) \, \mathrm{d}\boldsymbol{\xi}_1, \quad \boldsymbol{v}(0) = \boldsymbol{u}(0),
\tag{7.93}
$$

and let $\boldsymbol{w} := \boldsymbol{u} - \boldsymbol{v}$. We thus get

$$
\mathrm{d}\boldsymbol{w} = \mathcal{B}\boldsymbol{w} \, \mathrm{d}t + \mathbf{1}(|\cdot| \leq n^{\omega_A}) \, \mathrm{d}\boldsymbol{\xi}_1, \qquad \boldsymbol{w}(0) = 0.
\tag{7.94}
$$

To show that $w_i(t) \leq n^{-1-\omega}$ holds with very high probability for any $t \leq t_f$, for some $\omega > 0$, we do a standard ℓ^2–estimate. Define $F(t) := \sum_i |w_i(t)|^2$, then we have

$$
\mathrm{d}F = -\frac{1}{2} \sum_{i,j} \mathcal{B}_{ij} (w_i - w_j)^2 \, \mathrm{d}t + \sum_{|i| \leq n^{\omega_A}} w_i \, \mathrm{d}\boldsymbol{\xi}_1 + \frac{1}{n} \sum_{|i| \leq n^{\omega_A}} L_{ii} \, \mathrm{d}t.
\tag{7.95}
$$

Define the stopping time

$$
\tau := \inf\{t > 0 : F(t) > n^{\omega_A + \omega_f - \omega_Q/2 - 2}\} \wedge t_f.
$$

Using (7.47) and that $|w_i(t)| \lesssim n^{-1+\xi}$ by rigidity, the quadratic variation of the stochastic term in (7.95) is estimate by

$$\mathbf{E}\left[\sum_{|i| \leq n^{\omega_A}} w_i \, \mathrm{d}\boldsymbol{\xi}_1, \sum_{|i| \leq n^{\omega_A}} w_i \, \mathrm{d}\boldsymbol{\xi}_1 \Big| \mathcal{F}_t\right] = \frac{1}{2n} \sum_{|i|,|j| \leq n^{\omega_A}} w_i w_j L_{ij} \, \mathrm{d}t \leq n^{2\xi+2\omega_A-2\omega_Q-3} \, \mathrm{d}t.$$

By the Burkholder–Davis–Gundy (BDG) inequality, we thus conclude that

$$\sup_{0 \leq t \leq t_f} \left|\int_0^t \sum_{|i| \leq n^{\omega_A}} w_i(t) \, \mathrm{d}\boldsymbol{\xi}_1(t)\right| \lesssim n^{2\xi+\omega_A-\omega_Q-3/2} \sqrt{t_f} \tag{7.96}$$

with very high probability. Then, using that $|L_{ii}| \leq n^{-\omega_Q}$ to estimate the last term in (7.95) and that $F(0) = 0$, we obtain

$$\sup_{0 \leq t < \tau} F(t) \leq \frac{n^{\xi+\omega_A+\omega_f}}{n^{2+\omega_Q}}, \tag{7.97}$$

for any arbitrary small $\xi > 0$ with very high probability. This shows that $\tau = t_f$ and so that $|w_i(t_f)| \leq n^{-1-\omega}$, for any $i \in [n]$, for some sufficiently small $\omega > 0$. This follows from the relation among the various ω's in Section 7.2.3, as in our application $\omega_Q = \omega_E$.

We now turn to the estimate of the long range part, i.e. we look at (7.93). We fix an index $|p| \geq n^{\omega_A}$, and we will study the effect on the solution of each single $\mathrm{d}\xi_{1,p}$ separately. We thus consider $\boldsymbol{w} = \boldsymbol{w}^{(p)} := \boldsymbol{v} - \boldsymbol{v}^{(p)}$, with

$$\mathrm{d}\boldsymbol{v}^{(p)} = \mathcal{B}\boldsymbol{v}^{(p)} \, \mathrm{d}t + (1 - \mathbf{1}(\cdot = p)) \, \mathrm{d}\boldsymbol{\xi}_1, \qquad \boldsymbol{v}^{(p)}(0) = \boldsymbol{v}(0),$$

and so obtain

$$\mathrm{d}\boldsymbol{w} = \mathcal{B}\boldsymbol{w} \, \mathrm{d}t + \mathbf{1}(\cdot = p) \, \mathrm{d}\boldsymbol{\xi}_1, \qquad \boldsymbol{w}(0) = 0, \tag{7.98}$$

i.e. we just removed a single stochastic forcing term from $\boldsymbol{u}$. To show that for any $|p| \geq n^{\omega_A}$ we have $|w_i^{(p)}| \leq \exp(-n^{\omega_l/10})$ for $|i| \leq n^{\omega_l}$, with ω_l the short–range scale from (7.67), it is enough to use a "modified" ℓ^2–method as in the proof of finite speed of propagation in [26, Lemma 3.9], [59, Lemma 3.7], together with the changes presented in the proof of Lemma 7.21 to deal with the new additional stochastic term; we thus omit the details. Since the bound $|w_i^{(p)}| \leq \exp(-n^{\omega_l/10})$ holds for any fixed p, we then conclude that for $|i| \leq n^{\omega_l}$ can remove all the stochastic forcing terms in (7.93) at the price of an error which is smaller than $n \exp(-n^{\omega_l/10})$. This concludes the bound of the new term $\mathrm{d}\boldsymbol{\xi}_1$. The remainder of the proof of Proposition 7.14 proceeds exactly in the same way of [26, Eq. (3.86)–(3.99)], hence we omit it. Since $t_f = t_0 + t_1$, choosing $\omega = \omega_1/10$, $\widehat{\omega} \leq \omega/10$, the above computations conclude the proof of Proposition 7.14. □

Appendix A. Proof of Lemma 4.9

In order to prove Lemma 4.9 we have to compute

$$\frac{2}{\pi^2} \int_{\mathbf{C}} \mathrm{d}^2 z_1 \int_{\mathbf{C}} \mathrm{d}^2 z_2 \partial_1 \overline{\partial}_1 f(z_1) \partial_2 \overline{\partial}_2 \overline{g(z_2)} \Theta(z_1, z_2) \tag{A.1}$$

for compactly supported smooth functions f, g. We recall that

$$\Theta(z_1, z_2) = \Xi(z_1, z_2) + \Lambda(z_1, z_2), \qquad \Lambda(z_1, z_2) := -\frac{1}{2} \log|1 - z_1 \overline{z}_2|^2 \mathbf{1}(|z_1|, |z_2| > 1),$$

$$\Xi(z_1, z_2) := -\frac{1}{2} \log|z_1 - z_2|^2 \big[1 - \mathbf{1}(|z_1|, |z_2| > 1)\big] + \frac{1}{2} \log|z_1|^2 \mathbf{1}(|z_1| \geq 1) \tag{A.2}$$

$$+ \frac{1}{2} \log|z_2|^2 \mathbf{1}(|z_2| \geq 1).$$

In order to compute (A.1) we will perform integration by parts twice. For this purpose we split the integral in (A.1) for $\Xi(z_1, z_2)$ into the regimes $|z_1 - z_2| \geq \epsilon$ and its complement, and the integral of $\Lambda(z_1, z_2)$ into the regimes $|1 - z_1 \overline{z}_2| \geq \epsilon$ and its complement. We decided to perform two different cut-offs for Ξ and Λ as a consequence of the different kind of singularity of the logarithms in their definition.

By the explicit definitions in (A.2), it is easy to see that the integrals in the regimes $|z_1 - z_2| \leq \epsilon$, $|1 - z_1\overline{z}_2| \leq \epsilon$ go to zero as $\epsilon \to 0$, hence we have

$$2\mathcal{I} := \frac{2}{\pi^2} \int_{\mathbf{C}} \mathrm{d}^2 z_1 \int_{\mathbf{C}} \mathrm{d}^2 z_2 \partial_1 \overline{\partial}_1 f(z_1) \partial_2 \overline{\partial}_2 \overline{g(z_2)} \Theta(z_1, z_2)$$

$$= \lim_{\epsilon \to 0} \frac{2}{\pi^2} \int_{\mathbf{C}} \mathrm{d}^2 z_1 \int_{\mathbf{C}} \mathrm{d}^2 z_2 \partial_1 \overline{\partial}_1 f(z_1) \partial_2 \overline{\partial}_2 \overline{g(z_2)} \qquad \text{(A.3)}$$

$$\times \left[\Xi(z_1, z_2)\mathbf{1}(|z_1 - z_2| \geq \epsilon) + \Lambda(z_1, z_2)\mathbf{1}(|1 - z_1\overline{z}_2| \geq \epsilon) \right].$$

In order to prove Lemma 4.9 we write the l.h.s. of (A.3) as $\mathcal{I} + \mathcal{I}$ so that in the first integral we perform integration by parts with respect to $\partial_1, \overline{\partial}_2$ and in the second one with respect to $\overline{\partial}_1, \partial_2$. This split is motivated by the fact that

$$\overline{\partial}\overline{g}\partial f + \partial\overline{g}\overline{\partial}f = \frac{1}{2}\langle \nabla g, \nabla f \rangle,$$

which is the first term in the l.h.s. of (4.32) in Lemma 4.9. From now on we focus only on the integral for which we perform integration by parts with respect to $\partial_1, \overline{\partial}_2$. The computations for the other integral are exactly the same. It is well known that the distributional Laplacian of $\log|z_1 - z_2|$ is 2π the delta function in $z_1 = z_2$, more precisely, we have that

$$-\partial_1\partial_2 \log|z_1 - z_2| \, \mathrm{d}^2 z_1 \, \mathrm{d}^2 z_2 = \frac{\pi}{2}\delta(z_1 - z_2), \qquad \text{(A.4)}$$

in the sense of distributions. Hence, in the remainder of this section we focus on the computation of the integral of $\Lambda(z_1, z_2)$ and omit the ϵ-regularisation in the integral of Ξ.

Performing integration by parts in $\mathcal{I}$, which is defined in (A.5), with respect to $\partial_1, \overline{\partial}_2$ we get

$$\lim_{\epsilon \to 0} \frac{1}{\pi^2} \int_{\mathbf{C}} \mathrm{d}^2 z_1 \int_{\mathbf{C}} \mathrm{d}^2 z_2 \partial_1 \overline{\partial}_1 f(z_1) \partial_2 \overline{\partial}_2 \overline{g(z_2)} \left[\Xi(z_1, z_2) + \Lambda(z_1, z_2)\mathbf{1}(|1 - z_1\overline{z}_2| \geq \epsilon) \right]$$

$$= \lim_{\epsilon \to 0} \frac{1}{\pi^2} \int_{\mathbf{C}} \mathrm{d}^2 z_1 \int_{\mathbf{C}} \mathrm{d}^2 z_2 \overline{\partial}_1 f(z_1) \partial_2 \overline{g(z_2)} \left[\partial_1 \overline{\partial}_2 \Xi(z_1, z_2) + \partial_1 \overline{\partial}_2 \Lambda(z_1, z_2)\mathbf{1}(|1 - z_1\overline{z}_2| \geq \epsilon) \right]$$

$$+ \lim_{\epsilon \to 0} -\frac{\mathrm{i}}{2\pi^2} \int_{\mathbf{C}} \int_{\mathbf{C}} \mathrm{d}^2 z_2 \overline{\partial}_1 f \left[\overline{\partial}_2 \partial_2 \overline{g} \Lambda \mathbf{1}(|1 - z_1\overline{z}_2| = \epsilon) \, \mathrm{d}\overline{z}_1 - \partial_2 \overline{g} \partial_1 \Lambda \mathbf{1}(|1 - z_1\overline{z}_2| = \epsilon) \, \mathrm{d}z_2 \right]$$

$$=: \lim_{\epsilon \to 0} \left[J_{1,\epsilon} + J_{2,\epsilon} \right]. \qquad \text{(A.5)}$$

where in the fourth line we used Stokes theorem written symbolically in the form

$$\partial_z \mathbf{1}(|z - z_2| \geq \epsilon) \, \mathrm{d}^2 z = \frac{\mathrm{i}}{2}\mathbf{1}(|z - z_2| = \epsilon) \, \mathrm{d}\overline{z} \qquad \text{(A.6)}$$

for any fixed z_2. We remark that (A.6) is understood in the sense of distributions, i.e. the equality holds when tested again smooth compactly supported test functions f, i.e.

$$-\int_{\mathbf{C}} \partial_z f(z)\mathbf{1}(|z - z_2| \geq \epsilon) \, \mathrm{d}^2 z = \frac{\mathrm{i}}{2} \int_{|z - z_2| = \epsilon} f(z) \, \mathrm{d}\overline{z}.$$

Moreover, with a slight abuse of notation in (A.5)–(A.6) by $\mathbf{1}(|z - z_2| = \epsilon) \, \mathrm{d}\overline{z}$ we denoted the clockwise contour integral over the circle of radius ϵ around z_2. We use the notation above in the remainder of this section.

The second derivative (in the sense of the distributions) of $\Xi(z_1, z_2)$ in (A.5), using (A.4), is given by

$$\partial_1 \overline{\partial}_2 \Xi \, \mathrm{d}^2 z_1 \, \mathrm{d}^2 z_2$$

$$= \frac{\pi}{2}\delta(z_1 - z_2)\left[1 - \mathbf{1}(|z_1|, |z_2| > 1)\right] \mathrm{d}^2 z_1 \, \mathrm{d}^2 z_2 - \frac{1}{8}\log|z_1 - z_2|^2 \mathbf{1}(|z_1| = 1) \, \mathrm{d}\overline{z}_1 \mathbf{1}(|z_2| = 1) \, \mathrm{d}z_2$$

$$+ \frac{\mathrm{i}}{4}\frac{1}{z_1 - z_2}\mathbf{1}(|z_1| > 1) \, \mathrm{d}^2 z_1 \mathbf{1}(|z_2| = 1) \, \mathrm{d}z_2 - \frac{\mathrm{i}}{4}\frac{1}{\overline{z}_1 - \overline{z}_2}\mathbf{1}(|z_2| > 1) \, \mathrm{d}^2 z_2 \mathbf{1}(|z_1| = 1) \, \mathrm{d}\overline{z}_1,$$

$$\text{(A.7)}$$

whilst the second derivative of $\Lambda(z_1, z_2)$ by

$$\partial_1 \overline{\partial}_2 \Lambda \, \mathrm{d}^2 z_1 \, \mathrm{d}^2 z_2$$

$$= \frac{1}{2(1 - z_1 \overline{z}_2)^2} \mathbf{1}(|z_1|, |z_2| > 1) \, \mathrm{d}^2 z_1 \, \mathrm{d}^2 z_2 + \frac{1}{8} \log|1 - z_1 \overline{z}_2| \mathbf{1}(|z_1| = 1) \, \mathrm{d}\overline{z}_1 \mathbf{1}(|z_2| = 1) \, \mathrm{d}z_2$$

$$+ \frac{\mathrm{i}}{4} \frac{\overline{z}_2}{1 - z_1 \overline{z}_2} \mathbf{1}(|z_1| > 1) \, \mathrm{d}^2 z_1 \mathbf{1}(|z_2| = 1) \, \mathrm{d}z_2 + \frac{\mathrm{i}}{4} \frac{z_1}{1 - z_1 \overline{z}_2} \mathbf{1}(|z_2| > 1) \, \mathrm{d}^2 z_2 \mathbf{1}(|z_1| = 1) \, \mathrm{d}\overline{z}_1.$$

$$\tag{A.8}$$

Note that

$$\partial_1 \overline{\partial}_2 (\Xi + \Lambda) \, \mathrm{d}^2 z_1 \, \mathrm{d}^2 z_2 = \frac{\pi}{2} \delta(z_1 - z_2) \mathbf{1}(|z_1|, |z_2| \leq 1) \, \mathrm{d}^2 z_1 \, \mathrm{d}^2 z_2$$

$$+ \frac{1}{2(1 - z_1 \overline{z}_2)^2} \mathbf{1}(|z_1|, |z_2| > 1) \, \mathrm{d}^2 z_1 \, \mathrm{d}^2 z_2,$$

hence, by (A.7)–(A.8) we conclude that

$$\lim_{\epsilon \to 0} J_{1,\epsilon} = \frac{1}{2\pi} \int_{\mathbf{D}} \overline{\partial} f \partial \overline{g} \, \mathrm{d}^2 z + \lim_{\epsilon \to 0} \frac{1}{2\pi} \int_{|z_1| \geq 1} \mathrm{d}^2 z_1 \int_{|z_2| \geq 1} \mathrm{d}^2 z_2 \frac{\overline{\partial}_1 f(z_1) \partial_2 g(z_2)}{(1 - z_1 \overline{z}_2)^2} \mathbf{1}(|1 - z_1 \overline{z}_2| \geq \epsilon).$$

$$\tag{A.9}$$

On the other hand, the integration by parts with respect to $\overline{\partial}_1, \partial_2$ gives

$$\frac{1}{2\pi} \int_{\mathbf{D}} \overline{\partial} f \partial \overline{g} \, \mathrm{d}^2 z + \lim_{\epsilon \to 0} \frac{1}{2\pi} \int_{|z_1| \geq 1} \mathrm{d}^2 z_1 \int_{|z_2| \geq 1} \mathrm{d}^2 z_2 \frac{\overline{\partial}_1 f(z_1) \partial_2 g(z_2)}{(1 - z_1 \overline{z}_2)^2} \mathbf{1}(|1 - z_1 \overline{z}_2| \geq \epsilon). \tag{A.10}$$

Hence, summing (A.9)–(A.10) we get exactly the r.h.s. of (4.32) using that

$$\frac{1}{2\pi} \int_{\mathbf{D}} \left[\overline{\partial} \overline{g} \partial f + \partial \overline{g} \overline{\partial} f \right] \mathrm{d}^2 z = \frac{1}{4\pi} \int_{\mathbf{D}} \langle \nabla g, \nabla f \rangle \, \mathrm{d}^2 z.$$

In order to conclude the proof of Lemma 4.9 we prove that $|J_{2,\epsilon}| \to 0$ as $\epsilon \to 0$ in Lemma A.1 and that the limit in the r.h.s. of (A.9) exists in Lemma A.2.

Lemma A.1. *Let $J_{2,\epsilon}$ be defined in* (A.5)*, then*

$$\lim_{\epsilon \to 0} |J_{2,\epsilon}| = 0. \tag{A.11}$$

Proof. For the first integral in $J_{2,\epsilon}$, using the parametrization $z_2 = r_2 e^{\mathrm{i}\theta_2}$ and $z_1 = (1 + \epsilon e^{\mathrm{i}\theta_1})/\overline{z}_2$, for any fixed z_2, we get

$$\left| \int_1^\infty \mathrm{d}r_2 \int_0^{2\pi} \mathrm{d}\theta_1 \int_0^{2\pi} \mathrm{d}\theta_2 \, \epsilon e^{\mathrm{i}(\theta_1 + \theta_2)} \overline{\partial}_1 f \left(r_2^{-1} e^{\mathrm{i}\theta_2} [1 + \epsilon e^{\mathrm{i}\theta_1}] \right) \overline{\partial}_2 \partial_2 g(r_2 e^{\mathrm{i}\theta_2}) \log \epsilon \right| \lesssim \epsilon \log \epsilon,$$

$$\tag{A.12}$$

where we used that $\|\overline{\partial}_1 f\|_{L^\infty(\mathbf{C})}, \|\overline{\partial}_2 \partial_2 g\|_{L^1(\mathbf{C})} \lesssim 1$ as a consequence of $f, g \in H_0^{2+\delta}(\Omega)$, for an open set $\Omega \subset \mathbf{C}$ such that $\overline{\mathbf{D}} \subset \Omega$.

Furthermore, using the parametrizations $z_1 = r_1 e^{\mathrm{i}\theta_1}$ and $z_2 = (1 + \epsilon e^{\mathrm{i}\theta_2})/\overline{z}_1$ for the second integral in $J_{2,\epsilon}$, we have that

$$J_{2,\epsilon} = \left[\int_1^\infty \mathrm{d}r_1 \int_0^{2\pi} \mathrm{d}\theta_1 \int_0^{2\pi} \mathrm{d}\theta_2 \, \epsilon e^{\mathrm{i}(\theta_1 + \theta_2)} \overline{\partial}_1 f(r_1 e^{\mathrm{i}\theta_1}) \partial_2 \overline{g} \left(r_1^{-1} e^{\mathrm{i}\theta_1} [1 + \epsilon e^{\mathrm{i}\theta_2}] \right) \right.$$

$$\left. \times \frac{1 + \epsilon e^{-\mathrm{i}\theta_2}}{\epsilon r_1 e^{-\mathrm{i}\theta_2} e^{\mathrm{i}\theta_1}} \mathbf{1}(|1 + \epsilon e^{\mathrm{i}\theta_2}| > r_1) \right] + \mathcal{O}(\epsilon \log \epsilon),$$

$$\tag{A.13}$$

where the error term comes from the integral of $\partial_1 \mathbf{1}(|z_1|, |z_2| > 1)$ and the bound in (A.12). Note that $\mathbf{1}(|1 + \epsilon e^{\mathrm{i}\theta_1}| > r_1) = 0$ if $r_1 \geq 1 + 2\epsilon$, hence we can bound the first term in J_2 by

$$\int_1^{1+2\epsilon} \mathrm{d}r_1 \int_0^{2\pi} \mathrm{d}\theta_1 \int_0^{2\pi} \mathrm{d}\theta_2 \left| \overline{\partial}_1 f(r_1 e^{\mathrm{i}\theta_1}) \partial_2 \overline{g} \left(r_1^{-1} e^{\mathrm{i}\theta_1} [1 + \epsilon e^{\mathrm{i}\theta_2}] \right) \right| \lesssim \epsilon, \tag{A.14}$$

since $\|\partial_2 \overline{g}\|_{L^\infty(\mathbf{C})}, \|\overline{\partial}_1 f\|_{L^1(\mathbf{C})} \lesssim 1$. Hence, we conclude that

$$J_{2,\epsilon} = \mathcal{O}\left(\epsilon + \epsilon \log \epsilon \right).$$

This concludes the proof of (A.11). □

We conclude this section proving the existence of the limit of $J_{1,\epsilon}$ as $\epsilon \to 0$. More precisely, in Lemma A.2 we prove that $J_{1,\epsilon}$ is a Cauchy sequence.

Lemma A.2. *Let $J_{1,\epsilon}$ be defined in (A.5), then for any $0 < \epsilon' \le \epsilon$ we have that*

$$|J_{1,\epsilon} - J_{1,\epsilon'}| \lesssim \epsilon^{\delta}, \tag{A.15}$$

for some $\delta > 0$.

Proof. We only consider the integral with the second derivative of Λ. We dealt with the integral of the second derivative of $\Xi(z_1, z_2)$ already in (A.4). Define

$$I_{\epsilon} := \frac{1}{\pi^2} \int_{\mathbf{C}} d^2 z_1 \int_{\mathbf{C}} d^2 z_2 F(z_1, z_2) \Big[\partial_2 \overline{\partial}_1 \Lambda(z_1, z_2) \mathbf{1}(|1 - z_1 \overline{z}_2| \ge \epsilon)\Big], \tag{A.16}$$

where $F(z_1, z_2) := \overline{\partial}_1 f(z_1) \partial_2 \overline{g(z_2)}$ is a δ-Hölder continuous function. Then, for any $0 < \epsilon' < \epsilon$, using the change of variables $z_2 = r_2 e^{i\theta_2}$ and $z_1 = (1 + r_1 e^{i\theta_1})/\overline{z}_2$, we write

$$I_{\epsilon'} - I_{\epsilon} = \frac{1}{\pi^2} \int_{\mathbf{C}} d^2 z_1 \int_{\mathbf{C}} d^2 z_2 \big(F(z_1, z_2) - F(\overline{z}_2^{-1}, z_2)\big) \Big[\partial_2 \overline{\partial}_1 \Lambda(z_1, z_2) \mathbf{1}(\epsilon \ge |1 - z_1 \overline{z}_2| \ge \epsilon')\Big]$$
$$+ \frac{1}{\pi} \int_1^{\infty} dr_2 \int_0^{2\pi} d\theta_2 \int_0^{2\pi} d\theta_1 \int_{\epsilon'}^{\epsilon} dr_1 F(r_2^{-1} e^{-i\theta_2}, r_2 e^{i\theta_2}) \frac{e^{2i\theta_1}}{r_1 r_2}. \tag{A.17}$$

Note that the integral in the second line of (A.17) is exactly zero since $e^{2i\theta_1}$ the only term which depends on θ_1. On the other hand, we can bound the first integral in (A.17) by $\epsilon^{2\delta}$, with δ the Hölder exponent of F, using the fact that

$$\big|F(z_1, z_2) - F(\overline{z}_2^{-1}, z_2)\big| \le \left|\frac{1}{\overline{z}_2} + \frac{r_1 e^{i\theta_1}}{\overline{z}_2} - \frac{1}{\overline{z}_2}\right|^{2\delta} \lesssim \left(\frac{r_1}{r_2}\right)^{2\delta}.$$

This concludes the proof of this lemma. □

APPENDIX B. DERIVATION OF THE DBM FOR THE EIGENVALUES OF H^z

Let X be an $n \times n$ complex random matrix, let H^z be the Hermitisation of $X - z$ defined in (7.1), and define $Y^z := X - z$. We recall that $\{\lambda_i^z, -\lambda_i^z\}_{i=1}^n$ are the eigenvalues of H^z, and $\{w_i^z, w_{-i}^z\}_{i=1}^n$ are the corresponding orthonormal eigenvectors, i.e. for any $i, j \in [n]$ we have

$$H^z w_{\pm i}^z = \pm \lambda_i^z, \qquad (w_i^z)^* w_j^z = \delta_{i,j}, \qquad (w_i^z)^* w_{-j}^z = 0, \tag{B.1}$$

for any $i, j \in [n]$. For simplicity in the following derivation we assume that the eigenvalues are all distinct. In particular, for any $i \in [n]$, by the block structure of H^z it follows that

$$w_{\pm i}^z = (u_i^z, \pm v_i^z), \qquad Y^z v_i^z = \lambda_i^z u_i^z, \qquad (Y^z)^* u_i^z = \lambda_i^z v_i^z. \tag{B.2}$$

Moreover, since $\{w_{\pm i}^z\}_{i=1}^n$ is an orthonormal base, we conclude that

$$(u_i^z)^* u_i^z = (v_i^z)^* v_i^z = \frac{1}{2}. \tag{B.3}$$

In the following, for any fixed entry x_{ab} of X, we will use the notation

$$\dot{f} = \frac{\partial f}{\partial x_{ab}} \quad \text{or} \quad \dot{f} = \frac{\partial f}{\partial \overline{x}_{ab}}, \tag{B.4}$$

where $f = f(X)$ is a function of the matrix X. Then, we consider the flow

$$dX_t = \frac{dB_t}{\sqrt{n}}, \quad X_0 = X, \tag{B.5}$$

where B_t is a matrix valued complex standard Brownian motion.

From now on we only consider positive indices $1 \leq i \leq n$. We may also drop the z and t dependence to make our notation easier. For any $i, j \in [n]$, differentiating (B.1) we get

$$\dot{H}\boldsymbol{w}_i + H\dot{\boldsymbol{w}}_i = \dot{\lambda}_i \boldsymbol{w}_i + \lambda_i \dot{\boldsymbol{w}}_i, \tag{B.6}$$

$$\dot{\boldsymbol{w}}_i^* \boldsymbol{w}_j + \boldsymbol{w}_i^* \dot{\boldsymbol{w}}_j = 0, \tag{B.7}$$

$$\boldsymbol{w}_i^* \dot{\boldsymbol{w}}_i + \dot{\boldsymbol{w}}_i^* \boldsymbol{w}_i = 0. \tag{B.8}$$

Note that (B.8) implies that $\Re[\boldsymbol{w}_i^* \dot{\boldsymbol{w}}_i] = 0$. Hence, since the eigenvectors are defined modulo a phase, we can choose eigenvectors such that $\Im[\boldsymbol{w}_i^* \dot{\boldsymbol{w}}_i] = 0$ for any $t \geq 0$. Then, multiplying (B.6) by $\boldsymbol{w}_i^*$ we conclude that

$$\dot{\lambda}_i = \boldsymbol{u}_i^* \dot{Y} \boldsymbol{v}_i + \boldsymbol{v}_i^* \dot{Y}^* \boldsymbol{u}_i. \tag{B.9}$$

Moreover, multiplying (B.6) by $\boldsymbol{w}_j^*$, with $j \neq i$, and by $\boldsymbol{w}_{-j}*$, we get

$$(\lambda_i - \lambda_j)\boldsymbol{w}_j^* \dot{\boldsymbol{w}}_i = \boldsymbol{w}_j^* \dot{H} \boldsymbol{w}_i, \qquad (\lambda_i + \lambda_j)\boldsymbol{w}_{-j}^* \dot{\boldsymbol{w}}_i = \boldsymbol{w}_{-j}^* \dot{H} \boldsymbol{w}_i, \tag{B.10}$$

respectively. By (B.7)–(B.8) it follows that

$$\dot{\boldsymbol{w}}_i = \sum_{j \neq i} (\boldsymbol{w}_j^* \dot{\boldsymbol{w}}_i)\boldsymbol{w}_j + \sum_j (\boldsymbol{w}_{-j}^* \dot{\boldsymbol{w}}_i)\boldsymbol{w}_{-j}, \tag{B.11}$$

hence by (B.10) we conclude

$$\dot{\boldsymbol{w}}_i = \sum_{j \neq i} \frac{\boldsymbol{v}_j^* \dot{Y}^* \boldsymbol{u}_i + \boldsymbol{u}_j^* \dot{Y} \boldsymbol{v}_i}{\lambda_i - \lambda_j} \boldsymbol{w}_j + \sum_j \frac{\boldsymbol{u}_j^* \dot{Y} \boldsymbol{v}_i - \boldsymbol{v}_j^* \dot{Y}^* \boldsymbol{u}_i}{\lambda_i + \lambda_j} \boldsymbol{w}_{-j}. \tag{B.12}$$

By Ito's formula we have that

$$\mathrm{d}\lambda_i = \sum_{ab} \frac{\partial \lambda_i}{\partial x_{ab}} \mathrm{d}x_{ab} + \frac{\partial \lambda_i}{\partial \overline{x_{ab}}} \mathrm{d}\overline{x_{ab}} + \frac{1}{2} \sum_{ab} \sum_{kl} \frac{\partial^2 \lambda_i}{\partial x_{ab} \partial \overline{x_{kl}}} \mathrm{d}x_{ab} \, \mathrm{d}\overline{x_{kl}} + \frac{\partial^2 \lambda_i}{\partial \overline{x_{ab}} \partial x_{kl}} \mathrm{d}\overline{x_{ab}} \, \mathrm{d}x_{kl}. \tag{B.13}$$

Note that in (B.13) we used that $\mathrm{d}x_{ab} \, \mathrm{d}x_{ab} = \mathrm{d}\overline{x_{kl}} \, \mathrm{d}\overline{x_{kl}} = 0$. Then by (B.9)–(B.12) it follows that

$$\frac{\partial \lambda_i}{\partial x_{ab}} = u_i(a)^* v_i(b), \qquad \frac{\partial \lambda_i}{\partial \overline{x_{ab}}} = v_i(b)^* u_i(a), \tag{B.14}$$

and that

$$\frac{\partial w_i}{\partial x_{ab}}(k) = \sum_{j \neq i} \left[\frac{u_j^*(a)v_i(b)}{\lambda_i - \lambda_j} w_j(k) + \frac{u_j^*(a)v_i(b)}{\lambda_i + \lambda_j} w_{-j}(k) \right] + \frac{u_i(a)^* v_i(b)}{2\lambda_i} w_{-i}(k), \tag{B.15}$$

$$\frac{\partial w_i}{\partial \overline{x_{ab}}}(k) = \sum_{j \neq i} \left[\frac{v_j^*(b)u_i(a)}{\lambda_i - \lambda_j} w_j(k) - \frac{v_j^*(b)u_i(a)}{\lambda_i + \lambda_j} w_{-j}(k) \right] - \frac{v_i(b)^* u_i(a)}{2\lambda_i} w_{-i}(k). \tag{B.16}$$

Next, we compute

$$\frac{\partial^2 \lambda_i}{\partial x_{ab} \partial \overline{x_{kl}}} = \frac{\partial v_i^*}{\partial x_{ab}}(l)u_i(k) + v_i(l)^* \frac{\partial u_i}{\partial x_{ab}}(k)$$

$$= \sum_{j \neq i} \left[\frac{v_j(b)u_i^*(a)}{\lambda_i - \lambda_j} v_j(l)^* u_i(k) + \frac{v_j(b)u_i(a)^*}{\lambda_i + \lambda_j} v_j(l)^* u_i(k) \right] + \frac{v_i(b)u_i(a)^*}{2\lambda_i} v_i(l)^* u_i(k)$$

$$+ \sum_{j \neq i} \left[\frac{u_i^*(a)v_i(b)}{\lambda_i - \lambda_j} v_i(l)^* u_j(k) + \frac{u_j^*(a)v_i(b)}{\lambda_i + \lambda_j} v_i(l)^* u_j(k) \right] + \frac{u_i(a)^* v_i(b)}{2\lambda_i} v_\alpha(l)^* u_i(k). \tag{B.17}$$

Finally, combining (B.5), (B.14), (B.13) and (B.17), we conclude (cf. [39, Eq. (5.8)])

$$\mathrm{d}\lambda_i^z = \frac{\mathrm{d}b_i^z}{\sqrt{2n}} + \frac{1}{2n} \sum_{j \neq i} \left[\frac{1}{\lambda_i^z - \lambda_j^z} + \frac{1}{\lambda_i^z + \lambda_j^z} \right] \mathrm{d}t + \frac{\mathrm{d}t}{4n\lambda_i}, \tag{B.18}$$

where we defined

$$\mathrm{d}b_i^z := \sqrt{2}(\mathrm{d}B_{ii}^z + \mathrm{d}\overline{B_{ii}^z}), \quad \mathrm{d}B_{ij}^z := \sum_{ab} \overline{u_i^z(a)} \, \mathrm{d}B_{ab} v_j^z(b), \tag{B.19}$$

where B_t is the matrix values Brownian motion in (B.5). In particular, b_i^z is a standard real Brownian motion, indeed

$$\mathbf{E}(B_{ii}^z + \overline{B_{ii}^z})(B_{ii}^z + \overline{B_{ii}^z})^* = \mathbf{E}\left(\sum_{ab} \overline{u_i^z(a)}B_{ab}v_i^z(b) + u_i^z(a)\overline{B}_{ab}\overline{v_i^z(b)}\right)^2$$

$$= 2\sum_{abcd} \overline{u_i^z(a)}B_{ab}v_i^z(b)u_i^z(c)\overline{B}_{cd}\overline{v_i^z(d)}$$

$$= 2\sum_{abcd} \delta_{ac}\delta_{bd}\overline{u_i^z(a)}v_i^z(b)u_i^z(c)\overline{v_i^z(d)} = \frac{1}{2}.$$

References

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

[2] 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.

[3] O. H. Ajanki, L. Erdős, and T. Krüger, *Quadratic vector equations on complex upper half-plane*, Mem. Amer. Math. Soc. **261**, v+133 (2019), MR4031100.

[4] J. Alt, L. Erdős, and T. Krüger, *Local inhomogeneous circular law*, Ann. Appl. Probab. **28**, 148–203 (2018), MR3770875.

[5] J. Alt, L. Erdős, and T. Krüger, *Spectral radius of random matrices with independent entries*, Probab. Math. Phys. **2**, 221–280 (2021), MR4408013.

[6] J. Alt, L. Erdős, and T. Krüger, *The Dyson equation with linear self-energy: spectral bands, edges and cusps*, Doc. Math. **25**, 1421–1540 (2020), MR4164728.

[7] 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.

[8] Z. D. Bai, *Circular law*, Ann. Probab. **25**, 494–529 (1997), MR1428519.

[9] Z. D. Bai and J. Yao, *On the convergence of the spectral empirical process of Wigner matrices*, Bernoulli **11**, 1059–1092 (2005), MR2189081.

[10] Z. D. Bai and Y. Q. Yin, *Limiting behavior of the norm of products of random matrices and two problems of Geman-Hwang*, Probab. Theory Related Fields **73**, 555–569 (1986), MR863545.

[11] Z. Bao, K. Schnelli, and Y. Xu, *Central limit theorem for mesoscopic eigenvalue statistics of the free sum of matrices*, Int. Math. Res. Not. IMRN, 5320–5382 (2022), MR4403963.

[12] R. Bauerschmidt, P. Bourgade, M. Nikula, and H.-T. Yau, *The two-dimensional Coulomb plasma: quasi-free approximation and central limit theorem*, Adv. Theor. Math. Phys. **23**, 841–1002 (2019), MR4063572.

[13] F. Bekerman, T. Leblé, and S. Serfaty, *CLT for fluctuations of β-ensembles with general potential*, Electron. J. Probab. **23**, Paper no. 115, 31 (2018), MR3885548.

[14] F. Bekerman and A. Lodhia, *Mesoscopic central limit theorem for general β-ensembles*, Ann. Inst. Henri Poincaré Probab. Stat. **54**, 1917–1938 (2018), MR3865662.

[15] C. Bordenave, P. Caputo, D. Chafaï, and K. Tikhomirov, *On the spectral radius of a random matrix: An upper bound without fourth moment*, Ann. Probab. **46**, 2268–2286 (2018), MR3813992.

[16] C. Bordenave and D. Chafaï, *Around the circular law*, Probab. Surv. **9**, 1–89 (2012), MR2908617.

[17] G. Borot and A. Guionnet, *Asymptotic expansion of beta matrix models in the multi-cut regime*, preprint (2013), arXiv:1303.1045.

[18] P. Bourgade and H.-T. Yau, *The eigenvector moment flow and local quantum unique ergodicity*, Comm. Math. Phys. **350**, 231–278 (2017), MR3606475.

[19] P. Bourgade, *Extreme gaps between eigenvalues of Wigner matrices*, J. Eur. Math. Soc. (JEMS) **24**, 2823–2873 (2022), MR4416591.

[20] P. Bourgade, L. Erdős, H.-T. Yau, and J. Yin, *Fixed energy universality for generalized Wigner matrices*, Comm. Pure Appl. Math. **69**, 1815–1881 (2016), MR3541852.

[21] P. Bourgade, H.-T. Yau, and J. Yin, *Local circular law for random matrices*, Probab. Theory Related Fields **159**, 545–595 (2014), MR3230002.

[22] A. Boutet de Monvel and A. Khorunzhy, *Asymptotic distribution of smoothed eigenvalue density. I. Gaussian random matrices*, Random Oper. Stochastic Equations **7**, 1–22 (1999), MR1678012.

[23] A. Boutet de Monvel and A. Khorunzhy, *Asymptotic distribution of smoothed eigenvalue density. II. Wigner random matrices*, Random Oper. Stochastic Equations **7**, 149–168 (1999), MR1689027.

[24] D. Chafaï, *Around the circular law: An update*, (2018) http://djalil.chafai.net/blog/2018/11/04/around-the-circular-law-an-update/ (visited on 12/03/2019).

[25] Z. Che and B. Landon, *Local spectral statistics of the addition of random matrices*, Probab. Theory Related Fields **175**, 579–654 (2019), MR4009717.

[26] Z. Che and P. Lopatto, *Universality of the least singular value for sparse random matrices*, Electron. J. Probab. **24**, Paper No. 9, 53 (2019), MR3916329.

[27] Z. Che and P. Lopatto, *Universality of the least singular value for the sum of random matrices*, Electron. J. Probab. **26**, Paper No. 40, 38 (2021), MR4242226.

[28] 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.

[29] G. Cipolloni, L. Erdős, T. Krüger, and D. Schröder, *Cusp universality for random matrices, II: The real symmetric case*, Pure Appl. Anal. **1**, 615–707 (2019), MR4026551.

[30] G. Cipolloni, L. Erdős, and D. Schröder, *Edge universality for non-Hermitian random matrices*, Probab. Theory Related Fields **179**, 1–28 (2021), MR4221653.

[31] G. Cipolloni, L. Erdős, and D. Schröder, *Fluctuation around the circular law for random matrices with real entries*, Electron. J. Probab. **26**, Paper No. 24, 61 (2021), MR4235475.

[32] G. Cipolloni, L. Erdős, and D. Schröder, *Optimal lower bound on the least singular value of the shifted Ginibre ensemble*, Probab. Math. Phys. **1**, 101–146 (2020), MR4408004.

[33] O. Costin and J. L. Lebowitz, *Gaussian fluctuation in random matrices*, Phys. Rev. Lett. **75**, 69–72 (1995), MR3155254.

[34] N. Coston and S. O'Rourke, *Gaussian fluctuations for linear eigenvalue statistics of products of independent iid random matrices*, J. Theoret. Probab. **33**, 1541–1612 (2020), MR4125967.

[35] M. Duits and K. Johansson, *On mesoscopic equilibrium for linear statistics in Dyson's Brownian motion*, Mem. Amer. Math. Soc. **255**, v+118 (2018), MR3852256.

[36] 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.

[37] L. Erdős, T. Krüger, and D. Schröder, *Cusp universality for random matrices I: local law and the complex Hermitian case*, Comm. Math. Phys. **378**, 1203–1278 (2020), MR4134946.

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

[39] L. Erdős, B. Schlein, H.-T. Yau, and J. Yin, *The local relaxation flow approach to universality of the local statistics for random matrices*, Ann. Inst. Henri Poincaré Probab. Stat. **48**, 1–46 (2012), MR2919197.

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

[41] 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.

[42] L. Erdős and H.-T. Yau, *Gap universality of generalized Wigner and β-ensembles*, J. Eur. Math. Soc. (JEMS) **17**, 1927–2036 (2015), MR3372074.

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

[44] P. Erdős and A. Hajnal, *On chromatic number of graphs and set-systems*, Acta Math. Acad. Sci. Hungar. **17**, 61–99 (1966), MR193025.

[45] P. J. Forrester, *Fluctuation formula for complex random matrices*, J. Phys. A **32**, L159–L163 (1999), MR1687948.

[46] S. Geman, *The spectral radius of large random matrices*, Ann. Probab. **14**, 1318–1328 (1986), MR866352.

[47] V. L. Girko, *The circular law*, Teor. Veroyatnost. i Primenen. **29**, 669–679 (1984), MR773436.

[48] F. Götze and A. Tikhomirov, *The circular law for random matrices*, Ann. Probab. **38**, 1444–1491 (2010), MR2663633.

[49] Y. He, *Mesoscopic linear statistics of Wigner matrices of mixed symmetry class*, J. Stat. Phys. **175**, 932–959 (2019), MR3959983.

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

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

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

[53] K. Johansson, *On fluctuations of eigenvalues of random Hermitian matrices*, Duke Math. J. **91**, 151–204 (1998), MR1487983.

[54] O. Kallenberg, *Foundations of modern probability*, Second, Probability and its Applications (New York) (Springer-Verlag, New York, 2002), pp. xx+638, MR1876169.

[55] A. M. Khorunzhy, B. A. Khoruzhenko, and L. A. Pastur, *Asymptotic properties of large random matrices with independent entries*, J. Math. Phys. **37**, 5033–5060 (1996), MR1411619.

[56] P. Kopel, *Linear statistics of non-Hermitian matrices matching the real or complex Ginibre ensemble to four moments*, preprint (2015), arXiv:1510.02987.

[57] G. Lambert, M. Ledoux, and C. Webb, *Quantitative normal approximation of linear statistics of β-ensembles*, Ann. Probab. **47**, 2619–2685 (2019), MR4021234.

[58] B. Landon and P. Sosoe, *Applications of mesoscopic CLTs in random matrix theory*, Ann. Appl. Probab. **30**, 2769–2795 (2020), MR4187127.

[59] B. Landon, P. Sosoe, and H.-T. Yau, *Fixed energy universality of Dyson Brownian motion*, Adv. Math. **346**, 1137–1332 (2019), MR3914908.

[60] B. Landon and H.-T. Yau, *Convergence of local statistics of Dyson Brownian motion*, Comm. Math. Phys. **355**, 949–1000 (2017), MR3687212.

[61] T. Leblé and S. Serfaty, *Fluctuations of two dimensional Coulomb gases*, Geom. Funct. Anal. **28**, 443–508 (2018), MR3788208.

[62] T. Leblé and S. Serfaty, *Large deviation principle for empirical fields of log and Riesz gases*, Invent. Math. **210**, 645–757 (2017), MR3735628.

[63] 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.

[64] D. R. Lick and A. T. White, *k-degenerate graphs*, Canadian J. Math. **22**, 1082–1096 (1970), MR266812.

[65] E. H. Lieb and M. Loss, *Analysis*, Second, Vol. 14, Graduate Studies in Mathematics (American Mathematical Society, Providence, RI, 2001), pp. xxii+346, MR1817225.

[66] A. Lytova and L. Pastur, *Central limit theorem for linear eigenvalue statistics of random matrices with independent entries*, Ann. Probab. **37**, 1778–1840 (2009), MR2561434.

[67] M. L. Mehta, *Random matrices and the statistical theory of energy levels* (Academic Press, New York-London, 1967), pp. x+259, MR0220494.

[68] H. H. Nguyen and V. Vu, *Random matrices: law of the determinant*, Ann. Probab. **42**, 146–167 (2014), MR3161483.

[69] I. Nourdin and G. Peccati, *Universal Gaussian fluctuations of non-Hermitian matrix ensembles: from weak convergence to almost sure CLTs*, ALEA Lat. Am. J. Probab. Math. Stat. **7**, 341–375 (2010), MR2738319.

[70] S. O'Rourke and D. Renfrew, *Central limit theorem for linear eigenvalue statistics of elliptic random matrices*, J. Theoret. Probab. **29**, 1121–1191 (2016), MR3540493.

[71] G. Pan and W. Zhou, *Circular law, extreme singular values and potential theory*, J. Multivariate Anal. **101**, 645–656 (2010), MR2575411.

[72] B. Rider, *Deviations from the circular law*, Probab. Theory Related Fields **130**, 337–367 (2004), MR2095933.

[73] B. Rider and J. W. Silverstein, *Gaussian fluctuations for non-Hermitian random matrix ensembles*, Ann. Probab. **34**, 2118–2143 (2006), MR2294978.

[74] B. Rider and B. Virág, *The noise in the circular law and the Gaussian free field*, Int. Math. Res. Not. IMRN, Art. ID rnm006, 33 (2007), MR2361453.

[75] M. Shcherbina, *Central limit theorem for linear eigenvalue statistics of the Wigner and sample covariance random matrices*, Zh. Mat. Fiz. Anal. Geom. **7**, 176–192, 197, 199 (2011), MR2829615.

[76] M. Shcherbina, *Fluctuations of linear eigenvalue statistics of β matrix models in the multi-cut regime*, J. Stat. Phys. **151**, 1004–1034 (2013), MR3063494.

[77] S. Sheffield, *Gaussian free fields for mathematicians*, Probab. Theory Related Fields **139**, 521–541 (2007), MR2322706.

[78] N. J. Simm, *Central limit theorems for the real eigenvalues of large Gaussian random matrices*, Random Matrices Theory Appl. **6**, 1750002, 18 (2017), MR3612267.

[79] A. Soshnikov, *Gaussian limit for determinantal random point fields*, Ann. Probab. **30**, 171–187 (2002), MR1894104.

[80] P. Sosoe and P. Wong, *Regularity conditions in the CLT for linear eigenvalue statistics of Wigner matrices*, Adv. Math. **249**, 37–87 (2013), MR3116567.

[81] T. Tao and V. Vu, *Random matrices: the circular law*, Commun. Contemp. Math. **10**, 261–307 (2008), MR2409368.

[82] T. Tao and V. Vu, *Random matrices: universality of local spectral statistics of non-Hermitian matrices*, Ann. Probab. **43**, 782–874 (2015), MR3306005.

[83] T. Tao and V. Vu, *Smooth analysis of the condition number and the least singular value*, Math. Comp. **79**, 2333–2352 (2010), MR2684367.

[84] G. Téllez and P. J. Forrester, *Expanded Vandermonde powers and sum rules for the two-dimensional one-component plasma*, J. Stat. Phys. **148**, 824–855 (2012), MR2972857.

[85] H. Wang, *Quantitative universality for the largest eigenvalue of sample covariance matrices*, preprint (2019), arXiv:1912.05473.

[86] A. Zabrodin and P. Wiegmann, *Large-N expansion for the 2D Dyson gas*, J. Phys. A **39**, 8933–8963 (2006), MR2240466.