RANDOM MATRICES WITH SLOW CORRELATION DECAY

LÁSZLÓ ERDŐS† AND TORBEN KRÜGER† AND DOMINIK SCHRÖDER†‡

IST Austria, Am Campus 1, A-3400 Klosterneuburg, Austria

Abstract. We consider large random matrices with a general slowly decaying correlation among its entries. We prove universality of the local eigenvalue statistics and optimal local laws for the resolvent away from the spectral edges, generalizing the recent result of [1] to allow slow correlation decay and arbitrary expectation. The main novel tool is a systematic diagrammatic control of a multivariate cumulant expansion.

Contents

1. Introduction

In recent years it has been proven for increasingly general random matrix ensembles that their spectral measure converges to a deterministic measure up to the scale of individual eigenvalues as the size of the matrix tends to infinity, and that the fluctuation

E-mail address: dschroed@ist.ac.at, lerdos@ist.ac.at, tkrueger@ist.ac.at.

Date: June 1, 2020.

2010 *Mathematics Subject Classification.* 60B20, 15B52.

Key words and phrases. Local Law, Bulk Universality, Correlated Random Matrix, Multivariate Cumulant Expansion.

† Partially supported by ERC Advanced Grant No. 338804.

‡ Partially supported by the IST Austria Excellence Scholarship.

of the individual eigenvalues follows a universal distribution, independent of the specifics of the random matrix itself. The former is commonly called a *local law*, whereas the latter is known as the *Wigner-Dyson-Mehta (WDM) universality conjecture*, first envisioned by Wigner in the 1950's and formalized later by Dyson and Mehta in the 1960's [36]. In fact, the conjecture extends beyond the customary random matrix ensembles in probability theory and is believed to hold for any random operator in the delocalization regime of the Anderson metal-insulator phase transition. Given this profound universality conjecture for general disordered quantum systems, the ultimate goal of local spectral analysis of large random matrices is to prove the WDM conjecture for the largest possible class of matrix ensembles. In the current paper we complete this program for random matrices with a general, slow correlation decay among its matrix elements. Previous works covered only correlations with such a fast decay that, in a certain sense, they could be treated as a perturbation of the independent model. Here we present a new method that goes well beyond the perturbative regime. It relies on a novel multi-scale version of the cumulant expansion and its rigorous Feynman diagrammatic representation that can be useful for other problems as well. To put our work in context, we now explain the previous results.

In the last ten years a powerful new approach, the *three-step strategy* has been developed to resolve WDM universality problems, see [19] for a summary. In particular, the WDM conjecture in its classical form, stated for Wigner matrices with a general distribution of the entries, has been proven with this strategy in [14, 15, 21]; an independent proof for the Hermitian symmetry class was given in [42]. Recent advances have crystallized that the only model dependent step in this strategy is the first one, the local law. The other two steps, the fast relaxation to equilibrium of the Dyson Brownian motion and the approximation by Gaussian divisible ensembles, have been formulated as very general "black-box" tools whose only input is the local law [17, 31, 32]. Thus the proof of the WDM universality, at least for mean field ensembles, is automatically reduced to obtaining a local law.

Both local law and universality have first been established for *Wigner matrices*, which are real symmetric or complex Hermitian $N \times N$ matrices with mean-zero entries which are independent and identically distributed (i.i.d.) up to symmetry [15, 16]. For Wigner matrices it has long been known that the *limiting*, or *self-consistent* density is the *Wigner semicircle law*. In subsequent work the condition on the i.i.d. entries has been relaxed in several steps. First, it was proven in [21], that for *generalized Wigner ensembles*, i.e., for matrices with stochastic variance profile and uniform upper and lower bound on the variance of the matrix entries, the local law and universality also hold, with the self-consistent density still given by the semicircle law. Next, the condition of stochasticity was removed by introducing the *Wigner-type* ensemble [3], in which case the self-consistent density is, generally, not semicircular any more. Finally, the independence condition was dropped and in [1] both a local law on the optimal local scale and bulk universality were obtained for matrices with correlated entries with fast decaying general correlations. Special correlation structures were also considered before in [2, 11] on a local scale. We also mention that there exists an extensive literature on the global law for random matrices with correlated entries [6, 8, 9, 25, 26, 39, 40]. These results, however, either concern Gaussian random matrices or more specific correlation structures than considered in the present work. In a parallel development the zero-mean condition on the matrix elements has also been relaxed. First this was achieved for the *deformed Wigner ensembles* that have diagonal deterministic shifts in [34, 37] and more recently for i.i.d. Wigner matrices shifted by an arbitrary deterministic matrix in [28].

In this paper we prove a local law and bulk universality for random matrices with a slowly decaying correlation structure and arbitrary expectation, generalizing both [1, 28]. The main point is to considerably relax the condition on the decay of correlations compared to [1]: We allow for a polynomial decay of order two in a neighbourhood of size $\ll \sqrt{N}$ around every entry and we only have to assume a polynomial decay of a certain finite order outside these neighbourhoods. Another novelty is that our new concept of neighbourhoods is completely general, it is not induced by the product structure of the index set labelling the matrix elements. In particular, the improved correlation condition also includes many other matrix models of interest, for example, general block matrix type models, that have not been covered by [1].

Regarding strategy of proving the local law, the starting point is to find the deterministic approximation of the resolvent $G(z) = (H - z)^{-1}$ of the random matrix H with a complex spectral parameter z in the upper half plane $\mathbb{H} = \{ z \in \mathbb{C} \mid \Im z \geq 0 \}$. This approximation is given as the solution $M = M(z)$ to the *Matrix Dyson Equation (MDE)*

$$1 + (z - A + \mathcal{S}[M])M = 0,$$

where the expectation matrix $A \coloneqq \mathbf{E}\, H$ and the linear map $\mathcal{S}[V] \coloneqq \mathbf{E}(H - A)V(H - A)$ on the space of matrices R encode the first two moments of the random matrix. The resolvent approximately satisfies the MDE with an additive perturbation term

$$D \coloneqq (H - A)G + \mathcal{S}[G]G.$$

The smallness of D and stability of the MDE against small perturbations imply that G is indeed close to M. The necessary stability properties of the MDE have already been established in [1], so the main focus in this paper is to bound D in appropriate norms that can then be fed into the stability analysis. Most proofs of the previous local laws loosely follow a strategy of first reducing the problem to a smaller number of relevant variables, such as the diagonal entries of G. Instead, correlated ensembles require to carry out the analysis genuinely on the matrix level since G is not even approximately diagonal. This key feature distinguishes the current paper as well as [1] from all previous works, where the Dyson equation was only a scalar equation for the trace of the resolvent or a vector equation for its diagonal elements. Although adding a general expectation matrix A to a Wigner matrix already induces a non-diagonal resolvent, diagonalization of A reduced the analysis to the scalar level in [28]. A similar algebraic reduction is not possible for general correlations even if they decay as fast as in [1]. However, in [1] every matrix quantity, such as G or M, still had a very fast off-diagonal decay and thus it was sufficient to focus only on matrix elements very

close to the diagonal; the rest was treated as an irrelevant error. For the slow correlation decay considered in this paper such direct perturbative treatment for the off-diagonal elements is not possible. In fact, with our new method we can even handle the essentially optimal integrable correlation decay on a scale $\sqrt{N}$ near the diagonal.

To obtain a probabilistic bound on D, essentially two approaches are available. When G is approximately diagonal and when the columns of H are independent, one may use resolvent expansion formulas involving minors that lead to standard linear and quadratic large deviation bounds – a natural idea that first arose in the works of Girko and Pastur [24, 38], as well as in the works of Bai et. al., e.g. [7]. For correlated models the natural extension of this method requires a somewhat involved successive expansion of minors; this was the main technical tool in [1]. This approach is thus restricted to very fast correlation decay since it is essentially a perturbation around nearly diagonal matrices. The alternative method relies on the cumulant expansion of the form $\mathbf{E}\, hf(h) = \sum_k (\kappa_{k+1}/k!)\, \mathbf{E}\, f^{(k)}$, where κ_k is the k-th order cumulant of the random variable h. The power of this expansion in studying resolvents of random matrices was first recognized in [30] and it has been revived in several recent papers, e.g. [18, 27, 33]. It gives more flexibility than the minor expansion on two accounts. First, it can handle the stochastic effect of individual matrix elements instead of treating an entire column at the same time. This observation was essential in [28] to handle deformations of Wigner matrices with an arbitrary expectation matrix. Single entry expansions, as opposed to expansion by entire columns, also appeared in the proof of a version of the *fluctuation averaging theorem* [22], but in this context it did not have any major advantage over the row expansions. Secondly, a multivariate version of the cumulant expansion is inherently well suited to correlated models; it automatically keeps track of the correlation structure without artificial cut-offs and strong restrictions on the off-diagonal decay. This is the method we use to bound D in the current work to handle the slow correlation decay effectively.

After presenting our main results in Section 2, in Section 3 we first give a multivariate cumulant expansion formula with an explicit error term that is especially well suited for mean field random matrix models. The main ingredient is a novel *pre-cumulant decoupling identity*, Lemma 3.1. We were not able to find these formulas in the literature; related formulas, however, have probably been known. They are reminiscent to the Wick polynomials, their relationship is explained in Appendix B. Some consequences are collected in Section 3.3 via a toy model. When applying it to our problem, in order to bookkeep the numerous terms, we develop a graphical language which allows us to actually compute $\mathbf{E}\,|\Lambda(D)|^p$ up to a tiny error for arbitrary linear functionals Λ. The structure of D contains an essential cancellation: the term $(H-A)G$ is compensated by $\mathcal{S}[G]G$ that acts as a counter term or *self-energy renormalization* in the physics terminology. Our cumulant expansion automatically exploits this cancellation to all orders and the diagrammatic representation in Sections 4.1–4.4 conveniently visualizes this mechanism. Section 4 contains the main novel part of this paper, in Section 5 we combine the bounds on D with the stability argument for the MDE to prove the local law. Section 6 is devoted to the short proofs of bulk universality and other natural corollaries of the local law.

Acknowledgements. T.K. gratefully acknowledges private communications with Antti Knowles on the preliminary version of [28]. D.S. would like to thank Nikolaos Zygouras for raising the question how our novel pre-cumulants are related to Wick polynomials.

2. Main results

For a Hermitian $N \times N$ random matrix $H = H^{(N)}$ we denote its resolvent by

$$G(z) = G^{(N)}(z) = (H - z)^{-1},$$

where the spectral parameter z is assumed to be in the upper half plane $\mathbb{H}$. The first two moments of H determine the limiting behaviour of $G(z)$ in the large N limit. More specifically, let

$$A \coloneqq \mathbf{E}\, H, \qquad H =: A + \frac{1}{\sqrt{N}} W, \qquad \mathcal{S}[V] \coloneqq \frac{1}{N}\, \mathbf{E}\, WVW,$$

where $\mathcal{S}$ is a linear map on the space of $N \times N$ matrices and W is a random matrix with zero expectation. Then the unique, deterministic solution $M = M(z)$ to the matrix Dyson equation (MDE)

$$1 + (z - A + \mathcal{S}[M])M = 0 \qquad \text{under the constraint} \qquad \Im M \coloneqq \frac{1}{2\mathrm{i}}[M - M^*] > 0, \tag{2.1}$$

approximates the random matrix $G(z)$ increasingly well as N tends to ∞. Here $\Im M > 0$ indicates that the matrix $\Im M$ is positive definite. The properties of (2.1) and its solution have been comprehensively studied in [1]. In particular, it has been shown that

$$\frac{1}{N} \operatorname{Tr} M(z) = \int_{\mathbb{R}} \frac{1}{x - z}\, \mathrm{d}\mu(x) \tag{2.2}$$

is the Stieltjes transform of a measure μ on $\mathbb{R}$, which we call the *self-consistent density of states*, and whose support $\operatorname{supp}\mu$ we call the *self-consistent spectrum*. Under an additional flatness Assumption (see Assumption (E) later) it has also been shown that μ is absolutely continuous with compactly supported Hölder continuous probability density

$$\mathrm{d}\mu(x) = \rho(x)\, \mathrm{d}x \qquad \text{and that} \qquad \rho(z) \coloneqq \frac{1}{\pi N} \Im \operatorname{Tr} M(z) \tag{2.3}$$

is the harmonic extension of $\rho\colon \mathbb{R} \to [0,\infty)$. Moreover, (2.1) is stable with respect to small additive perturbations and therefore it is sufficient to show that the error matrix $D = D(z)$ defined by

$$D := 1 + (z - A + \mathcal{S}[G])G = (H - A + \mathcal{S}[G])G = \frac{W}{\sqrt{N}}G + \mathcal{S}[G]G \tag{2.4}$$

is small.

Choosing the correct norm to measure smallness of the error terms is a key technical ingredient. Similarly to the resolvent G, the error matrix D is very large in the usual induced $\ell^p \to \ell^q$ matrix norms, but its quadratic form $\langle \mathbf{x}, D\mathbf{y}\rangle$ is under control with very high probability for any fixed deterministic vectors $\mathbf{x}, \mathbf{y}$. Furthermore, to improve precision, we will distinguish two different concepts of measuring the size of D. We will show that D can be bounded in *isotropic sense* as $|\langle \mathbf{x}, D\mathbf{y}\rangle| \lesssim \|\mathbf{x}\|\,\|\mathbf{y}\|\,/\sqrt{N\Im z}$ for fixed deterministic vectors $\mathbf{x}, \mathbf{y}$ as well as in an *averaged sense* as $N^{-1}\,|\operatorname{Tr} BD| \lesssim \|B\|\,/N\Im z$ for fixed deterministic matrices B. Here $\|\mathbf{x}\|, \|\mathbf{y}\|, \|B\|$ denote the standard (Euclidean) vector norm $\|\mathbf{x}\|^2 = \sum_a |x_a|^2$ and (matrix) operator norm $\|B\| := \sup_{\|\mathbf{x}\|,\|\mathbf{y}\|\le 1} |\langle \mathbf{x}, B\mathbf{y}\rangle|$. The second step of the proof will be to show that because D is small, and (2.1) is stable under small additive perturbations, also $G - M$ is small in an appropriate sense.

2.1. **Notations and conventions**

An inequality with a subscript indicates that we allow for a constant in the bound depending only on the quantities in the subscript. For example, $A(N,\epsilon) \le_\epsilon B(N,\epsilon)$ means that there exists a constant $C = C(\epsilon)$, independent of N, such that $A(N,\epsilon) \le C(\epsilon)B(N,\epsilon)$ holds for all N and $\epsilon > 0$. In many statements we will implicitly assume that N is sufficiently large, depending on any other parameters of the model. Moreover, we will write $f \sim g$ if $f = \mathcal{O}(g)$ and $g = \mathcal{O}(f)$, if it is clear from the context in which regime we claim this comparability and how the implicit constant may depend on parameters.

An abstract index set J of size N labels the rows and columns of our matrix (generally one can think of $J = [N] := \{1, \dots, N\}$ but there is no need for having a (partial) order or a notion of distance on J). The elements of J will be denoted by letters $a, b, \dots$ and $i, j, \dots$ from the beginning of the alphabet. We will use boldfaced letters $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v}, \dots$ from the end of the alphabet to denote J-vectors with entries $\mathbf{x} = (x_a)_{a\in J}$. We will denote the set of ordered pairs of indices by $I := J \times J$ and will often call the elements of I *labels* to avoid confusion with other types of indices, and will denote them by Greek letters $\alpha = (a,b) \in I$. The matrix element w_{ab} will thus often be denoted by w_α. Summations of the form $\sum_a$ and $\sum_\alpha$ are always understood to sum over all $a \in J$ and $\alpha \in I$.

For indices $a, b \in J$ and vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^J$ we shall use the notations

$$A_{\mathbf{xy}} := \langle \mathbf{x}, A\mathbf{y}\rangle, \qquad A_{\mathbf{x}a} := \langle \mathbf{x}, Ae_a\rangle, \qquad A_{a\mathbf{x}} := \langle e_a, A\mathbf{x}\rangle,$$

where e_a is the a-th standard basis vector. We will frequently write $\Delta^{ab} = e_a e_b^t$ for the matrix of all zeros except a one in the (a,b) entry. The normalized trace of an $N \times N$ matrix is denoted by $\langle A\rangle := N^{-1}\operatorname{Tr} A$. Sometimes we will also use the notation $\langle z\rangle := 1 + |z|$ for the complex number z, but this should not create confusions as it will only be used for z. We will furthermore use the maximum norm and the normalized Hilbert-Schmidt norm

$$\|A\|_{\max} := \max_{a,b} |A_{ab}|, \qquad \|A\|_{\mathrm{hs}} := \Big[\frac{1}{N}\sum_{a,b} |A_{ab}|^2\Big]^{1/2}$$

for an $N \times N$ matrix A.

2.2. **Assumptions**

We now formulate our main assumptions on W and A.

Assumption (A) (Bounded expectation). *There exists some constant C such that $\|A\| \le C$ for all N.*

Assumption (B) (Finite moments). *For all $q \in \mathbb{N}$ there exists a constant μ_q such that $\mathbf{E}\,|w_\alpha|^q \le \mu_q$ for all α.*

Next, we formulate our conditions on the correlation decay conveniently phrased in terms of the multivariate cumulants κ of random variables of $\{ w_\alpha \mid \alpha \in I \}$. In Appendix A we recall the definition and some basic properties of multivariate cumulants. First we present a simple condition in terms of a tree type ρ-mixing decay of the cumulants with respect to the standard Euclidean metric on $[N]^2$. Later, in Section 2.5, we formulate weaker and more general conditions which we actually use for the proof of our results but their formulation is quite involved, so for the sake of clarity we first rather state simpler but more restrictive assumptions.

Consider $J = [N]$, $I = [N]^2$ equipped with the standard Euclidean distance modulo the Hermitian symmetry, i.e., for $\alpha, \beta \in I$ we set $d(\alpha,\beta) := \min\{|\alpha - \beta|, |\alpha^t - \beta|\}$ where $\alpha^t := (b,a)$ for $\alpha = (a,b)$. This distance naturally extends to subsets of I, i.e., $d(A,B) = \min\{ d(\alpha,\beta) \mid \alpha \in A, \beta \in B\}$ for any $A, B \subset I$.

Assumption (CD) (Polynomially decaying metric correlation structure). *For the $k = 2$ point correlation we assume a decay of the type*

$$|\kappa(f_1(W), f_2(W))| \le \frac{C}{1 + d(\operatorname{supp} f_1, \operatorname{supp} f_2)^s}\,\|f_1\|_2\,\|f_2\|_2, \tag{2.5a}$$

for some $s > 12$ and all square integrable functions f_1, f_2 on $N \times N$ matrices. For $k \geq 3$ we assume a decay condition of the form

$$|\kappa(f_1(W), \ldots, f_k(W))| \leq_k \prod_{e \in E(T_{min})} |\kappa(e)|\,, \tag{2.5b}$$

where T_{min} is the minimal spanning tree in the complete graph on the vertices $1, \ldots, k$ with respect to the edge length $d(\{i,j\}) = d(\operatorname{supp} f_i, \operatorname{supp} f_j)$, i.e., the tree for which the sum of the lengths $d(e)$ is minimal, and $\kappa(\{i,j\}) = \kappa(f_i, f_j)$.

A correlation decay of type (2.5b) is typical for various statistical physics models, see, e.g. [12]. Besides the assumptions on the decay of correlations we also impose a *flatness condition* to guarantee the stability of the Dyson equation:

Assumption (E) (Flatness)**.** *There exist constants $0 < c < C$ such that*

$$c \langle T \rangle \leq \mathcal{S}[T] \leq C \langle T \rangle$$

for any positive semi-definite matrix T.

Flatness is a certain *mean field* condition on the random matrix W. In particular, choosing T to be the diagonal matrix with a single nonzero entry in the (i,i) element, flatness implies that the variances of the matrix elements $\mathbf{E}\,|w_{ij}|^2$ are comparable for all $i, j = 1, \ldots, N$.

2.3. **Local law**

We now formulate our main theorem on the isotropic and averaged local laws. They compare the resolvent G with the (unique) solution to the MDE in (2.1) away from the spectral edges. To specify the range of spectral parameters z we define two spectral domains specified via any given parameters $\delta, \gamma > 0$. Outside of the self-consistent spectrum we will work on

$$\mathbb{D}^{\delta}_{\text{out}} := \left\{ z \in \mathbb{H} \;\middle|\; |z| \leq N^{C_0}, \operatorname{dist}(z, \operatorname{supp} \mu) \geq N^{-\delta} \right\}$$

for some arbitrary fixed $C_0 \geq 100$. Under Assumption (E), which guarantees the existence of a density ρ, we consider the spectral domains

$$\mathbb{D}^{\delta}_{\gamma} := \left\{ z \in \mathbb{H} \;\middle|\; |z| \leq N^{C_0},\ \Im z \geq N^{-1+\gamma},\ \rho(\Re z) + \operatorname{dist}(\Re z, \operatorname{supp} \mu) \geq N^{-\delta} \right\}$$

that will be used away from the edges of the self-consistent spectrum.

Theorem 2.1 (Local law outside of the spectrum and global law)**.** *Under Assumptions (A), (B) and (CD), the following statements hold: For any $\epsilon > 0$ there exists $\delta > 0$ such that for all $D > 0$ we have the isotropic law away from the spectrum,*

$$\mathbf{P}\left(|\langle \mathbf{x}, (G - M)\mathbf{y} \rangle| \leq \|\mathbf{x}\|\, \|\mathbf{y}\| \frac{N^{\epsilon}}{\langle z \rangle^2 \sqrt{N}} \quad \text{in} \quad \mathbb{D}^{\delta}_{\text{out}} \right) \geq 1 - CN^{-D} \tag{2.6a}$$

for all deterministic vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$ and we have the averaged law away from the spectrum,

$$\mathbf{P}\left(|\langle B(G - M) \rangle| \leq \|B\| \frac{N^{\epsilon}}{\langle z \rangle^2 N} \quad \text{in} \quad \mathbb{D}^{\delta}_{\text{out}} \right) \geq 1 - CN^{-D} \tag{2.6b}$$

for all deterministic matrices $B \in \mathbb{C}^{N \times N}$. In fact, for small ϵ, δ can be chosen such that $\delta = c\epsilon$ for some absolute constant $c > 0$. Here $G = G(z)$, $M = M(z)$ and $C = C(D, \epsilon)$ is some constant, depending only on its arguments and the constants in Assumptions (A)–(CD). Moreover, instead of Assumption (CD) it is sufficient to assume the more general Assumptions (C) (or (C)' for complex Hermitian matrices) and (D), as stated in Section 2.5.

If we additionally assume flatness in the form of Assumption (E), then we also obtain an optimal local law away from the spectral edges, especially in the bulk,

Theorem 2.2 (Local law in the bulk of the spectrum)**.** *Under Assumptions (A), (B), (CD) and (E), the following statements hold: For any $\gamma, \epsilon > 0$ there exists $\delta > 0$ such that for all $D > 0$ we have the isotropic law in the bulk,*

$$\mathbf{P}\left(|\langle \mathbf{x}, (G - M)\mathbf{y} \rangle| \leq \|\mathbf{x}\|\, \|\mathbf{y}\| \frac{N^{\epsilon}}{\sqrt{N \Im z}} \quad \text{in} \quad \mathbb{D}^{\delta}_{\gamma} \right) \geq 1 - CN^{-D} \tag{2.7a}$$

for all deterministic vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$ and we have the averaged law in the bulk,

$$\mathbf{P}\left(|\langle B(G - M) \rangle| \leq \|B\| \frac{N^{\epsilon}}{N \Im z} \quad \text{in} \quad \mathbb{D}^{\delta}_{\gamma} \right) \geq 1 - CN^{-D} \tag{2.7b}$$

for all deterministic matrices $B \in \mathbb{C}^{N \times N}$. In fact, δ can be chosen such that $\delta = c \min\{\epsilon, \gamma\}$ for some absolute constant $c > 0$. Here $C = C(D, \epsilon, \gamma)$ is some constant, depending only on its arguments and the constants in Assumptions (A)–(E). Moreover, as in the previous theorem, instead of Assumption (CD) it is sufficient to assume the more general Assumptions (C) (or (C)' for complex Hermitian matrices) and (D), as stated in Section 2.5.

Note that both theorems cover the regime where z is far away from the spectrum; in this case the estimates in Theorem 2.1 are stronger and require less conditions. Theorem 2.2 is really relevant when $\Re z$ is inside the bulk of the spectrum and $\Im z$ is very small; this is why we called it local law in the bulk. In the literature this regime is typically characterized by $\rho(\Re z) \geq \delta$ for some $\delta > 0$, but in Theorem 2.2 it is extended to $\rho(\Re z) \geq N^{-\delta}$ for some sufficiently small $\delta > 0$.

2.4. **Delocalization, rigidity and universality**

The local law is the main input for eigenvector delocalization, eigenvalue rigidity and universality, as stated below. We formulate them as corollaries since they follow from a general theory that has been developed recently. We explain how to adapt the general arguments to prove these corollaries in Sections 5.4 and 6.

Corollary 2.3 (No eigenvalues outside the support of the self-consistent density)**.** *Under the assumptions of Theorem 2.1 there exists a $\delta > 0$ such that for any $D > 0$,*

$$\mathbf{P}\left(\operatorname{Spec} H \not\subset (-N^{-\delta}, N^{-\delta}) + \operatorname{supp}\mu\right) \leq_D N^{-D},$$

where $\operatorname{supp}\mu \subset \mathbb{R}$ is the support of the self-consistent density of states μ.

Corollary 2.4 (Bulk delocalization)**.** *Under the assumptions of Theorem 2.2 it holds for an ℓ^2-normalized eigenvector $\boldsymbol{u}$ corresponding to a bulk eigenvalue λ of H that*

$$\mathbf{P}\left(\max_{a\in J}|u_a| \geq \frac{N^\epsilon}{\sqrt{N}},\ H\mathbf{u} = \lambda\mathbf{u},\ \rho(\lambda) \geq \delta\right) \leq_{\epsilon,\delta,D} N^{-D}$$

for any $\epsilon, \delta, D > 0$.

Corollary 2.5 (Bulk rigidity)**.** *Under the assumptions of Theorem 2.2 the following holds. Let $\lambda_1 \leq \cdots \leq \lambda_N$ be the ordered eigenvalues of H and denote the classical position of the eigenvalue close to energy $E \in \mathbb{R}$ by*

$$k(E) := \left\lceil N\int_{-\infty}^{E} \rho(x)\,\mathrm{d}x \right\rceil,$$

where $\lceil\cdot\rceil$ denotes the ceiling function. It then holds that

$$\mathbf{P}\left(\sup\left\{\ |\lambda_{k(E)} - E|\ \middle|\ E \in \mathbb{R},\ \rho(E) \geq \delta\ \right\} \geq \frac{N^\epsilon}{N}\right) \leq_{\epsilon,\delta,D} N^{-D}$$

for any $\epsilon, \delta, D > 0$.

For proving the bulk universality we replace the lower bound from Assumption (E) by the following, stronger, assumption:

Assumption (F) (Fullness)**.** *There exists a constant $\lambda > 0$ such that*

$$\mathbf{E}\,|\operatorname{Tr} BW|^2 \geq \lambda \operatorname{Tr} B^2$$

for any deterministic matrix B of the same symmetry class as H.

Fullness is a technical condition which ensures that the covariance matrix of W is bounded from below by that of a full GUE or GOE matrix with variance λ. Note this is the only condition that induces the difference between the complex Hermitian and real symmetric symmetry classes in the following universality statement.

Corollary 2.6 (Bulk universality)**.** *Under the assumptions of Theorem 2.2 and additionally Assumption (F) the following holds: Let $k \in \mathbb{N}$, $\delta > 0$, $E \in \mathbb{R}$ with $\rho(E) \geq \delta$ and let $\Phi\colon \mathbb{R}^k \to \mathbb{R}$ be a compactly supported smooth test function. Denote the k-point correlation function of the eigenvalues of H by ρ_k and denote the corresponding k-point correlation function of the GOE/GUE-point process by Υ_k. Then there exists a positive constant $c = c(\delta, k) > 0$ such that*

$$\left|\int_{\mathbb{R}^k} \Phi(\boldsymbol{t})\left[\frac{1}{\rho(E)}\rho_k\Big(E\mathbf{1} + \frac{\boldsymbol{t}}{N\rho(E)}\Big) - \Upsilon_k(\boldsymbol{t})\right]\mathrm{d}\boldsymbol{t}\right| \leq_{\Phi,\delta,k} N^{-c},$$
$$\left|\mathbf{E}\,\Phi\Big(\big(N\rho(\lambda_{k(E)})[\lambda_{k(E)+j} - \lambda_{k(E)}]\big)_{j=1}^k\Big) - \mathbf{E}_{GOE/GUE}\,\Phi\Big(\big(N\rho_{sc}(0)[\lambda_{\lceil N/2\rceil+j} - \lambda_{\lceil N/2\rceil}]\big)_{j=1}^k\Big)\right| \leq_{\Phi,\delta,k} N^{-c},$$

where $\mathbf{1}$ is the vector of k ones, $\mathbf{1} = (1, \ldots, 1)$, the expectation $\mathbf{E}_{GOE/GUE}$ is taken with respect to the Gaussian matrix ensemble in the same symmetry class as H, and ρ_{sc} denotes the semicircular density.

Remark 2.7. *We chose the standard Euclidean distance on J in the formulation of Assumption (CD) merely for convenience. In the context of [1] a similar key assumption was formulated in terms of a pseudometric δ on J which has sub-P dimensional volume, i.e.,*

$$\max_{a\in J}|\{\ b \in J\ |\ \delta(a,b) \leq \tau\ \}| \leq \tau^P$$

for all $\tau > 1$ and some $P > 0$. This pseudometric naturally extends to I as a product metric modulo the symmetry,

$$\delta_2((a,b),(c,d)) := \min\{\max\{\delta(a,c),\delta(b,d)\},\max\{\delta(a,d),\delta(b,c)\}\}$$

and to any two subsets A, B of I as $\delta_2(A,B) := \min\{\ \delta_2(\alpha,\beta)\ |\ \alpha \in A, \beta \in B\ \}$. All our results hold in this more general setup as well if d is replaced by δ_2 in Assumption (CD) and we require that $s > 12P$. We do not pursue the pseudometric formulation further in the present work since the relaxed decay conditions formulated in Section 2.5 are more general as they allow for further symmetries in the matrix, for which (CD) is not satisfied irrespective of the pseudometric. A typical example for such an additional symmetry is the fourfold model (see [4]).

2.5. **Relaxed assumption on correlation decay**

We now state the more general conditions on the correlation structure which are actually used in the proof of Theorem 2.2 and its corollaries, and are implied by Assumption (CD). For the more general conditions we split the correlation into two regimes. In the short range regime we express the correlation decay as a condition on cumulants, while in the long range regime, beyond neighbourhoods of size $\sqrt{N}$, we impose a mixing condition.

In the short range regime we assume the boundedness of certain norms on cumulants $\kappa(\alpha_1,\dots,\alpha_k) := \kappa(w_{\alpha_1},\dots,w_{\alpha_k})$ of matrix entries w_α, which are modifications of the usual ℓ^1-summability condition

$$\frac{1}{N^2}\sum_{\alpha_1,\dots,\alpha_k} |\kappa(\alpha_1,\dots,\alpha_k)| < \infty.$$

Cumulant norms. In order to formulate the conditions on the cumulants concisely, we from now on assume that W is real symmetric. We refer the reader to Appendix C for the necessary modifications for the complex Hermitian case. In Appendix A we will recall the equivalent analytical and combinatorial definitions of κ for the reader's convenience (see also [41]). We note that κ is invariant under any permutation of its arguments. Here we recall one central property of cumulants (which is also proved in the appendix): If $w_{\alpha_1},\dots,w_{\alpha_j}$ are independent from $w_{\alpha_{j+1}},\dots,w_{\alpha_k}$ for some $1\le j\le k-1$, then $\kappa(\alpha_1,\dots,\alpha_k)$ vanishes. Intuitively, the k-th order cumulant $\kappa(\alpha_1,\dots,\alpha_k)$ measures the part of the correlation of $w_{\alpha_1},\dots,w_{\alpha_k}$, which is truly of k-body type. For our results, cumulants of order four and higher require simple ℓ^1-type bounds, while the second and third order cumulants are controlled in specific, somewhat stronger norms. Finiteness of these norms imply a decay of correlation in a certain combinatorial sense even without a distance on the index set I. The isotropic and the averaged bound on D require slightly different norms, so we define two sets of norms distinguished by appropriate superscripts and we also define their sums without superscript.

We first introduce some custom notations which keep the definition of the cumulant norms relatively compact. If, in place of an index $a\in J$, we write a dot $(\cdot)$ in a scalar quantity then we consider the quantity as a vector indexed by the coordinate at the place of the dot. For example $\kappa(a_1\cdot,a_2b_2)$ is a J-vector, the i-th entry of which is $\kappa(a_1i,a_2b_2)$, and $\|\kappa(a_1\cdot,a_2b_2)\|$ is its (Euclidean) vector norm. Similarly, $\|A(*,*)\|$ refers to the operator norm of the matrix with matrix elements $A(i,j)$. We also define a combination of these conventions, in particular $\big\|\,\|\kappa(\mathbf{x}*,*\cdot)\|\,\big\|$ will denote the operator norm $\|A\|$ of the matrix A with matrix elements $A(i,j)=\|\kappa(\mathbf{x}i,j\cdot)\| = \big\|\sum_a x_a\kappa(ai,j\cdot)\big\|$. Since $\|A\|=\|A^t\|$ this does not introduce ambiguities with respect of the order of i,j. Notice that we use dot $(\cdot)$ for the dummy variable related to the inner norm and star $(*)$ for the outer norm.

For k-th order cumulants we set

$$|||\kappa|||_k := |||\kappa|||_k^{\mathrm{av}} + |||\kappa|||_k^{\mathrm{iso}}, \qquad |||\kappa|||^{\mathrm{av/iso}} = |||\kappa|||^{\mathrm{av/iso}}_{\le R} := \max_{2\le k\le R} |||\kappa|||_k^{\mathrm{av/iso}}, \tag{2.8a}$$

where the averaged norms are given by

$$\begin{aligned} |||\kappa|||_2^{\mathrm{av}} &:= \big\|\,|\kappa(*,*)|\,\big\|, \qquad |||\kappa|||_k^{\mathrm{av}} := N^{-2}\sum_{\alpha_1,\dots,\alpha_k}|\kappa(\alpha_1,\dots,\alpha_k)|, \qquad k\ge 4,\\ |||\kappa|||_3^{\mathrm{av}} &:= \Big\|\sum_{\alpha_1}|\kappa(\alpha_1,*,*)|\Big\| + \inf_{\kappa=\kappa_{dd}+\kappa_{dc}+\kappa_{cd}+\kappa_{cc}}\Big(|||\kappa_{dd}|||_{dd}+|||\kappa_{dc}|||_{dc}+|||\kappa_{cd}|||_{cd}+|||\kappa_{cc}|||_{cc}\Big) \end{aligned} \tag{2.8b}$$

and the infimum is taken over all decompositions of κ in four symmetric functions κ_{dd},κ_{cd}, etc. The letters d and c refer to "direct" and "cross", see Remark 2.8 below. The corresponding norms are given by

$$\begin{aligned} |||\kappa|||_{cc} = |||\kappa|||_{dd} &:= N^{-1}\sqrt{\sum_{b_2,a_3}\Big(\sum_{a_2,b_3}\sum_{\alpha_1}|\kappa(\alpha_1,a_2b_2,a_3b_3)|\Big)^2},\\ |||\kappa|||_{cd} &:= N^{-1}\sqrt{\sum_{b_3,a_1}\Big(\sum_{a_3,b_1}\sum_{\alpha_2}|\kappa(a_1b_1,\alpha_2,a_3b_3)|\Big)^2}, \quad |||\kappa|||_{dc} := N^{-1}\sqrt{\sum_{b_1,a_2}\Big(\sum_{a_1,b_2}\sum_{\alpha_3}|\kappa(a_1b_1,a_2b_2,\alpha_3)|\Big)^2}. \end{aligned}$$

For the isotropic bound we define

$$\begin{aligned} |||\kappa|||_2^{\mathrm{iso}} &:= \inf_{\kappa=\kappa_d+\kappa_c}\big(|||\kappa_d|||_d+|||\kappa_c|||_c\big) \qquad |||\kappa|||_d := \sup_{\|\mathbf{x}\|\le 1}\big\|\,\|\kappa(\mathbf{x}*,\cdot*)\|\,\big\| \qquad |||\kappa|||_c := \sup_{\|\mathbf{x}\|\le 1}\big\|\,\|\kappa(\mathbf{x}*,*\cdot)\|\,\big\|,\\ |||\kappa|||_k^{\mathrm{iso}} &:= \Big\|\sum_{\alpha_1,\dots,\alpha_{k-2}}|\kappa(\alpha_1,\dots,\alpha_{k-2},*,*)|\Big\|, \qquad k\ge 3, \end{aligned} \tag{2.8c}$$

where the inner norms in (2.8c) indicate vector norms and the outer norms operator norms, and the infimum is taken over all decomposition of κ into the sum of symmetric κ_c and κ_d.

Remark 2.8. *We remark that the particular form of the norms $|||\kappa|||_2^{iso}$ and $|||\kappa|||_3^{av}$ on κ is chosen to conform with the Hermitian symmetry. For example, in the case of Wigner matrices we have*

$$\kappa(a_1b_1,a_2b_2) = \delta_{a_1,a_2}\delta_{b_1,b_2}+\delta_{a_1,b_2}\delta_{b_1,a_2} =: \kappa_d(a_1b_1,a_2b_2)+\kappa_c(a_1b_1,a_2b_2), \tag{2.9}$$

i.e., the cumulant naturally splits into a direct and a cross part κ_d and κ_c. In general, the splitting $\kappa=\kappa_c+\kappa_d$ may not be unique but for the sharpest bound we can consider the most optimal splitting; this is reflected in the infimum in the definition of $|||\kappa|||_2^{iso}$. Note that in the example (2.9) $|||\kappa_d|||_d$ and $|||\kappa_c|||_c$ are bounded, but $|||\kappa_c|||_d$ would not be. A similar rationale stands behind the definition of $|||\kappa|||_3^{av}$.

We also remark that only the conditions on $|||\kappa|||_2^{iso}$ and $|||\kappa|||_3^{av}$ use the product structure $I = J \times J$. All other decay conditions are inherently conditions on index pairs $\alpha \in I$.

Assumption (C) (κ–correlation decay)**.** *There exists a constant C such that for all $R \in \mathbb{N}$ and $\epsilon > 0$*

$$|||\kappa|||_2^{iso} \le C, \qquad |||\kappa||| = |||\kappa|||_{\le R} := \max_{2 \le k \le R} |||\kappa|||_k \le_{\epsilon,R} N^{\epsilon}$$

where the norms $|||\cdot|||_k$ and $|||\cdot|||_2^{iso}$ on k-th order cumulants were defined in (2.8). If the matrix W is complex Hermitian we use Assumption (C)', as stated in Appendix C instead of Assumption (C).

Furthermore, in the long range regime beyond certain neighbourhoods of size $\ll \sqrt{N}$ we assume a finite polynomial decay of correlations that is reminiscent of the standard ρ-mixing condition in statistical physics (see, e.g. [10] for an overview of various mixing conditions). We will need this decay in a certain iterated sense that we now formulate precisely.

Assumption (D) (Higher order correlation decay)**.** *There exists $\mu > 0$ such that the following holds: For every $\alpha \in I$ and $q, R \in \mathbb{N}$ there exists a sequence of nested sets $\mathcal{N}_k = \mathcal{N}_k(\alpha)$ such that $\alpha \in \mathcal{N}_1 \subset \mathcal{N}_2 \subset \cdots \subset \mathcal{N}_R = \mathcal{N} \subset I$, $|\mathcal{N}| \le N^{1/2-\mu}$ and*

$$\kappa\Big(f(W_{I\setminus\bigcup_j \mathcal{N}_{n_j+1}(\alpha_j)}), g_1(W_{\mathcal{N}_{n_1}(\alpha_1)\setminus\bigcup_{j\neq 1}\mathcal{N}(\alpha_j)}), \dots, g_q(W_{\mathcal{N}_{n_q}(\alpha_q)\setminus\bigcup_{j\neq q}\mathcal{N}(\alpha_j)})\Big) \le_{R,q,\mu} N^{-3q} \|f\|_{q+1} \prod_{j=1}^{q} \|g_j\|_{q+1}$$

for any $n_1, \dots, n_q < R$, $\alpha_1, \dots, \alpha_q \in I$ and functions $f, g_1, \dots, g_q$. We will refer to these sets as "neighbourhoods" of α, although we do not assume any topological structure on I. For any $\mathcal{N} \subset I$, here $W_{\mathcal{N}}$ denotes the set of w_α indexed by $\alpha \in \mathcal{N}$.

Remark 2.9. *For the proof of Theorem 2.2 we need Assumptions (B), (C) and (D) only for finitely many values of q, R up to some threshold, depending only on the parameters D, γ, ϵ in the statement and μ from Assumption (D). This follows from the fact that the high moment bound from Theorem 4.1 is only needed for a finite value of p which relates to certain threshold on q, R.*

2.6. **Some examples**

We end this section by providing examples of correlated matrix models satisfying Assumptions (C)–(D). Our main example is the one already advertised in Assumption (CD). In Example 2.10 we check that Assumption (CD) indeed implies (C)–(D).

Example 2.10 (Polynomially decaying model)**.** *Recall the metric setting of Assumption (CD). Simple calculations show that Assumption (C) is satisfied even if we only request $s \ge 2$ in (2.5), independent of the chosen neighbourhood systems. As for Assumption (D), we define the neighbourhoods $\mathcal{N}_k(\alpha) := \{\beta \in I \mid d(\alpha,\beta) \le k N^{1/4-\mu}\}$ so that $d(\mathcal{N}_k(\alpha), \mathcal{N}_{k+1}(\alpha)^c) = N^{1/4-\mu}$. To ensure that*

$$\big|\kappa\big(f_1(W_{\mathcal{N}_n(\alpha)}), f_2(W_{\mathcal{N}_{n+1}(\alpha)^c})\big)\big| \le \frac{\|f_1\|_2 \|f_2\|_2}{N^3},$$

we thus have to choose $s \ge 12/(1-4\mu)$. The tree decay structure (2.5b) then ensures that Assumption (D) is satisfied for all q.

Example 2.11 (Block matrix)**.** *For $n, M, N \in \mathbb{N}$ with $nM = N$ we set $J = [N]$ and consider an $n \times n$-block matrix with identical copies of an $M \times M$ Wigner matrix in each block. We introduce an equivalence relation on $I = J \times J$ in such a way that we first identify $a \sim b \in J$ if $a = b \,(\mathrm{mod}\, M)$, and then $(a,b) \sim (c,d) \in I$ if $(a,b) = (c,d)$ or $(a,b) = (d,c)$ according to the Hermitian symmetry. Then the correlation structure is such that $\kappa(\alpha_1, \dots, \alpha_k) = \mathcal{O}(1)$ if $\alpha_1, \dots, \alpha_k$ all belong to the same equivalence class and $\kappa(\alpha_1, \dots, \alpha_k) = 0$ otherwise. Since every entry is correlated with at most $\mathcal{O}(n^2)$ other entries, Assumptions (C), (D) are clearly satisfied as long as n is bounded.*

The same correlation structure is obtained if the blocks contain possibly different random matrices with independent entries (respecting only the overall Hermitian symmetry, but possibly without symmetry within each block), see e.g. the ensemble discussed in [5]. Furthermore, one may combine the block matrix model with a polynomially decaying model from Example 2.10 to construct yet another example for which Theorem 2.2 is applicable. In this general model the matrices in each block should merely exhibit a polynomially decaying correlation instead of strictly independent elements.

Example 2.12 (Correlated Gaussian matrix models)**.** *Since all higher order cumulants for Gaussian random variables vanish, our method allows to prove the local law (and its corollaries) for correlated Gaussian random matrix models under even weaker conditions. In fact, besides Assumptions (A) and (E) (or (F) for universality) we only have to assume that*

$$|||\kappa|||_2^{av} + |||\kappa|||_2^{iso} \le_{\epsilon} N^{\epsilon}$$

for all $\epsilon > 0$. In particular, this includes the polynomially decaying model from Example 2.10 for $s \ge 2$. These statements can be directly proved by following our general proof, setting all higher order cumulants to zero and using neighbourhoods $\mathcal{N}(\alpha) = I$ for all α. The details are left to the reader.

Example 2.13 (Fourfold symmetry)**.** *A Wigner matrix W with fourfold symmetry is a matrix of independent entries w_α of unit variance up to the symmetries $w_{a,b} = w_{b,a} = w_{-a,-b} = w_{-b,-a}$ for all $a, b \in \mathbb{Z}/N\mathbb{Z}$. From the explicit formula*

$$\kappa(ab, cd) = \kappa_d(ab, cd) + \kappa_c(ab, cd) := (\delta_{a,c}\delta_{b,d} + \delta_{a,-c}\delta_{b,-d}) + (\delta_{a,d}\delta_{b,c} + \delta_{a,-d}\delta_{b,-c}),$$

and a similar one for the third order cumulants, Assumption (C) is straightforward to verify. By choosing the neighbourhoods $\mathcal{N}(\alpha)$ to contain the three other companions of α from the symmetry, it is obvious that also Assumption (D) is fulfilled. Strictly speaking, the flatness

condition (E) is violated by the fourfold symmetry, but as the resulting M is diagonal, there is an easy replacement for the flatness. For more details on the random matrix model with a fourfold symmetry we refer the reader to [4].

A similar argument shows that Assumptions (C)–(D) are also satisfied for other symmetries which naturally split in such a way that $w_{a,b}$ is identified with $w_{f_1(a),f_2(b)}$ and $w_{g_1(b),g_2(a)}$ for a finite collection of functions f_i, g_i. The appropriate replacement for the flatness condition (E), however, has to be checked on a case-by-case basis.

3. General multivariate cumulant expansion

The goal of this section is the derivation of a finite-order multivariate cumulant expansion with a precise control on the approximation error.

3.1. **Precumulants: Definition and relation to cumulants**

We begin by introducing the concept of *pre-cumulants* and establishing some of their important properties. For any collection of random variables $X, Y_1, \dots, Y_m$ we define the quantities

$$\begin{aligned} K(X) &:= X \\ K_{t_1,\dots,t_m}(X;\boldsymbol{Y}) &= K_{t_1,\dots,t_m}(X;Y_1,\dots,Y_m) := Y_m(\mathbb{1}_{t_m\le t_{m-1}}-\mathbf{E})Y_{m-1}(\mathbb{1}_{t_{m-1}\le t_{m-2}}-\mathbf{E})Y_{m-2}\dots Y_1(\mathbb{1}_{t_1\le 1}-\mathbf{E})X \end{aligned}$$

for $m \ge 1$, that depend on real parameters $t_1, \dots, t_m \in [0,1]$. We will call them *time ordered pre-cumulants*. We moreover introduce the integrated *symmetrized pre-cumulants*

$$K(X;\boldsymbol{Y}) := \sum_{\sigma\in S_{|\boldsymbol{Y}|}} \iiint_0^1 K_{t_1,\dots,t_{|\boldsymbol{Y}|}}(X;\sigma(\boldsymbol{Y}))\,\mathrm{d}\boldsymbol{t},$$

where $S_{|\boldsymbol{Y}|}$ is the group of permutations on a $|\boldsymbol{Y}|$-element set and $\mathrm{d}\boldsymbol{t} = \mathrm{d}t_1\dots\mathrm{d}t_m$ indicates integration over $[0,1]^{|\boldsymbol{Y}|}$. Note that the first variable X of $K(X;\boldsymbol{Y})$ plays a special role. Moreover, $K(X;\boldsymbol{Y})$ is invariant under permutations of the components of the vector $\boldsymbol{Y}$. These pre-cumulants are – other than the actual cumulants – random variables, but their expectations turn out to produce the traditional cumulants, justifying their name. While they appear to be very natural objects in the study of cumulants, we are not aware whether the pre-cumulants K have been previously studied, and whether the result of the following lemma is already known.

Lemma 3.1 (Pre-cumulant Lemma)**.** *Let X be a random variable and let $\boldsymbol{Y}$, $\boldsymbol{Z}$ be random vectors. Then we have*

$$\mathbf{E}\,K(X;\boldsymbol{Y}) = \kappa(X,\boldsymbol{Y}), \tag{3.1a}$$

$$K(X;\boldsymbol{Y}) = \kappa(X,\boldsymbol{Y}) + X(\Pi\boldsymbol{Y}) - \sum_{\boldsymbol{Y}'\subset\boldsymbol{Y}} (\Pi\boldsymbol{Y}')\kappa(X,\boldsymbol{Y}\setminus\boldsymbol{Y}'), \tag{3.1b}$$

and the pre-cumulant decoupling identity

$$K(X;\boldsymbol{Y}\sqcup\boldsymbol{Z}) - \kappa(X,\boldsymbol{Y}\sqcup\boldsymbol{Z}) = (\Pi\boldsymbol{Z})\big[K(X;\boldsymbol{Y}) - \kappa(X,\boldsymbol{Y})\big] - \sum_{\substack{\boldsymbol{Y}'\subset\boldsymbol{Y}\\ \boldsymbol{Z}'\subsetneq\boldsymbol{Z}}} (\Pi\boldsymbol{Y}')(\Pi\boldsymbol{Z}')\kappa(X,(\boldsymbol{Y}\setminus\boldsymbol{Y}')\sqcup(\boldsymbol{Z}\setminus\boldsymbol{Z}')), \tag{3.1c}$$

where $\boldsymbol{Y}'\subset\boldsymbol{Y}$ indicates that $\boldsymbol{Y}'$ is a sub-vector of $\boldsymbol{Y}$ (with $\boldsymbol{Y}'=\emptyset$ and $\boldsymbol{Y}'=\boldsymbol{Y}$ allowed) and $\boldsymbol{Y}\setminus\boldsymbol{Y}'$ is the vector of the remaining entries. By $\boldsymbol{Z}'\subsetneq\boldsymbol{Z}$ we denote all proper sub-vectors of $\boldsymbol{Z}$, i.e., not including $\boldsymbol{Z}$. By $\Pi\boldsymbol{Z}$ we mean the product of all entries of the vector $\boldsymbol{Z}$, while by $\boldsymbol{Z}\cup\boldsymbol{Y}$ we mean the concatenation of the two vectors $\boldsymbol{Z},\boldsymbol{Y}$. The order of the vector is of no importance as $K(X;\boldsymbol{Y})$ is symmetric with respect to the vector $\boldsymbol{Y}$ and κ is overall symmetric.

We note that (3.1c) is intentionally not symmetric in $\boldsymbol{Y}$, $\boldsymbol{Z}$, although an analogous formula holds with $\boldsymbol{Y}$ and $\boldsymbol{Z}$ interchanged. The relation (3.1c) should be interpreted as a refined version of the fact that centred precumulants factor independent random variables. Indeed, if $\boldsymbol{Z}$ was independent of $X,\boldsymbol{Y}$, then the sum in (3.1c) would vanish by independence properties of the cumulant and (3.1c) would simplify to

$$K(X;\boldsymbol{Y}\sqcup\boldsymbol{Z}) - \kappa(X,\boldsymbol{Y}\sqcup\boldsymbol{Z}) = (\Pi\boldsymbol{Z})\big[K(X;\boldsymbol{Y}) - \kappa(X,\boldsymbol{Y})\big].$$

In our applications $\boldsymbol{Z}$ will depend only very weakly on X and $\boldsymbol{Y}$, hence the sum in (3.1c) will be a small error term.

Proof. By the definition of the pre-cumulants, we have for $\boldsymbol{Y} = (Y_1,\dots,Y_m)$

$$K(X;\boldsymbol{Y}) = \sum_{\sigma\in S_m} \iiint_0^1 Y_{\sigma(m)}(\mathbb{1}_{t_m\le t_{m-1}}-\mathbf{E})Y_{\sigma(m-1)}(\mathbb{1}_{t_{m-1}\le t_{m-2}}-\mathbf{E})\dots(\mathbb{1}_{t_2\le t_1}-\mathbf{E})Y_{\sigma(1)}(\mathbb{1}_{t_1\le 1}-\mathbf{E})X\,\mathrm{d}\boldsymbol{t}. \tag{3.2}$$

Multiplying out the brackets and pulling the characteristic functions involving the t-variables out of the expectations, each term is a product of moments of $(X,\boldsymbol{Y})$-monomials. We rearrange the sum according to the number of moments in the form that

$K(X;\boldsymbol{Y}) = \sum_{b=0}^{m} \phi_b$, where ϕ_b contains exactly b moments. These terms are given by

$$\phi_b = (-1)^b \sum_{1\le j_1<\dots<j_b\le m} \sum_{\sigma\in S_m} \iiint_0^1 \mathbb{1}_{t_m\le\dots\le t_{j_b}} \mathbb{1}_{t_{j_b-1}\le\dots\le t_{j_{b-1}}} \dots \mathbb{1}_{t_{j_2-1}\le\dots\le t_{j_1}} \mathbb{1}_{t_{j_1-1}\le\dots\le t_1} \,\mathrm{d}\boldsymbol{t}$$
$$\times Y_{\sigma(m)}\dots Y_{\sigma(j_b)}(\mathbf{E}\, Y_{\sigma(j_b-1)}\dots Y_{\sigma(j_{b-1})})\dots(\mathbf{E}\, Y_{\sigma(j_2-1)}\dots Y_{\sigma(j_1)})(\mathbf{E}\, Y_{\sigma(j_1-1)}\dots Y_{\sigma(1)}X), \qquad b\ge 1 \tag{3.3}$$

and the integral in (3.3) can be computed to give

$$\iiint_0^1 [\cdots]\,\mathrm{d}\boldsymbol{t} = \frac{1}{(m-j_b+1)!}\frac{1}{(j_b-j_{b-1})!}\cdots\frac{1}{(j_2-j_1)!}\frac{1}{(j_1-1)!} =: V.$$

Here we introduced an additional variable $t_0 = 1$ for notational convenience and follow the convention that the last factor in (3.3) for $j_1 = 1$ reads $\mathbf{E}\, X$. For $b = 0$ the analogue of (3.3) is given by

$$\phi_0 = \Big(\sum_{\sigma\in S_m} \iiint_0^1 \mathbb{1}_{t_m\le\dots\le t_1}\,\mathrm{d}\boldsymbol{t}\Big) Y_1\dots Y_m X = Y_1\dots Y_m X.$$

Let the summation indices $1 \le j_1 < \cdots < j_b \le m$ be fixed and fix a labelled partition of $[m] = \pi_1 \sqcup \cdots \sqcup \pi_{b+1}$ into subsets of sizes $|\pi_1| = j_1 - 1$, $|\pi_2| = j_2 - j_1$, …, $|\pi_b| = j_b - j_{b-1}$ and $|\pi_{b+1}| = m - j_b + 1$. Those permutations σ in (3.3) for which $\sigma([1, j_1 - 1]) = \pi_1, \sigma([j_1, j_2 - 1]) = \pi_2, \dots, \sigma([j_{b-1}, j_b - 1]) = \pi_b$ and $\sigma([j_b, m]) = \pi_{m+1}$ all produce the same term $(-1)^b V \Pi\boldsymbol{Y}_{\pi_{b+1}} \dots (\mathbf{E}\,\Pi\boldsymbol{Y}_{\pi_2})(\mathbf{E}\, X\Pi\boldsymbol{Y}_{\pi_1})$, where $\boldsymbol{Y}_\pi = (\, Y_k \mid k \in \pi\,)$. We note that π_1 plays a special role since it is explicitly allowed to be the empty set, in which the last factor is just X. The combinatorial factor V is precisely cancelled by the number of such permutations, i.e., $1/V$. Thus (3.3) can be rewritten as

$$\phi_b = (-1)^b \sum_{\substack{\pi_1\sqcup\dots\sqcup\pi_{b+1}=[m]\\ |\pi_j|\ge 1 \text{ for } j\ge 2}} \Pi\boldsymbol{Y}_{\pi_{b+1}}(\mathbf{E}\,\Pi\boldsymbol{Y}_{\pi_b})\dots(\mathbf{E}\,\Pi\boldsymbol{Y}_{\pi_2})(\mathbf{E}\, X\Pi\boldsymbol{Y}_{\pi_1}), \tag{3.4a}$$

and therefore

$$K(X;\boldsymbol{Y}) = \sum_{b=0}^{m}\phi_b = \sum_{b=0}^{m}(-1)^b \sum_{\substack{\pi_1\sqcup\dots\sqcup\pi_{b+1}=[m]\\ |\pi_j|\ge 1 \text{ for } j\ge 2}} \Pi\boldsymbol{Y}_{\pi_{b+1}}(\mathbf{E}\,\Pi\boldsymbol{Y}_{\pi_b})\dots(\mathbf{E}\,\Pi\boldsymbol{Y}_{\pi_2})(\mathbf{E}\, X\Pi\boldsymbol{Y}_{\pi_1}), \qquad b\ge 1. \tag{3.4b}$$

We recognize the expectation of (3.4a) as the sum over all unlabelled partitions $\mathcal{P} \vdash (X, \boldsymbol{Y})$ with $|\mathcal{P}| = b + 1$ blocks, undercounting by a factor of $b!$ as the first b factors on the rhs. of(3.4a) after taking the expectation are interchangeable (the last factor is special due to X). We can thus conclude that $\mathbf{E}\, K(X;\boldsymbol{Y})$ reads

$$\mathbf{E}\, K(X;\boldsymbol{Y}) = \sum_{b=0}^{m}(-1)^b b! \sum_{\substack{\mathcal{P}\vdash(X,\boldsymbol{Y})\\ |\mathcal{P}|=b+1}} \prod_{A\in\mathcal{P}} \mathbf{E}\,\Pi(X,\boldsymbol{Y})_A = \kappa(X,\boldsymbol{Y}), \tag{3.5}$$

where we used (A.4) in the ultimate step, an identity that is equivalent to the analytical definition of the cumulant, see Appendix A for more details. This completes the proof of (3.1a). Now (3.1b) follows from first separating $b = 0$ to produce the $X(\Pi\boldsymbol{Y})$ term and then separating the π_{b+1} summation in (3.4b) so that $\boldsymbol{Y}_{\pi_{b+1}}$ plays the role of $\boldsymbol{Y}'$ for $\boldsymbol{Y}' \neq \emptyset$. The sum over the remaining moments is exactly the cumulant $\kappa(X, \boldsymbol{Y} \setminus \boldsymbol{Y}')$, see (3.5). Finally, the term $\boldsymbol{Y}' = \emptyset$ in (3.1b) cancels the first $\kappa(X, \boldsymbol{Y})$ term, completing the proof of (3.1b). The identity (3.1c) follows from (3.1b) where $\boldsymbol{Y}$ plays the role of $\boldsymbol{Y} \sqcup \boldsymbol{Z}$. The $\boldsymbol{Z}' = \boldsymbol{Z}$ term is considered separately, and then the identity (3.1b) is used again, this time for X and $\boldsymbol{Y}$. □

3.2. **Precumulant expansion formula**

We consider a random vector $\boldsymbol{w} \in \mathbb{R}^{\mathcal{I}}$, indexed by an abstract set $\mathcal{I}$, and a sufficiently often differentiable function $f\colon \mathbb{R}^{\mathcal{I}} \to \mathbb{C}$. The goal is to derive an expansion for $\mathbf{E}\, w_{i_0} f(w)$ in the variables indexed by a fixed subset $\mathcal{N} \subset \mathcal{I}$ that contains a distinguished element $i_0 \in \mathcal{N}$. The expansion will be in terms of cumulants $\kappa(w_{i_1}, \dots, w_{i_m})$ and expectations $\mathbf{E}\,\partial_{\boldsymbol{i}} f$ of derivatives $\partial_{\boldsymbol{i}} f := \partial_{i_1}\dots\partial_{i_m} f$, where we identify $\partial_i = \partial_{w_i}$ and $\boldsymbol{i} = \{i_1, \dots, i_m\}$. To state the expansion formula compactly we first introduce some notations and definitions. We recall that a multiset is an unordered set with possible multiple appearances of the same element. For a given tuple $\boldsymbol{i} = (i_1, \dots, i_m) \in \mathcal{N}^m$ we define the multisets

$$\underline{w}_{\boldsymbol{i}} := \{w_{i_1}, \dots, w_{i_m}\} \quad \text{and the augmented multiset} \quad \underline{w}_{i_0\boldsymbol{i}} := \{w_{i_0}\} \sqcup \underline{w}_{\boldsymbol{i}}, \tag{3.6}$$

where we consider $\sqcup$ as a disjoint union in the sense that $\underline{w}_{i_0\boldsymbol{i}}$ has $m + 1$ elements (counting repetitions), regardless of whether $i_0 = i_k$ for some $k \in [m]$. Similarly, we write $\underline{w}_* \subset \underline{w}$ to indicate that $\underline{w}_*$ is a sub-multiset of a multiset $\underline{w}$. As cumulants are invariant under permutations of their entries we will write $\kappa(\underline{w})$ for multisets $\underline{w}$ of random variables. We will also write $\Pi\underline{w} := \prod_{j=1}^{m} w_{i_j}$ for the product of elements of a multiset $\underline{w} = \{\, w_{i_j} \mid j \in [m]\,\}$.

Equipped with Lemma 3.1 we can now state and prove the version of the multivariate cumulant expansion with a remainder that is best suitable for our application. Recall from (3.1a) that $\mathbf{E}\, K(w_{i_0}; \underline{w}_{\boldsymbol{i}}) = \kappa(\underline{w}_{i_0\boldsymbol{i}})$.

Proposition 3.2 (Multivariate cumulant expansion). *Let $f\colon \mathbb{R}^{\mathcal{I}} \to \mathbb{C}$ be R times differentiable with bounded derivatives and let $\boldsymbol{w} \in \mathbb{R}^{\mathcal{I}}$ be a random vector with finite moments up to order R. Fix a subset $\mathcal{N} \subset \mathcal{I}$ and an element $i_0 \in \mathcal{N}$, then it holds that*

$$\mathbf{E}\, w_{i_0} f(w) = \sum_{m=0}^{R-1} \sum_{\boldsymbol{i} \in \mathcal{N}^m} \left[\mathbf{E}\, \frac{\kappa(\underline{w}_{i_0 \boldsymbol{i}})}{m!} \partial_{\boldsymbol{i}} f + \mathbf{E}\, \frac{K(w_{i_0}; \underline{w}_{\boldsymbol{i}}) - \kappa(\underline{w}_{i_0 \boldsymbol{i}})}{m!} \partial_{\boldsymbol{i}} f \big|_{\boldsymbol{w}_{\mathcal{N}}=0} \right] + \Omega(f, i_0, \mathcal{N}), \quad \textit{where} \tag{3.7a}$$

$$\Omega(f, i_0, \mathcal{N}) := \sum_{\boldsymbol{i} \in \mathcal{N}^R} \mathbf{E} \iiint_0^1 K_{t_1,\dots,t_R}(w_{i_0}, \dots, w_{i_R})\, \mathrm{d}t_1 \dots \mathrm{d}t_{R-1} \int_0^1 (\partial_{\boldsymbol{i}} f)(t_R \boldsymbol{w}', \boldsymbol{w}'')\, \mathrm{d}t_R, \tag{3.7b}$$

where for $m = 0$ the derivative should be considered as the 0-th derivative, i.e. as the function itself. Here we introduced a decomposition $\boldsymbol{w} = (\boldsymbol{w}', \boldsymbol{w}'')$ of all random variables $\boldsymbol{w} = \boldsymbol{w}_{\mathcal{I}}$ such that $\boldsymbol{w}' = \boldsymbol{w}_{\mathcal{N}} = (w_i \,|\, i \in \mathcal{N})$ and $\boldsymbol{w}'' = \boldsymbol{w}_{\mathcal{N}^c} = (w_i \,|\, i \in \mathcal{I} \setminus \mathcal{N})$ and we write $f(\boldsymbol{w}) = f(\boldsymbol{w}', \boldsymbol{w}'')$. Moreover, if $\mathbf{E}\,|w_i|^{2R} \le \mu_{2R}$ for all $i \in \mathcal{I}$, then

$$|\Omega(f, i_0, \mathcal{N})| \lesssim_R \mu_{2R}^{1/2} \sum_{\boldsymbol{i} \in \mathcal{N}^R} \int_0^1 \left(\mathbf{E}\, \big|(\partial_{\boldsymbol{i}} f)(t_R \boldsymbol{w}', \boldsymbol{w}'')\big|^2 \right)^{1/2} \mathrm{d}t_R. \tag{3.8}$$

Proof. For functions $f = f(\boldsymbol{w})$, $g = g(\boldsymbol{w})$ a Taylor expansion yields, for any $s \ge 0$,

$$\mathbf{E}\, g(\boldsymbol{w}) f(s\boldsymbol{w}', \boldsymbol{w}'') = (\mathbf{E}\, g)(\mathbf{E}\, f(0, \boldsymbol{w}'')) + \mathbf{Cov}\left(g, f(0, \boldsymbol{w}'')\right) + \sum_{i \in \mathcal{N}} \int_0^s \mathbf{E}\, g(\boldsymbol{w}) w_i (\partial_i f)(t\boldsymbol{w}', \boldsymbol{w}'')\, \mathrm{d}t$$

and after another Taylor expansion to restore $f(\boldsymbol{w}', \boldsymbol{w}'')$ in the first term we find

$$\mathbf{E}\, g(\boldsymbol{w}) f(s\boldsymbol{w}', \boldsymbol{w}'') = (\mathbf{E}\, g)(\mathbf{E}\, f) + \mathbf{Cov}\left(g, f(0, \boldsymbol{w}'')\right) + \sum_{i \in \mathcal{N}} \int_0^1 \mathbf{E}\, w_i [\mathbb{1}_{t \le s} g - (\mathbf{E}\, g)](\partial_i f)(t\boldsymbol{w}', \boldsymbol{w}'')\, \mathrm{d}t. \tag{3.9}$$

Here we follow the convention that if no argument is written, then $\mathbf{E}\, g = \mathbf{E}\, g(\boldsymbol{w})$. Starting with $g(\boldsymbol{w}) = w_{i_0}$, the last term in (3.9) requires to compute $\mathbf{E}\, K_t(w_{i_0}; w_i)(\partial_i f)(t\boldsymbol{w}', \boldsymbol{w}'')$ with $t = t_1, i = i_1$. So this has the structure $\mathbf{E}\, \widetilde{g} \widetilde{f}(t\boldsymbol{w}', \boldsymbol{w}'')$ with $\widetilde{g} = K_{t_1}$ and $\widetilde{f} = \partial_{i_1} f$ and we can use (3.9) again. Iterating this procedure with

$$(g(\boldsymbol{w}), s, i, t) = (K_{t_1,\dots,t_{m-1}}(w_{i_0}; w_{i_1} \dots, w_{i_{m-1}}), t_{m-1}, i_m, t_m)$$

for $m = 1, \dots, R$, we arrive at

$$\begin{aligned} \mathbf{E}\, w_{i_0} f = &\sum_{m=0}^{R-1} \sum_{i_1,\dots,i_m \in \mathcal{N}} \left(\mathbf{E} \iiint_0^1 K_{t_1,\dots,t_m}\, \mathrm{d}\boldsymbol{t} \right) (\mathbf{E}\, \partial_{\boldsymbol{i}} f) + \sum_{m=0}^{R-1} \sum_{i_1,\dots,i_l \in \mathcal{N}} \mathbf{E} \left(\iiint_0^1 K_{t_1,\dots,t_m}\, \mathrm{d}\boldsymbol{t} - \mathbf{E} \iiint_0^1 K_{t_1,\dots,t_m}\, \mathrm{d}\boldsymbol{t} \right) (\partial_{\boldsymbol{i}} f)(0, \boldsymbol{w}'') \\ &+ \sum_{i_1,\dots,i_R \in \mathcal{N}} \mathbf{E} \iiint_0^1 K_{t_1,\dots,t_R}\, \mathrm{d}t_1 \dots \mathrm{d}t_{R-1} \int_0^1 (\partial_{\boldsymbol{i}} f)(t_R \boldsymbol{w}', \boldsymbol{w}'')\, \mathrm{d}t_R, \end{aligned} \tag{3.10}$$

where $K_{t_1,\dots,t_m} = K_{t_1,\dots,t_m}(w_{i_0}, \dots, w_{i_m})$ and $\mathrm{d}\boldsymbol{t} = \mathrm{d}t_1 \dots \mathrm{d}t_m$. We note that (3.10) does not include the sum over permutations, but since the summation over all $i_1, \dots, i_m$ is taken we can artificially insert the permutation as in

$$\sum_{i_1,\dots,i_m} \phi(i_1, \dots, i_m) = \frac{1}{m!} \sum_{i_1,\dots,i_m} \sum_{\sigma \in S_m} \phi(i_{\sigma(1)}, \dots, i_{\sigma(m)}).$$

Now (3.7a) follows from combining (3.10) with (3.1a). Finally, (3.8) follows directly from a simple application of the Hölder inequality. □

3.3. Toy model

Proposition 3.2 will be the main ingredient for the probabilistic part of the proofs of Theorems 2.1 and 2.2. For pedagogical reasons we first demonstrate the multiplicative cancellation effect of *self-energy renormalization* through iterated cumulant expansion in a toy model.

Let f and $\boldsymbol{w}$ be as in Proposition 3.2 and let us suppose that $\mathcal{I}$ is equipped with a metric d. We furthermore assume that $\mathbf{E}\, \boldsymbol{w} = 0$ and that the multivariate cumulants of $\boldsymbol{w}$ follow a tree-like mixing decay structure as in Example 2.10, i.e.,

$$\kappa(f_1(\boldsymbol{w}), \dots, f_k(\boldsymbol{w})) \lesssim \prod_{\{i,j\} \in E(T_{\min})} \frac{1}{1 + d(\operatorname{supp} f_i, \operatorname{supp} f_j)^s} \tag{3.11}$$

for some $s > 0$, where $T_{\min}$ is the tree such that the sum of $d(\operatorname{supp} f_i, \operatorname{supp} f_j)$ along its edges $\{i, j\} \in E(T_{\min})$ is minimal. Fix now a finite positive integer parameter R and a large length scale $l > 0$. Around every $i \in \mathcal{I}$ we use the metric d to define neighbourhoods $\mathcal{N}(i) := \{\, j \in \mathcal{I} \mid d(i, j) \le lR \,\}$ and $\mathcal{N}_k(i) := \{\, j \in I \mid d(i, j) \le lk \,\}$, as in Assumption (D). For definiteness we furthermore assume that $\mathcal{I}$ has dimension two in the sense that $|\mathcal{N}| \sim l^2 R^2$ as for the standard labelling of a matrix where $\mathcal{I} = [N]^2$. We now assume that f does not depend strongly on any single w_i, more specifically, for an multi-index $\boldsymbol{i}$ we assume

$$|\partial_{\boldsymbol{i}} f| \lesssim |\mathcal{N}|^{-(1+\epsilon)|\boldsymbol{i}|}, \qquad \boldsymbol{i} = (i_1, \dots, i_p), \qquad |\boldsymbol{i}| = p. \tag{3.12}$$

This bound ensures that the size of the derivative in the Taylor expansion in the neighbourhood $\mathcal{N}$ compensates for the combinatorics.

3.3.1. **Expansion of a weakly dependent function.** For this setup we want to study the size of the expression

$$\mathbf{E}\, w_{i_1} \dots w_{i_p} f(\boldsymbol{w})$$

where $i_1, \dots, i_p$ are in general position in the sense that their $\mathcal{N}(i_k)$ neighbourhoods do not intersect. If f were constant we could use the following lemma:

Lemma 3.3. *Assume that $\boldsymbol{w}$ has a tree-like correlation decay as in (3.11) and assume that the random variables $g_0(\boldsymbol{w}), \dots, g_p(\boldsymbol{w})$ have mutually l-separated supports, i.e., that $d(\operatorname{supp} g_i, \operatorname{supp} g_j) \gtrsim l$ for all $i \neq j$. If furthermore $\mathbf{E}\, g_k = 0$ for $k = 1, \dots, p$, then it holds that*

$$|\mathbf{E}\, g_0 \dots g_p| \lesssim l^{-s\lceil p/2 \rceil}.$$

Proof. Due to a basic identity on cumulants, see (A.2), we have that

$$\mathbf{E}\, g_0 \dots g_p = \sum_{A_1 \sqcup \dots \sqcup A_k = [0,p]} \kappa(g_{A_1}) \dots \kappa(g_{A_k}),$$

where the sum goes over all partitions $[0, p]$ and $g_A = \{\, g_k \mid k \in A \,\}$. From (3.11) it follows that

$$|\kappa(g_{A_k})| \lesssim l^{-s(|A_k|-1)}$$

and due to the assumption of zero mean $\mathbf{E}\, g_k = 0$ for $k \in [p]$ we have that $\kappa(g_A) = 0$ whenever $A = \{k\}$ for some $k \in [p]$. It follows that the worst case is given by pair partitions with $|A_k| = 2$ for all A_k not containing 0 which completes the proof. □

From this lemma with $g_0 = 1$ and $g_k = w_{i_k}$ for $k = 1, \dots, p$ we conclude that for constant f we have the asymptotic bound $\left| f\, \mathbf{E}\, w_{i_1} \dots w_{i_p} \right| \lesssim l^{-s\lceil p/2 \rceil}$ by the zero mean assumption $\kappa(w_i) = \mathbf{E}\, w_i = 0$. We now want to argue that for weakly dependent f as in (3.12) a similar bound still holds true although f depends on all variables. Note that the weak dependence renders the minimal spanning tree distance trivial and a direct application of (3.11) would not give any decay. For brevity, we introduce the notations

$$\kappa(i, \boldsymbol{j}) \coloneqq \kappa(w_i, w_{\boldsymbol{j}}), \qquad K(i; \boldsymbol{j}) \coloneqq K(w_i; w_{\boldsymbol{j}}),$$

i.e. we identify cumulants and precumulants as functions of indices rather than random variables. We begin by expanding the first w_{i_1} to obtain from (3.7a)

$$\mathbf{E}\, w_{i_1} \dots w_{i_p} f = \sum_{\boldsymbol{j}_1}^{\mathcal{N}(i_1)} \mathbf{E} \left[\frac{\kappa(i_1, \boldsymbol{j}_1)}{|\boldsymbol{j}_1|!} + \frac{K(i_1; \boldsymbol{j}_1) - \mathbf{E}\, K(i_1; \boldsymbol{j}_1)}{|\boldsymbol{j}_1|!} \Big|^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}(i_1)} = 0} \right] w_{i_2} \dots w_{i_p} \partial_{\boldsymbol{j}_1} f + \mathcal{O}\left(l^{-2\epsilon R} \right), \tag{3.13}$$

where we set $\sum_{\boldsymbol{j}}^{\mathcal{N}} \coloneqq \sum_{0 \leq m < R} \sum_{\boldsymbol{j} \in \mathcal{N}^m}$ and the parameter R, the maximal order of the expansion, is omitted for brevity. The notation $|^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}} = 0}$ means that in all expressions to the right, the argument $\boldsymbol{w}$ is set to zero in the set $\mathcal{N}$, i.e. $\boldsymbol{w}_{\mathcal{N}} = 0$. This effect includes expectation values and cumulants. Note that $|^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}_1} = 0} |^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}_2} = 0} = |^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}_1 \cup \mathcal{N}_2} = 0}$, i.e. the effects of multiple $|^{\rightarrow}$ operators accumulate. For example,

$$f(w_1, w_2)|^{\rightarrow}_{w_1 = 0}\, g(w_1, w_2)|^{\rightarrow}_{w_2 = 0}\, h(w_1, w_2) = f(w_1, w_2) g(0, w_2) h(0, 0). \tag{3.14}$$

However, the order of $|^{\rightarrow}_{w_1 = 0}$ and $|^{\rightarrow}_{w_2 = 0}$ matters as long as there is a nontrivial function in between, clearly

$$g(w_1, w_2)|^{\rightarrow}_{w_2 = 0}\, f(w_1, w_2)|^{\rightarrow}_{w_1 = 0}\, h(w_1, w_2) = g(w_1, w_2) f(0, w_2) h(0, 0),$$

which is different from (3.14). Finally, the error term in (3.13) was estimated using (3.8), and by comparing the combinatorics $|\mathcal{N}|^R$ of the summation to the size of the R-th derivative, $|\partial_{i_1} \dots \partial_{i_R} f| \leq |\mathcal{N}|^{-(1+\epsilon)R}$. We will choose $R \approx ps/4\epsilon$ large, so that the error term is negligible.

Iterating this procedure, we find

$$\mathbf{E}\, w_{i_1} \dots w_{i_p} f = \left(\prod_{k \in [p]} \sum_{\boldsymbol{j}_k}^{\mathcal{N}(i_k)} \right) \mathbf{E} \prod_{k \in [p]}^{\rightarrow} \left[\frac{\kappa(i_k, \boldsymbol{j}_k)}{|\boldsymbol{j}_k|!} + \frac{K(i_k; \boldsymbol{j}_k) - \mathbf{E}\, K(i_k; \boldsymbol{j}_k)}{|\boldsymbol{j}_k|!} \Big|^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}(i_k)} = 0} \right] \partial_{\boldsymbol{j}_1} \dots \partial_{\boldsymbol{j}_p} f + \mathcal{O}\left(l^{-sp/2} \right) \tag{3.15}$$

where $\prod^{\rightarrow}_{k \in [p]} a_k$ indicates that the order of the factors a_k is taken to be increasing in k, i.e., as $a_1 \dots a_p$. This is important due to the noncommutativity of the effect of the $|^{\rightarrow}$ operation on subsequent factors. We now open the bracket in (3.15) and first consider the extreme case, where we take the product all the first terms from each bracket, i.e., the product of p factors with κ. In this case the summation is of order 1 as the cumulant assumption (3.11) implies that $\sum_{\boldsymbol{j} \in \mathcal{I}^k} |\kappa(i, \boldsymbol{j})| \lesssim 1$ for any fixed i_1 if $s \geq 2$. Therefore the worst case is when the least total number of derivatives is taken, i.e., when $|\boldsymbol{j}_l| = 1$ for all l, in which case $\left| \partial_{\boldsymbol{j}_1} \dots \partial_{\boldsymbol{j}_p} f \right| \lesssim |\mathcal{N}|^{-(1+\epsilon)p} \lesssim l^{-2p}$. Now we consider the other extreme case where all the $(K - \mathbf{E}\, K) = (K - \kappa)$ factors are multiplied. There we a priori do not see the smallness as the summation size $|\mathcal{N}|^{|\boldsymbol{j}_1| + \dots + |\boldsymbol{j}_p|}$ roughly cancels the derivative size $|\mathcal{N}|^{-(1+\epsilon)(|\boldsymbol{j}_1| + \dots + |\boldsymbol{j}_p|)}$. The desired smallness thus has to come from the correlation decay (3.11). We can, however not directly apply the tree-like decay structure since there does not have to be a "security distance" between the supports of $\boldsymbol{w}_{\boldsymbol{j}_k}$ and f. For those k with such a security we can apply the tree-like decay immediately, and for those k where there is no such security distance we instead use (3.1c) to write $K - \kappa$ approximately as the product of two functions whose supports are separated by a security distance of scale l. Indeed, if $\boldsymbol{j}_k$ is not separated from $\operatorname{supp} f$ at least by l, then by the pigeon hole principle of placing less than

R labels into R nested layers, it splits into two groups $\boldsymbol{\jmath}_k^{(i)}$ and $\boldsymbol{\jmath}_k^{(o)}$ of "inside" and "outside" indices such that $\operatorname{dist}(\boldsymbol{\jmath}_k^{(i)}, \boldsymbol{\jmath}_k^{(o)}) \gtrsim l$. Now by (3.1c) we have that

$$K(i_k;\boldsymbol{\jmath}_k) - \kappa(i_k;\boldsymbol{\jmath}_k) = (\Pi\boldsymbol{\jmath}_k^{(o)})\big[K(i_k;\boldsymbol{\jmath}_k^{(i)}) - \kappa(i_k,\boldsymbol{\jmath}_k^{(i)})\big] - \sum_{\boldsymbol{n}_k^{(o)}\subsetneq\boldsymbol{\jmath}_k^{(o)}} \sum_{\boldsymbol{n}_k^{(i)}\subset\boldsymbol{\jmath}_k^{(i)}} (\Pi\boldsymbol{n}_k^{(i)})(\Pi\boldsymbol{n}_k^{(o)})\kappa(i_k,\boldsymbol{\jmath}_k^{(i)}\setminus\boldsymbol{n}_k^{(i)},\boldsymbol{\jmath}_k^{(o)}\setminus\boldsymbol{n}_k^{(o)}), \tag{3.16}$$

where $\Pi\boldsymbol{\jmath} \coloneqq \Pi\boldsymbol{w}_{\boldsymbol{\jmath}}$. When multiplying (3.16) for all k, in the product of the second terms we (multiplicatively) collect p decay factors l^{-s}, resulting in l^{-sp}. For the product of the first terms we have to estimate a term of the type $\mathbf{E}\, g_1 \dots g_p \widetilde{f}$ with g_k being zero mean random variables such that all factors have mutually l-separated support. Here we set $g_k \coloneqq K(i_k;\boldsymbol{\jmath}_k^{(i)}) - \kappa(i_k,\boldsymbol{\jmath}_k^{(i)})$ and absorbed the $\Pi\boldsymbol{\jmath}_k^{(o)}$ factors into $\widetilde{f}$. It follows that

$$|\mathbf{E}\, g_1 \dots g_p \widetilde{f}| \lesssim l^{-s\lceil p/2\rceil}, \tag{3.17}$$

from Lemma 3.3. In this argument we only considered the two extreme cases when we opened the bracket in (3.15) and even in the product $\Pi(K-\kappa)$, after using (3.16) for each factor we only considered the two extreme cases. There are many mixed terms in both steps but they can be estimated similarly and altogether we have

$$\big|\mathbf{E}\, w_{i_1} \dots w_{i_p} f\big| \lesssim l^{-2p} + l^{-sp/2},$$

i.e., a power law decay whose exponent is proportional to the number of factors.

3.3.2. **Expansion of a product of weakly dependent functions and self-energy renormalization.** Now we generalize the expansion from Section 3.3.1 and consider another simple example: the iterated expansion of multipole weakly dependent functions. In particular, we will introduce the concept of *self-energy renormalization*.

Let $f_1,\dots,f_p$ be some functions of $\boldsymbol{w}$ which also depend weakly on each single w_i in such a way that $|\partial_{\boldsymbol{\jmath}} f| \lesssim |\mathcal{N}|^{-(1+\epsilon)|\boldsymbol{\jmath}|}$, and let $i_1,\dots,i_p$ be in general position as in the previous example. We want to study

$$\mathbf{E} \prod_{k\in[p]} w_{i_k} f_k,$$

which, by (3.15) with f replaced by $\prod f_k$, can be expanded to

$$\mathbf{E} \prod_{k\in[p]} w_{i_k} f_k = \prod_{k\in[p]} \Bigg(\sum_{\boldsymbol{\jmath}_k}^{\mathcal{N}(i_k)} \sum_{(\boldsymbol{\jmath}_k^l)_{l\in[p]}=\boldsymbol{\jmath}_k} \Bigg) \mathbf{E} \overrightarrow{\prod_{k\in[p]}} \left[\frac{\kappa(i_k,\boldsymbol{\jmath}_k)}{|\boldsymbol{\jmath}_k|!} + \frac{K(i_k;\boldsymbol{\jmath}_k) - \mathbf{E}\, K(i_k;\boldsymbol{\jmath}_k)}{|\boldsymbol{\jmath}_k|!} \right|_{\boldsymbol{w}_{\mathcal{N}(i_k)}=0}^{\rightarrow} \Bigg] \prod_{n\in[p]} (\partial_{\boldsymbol{\jmath}^n} f_n) + \mathcal{O}\left(l^{-sp/2}\right).$$

Here the second sum is the sum over all partitions $\boldsymbol{\jmath}_k^1 \sqcup \dots \sqcup \boldsymbol{\jmath}_k^p = \boldsymbol{\jmath}_k$ of the multi-index $\boldsymbol{\jmath}_k$, the multi-index $\boldsymbol{\jmath}^n$ is given by the disjoint union $\boldsymbol{\jmath}^n = \boldsymbol{\jmath}_1^n \sqcup \dots \sqcup \boldsymbol{\jmath}_p^n$, and we choose $R \approx ps/4\epsilon$, as in the previous example (recall that R is the maximal order of expansion, i.e. $|\boldsymbol{\jmath}_k| \le R$). Thus $\boldsymbol{\jmath}_k^n$ encodes those derivatives hitting f_n which originate from the expansion according to w_{i_k}. By expanding the product we can rewrite this expression as

$$\begin{aligned}\mathbf{E} \prod_{k\in[p]} w_{i_k} f_k = \sum_{L_1\sqcup L_2=[p]} \mathbf{E} \prod_{k\in L_1} \Bigg[\sum_{\boldsymbol{\jmath}_k}^{\mathcal{N}(i_k)} \frac{\kappa(i_k,\boldsymbol{\jmath}_k)}{|\boldsymbol{\jmath}_k|!} \sum_{(\boldsymbol{\jmath}_k^n)_{n\in[p]}=\boldsymbol{\jmath}_k} \Bigg] \\ \times \overrightarrow{\prod_{k\in L_2}} \left[\sum_{\boldsymbol{\jmath}_k}^{\mathcal{N}(i_k)} \frac{K(i_k;\boldsymbol{\jmath}_k) - \mathbf{E}\, K(i_k;\boldsymbol{\jmath}_k)}{|\boldsymbol{\jmath}_k|!} \right|_{\boldsymbol{w}_{\mathcal{N}(i_k)}=0}^{\rightarrow} \sum_{(\boldsymbol{\jmath}_k^n)_{n\in[p]}=\boldsymbol{\jmath}_k} \Bigg] \prod_{n\in[p]} (\partial_{\boldsymbol{\jmath}^n} f_n) + \mathcal{O}\left(l^{-sp/2}\right).\end{aligned}$$

It turns out that in many relevant cases, in particular after the summation over $i_1,\dots,i_k$, the leading contribution comes from those $k\in L_1$ for which $|\boldsymbol{\jmath}_k| = 1$ and $\big|\boldsymbol{\jmath}_k^k\big| = 1$. To counteract these leading terms we subtract this contribution from each factor $w_{i_k} f_k$ and instead compute

$$\begin{aligned}\mathbf{E} \prod_{k\in[p]} \big[w_{i_k} f_k - \sum_{j\in\mathcal{N}(i_k)} \kappa(i_k,j)\partial_j f_k\big] = \sum_{L_1\sqcup L_2=[p]} \mathbf{E} \prod_{k\in L_1} \Bigg[\sum_{\boldsymbol{\jmath}_k}^{\mathcal{N}(i_k)} \frac{\kappa(i_k,\boldsymbol{\jmath}_k)}{|\boldsymbol{\jmath}_k|!} \sum_{(\boldsymbol{\jmath}_k^n)_{n\in[p]}=\boldsymbol{\jmath}_k} \mathbb{1}\big(\big|\boldsymbol{\jmath}_k^k\big| = 0 \text{ if } |\boldsymbol{\jmath}_k| = 1\big) \Bigg] \\ \times \overrightarrow{\prod_{k\in L_2}} \left[\sum_{\boldsymbol{\jmath}_k}^{\mathcal{N}(i_k)} \frac{K(i_k;\boldsymbol{\jmath}_k) - \mathbf{E}\, K(i_k;\boldsymbol{\jmath}_k)}{|\boldsymbol{\jmath}_k|!} \right|_{\boldsymbol{w}_{\mathcal{N}(i_k)}=0}^{\rightarrow} \sum_{(\boldsymbol{\jmath}_k^n)_{n\in[p]}=\boldsymbol{\jmath}_k} \Bigg] \prod_{n\in[p]} (\partial_{\boldsymbol{\jmath}^n} f_n) + \mathcal{O}\left(l^{-sp/2}\right).\end{aligned} \tag{3.18}$$

We note that this substraction or *self-energy renormalization* does not affect the power counting bound of $l^{-2p} + l^{-sp/2}$ because it does not change the order of the terms but only excludes certain allocations of derivatives. However, beyond power counting, this exclusion can still reduce the effective size of the term considerably, see Section 4 where f is the resolvent of a random matrix.

4. Bound on the error matrix D through a multivariate cumulant expansion

In this section we prove an isotropic and averaged bound on the error matrix D defined in (2.4), in the form of high-moment estimates using the multivariate cumulant expansion. To formalize the bounds, we define the high-moment norms for random variables X and random matrices A by

$$\|X\|_p := (\mathbf{E}\,|X|^p)^{1/p}, \quad \|A\|_p := \sup_{\|\mathbf{x}\|,\|\mathbf{y}\|\le 1} \|\langle \mathbf{x}, A\mathbf{y}\rangle\|_p = \Big[\sup_{\|\mathbf{x}\|,\|\mathbf{y}\|\le 1} \mathbf{E}\,|\langle \mathbf{x}, A\mathbf{y}\rangle|^p\Big]^{1/p},$$

where the supremum is taken over deterministic vectors $\mathbf{x}, \mathbf{y}$.

Theorem 4.1 (Bound on the Error)**.** *Under Assumptions (A), (B) and (D), there exist a constant C_* such that for any $p \ge 1, \epsilon > 0$, z with $\Im z \ge N^{-1}$, $B \in \mathbb{C}^{N\times N}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$ it holds that*

$$\frac{\|\langle \mathbf{x}, D\mathbf{y}\rangle\|_p}{\|\mathbf{x}\|\,\|\mathbf{y}\|} \le_{\epsilon,p} (1 + |||\mathcal{S}||| + |||\kappa|||^{iso}_{\le R}) N^{\epsilon} \sqrt{\frac{\|\Im G\|_q}{N\Im z}} \Big(1 + \|G\|_q\Big)^{\frac{C_*}{\mu}} \Big(1 + \frac{\|G\|_q}{N^{\mu}}\Big)^{\frac{C_* p}{\mu}} \tag{4.1a}$$

$$\frac{\|\langle BD\rangle\|_p}{\|B\|} \le_{\epsilon,p} (1 + |||\mathcal{S}||| + |||\kappa|||^{av}_{\le R}) N^{\epsilon} \langle z\rangle \frac{\|\Im G\|_q}{N\Im z} \Big(1 + \|G\|_q\Big)^{\frac{C_*}{\mu}} \Big(1 + \frac{\|G\|_q}{N^{\mu}}\Big)^{\frac{C_* p}{\mu}}, \tag{4.1b}$$

where $q = C_* p^4/(\mu^2\epsilon)$, $R = 4p/\mu$, and for convenience we separately defined

$$|||\mathcal{S}||| := |||\kappa|||^{iso}_2 . \tag{4.2}$$

Remark 4.2. *We remark that the size of $\mathcal{S}$ can be effectively controlled by $|||\kappa|||^{iso}_2$, justifying the definition of $|||\mathcal{S}|||$. To see this we note that due to*

$$\mathcal{S}[V] = \frac{1}{N} \sum_{\alpha_1,\alpha_2} \kappa(\alpha_1, \alpha_2) \Delta^{\alpha_1} V \Delta^{\alpha_2} \tag{4.3}$$

an arbitrary partition of $\kappa = \kappa_c + \kappa_d$ naturally induces a partition $\mathcal{S} = \mathcal{S}_c + \mathcal{S}_d$. Furthermore, it is easy to see that $\|\mathcal{S}_c[V]T\|_p \le |||\kappa_c|||_c \|V\|_{2p} \|T\|_{2p}$ and $\|\mathcal{S}_d[V]T\|_p \le |||\kappa_d|||_d \|V\|_{2p} \|T\|_{2p}$, c.f. Lemma D.2, thus

$$\|\mathcal{S}[V]T\|_p \le |||\kappa|||^{iso}_2 \|V\|_{2p} \|T\|_{2p} .$$

Here we recall that the double-index α stands for a pair $\alpha = (a, b)$ of single indices, and that the matrix Δ^{α} is a matrix of 0's except for a 1 in the (a, b)-entry.

Remark 4.3. *We point out an additional feature of the estimates (4.1a)–(4.1b): they not only provide the optimal power of $\|\Im G\|_q/(N\Im z)$, but the power of $\|G\|_q$, without an extra smallness factor $N^{-\mu}$, is independent of p. This will be essential in the second part of the proof of the local law, see (5.12) later.*

The main tool for proving Theorem 4.1 is the multivariate cumulant expansion from Proposition 3.2. To connect to the toy model considered in Section 3.3, we note that the *self-energy renormalization* of $N^{-1/2}WG$ is $-\mathcal{S}[G]G$, up to an irrelevant contribution from indices $j \notin \mathcal{N}(i_k)$ in (3.18). In this sense the error term $D = N^{-1/2}WG + \mathcal{S}[G]G$ is the difference of $N^{-1/2}WG$ and its self-energy renormalization. As already noted in the context of the toy model we recall that this substraction does not change the power counting of the resulting terms. It does, however, exclude certain allocations of derivatives which in the case of $N^{-1/2}WG$ means that the main contributions coming from the diagonal elements of the form G_{aa} are absent. Off-diagonal elements G_{ab} are smaller on average, in fact the main gain comes from the key formula about resolvents of Hermitian matrices

$$\sum_b |G_{ab}|^2 = \frac{\Im G_{bb}}{\eta},$$

where $\eta = \Im z$. This identity follows directly from the spectral theorem. In the physics literature it is often called *Ward identity* and we will refer to it with this name. Notice that a sum of order N is reduced to a $1/\eta$ factor, so the Ward identity effectively gains a factor of $1/(N\eta)$ over the naive power counting. However, this effect is available only if off-diagonal elements of the resolvent are summed up, the same reduction would not take place in the sum $\sum_a |G_{aa}|^2$ which remains of order N. So the precise index structure is important. The next calculation shows this effect in the simplest case.

Exemplary gain through self-energy renormalization. We now give a short calculation to demonstrate the role of self-energy renormalization term $\mathcal{S}[G]G$ while computing $\mathbf{E}\,\langle D\rangle^2$. Notice that

$$\langle D\rangle = \frac{1}{N}\sum_a \Big[\sum_b \frac{w_{ab}}{\sqrt{N}} G_{ba} + (\mathcal{S}[G]G)_{aa}\Big] = \frac{1}{N}\sum_{a,b} \Big[\frac{w_{ab}}{\sqrt{N}} G_{ba} - \sum_{c,d} \frac{\kappa(ab, cd)}{N} \partial_{cd} G_{ba}\Big] \tag{4.4}$$

is the sum of terms of the form $w_i f$ plus their self-energy renormalization $-N^{-1}\sum_{c,d} \kappa(ab, cd)\partial_{cd} G_{ba}$ where $i = (a, b)$ and $f = G_{ab}$. We note that (4.4) is the direct analogue of the self-energy renormalization in the toy-model discussed in Section 3.3, see (3.18). In (4.4) we expanded $\mathcal{S}[V] = \sum_{\alpha,\beta} N^{-1}\kappa(\alpha, \beta)\Delta^{\alpha} V \Delta^{\beta}$ and used the fact that the resolvent derivative reads $\Delta_{\alpha} G = -G\Delta^{\alpha} G$. Thus one should think of $\mathcal{S}[G]G$ as being the matrix self-energy renormalization of $N^{-1/2}WG$. To present

this example in the simplest form, we assume that W is a Gaussian random matrix which automatically makes all higher order cumulants vanish. We find

$$\mathbf{E}\,\langle D\rangle^2 = N^{-1}\sum_{\alpha_1,\beta_1}\kappa(\alpha_1,\beta_1)\,\mathbf{E}\,\langle\Delta^{\alpha_1}G\rangle\,\langle\Delta^{\beta_1}G\rangle + N^{-2}\sum_{\alpha_1,\beta_1}\kappa(\alpha_1,\beta_1)\sum_{\alpha_2,\beta_2}\kappa(\alpha_2,\beta_2)\,\mathbf{E}\,\langle\Delta^{\alpha_1}G\Delta^{\beta_2}G\rangle\,\langle\Delta^{\alpha_2}G\Delta^{\beta_1}G\rangle\,,$$

the first term of which can be further bounded by

$$N^{-1}\sum_{\alpha_1,\beta_1}\Big|\kappa(\alpha_1,\beta_1)\,\langle\Delta^{\alpha_1}G\rangle\,\langle\Delta^{\beta_1}G\rangle\Big| \le \frac{|||\kappa|||_2^{\mathrm{av}}}{N}\sum_{\alpha}|\langle\Delta^{\alpha}G\rangle|^2 = \frac{|||\kappa|||_2^{\mathrm{av}}}{N^3}\sum_{a,b}|G_{ba}|^2 = \frac{|||\kappa|||_2^{\mathrm{av}}}{N^2}\frac{\langle\Im G\rangle}{\eta}.$$

For the second term we instead compute

$$\begin{aligned}\sum_{\alpha_1,\beta_1}\sum_{\alpha_2,\beta_2}\Big|\frac{\kappa(\alpha_1,\beta_1)\kappa(\alpha_2,\beta_2)}{N^2}\,\langle\Delta^{\alpha_1}G\Delta^{\beta_2}G\rangle\,\langle\Delta^{\alpha_2}G\Delta^{\beta_1}G\rangle\Big| &\le \frac{(|||\kappa|||_2^{\mathrm{av}})^2}{N^2}\sum_{\alpha_1,\alpha_2}|\langle\Delta^{\alpha_2}G\Delta^{\alpha_1}G\rangle|^2\\ &= \frac{(|||\kappa|||_2^{\mathrm{av}})^2}{N^4}\sum_{a_1,b_1,a_2,b_2}|G_{b_2a_1}|^2\,|G_{b_1a_2}|^2 = (|||\kappa|||_2^{\mathrm{av}})^2\frac{\langle\Im G\rangle^2}{(N\eta)^2}\end{aligned}$$

and we conclude that

$$\mathbf{E}\,|\langle D\rangle|^2 \le \frac{1}{N^2}\,\mathbf{E}\left[\frac{|||\kappa|||_2^{\mathrm{av}}\,\langle\Im G\rangle}{\eta} + \left(\frac{|||\kappa|||_2^{\mathrm{av}}\,\langle\Im G\rangle}{\eta}\right)^2\right],$$

which is small if $\eta \gg 1/N$. Without self-energy renormalization, however, i.e., for $\mathbf{E}\,\langle N^{-1/2}WG\rangle^2$ we, for example, also encounter a term of the type

$$N^{-2}\sum_{\alpha_1,\beta_1}\kappa(\alpha_1,\beta_1)\sum_{\alpha_2,\beta_2}\kappa(\alpha_2,\beta_2)\,\mathbf{E}\,\langle\Delta^{\alpha_1}G\Delta^{\beta_1}G\rangle\,\langle\Delta^{\alpha_2}G\Delta^{\beta_2}G\rangle\,,$$

which is of order 1 because it lacks the gain from the Ward identity.

4.1. **Computation of high moments of D through cancellation identities**

Before going into the proof of Theorem 4.1, we sketch the strategy. For simplicity, we first present the proof for the case of bounded spectral parameter $\langle z\rangle \le C$ and we will comment on the trivial modification for the general case at the end of Section 4. We will also temporarily assume that $\|H\| \le C$ with some large constant C. For arbitrary linear (or conjugate linear in the sense that $\Lambda(\lambda\cdot) = \overline{\lambda}\Lambda(\cdot)$ for $\lambda \in \mathbb{C}$) functionals $\Lambda^{(1)},\dots,\Lambda^{(k)}$ we derive an explicit expansion for

$$\mathbf{E}\,\Lambda^{(1)}(D)\dots\Lambda^{(k)}(D) \tag{4.5}$$

in terms of joint cumulants $\kappa(\alpha_1,\dots,\alpha_k)$ of the entries of W and expectations of products of factors of the form

$$\Lambda(\Delta^{\alpha_1}G\Delta^{\alpha_2}G\dots G\Delta^{\alpha_k}G).$$

In other words, we express (4.5) solely in terms of matrix elements of G, which allows for a very systematic estimate. For the main part of the expansion we will then specialize to $\Lambda^{(k)}(D) = \langle BD\rangle$, $\Lambda^{(k)}(D) = \langle\mathbf{x}, D\mathbf{y}\rangle$ or their complex conjugates, and develop a graphical representation of the expansion. In this framework both the averaged and the isotropic bound on D reduce to a sophisticated power counting argument which – with the help of Ward estimates – directly gives the desired size of the averaged and isotropic error.

Equipped with the cumulant expansion from Proposition 3.2, we now aim at expressing $\mathbf{E}\,\Lambda^{(1)}(D)\dots\Lambda^{(p)}(D)$ for linear and conjugate linear functions $\Lambda^{(j)}$, purely in terms of the expectation of products of G's in the form

$$\Lambda_{\alpha_1,\dots,\alpha_k} := -(-1)^k N^{-k/2}\begin{cases}\Lambda(\Delta^{\alpha_1}G\dots\Delta^{\alpha_k}G) & \text{if } \Lambda \text{ is linear}\\ \Lambda(\Delta^{\alpha_1^t}G\dots\Delta^{\alpha_k^t}G) & \text{if } \Lambda \text{ is conjugate linear}\end{cases} \tag{4.6}$$

for double indices $\alpha_1,\dots,\alpha_k \in I = J\times J$, where we recall that for $\alpha = (a,b)$ the transpose α^t denotes $\alpha^t = (b,a)$. The sign choice will make the subsequent expansion sign-free. The reason for the $N^{-k/2}$ pre-factor is that the $\Lambda_{\alpha_1,\dots,\alpha_k}$ terms appear through k derivatives of G's each of which carries a $N^{-1/2}$ from the scaling $H = A + N^{-1/2}W$. Since the derivatives of G naturally come with many permutations from the Leibniz rule, we will also use the notations

$$\Lambda_{\{\alpha_1,\dots,\alpha_m\}} := \sum_{\sigma\in S_m}\Lambda_{\alpha_{\sigma(1)},\dots,\alpha_{\sigma(m)}}, \quad \Lambda_{\alpha,\{\alpha_1,\dots,\alpha_m\}} := \sum_{\sigma\in S_m}\Lambda_{\alpha,\alpha_{\sigma(1)},\dots,\alpha_{\sigma(m)}}, \quad \Lambda_{\underline{\alpha},\underline{\beta}} := \sum_{\alpha\in\underline{\alpha}}\Lambda_{\alpha,\underline{\alpha}\cup\underline{\beta}\setminus\{\alpha\}} \tag{4.7}$$

for multisets $\{\alpha_1,\dots,\alpha_m\}$, $\underline{\alpha}$, $\underline{\beta}$. We will follow the convention that underlined Greek letters denote multisets of labels from I, while non-underlined Greek letters still denote single labels from I. By convention we set $\Lambda_\emptyset = \Lambda_{\emptyset,\underline{\beta}} = 0$. The last two definitions in (4.7) reflect the fact that the first index of Λ will often play a special role since derivatives of $\Lambda_{\alpha_1,\dots,\alpha_k}$ will all keep α_1 as their first index. With these notations, we note that

$$\Lambda_{\underline{\alpha}} = -\mathbb{1}(|\alpha|>0)\Lambda(G^{-1}\partial_{\underline{\alpha}}G), \qquad \Lambda_{\underline{\alpha},\underline{\beta}} = \partial_{\underline{\beta}}\Lambda_{\underline{\alpha}}$$

hold for arbitrary multisets $\underline{\alpha}$, where $|\underline{\alpha}|$ denotes the number of elements (counting multiplicity) in the multiset.

Expansion of a single factor of D. We now use Proposition 3.2 to compute $\mathbf{E}\,\Lambda(D)f$ for any random variable f (later f will be the product of the other Λ's). In the remainder of Section 4 the neighbourhoods $\mathcal{N} = \mathcal{N}(\alpha)$ are those from Assumption (D). The analogue of the length scale l from Section 3.3 is thus $N^{1/4-\mu/2}$, while the parameter R is still a large integer, depending only on p and μ. We expand

$$\mathbf{E}\,\Lambda(D)f = \mathbf{E}\,\frac{1}{\sqrt{N}}\sum_\alpha w_\alpha \Lambda(\Delta^\alpha G)f + \mathbf{E}\,\Lambda(\mathcal{S}[G]G)f = \mathbf{E}\sum_\alpha w_\alpha \Lambda_\alpha f + \mathbf{E}\,\Lambda(\mathcal{S}[G]G)f$$

and from (3.7a) we obtain

$$\mathbf{E}\,\Lambda(D)f = \sum_\alpha \sum_{0\le m<R}\sum_{\boldsymbol{\beta}\in\mathcal{N}^m} \mathbf{E}\left[\frac{\kappa(\alpha,\underline{\beta})}{m!} + \frac{K(\alpha;\underline{\beta})-\kappa(\alpha,\underline{\beta})}{m!}\Big|^{\to}_{W_{\mathcal{N}}=0}\right]\partial_{\underline{\beta}}\Lambda_\alpha f + \mathbf{E}\,\Lambda(\mathcal{S}[G]G)f + \sum_\alpha \Omega(\Lambda_\alpha f,\alpha,\mathcal{N}). \quad (4.8)$$

Here we follow the convention that $\boldsymbol{\beta}$ is the tuple with elements $(\beta_1,\dots,\beta_m)$ and $\underline{\beta}$ is the multiset obtained from the entries $\underline{\beta} = \{\beta_1,\dots,\beta_m\}$, and we recall that for $\mathcal{I} = I$ we denote $\kappa(w_{\alpha_1},\dots,w_{\alpha_k})$ and $K(w_{\alpha_1};w_{\alpha_2},\dots,w_{\alpha_k})$ by $\kappa(\alpha_1,\dots,\alpha_k)$ and $K(\alpha_1;\alpha_2,\dots,\alpha_k)$ (in contrast to the general setting of Section 3 where κ was viewed as a function of the random variables). For $m=0$ the first term in the first bracket of (4.8) vanishes due to $\kappa(\alpha) = \mathbf{E}\,w_\alpha = 0$; for $m=1$ its contribution is given by

$$\sum_{\alpha\in I,\beta\in\mathcal{N}} \kappa(\alpha,\beta)\partial_\beta(\Lambda_\alpha f) = \sum_{\alpha\in I,\beta\in\mathcal{N}} \kappa(\alpha,\beta)\Lambda_{\alpha,\beta} f + \sum_{\alpha\in I,\beta\in\mathcal{N}} \kappa(\alpha,\beta)\Lambda_\alpha\partial_\beta f,$$

where we observe that the first term almost cancels the $\mathbf{E}\,\Lambda(\mathcal{S}[G]G)f = -\sum_{\alpha,\beta\in I}\kappa(\alpha,\beta)\Lambda_{\alpha,\beta}f$ term except for the small contribution from $\beta\notin\mathcal{N}$. We thus rewrite (4.8) in the form

$$\begin{aligned}\mathbf{E}\,\Lambda(D)f = \mathbf{E}\sum_{\alpha\in I,\beta\in\mathcal{N}}\kappa(\alpha,\beta)\Lambda_\alpha\partial_\beta f + \mathbf{E}\sum_{\alpha\in I}\sum_{m<R}\sum_{\boldsymbol{\beta}\in\mathcal{N}^m}\left[\frac{\kappa(\alpha,\underline{\beta})}{l!}\mathbb{1}_{m\ge 2} + \frac{K(\alpha;\underline{\beta})-\kappa(\alpha,\underline{\beta})}{m!}\Big|^{\to}_{W_{\mathcal{N}}=0}\right]\partial_{\underline{\beta}}\big(\Lambda_\alpha f\big)\\ + \mathbf{E}\left(-\sum_{\alpha,\beta\in I}\kappa(\alpha,\beta) + \sum_{\alpha\in I,\beta\in\mathcal{N}}\kappa(\alpha,\beta)\right)\Lambda_{\alpha,\beta}f + \sum_\alpha\Omega(\Lambda_\alpha f,\alpha,\mathcal{N}). \quad (4.9a)\end{aligned}$$

In the above derivation of (4.9) we used directly that Λ is linear. In the case of conjugate linear we replace $\Lambda(D)$ by $\Lambda(D^*)$ which is linear again. This replacement is remedied by the fact that in the definition of $\Lambda_{\alpha_1,\dots,\alpha_k}$ in (4.6) we consider transposed double indices. More generally, following the same computation, we have

$$\begin{aligned}\mathbf{E}\,\Lambda(\partial_{\underline{\gamma}}D)f = \mathbf{E}\,\Lambda_{\underline{\gamma}}f + \mathbf{E}\sum_{\alpha\in I,\beta\in\mathcal{N}}\kappa(\alpha,\beta)\Lambda_{\alpha,\underline{\gamma}}\partial_\beta f + \mathbf{E}\sum_{\alpha\in I}\sum_{m<R}\sum_{\boldsymbol{\beta}\in\mathcal{N}^m}\left[\frac{\kappa(\alpha,\underline{\beta})}{m!}\mathbb{1}_{m\ge 2} + \frac{K(\alpha;\underline{\beta})-\kappa(\alpha,\underline{\beta})}{m!}\Big|^{\to}_{W_{\mathcal{N}}=0}\right]\partial_{\underline{\beta}}\big(\Lambda_{\alpha,\underline{\gamma}}f\big)\\ + \mathbf{E}\left(-\sum_{\alpha,\beta\in I}\kappa(\alpha,\beta) + \sum_{\alpha\in I,\beta\in\mathcal{N}}\kappa(\alpha,\beta)\right)\Lambda_{\alpha,\{\beta\}\sqcup\underline{\gamma}}f + \sum_\alpha\Omega(\Lambda_{\alpha,\underline{\gamma}}f,\alpha,\mathcal{N}). \quad (4.9b)\end{aligned}$$

We think of the first two terms and the first term of the square bracket in the third term (4.9b) as the leading order terms. The second summand in the third term will be small due to the structure of the pre-cumulants and the fact that the subsequent function $\partial\Lambda f$ has the $\mathcal{N}$-randomness removed. The fourth term is small because the two sums in the parenthesis almost cancel; and finally the fifth term will be small by choosing R sufficiently large. We call (4.9) (approximate) *cancellation identities* as they exhibit the cancellation of second order statistics due to the definition of $\mathcal{S}$ and D.

Iterated expansion of multiple factors of D. We now use (4.9b) repeatedly to compute $\mathbf{E}\prod_{k\in[p]}\Lambda^{(k)}(D)$. As a first step we expand the D in the $\Lambda^{(1)}$ factor, for which the special case (4.9a) is sufficient and we find

$$\begin{aligned}\mathbf{E}\,\Lambda^{(1)}(D)\prod_{k\ge2}\Lambda^{(k)}(D) = \sum_{\alpha_1\in I}\Omega\Big(\Lambda^{(1)}_{\alpha_1}\prod_{k\ge2}\Lambda^{(k)}(D),\alpha_1,\mathcal{N}(\alpha_1)\Big)\\ + \mathbf{E}\sum_{\substack{\alpha_1\in I\\ \beta_1\in\mathcal{N}(\alpha_1)}}\kappa(\alpha_1,\beta_1)\Lambda^{(1)}_{\alpha_1}\partial_{\beta_1}\Big(\prod_{k\ge2}\Lambda^{(k)}(D)\Big) + \mathbf{E}\left(-\sum_{\alpha_1,\beta_1\in I}\kappa(\alpha_1,\beta_1) + \sum_{\substack{\alpha_1\in I\\ \beta_1\in\mathcal{N}(\alpha_1)}}\kappa(\alpha_1,\beta_1)\right)\Lambda^{(1)}_{\alpha_1,\beta_1}\prod_{k\ge2}\Lambda^{(k)}(D)\\ + \mathbf{E}\sum_{\alpha_1\in I}\sum_{m<R}\sum_{\boldsymbol{\beta}_1\in\mathcal{N}(\alpha_1)^m}\left[\frac{\kappa(\alpha_1,\underline{\beta}_1)}{m!}\mathbb{1}_{m\ge2} + \frac{K(\alpha_1;\underline{\beta}_1)-\kappa(\alpha_1,\underline{\beta}_1)}{m!}\Big|^{\to}_{W_{\mathcal{N}(\alpha_1)}=0}\right]\partial_{\underline{\beta}_1}\Big(\Lambda^{(1)}_{\alpha_1}\prod_{k\ge2}\Lambda^{(k)}(D)\Big). \quad (4.10)\end{aligned}$$

We now distribute the $\underline{\beta}_1$-derivatives in the last term among the $\Lambda^{(1)}_{\alpha_1}$ and $\Lambda^{(k)}(D)$ factors according to the Leibniz rule. We handle the ∂_{β_1} derivative in the second term similarly but observe that this is slightly different in the sense that the ∂_{β_1} derivative does not hit the $\Lambda^{(1)}_{\alpha_1}$ factor. In other words, terms involving second order cumulants ($m=1$) come with the restriction that $\partial_{\beta_1}\Lambda^{(1)}_{\alpha_1}$ derivative is absent. This is precisely the effect we already encountered in Section 3.3; the self-energy normalization does not cancel all second order terms, it merely puts a restriction on the index-allocations in such a way that gains through

Ward estimates are guaranteed in all remaining terms. In order to write (4.10) more concisely we introduce the notations

$$\begin{aligned}
\sum_{\alpha_l,\boldsymbol{\beta}_l}^{\sim(l)} &:= \sum_{\alpha_l\in I}\sum_{1\le m<R}\sum_{\boldsymbol{\beta}_l\in\mathcal{N}(\alpha_l)^m}\frac{\kappa(\alpha_l,\underline{\beta}_l)}{m!}\sum_{\underline{\beta}_l^1\sqcup\cdots\sqcup\underline{\beta}_l^p=\underline{\beta}_l}\mathbb{1}\left(|\underline{\beta}_l^l|=0 \text{ if } |\underline{\beta}_l|=1\right),\\
\sum_{\alpha_l,\boldsymbol{\beta}_l}^{*} &:= \sum_{\alpha_l\in I}\sum_{0\le m<R}\sum_{\boldsymbol{\beta}_l\in\mathcal{N}(\alpha_l)^m}\sum_{\underline{\beta}_l^1\sqcup\cdots\sqcup\underline{\beta}_l^p=\underline{\beta}_l}\frac{K(\alpha_l;\underline{\beta}_l)-\kappa(\alpha_l,\underline{\beta}_l)}{m!},\\
\sum_{\alpha_l,\beta_l^l}^{\#} &:= \Big[-\sum_{\alpha_l,\beta_l^l\in I}\kappa_{\mathcal{S}}(\alpha_l,\beta_l^l)+\sum_{\alpha_l\in I}\sum_{\beta_l^l\in\mathcal{N}(\alpha_l)}\kappa(\alpha_l,\beta_l^l)\Big],
\end{aligned}\tag{4.11}$$

where $\kappa_{\mathcal{S}}(\alpha_1,\dots,\alpha_k) := \kappa(\widetilde{w}_{\alpha_1},\dots,\widetilde{w}_{\alpha_k})$ and where $\widetilde{W}=(\widetilde{w}_\alpha)_{\alpha\in I}$ is an identical copy of W. The reason for introducing this identical copy will become apparent in the next step. We furthermore follow the convention that $\underline{\beta}_l^k=\emptyset$ if $\underline{\beta}_l^k$ does not appear in the summation (which is the case for all $k\neq l$ in $\sum_{\alpha_l,\beta_l^l}^{\#}$ in (4.12)). Using these notations we can write (4.10) as

$$\mathbf{E}\prod_{k\in[p]}\Lambda^{(k)}(D)=\mathbf{E}\left(\sum_{\alpha_1,\boldsymbol{\beta}_1}^{\sim(1)}+\sum_{\alpha_1,\boldsymbol{\beta}_1}^{*}\Big|^{\rightarrow}_{W_{\mathcal{N}(\alpha_1)}=0}+\sum_{\alpha_1,\beta_1^1}^{\#}\right)\Lambda_{\alpha_1,\underline{\beta}_1^1}\prod_{k=2}^{p}\Lambda^{(k)}\big(\partial_{\underline{\beta}_1^k}D\big)+\Omega,\tag{4.12}$$

where the error term Ω collects all other terms and is defined in (4.13) below. We point out that the notations introduced in (4.11) implicitly depend on the parameter R determining the order of expansion.

Estimate of error term Ω. It remains to estimate the error term Ω which is bounded by

$$\Omega := \sum_{\alpha_1\in I}\Omega\Big(\Lambda_{\alpha_1}^{(1)}\prod_{k\ge2}\Lambda^{(k)}(D),\alpha_1,\mathcal{N}(\alpha_1)\Big)\le_R\sum_{\alpha_1,\boldsymbol{\beta}_1\in\mathcal{N}(\alpha_1)^R}\Big\|\partial_{\underline{\beta}_1}\Big(\Lambda_{\alpha_1}^{(1)}\prod_{k\ge2}\Lambda^{(k)}(D)\Big)\Big|_{\widehat{W}_t}\Big\|_2\tag{4.13}$$

for some $t\in[0,1]$, where $\widehat{W}_t=\widehat{W}_t^{(\alpha_1)}=tW_{\mathcal{N}(\alpha_1)}+W_{\mathcal{N}(\alpha_1)^c}$, where we recall the definition of $\Omega(\Lambda,\alpha,f)$ in (3.7a) and its bound in (3.8). To further estimate this expression, we first distribute the $\partial_{\underline{\beta}_1}$ derivative to the p factors involving $\Lambda^{(1)},\dots,\Lambda^{(p)}$ following the Leibniz rule, and then separate those factors by a simple application of Hölder inequality into p factors of $\|\cdot\|_{2p}$ norms. Each of these factors can be written as a sum of terms of the type $\big\|\Lambda^{(k)}(\partial_{\underline{\gamma}}G|_{\widehat{W}_t})\big\|_{2p}$ or $\big\|\Lambda^{(k)}(\partial_{\underline{\gamma}}D|_{\widehat{W}_t})\big\|_{2p}$ for some derivative operator $\partial_{\underline{\gamma}}$. We can then estimate these norms using

$$\|\Lambda(R)\|_q\le\|\Lambda\|\,\|R\|_q,\quad\text{and}\quad\big\|\partial_{\underline{\gamma}}G|_{\widehat{W}_t}\big\|_q+\big\|\partial_{\underline{\gamma}}D|_{\widehat{W}_t}\big\|_q\le_{|\underline{\gamma}|}N^{-|\underline{\gamma}|/2}(1+|||S|||)(1+\|G\|_{Cq|\underline{\gamma}|/\mu})^{(|\underline{\gamma}|+5)/\mu},\tag{4.14}$$

where the second inequality follows from Lemma D.3, and we note that $Cp|\underline{\gamma}|\le CRp^2$. We now count the total number of derivatives: There are $R+1$ derivatives from $|\underline{\beta}_1|$ and α_1, each providing a factor of $N^{-1/2}$. It remains to account for the $\alpha_1,\boldsymbol{\beta}_1$-sums which is at most of size $\sum_{\alpha_1}|\mathcal{N}(\alpha_1)|^R\le N^{2+R/2-\mu R}$. We now choose R large enough so that

$$N^{2-(R+1)/2+R/2-\mu R}\le N^{-p},$$

which is satisfied if we choose $R\ge3p/\mu$. Combining these rough bounds we have shown that, up to irrelevant combinatorial factors,

$$\Omega\le_{p,\mu}N^{-p}\Big[\prod_{k=1}^{p}\|\Lambda^{(k)}\|\Big]\Big(1+|||\mathcal{S}|||\Big)^p\Big(1+\|G\|_{Cp^3/\mu^2}\Big)^{Cp/\mu^2}.\tag{4.15}$$

Main expansion formula for multiple factors of D. Formula (4.12) with the bound (4.15) on the error term is the first step where the cumulant expansion was used in the $\Lambda^{(1)}(D)$ factor. Now we iterate this procedure for the $\Lambda^{(2)}(D),\Lambda^{(3)}(D),\dots$ inductively. We arrive at the following proposition modulo the claimed bound on the overall error which we will prove after an extensive explanation.

Proposition 4.4. *Let $\Lambda^{(1)},\dots,\Lambda^{(p)}$ be linear (or conjugate linear) functionals and let $p\in\mathbb{N}$ be given. Then we have*

$$\mathbf{E}\prod_{k\in[p]}\Lambda^{(k)}(D)=\mathbf{E}\prod_{l\in[p]}^{\rightarrow}\left(1+\sum_{\alpha_l,\boldsymbol{\beta}_l}^{\sim(l)}+\sum_{\alpha_l,\boldsymbol{\beta}_l}^{*}\Big|^{\rightarrow}_{W_{\mathcal{N}(\alpha_l)}=0}+\sum_{\alpha_l,\beta_l^l}^{\#}\right)\prod_{k\in[p]}\begin{cases}\Lambda^{(k)}_{\alpha_k,\bigsqcup_{l\in[p]}\underline{\beta}_l^k} & \text{if } \sum_{\alpha_k}\\ \Lambda^{(k)}_{\bigsqcup_{l<k}\underline{\beta}_l^k,\bigsqcup_{l>k}\underline{\beta}_l^k} & \text{else}\end{cases}+\Omega,\tag{4.16}$$

where "if $\sum_{\alpha_k}$" means cases where after multiplying out the first product $\prod_l$ the summation over the index α_k is performed. Under Assumptions (A), (B) and (D), the error term Ω is bounded by

$$|\Omega|\le_{p,\mu}N^{-p}\langle z\rangle^{-p}\Big[\prod_{k=1}^{p}\|\Lambda^{(k)}\|\Big](1+|||\mathcal{S}|||)^p\Big(1+\|G\|_q\Big)^{\frac{Cp}{\mu^2}}\left(1+\frac{\|G\|_q}{N^\mu}\right)^{\frac{Cp^2}{\mu^2}},\tag{4.17}$$

if we choose $R=4p/\mu$ to be order of expansion in the summations, see (4.11). Furthermore, we set $q:=Cp^3/\mu^2$ for some constant C, and $\|\Lambda^{(k)}\|$ denotes the operator norm of the linear functional $\Lambda^{(k)}$.

For (4.16) we recall the convention that $\underline{\beta}_l^k = \emptyset$ whenever $\underline{\beta}_l^k$ is not summed, i.e., for the contribution from the 1 in the l-th factor, or the contribution from $\sum^{\#}$ in the l-th factor for $k \neq l$. Moreover, we remind the reader that the custom notation $|^{\rightarrow}_{W_{\mathcal{N}}=0}$ was introduced right after (3.15). We also note that the terms with a 1 from the first factor vanish as they contain $\Lambda^{(1)}_{\emptyset,\sqcup_{l>1}\underline{\beta}_l^1} = 0$. Moreover, we can now explain why we introduced the identical copy $\widetilde{W}$ of W in the definition of $\kappa_{\mathcal{S}}$ in (4.11). The cumulants in the representation of the term $\mathcal{S}[G]G = -\sum_{\alpha,\beta\in I}\kappa_{\mathcal{S}}(\alpha,\beta)\Lambda_{\alpha,\beta}$ should not be affected by the restriction imposed by the operation $|^{\rightarrow}_{W_{\mathcal{N}}=0}$. Changing W to $\widetilde{W}$ within the definition of κ_S protects it from the action of $|^{\rightarrow}_{W_{\mathcal{N}}=0}$ that turns all subsequent W variables zero. This non-restriction of the particular sum is formally achieved by writing $\mathcal{S}$ in terms of $\kappa_{\mathcal{S}}$ instead of κ. This is only a notational pedantry, in the next step where we multiply (4.16) out, it will disappear. We remark that because of the effect of $|^{\rightarrow}_{W_{\mathcal{N}}=0}$ the order in which the product in (36) is performed matters. It starts with $l = 1$ and ends with $l = p$.

We point out that the estimate (4.17) not only provides the necessary N^{-p} factor, but it also involves at most $O(p)$ power of $\|G\|_q$ without an extra smallness factor $N^{-\mu}$, see Remark 4.3. While from the perspective of an N-power counting, any factor $\|G\|_q$ is neutral, of order one, we need to track that its power is not too big. Factors of $\|G\|_q$ that come with a factor $N^{-\mu}$ can be handled much easier and are not subject to the restriction of their power.

Reformulation of the main expansion formula. We now derive an alternative, less compact formula (4.18) for (4.16) which avoids the provisional $|^{\rightarrow}$ notation. By expanding the first product in (4.16) we can rearrange (4.16) according to partitions $[p] = L_1 \sqcup \cdots \sqcup L_4$, where L_i contains those indices l for which the l-th factor in the product contributes with its i-th term. In particular $L := L_2 \sqcup L_3 \sqcup L_4 \subset [p]$ contains those indices l, for which $\alpha_l, \boldsymbol{\beta}_l$ are summed. We shall use the nomenclature that labels α_l and the elements of $\underline{\beta}_l$ are *type-l* labels. These labels have been generated in the l-th application of the cancellation identities (4.9). The partition $\underline{\beta}_l^1 \sqcup \cdots \sqcup \underline{\beta}_l^p = \underline{\beta}_l$ encodes how these labels have been distributed among the p factors via the Leibniz rule. Thus labels $\underline{\beta}_l^k$ have been generated on $\Lambda^{(k)}$ at the l-th application of (4.9). Thus L encodes the types of labels present in the different parts of the expansion. To specify the number of type–l labels we introduce the notations

$$M_l := |\underline{\beta}_l|, \quad M_l^k := |\underline{\beta}_l^k|.$$

Thus the number of labels of *type l* is $M_l + 1$ and the number of type l-labels in $\Lambda^{(k)}$ is $M_l^k + \delta_{lk}$. We observe that in all non-zero terms of (4.16) the labels α_l, $\underline{\beta}_l$ for $l \in L$ are distributed to the $\Lambda^{(1)}, \ldots, \Lambda^{(p)}$ in such a way that

(a) there are p factors $\Lambda^{(1)}, \ldots, \Lambda^{(p)}$,
(b) every $\Lambda^{(k)}$ carries at least one label (that is for all k, $\sum_{l\in L}(M_l^k + \delta_{kl}) \geq 1$),
(c) for every $l \in L$, there exist at least two and at most $R-1$ *type-l* labels (that is for all $l \in L$, $M_l \geq 1$), for $l \in L_4$ there exist exactly two *type-l* labels in such a way that $M_l = M_l^l = 1$,
(d) if for some $l \in L_2$ there are exactly two *type-l* labels, then these two labels must occur in distinct $\Lambda's$ (that is, if $l \in L_2$ and $M_l = 1$, then $M_l^l = 0$).
(e) for every $l \in L$, the first index of $\Lambda^{(l)}$ is α_l.

We now reformulate (4.16) in such a way that we first sum up over the partitions $L_1 \sqcup L_2 \sqcup L_3 \sqcup L_4 = [p]$, the collection of multiplicities $M = (\, M_l^k \mid l \in L, k \in [p]\,)$ and the permutations of indices, and only then perform the actual summation over the labels from I. As the first three sums carry no N, they are irrelevant for the N-power counting. From (4.16) we find

$$\mathbf{E}\prod_{k\in[p]}\Lambda^{(k)}(D) = \mathbf{E}\sum_{\sqcup L_i=[p]}\sum_{M}^{\sim(L)} C_M \sum_{\sigma}^{\sim(M)}\left[\prod_{l\in L_3}\sum_{\alpha_l,\boldsymbol{\beta}_l\notin\mathcal{N}_{L_3}^{<l}}^{(M,l)}\frac{K(\alpha_l;\boldsymbol{\beta}_l)-\kappa(\alpha_l,\boldsymbol{\beta}_l)}{|\boldsymbol{\beta}_l|!}\right]\mathcal{M}' + \mathcal{O}_{p,\mu}\left(N^{-p}\right), \tag{4.18}$$

$$\mathcal{M}' := \left[\prod_{l\in L_4}\left(-\sum_{\alpha_l,\beta_l^l\in I} + \sum_{\alpha_l\in I\setminus\mathcal{N}_{L_3}^{<l}}\sum_{\beta_l^l\in\mathcal{N}(\alpha_l)\setminus\mathcal{N}_{L_3}^{<l}}\right)\frac{\kappa(\alpha_l,\beta_l^l)}{1!}\right]\mathcal{M},$$

$$\mathcal{M} := \left[\prod_{l\in L_2}\sum_{\alpha_l,\boldsymbol{\beta}_l\notin\mathcal{N}_{L_3}^{<l}}^{(M,l)}\frac{\kappa(\alpha_l,\boldsymbol{\beta}_l)}{|\boldsymbol{\beta}_l|!}\right]\left[\left(\prod_{k\in L}\Lambda^{(k)}_{\alpha_k,\sigma_k(\boldsymbol{\beta}^k)}\right)\left(\prod_{k\notin L}\Lambda^{(k)}_{\sigma_k(\boldsymbol{\beta}^k)}\right)\right]\Big|_{W_{\mathcal{N}_{L_3}}=0},$$

where $\sum_M^{\sim(L)}$ is the sum over all arrays M fulfilling (a)–(e) above and C_M are purely combinatorial constants bounded by a function of p, R; $C_M \leq C(p,R)$, in which we also absorbed the (-1)'s from the L_4 terms. Moreover, $\sum_\sigma^{\sim(M)}$ is the sum over all permutations $\sigma_1,\ldots,\sigma_p$ in the permutation groups $S_{M^1},\ldots,S_{M^p}$ (where $M^k := \sum_{l\in L} M_l^k$) such that for $k \notin L$ the first element of $\sigma_k(\boldsymbol{\beta}^k)$ is from $(\,\boldsymbol{\beta}_l^k \mid l \in L \cap [k])$. Furthermore, for any $\mathcal{N} \subset I$ we set

$$\sum_{\alpha_l,\boldsymbol{\beta}_l\notin\mathcal{N}}^{(M,l)} := \sum_{\alpha_l\in I\setminus\mathcal{N}}\prod_{k\in[p]}\sum_{\boldsymbol{\beta}_l^k\in(\mathcal{N}(\alpha_l)\setminus\mathcal{N})^{M_l^k}}.$$

Finally, we introduced the notations $\mathcal{N}_{L_3}^{<l} := \bigcup_{l>k\in L_3}\mathcal{N}(\alpha_k)$, and $\mathcal{N}_{L_3} := \bigcup_{k\in L_3}\mathcal{N}(\alpha_k)$. Here the $\boldsymbol{\beta}_l^k$ are actual (ordered) tuples and not multisets, which is why we denote them by boldfaced Greek letters to avoid possible confusion with the previously used $\underline{\beta}_l^k$. In (4.18) we furthermore used the short-hand notation $\boldsymbol{\beta}^k = (\boldsymbol{\beta}_l^k)_{l\in L}$ for the tuple (ordered according to the natural order on $L \subset [p] \subset \mathbb{N}$) of $\boldsymbol{\beta}_l^k$. We note that the artificial $\kappa_{\mathcal{S}}$ from (4.16) has been removed in (4.18) since we "pushed" the $|^{\rightarrow}$-operator

all the way to the end. In the following we will establish bounds on (4.18) for fixed L and M and fixed permutations $\sigma_1, \dots, \sigma_p$. Since the number of possible choices for M, L and permutations is finite, depending on R and p only, this will be sufficient for bounding $\mathbf{E}\prod \Lambda^{(k)}(D)$. We also stress that the (multi)labels $\boldsymbol{\beta}_l^k$ themselves are not important, but only their type l.

Proof of the error bound in Proposition 4.4. We now turn to the proof of the claimed error bound (4.17). So far this was only done for the error from the first cumulant expansion in (4.15).

Proof of the error bound in Proposition 4.4. The error Ω in (4.16) is a sum over p terms, where the j-th term is the error from the expansion of $\Lambda^{(j)}(D)$. Recalling the definition of $\Omega(f, i, \mathcal{N})$ from (3.7b), this j-th expansion error is given by

$$\Omega_j := \sum_{\alpha_j} \Omega\left(\prod_{l<j} \Big(1 + \sum_{\alpha_l, \boldsymbol{\beta}_l}^{\sim(l)} + \sum_{\alpha_l, \boldsymbol{\beta}_l}^{*} \Big|_{W_{\mathcal{N}(\alpha_l)}=0}^{\rightarrow} + \sum_{\alpha_l, \beta_l^l}^{\#} \Big) \prod_{k=1}^{p} \begin{Bmatrix} \Lambda^{(k)}_{\alpha_k, \bigsqcup_{l\in[p]} \underline{\beta}_l^k} & \text{if } k = j \text{ or } \big(k<j, \textstyle\sum_{\alpha_k}\big) \\ \Lambda^{(k)}\big(\partial_{\bigsqcup_{l<k} \underline{\beta}_l^k} D\big) & \text{if } k > j \\ \Lambda^{(k)}_{\bigsqcup_{l<k} \underline{\beta}_l^k, \bigsqcup_{l>k} \underline{\beta}_l^k} & \text{else} \end{Bmatrix}, \alpha_j, \mathcal{N}(\alpha_j) \right),$$

where "if $(k<j, \sum_{\alpha_k})$" means "if $k<j$ and α_k is summed". This j-th error Ω_j can be estimated through (3.8) and Assumption (B) by the sum of

$$\Big[\prod_{l\in L_2 \sqcup L_3} \sum_{\alpha_l, \boldsymbol{\beta}_l}^{(M,l)} \Big] \Big[\prod_{l\in L_4} \sum_{\alpha_l, \beta_l^l \in I} \Big] \sum_{\alpha_j} \sum_{\boldsymbol{\beta}_j \in \mathcal{N}(\alpha_j)^R} \Big\| \Big(\prod_{k\in L} \Lambda^{(k)}_{\alpha_k, \sigma_k(\boldsymbol{\beta}^k)} \Big) \Big(\prod_{k\in[j]\setminus L} \Lambda^{(k)}_{\sigma_k(\boldsymbol{\beta}^k)} \Big) \Big(\prod_{k>j} \Lambda^{(k)}(\partial_{\sigma_k(\boldsymbol{\beta}^k)} D)\Big|_{\widehat{W}} \Big) \Big\|_2, \tag{4.19}$$

over partitions $L = L_2 \sqcup L_3 \sqcup L_4 \subset [j-1]$, arrays M fulfilling (a)–(e) above and partitions σ_k. In all terms $\widehat{W}$ is a modification of W which differs from W in at most $C\sqrt{N}$ entries. The previously studied error from (4.13) for example corresponds to $j=1$, $L_2 = L_3 = L_4 = \emptyset$. The combinatorics of all these summations are independent of N, hence can be neglected. So we can focus on a single term of the form (4.19). The norm in (4.19) will first be estimated by Hölder and then by (4.14) to reduce it to many factor of $\|G\|_q$. We now have to count the size of the sums, the number of $N^{-1/2}$ factors from the derivatives, and the number of $\|G\|_q$'s we collect in the bound. We start with the sums which are at most of size

$$N^{2|L_2\sqcup L_3|}(N^{1/2-\mu})^{M_{L_2\sqcup L_3}}(N^2\cdot N^2)^{|L_4|}N^2(N^{1/2-\mu})^R = N^{2|L_2\sqcup L_3|+(M_{L_2\sqcup L_3}+R)(1/2-\mu)+4|L_4|+2}. \tag{4.20}$$

Here the first factor comes from the α_l summations for $l \in L_2 \sqcup L_3$, while the second term comes from the corresponding $\boldsymbol{\beta}_l$ summations. The third factor comes from the α_l, β_l^l-summations for $l \in L_4$, and finally the fifth and sixth factor correspond to the α_j and $\boldsymbol{\beta}_j$ summations. Next, we count the total number of derivatives. Every index α_l and β_l^k accounts for a derivative, and each derivative contributes a factor of $N^{-1/2}$. So we have

$$(N^{-1/2})^{|L_2\sqcup L_3|+M_{L_2\sqcup L_3}+2|L_4|+(R+1)} = N^{-|L_2\sqcup L_3|/2-M_{L_2\sqcup L_3}/2-|L_4|-(R+1)/2}, \tag{4.21}$$

so that altogether from (4.20) and (4.21) we have an N-power of

$$N^{3/2(|L_2\sqcup L_3|+1)+3|L_4|-R\mu}N^{-\mu M_{L_2\sqcup L_3}} \le N^{-p}N^{-\mu M_{L_2\sqcup L_3}}.$$

It remains to count the number of $\|G\|_{CRp^2} = \|G\|_q$ coming from the application of (4.14), which in total provides

$$\sum_{k\in L}(1+|\boldsymbol{\beta}^k|+5) + \sum_{k\in[j]\setminus L}(|\boldsymbol{\beta}^k|+5) + \sum_{k>j}(|\boldsymbol{\beta}^k|+5) = 5p + |L_2\sqcup L_3| + M_{L_2\sqcup L_3} + 2|L_4| + R + 1 \le Cp/\mu + M_{L_2\sqcup L_3} \tag{4.22}$$

factors of $\|G\|_q$. The claim (4.17) now follows from the trivial estimate $M_{L_2\sqcup L_3} \le Rp \le Cp^2/\mu$. □

Subsequently we establish a bound on the rhs. of (4.18), by first estimating it in terms of $\|\mathcal{M}'\|_p$, then estimating $\|\mathcal{M}'\|_p$ in terms of $\|\mathcal{M}\|_p$ and finally bounding the leading contribution $\mathcal{M}$. We consider the first two steps in this procedure as errors stemming from the neighbourhood structure of the expansion, while the third step is concerned with the leading order contribution from the expansions. In Section 4.2 we consider the errors stemming from the neighbourhood structure, while in Sections 4.3 and 4.4 we derive bounds on $\|\mathcal{M}\|_p$ for the averaged and isotropic case, separately. For simplicity we first carry out the technically most involved argument from Sections 4.3–4.4 in the extreme case $L_3 = L_4 = \emptyset$ where the neighbourhood errors are absent. Finally, we explain the necessary modifications for the general case in Section 4.5.

4.2. Bound on neighbourhood errors

We start with the bound on the L_3-factors in (4.18). Neglecting the irrelevant combinatorial factors $|\boldsymbol{\beta}_l|!$ and the summations over L_i, M and σ, we have to estimate

$$\mathcal{E} := \Big[\prod_{l\in L_3} \sum_{\alpha_l, \boldsymbol{\beta}_l \notin \mathcal{N}_{L_3}^{<l}}^{(M,l)} \Big] \mathcal{E}(\alpha_{L_3}, \boldsymbol{\beta}_{L_3}) := \Big[\prod_{l\in L_3} \sum_{\alpha_l, \boldsymbol{\beta}_l \notin \mathcal{N}_{L_3}^{<l}}^{(M,l)} \Big] \mathbf{E}\,\mathcal{M}' \prod_{l\in L_3} \big[K(\alpha_l; \boldsymbol{\beta}_l) - \kappa(\alpha_l, \boldsymbol{\beta}_l) \big]. \tag{4.23}$$

By the pigeon hole-principle we find that for every $l \in L_3$ and any assignment of $\alpha_l, \boldsymbol{\beta}_l$ there exist some $n_l < R$ such that we have a partition $\underline{\beta}_l = \underline{\beta}_l^{(i)} \sqcup \underline{\beta}_l^{(o)}$ into *inside* and *outside* elements with $\underline{\beta}_l^{(i)} \subset \mathcal{N}_{n_l}(\alpha_l)$ and $\underline{\beta}_l^{(o)} \subset \mathcal{N}_{n_l+1}(\alpha_l)^c$ since $|\underline{\beta}_l| = M_l < R$ (see rule (c)). We recall the nested structure of the neighbourhoods as stated in Assumption (D), and provide an illustration of the "security layers" in Figure 1. According to (3.1c) we can then write (L_3' collects those indices where we took the middle term

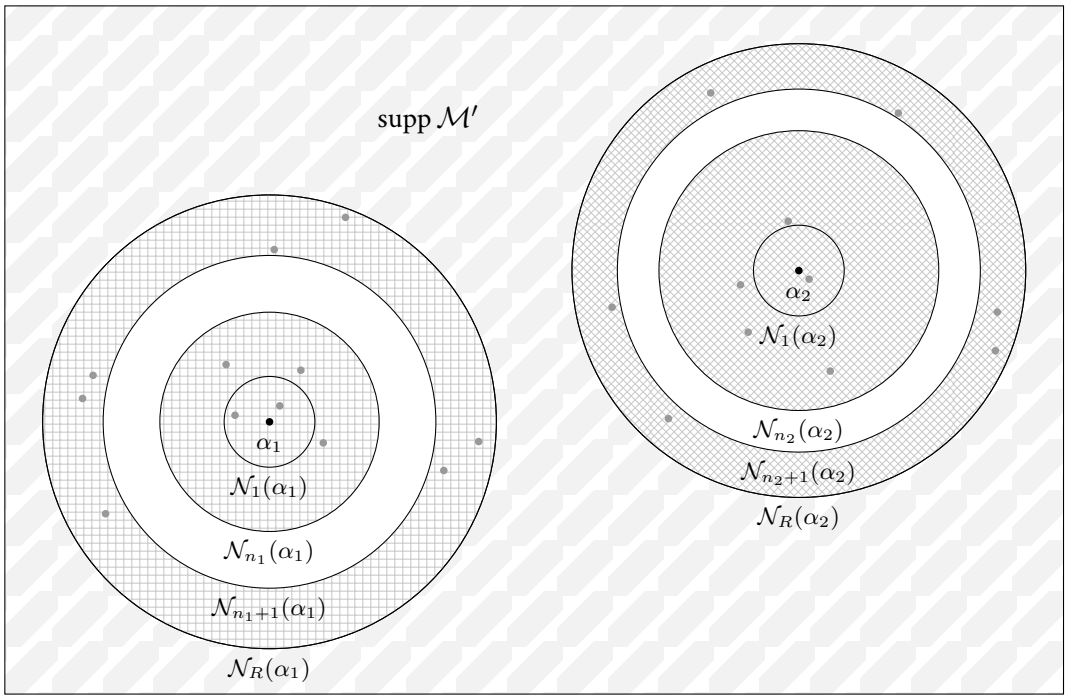


Figure 1. Illustration for the bound on $\mathcal{E}$ (4.23). Gray dots • denote the $\underline{\beta}_1, \underline{\beta}_2$ labels. Since there are $|\underline{\beta}_i| < R$ labels and R rings, there is always one empty ring by the pigeon-hole principle.

of (3.1c) in the l factor)

$$\mathcal{E}(\alpha_{L_3}, \boldsymbol{\beta}_{L_3}) = \sum_{L_3 = L_3' \sqcup L_3''} (-1)^{|L_3''|} \prod_{l \in L_3''} \Bigg[\sum_{\gamma_l^{(i)} \subset \beta_l^{(i)}} \sum_{\gamma_l^{(o)} \subsetneq \beta_l^{(o)}} \kappa(\alpha_l, \underline{\beta}_l^{(i)} \setminus \underline{\gamma}_l^{(i)}, \underline{\beta}_l^{(o)} \setminus \underline{\gamma}_l^{(o)}) \Bigg] \mathbf{E}\, f \prod_{l \in L_3'} \big[K(\alpha; \underline{\beta}_l^{(i)}) - \kappa(\alpha, \underline{\beta}_l^{(i)}) \big],$$

where

$$f \coloneqq \mathcal{M}' \prod_{l \in L_3'} (\Pi \underline{\beta}_l^{(o)}) \prod_{l \in L_3''} \Big[(\Pi \underline{\gamma}_l^{(i)}) (\Pi \underline{\gamma}_l^{(o)}) \Big]$$

is a random variable supported in $\bigcap_{l \in L_3'} \mathcal{N}_{n_l+1}(\alpha_l)^c$, i.e., well separated from the variables $K(\alpha_l; \underline{\beta}_l^{(i)})$ for $l \in L_3'$. It remains to estimate a quantity of the type $\mathbf{E}\, f g_1 \dots g_k$, where $f, g_1, \dots, g_k$ are random variables whose supports are pairwise separated by "security layers" and where each g_i is of the form $K - \kappa$ with $\mathbf{E}\, g_i = 0$. Here $k = |L_3'|$ and from Lemma 3.3 and Assumption (D) it follows that $\mathbf{E}\, f g_1 \dots g_k \lesssim_k \|f\|_{k+1} N^{-3\lceil k/2 \rceil}$. According to Lemma A.1 the $\kappa(\alpha_l, \underline{\beta}_l^{(i)} \setminus \underline{\gamma}_l^{(i)}, \underline{\beta}_l^{(o)} \setminus \underline{\gamma}_l^{(o)})$ factors are also at least N^{-3} small and we can conclude that

$$|\mathcal{E}(\alpha_{L_3}, \boldsymbol{\beta}_{L_3})| \lesssim_{p,R} N^{-3\lceil |L_3|/2 \rceil} \left\| \mathcal{M}' \right\|_p . \tag{4.24}$$

Next, we use the triangle inequality to pull the L_4 summation out of $\|\mathcal{M}'\|_p$ to achieve a bound in terms of $\|\mathcal{M}\|_p$. We have

$$\begin{aligned} \Bigg| \Bigg(- \sum_{\alpha_l, \beta_l^l \in I} + \sum_{\alpha_l \in I \setminus \mathcal{N}_{L_3}^{<l}} \sum_{\beta_l^l \in \mathcal{N}(\alpha_l) \setminus \mathcal{N}_{L_3}^{<l}} \Bigg) \kappa(\alpha_l, \beta_l^l) \Bigg| &\le \Bigg(\sum_{\alpha_l \in I \setminus \mathcal{N}_{L_3}^{<l}} \sum_{\beta_l^l \in \mathcal{N}_{L_3}^{<l}} + \sum_{\alpha_l \in I \setminus \mathcal{N}_{L_3}^{<l}} \sum_{\beta_l^l \in I \setminus \mathcal{N}(\alpha_l)} + \sum_{\alpha_l \in \mathcal{N}_{L_3}^{<l}} \sum_{\beta_l^l \in I} \Bigg) |\kappa(\alpha_l, \beta_l^l)| \\ &\le \Bigg(\sum_{\beta_l^l \in \mathcal{N}_{L_3}^{<l}} \sum_{\alpha_l \in \mathcal{N}(\beta_l^l)} + \sum_{\beta_l^l \in \mathcal{N}_{L_3}^{<l}} \sum_{\alpha_l \in I \setminus \mathcal{N}(\beta_l^l)} + \sum_{\alpha_l \in I} \sum_{\beta_l^l \in I \setminus \mathcal{N}(\alpha_l)} + \sum_{\alpha_l \in \mathcal{N}_{L_3}^{<l}} \sum_{\beta_l^l \in \mathcal{N}(\alpha_l)} + \sum_{\alpha_l \in \mathcal{N}_{L_3}^{<l}} \sum_{\beta_l^l \in I \setminus \mathcal{N}(\alpha_l)} \Bigg) |\kappa(\alpha_l, \beta_l^l)| \le CN, \end{aligned}$$

where we estimated the first and the fourth term with two small summations purely by size $(CN^{1/2-\mu})^2 \le CN$ and the other terms using the fact that $|\kappa(\alpha, \beta)| \lesssim N^{-3}$ for $\beta \in I \setminus \mathcal{N}(\alpha)$. Summarizing, we thus have that

$$\Big| \mathbf{E} \prod \Lambda^{(k)}(D) \Big| \lesssim_{p,\mu} N^{-p} + \sum_{\bigsqcup L_i = [p]} \sum_{M}^{\sim (L)} \frac{N^{|L_4|}}{N^{3\lceil |L_3|/2 \rceil}} \sum_{\sigma}^{\sim (M)} \Bigg[\prod_{l \in L_3} \sum_{\alpha_l, \boldsymbol{\beta}_l \notin \mathcal{N}_{L_3}^{<l}}^{(M,l)} \Bigg] \Bigg[\prod_{l \in L_4} \max_{\alpha_l, \beta_l^l \in I} \Bigg] \|\mathcal{M}\|_p , \tag{4.25}$$

and it only remains to estimate the leading order term $\mathcal{M}$, as defined in (4.18). This has to be done separately for averaged and isotropic bound and should be considered as the main part of the proof. To simplify notations we will first prove the bound on $\mathcal{M}$ for the case that $L_3 = L_4 = \emptyset$ and $\mathcal{N}(\alpha) = I$. In particular $L_3 = \emptyset$ implies that $\mathcal{N}_{L_3} = \emptyset$ and therefore in the next two

Sections 4.3 and 4.4 we now aim at deriving a bound on $\big\|\mathcal{M}((\Lambda^{(k)})_{k\in[p]};L,M,\sigma)\big\|_p$, where

$$\mathcal{M}((\Lambda^{(k)})_{k\in[p]};L,M,\sigma) \coloneqq \left[\prod_{l\in L}\sum_{\alpha_l,\boldsymbol{\beta}_l}^{(M,l)}\frac{\kappa(\alpha_l,\boldsymbol{\beta}_l)}{|\boldsymbol{\beta}_l|!}\right]\left(\prod_{k\in L}\Lambda^{(k)}_{\alpha_k,\sigma_k(\boldsymbol{\beta}^k)}\right)\left(\prod_{k\notin L}\Lambda^{(k)}_{\sigma_k(\boldsymbol{\beta}^k)}\right), \quad \sum_{\alpha_l,\boldsymbol{\beta}_l}^{(M,l)} \coloneqq \sum_{\alpha_l\in I}\prod_{k\in[p]}\sum_{\boldsymbol{\beta}_l^k\in I^{M_l^k}}. \tag{4.26}$$

The definition of $\mathcal{M}$ in (4.26) agrees with the one in (4.18) in the special case $L_3=L_4=\emptyset$, except for a tiny contribution from $\boldsymbol{\beta}_l\not\subset\mathcal{N}(\alpha_l)$. The reason for extending the sum here to the whole index set is twofold: First, we do not have to keep track of the summation ranges of individual indices, and, second, we demonstrate that for the main terms separating the contribution outside of the neighbourhoods $\mathcal{N}$ is not necessary, all estimates on $\mathcal{M}$ would also hold for the unrestricted sum. In particular, the neighbourhood decay condition is not necessary for the main terms, they are used only for bounding $\mathcal{M}'$ in terms of $\mathcal{M}$ in Section 4.2. This fact was already advertised in Example 2.12 where we claimed that in the Gaussian case we can considerably relax our decay conditions. Later, in Section 4.5 we will explain how to elevate the proof for the special case $L_3=L_4=\emptyset$ with extended index sets to the general case.

4.3. **Averaged bound on** D

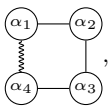
To treat (4.26) systematically, we introduce a graphical representation for any M, L and permutations σ in (4.26). For the averaged local law we need averaged estimates on D, so we set

$$\Lambda^{(k)}(D)\coloneqq\langle BD\rangle \qquad \text{or} \qquad \Lambda^{(k)}(D)\coloneqq\overline{\langle BD\rangle},$$

where B is a generic norm-bounded matrix, $\|B\|\lesssim 1$ and we recall that $\langle\cdot\rangle=N^{-1}\operatorname{Tr}$ denotes the normalized trace. A factor $\Lambda_{\alpha_1,\dots,\alpha_n}$ can be represented as a directed cyclic graph on the vertex set $\{\alpha_1,\dots,\alpha_n\}$. Up to sign we have

$$|\Lambda_{\alpha_1,\dots,\alpha_n}|=N^{-n/2}\langle B\Delta^{\alpha_1}G\Delta^{\alpha_2}G\dots\Delta^{\alpha_n}G\rangle=N^{-1-n/2}G_{b_1a_2}G_{b_2a_3}\dots G_{b_{n-1}a_n}(GB)_{b_na_1}, \tag{4.27}$$

which we represent as a cyclic graph in such a way that the vertices represent labels $\alpha_i=(a_i,b_i)$ and a directed edge from $\alpha_i=(a_i,b_i)$ to $\alpha_j=(a_j,b_j)$ represents $G_{b_ia_j}$. Since we will always draw the graphs in a clockwise orientation we will not indicate the direction of the edges specifically. The specific GB factor will be denoted by a wiggly line instead of a straight line used for the G factors. As an example, we have the correspondences

$\Lambda_{\alpha_1,\alpha_2,\alpha_3,\alpha_4}\leftrightarrow$ 

, $\Lambda_{\alpha_1,\alpha_2}\leftrightarrow$ 

and $\Lambda_{\alpha_1}\leftrightarrow$ 

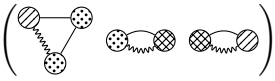
.

In (4.26) the labels of type l are connected through the $\kappa(\alpha_l,\boldsymbol{\beta}_l)$ factor which strongly links those labels due to the decay properties of the cumulants. We represent this fact graphically as a vertex colouring of the graph in which label types correspond to colours. The set of colours representing the label types L will be denoted by C. The M_l+1 vertices of a given type l will be denoted by V_c, where c is the colour corresponding to l.

We define $\operatorname{Val}(\Gamma)$, the *value* of a graph Γ, as summation over all labels consistent with the colouring, such that equally coloured labels are linked through a cumulant, of the product of the corresponding Λ's, just as in (4.26). For example, we have

$$\sum_{\alpha_1,\beta_1^2(1)}\kappa(\alpha_1,\beta_1^2(1))\sum_{\alpha_2,\beta_2^1(1)}\kappa(\alpha_2,\beta_2^1(1))\Lambda_{\alpha_1,\beta_2^1(1)}\Lambda_{\alpha_2,\beta_1^2(1)}=\operatorname{Val}\left(\right. \tag{4.28}$$


or

$$\sum_{\alpha_1,\beta_1^2(1)}\kappa(\alpha_1,\beta_1^2(1))\sum_{\alpha_2,\beta_2^1(1),\beta_2^1(2)}\frac{\kappa(\alpha_2,\beta_2^1(1),\beta_2^1(2))}{2!}\sum_{\alpha_3,\beta_3^2(1)}\kappa(\alpha_3,\beta_3^2(1))\Lambda_{\alpha_1,\beta_2^1(2),\beta_2^1(1)}\Lambda_{\alpha_2,\beta_1^2(1)}\Lambda_{\alpha_3,\beta_3^3(1)}$$
$$=\operatorname{Val}\left(\right.$$


,

where we choose the variable names for the labels in accordance with (4.26) following the convention that the elements of the tuple $\boldsymbol{\beta}_l^k$ are denoted by $(\beta_l^k(1),\beta_l^k(2),\dots)$. We warn the reader that $\operatorname{Val}(\Gamma)$, the value of a diagram itself is a random variable unlike in customary Feynman diagrammatic expansion theory. In the following we will derive bounds on the value of diagrams. To separate the conceptual from the technical difficulties we first derive those bounds in a vague $\lesssim$ sense which ignores a technical subtlety: The entries G_{ab} of the resolvent are bounded with overwhelming probability, but usually not almost surely. In the first conceptual step we will tacitly assume such an almost sure bound and write $|G_{ab}|\lesssim 1$. Later in Section 4.3.1 we will make the bounds rigorous in a high-moment sense. We note that if $\Lambda(D)=\overline{\langle BD\rangle}$, then the edges would represent G^* and $(GB)^*$ instead of G and GB and the order would be reversed (recall that the double indices are transposed in (4.6)) but the counting argument is not sensitive to these nuances, so we omit these distinctions in our graphs.

We now rephrase the rules on M in this graphical representation. They dictate that we need to consider the set of all vertex coloured graphs Γ with cyclic components such that

(a) there exist p connected components, all of which are cycles,
(b) each connected component contains at least one vertex,
(c) each colour colours at least two vertices,
(d) if a colour colours exactly two vertices, then these vertices are in different components.

(e) for each colour there exists a component in which the vertex after the wiggled edge (in clockwise orientation) is of that colour.

We note that these rules, compared to (4.26), disregarded the restrictions on the permutations σ_k for $k \notin L$ as these are not relevant for the averaged bound. The set of graphs satisfying (a)–(e) will be denoted by $\mathcal{G}^{\mathrm{av}}(p,R)$ and for each L, M, σ the main term $\mathcal{M}$ from (4.26) is given by the value of some graph $\Gamma \in \mathcal{G}^{\mathrm{av}(p,R)}$.

$$\mathcal{M}\Big(\big(\langle B\cdot\rangle^{[p/2]}, \overline{\langle B\cdot\rangle}^{[p/2]} \big); L, M, \sigma \Big) = \mathrm{Val}(\Gamma), \qquad \Gamma = \Gamma(L,M,\sigma) \in \mathcal{G}^{\mathrm{av}(p,R)} \tag{4.29}$$

where $\langle B\cdot\rangle^{[p/2]}$ denotes the tuple of $p/2$ functionals mapping $D \mapsto \langle BD\rangle$ and similarly for $\overline{\langle B\cdot\rangle}$. As the number of such graphs is finite for given p, R it follows that it is sufficient to prove the required bound for every single graph.

As for any fixed colour $\sum_{\oslash} \le N^2 \,|||\kappa|||^{\mathrm{av}}$, the naive size of the value $\mathrm{Val}(\Gamma)$ is bounded by

$$\mathrm{Val}(\Gamma) \lesssim N^{-p} \prod_{c\in C} N^{2-|V_c|/2} \le 1 \tag{4.30}$$

since according to (4.27) every component contributes a factor N^{-1} and every label contributes a factor $N^{-1/2}$, and where the ultimate inequality followed from $|V_c| \ge 2$ and $|C| \le p$. We now demonstrate that using Ward identities of the form

$$\sum_a |G_{ab}|^2 = \frac{(\Im G)_{bb}}{\eta}$$

we can improve upon this naive size by a factor of ψ^{2p}, where $\psi \approx 1/\sqrt{N\eta}$ and $\eta := \Im z$. We will often use the Ward identity in the form

$$\sum_b |G_{ab}| \le \sqrt{N}\sqrt{\sum_b |G_{ab}|^2} = N\sqrt{\frac{(\Im G)_{aa}}{N\eta}} \lesssim N\psi, \qquad \sum_b |(GB)_{ab}| \lesssim \|B\|\, N\psi \tag{4.31a}$$

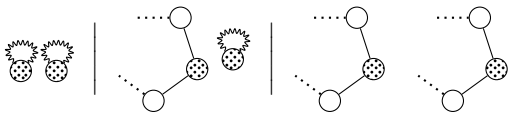
which explicitly exhibits a gain of a factor ψ over the trivial bound of order N. Together with the previous bound

$$\sum_b |G_{ab}|^2 \le N\psi^2, \qquad \sum_b |(GB)_{ab}|^2 \lesssim \|B\|^2\, N\psi^2 \tag{4.31b}$$

we will call (4.31a)–(4.31b) *Ward estimates*. Here we used the trivial bound $|G| \lesssim 1$ and we set $\psi := \sqrt{\Im G/N\eta}$ (where $\Im G$ is meant in an isotropic sense which we will define rigorously later).

We consider the subset of colours $C' := \{\, c \in C \mid |V_c| \le 3 \,\} \subset C$ which colour either two or three vertices and we intend to use Ward identities only when summing up vertices with those colours. However, one may not use Ward estimates for every such summation, e.g. even if both a and b were indices of eligible labels, one cannot gain from both of them in the sum $\sum_{a,b} |G_{ab}|$. We thus need a systematic procedure to identify sufficiently many labels so that each summation over them can be performed by using Ward estimates. In the following, we first describe a procedure how to *mark* those edges we can potentially use for Ward estimates. Secondly, we will show that for sufficiently many marked edges the Ward estimates can be used in parallel.

Procedure for colours appearing twice in Γ. If a colour $\circledcirc$ appears twice, then it appears in two different components of Γ, i.e., in one of the following forms



where the white vertices can be of any colour other than $\circledcirc$ (and may even coincide), the dotted edges indicate an arbitrary continuation of the component and some additional edges may be wiggled. The picture only shows those two components with colour $\circledcirc$, the other components of Γ are not drawn. Vertical lines separate different cases. When summing up the $\circledcirc$-coloured labels, we can use the Ward estimates on all edges adjacent to $\circledcirc$ using the operator norm $|||\kappa|||_2^{\mathrm{av}} = \big\| \,|\kappa(*,*)|\, \big\|$ on κ. To see this we note that

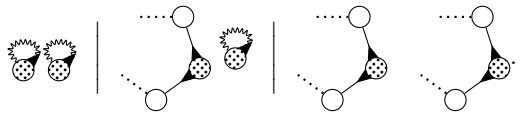
$$\sum_{\alpha_1,\alpha_2} |\kappa(\alpha_1,\alpha_2) A_{\alpha_1} B_{\alpha_2}| \le |||\kappa|||_2^{\mathrm{av}} \sqrt{\sum_{\alpha_1} |A_{\alpha_1}|^2}\sqrt{\sum_{\alpha_2} |B_{\alpha_2}|^2}, \tag{4.32}$$

after which (4.31b) with $A_{\alpha_1}, B_{\alpha_1} \in \{G_{b_1a_1}, (GB)_{b_1a_1}, G_{ca_1}G_{b_1d}, (GB)_{ca_1}G_{b_1d}, G_{ca_1}(GB)_{b_1d}\}$ and arbitrary fixed indices c, d is applicable.

Remark 4.5. *In the sequel we will not write up separate estimates for edges representing GB instead of G as the same Ward estimates* (4.31a)–(4.31b) *hold true and the bound is automatic in the sense that there are in total p wiggly edges in Γ, each of which will contribute a factor of $\|B\|$ to the final estimate, regardless of whether the corresponding edge has been bounded trivially $|(GB)_\alpha| \lesssim \|B\|$ or by* (4.31a)–(4.31b).

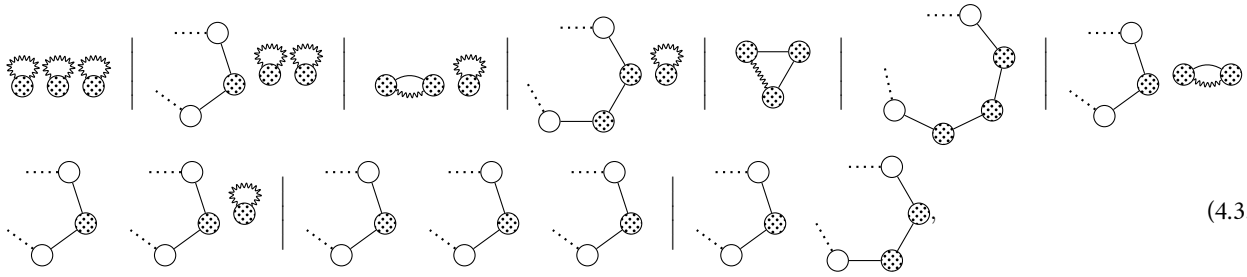
We find that for every edge connected to $\circledcirc$ we can gain a factor ψ compared to the naive size of the $\circledcirc$-sum, using only the trivial bound $|G| \lesssim 1$. We will indicate visually that an edge has potential for a gain of ψ through some colour by putting a mark

(a small arrow) pointing from the vertex towards the edge. Thus in the case where ⊛ appears twice we mark all edges adjacent to ⊛ to obtain the following marked graphs



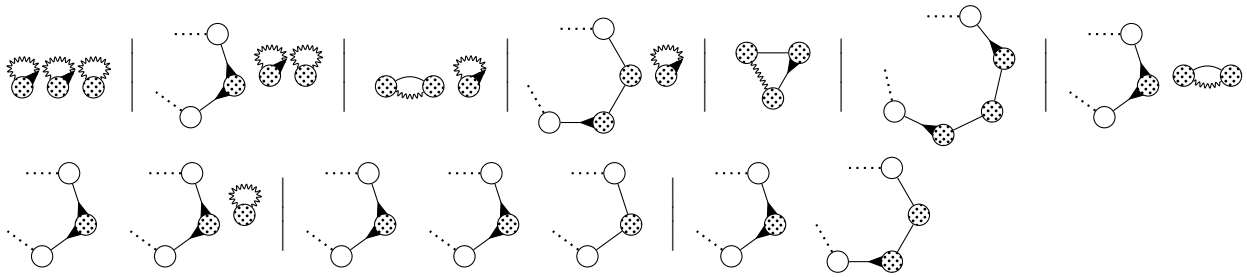
We note that these marks indicate that we can use a Ward estimate for every marked edge, when performing the ⊛-summation, while keeping all other labels fixed. When simultaneously summing over labels from different colours it is not guaranteed any more that we can perform a Ward estimate for every marked edge. We will later resolve this possible issue by introducing the concept of *effective* and *ineffective* marks.

Procedure for colours appearing three times in Γ. If a colour ⊛ appears three times, then the following ten setups are possible



(4.33)

where we explicitly allow components with open continuations to be connected (unlike in the previous case, where rule (d) applied). We now mark the edges adjacent to ⊛ as follows and observe that at most two remain unmarked. Explicitly we choose the markings



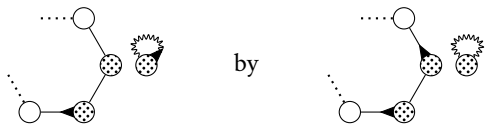
and observe that in all but the fifth graph we can gain a factor of ψ for every marked edge using the first term in the norm $|||\kappa|||_3^{\mathrm{av}}$. For example, in the second graph this follows from

$$\sum_{\alpha_1,\alpha_2,\alpha_3} |\kappa(\alpha_1,\alpha_2,\alpha_3) G_{ca_3} G_{b_3 d} G_{b_1 a_1} G_{b_2 a_2}| \lesssim |||\kappa|||_3^{\mathrm{av}} \sqrt{\sum_{\alpha_2} |G_{b_2 a_2}|^2} \sqrt{\sum_{\alpha_3} |G_{ca_3} G_{b_3 d}|^2} \lesssim |||\kappa|||_3^{\mathrm{av}} N^2 \psi^3$$

and in third graph from

$$\sum_{\alpha_1,\alpha_2,\alpha_3} |\kappa(\alpha_1,\alpha_2,\alpha_3) G_{b_1 a_2} G_{b_2 a_1} G_{b_3 a_3}| \lesssim \sum_{\alpha_2,\alpha_3} |G_{b_3 a_3}| \sum_{\alpha_1} |\kappa(\alpha_1,\alpha_2,\alpha_3)| \lesssim |||\kappa|||_3^{\mathrm{av}} N^2 \psi,$$

where c and d are the connected indices from the white vertices in the graph. The computations for the other graphs are identical. We note that the markings we chose above are not the only ones possible. For example we could have replaced



(4.34)

For the fifth graph in (4.33) the second term in the $|||\kappa|||_3^{\mathrm{av}}$ is necessary. The norms in (2.8c) ensure that we can perform at least one Ward estimate and we have

$$\sum_{\alpha_1,\alpha_2,\alpha_3} |\kappa(\alpha_1,\alpha_2,\alpha_3) G_{b_1 a_2} G_{b_2 a_3} G_{b_3 a_1}| \lesssim |||\kappa|||_3^{\mathrm{av}} N^2 \psi.$$

Indeed, for example

$$\sum_{\alpha_1,\alpha_2,\alpha_3} |\kappa_{cd}(\alpha_1,\alpha_2,\alpha_3) G_{b_1 a_2} G_{b_2 a_3} G_{b_3 a_1}| \lesssim \sum_{\alpha_1,\alpha_2,\alpha_3} |\kappa_{cd}(\alpha_1,\alpha_2,\alpha_3) G_{b_3 a_1}| \le |||\kappa_{cd}|||_{cd} N \sqrt{\sum_{b_3,a_1} |G_{b_3 a_1}|^2} \lesssim |||\kappa_{cd}|||_{cd} N^2 \psi$$

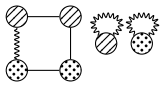
and the other three cases are similar.

Procedure for all other colours in Γ. For colours in $C \setminus C'$, i.e., those which appear four times or more, we do not intend to use any Ward estimates and therefore we do not place any additional markings. Thus we only have to control the size of the summation over any fixed colour, as is guaranteed by the finiteness of $\|\!|\!|\kappa|\!|\!|_k^{\mathrm{av}}$.

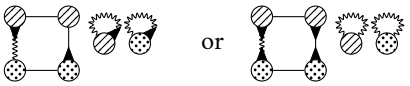
Counting of markings. After we have chosen all markings, we select the "useful" ones. We call an edge *ineffectively marked* if it only carries one mark and joins two distinctly C'-coloured vertices. All other marked edges we call *effectively marked* because the parallel gain through a Ward estimate is guaranteed for all those edges. In total, there are at least $\sum_{c\in C'}|V_c|$ edges adjacent to C' (i.e., adjacent to a C'-coloured vertex). After the above marking procedure there are at most $2\sum_{c\in C'}(|V_c|-2)$ unmarked or ineffectively marked edges adjacent to C'. To see this we note that edges between two C'-colours with only one marking are counted as unmarked from the perspective of exactly one of the two colours. Thus we find that there are at least

$$\sum_{c\in C'}|V_c| - 2\sum_{c\in C'}(|V_c|-2) = \sum_{c\in C'}(4-|V_c|) \tag{4.35}$$

effectively marked edges adjacent to C' after the marking procedure. We illustrate this counting in an example. In the graph



we have $V_{\oslash} = V_{\otimes} = 3$ and there are six edges adjacent to $C' = \{\oslash, \otimes\}$. After the marking procedure we could for example obtain the graphs



where the second graph would result from the replaced marking in (4.34). In both cases there are two effectively marked edges, in accordance with (4.35); in the first example there are also two ineffectively marked edges.

Power counting estimate. The strategy now is that we iteratively perform the Ward estimates colour by colour in C' in no particular order. In each step we thus remove all the edges adjacent to some given colour, either through Ward estimates (if the edge was marked in that colour), or through the trivial bound $|G_\alpha| \lesssim 1$. If some edge is missing because it already was removed in a previous step, then the corresponding G is replaced by 1 in that estimate (e.g. in (4.32)). This might reduce the number of available Ward estimates in some steps, but the concept of effective markings ensures that whenever an effectively marked edge is removed, then a gain through a Ward estimate is guaranteed. After the summation over all colours from C' we have thus performed Ward estimates in all the effectively marked edges, which amounts to at least

$$\sum_{c\in C'}(4-|V_c|)$$

gains of the factor ψ. We note that ineffectively marked edges may not be estimated by a Ward estimates, as it might be necessary to bound the corresponding G trivially while performing the sum over another colour. Using only the gains from the effective marks, we can improve on the naive power counting (4.30) to conclude that the value of Γ is bounded by

$$\mathrm{Val}(\Gamma) \lesssim N^{-p}\prod_{c\in C\setminus C'} N^{2-|V_c|/2}\prod_{c\in C'}(N\psi^2)^{2-|V_c|/2} \le \psi^{2p}N^{2|C\setminus C'|-|V_{C\setminus C'}|/2}, \tag{4.36}$$

where we used that $|C'| \le |C| \le p$, $|V_c| = 2, 3$ for $c \in C'$ and $|V_c| \ge 4$ otherwise, and that $N\psi^2 \ge 1$. Note that this last bound used $\langle z\rangle \le C$ and $\|H\| \le C$, without these conditions we have $N\psi^2 \ge (\|H\|^2 + \langle z\rangle^2)^{-1} \ge \langle\|H\|\rangle^{-2}\langle z\rangle^{-2}$.

4.3.1. **Detailed bound.** The argument above tacitly assumed bounds of the form $|G_\alpha| \lesssim 1$ and $\sum_\alpha |G_\alpha|^2 \lesssim N^2\psi^2$. Apart from unspecified and irrelevant constants, these bounds are not available almost surely, they hold only in the sense of high moments, e.g. $\mathbf{E}\,|G_\alpha|^q \le_q 1$. Secondly, the definition of ψ intentionally left the role of $\Im G$ in it vague. The precise definition of ψ will involve high L^q norms of $\Im G$. Moreover, different G-factors in the monomials Λ are not independent. All these difficulties can be handled by the following general Hölder inequality. Suppose, we aim at estimating

$$\mathbf{E}\sum_A X_A \sum_B Y_{A,B}$$

for random variables X_A, $Y_{A,B}$, then we use the Hölder inequality to estimate

$$\Big\|\sum_A X_A\sum_B Y_{A,B}\Big\|_q \le \Big(\sum_A\Big)^\epsilon \Big\|\sum_A X_A\Big\|_{2q}\max_A\Big\|\sum_B Y_{A,B}\Big\|_{1/\epsilon} \tag{4.37}$$

for $0 < \epsilon \le 1/2q$. In our procedure (4.37) enables us to iteratively bound the graphs colour by colour at the expense of an additional factor $N^{2pR\epsilon}$ in every colour step of the bound, as the total sum is at most of size N^{2pR}. To estimate a G or an $\Im G$ directly we use the Hölder inequality and note that there are at most $|V| = \sum_c |V_c| \le pR$ factors of the form G or GB, so that we can estimate those terms isotropically by $\|G\|_{pR/\epsilon}$, $\|B\|\,\|G\|_{pR/\epsilon}$ and $\|\Im G\|_{pR/\epsilon}$. We use (4.37) at most with $q \in$

$\{1, 2, 4, \dots, 2^{p-1}\}$ and thus have a restriction of $0 < \epsilon \le 2^{-p}$. Thus, combining the power counting above with the iterated application of the Hölder inequality, we have shown that

$$\|\mathrm{Val}(\Gamma)\|_p \le_{p,R,\epsilon} N^{2p^2R\epsilon}\Big(1 + |||\kappa|||^{\mathrm{av}}\Big)^p \|B\|^p \,(1 + \|G\|_{pR/\epsilon}^{|V|})\psi_{pR/\epsilon}^{2p} N^{2|C\setminus C'|-|V_{C\setminus C'}|/2}, \quad \psi_q := \sqrt{\frac{\|\Im G\|_q}{N\eta}} \tag{4.38}$$

for all $\Gamma \in \mathcal{G}^{\mathrm{av}}(p, R)$ and $0 < \epsilon \le 1/2^p$. Therefore, together with (4.29) we conclude the bound

$$\Big\|\mathcal{M}\Big(\big(\langle B\cdot\rangle^{[p/2]}, \overline{\langle B\cdot\rangle}^{[p/2]}\big); L, M, \sigma\Big)\Big\|_p \le_{p,R,\epsilon} N^{2p^2R\epsilon}\Big(1 + |||\kappa|||^{\mathrm{av}}\Big)^p \|B\|^p \,(1 + \|G\|_{pR/\epsilon}^{|V|})\psi_{pR/\epsilon}^{2p} N^{2|C\setminus C'|-|V_{C\setminus C'}|/2} \tag{4.39}$$

on (4.26).

4.4. **Isotropic bound on** D

We turn to the isotropic bound on D, i.e. we give bounds on (4.26) with functionals Λ of the following type. We consider fixed vectors $\mathbf{x}, \mathbf{y}$ and set $\Lambda(D) = D_{\mathbf{xy}}$ or $\Lambda(D) = \overline{D_{\mathbf{xy}}}$. Up to sign we then have

$$|\Lambda_{\alpha_1,\dots,\alpha_n}| = N^{-n/2}(\Delta^{\alpha_1} G \dots \Delta^{\alpha_n} G)_{\mathbf{xy}} = N^{-n/2}\mathbf{x}_{a_1} G_{b_1 a_2} \dots G_{b_{n-1}a_n} G_{b_n \mathbf{y}}. \tag{4.40}$$

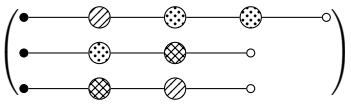
The graph component representing $\Lambda_{\alpha_1,\dots,\alpha_n}$ is a chain in contrast to the cycles in the averaged case. We also have additional *edges* representing the first $\mathbf{x}_{a_1}$ and last $G_{b_n\mathbf{y}}$ factor which we will picture as •——— and ———○, respectively. These are special edges that are adjacent to one vertex only (the dots • and ○ are not considered as vertices). We will call them *initial* and *final* edge. Due to these special edges we should, strictly speaking, talk about a special class of hypergraphs consisting of a union of chains each of them starting and ending with such a special edge, but for simplicity we continue to use the term *graph*. For example we have the correspondence

$$\Lambda_{\alpha_1,\alpha_2} \leftrightarrow \bullet\!\!-\!\!\!-(\alpha_1)-\!\!\!-(\alpha_2)-\!\!\!-\circ. \tag{4.41}$$

For $\Lambda(D) = \overline{D_{\mathbf{xy}}}$ the edges represent $\overline{x_{a_1}}$, $G^*_{b_k\mathbf{y}}$ and $G^*_{b_k a_{k+1}}$ but we do not indicate complex and Hermitian conjugate visually as they have no consequences on the argument. We follow the same convention regarding the colouring, as we did in the averaged case and for example have the representation

$$\sum_{\alpha_1,\beta_1^2(1)} \kappa(\alpha_1,\beta_1^2(1)) \sum_{\alpha_2,\beta_2^1(1),\beta_2^1(2)} \frac{\kappa(\alpha_2,\beta_2^1(1),\beta_2^1(2))}{2!} \sum_{\alpha_3,\beta_3^2(1)} \kappa(\alpha_3,\beta_3^2(1))\Lambda_{\alpha_1,\beta_2^1(2),\beta_2^1(1)}\Lambda_{\alpha_2,\beta_3^2(1)}\Lambda_{\alpha_3,\beta_1^3(1)}$$

$$= \mathrm{Val}$$



We again rephrase the rules on M as rules on the graph Γ. We consider all vertex coloured graphs Γ such that the connected components are chains with an initial edge of type •——— and a final edge of type ———○ such that

(a) there exist p connected components, all of which are chains,
(b) every component contains at least one vertex,
(c) every colour occurs at least once on a vertex adjacent to •——— ,
(d) every colour occurs at least twice,
(e) if a colour occurs exactly twice, then it occurs in two different chains.

The set of graphs satisfying (a)–(e) will be denoted by $\mathcal{G}^{\mathrm{iso}}(p, R)$ and for each L, M, σ in (4.26) we can write the main term $\mathcal{M}$ as

$$\mathcal{M}\Big(\big(\langle \mathbf{x}, \cdot\mathbf{y}\rangle^{[p/2]}, \overline{\langle \mathbf{x}, \cdot\mathbf{y}\rangle}^{[p/2]}\big); L, M, \sigma\Big) = \mathrm{Val}(\Gamma), \qquad \Gamma = \Gamma(L, M, \sigma) \in \mathcal{G}^{\mathrm{iso}(p,R)} \tag{4.42}$$

where $\langle \mathbf{x}, \cdot\mathbf{y}\rangle^{[p/2]}$ denotes the tuple of $p/2$ functionals mapping $D \mapsto \langle \mathbf{x}, D\mathbf{y}\rangle$ and similarly for $\overline{\langle \mathbf{x}, \cdot\mathbf{y}\rangle}$. As the number of such graphs is finite for given p, R it follows that it sufficient to prove the required bound for every single graph.

In contrast to the averaged case, where each Λ carried a factor $1/N$ from the definition of $\Lambda(D) = N^{-1}\operatorname{Tr} BD$, now the naive size of the sum over Γ is not of order 1, but of order

$$\mathrm{Val}(\Gamma) \lesssim \prod_{c\in C} N^{2-|V_c|/2} = N^{2|C|-|V|/2}, \tag{4.43}$$

which can be large. Consequently we have to be more careful in our bound and first make use of a cancellation.

Step 1: Improved naive size. We first observe that we can reduce the naive size (4.43) to order 1, without using any Ward estimates, yet. The improvement comes from the fact that sums of the type

$$\sum_a v_a G_{ab} = G_{\mathbf{v}b}$$

can be directly bounded via the right hand side by $|G_{\mathbf{v}b}| \lesssim \|\mathbf{v}\|$ using the isotropic bound. Note that the naive estimate on the left hand side would be

$$\Big|\sum_a v_a G_{ab}\Big| \lesssim \sum_a |v_a| \le \sqrt{N}\,\|\mathbf{v}\|$$

and even with a Ward estimate it can only be improved to

$$\left|\sum_a v_a G_{ab}\right| \le \|\mathbf{v}\| \sqrt{\sum_a |G_{ab}|^2} \le \sqrt{N\psi}\,\|\mathbf{v}\|$$

So the procedure "summing up a vector $\mathbf{v}$ into the argument of G" is much more efficient than a Ward estimate. The limitation of this idea is that only deterministic vectors $\mathbf{v}$ can be summed up, since isotropic bounds on $G_{\mathbf{uv}}$ hold only for fixed vectors $\mathbf{u}, \mathbf{v}$.

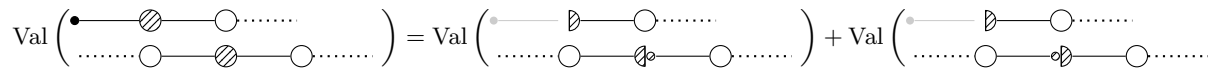
Improvement for colours occurring twice in Γ. For colours which occur exactly twice we can sum up the $\mathbf{x}$ into a G factor without paying the price of an N factor from this summation. To do so, we consider an arbitrary partition of $\kappa = \kappa_c + \kappa_d$, where one should think of that $\kappa_d(\alpha_1, \alpha_2)$ forces $\alpha_1 = (a_1, b_1)$ to be close to $\alpha_2 = (a_2, b_2)$, whereas $\kappa_d(\alpha_1, \alpha_2)$ forces (a_1, b_1) to be close to (b_2, a_2). In both cases we can, according to rule (b), perform two single index summations as follows. First, we sum up the index a_1 of $\mathbf{x}$ as

$$\sum_{a_1} \kappa(a_1 b_1, a_2 b_2) x_{a_1} = \kappa(\mathbf{x} b_1, a_2 b_2).$$

Then we sum up its companion b_2 or a_2, depending on whether we consider the cross or direct term:

$$\sum_{b_2} \kappa_c(\mathbf{x} b_1, a_2 b_2) G_{b_2 \mathbf{v}} = G_{\kappa_c(\mathbf{x} b_1, a_2 \cdot)\mathbf{v}} \quad \text{or} \quad \sum_{a_2} \kappa_d(\mathbf{x} b_1, a_2 b_2) G_{\mathbf{v} a_2} = G_{\mathbf{v}\kappa_d(\mathbf{x} b_1, \cdot b_2)},$$

where $\mathbf{v}$ can be any vector or index. Thus we effectively performed a single label (two index) summation into a single G factor that will be estimated by a constant in the isotropic norm. We indicate this summation graphically by introducing half-vertices and representing the single leftover indices a and b corresponding to a label $\alpha = (a, b)$ and new (half)edges and representing the $G_{\kappa_c(\mathbf{x} b_1, a_2 \cdot)\mathbf{v}}$ and $G_{\mathbf{v}\kappa_d(\mathbf{x} b_1, \cdot b_2)}$ factors. To indicate that the half-edges representing $\mathbf{x}$ have been summed, we grey them out. This partial summation can thus be graphically represented as



since

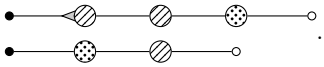
$$\sum_{a_1, b_1, a_2, b_2} \kappa(a_1 b_1, a_2 b_2) \left(\mathbf{x}_{a_1} G_{b_1 \mathbf{u}}\right) \left(G_{\mathbf{v} a_2} G_{b_2 \mathbf{w}}\right) = \sum_{b_1, a_2} \left(G_{b_1 \mathbf{u}}\right) \left(G_{\mathbf{v} a_2} G_{\kappa_c(\mathbf{x} b_1, a_2 \cdot)\mathbf{w}}\right) + \sum_{b_1, b_2} \left(G_{b_1 \mathbf{u}}\right) \left(G_{\mathbf{v}\kappa_d(\mathbf{x} b_1, \cdot b_2)} G_{b_2 \mathbf{w}}\right)$$

where $\mathbf{u}, \mathbf{v}, \mathbf{w}$ are the connecting indices from the white vertices.

Improvement for colours occurring three times in Γ. For colours which appear exactly three times we cannot perform the summation of $\mathbf{x}$ directly. We can, however use a Cauchy-Schwarz in the vertex adjacent to the $\mathbf{x}$–edge to improve the naive size of the –sum to $N^{3/2}$ from N^2. Explicitly, for any index or vector $\mathbf{v}$ we use that

$$\sum_{a_i, b_i} |\mathbf{x}_{a_i} G_{b_i \mathbf{v}}| \lesssim \sum_{a_i, b_i} |\mathbf{x}_{a_i}| \le N^{3/2} \left(\sum_{a_i} |\mathbf{x}_{a_i}|^2\right)^{1/2} = N^{3/2} \|\mathbf{x}\|.$$

To indicate the intend to use the Cauchy-Schwarz improvement on a specific $\mathbf{x}$ edge, we mark the corresponding edge with a marking originating in the adjacent vertex, very much similar to the marking procedure in the averaged case. To differentiate this marking from those indicating the potential for a Ward estimate we use a grey marking . As an example we would indicate



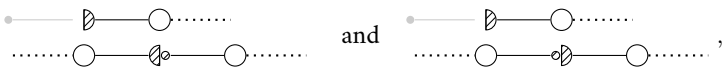
After these two improvements over (4.43) the naive size (naive in the sense without any Ward estimates, yet) of the summed graph is

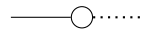
$$\mathrm{Val}(\Gamma) \lesssim \left(\prod_{c \in C, |V_c| = 2} N^{1 - |V_c|/2}\right) \left(\prod_{c \in C, |V_c| = 3} N^{3/2 - |V_c|/2}\right) \left(\prod_{c \in C, |V_c| \ge 4} N^{2 - |V_c|/2}\right) \le 1. \tag{4.44}$$

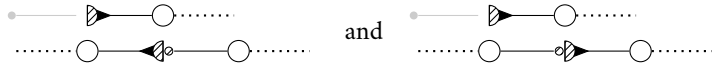
Notice that the first two factors give 1, so the improved power counting for colours with two or three occurrences is neutral. We thus restored the order 1 bound and can now focus on the counting of Ward estimates, with which we can further improve the bound.

Step 2: Further improvements through Ward estimates. The counting procedure is very similar to what we used in the averaged law in the sense that we mark potential edges for Ward estimates colour by colour. To be consistent with the improved naive bound we count the grey initial edges (those from the summation of colours occurring twice) and the initial edges with a grey arrow (those from the summation of colours appearing three times) as unmarked, since they will not be available for Ward estimates.

Marking procedure for colours occurring twice. Colours occurring twice can, after Step 1, only occur in the reduced forms



where we allow  and  to stand for an arbitrary continuation of the graph, as well as the initial  and final edge . In both cases we mark the edges adjacent to the remaining two half-vertices to obtain:



Thus for colours appearing twice we always leave two edges unmarked (which includes the greyed out initial edge). Using the $|||\kappa|||_2^{\text{iso}}$ norm we indeed find that the solid edges in the two graphs above can be bounded by

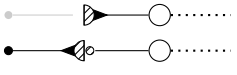
$$\sum_{b_1,a_2} \left|G_{b_1\mathbf{u}} G_{\mathbf{v}a_2} G_{\kappa_c(\mathbf{x}b_1,a_2\cdot)\mathbf{w}}\right| \lesssim \|\mathbf{w}\| \sum_{b_1,a_2} \left|G_{b_1\mathbf{u}} G_{\mathbf{v}a_2}\right| \|\kappa_c(\mathbf{x}b_1,a_2\cdot)\|\| \lesssim N^2\psi^2 \,|||\kappa|||_2^{\text{iso}} \,\|\mathbf{x}\|\,\|\mathbf{u}\|\,\|\mathbf{v}\|\,\|\mathbf{w}\| \tag{4.45}$$

and

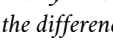
$$\sum_{b_1,b_2} \left|G_{b_1\mathbf{u}} G_{\mathbf{v}\kappa_d(\mathbf{x}b_1,\cdot b_2)} G_{b_2\mathbf{w}}\right| \lesssim \|\mathbf{v}\| \sum_{b_1,b_2} \left|G_{b_1\mathbf{u}} G_{b_2\mathbf{w}}\right| \|\kappa_d(\mathbf{x}b_1,\cdot b_2)\|\| \lesssim N^2\psi^2 \,|||\kappa|||_2^{\text{iso}} \,\|\mathbf{x}\|\,\|\mathbf{u}\|\,\|\mathbf{v}\|\,\|\mathbf{w}\|$$

where the vectors (which are allowed to be indices, as well) $\mathbf{u}, \mathbf{v}, \mathbf{w}$ are the endpoints of the edges in the three white vertices.

Remark 4.6. *In the case that*  *stands for the initial edge*  *, we cannot use the Ward estimate, but use instead that* $\mathbf{x}$ *is a vector of finite norm, providing a gain of* $N^{-1/2} \ll \psi$*. For example, we could bound the graph*

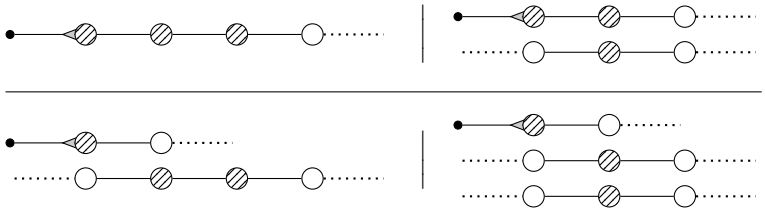
 *by* $\displaystyle\sum_{b_1,a_2} \left|G_{b_1\mathbf{u}} \mathbf{x}_{a_2} G_{\kappa_c(\mathbf{x}b_1,a_2\cdot)\mathbf{w}}\right| \lesssim N^2 \frac{\psi}{\sqrt{N}} \,|||\kappa|||_2^{iso} \,\|\mathbf{x}\|^2 \,\|\mathbf{u}\|\,\|\mathbf{w}\|\,,$

which is better than (4.45) *as* $\sqrt{N}\psi \geq 1$*. In the sequel we will not specifically distinguish this case when instead of a Ward estimate we have to use the finite norm of* $\mathbf{x}$*, as the procedure is identical and the resulting bound is always smaller in the latter case.*

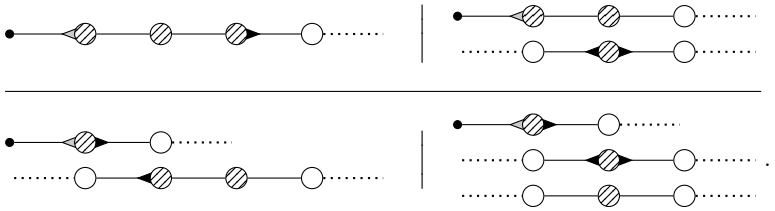
We also will not separately consider the case when  *stands for the final edge* *, as we can use the same Ward estimate as before, with the difference that* $\mathbf{u}$ *and/or* $\mathbf{w}$ *are replaced by* $\mathbf{y}$*.*

We will not manually keep track of the number of $\|\mathbf{x}\|, \|\mathbf{y}\|$ *in the bound as it is automatic in the sense that there are* p *initial and* p *final edges in* Γ*, each contributing a factor of* $\|\mathbf{x}\|, \|\mathbf{y}\|$ *to the final estimate.*

Marking procedure for colours occurring three times. Colours appearing three times occur in one of the following four forms

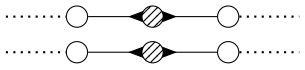


and in all cases we mark the edges adjacent to in such a way that at most three edges (including the initial edge with the grey mark) remain unmarked. Indeed, we mark the edges as follows.

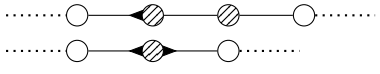


Very similar to the bound using $|||\kappa|||_3^{\text{av}}$, we find that using the norm $|||\kappa|||_3^{\text{iso}}$ we can perform Ward estimates on all marked edges.

Marking procedure for colours occurring more than three times. For any colour c occurring more than three times we claim that we can always mark edges in such a way that at most $2|V_c| - 4$ edges adjacent to V_c remain unmarked. Indeed, if we call an edge connected to two c–coloured vertices *c–internal* and denote their set E_c^{int}, then there are $2|V_c| - |E_c^{\text{int}}|$ edges adjacent to c. Out of this set of all c-adjacent edges, we mark any two and thus the claim is trivially fulfilled if $|E_c^{\text{int}}| \geq 2$. If $|E_c^{\text{int}}| = 0$, then the graph contains two single vertices of colour c, for which we mark all four adjacent edges, i.e.



also confirming the claim in this case. Finally, if $|E_c^{\text{int}}| = 1$, then the graph has to contain



for which we mark the three indicated edges, confirming the claim also in this final case. We note (again similarly to $|||\kappa|||_3^{\text{av}}$) that the norm $|||\kappa|||_k^{\text{iso}}$ allows to perform Ward estimates on all marked edges.

Counting of markings. In contrast to the averaged case, we now call an edge *ineffectively marked* if it only carries one mark and connects any two distinctly coloured vertices (in the averaged case the analogous definition was restricted to C'-coloured vertices). All other marked edges we call *effectively marked*. In particular the initial and final edge are always effectively marked, once they are marked. By construction, all effectively marked edges can be summed up by Ward estimates. In total, there are exactly $p + \sum_{c\in C} |V_c|$ edges in Γ. After the marking procedure there are at most

$$\sum_{c\in C,|V_c|=2} 2 + \sum_{c\in C,|V_c|=3} 3 + \sum_{c\in C,|V_c|\geq 4} (2\,|V_c| - 4)$$

unmarked or ineffectively marked edges in Γ. Thus there are at least

$$\Big(p + \sum_{c\in C} |V_c|\Big) - \Big(\sum_{c\in C,|V_c|=2} 2 + \sum_{c\in C,|V_c|=3} 3 + \sum_{c\in C,|V_c|\geq 4} (2\,|V_c| - 4)\Big) = p + \sum_{c\in C,|V_c|\geq 4} (4 - |V_c|) \tag{4.46}$$

effectively marked edges in Γ, which can be negative, but it turns out that in this case the (improved) naive size already is sufficiently small.

Power counting estimate. The strategy for performing the Ward estimates is identical to that in the averaged case; we perform them colour by colour in an arbitrary order. According to the improved naive bound from Step 1, and recalling that the power counting for $|V_c| = 2$ and $|V_c| = 3$ gives 1, i.e. is neutral, and the counting of additional effective markings we find that the summed value of Γ is bounded by

$$\mathrm{Val}(\Gamma) \lesssim N^{2|C\setminus C'|-|V_{C\setminus C'}|/2} \psi^{(p+4|C\setminus C'|-|V_{C\setminus C'}|)_+},$$

where C' are those colours c with $|V_c| = 2, 3$.

Detailed estimate. Finally, this power counting is performed with the procedure of iterated Hölder inequalities, exactly as in the averaged case to obtain

$$\|\mathrm{Val}(\Gamma)\|_p \leq_{\epsilon,R,p} N^{2p^2R\epsilon}\Big(1 + |||\kappa|||^{\mathrm{iso}}\Big)^p \|\mathbf{x}\|^p \|\mathbf{y}\|^p (1 + \|G\|^{|V|}_{pR/\epsilon})\psi^{(p+4|C\setminus C'|-|V_{C\setminus C'}|)_+}_{pR/\epsilon} N^{2|C\setminus C'|-|V_{C\setminus C'}|/2} \tag{4.47}$$

for all $\Gamma \in \mathcal{G}^{\mathrm{iso}}(p,R)$ and $0 < \epsilon \leq 1/2^p$. Therefore we conclude together with (4.42) that

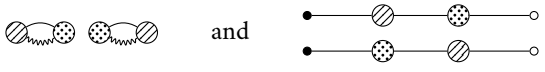
$$\begin{aligned}\Big\|\mathcal{M}\Big(\big(\langle \mathbf{x},\cdot\mathbf{y}\rangle^{[p/2]}, \overline{\langle \mathbf{x},\cdot\mathbf{y}\rangle}^{[p/2]}\big); L, M, \sigma\Big)\Big\|_p \leq_{\epsilon,R,p} N^{2p^2R\epsilon}\Big(1 + |||\kappa|||^{\mathrm{iso}}\Big)^p \|\mathbf{x}\|^p \|\mathbf{y}\|^p \\ \times (1 + \|G\|^{|V|}_{pR/\epsilon})\psi^{(p+4|C\setminus C'|-|V_{C\setminus C'}|)_+}_{pR/\epsilon} N^{2|C\setminus C'|-|V_{C\setminus C'}|/2}.\end{aligned} \tag{4.48}$$

4.5. **Modifications for general case**

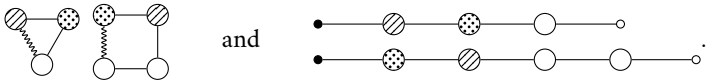
In the previous Sections 4.3 and 4.4 we estimated $\mathcal{M}$ defined in (4.26) under the simplifying assumptions $L_3 = L_4 = \emptyset$ and $\mathcal{N}(\alpha_l) = I$. For the final bound in (4.25) we need to treat all other cases. In this section we now demonstrate that these simplifying assumptions are not substantial and that the results from (4.35) and (4.46) on the number of available Ward estimates remain valid in the more general setting. By definition, $\mathcal{M}$ depends on the labels of types L_3 and L_4, which are considered fixed in the subsequent discussion. The graphs we introduced to systematically bound $\mathcal{M}$ do not change in their form for the general case, but only have additional *fixed* vertices α_l, β_l for $l \in L_3 \cup L_4$, which we consider as uncoloured. Thus we enlarge the set graphs $\mathcal{G}^{\mathrm{av}}$ and $\mathcal{G}^{\mathrm{iso}}$ to $\widetilde{\mathcal{G}}^{\mathrm{av}}$ and $\widetilde{\mathcal{G}}^{\mathrm{iso}}$, which are defined by the previously stated rules (a)-(e) with the addition of

(f) certain vertices may be uncoloured.

These uncoloured vertices represent exactly those labels of types L_3 and L_4, which are parameters of $\mathcal{M}$, as defined in (4.18). For example, the previously studied graphs

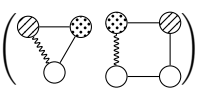


can be extended to



The definition of the value naturally extends to these larger classes of graphs, but without a summation over the uncoloured vertices. In the above example (4.28) is then replaced by

$$\sum_{\alpha_1,\beta_1^2(1)} \kappa(\alpha_1,\beta_1^2(1)) \sum_{\alpha_2,\beta_2^1(1)} \kappa(\alpha_2,\beta_2^1(1))\Lambda_{\alpha_1,\beta_2^1(1),\gamma(1)}\Lambda_{\alpha_2,\beta_1^2(1),\gamma(2),\gamma(3)} = \mathrm{Val}$$



where $\gamma(1),\gamma(2),\gamma(3)$ are the fixed labels and the value of the graph depends on them. The isotropic case is analogous.

The argument in Sections 4.3 and 4.4, however, only concern those labels which are actually summed over, i.e., those of type l for $l \in L_2$. In other words, we only aim at improving the L_2-summation by Ward estimates. The presence of additional fixed labels do neither change the naive bounds, the improvement through Ward estimates, nor the counting of those Ward estimates.

Next, the restricted summations due to the neighbourhood sets $\mathcal{N}(\alpha) \subset I$ do also not change the argument. In fact, Ward estimates stay true for restricted summations since

$$\sum_{a\in\mathcal{J}} |G_{a\mathbf{x}}|^2 \le \sum_{a\in J} |G_{a\mathbf{x}}|^2 = \frac{\Im G_{\mathbf{xx}}}{\eta}$$

for arbitrary $\mathcal{J} \subset J$. Also the procedure for improving the naive size in Section 4.4 holds true if only summed over subsets, i.e.,

$$\sum_{a_1\in\mathcal{J}} \kappa(a_1b_1, a_2b_2)x_{a_1} = \kappa(\mathbf{x}_{\mathcal{J}}b_1, a_2b_2),$$

where the sub-vector $\mathbf{x}_{\mathcal{J}}$ has bounded norm $\|\mathbf{x}_{\mathcal{J}}\| \le \|\mathbf{x}\|$.

Finally, the modification of $\mathcal{M}$ by setting $W_{\mathcal{N}_{L_3}} = 0$ also does not change the substance of the argument as the bound verbatim also covers this modified W, and the final bounds can be rephrased in terms of $\|G\|$ as $\|\widehat{G}\|_q \le_q 1 + \|G\|^{C/\mu}_{Cq/\mu}$, as demonstrated in Lemma D.3.

4.6. **Proof of Theorem 4.1**

We now have all the ingredients to complete the proof of Theorem 4.1 starting from (4.25), where we recall that $\mathcal{M}$ was defined in (4.18).

Proof of the averaged bound. We recall from (4.30) that for the averaged bound the naive size of $\mathcal{M}$ is given by

$$\mathcal{M} \lesssim N^{-p}N^{-|L|/2-M_L/2}N^{2|L_2|},$$

where the first factor comes from the normalized trace, the second from the derivatives and the third from the L_2 summations. We demonstrated in Section 4.3 (see (4.36) and the counting estimate (4.35)) that through Ward estimates we can improve the naive size $N^{2|L_2|}$ of the L_2 summation to

$$\begin{aligned}\mathcal{M} &\lesssim N^{-p}N^{-|L_3\sqcup L_4|/2-M_{L_3\sqcup L_4}/2}\prod_{\substack{l\in L_2\\ M_l\ge 3}} N^{3/2-M_l/2}\prod_{\substack{l\in L_2\\ M_l\le 2}} (N\psi^2)^{3/2-M_l/2}\\ &\le N^{-|L_1|}N^{-3|L_3|/2-M_{L_3}/2}N^{-2|L_4|}\psi^{2|L_2|}\prod_{\substack{l\in L_2\\ M_l\ge 3}} N^{(3-M_l)/2},\end{aligned}$$

where we used that $N\psi^2 \ge 1$ and that $M_{L_4} = |L_4|$ and we recall that $p = |L_1| + |L_2| + |L_3| + |L_4|$. Consequently we have from (4.25) that

$$\begin{aligned}\mathbf{E}\,|\langle BD\rangle|^p &\lesssim_{p,\mu} N^{-p} + \sum_{\sqcup L_i=[p]} N^{-|L_1|}\psi^{2|L_2|}\Bigg[\prod_{\substack{l\in L_2\\ M_l\ge 3}} N^{(3-M_l)/2}\Bigg]N^{-|L_3|-\mu M_{L_3}}N^{-|L_4|}, \qquad (4.49)\\ &\lesssim_{p,\mu} N^{-p} + \psi^{2p}\sum_{\sqcup L_i=[p]}\Bigg[\prod_{\substack{l\in L_2\\ M_l\ge 3}} N^{(3-M_l)/2}\Bigg]N^{-\mu M_{L_3}} \lesssim_{p,\mu} \psi^{2p}\sum_{\sqcup L_i=[p]} N^{-\frac{1}{2}(M_{L_2}-3p)_+-\mu M_{L_3}},\end{aligned}$$

where we bounded the L_3-summation in (4.25) by $N^{2|L_3|}(N^{1/2-\mu})^{M_{L_3}} = N^{2|L_3|+M_{L_3}/2}N^{-\mu M_{L_3}}$ in the first line, and used $N^{-1} \le \psi^2$ in the second. To conclude the moment bound (4.1b) from (4.49) we have to count the number of $\|G\|_q$'s just as in the proof of (4.39). The key point is to collect enough $N^{-\mu}$ factors so that all but maybe $O(p)$ factors $\|G\|_q$ could be compensated by an $N^{-\mu}$. Since all $|L_i|$ and $M_{L_4} = |L_4|$ are of order p, the only way of collecting more than Cp factors of $\|G\|_q$ is having M_{L_2} or M_{L_3} bigger than a constant times p. But in this case we collect the same order of factors of the type $N^{-1/2}$ or $N^{-\mu}$ from (4.49) and the claim follows since $N^{-1/2} \le N^{-\mu}$.

Proof of the isotropic bound. We recall from (4.44) that for the isotropic law the improved naive size of $\mathcal{M}$ is given by

$$\mathcal{M} \lesssim N^{-|L_3\sqcup L_4|/2-M_{L_3\sqcup L_4}/2}\prod_{\substack{l\in L_2\\ M_l\ge 3}} N^{3/2-M_l/2}$$

and from (4.46) that we can always perform at least $\big(p+\sum_{l\in L_2,\,M_l\ge 3}(3-M_l)\big)_+$ Ward estimates. Consequently, with Proposition 4.4 and (4.25) we obtain

$$\mathbf{E}\,|D_{\mathbf{xy}}|^p \lesssim_{p,\mu} N^{-p} + \sum_{\sqcup L_i=[p]} N^{-\mu M_{L_3}}\psi^{\left(p+\sum_{l\in L_2,\,M_l\ge 3}(3-M_l)\right)_+}\prod_{\substack{l\in L_2\\ M_l\ge 3}} N^{3/2-M_l/2}, \qquad (4.50)$$

where we again bounded the L_3 summation in (4.25) by $N^{2|L_3|+M_{L_3}/2}N^{-\mu M_{L_3}}$. The rhs. of (4.50) is bounded by ψ^p since every missing ψ power is compensated by an $N^{-1/2} \ll \psi$. To conclude the moment bound (4.1a) from (4.50) we again have to count the number of $\|G\|_q$-factors as in the proof of (4.48). This is very similar to the averaged case above and completes the proof of Theorem 4.1.

Modifications for large $|z|$. Our proof so far assumed $\langle z\rangle \lesssim 1$ and $\|H\| \lesssim 1$. The general case follows exactly along the same lines but carrying additional $\langle z\rangle$-factors. The condition $\|H\| \lesssim 1$ is relaxed to $\|H\| \lesssim N^\epsilon$ for any fixed $\epsilon > 0$ with very high probability. This upper bound directly follows by applying Chebysev's inequality to the moment bound $\mathbf{E}\langle H^k\rangle \le C_k$, for any $k \in \mathbb{N}$, which is obtained by a cumulant expansion using the assumed decays of the cumulants. We therefore redefine $\psi := N^\epsilon \langle z\rangle \sqrt{\Im G/N\eta}$ and with this definition $N\psi^2 \gtrsim 1$, used in (4.38) and (4.47) still holds. The other two key bounds we used, $|G_\alpha| \le 1$ and $\sum_\alpha |G_\alpha|^2 \lesssim N^2\psi^2$, naturally change to $|G_\alpha| \lesssim N^\epsilon/\langle z\rangle$ and $\sum_\alpha |G_\alpha|^2 \lesssim N^2\psi^2/\langle z\rangle^2$. Ignoring the irrelevant N^ϵ factors, these obvious bounds express the correct scaling of G in the large z regime, thus every G factor naturally comes with a $1/\langle z\rangle$-factor on top of any previous bounds we used so far. Since each D carries one G-factor, and for the main term of the cumulant expansion of $\mathbf{E}|\Lambda(D)|^p$ the number of G-factors does not decrease, we obtain at least an additional $\langle z\rangle^{-p}$-factor. For the error term of the cumulant expansion the $\langle z\rangle^{-p}$ decay in (4.17) follows directly from the corresponding decays in Lemma D.3. Combining this with the redefined ψ and adjusting the ϵ-exponent, we obtain Theorem 4.1 in the general case.

5. Proof of the stability of the MDE and proof of the local law

Before going into the proof of Theorem 2.2, we collect some facts from [1, 5, 29] about the deterministic MDE (2.1) and its solution.

Proposition 5.1 (Stability of MDE and properties of the solution)**.** *The following hold true under Assumption (A).*

(i) The MDE (2.1) has a unique solution $M = M(z)$ for all $z \in \mathbb{H}$ and moreover the map $z \mapsto M(z)$ is holomorphic.

(ii) The holomorphic function $\langle M\rangle : \mathbb{H} \to \mathbb{H}$ is the Stieltjes transform of a probability measure μ on $\mathbb{R}$.

(iii) There exists a constant $c > 0$ such that we have the bounds

$$\frac{c}{\langle z\rangle + |||\mathcal{S}||| \operatorname{dist}(z, \operatorname{supp}\mu)^{-1}} \le \|M(z)\| \le \frac{1}{\operatorname{dist}(z,\operatorname{supp}\mu)} \quad \text{and} \quad \|\Im M\| \le \frac{\eta}{\operatorname{dist}(z,\operatorname{supp}\mu)^2},$$

where we recall the definition of $|||\mathcal{S}|||$ in (4.2).

(iv) There exist constants $c, C > 0$ such that

$$\big\|(1 - \mathcal{C}_{M(z)}\mathcal{S})^{-1}\big\|_{hs\to hs} \le c\Big[\frac{\langle z\rangle}{\operatorname{dist}(z,\operatorname{supp}\mu)} + \frac{|||\mathcal{S}|||}{\operatorname{dist}(z,\operatorname{supp}\mu)^2}\Big]^C,$$

where $\mathcal{C}$ is the sandwiching *operator $\mathcal{C}_R[T] := RTR$. The norm on the lhs. is the operator norm where $1 - \mathcal{C}_M\mathcal{S}$ is viewed as a linear map on the space of matrices equipped with the Hilbert-Schmidt norm.*

If, in addition, Assumption (E) is also satisfied, then the following statements hold true, as well.

(v) The measure μ from (ii) is absolutely continuous with respect to the Lebesgue measure and has a continuous density $\rho\colon \mathbb{R} \to [0,\infty)$, called the self-consistent density of states, which is also real analytic on the open set $\{\rho > 0\}$.

(vi) There exist constants $c, C > 0$ such that we have the bounds

$$\frac{c}{\langle z\rangle} \le \|M(z)\| \le \frac{C}{\rho(z) + \operatorname{dist}(z,\operatorname{supp}\rho)} \quad \text{and} \quad c\,\rho(z) \le \Im M(z) \le C\langle z\rangle^2 \|M(z)\|^2 \rho(z)$$

in terms of the harmonic extension $\rho(z) := \pi^{-1}\Im\langle M(z)\rangle$ of the self-consistent density of states to the upper half plane $\mathbb{H}$.

(vii) There exist constants $c, C > 0$ such that

$$\big\|(1 - \mathcal{C}_{M(z)}\mathcal{S})^{-1}\big\|_{hs\to hs} \le c\left(1 + \big[\rho(z) + \operatorname{dist}(z,\operatorname{supp}\rho)\big]^{-C}\right).$$

Proof. Parts (i)–(ii) follow from [29, Thm. 2.1]. Parts (iii)–(iv) follow from [5, Section 3] and $\|M\| \ge \|M^{-1}\|^{-1}$. Finally, parts (v)–(vii) follow from [1, Prop. 2.2, 4.2, 4.4]. □

Due to Assumption (C), (4.2) and (5.7) below we have $|||\mathcal{S}||| \le C$. Therefore parts (iii),(iv),(vi) and (vii) show that we have

$$\langle z\rangle\,\|M(z)\| \le_\epsilon N^\epsilon \quad \text{and} \quad \big\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\big\|_{\mathrm{hs}\to\mathrm{hs}} \le_\epsilon N^\epsilon \qquad \text{in} \quad \mathbb{D}^\delta_{\mathrm{out}} \quad \text{and also in} \quad \mathbb{D}^\delta_0 \quad \text{under Assump. (E)} \tag{5.1}$$

for some $\delta = \delta(\epsilon) > 0$. Similarly to (5.1), we will often state estimates that hold both in the spectral domain $\mathbb{D}^\delta_{\mathrm{out}}$ without Assumption (E) as well as in the spectral domain $\mathbb{D}^\delta_\gamma$ but under Assumption (E). We recall that according to our convention about $\le_\epsilon$, (5.1) implies the existence of a constant $C(\epsilon)$ such that the inequalities hold true with that constant for all z in the given ϵ-dependent domains.

5.1. Definition of an isotropic norm suitable for the stability analysis

For a fixed $z \in \mathbb{H}$ define the map

$$\mathcal{J}_z[G, D] := 1 + (z - A + \mathcal{S}[G])G - D$$

on arbitrary matrices G and D. From the definition of $D = D(z)$ (2.4) and the solution $M = M(z)$ of the MDE (2.1) it follows that $\mathcal{J}_z[M(z), 0] = 0$ and $\mathcal{J}_z[G(z), D(z)] = 0$. Throughout this discussion we will fix z and we omit it from the notation, i.e. $\mathcal{J} = \mathcal{J}_z$. We will consider G as a function $G = G(D)$ of an arbitrary error matrix D satisfying $\mathcal{J}[G(D), D] = 0$. Via the implicit function theorem, this relation defines a unique function $G(D)$ for sufficiently small D and $G(D)$ will be analytic as long as $\mathcal{J}$ is stable. The stability will be formulated in a specific norm that takes into account that the smallness of D can only be established

in isotropic sense, i.e. in the sense of high moment bound on $D_{\mathbf{xy}}$ for any fixed deterministic vectors $\mathbf{x}, \mathbf{y}$. To define this special norm, we fix vectors $\mathbf{x}, \mathbf{y}$ and define sets of vectors containing the standard basis vectors $e_a, a \in J$, recursively by

$$I_0 := \{ \mathbf{x}, \mathbf{y} \} \cup \{ e_a \mid a \in J \}, \qquad I_{k+1} := I_k \cup \{ M\mathbf{u} \mid \mathbf{u} \in I_k \} \cup \{ \kappa_c((M\mathbf{u})a, b\cdot), \kappa_d((M\mathbf{u})a, \cdot b) \mid \mathbf{u} \in I_k,\ a, b \in J \},$$

which give rise to the norm

$$\|G\|_* = \|G\|_*^{K,\mathbf{x},\mathbf{y}} := \sum_{0 \le k < K} N^{-k/2K} \|G\|_{I_k} + N^{-1/2} \max_{\mathbf{u} \in I_K} \frac{\|G_{\cdot\mathbf{u}}\|}{\|\mathbf{u}\|}, \qquad \|G\|_I := \max_{\mathbf{u},\mathbf{v} \in I} \frac{|G_{\mathbf{uv}}|}{\|\mathbf{u}\| \|\mathbf{v}\|},$$

where we will choose K later.

Theorem 5.2. *Let $K \in \mathbb{N}$, $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$, and denote the open ball of radius δ around M in $(\mathbb{C}^{N\times N}, \|\cdot\|_*^{K,\mathbf{x},\mathbf{y}})$ by $B_\delta(M)$. Then for*

$$\epsilon_1 := \frac{\left[1 + |||\mathcal{S}||| \, \|M\|^2 + |||\mathcal{S}|||^2 \, \|M\|^4 \left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{hs\to hs}\right]^{-2}}{10 N^{1/K} \|M\|^2 |||\mathcal{S}|||}, \qquad \epsilon_2 := \sqrt{\frac{\epsilon_1}{10 |||\mathcal{S}|||}} \tag{5.2}$$

there exists a unique function $G\colon B_{\epsilon_1}(0) \to B_{\epsilon_2}(M)$ with $G(0) = M$ that satisfies $\mathcal{J}[G(D), D] = 0$. Moreover, the function G is analytic and satisfies

$$\|G(D_1) - G(D_2)\|_* \le 10 N^{1/2K} \left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{*\to*} \|M\| \, \|D_1 - D_2\|_* . \tag{5.3}$$

for any $D_1, D_2 \in B_{\epsilon_1}(0)$.

Proof. First, we rewrite the equation $\mathcal{J}[G, D] = 0$ in the form $\widetilde{\mathcal{J}}[V, D] = 0$, where

$$\widetilde{\mathcal{J}}[V, D] := (1 - \mathcal{C}_M\mathcal{S})V - M\mathcal{S}[V]V + MD, \qquad V := G - M$$

and for arbitrary V and D we claim the bounds

$$\|M\mathcal{S}[V]V\|_* \le N^{1/2K} |||\mathcal{S}||| \, \|M\| \, \|V\|_*^2 , \tag{5.4a}$$

$$\|MD\|_* \le N^{1/2K} \|M\| \, \|D\|_* , \tag{5.4b}$$

$$\left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{*\to*} \le 1 + |||\mathcal{S}||| \, \|M\|^2 + |||\mathcal{S}|||^2 \, \|M\|^4 \left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{\mathrm{hs}\to\mathrm{hs}} . \tag{5.4c}$$

We start with the proof of (5.4a). Let $\kappa = \kappa_c + \kappa_d$ be an arbitrary partition which induces a partition of $\mathcal{S} = \mathcal{S}_c + \mathcal{S}_d$ (as in Remark 4.2). Then for $\mathbf{u}, \mathbf{v} \in I_k$ we compute

$$\frac{|(M\mathcal{S}_c[V]V)_{\mathbf{uv}}|}{\|\mathbf{u}\| \|\mathbf{v}\|} \le \frac{1}{N} \sum_{a,b} \frac{\left|V_{ab} V_{\kappa_c((M\mathbf{u})a,b\cdot)\mathbf{v}}\right|}{\|\mathbf{u}\| \|\mathbf{v}\|} \le |||\kappa_c|||_c \, \|V\|_{\max} \, \|M\| \min\left\{ \|V\|_{I_{k+1}} , \frac{\|V_{\cdot\mathbf{v}}\|}{\|\mathbf{v}\|} \right\}, \tag{5.5a}$$

$$\frac{|(M\mathcal{S}_d[V]V)_{\mathbf{uv}}|}{\|\mathbf{u}\| \|\mathbf{v}\|} \le \frac{1}{N} \sum_{a,b} \frac{\left|V_{a\kappa_d((M\mathbf{u})a,\cdot b)} V_{b\mathbf{v}}\right|}{\|\mathbf{u}\| \|\mathbf{v}\|} \le |||\kappa_d|||_d \, \|M\| \min\left\{ \|V\|_{I_{k+1}} \frac{\|V_{\cdot\mathbf{v}}\|}{\sqrt{N}\|\mathbf{v}\|}, \|V\|_{\max} \frac{\|V_{\cdot\mathbf{v}}\|}{\|\mathbf{v}\|} \right\}, \tag{5.5b}$$

where we used $|V_{a\mathbf{w}}| \le \sqrt{N} \|V\|_{\max} \|\mathbf{w}\|$ in the second bound of (5.5b), so that

$$\|M\mathcal{S}_e[V]V\|_* = \sum_{0 \le k < K} \frac{\|M\mathcal{S}_e[V]V\|_{I_k}}{N^{k/2K}} + \max_{\mathbf{u} \in I_K} \frac{\|(M\mathcal{S}_e[V]V)_{\cdot\mathbf{u}}\|}{\sqrt{N}\|\mathbf{u}\|} \le N^{1/2K} |||\kappa_e|||_e \, \|M\| \, \|V\|_*^2$$

for $e \in \{c, d\}$ and (5.4a) follows immediately, recalling (4.2). We continue with the proof of (5.4b), which follows from the fact that for $\mathbf{u}, \mathbf{v} \in I_k$ we have

$$\frac{|(MD)_{\mathbf{uv}}|}{\|\mathbf{u}\| \|\mathbf{v}\|} \le \|M\| \min\left\{ \|D\|_{I_{k+1}} , \frac{\|D_{\cdot\mathbf{v}}\|}{\|\mathbf{v}\|} \right\}.$$

Finally, we show (5.4c). We use a three term geometric expansion to obtain

$$\begin{aligned} \left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{*\to*} &\le 1 + \|\mathcal{C}_M\mathcal{S}\|_{*\to*} + \|\mathcal{C}_M\mathcal{S}\|_{*\to\mathrm{hs}} \left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{\mathrm{hs}\to\mathrm{hs}} \|\mathcal{C}_M\mathcal{S}\|_{\mathrm{hs}\to*} \\ &\le 1 + \|M\|^2 \|\mathcal{S}\|_{\max\to\|\cdot\|} + \|M\|^4 \|\mathcal{S}\|_{\max\to\|\cdot\|} \left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{\mathrm{hs}\to\mathrm{hs}} \|\mathcal{S}\|_{\mathrm{hs}\to\|\cdot\|} \end{aligned} \tag{5.6}$$

and it only remains to derive bounds on $\|\mathcal{S}\|_{\max\to\|\cdot\|}$ and $\|\mathcal{S}\|_{\mathrm{hs}\to\|\cdot\|}$. We begin to compute for the cross part κ_c and arbitrary normalized vectors $\mathbf{v}, \mathbf{u} \in \mathbb{C}^N$ that

$$|\mathcal{S}_c[V]_{\mathbf{vu}}| = \left| \frac{1}{N} \sum_{b_1, a_2} \langle \kappa_c(\mathbf{v}b_1, a_2\cdot), \mathbf{u} \rangle \, V_{b_1 a_2} \right| \le \frac{\|V\|_{\max}}{N} \sum_{b_1, a_2} \|\kappa_c(\mathbf{v}b_1, a_2\cdot)\| \le |||\kappa_c|||_c \, \|V\|_{\max} ,$$

and

$$\begin{aligned} |\mathcal{S}_c[V]_{\mathbf{vu}}| &= \left| \frac{1}{N} \sum_{a_1, a_2, b_2} v_{a_1} \langle \kappa_c(a_1\cdot, a_2 b_2), V_{\cdot a_2} \rangle \, u_{b_2} \right| \le \frac{1}{N} \sum_{a_1, a_2, b_2} |v_{a_1}| \, \|\kappa_c(a_1\cdot, a_2 b_2)\| \, |u_{b_2}| \, \|V_{\cdot a_2}\| \le \frac{|||\kappa_c|||_c}{N} \sum_{a_2} \|V_{\cdot a_2}\| \\ &\le |||\kappa_c|||_c \sqrt{\frac{1}{N} \sum_{b_1, a_2} |V_{b_1 a_2}|^2} = |||\kappa_c|||_c \, \|V\|_{\mathrm{hs}} . \end{aligned}$$

Next, we estimate for the direct part κ_d that

$$|\mathcal{S}_d[V]_{\mathbf{vu}}| = \Big|\frac{1}{N}\sum_{b_1,b_2}\langle\kappa_d(\mathbf{v}b_1,\cdot b_2),V_{b_1\cdot}\rangle\,u_{b_2}\Big| \le \frac{1}{N}\sum_{b_1,b_2}\|V_{b_1\cdot}\|\,\|\kappa_d(\mathbf{v}b_1,\cdot b_2)\|\,|u_{b_2}| \le \frac{|||\kappa_d|||_d}{N}\sqrt{\sum_{b_1}\|V_{b_1\cdot}\|^2}$$
$$\le \frac{|||\kappa_d|||_d}{N}\sqrt{\sum_{b_1,a_2}|V_{b_1a_2}|^2} \le |||\kappa_d|||_d\min\left\{\frac{\|V\|_{\mathrm{hs}}}{\sqrt{N}},\|V\|_{\max}\right\},$$

so that it follows that, using (4.2),

$$\|\mathcal{S}[V]\| = \sup_{\|\mathbf{v}\|,\|\mathbf{u}\|\le1}|\mathcal{S}[V]_{\mathbf{vu}}| \le |||\mathcal{S}|||\min\left\{\|V\|_{\mathrm{hs}},\|V\|_{\max}\right\},\quad \max\left\{\|S\|_{\max\to\|\cdot\|},\|S\|_{\mathrm{hs}\to\|\cdot\|}\right\} \le |||\mathcal{S}||| \tag{5.7}$$

and therefore (5.4c) follows from (5.6) with (5.7). Now the statement (5.3) follows from the implicit function theorem as formulated in Lemma D.1 applied to the equation $\widetilde{\mathcal{J}}[G-M,D]=0$ written in the form

$$(1-\mathcal{C}_M\mathcal{S})V - M\mathcal{S}[V]V = -MD$$

with $A = 1-\mathcal{C}_M\mathcal{S}$, $B=M$ and $d=D$ in the notation of Lemma D.1. □

This general stability result will be used in the following form

$$\|G-M\|_* \le_\epsilon N^{\epsilon+1/2K}\frac{\|D\|_*}{\langle z\rangle} \qquad \text{in} \quad \mathbb{D}^\delta_{\mathrm{out}} \quad \text{and in} \quad \mathbb{D}^\delta_0 \quad \text{under Assump. (E)} \tag{5.8}$$

for some $\delta=\delta(\epsilon)>0$, as long as $\|D\|_* \le N^{-1/2K}\langle z\rangle^2$ by applying it to $D_1=0, D_2=D(z)$ and using (5.1) and (5.4c).

5.2. Stochastic domination and relation to high moment bounds

In order to keep the notation compact, we now introduce a commonly used (see, e.g., [13]) notion of high-probability bound.

Definition 5.3 (Stochastic Domination). *If*

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

are families of 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

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

for large enough $N\ge N_0(\epsilon,D)$. In this case we use the notation $X\prec Y$.

It can be checked (see [13, Lemma 4.4]) that $\prec$ satisfies the usual arithmetic properties, e.g. if $X_1\prec Y_1$ and $X_2\prec Y_2$, then also $X_1+X_2\prec Y_1+Y_2$ and $X_1X_2\prec Y_1Y_2$. We will say that a (sequence of) events $A=A^{(N)}$ holds with *overwhelming probability* if $\mathbf{P}(A^{(N)})\ge 1-N^{-D}$ for any $D>0$ and $N\ge N_0(D)$. In particular, under Assumption (B), we have $w_{ij}\prec 1$.

In the following lemma we establish that a control of the $\|\cdot\|_*^{K,\mathbf{x},\mathbf{y}}$-norm for all $\mathbf{x},\mathbf{y}$ in a high probability sense is essentially equivalent to a control of the $\|\cdot\|_p$-norm for all p.

Lemma 5.4. *Let R be a random matrix and Φ a deterministic control parameter. Then the following implications hold:*

(i) If $\Phi\ge N^{-C}$, $\|R\|\le N^C$ and $|R_{\mathbf{xy}}|\prec\Phi$ for all normalized $\mathbf{x},\mathbf{y}$ and some C, then also $\|R\|_p\le_{p,\epsilon}N^\epsilon\Phi$ for all $\epsilon>0, p\ge1$.

(ii) Conversely, if $\|R\|_p\le_{p,\epsilon}N^\epsilon\Phi$ for all $\epsilon>0,p\ge1$, then $\|R\|_*^{K,\mathbf{x},\mathbf{y}}\prec\Phi$ for any fixed $K\in\mathbb{N}$, $\mathbf{x},\mathbf{y}\in\mathbb{C}^N$.

Proof. We begin with the proof of (ii) and infer from Markov's inequality and Hölder's inequality (as in (4.37)) that

$$\mathbf{P}\left(\|R\|_*>N^\sigma\Phi\right)\le\mathbf{P}\left(2\|R\|_{I_K}>N^\sigma\Phi\right)\le_p\frac{\mathbf{E}\,\|R\|_{I_K}^p}{N^{\sigma p}\Phi^p}\le_p|I_K|^{2/r}\frac{\mathbf{E}\,\|R\|_{pr}^p}{N^{\sigma p}\Phi^p}\le_{p,r,\epsilon}|I_K|^{2/r}N^{\epsilon p-\sigma p}, \tag{5.9}$$

and since $|I_K|\le 4^KN^{K+2}$ we conclude that $\|R\|_*\prec\Phi$ by choosing ϵ sufficiently small and p,r sufficiently large. On the other hand, (i) directly follows from

$$\|R\|_p \le N^\epsilon\Phi + \sup_{\|\mathbf{x}\|,\|\mathbf{y}\|\le1}\left(|R_{\mathbf{xy}}|\,\mathbf{P}[|R_{\mathbf{xy}}|\ge N^\epsilon\Phi]^{1/p}\right).$$ □

5.3. Bootstrapping step

The proof of the local law follows a *bootstrapping procedure*: First, we prove the local law for $\eta\ge N$, and afterwards we iteratively show that if the local law holds for $\eta\ge N^{\gamma_0}$, then it also holds for $\eta\ge N^{\gamma_1}$ for some $\gamma_1<\gamma_0$. We now formulate the iteration step.

Proposition 5.5. *The following holds true under the assumptions of Theorem 2.2: Let $\delta,\gamma>0$ and $\gamma_0>\gamma_1\ge\gamma$ with $4(2C_*/\mu+1)(\gamma_0-\gamma_1)<\gamma<1/2$ and suppose that*

$$\|G-M\|_p \le_{\gamma,p}\frac{N^{-\gamma/6}}{\langle z\rangle} \quad \text{in} \quad \mathbb{D}^\delta_{\gamma_0}, \tag{5.10}$$

holds for all $p\ge1$, where C_* is the constant from Theorem 4.1. Then the same inequality (5.10) (with a possibly different implicit constant depending on γ,δ,p) holds also true in $\mathbb{D}^{\delta'}_{\gamma_1}$ for some $\delta'=\delta'(\gamma,\delta)>0$. Furthermore, the same statement holds true under the assumptions of Theorem 2.1 if we replace $\mathbb{D}^\delta_{\gamma_0}$ and $\mathbb{D}^\delta_{\gamma_1}$ by $\mathbb{D}^\delta_{\gamma_0}\cap\mathbb{D}^\delta_{out}$ and $\mathbb{D}^\delta_{\gamma_1}\cap\mathbb{D}^\delta_{out}$, respectively, in the above sentence.

Proof. We first prove the assertion under the assumptions of Theorem 2.2. In the proof we will abbreviate the step size from γ_0 to γ_1 by $\gamma_s := \gamma_0 - \gamma_1$. We will suppress the dependence of the constants on δ, γ in our notation. In particular, (5.10) and (5.1) imply $\|G\|_p \leq_{p,\gamma} N^{\gamma_s} \langle z\rangle^{-1}$ in $\mathbb{D}^{\delta'}_{\gamma_0}$ with $\delta' = \delta'(\gamma)$. For fixed E the function $\eta \mapsto f(\eta) := \eta \|G(E+\mathrm{i}\eta)\|_p$ satisfies

$$\begin{aligned} \liminf_{\epsilon\to 0} \frac{f(\eta+\epsilon)-f(\eta)}{\epsilon} &\geq \|G(E+\mathrm{i}\eta)\|_p - \eta \left\| \lim_{\epsilon\to 0} \frac{G(E+\mathrm{i}(\eta+\epsilon))-G(E+\mathrm{i}\eta)}{\epsilon} \right\|_p \\ &= \|G(E+\mathrm{i}\eta)\|_p - \eta \left\| G(E+\mathrm{i}\eta)^2 \right\|_p \geq 0, \end{aligned} \tag{5.11}$$

where we used

$$\eta \left| \langle \mathbf{x}, G^2 \mathbf{y}\rangle \right| \leq \frac{\eta}{2} \left(\langle \mathbf{x}, |G|^2 \mathbf{x}\rangle + \langle \mathbf{y}, |G|^2 \mathbf{y}\rangle \right) \leq \frac{1}{2} \left(\langle \mathbf{x}, \Im G \mathbf{x}\rangle + \langle \mathbf{y}, \Im G \mathbf{y}\rangle \right)$$

in the last step. We thus know that $\eta \mapsto \eta \|G(E+\mathrm{i}\eta)\|_p$ is monotone and we can conclude that $\langle z\rangle \|G\|_p \leq_{p,\gamma} N^{2\gamma_s}$ in $\mathbb{D}^{\delta'}_{\gamma_1}$. From (4.1a) and $\gamma_s < \mu$ it thus follows that

$$\|D\|_p \leq_{p,\gamma,\epsilon} N^{\epsilon + 2(C_*/\mu + 1/2)\gamma_s - \gamma/2} \leq N^{\epsilon - \gamma/4} \quad \text{in} \quad \mathbb{D}^{\delta'}_{\gamma_1}. \tag{5.12}$$

Note that the exponent in the right hand side is independent of p; this was possible because the power of $\|G\|_q$ in (4.1a) was linear in p.

We now relate these high moment bounds to high probability bounds in the $\|\cdot\|_*$ norm, as defined before Theorem 5.2 and find for any fixed $\mathbf{x}, \mathbf{y}$ and K that $\|D\|_* \prec N^{-\gamma/4}$ from Lemma 5.4(ii) (we recall that the $\|\cdot\|_*$ implicitly depends on $\mathbf{x}, \mathbf{y}$ and K). Next, we apply (5.8) to obtain

$$\|G-M\|_* \chi(\|G-M\|_* \leq N^{-\gamma/9}) \prec \frac{N^{-\gamma/5}}{\langle z\rangle} \quad \text{in} \quad \mathbb{D}^{\delta'}_{\gamma_1}, \tag{5.13}$$

provided $K \geq 10/\gamma$. The bound (5.13) shows that there is a gap in the set of possible values for $\|G-M\|_*$. The extension of (5.10) to $\mathbb{D}^{\delta'}_{\gamma_1}$ then follows from a standard continuity argument using a fine grid of intermediate values of η: Suppose that (5.13) were true as a deterministic inequality. Since $\eta \mapsto \|(G-M)(E+\mathrm{i}\eta)\|_*$ is continuous, and for $\eta = N^{-1+\gamma_0}$ we know that $\|(G-M)(E+\mathrm{i}\eta)\|_* \leq N^{-\gamma/6}$ by (5.10) and Lemma 5.4(ii), we would conclude the same bound for $\eta = N^{-1+\gamma_1}$. Going back to the $\|\cdot\|_p$-norm by Lemma 5.4(i) we could conclude (5.10) in $\mathbb{D}^{\delta}_{\gamma_1}$. Since (5.13) may not control $\|G-M\|_*$ on a set of very small probability, and we cannot exclude a "bad" set for every $\eta \in [N^{-1+\gamma_1}, N^{-1+\gamma_0}]$, we use a fine N^{-3}-grid. The relation (5.13) is only used for a discrete set of η's and intermediate values are controlled by the η^{-1}-Lipschitz continuity of $\|G-M\|_*$ in the continuity argument above. This completes the proof of Proposition 5.5 in the setup of Theorem 2.2. The proof in the setup of Theorem 2.1 is identical except for the fact that the inequalities (5.1) and (5.8) only hold true in the restricted set $\mathbb{D}^{\delta}_{\mathrm{out}}$ without Assumption (E). □

5.4. **Proof of the local law and the absence of eigenvalues outside of the support**

We now have all the ingredients to complete the proof of Theorems 2.1 and 2.2.

Proof of Theorems 2.1, 2.2 and Corollary 2.3. We will first prove Theorem 2.2 and then remark in the end how to adapt it to prove Theorem 2.1. The proof involves five steps. In the first step we derive a weak initial isotropic bound, which we improve in the second step to obtain the isotropic local law. In the third step we use the isotropic local law to obtain the averaged local law in the bulk, which we use in the fourth step to establish that with very high probability there are no eigenvalues outside the support of ρ, also proving Corollary 2.3. Finally, in the fifth step we use the fact that there are no eigenvalues outside the support of ρ to improve the isotropic and averaged law outside the support.

Step 1: Initial isotropic bound. We claim the initial bound

$$\|G-M\|_p \leq_{p,\gamma} \frac{N^{-\gamma/6}}{\langle z\rangle} \quad \text{in} \quad \mathbb{D}^{\delta}_{\gamma} \tag{5.14}$$

for some $\delta = \delta(\gamma)$. First, we aim at proving (5.14) for large $\eta \geq N$, i.e., in $\mathbb{D}^{\delta}_{\gamma=2} = \mathbb{D}^{\delta}_{2}$ for arbitrary δ. We use that

$$\|H\| = \max_k |\lambda_k| \leq \sqrt{\operatorname{Tr}|H|^2} \leq \sqrt{\operatorname{Tr}|A|^2} + \sqrt{N^{-1}\operatorname{Tr}|W|^2} \prec \sqrt{N},$$

as follows from Assumptions (A) and (B). In fact, by computing traces of higher moments W^k and using the summable decay of the cumulants, this bound can easily be improved to $\|H\| \prec 1$. Since $|z| \geq N$ and $\|H\| \prec \sqrt{N}$, we have $\|G\|_p \leq_p \langle z\rangle^{-1}$ and $\|\Im G\|_p \leq_p \langle z\rangle^{-2} \eta$ and thus from Theorem 4.1 it follows that that

$$\|D\|_p \leq_{p,\epsilon} \frac{N^{\epsilon}}{\langle z\rangle \sqrt{N}} \quad \text{in} \quad \mathbb{D}^{\delta}_{2}.$$

We now fix normalized vectors $\mathbf{x}, \mathbf{y}$ and any $K \geq 10/\gamma$ in the norm $\|\cdot\|_* = \|\cdot\|_*^{K,\mathbf{x},\mathbf{y}}$ and translate these p norm bounds into high-probability bounds using Lemma 5.4 to infer $\|D\|_* \prec \langle z\rangle^{-1}/\sqrt{N}$ and $\|G\|_* \prec \langle z\rangle^{-1}$. Using the stability in the form of (5.8) and absorbing N^{ϵ} factors into $\prec$ we conclude

$$\|G-M\|_* \prec \frac{N^{1/2K}}{\langle z\rangle^2 \sqrt{N}} \quad \text{in} \quad \mathbb{D}^{\delta}_{2}.$$

Now (5.14) in $\mathbb{D}_2^\delta$ follows from 5.4(i) since $\mathbf{x}, \mathbf{y}$ and K were arbitrary. By applying Proposition 5.5 iteratively starting from $\gamma_0 = 2$ and (possibly) reducing δ in every step we can then conclude that (5.14) holds in all of $\mathbb{D}_\gamma^\delta$ for some $\delta = \delta(\gamma) > 0$.

Step 2: Iterative improvement of the isotropic bound. We now iteratively improve the initial bound (5.14) until we reach the intermediate bound

$$\|G - M\|_p \leq_{p,\epsilon} \frac{N^\epsilon}{\langle z\rangle}\left(\sqrt{\frac{\|\Im M\|}{N\eta}} + \frac{1}{\langle z\rangle}\frac{1}{N\eta}\right) \quad \text{in} \quad \mathbb{D}_\gamma^\delta \tag{5.15}$$

for $\delta = \delta(\epsilon) > 0$. From (5.14) and the bound on $\langle z\rangle \|M\|$ from (5.1) we conclude that $\langle z\rangle \|G\|_p$ is N^ϵ-bounded in $\mathbb{D}_\gamma^\delta$ for some $\delta = \delta(\epsilon) > 0$. Then from Theorem 4.1 and (5.14), again, it follows that

$$\|D\|_p \leq_{p,\epsilon} N^\epsilon \sqrt{\frac{\|\Im G\|_q}{N\eta}} \quad \text{and} \quad \|G - M\|_* + \|D\|_* \prec N^{-\gamma/6} \quad \text{in} \quad \mathbb{D}_\gamma^\delta. \tag{5.16}$$

From now on all claimed bounds hold true uniformly in all of $\mathbb{D}_\gamma^\delta$; we will therefore suppress this qualifier in the following steps. In order to prove (5.15), we show inductively

$$\|G - M\|_p \leq_{p,\epsilon} N^\epsilon \Psi_l, \tag{5.17}$$

where we define successively improving control parameters $(\Psi_l)_{l=0}^L$ through $\Psi_0 := 1$ and $\Psi_{l+1} := N^{-\sigma}\Psi_l = N^{-(l+1)\sigma}$, where $\sigma \in (0,1)$ is arbitrary. The final iteration step L is chosen to be the largest integer such that

$$\Psi_L \geq \frac{N^\sigma}{\langle z\rangle}\left(\sqrt{\frac{\|\Im M\|}{N\eta}} + \frac{1}{\langle z\rangle}\frac{N^\sigma}{N\eta}\right). \tag{5.18}$$

For the induction step from l to $l+1$, we write $\Im G = \Im M + \Im(G - M)$ and we continue from (5.16) and (5.17) and estimates that

$$\|D\|_p \leq_{p,\epsilon} N^\epsilon \left(\sqrt{\frac{\|\Im M\|}{N\eta}} + \sqrt{\frac{\Psi_l}{N\eta}}\right) \leq_{p,\epsilon} N^\epsilon \left(\sqrt{\frac{\|\Im M\|}{N\eta}} + \frac{1}{\langle z\rangle}\frac{N^\sigma}{N\eta} + \langle z\rangle N^{-\sigma}\Psi_l\right).$$

Thus we also have, for any normalized $\mathbf{x}, \mathbf{y}$,

$$\|D\|_* = \|D\|_*^{K,\mathbf{x},\mathbf{y}} \prec \sqrt{\frac{\|\Im M\|}{N\eta}} + \frac{1}{\langle z\rangle}\frac{N^\sigma}{N\eta} + \langle z\rangle N^{-\sigma}\Psi_l$$

and from (5.8) we conclude

$$\|G - M\|_* \prec \frac{N^{1/2K}}{\langle z\rangle}\left(\sqrt{\frac{\|\Im M\|}{N\eta}} + \frac{1}{\langle z\rangle}\frac{N^\sigma}{N\eta}\right) + N^{1/2K-\sigma}\Psi_l$$

provided $K \geq 7/\gamma$ (c.f. the bound on $\|D\|_*$ from (5.16) and the definition of ϵ-neighbourhoods in (5.2)). In particular, since K can be chosen arbitrarily large, we find, for any normalized $\mathbf{x}, \mathbf{y}$ that

$$|(G - M)_{\mathbf{xy}}| \prec \frac{1}{\langle z\rangle}\left(\sqrt{\frac{\|\Im M\|}{N\eta}} + \frac{1}{\langle z\rangle}\frac{N^\sigma}{N\eta}\right) + N^{-\sigma}\Psi_l \leq 2N^{-\sigma}\Psi_l,$$

where we used $l < L$ and (5.18) in the last step. By the definition of Ψ_{l+1} we infer

$$\|G - M\|_p \leq_{p,\epsilon} N^\epsilon \Psi_{l+1},$$

completing the induction step, and thereby the proof of (5.15).

Finally, in order to obtain (2.7a) from (5.15), we recall

$$\|\Im M\| \leq \|M\| \leq_\epsilon N^\epsilon \tag{5.19}$$

from Proposition 5.1(vi) and (2.7a) follows.

Step 3: Averaged bound. First, it follows from (2.1) and (2.4) or equivalently from $\widetilde{\mathcal{J}}[G - M, D] = 0$ that $G - M$ satisfies the following quadratic relation

$$G - M = (1 - \mathcal{C}_M\mathcal{S})^{-1}\big[-MD + M\mathcal{S}[G - M](G - M)\big]$$

and therefore

$$\|\langle B(G - M)\rangle\|_p \leq \big\|\langle B(1 - \mathcal{C}_M\mathcal{S})^{-1}[MD]\rangle\big\|_p + \big\|\langle B(1 - \mathcal{C}_M\mathcal{S})^{-1}\big[M\mathcal{S}[G - M](G - M)\big]\rangle\big\|_p.$$

By geometric expansion, as in (5.6), it follows that

$$\big\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\big\|_{\|\cdot\|\to\|\cdot\|} \leq 1 + \|M\|^2 \|\mathcal{S}\| + \|M\|^4 \|\mathcal{S}\|^2 \big\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\big\|_{\mathrm{hs}\to\mathrm{hs}}$$

and thus that $\big\|\big((1 - \mathcal{C}_M S)^{-1}\big)^*[B^*]\big\| \leq_\epsilon N^\epsilon \|B\|$ by (5.1). Using (4.1b), where $\big((1 - \mathcal{C}_M S)^{-1}\big)^*[B^*]$ plays the role of B, and writing $\|\Im G\|_q \leq \|\Im M\| + \|G - M\|_q$ and using (5.15) we can conclude that

$$\|\langle B(G - M)\rangle\|_p \leq_{p,\epsilon,\gamma} \frac{\|B\| N^\epsilon}{\langle z\rangle}\left[\langle z\rangle \frac{\|\Im M\|}{N\eta} + \sqrt{\frac{\|\Im M\|}{N\eta}}\frac{1}{N\eta} + \frac{1}{(N\eta)^2}\right] \tag{5.20}$$

from Lemma D.2. Now (2.7b) follows directly from (5.20) and (5.19).

The proof of Theorem 2.2 is now complete. For the proof of Theorem 2.1 the first three steps are identical except that we only work in the resticted domains $\mathbb{D}^\delta_\gamma \cap \mathbb{D}^\delta_{\text{out}}$. Due to (5.1) and (5.8), it then follows that in $\mathbb{D}^\delta_{\text{out}}$ the only place where the above proof used Assumption (E) is (5.19). In the absence of Assumption (E) we replace (5.19) by the bound $\|\Im M\| \le \eta \operatorname{dist}(z, \operatorname{supp}\mu)^{-2}$ from Proposition 5.1 in (5.15) and (5.20), which only adds another negligible N^ϵ factor. This proves

$$\|G-M\|_p \le_{p,\epsilon} \frac{N^\epsilon}{\langle z\rangle}\left(\sqrt{\frac{1}{N}}+\frac{1}{\langle z\rangle}\frac{1}{N\eta}\right), \quad \|\langle B(G-M)\rangle\|_p \le_{p,\epsilon,\gamma} \frac{\|B\|\,N^\epsilon}{\langle z\rangle}\left[\frac{1}{N}+\sqrt{\frac{1}{N}}\frac{1}{N\eta}+\frac{1}{(N\eta)^2}\right] \tag{5.21}$$

in the restricted domain $\mathbb{D}^\delta_\gamma \cap \mathbb{D}^\delta_{\text{out}}$. We now need two additional steps to prove Theorem 2.1 in all of $\mathbb{D}^\delta_{\text{out}}$.

Step 4: Absence of eigenvalues outside of the support. For $B = 1$ it follows from (5.21) and a spectral decomposition of H that with very high probability in the sense of Corollary 2.3 there are no eigenvalues outside the support of μ. Indeed, if there is an eigenvalue λ with $\operatorname{dist}(\lambda, \operatorname{supp}\mu) \ge N^{-\delta}$, then $|\langle G(\lambda+i\eta)\rangle| \ge |\langle \Im G(\lambda+i\eta)\rangle| \ge 1/N\eta$. From (5.21) with $\epsilon = 1/4$ and $\gamma = 1/2$ we have

$$\mathbf{P}\left(\exists\lambda \text{ with } \operatorname{dist}(\lambda,\operatorname{supp}\mu)\ge N^{-\delta}\right) \le \mathbf{P}\left(|\langle G-M\rangle| \ge c/N\eta \text{ in } \mathbb{D}^\delta_{\text{out}}\cap\mathbb{D}^\delta_{1/2}\right) \lesssim \inf_{\eta\ge N^{-1/2}}\left(N^\epsilon\left[\eta+\frac{1}{\sqrt{N}}+\frac{1}{N\eta}\right]\right)^p \lesssim N^{-p/4}.$$

Now Corollary 2.3 follows from the remark about the dependence of δ on ϵ in Theorem 2.1.

Step 5: Improved bounds outside of the support. Now we fix z such that $\operatorname{dist}(z,\operatorname{supp}\rho) \ge N^{-\delta}$ and $\eta \ge N^{-1+\gamma}$. Then we have $\|\Im G\| \prec \eta\langle z\rangle^{-2}$ and $\|G\| \prec \langle z\rangle^{-1}$ and also $\|\Im G\|_p \le_{p,\epsilon} N^\epsilon\eta\langle z\rangle^{-2}$ and $\|G\|_p \le_{p,\epsilon} N^\epsilon\langle z\rangle^{-1}$ and we infer from Theorem 4.1 that

$$\|D\|_p \le_{p,\epsilon} \frac{N^\epsilon}{\langle z\rangle\sqrt{N}} \quad \text{and therefore} \quad \|D\|_* \prec \frac{1}{\langle z\rangle\sqrt{N}}.$$

Again using stability in the form of (5.8) we find

$$\|G-M\|_* \prec \frac{N^{1/2K}}{\langle z\rangle^2\sqrt{N}}$$

and since K was arbitrary we also have

$$\|G-M\|_p \le_{p,\epsilon} \frac{N^\epsilon}{\langle z\rangle^2\sqrt{N}}.$$

By Lipschitz-continuity of G and M with Lipschitz constant of order one we can extend the regime of validity of this bound from $\eta \ge N^{-1+\gamma}$ to $\eta \ge 0$ to conclude (2.6a). The improvement on the averaged law outside of the support of the ρ then follows immediately from the improved isotropic law and the fact that with very high probability there are no eigenvalues outside of the support of ρ. □

6. Delocalization, rigidity and universality

In this section we infer eigenvector delocalization, eigenvalue rigidity and universality in the bulk from the local law in Theorem 2.2. These proofs are largely independent of the correlation structure of the random matrix, so arguments that have been developed for Wigner matrices over the last few years can be applied with minimal modifications. Especially the *three step strategy* for proving bulk universality (see [20] for a short summary) has been streamlined recently [17, 31, 32] so that the only model-dependent input is the local law. The small modifications required for the correlated setup have been presented in detail in [1] and we will not repeat them. Here we only explain why the proofs in [1] work under the more general conditions imposed in the current paper. In fact, the proof of the eigenvector delocalization and eigenvector rigidity from [1] holds *verbatim* in the current setup as well. The proof of the bulk universality in [1] used that the correlation length was N^ϵ at a technical step that can be easily modified for our weaker assumptions. In the following we will highlight which arguments of [1] have to be modified in the current, more general, setup.

Proof of Corollary 2.4 on bulk eigenvector delocalization. As usual, delocalization of eigenvectors corresponding to eigenvalues in the bulk is an immediate corollary of the local law since for the eigenvectors $\boldsymbol{u}_k = \big(u_k(i)\big)_{i\in J}$ and eigenvalues λ_k of H and $i \in J$ we find from the spectral decomposition

$$C \gtrsim \Im G_{ii} = \eta\sum_k \frac{|u_k(i)|^2}{(E-\lambda_k)^2+\eta^2} \ge \frac{|u_k(i)|^2}{\eta} \quad \text{for} \quad z = E+\mathrm{i}\eta,$$

where the first inequality is meant in a high-probability sense and follows from the boundedness of M and Theorem 2.2, and the last inequality followed assuming that E is η-close to λ_k. □

Proof of Corollary 2.5 on bulk eigenvalue rigidity. Rigidity of bulk eigenvalues follows, verbatim as in [1, Corollary 2.9], from the improved local law away from the spectrum and [3, Lemma 5.1]. □

Proof of Corollary 2.6 on bulk universality. Bulk universality follows from the *three step strategy*, out of which only the third step requires a minor modification, compared to [1]. Since in [1] arbitrarily high polynomial decay outside of N^ϵ neighbourhoods was assumed, we have to replace to three term Taylor expansion in [1, Lemma 7.5] by an $2/\mu$-term cumulant expansion to accommodate for neighbourhoods of sizes $N^{1/2-\mu}$.

The key input for the universality proof through Dyson Brownian motion is the Ornstein Uhlenbeck (OU) process, which creates a family $H(t)$ of interpolating matrices between the original matrix $H = H(0)$ and a matrix with sizeable Gaussian component, for which universality is known from the second step of the three step strategy. The OU process is defined via

$$\mathrm{d}H(t) = -\frac{1}{2}(H(t) - A)\,\mathrm{d}t + \Sigma^{1/2}[\mathrm{d}B(t)], \qquad \text{where} \quad \Sigma[R] := \mathbf{E}\,\langle W^* R\rangle\, W, \tag{6.1}$$

where $B(t)$ is a matrix of independent (real, or complex according to the symmetry class of H) Brownian motions. It is designed in a way which preserves mean and covariances along the flow, i.e., $H(t) = A + N^{-1/2}W(t)$ and it is easy to check that $\mathbf{E}\,W(t) = 0$ and $\mathbf{Cov}\,(w_\alpha(t), w_\beta(t)) = \mathbf{Cov}\,(w_\alpha(0), w_\beta(0))$, where $W(t) = (w_\alpha(t))_{\alpha\in I}$. Furthermore, Assumptions (C), (D) hold also, uniformly in t, for $W(t)$. Indeed, adding an independent Gaussian vector $\boldsymbol{g} = (g_{\alpha_1}, \dots, g_{\alpha_k})$ to $(w_{\alpha_1}, \dots, w_{\alpha_k})$ leaves the cumulant invariant by additivity

$$\kappa(w_{\alpha_1} + g_{\alpha_1}, \dots, w_{\alpha_k} + g_{\alpha_k}) = \kappa(w_{\alpha_1}, \dots, w_{\alpha_k}) + \kappa(g_{\alpha_1}, \dots, g_{\alpha_k})$$

and the fact that cumulants of Gaussian vectors vanish for $k \geq 3$ (for $k \geq 2$ we already noticed that, by design, the expectation and the covariance is invariant under t). We now estimate

$$\mathbf{E}\,f(N^{-1/2}W(t)) - \mathbf{E}\,f(N^{-1/2}W(0))$$

for smooth functions f. For notational purposes we set $v_\alpha(t) = N^{-1/2}w_\alpha(t)$ and $V(t) = N^{-1/2}W(t)$ and will often suppress the t-dependence. It follows from Ito's formula that

$$2\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{E}\,f(V) = -\,\mathbf{E}\sum_\alpha v_\alpha(\partial_\alpha f)(V) + \sum_{\alpha,\beta}\mathbf{Cov}\,(v_\alpha, v_\beta)\,\mathbf{E}(\partial_\alpha\partial_\beta f)(V).$$

We now apply Proposition 3.2 to the first term and obtain

$$\begin{aligned} 2\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{E}\,f = &-\sum_{2\leq m<R}\sum_\alpha\sum_{\boldsymbol{\beta}\in\mathcal{N}^m}\frac{\kappa(v_\alpha, v_{\boldsymbol{\beta}})}{m!}(\mathbf{E}\,\partial_\alpha\partial_{\boldsymbol{\beta}}f) - \sum_{m<R}\sum_\alpha\sum_{\boldsymbol{\beta}\in\mathcal{N}^m}\mathbf{E}\,\frac{K(v_\alpha; v_{\boldsymbol{\beta}}) - \kappa(v_\alpha, v_{\boldsymbol{\beta}})}{m!}\partial_\alpha\partial_{\boldsymbol{\beta}}f\big|_{W_{\mathcal{N}}=0} \\ &-\sum_\alpha\Omega(\partial_\alpha f, \alpha, \mathcal{N}) + \sum_\alpha\sum_{\beta\in I\setminus\mathcal{N}}\kappa(v_\alpha, v_\beta)\,\mathbf{E}\,\partial_\alpha\partial_\beta f, \end{aligned}$$

where we used a cancellation for the $m = 1$ term in $\beta \in \mathcal{N}$ and the fact that $\kappa(v_\alpha) = \mathbf{E}\,v_\alpha = 0$ for the $m = 0$ term. We now estimate the four terms separately. The sum in the last term is of size N^4, the derivative contributes an N^{-1} and the covariance is assumed to be N^{-3} small, i.e., the last term is of order 1. The first term for fixed m is of size $|||\kappa|||^{\mathrm{av}}\,N^{2-(m+1)/2}$ and therefore altogether of size $|||\kappa|||^{\mathrm{av}}\,\sqrt{N}$. Estimating the sums by their size, and the derivative by its prefactor $N^{-(R+1)/2}$, we find from (3.8) that the third term is of size

$$N^2\,|\mathcal{N}|^R\,N^{-(R+1)/2} \leq N^{3/2-\mu R},$$

which can be made smaller than $\sqrt{N}$ by choosing $R = 2/\mu$. Finally, the second term is naively of size $N^{3/2}$, but using (3.1c), the security layers and the pigeon-hole principle as in (3.16) or in (4.24), this can be improved to $N^{-3/2}$. We can conclude that

$$\left|\mathbf{E}\,\frac{\mathrm{d}}{\mathrm{d}t}f(V(t))\right| \lesssim \sqrt{N} \quad \text{and therefore} \quad |\mathbf{E}\,f(V(t)) - \mathbf{E}\,f(V(0))| \lesssim t\sqrt{N}. \tag{6.2}$$

The remaining argument of [1, Section 7.2] can be, assuming fullness as in Assumption (F), followed verbatim to conclude bulk universality. □

Appendix A. Cumulants

In this section we provide some results on cumulants which we refer to in the main part of the proof. The section largely follows the approach of [35, 41], but our application requires a more quantitative version of the independence property exhibited by cumulants, which we work out here.

Cumulants $\kappa_{\boldsymbol{m}}$ of a random vector $\boldsymbol{w} = (w_1, \dots, w_l)$ are traditionally defined as the coefficients of log-characteristic function

$$\log\mathbf{E}\,e^{i\boldsymbol{t}\cdot\boldsymbol{w}} = \sum_{\boldsymbol{m}}\kappa_{\boldsymbol{m}}\frac{(i\boldsymbol{t})^{\boldsymbol{m}}}{\boldsymbol{m}!},$$

while the (mixed) moments of $\boldsymbol{w}$ are the coefficients of the characteristic function

$$\mathbf{E}\,e^{i\boldsymbol{t}\cdot\boldsymbol{w}} = \sum_{\boldsymbol{m}}(\mathbf{E}\,\boldsymbol{w}^{\boldsymbol{m}})\frac{(i\boldsymbol{t})^{\boldsymbol{m}}}{\boldsymbol{m}!},$$

where $\sum_{\boldsymbol{m}}$ is the sum over all multi-indices $\boldsymbol{m} = (m_1, \dots, m_l)$. Thus

$$\exp\left(\sum_{\boldsymbol{m}}\kappa_{\boldsymbol{m}}\frac{(i\boldsymbol{t})^{\boldsymbol{m}}}{\boldsymbol{m}!}\right) = \sum_{\boldsymbol{m}}(\mathbf{E}\,\boldsymbol{w}^{\boldsymbol{m}})\frac{(i\boldsymbol{t})^{\boldsymbol{m}}}{\boldsymbol{m}!}. \tag{A.1}$$

It is easy to check that for a set $A \subset [l]$ the coefficient of $\prod_{a\in A} t_a$ in (A.1) is given by

$$\mathbf{E}\,\Pi\underline{w}_A = \Big(\prod_{a\in A}\partial_{t_a}\Big)\exp\left(\sum_{\boldsymbol{m}}\kappa_{\boldsymbol{m}}\frac{\boldsymbol{t}^{\boldsymbol{m}}}{\boldsymbol{m}!}\right)\Big|_{\boldsymbol{t}=0} = \sum_{\mathcal{P}\vdash A}\kappa^{\mathcal{P}},$$

where $\mathcal{P}\vdash A$ indicates the summation over all partitions of the (multi)set A, and where for partitions $\mathcal{P}=\{\mathcal{P}_1,\dots,\mathcal{P}_b\}$ of A we defined $\kappa^{\mathcal{P}}=\prod_{k=1}^{b}\kappa_{\chi(\mathcal{P}_k)}$ with $\chi(\mathcal{P}_k)$ being the characteristic multi-index of the set $\mathcal{P}_k$. Thus for a partition $\mathcal{Q}$ of $[l]$ it follows that

$$M^{\mathcal{Q}} := \prod_{\mathcal{Q}_i\in\mathcal{Q}}\mathbf{E}\,\Pi\underline{w}_{\mathcal{Q}_i} = \prod_{\mathcal{Q}_i\in\mathcal{Q}}\sum_{\mathcal{P}\vdash\mathcal{Q}_i}\kappa^{\mathcal{P}} = \sum_{\mathcal{P}\le\mathcal{Q}}\kappa^{\mathcal{P}}, \tag{A.2}$$

where $\mathcal{P}\le\mathcal{Q}$ indicates that $\mathcal{P}$ is a finer partition than $\mathcal{Q}$.

Now we establish the inverse of the relation (A.2), i.e., express cumulants in terms of products of moments. To do so, we notice that the set of partitions $\mathcal{P}$ on $[l]$ (or, in fact, any finite set) is a partially ordered set with respect to the relation $\le$. It is, in fact, also a lattice, as any two partitions $\mathcal{P},\mathcal{Q}$ have both a unique greatest lower bound $\mathcal{P}\wedge\mathcal{Q}$ and a unique least upper bound $\mathcal{P}\vee\mathcal{Q}$. One then defines the *incidence algebra* as the algebra of scalar functions f mapping intervals $[\mathcal{P},\mathcal{Q}]=\{\,\mathcal{R}\mid\mathcal{P}\le\mathcal{R}\le\mathcal{Q}\,\}$ to scalars $f(\mathcal{P},\mathcal{Q})$ equipped with point-wise addition and scalar multiplication and the product $*$

$$(f*g)(\mathcal{P},\mathcal{Q}) = \sum_{\mathcal{P}\le\mathcal{R}\le\mathcal{Q}} f(\mathcal{P},\mathcal{R})g(\mathcal{R},\mathcal{Q}).$$

There are three special elements in the incidence algebra; the δ function mapping $[\mathcal{P},\mathcal{Q}]$ to $\delta(\mathcal{P},\mathcal{Q})=1$ if $\mathcal{P}=\mathcal{Q}$ and $\delta(\mathcal{P},\mathcal{Q})=0$ otherwise, the ζ function mapping all intervals $[\mathcal{P},\mathcal{Q}]$ to $\zeta(\mathcal{P},\mathcal{Q})=1$, and finally the Möbius function defined inductively via

$$\mu(\mathcal{P},\mathcal{Q}) = \begin{cases} 1, & \text{if } \mathcal{P}=\mathcal{Q},\\ -\sum_{\mathcal{P}\le\mathcal{R}<\mathcal{Q}}\mu(\mathcal{P},\mathcal{R}), & \text{if } \mathcal{P}<\mathcal{Q}.\end{cases}$$

The δ function is the unit element of the incidence algebra. It is well known (and easy to check) that the multiplicative inverse of the zeta function is the Möbius function, and vice versa, i.e., that $\mu*\zeta=\zeta*\mu=\delta$. Thus it follows that for any functions F and G on the partitions, we have

$$F(\mathcal{P})=\sum_{\mathcal{Q}\le\mathcal{P}}G(\mathcal{Q}) \qquad \text{if and only if} \qquad G(\mathcal{Q})=\sum_{\mathcal{P}\le\mathcal{Q}}\mu(\mathcal{P},\mathcal{Q})F(\mathcal{P}).$$

Applying this equivalence to (A.2) yields

$$\kappa^{\mathcal{P}} = \sum_{\mathcal{Q}\le\mathcal{P}}\mu(\mathcal{Q},\mathcal{P})M^{\mathcal{Q}} \tag{A.3}$$

and thus it only remains to identify μ. One can check that for $\mathcal{P}\le\mathcal{Q}$, $\mu(\mathcal{P},\mathcal{Q})$ is given by

$$\mu(\mathcal{P},\mathcal{Q}) = (-1)^{n-r}0!^{r_1}1!^{r_2}\dots(n-1)!^{r_n},$$

where n is the number of blocks of $\mathcal{P}$, r is the number of blocks of $\mathcal{Q}$ and r_i is the number of blocks of $\mathcal{Q}$ which contain exactly i blocks of $\mathcal{P}$. For the particular choice of the trivial partition $\{[l]\}$ of $[l]$ it follows that

$$\kappa(w_1,\dots,w_l) := \kappa_{(1,\dots,1)} = \kappa^{\{[l]\}} = \sum_{\mathcal{P}}(-1)^{|\mathcal{P}|-1}(|\mathcal{P}|-1)!M^{\mathcal{P}} = \sum_{\mathcal{P}}(-1)^{|\mathcal{P}|-1}(|\mathcal{P}|-1)!\prod_{\mathcal{P}_i\in\mathcal{P}}\mathbf{E}\,\Pi\underline{w}_{\mathcal{P}_i}, \tag{A.4}$$

providing an alternative (purely combinatorial) definition of cumulants.

Lemma A.1. *If for a partition of the index set $[n]=A\sqcup B$ with $|A|,|B|>0$ the random variables $\underline{w}_A$ and $\underline{w}_B$ are independent, then $\kappa(\underline{w}_{[n]})=\kappa(\underline{w}_A,\underline{w}_B)=0$. If, instead of independence, we merely assume that*

$$\mathbf{Cov}\left(f(w_i\mid i\in A), g(w_j\mid j\in B)\right)\le\epsilon\left\|f\right\|_2\left\|g\right\|_2 \tag{A.5}$$

for all f,g, and that the random variables w_i have finite $2n$-th moments $\max_i\mathbf{E}\left|w_i\right|^{2n}\le\mu_{2n}$, then we still have

$$\left|\kappa(\underline{w}_{[n]})\right|\le\epsilon\,C(n,\mu_{2n}). \tag{A.6}$$

Proof. We first recall the well known proof, based on the relations (A.2)–(A.3), that the cumulant of independent $\underline{w}_A,\underline{w}_B$ vanishes. Let $\mathcal{P}$ be a partition on $[n]$, $\mathcal{Q}$ a partition on A and $\mathcal{R}$ a partition on B. $\mathcal{P}$ naturally induces partitions $\mathcal{P}\cap A$ and $\mathcal{P}\cap B$ on A and B; conversely $\mathcal{Q}$ and $\mathcal{R}$ naturally induce a partition $\mathcal{Q}\cup\mathcal{R}$ on $[n]$. We observe that $\mathcal{Q}\le\mathcal{P}\cap A$ and $\mathcal{R}\le\mathcal{P}\cap B$ if and only if $\mathcal{Q}\cup\mathcal{R}\le\mathcal{P}$. We then compute

$$\begin{aligned}\kappa(\underline{w}_{[n]}) &= \sum_{\mathcal{P}}\mu(\mathcal{P},\{[n]\})M^{\mathcal{P}} = \sum_{\mathcal{P}}\mu(\mathcal{P},\{[n]\})M^{\mathcal{P}\cap A}M^{\mathcal{P}\cap B} = \sum_{\mathcal{P}}\mu(\mathcal{P},\{[n]\})\Big(\sum_{\mathcal{Q}\le\mathcal{P}\cap A}\kappa^{\mathcal{Q}}\Big)\Big(\sum_{\mathcal{R}\le\mathcal{P}\cap B}\kappa^{\mathcal{R}}\Big)\\ &= \sum_{\mathcal{Q}\vdash A}\sum_{\mathcal{R}\vdash B}\sum_{\mathcal{P}\vdash[n]}\zeta(\mathcal{Q}\cup\mathcal{R},\mathcal{P})\mu(\mathcal{P},\{[n]\})\kappa^{\mathcal{Q}}\kappa^{\mathcal{R}} = \sum_{\mathcal{Q}\vdash A}\sum_{\mathcal{R}\vdash B}\delta(\mathcal{Q}\cup\mathcal{R},\{[n]\})\kappa^{\mathcal{Q}}\kappa^{\mathcal{R}} = 0,\end{aligned} \tag{A.7}$$

where the first equality followed from (A.3), the second equality from independence, the third equality from (A.2), the fourth equality from the previous observation, the fifth equality from $\delta=\zeta*\mu$ and the ultimate equality from the fact that the trivial partition cannot be decomposed into two partitions on smaller sets, using that $|A|,|B|>0$.

If $\underline{w}_A$ and $\underline{w}_B$ are not independent but merely (A.5) holds, then there is an additional covariance term in the second step in the above equation. We write

$$M^{\mathcal{P}} = \prod_{\mathcal{P}_i \in \mathcal{P}} \mathbf{E}\, \Pi \underline{w}_{\mathcal{P}_i} = \prod_{\mathcal{P}_i \in \mathcal{P}} \Big[(\mathbf{E}\, \Pi \underline{w}_{\mathcal{P}_i \cap A})(\mathbf{E}\, \Pi \underline{w}_{\mathcal{P}_i \cap B}) + \mathbf{Cov}\left(\Pi \underline{w}_{\mathcal{P}_i \cap A}, \Pi \underline{w}_{\mathcal{P}_i \cap B}\right) \Big], \tag{A.8}$$

and thus the claim follows from (A.5). □

Appendix B. Precumulants and Wick polynomials

The precumulants defined in Section 3 are structurally similar to the well known *Wick polynomials* (which are also known as *Appell polynomials*). We first recall some basic definitions and facts about Wick polynomials from [23]. For a random vector $\boldsymbol{X}$ of length $|\boldsymbol{X}|$ we can define the Wick polynomial $:\boldsymbol{X}:$ as the derivative

$$:\boldsymbol{X}: := \partial_{t_1} \dots \partial_{t_{|\boldsymbol{X}|}} \frac{e^{\boldsymbol{t}\cdot\boldsymbol{X}}}{\mathbf{E}\, e^{\boldsymbol{t}\cdot\boldsymbol{X}}}\Big|_{\boldsymbol{t}=0}.$$

Alternatively, we can define $:\boldsymbol{X}:$ combinatorially as

$$:\boldsymbol{X}: = \sum_{\boldsymbol{X}' \subset \boldsymbol{X}} (\Pi \boldsymbol{X}') \sum_{\mathcal{P} \vdash \boldsymbol{X} \setminus \boldsymbol{X}'} (-1)^{|\mathcal{P}|} \prod_{\mathcal{P}_i \in \mathcal{P}} \kappa(\mathcal{P}_i). \tag{B.1a}$$

or indirectly via

$$\Pi \boldsymbol{X} = \sum_{\boldsymbol{X}' \subset \boldsymbol{X}} :\boldsymbol{X}': \big(\mathbf{E}\, \Pi(\boldsymbol{X} \setminus \boldsymbol{X}')\big). \tag{B.1b}$$

One useful property of Wick polynomials is that for any random variable Y we have

$$\mathbf{E}\, Y : \boldsymbol{X}_1 \sqcup \boldsymbol{X}_2 : = 0 \qquad \text{whenever} \quad \boldsymbol{X}_1 \quad \text{is independent of} \quad \{\boldsymbol{X}_2, Y\} \tag{B.2}$$

and $\boldsymbol{X}_1$ is not empty. Eq. (B.2) follows, for example, immediately from the analytical definition since

$$\mathbf{E}\, Y : \boldsymbol{X}_1 \sqcup \boldsymbol{X}_2 : = \partial_{\boldsymbol{t}} \frac{\mathbf{E}\, Y e^{\boldsymbol{t}_1 \cdot \boldsymbol{X}_1 + \boldsymbol{t}_2 \cdot \boldsymbol{X}_2}}{\mathbf{E}\, e^{\boldsymbol{t}_1 \cdot \boldsymbol{X}_1 + \boldsymbol{t}_2 \cdot \boldsymbol{X}_2}}\Big|_{\boldsymbol{t}=0} = \partial_{\boldsymbol{t}} \frac{\mathbf{E}\, Y e^{\boldsymbol{t}_2 \cdot \boldsymbol{X}_2}}{\mathbf{E}\, e^{\boldsymbol{t}_2 \cdot \boldsymbol{X}_2}}\Big|_{\boldsymbol{t}=0}$$

by independence and the remaining derivative vanishes as the function is constant with respect to $\boldsymbol{t}_1$.

Our pre-cumulants $K(X; \boldsymbol{Y})$ and their centered versions $K(X; \boldsymbol{Y}) - \kappa(X, \boldsymbol{Y})$ are inherently non-symmetric functions due to the special role of X. After symmetrization, however, we can express them through Wick polynomials as

$$\sum_{X \in \boldsymbol{X}} \big[K(X; \boldsymbol{X} \setminus \{X\}) - \kappa(\boldsymbol{X})\big] = |\boldsymbol{X}|\, \Pi \boldsymbol{X} - \sum_{\boldsymbol{X}' \subset \boldsymbol{X}} |\boldsymbol{X}'| \, \big(\mathbf{E}\, \Pi \boldsymbol{X}'\big) : \boldsymbol{X} \setminus \boldsymbol{X}' : . \tag{B.3}$$

In order to prove (B.3) we start from (3.1b) and compute

$$\begin{aligned}\sum_{X \in \boldsymbol{X}} \big[K(X; \boldsymbol{X} \setminus \{X\}) - \kappa(\boldsymbol{X})\big] &= |\boldsymbol{X}|\, \Pi \boldsymbol{X} - \sum_{\boldsymbol{X}' \subset \boldsymbol{X}} |\boldsymbol{X} \setminus \boldsymbol{X}'| \, (\Pi \boldsymbol{X}') \kappa(\boldsymbol{X} \setminus \boldsymbol{X}') \\ &= |\boldsymbol{X}|\, \Pi \boldsymbol{X} - \sum_{\boldsymbol{X}'' \subset \boldsymbol{X}' \subset \boldsymbol{X}} |\boldsymbol{X} \setminus \boldsymbol{X}'| : \boldsymbol{X}'' : \big(\mathbf{E}\, \Pi(\boldsymbol{X}' \setminus \boldsymbol{X}'')\big) \kappa(\boldsymbol{X} \setminus \boldsymbol{X}'),\end{aligned}$$

where the second inequality followed from (B.1b). We now relabel the summation indices to obtain

$$\sum_{X \in \boldsymbol{X}} \big[K(X; \boldsymbol{X} \setminus \{X\}) - \kappa(\boldsymbol{X})\big] = |\boldsymbol{X}|\, \Pi \boldsymbol{X} - \sum_{\boldsymbol{X}'' \subset \boldsymbol{X}' \subset \boldsymbol{X}} |\boldsymbol{X}''| : \boldsymbol{X} \setminus \boldsymbol{X}' : \big(\mathbf{E}\, \Pi(\boldsymbol{X}' \setminus \boldsymbol{X}'')\big) \kappa(\boldsymbol{X}''),$$

from which (B.3) follows using the well known cumulant identity

$$|\boldsymbol{X}'| \, \mathbf{E}\, \Pi \boldsymbol{X}' = \sum_{\boldsymbol{X}'' \subset \boldsymbol{X}'} |\boldsymbol{X}''| \, \big(\mathbf{E}\, \Pi(\boldsymbol{X}' \setminus \boldsymbol{X}'')\big) \kappa(\boldsymbol{X}''). \tag{B.4}$$

In order to prove (B.4), we use (A.2) on the rhs. to obtain

$$\sum_{\boldsymbol{X}'' \subset \boldsymbol{X}'} |\boldsymbol{X}''| \, \big(\mathbf{E}\, \Pi(\boldsymbol{X}' \setminus \boldsymbol{X}'')\big) \kappa(\boldsymbol{X}'') = \sum_{\boldsymbol{X}'' \subset \boldsymbol{X}'} |\boldsymbol{X}''| \, \kappa(\boldsymbol{X}'') \sum_{\mathcal{P} \vdash \boldsymbol{X}' \setminus \boldsymbol{X}''} \kappa^{\mathcal{P}} = \sum_{\mathcal{P} \vdash \boldsymbol{X}'} \kappa^{\mathcal{P}} \sum_{\substack{\boldsymbol{X}'' \subset \boldsymbol{X}' \\ \boldsymbol{X}'' \in \mathcal{P}}} |\boldsymbol{X}''| = |\boldsymbol{X}'| \sum_{\mathcal{P} \vdash \boldsymbol{X}'} \kappa^{\mathcal{P}},$$

from which (B.4) follows by another application of (A.2).

Finally we remark that a quantitative variant of (B.2) for the pre-cumulants was centrally used in our proof in Section 4.2. Qualitatively the analogue of (B.2) for pre-cumulants reads

$$\mathbf{E}\, Y \big[K(X; \boldsymbol{X}_1, \boldsymbol{X}_2) - \kappa(X, \boldsymbol{X}_1, \boldsymbol{X}_2)\big] = 0 \qquad \text{whenever} \quad \{X, \boldsymbol{X}_1\} \quad \text{is independent of} \quad \{\boldsymbol{X}_2, Y\} \tag{B.5}$$

and $\boldsymbol{X}_2$ is non-empty. Indeed, from the pre-cumulant decoupling identity (3.1c) we have that

$$\mathbf{E}\, Y \big[K(X; \boldsymbol{X}_1, \boldsymbol{X}_2) - \kappa(X, \boldsymbol{X}_1, \boldsymbol{X}_2)\big] = \mathbf{E}\, Y (\Pi \boldsymbol{X}_2) \big[K(X; \boldsymbol{X}_1) - \kappa(X, \boldsymbol{X}_1)\big] - \sum_{\substack{\boldsymbol{X}_1' \subset \boldsymbol{X}_1 \\ \boldsymbol{X}_2' \subsetneq \boldsymbol{X}_2}} \mathbf{E}\, Y (\Pi \boldsymbol{X}_1')(\Pi \boldsymbol{X}_2') \kappa(X, \boldsymbol{X}_1 \setminus \boldsymbol{X}_1', \boldsymbol{X}_2 \setminus \boldsymbol{X}_2')$$

and the first term vanishes due to independence and (3.1c), and the second term vanishes due to Lemma A.1 because the argument of κ splits into two independent groups.

Appendix C. Modifications for complex Hermitian W

Our main arguments were carried out for the real symmetric case. We now explain how to modify our proofs if W is complex Hermitian. A quick inspection of the proofs shows that the only modification concerns Proposition 3.2 where we have to replace the cumulant expansion by its complex variant. We reduce the problem to the real case by considering real and imaginary parts of each variable separately. Another option would have been to consider w and $\overline{w}$ independent variables, but our choice seems to require the least modifications. In order to compute $\mathbf{E}\, w_{i_0} f(\boldsymbol{w})$ for a random vector $\boldsymbol{w} \in \mathbb{C}^{\mathcal{I}}$, $w_{i_0} \in \mathbb{C}$ and a function $f\colon \mathbb{C}^{\mathcal{I}} \to \mathbb{C}$, we can define $\widetilde{f}\colon \mathbb{R}^{\mathcal{I}\sqcup\mathcal{I}} \to \mathbb{C}$ by mapping $(\boldsymbol{w}^{\Re}, \boldsymbol{w}^{\Im}) \mapsto f(\boldsymbol{w}^{\Re} + i\boldsymbol{w}^{\Im})$, where the new index set $\mathcal{I}\sqcup\mathcal{I}$ should be understood as two copies of $\mathcal{I}$ in the sense that $\mathcal{I}\sqcup\mathcal{I} = \{\,(i,\Re),(i,\Im) \mid i\in\mathcal{I}\,\}$. If we want to expand $w_{i_0} f(\boldsymbol{w})$ in the variables of some fixed index set $\mathcal{N} \subset \mathcal{I}$, we separately apply Proposition 3.2 to $\mathbf{E}\, \widetilde{w}_{(i_0,\Re)} \widetilde{f}(\widetilde{\boldsymbol{w}})$ and $\mathbf{E}\, \widetilde{w}_{(i_0,\Im)} \widetilde{f}(\widetilde{\boldsymbol{w}})$ in $\mathcal{N}\sqcup\mathcal{N}$, where $\widetilde{\boldsymbol{w}} = (\Re\boldsymbol{w}, \Im\boldsymbol{w})$ and $\widetilde{w}_{(i,\Re)} = \Re w_i$, $\widetilde{w}_{(i,\Im)} = \Im w_i$. It follows that

$$\mathbf{E}\, w_{i_0} \widetilde{f}(\widetilde{\boldsymbol{w}}) = \sum_{l>0} \sum_{\widetilde{\boldsymbol{i}}\in(\mathcal{N}\sqcup\mathcal{N})^l} \frac{\kappa(\widetilde{w}_{(i_0,\Re)}, \widetilde{\boldsymbol{w}}_{\widetilde{\boldsymbol{i}}}) + \kappa(\mathrm{i}\widetilde{w}_{(i_0,\Im)}, \widetilde{\boldsymbol{w}}_{\widetilde{\boldsymbol{i}}})}{l!} \partial_{\widetilde{\boldsymbol{i}}}(\mathbf{E}\, \widetilde{f}) + \widetilde{\Omega}^1 + \widetilde{\Omega}^2, \tag{C.1}$$

where the error terms are those from two applications of (3.7a). We note that we can make sense of κ with complex arguments directly through Definition (A.4). We now want to go back to a summation over our initial index set $\mathcal{N}$ and therefore regroup the terms in (C.1) according to the first indices of $\widetilde{\boldsymbol{i}}$. To formulate the result compactly we introduce the tensors

$$\widetilde{\kappa}(w_{i_0},\dots,w_{i_l}) \coloneqq \kappa\Big[\begin{pmatrix} \Re w_{i_0} \\ \mathrm{i}\Im w_{i_0} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} \Re w_{i_l} \\ \mathrm{i}\Im w_{i_l} \end{pmatrix} \Big] \in (\mathbb{R}\times\mathrm{i}\mathbb{R})^{\otimes(l+1)} \quad \text{and} \quad \widetilde{\partial}_{\boldsymbol{i}} \coloneqq \begin{pmatrix} \partial_{\Re w_{i_1}} \\ \partial_{\Im w_{i_1}} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} \partial_{\Re w_{i_l}} \\ \partial_{\Im w_{i_l}} \end{pmatrix},$$

where the application of κ is understood in an entrywise sense and the derivative tensor has dimension $(\mathbb{C}^2)^{\otimes l}$. By saying that κ is understood in an entrywise sense, we mean, by slight abuse of notation that, for example,

$$\kappa\left(\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \otimes \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \right) = \kappa\Big(\sum_{i,j=1}^{2} v_i w_j\, e_i \otimes e_j \Big) \coloneqq \sum_{i,j=1}^{2} \kappa(v_i, w_j)\, e_i \otimes e_j,$$

where e_1, e_2 is the standard basis of $\mathbb{R}\times\mathrm{i}\mathbb{R}$. Due to the special nature of the index i_0 we see from (C.1) that $\Re w_{i_0}$ and $\mathrm{i}\Im w_{i_0}$ always occur in a sum of two and the rhs. of (C.1) can be expressed in terms of the partial trace $\operatorname{Tr}_1 \widetilde{\kappa}(w_{i_0},\dots,w_{i_l}) \in (\mathbb{R}\times i\mathbb{R})^{\otimes l}$ along the first dimension, which corresponds to i_0. Thus we can compactly write (C.1) as

$$\mathbf{E}\, w_{i_0} f(\boldsymbol{w}) = \sum_{0\le l<R} \sum_{\boldsymbol{i}\in\mathcal{N}^l} \frac{\langle \operatorname{Tr}_1 \widetilde{\kappa}(w_{i_0}, \boldsymbol{w}_{\boldsymbol{i}}), \mathbf{E}(\widetilde{\partial}_{\boldsymbol{i}} f)\rangle}{l!} + \widetilde{\Omega}^1 + \widetilde{\Omega}^2, \tag{C.2}$$

where the scalar product is taken between two tensors of size 2^l. For example, the $l=1$ term from (C.2) reads

$$\sum_{i_1\in\mathcal{N}} \left(\frac{\kappa(\Re w_{i_0}, \Re w_{i_1}) + \kappa(\mathrm{i}\Im w_{i_0}, \Re w_{i_1})}{1!} (\mathbf{E}\, \partial_{\Re w_{i_1}} f) + \frac{\kappa(\Re w_{i_0}, \mathrm{i}\Im w_{i_1}) + \kappa(\mathrm{i}\Im w_{i_0}, \mathrm{i}\Im w_{i_1})}{1!} (\mathbf{E}\, \partial_{\Im w_{i_1}} f) \right).$$

The rest of the argument in Section 4 can be carried out verbatim for any specific choice of distribution of $\Re$, $\Im$ to the entries of κ. We only have to replace the norms $|||\kappa|||^{\mathrm{av}}$ and $|||\kappa|||^{\mathrm{iso}}$ in Assumption (C) by applying them entrywise to $\widetilde{\kappa}$, i.e.,

$$|||\widetilde{\kappa}(w_{\alpha_1},\dots,w_{\alpha_k})|||^{\mathrm{av}} \coloneqq \sum_{\mathfrak{x}_1,\dots,\mathfrak{x}_k\in\{\Re,\Im\}} |||\kappa(\mathfrak{x}_1 w_{\alpha_1},\dots,\mathfrak{x}_k w_{\alpha_k})|||^{\mathrm{av}}, \tag{C.3a}$$

$$|||\widetilde{\kappa}(w_{\alpha_1},\dots,w_{\alpha_k})|||^{\mathrm{iso}} \coloneqq \sum_{\mathfrak{x}_1,\dots,\mathfrak{x}_k\in\{\Re,\Im\}} |||\kappa(\mathfrak{x}_1 w_{\alpha_1},\dots,\mathfrak{x}_k w_{\alpha_k})|||^{\mathrm{iso}}. \tag{C.3b}$$

Assumption (C)' (Hermitian κ-correlation decay)**.** *We assume that for all $R\in\mathbb{N}$ and $\epsilon>0$*

$$|||\widetilde{\kappa}|||^{av} \le_{\epsilon,R} N^{\epsilon} \quad \textit{and} \quad |||\widetilde{\kappa}|||^{iso} \le_{\epsilon,R} N^{\epsilon}.$$

Since there are at most 2^R such choices this change has no impact on any of the claimed bounds which always implicitly allow for an R–dependent constant.

Appendix D. Proofs of auxiliary results

Lemma D.1 (Quadratic Implicit Function Theorem)**.** *Let $\|\cdot\|$ be a norm on $\mathbb{C}^d$, $A, B \in \mathbb{C}^d$ and $Q\colon \mathbb{C}^d\times\mathbb{C}^d \to \mathbb{C}^d$ a bounded $\mathbb{C}^d$-valued quadratic form, i.e.,*

$$\|Q\| = \sup_{x,y} \frac{\|Q(x,y)\|}{\|x\|\,\|y\|} < \infty.$$

Suppose that A is invertible. Then for $\epsilon_2 \coloneqq \big[2\,\|A^{-1}\|\,\|Q\|\,\big]^{-1}$ and $\epsilon_1 \coloneqq \epsilon_2\big[2\,\|A^{-1}\|\,\|B\|\,\big]^{-1}$ there is a unique function $X\colon B_{\epsilon_1} \to B_{\epsilon_2}$ such that

$$AX(d) + Q(X(d), X(d)) = Bd,$$

where B_ϵ denotes the open ϵ-ball around 0. Moreover, the function X is analytic and satisfies

$$\|X(d_1) - X(d_2)\| \le 2\,\|A^{-1}\|\,\|B\|\,\|d_1 - d_2\| \quad \textit{for all} \quad d_1, d_2 \in B_{\epsilon_1/2}.$$

Proof. A simple application of the Banach fixed point theorem. □

Lemma D.2. *For random matrices R, T and $p \geq 1$ it holds that $\|\mathcal{S}[V]T\|_p \leq |||\mathcal{S}||| \, \|V\|_{2p} \, \|T\|_{2p}$.*

Proof. Let $\kappa = \kappa_c + \kappa_d$ be an arbitrary partition, which induces a partition of $\mathcal{S}$ since

$$\mathcal{S}[V] = \frac{1}{N} \sum_{\alpha_1, \alpha_2} \kappa(\alpha_1, \alpha_2) \Delta^{\alpha_1} V \Delta^{\alpha_2}.$$

For vectors $\mathbf{x}, \mathbf{y}$ with $\|\mathbf{x}\|, \|\mathbf{y}\| \leq 1$ we compute

$$\begin{aligned} \|(\mathcal{S}[V]T)\mathbf{x}\mathbf{y}\|_p &= \Big\| \frac{1}{N} \sum_{b_1, a_2, b_2} \kappa(\mathbf{x}b_1, a_2 b_2) V_{b_1 a_2} T_{b_2 \mathbf{y}} \Big\|_p \leq \Big\| \frac{1}{N} \sum_{b_1, b_2} V_{b_1 \kappa_c(\mathbf{x}b_1, \cdot b_2)} T_{b_2 \mathbf{y}} \Big\|_p + \Big\| \frac{1}{N} \sum_{b_1, a_2} R_{b_1 a_2} T_{\kappa_d(\mathbf{x}b_1, a_2 \cdot)\mathbf{y}} \Big\|_p \\ &\leq \frac{\|V\|_{2p} \, \|T\|_{2p}}{N} \Big[\sum_{b_1, b_2} \|\kappa_d(\mathbf{x}b_1, \cdot b_2)\| + \sum_{b_1, a_2} \|\kappa_c(\mathbf{x}b_1, a_2 \cdot)\| \Big] \leq \Big[|||\kappa_d|||_d + |||\kappa_c|||_c \Big] \|V\|_{2p} \, \|T\|_{2p} \end{aligned}$$

and the result follows from optimizing over the decompositions of κ and recalling the definition (4.2). □

Lemma D.3. *For any $t \in [0,1]$, $q \geq 1$, $\epsilon > 0$ and multi-set $\underline{\beta} \subset I$ we have under Assumption (A) that*

$$\big\| \partial_{\underline{\beta}} G \big|_{\widehat{W}} \big\|_q \leq_{|\underline{\beta}|} N^{-|\underline{\beta}|/2} N^{\epsilon} \langle z \rangle^{-1} \Big(1 + \|G\|_{2q(|\underline{\beta}|+1)} \Big)^{(|\underline{\beta}|+1)/\mu} \tag{D.1a}$$

$$\big\| \partial_{\underline{\beta}} D \big|_{\widehat{W}} \big\|_q \leq_{|\underline{\beta}|} N^{-|\underline{\beta}|/2} (1 + |||\mathcal{S}|||) N^{\epsilon} \langle z \rangle^{-1} \Big(1 + \|G\|_{6q(|\underline{\beta}|+2)} \Big)^{(|\underline{\beta}|+2)/\mu}, \tag{D.1b}$$

where $\widehat{W}_\alpha = t w_\alpha$ for $\alpha \in \mathcal{N}$ and $\widehat{W}_\alpha = w_\alpha$ otherwise for a set $\mathcal{N} \subset I$ of size $|\mathcal{N}| \leq N^{1/2-\mu}$.

Proof. We write $\underline{\beta} = \{\beta_1, \ldots, \beta_n\}$ and its easy to see inductively that

$$\partial_{\underline{\beta}} G \big|_{\widehat{W}} = \frac{(-1)^n}{N^{n/2}} \sum_{\sigma \in S_n} \widehat{G} \Delta^{\beta_{\sigma(1)}} \widehat{G} \Delta^{\beta_{\sigma(2)}} \widehat{G} \ldots \widehat{G} \Delta^{\beta_{\sigma(n)}} \widehat{G}, \tag{D.2}$$

where $\widehat{G} = G(\widehat{W})$. From the resolvent identity it follows that

$$\widehat{G} - G = \sum_{k=1}^{R-1} G \Big(\frac{W - \widehat{W}}{\sqrt{N}} G \Big)^k + \widehat{G} \Big(\frac{W - \widehat{W}}{\sqrt{N}} G \Big)^R$$

and therefore by the trivial bound $\|\widehat{G}\| \leq 1/\eta$ and Assumption (B) it follows that for $R := \lceil 1/\mu \rceil$ and $\eta \geq N^{-1}$ we have

$$\big\| \widehat{G} - G \big\|_q \leq \sum_{k=1}^{R-1} \frac{|\mathcal{N}|^k \, \|G\|_{(2k+1)q}^{k+1} \max_\alpha \|w_\alpha\|_{(2k+1)q}^k}{N^{k/2}} + \frac{|\mathcal{N}|^R \, \|G\|_{2Rq}^R \max_\alpha \|w_\alpha\|_{(2R+1)q}^R}{N^{R/2} \eta} \leq_q \|G\|_{2Rq} \, (1 + \|G\|_{2Rq})^R. \tag{D.3}$$

Since $\|H\| \leq N^{\epsilon/2}$ with very high probability for sufficiently large N it follows that $\|G\|_r \leq \|G\| \lesssim N^{\epsilon/2} / \langle z \rangle$ for $|z| \gg N^{\epsilon/2}$ which immediately implies (D.1a).

Similarly, (D.1b) follows from the easily verifiable identities

$$\partial_{\underline{\beta}} D \big|_{\widehat{W}} = \frac{(-1)^n}{N^{n/2}} \sum_{\sigma \in S_n} \Big[\widehat{D} \Delta^{\beta_{\sigma(1)}} \widehat{G} \ldots \Delta^{\beta_{\sigma(n)}} \widehat{G} + \sum_{k=1}^{n} \mathcal{S}[\widehat{G} \Delta^{\beta_{\sigma(1)}} \widehat{G} \ldots \Delta^{\beta_{\sigma(k)}} \widehat{G}] \widehat{G} \Delta^{\beta_{\sigma(k+1)}} \widehat{G} \ldots \Delta^{\beta_{\sigma(n)}} \widehat{G} \Big] \tag{D.4}$$

and

$$\widehat{D} = D - \mathcal{S}[G]G + (D - \mathcal{S}[G]G) \frac{W - \widehat{W}}{\sqrt{N}} \widehat{G} + \frac{\widehat{W} - W}{\sqrt{N}} \widehat{G} + \mathcal{S}[\widehat{G}]\widehat{G} \tag{D.5}$$

together with Lemma D.2, (D.1a), (D.3) and

$$\|D\|_r \leq_r N^{\epsilon} \langle z \rangle^{-1} (1 + \|G\|_r) + \|\mathcal{S}[G]G\|_r. \tag{D.6}$$

To see why (D.6) holds we write $D = HG - AG + \mathcal{S}[G]G$ and use $\|AG\|_r \lesssim \|G\|_r$ while $\|HG\|_r = \|1 + zG\|_r \lesssim 1 + N^{\epsilon/2} \|G\|_r$ for $|z| \lesssim N^{\epsilon/2}$. For large $|z| \gg N^{\epsilon/2}$ we estimate that $\|HG\|_r \leq \|HG\| \leq N^{\epsilon} \langle z \rangle^{-1}$ since $\|H\| \leq N^{\epsilon/2}$. □

References

[1] O. Ajanki, L. Erdős, and T. Krüger, *Stability of the matrix Dyson equation and random matrices with correlations*, to appear in Probab. Theory Related Fields (2016), arXiv:1604.08188.

[2] O. H. Ajanki, L. Erdős, and T. Krüger, *Local spectral statistics of Gaussian matrices with correlated entries*, J. Stat. Phys. **163**, 280–302 (2016), MR3478311.

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

[4] J. Alt, *The local semicircle law for random matrices with a fourfold symmetry*, J. Math. Phys. **56**, 103301, 2 0 (2015), MR3406427.

[5] J. Alt, L. Erdős, T. Krüger, and Y. Nemish, *Location of the spectrum of Kronecker random matrices*, to appear in Ann. Inst. Henri Poincaré Probab. Stat. (2017), arXiv:1706.08343.

[6] G. W. Anderson and O. Zeitouni, *A law of large numbers for finite-range dependent random matrices*, Comm. Pure Appl. Math. **61**, 1118–1154 (2008), MR2417889.

[7] Z. D. Bai and Y. Q. Yin, *Convergence to the semicircle law*, Ann. Probab. **16**, 863–875 (1988), MR929083.

[8] M. Banna, F. Merlevède, and M. Peligrad, *On the limiting spectral distribution for a large class of symmetric random matrices with correlated entries*, Stochastic Process. Appl. **125**, 2700–2726 (2015), MR3332852.

[9] A. Boutet de Monvel, A. Khorunzhy, and V. Vasilchuk, *Limiting eigenvalue distribution of random matrices with correlated entries*, Markov Process. Related Fields **2**, 607–636 (1996), MR1431189.

[10] R. C. Bradley, *Basic properties of strong mixing conditions. A survey and some open questions*, Probab. Surv. **2**, Update of, and a supplement to, the 1986 original, 107–144 (2005), MR2178042.

[11] Z. Che, *Universality of random matrices with correlated entries*, Electron. J. Probab. **22**, Paper No. 30, 38 (2017), MR3629874.

[12] M. Duneau, D. Iagolnitzer, and B. Souillard, *Decrease properties of truncated correlation functions and analyticity properties for classical lattices and continuous systems*, Comm. Math. Phys. **31**, 191–208 (1973), MR0337229.

[13] 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.

[14] L. Erdős, S. Péché, J. A. Ramírez, B. Schlein, and H.-T. Yau, *Bulk universality for Wigner matrices*, Comm. Pure Appl. Math. **63**, 895–925 (2010), MR2662426.

[15] L. Erdős, B. Schlein, and H.-T. Yau, *Universality of random matrices and local relaxation flow*, Invent. Math. **185**, 75–119 (2011), MR2810797.
[16] 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.
[17] L. Erdős and K. Schnelli, *Universality for random matrix flows with time-dependent density*, Ann. Inst. Henri Poincaré Probab. Stat. **53**, 1606–1656 (2017), MR3729630.
[18] L. Erdős and D. Schröder, *Fluctuations of rectangular Young diagrams of interlacing Wigner eigenvalues*, Int. Math. Res. Not. IMRN, 3255–3298 (2018), MR3805203.
[19] L. Erdős and H.-T. Yau, *Dynamical approach to random matrix theory* (To appear in: Courant Lecture Notes, American Mathematical Society, 2017).
[20] L. Erdős and H.-T. Yau, *Universality of local spectral statistics of random matrices*, Bull. Amer. Math. Soc. (N.S.) **49**, 377–414 (2012), MR2917064.
[21] L. Erdős, H.-T. Yau, and J. Yin, *Bulk universality for generalized Wigner matrices*, Probab. Theory Related Fields **154**, 341–407 (2012), MR2981427.
[22] L. Erdős, H.-T. Yau, and J. Yin, *Universality for generalized Wigner matrices with Bernoulli distribution*, J. Comb. **2**, 15–81 (2011), MR2847916.
[23] L. Giraitis and D. Surgailis, "Multivariate Appell polynomials and the central limit theorem", *Dependence in probability and statistics*, Vol. 11, edited by E. Eberlein and M. S. Taqqu, Progress in Probability and Statistics, A survey of recent results, Papers from the conference held at the Mathematical Research Institute Oberwolfach, Oberwolfach, April 1985 (Birkhäuser Boston, Inc., Boston, MA, 1986), pp. 21–71, MR899982.
[24] V. L. Girko, *Asymptotics of the distribution of the spectrum of random matrices*, Uspekhi Mat. Nauk **44**, 7–34, 256 (1989), MR1023102.
[25] V. L. Girko, *Theory of stochastic canonical equations. Vol. I*, Vol. 535, Mathematics and its Applications (Kluwer Academic Publishers, Dordrecht, 2001), pp. xxiv+497, MR1887675.
[26] W. Hachem, P. Loubaton, and J. Najim, *The empirical eigenvalue distribution of a gram matrix: from independence to stationarity*, Markov Process. Related Fields **11**, 629–648 (2005), MR2191967.
[27] Y. He and A. Knowles, *Mesoscopic eigenvalue statistics of Wigner matrices*, Ann. Appl. Probab. **27**, 1510–1550 (2017), MR3678478.
[28] Y. He, A. Knowles, and R. Rosenthal, *Isotropic self-consistent equations for mean-field random matrices*, Probab. Theory Related Fields **171**, 203–249 (2018), MR3800833.
[29] J. W. Helton, R. Rashidi Far, and R. Speicher, *Operator-valued semicircular elements: solving a quadratic matrix equation with positivity constraints*, Int. Math. Res. Not. IMRN, Art. ID rnm086, 15 (2007), MR2376207.
[30] 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.
[31] B. Landon, P. Sosoe, and H.-T. Yau, *Fixed energy universality for Dyson Brownian motion*, preprint (2016), arXiv:1609.09011.
[32] B. Landon and H.-T. Yau, *Convergence of local statistics of Dyson Brownian motion*, Comm. Math. Phys. **355**, 949–1000 (2017), MR3687212.
[33] J. O. Lee and K. Schnelli, *Edge universality for deformed Wigner matrices*, Rev. Math. Phys. **27**, 1550018, 94 (2015), MR3405746.
[34] J. O. Lee, K. Schnelli, B. Stetler, and H.-T. Yau, *Bulk universality for deformed Wigner matrices*, Ann. Probab. **44**, 2349–2425 (2016), MR3502606.
[35] P. McCullagh, *Tensor notation and cumulants of polynomials*, Biometrika **71**, 461–476 (1984), MR775392.
[36] M. L. Mehta, *Random matrices and the statistical theory of energy levels* (Academic Press, New York-London, 1967), pp. x+259, MR0220494.
[37] S. O'Rourke and V. Vu, *Universality of local eigenvalue statistics in random matrices with external source*, Random Matrices Theory Appl. **3**, 1450005, 37 (2014), MR3208886.
[38] L. A. Pastur, *Spectra of random selfadjoint operators*, Uspehi Mat. Nauk **28**, 3–64 (1973), MR0406251.
[39] R. Rashidi Far, T. Oraby, W. Bryc, and R. Speicher, *On slow-fading MIMO systems with nonseparable correlation*, IEEE Trans. Inform. Theory **54**, 544–553 (2008), MR2444540.
[40] J. H. Schenker and H. Schulz-Baldes, *Semicircle law and freeness for random matrices with symmetries or correlations*, Math. Res. Lett. **12**, 531–542 (2005), MR2155229.
[41] T. P. Speed, *Cumulants and partition lattices*, Austral. J. Statist. **25**, 378–388 (1983), MR725217.
[42] T. Tao and V. Vu, *Random matrices: universality of local eigenvalue statistics*, Acta Math. **206**, 127–204 (2011), MR2784665.