

Isotropic self-consistent equations for mean-field random matrices

Yukun He*

Antti Knowles†

Ron Rosenthal‡

April 12, 2017

We present a simple and versatile method for deriving (an)isotropic local laws for general random matrices constructed from independent random variables. Our method is applicable to mean-field random matrices, where all independent variables have comparable variances. It is entirely insensitive to the expectation of the matrix. In this paper we focus on the probabilistic part of the proof – the derivation of the self-consistent equations. As a concrete application, we settle in complete generality the local law for Wigner matrices with arbitrary expectation.

1. Introduction

1.1. Overview. The empirical eigenvalue measure of a large random matrix is typically well approximated by a deterministic asymptotic measure. For instance, for Wigner matrices this measure is the celebrated Wigner semicircle law [40]. This approximation is best formulated using the Green function and Stieltjes transforms. Let W be an $N \times N$ Hermitian random matrix normalized so that its typical eigenvalue spacing is of order N^{-1} , and denote by

$$G(z) := (W - z)^{-1}$$

the associated Green function. Here $z = E + i\eta$ is a spectral parameter with positive imaginary part η . Then the Stieltjes transform of the empirical eigenvalue measure is equal to $N^{-1} \operatorname{Tr} G(z)$, and the approximation mentioned above may be written informally as

$$\frac{1}{N} \operatorname{Tr} G(z) \approx m(z) \tag{1.1}$$

for large N and with high probability. Here $m(z)$ is the Stieltjes transform of the asymptotic eigenvalue measure, which we denote by ϱ . We call an estimate of the form (1.1) an *averaged law*.

As may be easily seen by taking the imaginary part of (1.1), control of the convergence of $N^{-1} \operatorname{Tr} G(z)$ yields control of an order ηN eigenvalues around the point E . A *local law* is an estimate of the form (1.1) for all $\eta \gg N^{-1}$. Note that the approximation (1.1) cannot be correct at or below the scale $\eta \asymp N^{-1}$, at which the behaviour of the left-hand side of (1.1) is governed by the fluctuations of individual eigenvalues.

*University of Geneva, Section of Mathematics, yukun.he@unige.ch.

†University of Geneva, Section of Mathematics, antti.knowles@unige.ch.

‡Technion - Israel Institute of Technology, Department of Mathematics, ron.ro@tx.technion.ac.il.

Local laws have become a cornerstone of random matrix theory, starting from the work [22] where a local law was first established for Wigner matrices. Local laws constitute the main tool needed to analyse the distribution of the eigenvalues of random matrices, including the universality of the local spectral statistics and the proof of the celebrated Wigner-Dyson-Mehta conjecture [36], and the eigenvectors of random matrices, including eigenvector delocalization and the distribution of the eigenvector components.

In fact, for the aforementioned applications to the distribution of eigenvalues and eigenvectors, the averaged local law from (1.1) is not sufficient. One has to control not only the normalized trace of G but the matrix G itself, by showing that G is close to some deterministic matrix, M , provided that $\eta \gg N^{-1}$. As the dimension N of these matrices is very large, which notion of closeness to use is a nontrivial question. The control of $G - M$ as a matrix was initiated in [24] in the context of Wigner matrices, where the individual matrix entries $G_{ij} - M_{ij}$ are controlled. We call such a result an *entrywise local law*. Subsequently, the entrywise local law of [24] has been refined and significantly generalized [1–5, 7, 9, 11, 15, 19–21, 23, 25, 30, 31, 33–35, 38].

More generally, canonical notions are closeness in the *weak operator sense*, *strong operator sense*, and *norm sense*, which entail control of

$$\langle \mathbf{w}, (G - M)\mathbf{v} \rangle, \quad |(G - M)\mathbf{v}|, \quad \|G - M\|$$

respectively, for deterministic $\mathbf{v}, \mathbf{w} \in \mathbb{C}^N$. It is easy to see that already convergence in the strong operator sense must fail, as one typically has $|(G - M)\mathbf{v}| \asymp \frac{1}{\sqrt{N}\eta}$; we refer to [10, Section 12.1] for more details. We conclude that control in the weak operator sense, i.e. of $\langle \mathbf{w}, (G - M)\mathbf{v} \rangle$, is the strongest form of control that one may expect to hold for the error $G - M$. Following [11, 30, 31] we call such a result an *(an)isotropic local law*. (An)isotropic local laws have played a central role in the study of deformed matrix ensembles [11, 12, 30, 31] and the distribution of eigenvector components of random matrices [12–14].

All proofs of local laws consist of at least the two following major steps, which are completely decoupled.

- (A) *A stochastic step* establishing a self-consistent equation for G with a random error term.
- (B) *A deterministic step* analysing the stability of the self-consistent equation.

Step (A) may be formulated in general as follows. There is a map $\Pi : \mathbb{C}^{N \times N} \rightarrow \mathbb{C}^{N \times N}$ such that M is uniquely characterized as the solution of the equation $\Pi(M) = 0$ with positive imaginary part. For a large class of random matrices, including those considered in this paper, Π is a nonlinear quadratic function. Step (A) provides high-probability bounds on the matrix $\Pi(G)$ (in some suitable sense).

In Step (B) one shows that if $\Pi(G) = \Pi(G) - \Pi(M)$ is small then so is $G - M$.

In general, both steps are nontrivial. In addition, steps (A) and (B) have up to now only been used to derive an entrywise local law; to obtain a full (an)isotropic local law, a further nontrivial argument was necessary; three different techniques for deriving (an)isotropic laws from entrywise laws were developed in [11, 30, 31].

In this paper we develop a new method for dealing with Step (A), which has the following advantages as compared to previous approaches.

- (i) It can handle very general mean-field random matrices W , and is in particular completely insensitive to the expectation $\mathbb{E}W$. (Previous methods heavily rely on the assumption that $\mathbb{E}W$ is diagonal or close to diagonal.)

- (ii) It automatically yields the (an)isotropic law in one step, without requiring a further intermediate step via an entrywise law, as in [11, 30, 31].
- (iii) It is simple and very versatile. It is applicable to many matrix models where previous methods either fail or are very cumbersome.

For conciseness, in this paper we focus on mean-field Hermitian random matrices whose upper-triangular entries are independent. By *mean-field* we mean that $\text{Var}(W_{ij}) = O(N^{-1})$ for all i, j . We make no assumption on $\mathbb{E}W$. We remark that our method extends to very general matrices built from independent random variables, such as sample covariance matrices built from populations with arbitrary expectations and covariances.

We now outline our main results. Throughout the paper we split

$$W = H + A, \quad A := \mathbb{E}W.$$

We define the map

$$\Pi(M) := I + zM + \mathcal{S}(M)M - AM, \quad \mathcal{S}(M) := \mathbb{E}[HMH]. \quad (1.2)$$

Then it is not hard to show that for $z \in \mathbb{C}_+$, the equation $\Pi(M) = 0$ has a unique solution M with positive imaginary part (see Lemma 1.4 below). Our main result (Theorem 1.5 below) is the derivation of the self-consistent equation $\Pi(G) \approx 0$: we establish high-probability bounds on the quantities

$$\langle \mathbf{v}, \Pi(G)\mathbf{w} \rangle, \quad \text{Tr}[B\Pi(G)],$$

where B is a bounded deterministic matrix and $\mathbf{v}, \mathbf{w} \in \mathbb{C}^N$ are deterministic. These bounds are sharp (up to some technicalities in the definition of a high probability bound) throughout the spectrum, in both the bulk and in the vicinity of regular edges where the density of the limiting measure ϱ exhibits the characteristic square-root decay of random matrix theory. We emphasize that we make no assumption on the expectation $A = \mathbb{E}W$.

Our main result performs Step (A) for the general class of mean-field random matrices introduced above. To illustrate how this result may be combined with Step (B) to complete the proof of the (an)isotropic local law, we use our main result to settle in complete generality the local law for Wigner matrices with arbitrary expectation (Theorem 1.9 below), i.e. $\mathbb{E}|H_{ij}|^2 = N^{-1}$ with arbitrary expectation A . Previously, the local law for Wigner matrices with nonzero expectation was derived under the assumption that A is diagonal [35] or that all third moments of H vanish [30]. Even the averaged local law for general A was not accessible using previous methods.

In this paper we do not perform Step (B) (the stability analysis of the self-consistent equation $\Pi(G) \approx 0$) for the aforementioned general class of mean-field matrices. For general W this is a nontrivial task, and we refer to the works [1–5], where a related analysis is performed to obtain an entrywise law for the case of almost diagonal A .

We remark that the map Π defined in (1.2) has been extensively studied in the random matrix literature, starting with [26, 37], where its stability was analysed on the global spectral scale. Recently, it has been the focus of the seminal series of works [1–5] mentioned above, where the stability of the matrix equation $\Pi(M) = 0$ was analysed in great detail in order to establish local laws for a general class of random matrix ensembles.

We conclude this introductory section with a few words about the proof. Traditionally, in all of the works [1–5, 7, 11, 15, 19–21, 23–25, 30, 31, 34, 35, 38], the proof of Step (A) for matrices with

independent entries relied on Schur’s complement formula. When applicable, Schur’s complement formula provides a direct approach to proving Step (A). However, its applicability to random matrix models is heavily dependent on the independence of the entries of H and on the fact that A is diagonal. If the mapping between the independent random variables and the entries of W is nontrivial or if A is not diagonal, the use of Schur’s complement formula becomes more cumbersome and possibly even fruitless. Moreover, Schur’s complement formula is only effective when deriving entrywise estimates, and it is ill-suited for deriving isotropic estimates. The first proof of a local law without using Schur’s complement formula was [9], where a local law was established for random regular graphs (which are not built from independent random variables). In [9], Schur’s complement formula was replaced by a much more robust expansion using the resolvent identity. Our current approach is in part motivated by this philosophy: instead of working with entire rows and columns, as dictated by Schur’s complement formula, we work with individual matrix entries. A natural way of achieving this is to perform a series of expansions using the resolvent identity, where G is expanded in a single entry of H . We remark that such ideas have previously been used in the somewhat different context of proving the so-called fluctuation averaging result in random matrix theory; see [25].

In practice, we choose to replace the resolvent expansion with a cumulant expansion inspired by [27, 29, 32] (see Lemma 2.4 below), which effectively performs the same steps but in a more streamlined fashion. A drawback of the cumulant expansion is that it requires working with expectations of random variables instead of just random variables. This leads us to consider high moments of error terms, which are estimated self-consistently. An analogous strategy was first used in [31] to prove the isotropic local law for Wigner matrices, although there the moments were estimated by the Green function comparison strategy instead of the cumulant expansion used in this paper. The first use of recursive self-consistent estimates for high moments using the cumulant expansion, including exploiting a cancellation among two leading terms, was [27]; this strategy constitutes the core of our method. The cumulant expansion was first used to derive a local law in the recent work [33], where a precise averaged law was derived for sparse random matrices with $A = 0$. Analogous ideas for the Haar measure on the unitary group were recently exploited in [8] to obtain optimal bounds on the convergence rate for the spectral distribution of the sum of random matrices. Historically, the first use of the cumulant expansion for local spectral statistics is [32], where edge universality and the Tracy-Widom limit for deformed Wigner matrices was proved, along with a streamlined proof of the edge universality for Wigner matrices.

Aside from the novel idea of using the resolvent/cumulant expansion to prove (an)isotropic local laws for random matrices with independent entries, the key new ingredients of our method may be summarized as follows.

- (i) We bootstrap isotropic bounds on the Green function, i.e. our proof makes use of a priori bounds of the form $\langle \mathbf{v}, G\mathbf{w} \rangle \lesssim |\mathbf{v}||\mathbf{w}|$ with high probability. When using the cumulant expansion (or, alternatively, the resolvent expansion), we are naturally led to sums of the form $\sum_i v_i G_{ij} \cdots$. It is crucial to first perform the sum over i to exploit the isotropic a priori bounds on G ; simply estimating the sum term by term leads to errors which are not affordable. Reducing such sums to isotropic estimates on the Green function requires a certain amount of care.
- (ii) A direct application of the cumulant expansion to estimating high moments of the error results in terms that cannot be controlled with a high enough precision. To circumvent this

issue, we use multiple self-improving matrix bounds at several stages of the proof. We assume a rough a priori bound on an error matrix and derive a self-improving bound on it, which may be iterated to obtain the optimal bound. See e.g. (3.26) or (3.61) below for an example. For the estimate of $\text{Tr}[B\Pi(G)]$, we apply this approach to a matrix that we call Q , defined in (3.48) below. The identification of Q as a central object of the proof and the derivation of optimal bounds on it by means of a self-improving scheme obtained from a second cumulant expansion is a key idea of our proof.

Conventions. The central objects of this paper, such as the Green function G , depend on $N \in \mathbb{N}$ and the spectral parameter $z \in \mathbb{C}_+ := \{z \in \mathbb{C} : \text{Im } z > 0\}$. We frequently omit both arguments, writing for instance G instead of $G_N(z)$. If some quantity does not depend on N , we indicate this explicitly by calling it *constant* or *fixed*. We always use the notation $z = E + i\eta$ for the real and imaginary parts of z . We use the usual $O(\cdot)$ notation, and write $O_\alpha(\cdot)$ if the implicit constant depends on the parameter α . The parameter α can never depend on N or z .

We use the notation $\mathbf{v} = (v_i)_{1 \leq i \leq N} \in \mathbb{C}^N$ for vectors, $\langle \cdot, \cdot \rangle$ for the scalar product $\langle \mathbf{v}, \mathbf{w} \rangle := \sum_{i=1}^N \bar{v}_i w_i$ and $|\mathbf{v}| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$ for the Euclidean norm. We denote by $\|B\|$ the Euclidean operator norm of an $N \times N$ matrix B . We write $\underline{B} := \frac{1}{N} \text{Tr } B$ for the normalized trace of B .

Acknowledgements. We thank Torben Krüger for drawing our attention to the importance of the equation (1.3) for general mean-field matrix models. In a private communication, Torben Krüger also informed us that the idea of organizing proofs of local laws by bootstrapping is being developed independently for random matrices with general correlated entries in [17].

1.2. Basic definitions. We consider $N \times N$ matrices of the form $W = H + A \in \mathbb{C}^{N \times N}$, where $H = H^*$ is random and $A = A^*$ is deterministic. We always make the following assumption on H .

Assumption 1.1. The upper-triangular entries $(H_{ij} : 1 \leq i \leq j \leq N)$ are independent mean-zero random variables satisfying $E[|\sqrt{N}H_{ij}|^p] = O_p(1)$ for all i, j and $p \in \mathbb{N}$.

Let $z = E + i\eta \in \mathbb{C}_+$ be a spectral parameter, and define the Green function

$$G(z) := (H + A - z)^{-1}.$$

We use the following notion of high-probability bounds, the first version of which appeared in [18].

Definition 1.2 (Stochastic domination).

(i) Let

$$X = (X^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}), \quad Y = (Y^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)})$$

be two families of nonnegative random variables, where $U^{(N)}$ is a possibly N -dependent parameter set.

We say that X is *stochastically dominated by Y , uniformly in u* , if for all (small) $\varepsilon > 0$ and (large) $D > 0$ we have

$$\sup_{u \in U^{(N)}} \mathbb{P}\left(X^{(N)}(u) > N^\varepsilon Y^{(N)}(u)\right) \leq N^{-D}$$

for large enough $N \geq N_0(\varepsilon, D)$. Throughout this paper the stochastic domination will always be uniform in all parameters (such as matrix indices, deterministic vectors, and spectral parameters z) that are not explicitly fixed. Note that $N_0(\varepsilon, D)$ may depend on quantities that are explicitly constant.

- (ii) If X is stochastically dominated by Y , uniformly in u , we use the notation $X \prec Y$. Moreover, if for some complex family X we have $|X| \prec Y$ we also write $X = O_{\prec}(Y)$. (Note that for deterministic X and Y , $X = O_{\prec}(Y)$ means $X = O_{\varepsilon}(N^{\varepsilon}Y)$ for any $\varepsilon > 0$.)
- (iii) We extend the definition of $O_{\prec}(\cdot)$ to matrices in the weak operator sense as follows. Let X be a family of complex $N \times N$ random matrices and Y a family of nonnegative random variables. Then we write $X = O_{\prec}(Y)$ to mean $|\langle \mathbf{v}, X \mathbf{w} \rangle| \prec |\mathbf{v}| |\mathbf{w}| Y$ uniformly for all deterministic vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^N$.

1.3. Derivation of the self-consistent equation. We use the notation $A \geq 0$ to mean that $A \in \mathbb{C}^{N \times N}$ is a positive semidefinite matrix, i.e. it is Hermitian and all of its eigenvalues are nonnegative.

Definition 1.3. Define $\text{Im } M := \frac{1}{2i}(M - M^*)$ and $\mathcal{M}_+ := \{M \in \mathbb{C}^{N \times N} : \text{Im } M \geq 0\}$. Let $\mathcal{S} : \mathbb{C}^{N \times N} \rightarrow \mathbb{C}^{N \times N}$ be a linear map that maps \mathcal{M}_+ to itself. For $z \in \mathbb{C}_+$ define the function $\Pi \equiv \Pi(\cdot, z) : \mathbb{C}^{N \times N} \rightarrow \mathbb{C}^{N \times N}$ through

$$\Pi(M) \equiv \Pi(M, z) := I + zM + \mathcal{S}(M)M - AM.$$

The following deterministic result was proved in [28].

Lemma 1.4. *Let \mathcal{S} and Π be as in Definition 1.3. Then for any $z \in \mathbb{C}_+$ the equation*

$$\Pi(M) = 0 \tag{1.3}$$

has a unique solution $M \in \mathcal{M}_+$. We denote this solution by $M(z)$.

We now choose \mathcal{S} to be

$$\mathcal{S}(M) := \mathbb{E}[HMH]. \tag{1.4}$$

Fix a constant $\tau > 0$ and define the fundamental domain

$$\mathbf{D} \equiv \mathbf{D}_N(\tau) := \{E + i\eta : |E| \leq \tau^{-1}, N^{-1+\tau} \leq \eta \leq \tau^{-1}\}. \tag{1.5}$$

Our first main result gives optimal high-probability bounds on $\Pi(G)$.

Theorem 1.5 (Self-consistent equation for G). *Suppose that Assumption 1.1 holds. Denote by $M \in \mathcal{M}_+$ the solution of (1.3) with respect to the linear map (1.4). Let $z \in \mathbf{D}$ and suppose that $\|M\| = O(1)$ and $G - M = O_{\prec}(\phi)$ for some deterministic $\phi \in [N^{-1}, N^{\tau/10}]$ hold at z . Then the following estimates hold at z .*

(i) *We have*

$$\Pi(G) = O_{\prec} \left((1 + \phi)^3 \sqrt{\frac{\|\text{Im } M\| + \phi + \eta}{N\eta}} \right).$$

(ii) *For any deterministic $B \in \mathbb{C}^{N \times N}$ satisfying $\|B\| = O(1)$ we have*

$$\underline{B\Pi(G)} = O_{\prec} \left((1 + \phi)^6 \frac{\|\text{Im } M\| + \phi + \eta}{N\eta} \right).$$

Remark 1.6. The expression $\underline{B\Pi(G)}$ is the most general linear function of $\Pi(G)$, in the sense that any linear function $\Phi : \mathbb{C}^{N \times N} \rightarrow \mathbb{C}$ can be written in the form $\Phi(\Pi) = \underline{B\Pi}$ for some $B \in \mathbb{C}^{N \times N}$.

1.4. Application to Wigner matrices. We make the following assumption about Wigner matrices.

Assumption 1.7. For all $1 \leq i, j \leq N$ we have $\mathbb{E}|H_{ij}|^2 = N^{-1}(1 + O(\delta_{ij}))$.

For Wigner matrices satisfying Assumption 1.7, instead of (1.4) it is more convenient to use the slightly different expression

$$\mathcal{S}(M) := mI, \quad m := \underline{M}. \quad (1.6)$$

It is immediate that, under the Assumption $\|M\| = O(1)$, the entries of the quantities (1.4) and (1.6) differ by $O(N^{-1})$.

The equations (1.3) and (1.6) have been studied in detail in the literature on deformed Wigner matrices. They may be rewritten as

$$M = (A - z - m)^{-1},$$

where m solves the equation

$$m = \int \frac{\nu(da)}{a - m - z} \quad (1.7)$$

and

$$\nu := \frac{1}{N} \sum_{a \in \text{spec}(A)} \delta_a \quad (1.8)$$

is the empirical spectral measure of A . The function m is a holomorphic function from \mathbb{C}_+ to \mathbb{C}_+ , and we find immediately that $zm(z) \rightarrow -1$ as $z \rightarrow \infty$ in \mathbb{C}_+ . By the integral representation of Nevanlinna functions, we conclude that m is the Stieltjes transform of a probability measure ϱ on \mathbb{R} :

$$m(z) = \int \frac{\varrho(dx)}{x - z}. \quad (1.9)$$

We remark that the measure ϱ is the *free additive convolution* of ν and the semicircle law μ , written $\varrho = \mu \boxplus \nu$. For more details, we refer to [39] for a review of free probability theory. For recent results on the local behaviour of the free additive convolution, we refer to [34, 35].

Define the Stieltjes transform of the empirical spectral measure of $H + A$ as

$$g := \underline{G}.$$

Definition 1.8. We call a subset $\mathbf{S} \equiv \mathbf{S}_N \subset \mathbf{D}$ a *spectral domain* if for each $z \in \mathbf{S}$ we have

$$\{w \in \mathbf{D} : \text{Re } w = \text{Re } z, \text{Im } w \geq \text{Im } z\} \subset \mathbf{S}.$$

Theorem 1.9. *Suppose that Assumptions 1.1 and 1.7 hold and that $\|A\| = O(1)$. Let $M \in \mathcal{M}_+$ and m be the solutions of (1.3) and (1.6). Let \mathbf{S} be a spectral domain. Suppose that $\|M\| = O(1)$ on \mathbf{S} and that (1.7) is stable on \mathbf{S} in the sense of Definition 4.9 below. Then we have for $z \in \mathbf{S}$*

$$G - M = O_{\prec} \left(\sqrt{\frac{\text{Im } m}{N\eta}} + \frac{1}{N\eta} \right) \quad (1.10)$$

and

$$g - m = O_{\prec} \left(\frac{1}{N\eta} \right). \quad (1.11)$$

Remark 1.10. The assumptions that $\|M\| = O(1)$ on \mathbf{S} and that (1.7) is stable on \mathbf{S} in the sense of Definition 4.9 only pertain to the deterministic measure ν from (1.8), and have been extensively studied in the literature, starting with the work [34]. For instance, a sufficient condition is that

$$\inf_{x \in I_\nu} \int \frac{\nu(da)}{(x-a)^2} \geq 1 + c$$

for some constant $c > 0$, where I_ν is the smallest closed interval containing the support of ν . See [35] for more details.

Remark 1.11. In Theorem 1.9, we make the assumption $\|A\| = O(1)$ for convenience in order to simplify the analysis of the self-consistent equation (1.7). This assumption can be easily relaxed; in particular, it is not needed to apply Theorem 1.5.

Remark 1.12. A corollary of Theorem 1.9 is that the Wigner-Dyson-Mehta conjecture holds for Wigner matrices with arbitrary expectation. In other words, $H + A$ satisfying the assumptions of Theorem 1.9 has universal (sine kernel) local spectral statistics. This follows immediately by combining Theorem 1.9 with [35].

2. Preliminaries

The rest of the paper is devoted to the proofs of Theorems 1.5 and 1.9. To simplify notation, from now on we only consider the case where all matrix and vector entries are real. The complex case is a simple extension, whose details are given in Appendix B.

In this section we collect notation and various tools which are used throughout the paper.

Additional notation. Let $\llbracket n \rrbracket := \{1, 2, \dots, n\}$ and denote by $\mathbb{S} := \{\mathbf{v} \in \mathbb{R}^N : |\mathbf{v}| = 1\}$ the unit sphere of \mathbb{R}^N . For an $N \times N$ matrix $X \in \mathbb{R}^{N \times N}$, vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^N$, and $i, j \in \llbracket N \rrbracket$, we use the notations

$$X_{\mathbf{v}\mathbf{w}} := \langle \mathbf{v}X, \mathbf{w} \rangle, \quad X_{i\mathbf{w}} := \langle \mathbf{e}_i X, \mathbf{w} \rangle, \quad X_{\mathbf{v}j} := \langle \mathbf{v}X, \mathbf{e}_j \rangle, \quad X_{ij} := \langle \mathbf{e}_i X, \mathbf{e}_j \rangle,$$

where \mathbf{e}_i is the standard i -th basis vector of \mathbb{R}^N . We use the abbreviation $G_{\mathbf{v}\mathbf{w}}^k := (G_{\mathbf{v}\mathbf{w}})^k$ for $k \in \mathbb{N}$.

Basic properties of stochastic domination. The following lemma collects some of the basic properties of stochastic domination. We use the lemma tacitly throughout the paper.

Lemma 2.1 (Basic properties of \prec).

- (i) Suppose that $X(v) \prec Y(v)$ for all $v \in V$. If $|V| \leq N^C$ for some constant C then $\sum_{v \in V} X(v) \prec \sum_{v \in V} Y(v)$.
- (ii) Suppose that $X_1 \prec Y_1$ and $X_2 \prec Y_2$. Then $X_1 X_2 \prec Y_1 Y_2$.
- (iii) Suppose that $X \leq N^C$ and $Y \geq N^{-C}$ for some constant $C > 0$. Then $X \prec Y$ implies $\mathbb{E}[X] \prec \mathbb{E}[Y]$.

If the above random variables depend on some parameter u and the hypotheses are uniform in u then so are the conclusions.

Proof. The proof follows from the definition of stochastic domination together with a union bound argument. \square

Lemma 2.2. *Under Assumption 1.1 we have $\|H\| \prec 1$.*

Proof. This is a standard application of the Füredi-Komlós argument (see e.g. [6, Section 2.1.6]). \square

Ward identity. An important ingredient of the proofs is the following well-known identity.

Lemma 2.3 (Ward identity). *For every vector $\mathbf{x} \in \mathbb{R}^N$*

$$\sum_j |G_{\mathbf{x}j}|^2 = \frac{\operatorname{Im} G_{\mathbf{x}\mathbf{x}}}{\eta}.$$

Proof. This follows immediately from the resolvent identity $G - G^* = 2i\eta GG^*$. \square

As a consequence, for any $B \in \mathbb{R}^{N \times N}$ we have the estimate

$$\sum_j |(GB)_{\mathbf{x}j}|^2 = (GBB^*G^*)_{\mathbf{x}\mathbf{x}} \leq \|B\|^2 (GG^*)_{\mathbf{x}\mathbf{x}} = \|B\|^2 \frac{\operatorname{Im} G_{\mathbf{x}\mathbf{x}}}{\eta}. \quad (2.1)$$

Cumulant expansion. Recall that for a real random variable h , all of whose moments are finite, the k -cumulant of h is

$$\mathcal{C}_k(h) := (-i)^k \left(\frac{d^k}{dt^k} \log \mathbb{E}[e^{ith}] \right) \Big|_{t=0}.$$

We shall use a standard cumulant expansion from [16, 27, 29], which we record in Lemma 2.4 below. We require slightly stronger bounds on the remainder term than in previous applications, and for completeness we give a proof of Lemma 2.4 in Appendix A.

Lemma 2.4 (Cumulant expansion). *Let $f : \mathbb{R} \rightarrow \mathbb{C}$ be a smooth function, and denote by $f^{(k)}$ its k -th derivative. Then, for every fixed $\ell \in \mathbb{N}$, we have*

$$\mathbb{E}[h \cdot f(h)] = \sum_{k=0}^{\ell} \frac{1}{k!} \mathcal{C}_{k+1}(h) \mathbb{E}[f^{(k)}(h)] + \mathcal{R}_{\ell+1}, \quad (2.2)$$

assuming that all expectations in (2.2) exist, where $\mathcal{R}_{\ell+1}$ is a remainder term (depending on f and h), such that for any $t > 0$,

$$\mathcal{R}_{\ell+1} = O(1) \cdot \left(\mathbb{E} \sup_{|x| \leq |h|} |f^{(\ell+1)}(x)|^2 \cdot \mathbb{E} |h|^{2\ell+4} \mathbf{1}_{|h|>t} \right)^{1/2} + O(1) \cdot \mathbb{E} |h|^{\ell+2} \cdot \sup_{|x| \leq t} |f^{(\ell+1)}(x)|. \quad (2.3)$$

The following result gives bounds on the cumulants of the entries of H .

Lemma 2.5. *If H satisfies Assumption 1.1 then for every $i, j \in \llbracket N \rrbracket$ and $k \in \mathbb{N}$ we have*

$$\mathcal{C}_{k+1}(H_{ij}) = O_k(N^{-(k+1)/2})$$

and $\mathcal{C}_1(H_{ij}) = 0$.

Proof. This follows easily by the homogeneity of the cumulants. \square

Self-improving high-probability bounds. Throughout the proof, we shall repeatedly use self-improving high-probability bounds, of which the following lemma is a prototype.

Lemma 2.6. *Let $C > 0$ be a constant, $X \geq 0$, and $Y \in [N^{-C}, N^C]$. Suppose there exists a constant $q \in [0, 1)$ such that for any $Z \in [Y, N^C]$, we have the implication*

$$X \prec Z \quad \implies \quad X \prec Z^q Y^{1-q}, \quad (2.4)$$

Then we have $X \prec Y$ provided that $X \prec N^C$.

Proof. By iteration of (2.4) k times, starting from $X \prec N^C$, we find

$$X \prec N^{Cq^k} Y^{1-q^k} \leq N^{2Cq^k} Y.$$

Since $2Cq^k$ can be made arbitrarily small by choosing k large enough, the claim follows. \square

3. Proof of Theorem 1.5

3.1. Preliminary estimates. In this subsection we collect some tools and basic estimates that are used in the proof of Theorem 1.5. Throughout the following we consider two fixed deterministic vectors $\mathbf{v}, \mathbf{w} \in \mathbb{S}$.

Define

$$s_{ij} := (1 + \delta_{ij})^{-1} N \mathbb{E} |H_{ij}|^2, \quad (3.1)$$

and for a vector $\mathbf{x} = (x_i)_{i \in [N]} \in \mathbb{R}^N$ denote

$$\mathbf{x}^j = (x_i^j)_{i \in [N]}, \quad \text{where} \quad x_i^j := x_i s_{ij}. \quad (3.2)$$

We define the set of vectors

$$\mathbb{X} \equiv \mathbb{X}(\mathbf{v}, \mathbf{w}) := \{\mathbf{v}, \mathbf{w}, \mathbf{v}^1, \dots, \mathbf{v}^N, \mathbf{e}_1, \dots, \mathbf{e}_N\}.$$

Note that, due to Assumption 1.1, we have $s_{ij} = O(1)$, so that $|\mathbf{x}^j| = O(|\mathbf{x}|)$ for all $\mathbf{x} \in \mathbb{R}^N$ and $j \in [N]$, and in particular $|\mathbf{x}| = O(1)$ for all $\mathbf{x} \in \mathbb{X}$. Furthermore, under the assumptions $\|M\| = O(1)$ and $G - M = O_{\prec}(\phi)$ of Theorem 1.5 we have, by a union bound,

$$\max_{\mathbf{x}, \mathbf{y} \in \mathbb{X}} |G_{\mathbf{xy}}| \leq \max_{\mathbf{x}, \mathbf{y} \in \mathbb{X}} |M_{\mathbf{xy}}| + \max_{\mathbf{x}, \mathbf{y} \in \mathbb{X}} |(G - M)_{\mathbf{xy}}| \prec 1 + \phi. \quad (3.3)$$

We abbreviate

$$\zeta := \sqrt{\frac{\|\operatorname{Im} M\| + \phi + \eta}{N\eta}}, \quad \text{and} \quad \tilde{\zeta} := (1 + \phi)^3 \zeta. \quad (3.4)$$

Since $N^{-1} \leq \phi \leq N^{\tau/10}$, $N^{-1+\tau} \leq \eta \leq \tau^{-1}$, and $\|M\| = O(1)$ by the assumption of Theorem 1.5, we have

$$\zeta \leq \tilde{\zeta} \leq (1 + \phi)^3 O\left(\sqrt{\frac{1 + \phi}{N\eta} + \frac{1}{N}}\right) \leq (1 + \phi)^3 O\left(\sqrt{\frac{1 + \phi}{N\eta}}\right) \leq O(N^{-3\tau/20}), \quad (3.5)$$

and in particular $\tilde{\zeta} \leq 1$. In addition, it follows from the definition of ζ that $\zeta \geq N^{-1/2}$.

Using the Ward identity. The proof of Theorem 1.5 makes constant use of the Ward identity (Lemma 2.3) and its consequence (2.1), usually combined with the Cauchy-Schwarz inequality. For future reference, we collect here several representative examples of such uses, as they appear in the proof. To streamline the presentation, in the actual proof of Theorem 1.5 below, we do not give the details of the applications of the Ward identity. Instead, we tacitly use estimates of the following type whenever we mention the use of the Ward identity. In each of the examples, $\mathbf{v}, \mathbf{w} \in \mathbb{S}$ are deterministic and B is a deterministic $N \times N$ matrix satisfying $\|B\| = O(1)$.

(i) By (2.1) and the estimate $G - M = O_{\prec}(\phi)$ we have

$$\frac{1}{N} \sum_i |(GB)_{\mathbf{v}i}|^2 \leq O(1) \frac{\text{Im } G_{\mathbf{v}\mathbf{v}}}{N\eta} \prec \frac{\text{Im } M_{\mathbf{v}\mathbf{v}} + \phi}{N\eta} \leq \frac{\|\text{Im } M\| + \phi}{N\eta} \leq \zeta^2. \quad (3.6)$$

Similarly,

$$\frac{1}{N} \sum_i |(GB)_{\mathbf{v}i}| \leq \left(\frac{1}{N} \sum_i |(GB)_{\mathbf{v}i}|^2 \right)^{1/2} \prec \zeta. \quad (3.7)$$

(ii) Using that $s_{ij} = O(1)$ and (3.7) we find

$$\begin{aligned} \frac{1}{N^2} \sum_{i,j} |G_{\mathbf{v}\mathbf{e}_i^j} (GB)_{ji} G_{j\mathbf{w}}| &\leq \frac{O(1)}{N} \sum_i |G_{\mathbf{v}i}| \left(\frac{1}{N} \sum_j |(GB)_{ji} G_{j\mathbf{w}}| \right) \\ &\leq \frac{O(1)}{N} \sum_i |G_{\mathbf{v}i}| \left(\frac{1}{N} \sum_j |(GB)_{ji}|^2 \right)^{1/2} \left(\frac{1}{N} \sum_j |G_{j\mathbf{w}}|^2 \right)^{1/2} \prec \frac{1}{N} \sum_i |G_{\mathbf{v}i}| \zeta^2 \prec \zeta^3. \end{aligned} \quad (3.8)$$

(iii) We have

$$\begin{aligned} &\frac{1}{N^4} \sum_{i,j,a,b} |(GB)_{\mathbf{v}\mathbf{e}_i^j} G_{j\mathbf{w}} G_{\mathbf{v}a} G_{a\mathbf{w}} G_{\mathbf{e}_b^a} (GB)_{jb}| \\ &= \frac{O(1)}{N^3} \sum_{i,j,a} |(GB)_{\mathbf{v}i} G_{j\mathbf{w}} G_{\mathbf{v}a} G_{a\mathbf{w}}| \left(\frac{1}{N} \sum_b |G_{b\mathbf{e}_i^a} (GB)_{jb}| \right) \prec \frac{1}{N^3} \sum_{i,j,a} |(GB)_{\mathbf{v}i} G_{j\mathbf{w}} G_{\mathbf{v}a} G_{a\mathbf{w}}| \zeta^2 \\ &\leq \left(\frac{1}{N} \sum_i |(GB)_{\mathbf{v}i}| \right) \left(\frac{1}{N} \sum_j |G_{j\mathbf{w}}| \right) \left(\frac{1}{N} \sum_a |G_{\mathbf{v}a} G_{a\mathbf{w}}| \right) \zeta^2 \prec \zeta^6. \end{aligned} \quad (3.9)$$

(iv) Let Q be a random matrix satisfying $Q = O_{\prec}(\lambda)$. Then

$$\frac{1}{N^2} \sum_{i,j} |s_{ij} (GB)_{ji} Q_{ij}| \leq \frac{O(1)}{N^2} \sum_{i,j} |(GB)_{ji} Q_{ij}| \prec \frac{\lambda}{N^2} \sum_{i,j} |(GB)_{ji}| \leq \lambda \zeta. \quad (3.10)$$

Reduction: removal of \mathcal{J} . In order to simplify the estimation of $\Pi(G)$ and $\underline{B}\Pi(G)$ we subtract from $\Pi(G)$ a term, denoted below by \mathcal{J} , which can easily be shown to be sufficiently small. To this end, we split

$$(\mathcal{S}(G)G)_{\mathbf{v}\mathbf{w}} = \mathcal{J}_{\mathbf{v}\mathbf{w}} + \mathcal{K}_{\mathbf{v}\mathbf{w}}, \quad \mathcal{J}_{\mathbf{v}\mathbf{w}} := \frac{1}{N} \sum_j G_{j\mathbf{v}^j} G_{j\mathbf{w}}, \quad \mathcal{K}_{\mathbf{v}\mathbf{w}} := \frac{1}{N} \sum_j G_{jj} G_{\mathbf{v}^j\mathbf{w}}, \quad (3.11)$$

and denote

$$\mathcal{D}_{\mathbf{vw}} := \Pi(G)_{\mathbf{vw}} - \mathcal{J}_{\mathbf{vw}} = (I + zG - AG + \mathcal{K})_{\mathbf{vw}} = (HG + \mathcal{K})_{\mathbf{vw}}, \quad (3.12)$$

where we used the identity $I + zG - AG = HG$.

Due to (3.3) and the Ward identity, we know that

$$|\mathcal{J}_{\mathbf{vw}}| \prec (1 + \phi) \cdot \frac{1}{N} \sum_j |G_{j\mathbf{w}}| \prec (1 + \phi)\zeta. \quad (3.13)$$

Similarly, using the Ward identity we get

$$|\underline{\mathcal{J}B}| = \frac{1}{N^2} \left| \sum_{i,j} (GB)_{ji} G_{ji} s_{ij} \right| \prec \zeta^2. \quad (3.14)$$

Because of (3.13) and (3.14), for the proof of Theorem 1.5 it suffices to prove the following two results.

Proposition 3.1. *Under the assumptions of Theorem 1.5 we have $\mathcal{D} = O_{\prec}(\tilde{\zeta})$.*

Proposition 3.2. *Under the assumptions of Theorem 1.5 we have $\underline{B}\mathcal{D} = O_{\prec}(\tilde{\zeta}^2)$ provided that $\|B\| = O(1)$.*

In both proofs we expand the term HG on the right-hand side of (3.12) using the cumulant expansion (Lemma 2.4). Since H is symmetric, for any differentiable function $f = f(H)$ we set

$$\partial_{ij} f(H) := \frac{\partial}{\partial H_{ij}} f(H) = \frac{\partial}{\partial H_{ji}} f(H) := \frac{d}{dt} \Big|_{t=0} f(H + t \Delta^{ij}), \quad (3.15)$$

where Δ^{ij} denotes the matrix whose entries are zero everywhere except at the sites (i, j) and (j, i) where they are one: $\Delta_{kl}^{ij} = (\delta_{ik}\delta_{jl} + \delta_{jk}\delta_{il})(1 + \delta_{ij})^{-1}$. By differentiation of $G = (H + A - z)^{-1}$ we get

$$\partial_{ij} G_{\mathbf{xy}} = -(1 + \delta_{ij})^{-1} (G_{\mathbf{x}i} G_{j\mathbf{y}} + G_{\mathbf{x}j} G_{i\mathbf{y}}). \quad (3.16)$$

Combining (3.16), the assumption $\|B\| = O(1)$, and (3.3), a simple induction gives

$$|\partial_{ij}^m G_{\mathbf{xy}}| \prec (1 + \phi)^{m+1}, \quad |\partial_{ij}^m (GB)_{\mathbf{xy}}| \prec (1 + \phi)^{m+1} \quad (3.17)$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{X}$ and $m \geq 0$.

3.2. Proof of Proposition 3.1. Before starting the proof, we record a preliminary estimate on the derivatives of \mathcal{D} .

Lemma 3.3. *Suppose that the assumptions of Theorem 1.5 hold. Then for any fixed $l \geq 0$ we have*

$$\partial_{ij}^l \mathcal{D} = O_{\prec}(N^{1/2}(1 + \phi)^{l+2}\zeta).$$

Proof. We proceed by induction. Fix $\mathbf{v}, \mathbf{w} \in \mathbb{S}$. Using (3.3), we obtain

$$|\mathcal{D}_{\mathbf{vw}}| \leq |(HG)_{\mathbf{vw}}| + \frac{1}{N} \sum_j |G_{jj} G_{\mathbf{v}j\mathbf{w}}| \prec |(HG)_{\mathbf{vw}}| + (1 + \phi)^2.$$

Using the bound (3.3), the Ward identity, and Lemma 2.2, we therefore get

$$|(HG)_{\mathbf{vw}}| \leq \|H\| |G_{\mathbf{w}}| \prec |G_{\mathbf{w}}| = \sqrt{\frac{\text{Im } G_{\mathbf{ww}}}{\eta}} \prec N^{1/2} \zeta.$$

Since $\zeta \geq N^{-1/2}$ we conclude that

$$|\mathcal{D}_{\mathbf{vw}}| \prec N^{1/2} (1 + \phi)^2 \zeta.$$

This completes the proof for the case $l = 0$.

Next, let $l \geq 1$ and suppose that $\partial_{ij}^m \mathcal{D} = O_{\prec}(N^{1/2} (1 + \phi)^{m+2} \zeta)$ for $0 \leq m \leq l - 1$. Using (3.16) we find

$$(1 + \delta_{ij}) \partial_{ij} \mathcal{D}_{\mathbf{vw}} = -(\mathcal{D}_{\mathbf{vi}} G_{j\mathbf{w}} + \mathcal{D}_{\mathbf{vj}} G_{i\mathbf{w}}) + U_{\mathbf{vw}}^{ij}, \quad (3.18)$$

where

$$U_{\mathbf{vw}}^{ij} := v_i G_{j\mathbf{w}} + v_j G_{i\mathbf{w}} - \frac{1}{N} \sum_r G_{ri} G_{jr} G_{v\mathbf{r}\mathbf{w}} - \frac{1}{N} \sum_r G_{rj} G_{ir} G_{v\mathbf{r}\mathbf{w}}. \quad (3.19)$$

Thus

$$(1 + \delta_{ij}) \partial_{ij}^l \mathcal{D}_{\mathbf{vw}} = -\partial_{ij}^{l-1} (\mathcal{D}_{\mathbf{vi}} G_{j\mathbf{w}} + \mathcal{D}_{\mathbf{vj}} G_{i\mathbf{w}}) + \partial_{ij}^{l-1} U_{\mathbf{vw}}^{ij}. \quad (3.20)$$

We deal with each of the terms on the right-hand side separately. The term $\partial_{ij}^{l-1} (\mathcal{D}_{\mathbf{vi}} G_{j\mathbf{w}})$ is estimated using (3.17) as

$$\begin{aligned} |\partial_{ij}^{l-1} (\mathcal{D}_{\mathbf{vi}} G_{j\mathbf{w}})| &\leq \sum_{m=0}^{l-1} \binom{l-1}{m} |(\partial_{ij}^{l-1-m} \mathcal{D}_{\mathbf{vi}}) (\partial_{ij}^m G_{j\mathbf{w}})| \prec \sum_{m=0}^{l-1} (1 + \phi)^{m+1} |\partial_{ij}^{l-1-m} \mathcal{D}_{\mathbf{vi}}| \\ &\prec \sum_{m=0}^{l-1} (1 + \phi)^{m+1} N^{1/2} (1 + \phi)^{l-m+1} \zeta = O(N^{1/2} (1 + \phi)^{l+2} \zeta), \end{aligned}$$

where in the third step we used the induction assumption. The second term of (3.20) is estimated analogously. Finally, by (3.16), (3.17), (3.19), and the bound $|v_i| \leq 1$, we can estimate the last term of (3.20) as

$$\partial_{ij}^{l-1} U_{\mathbf{vw}}^{ij} \prec (1 + \phi)^{l-1+3} = (1 + \phi)^{l+2} \leq N^{1/2} (1 + \phi)^{l+2} \zeta.$$

This concludes the proof. \square

Now we turn to the proof of Proposition 3.1. Fix a constant $p \in \mathbb{N}$. Then

$$\mathbb{E}[|\mathcal{D}_{\mathbf{vw}}|^{2p}] = \mathbb{E}[(\mathcal{K} + HG)_{\mathbf{vw}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p] = \mathbb{E}[\mathcal{K}_{\mathbf{vw}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p] + \sum_{i,j} v_i \mathbb{E}[H_{ij} G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p]. \quad (3.21)$$

Using the cumulant expansion (Lemma 2.4) for the second term in (3.21) with $h = H_{ij}$ and

$$f = f_{ij}(H) = G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p, \quad (3.22)$$

we obtain

$$\mathbb{E}|\mathcal{D}_{\mathbf{vw}}|^{2p} = \mathbb{E}[\mathcal{K}_{\mathbf{vw}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p] + \sum_{k=1}^{\ell} X_k + \sum_{i,j} v_i \mathcal{R}_{\ell+1}^{ij}, \quad (3.23)$$

where we used the notation

$$X_k := \sum_{i,j} v_i \left(\frac{1}{k!} \mathcal{C}_{k+1}(H_{ij}) \mathbb{E}[\partial_{ij}^k (G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p)] \right). \quad (3.24)$$

Note that the sum in (3.23) begins from $k = 1$ because $\mathcal{C}_1(H_{ij}) = 0$. Here ℓ is a fixed positive integer to be chosen later, and $\mathcal{R}_{\ell+1}^{ij}$ is a remainder term defined analogously to $\mathcal{R}_{\ell+1}$ in (2.3). More precisely, for any $t > 0$ we have the bound

$$\begin{aligned} \mathcal{R}_{\ell+1}^{ij} &= O(1) \cdot \left(\mathbb{E} \sup_{|x| \leq |H_{ij}|} |\partial_{ij}^{\ell+1} f(H^{ij} + x\Delta^{ij})|^2 \cdot \mathbb{E} |H_{ij}^{2\ell+4} \mathbf{1}_{|H_{ji}| > t}| \right)^{1/2} \\ &\quad + O(1) \cdot \mathbb{E} |H_{ij}|^{\ell+2} \cdot \mathbb{E} \left[\sup_{|x| \leq t} |\partial_{ij}^{\ell+1} f(H^{ij} + x\Delta^{ij})| \right], \end{aligned} \quad (3.25)$$

where we defined $H^{ij} := H - H_{ij}\Delta^{ij}$, so that the matrix H^{ij} has zero entries at the positions (i, j) and (j, i) . The proof of Proposition 3.1 is broken down into the following lemma.

Lemma 3.4. *Suppose that $\mathcal{D} = O_{\prec}(\lambda)$ for some deterministic $\lambda \geq \tilde{\zeta}$. Fix $p \geq 2$. Under the assumptions of Theorem 1.5, we have the following estimates.*

- (i) $\mathbb{E}[\mathcal{K}_{\mathbf{vw}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p] + X_1 = O_{\prec}(\tilde{\zeta}) \cdot \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-1} + O_{\prec}(\tilde{\zeta}\lambda) \cdot \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-2}$.
- (ii) For $k \geq 2$, $X_k = \sum_{n=1}^{2p} O_{\prec}((\tilde{\zeta}\lambda^2)^{n/3}) \cdot \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-n}$.
- (iii) For any $D > 0$ there exists $\ell \equiv \ell(D) \geq 1$ such that $\mathcal{R}_{\ell+1}^{ij} = O(N^{-D})$ uniformly for all $i, j \in \llbracket N \rrbracket$.

Indeed, combining Lemma 3.4 for $\ell \equiv \ell(2p+2)$ together with (3.23) we obtain

$$\begin{aligned} \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p} &= \sum_{n=1}^{2p} O_{\prec}((\tilde{\zeta}\lambda^2)^{n/3}) \cdot \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-n} + O(N^{-2p}) \\ &\leq \sum_{n=1}^{2p} O_{\prec}((\tilde{\zeta}\lambda^2)^{n/3}) \cdot (\mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p})^{(2p-n)/2p} + O(N^{-2p}), \end{aligned}$$

where in the second step we used Hölder's inequality. Since $\lambda \geq \tilde{\zeta} \geq N^{-1/2}$, we conclude that $N^{-2p} \leq (\tilde{\zeta}\lambda^2)^{2p/3}$, and therefore $\mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p} = O_{\prec}((\tilde{\zeta}\lambda^2)^{2p/3})$. Since p was arbitrary, we conclude from Markov's inequality the implication

$$\mathcal{D} = O_{\prec}(\lambda) \quad \implies \quad \mathcal{D} = O_{\prec}((\tilde{\zeta}\lambda^2)^{1/3}), \quad (3.26)$$

for any deterministic $\lambda \geq \tilde{\zeta}$. Moreover, by Lemma 3.3 we have the a priori bound $\mathcal{D} = O_{\prec}((1 + \phi)^2 N^{1/2} \zeta)$. Hence, an iteration of (3.26), analogous to Lemma 2.6, yields $\mathcal{D} = O_{\prec}(\tilde{\zeta})$. This concludes the proof of Proposition 3.1.

What remains, therefore, is the proof of Lemma 3.4.

Proof of Lemma 3.4 (i). Note that $\mathcal{C}_2(H_{ij}) = \mathbb{E}[H_{ij}^2]$ since $\mathbb{E}H_{ij} = 0$. Hence,

$$\begin{aligned} X_1 &= \sum_{i,j} v_i \mathcal{C}_2(H_{ij}) \mathbb{E}[\partial_{ij}(G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p)] \\ &= - \sum_{i,j} v_i \mathbb{E}[H_{ij}^2] \cdot \mathbb{E}[(G_{ji} G_{j\mathbf{w}} + G_{jj} G_{i\mathbf{w}})(1 + \delta_{ij})^{-1} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p] + \sum_{i,j} v_i \mathbb{E}[H_{ij}^2] \mathbb{E}[G_{j\mathbf{w}} \partial_{ij}(\mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p)] \\ &= -\mathbb{E}[(\mathcal{S}(G)G)_{\mathbf{vw}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p] + \sum_{i,j} v_i \mathbb{E}[H_{ij}^2] \mathbb{E}[G_{j\mathbf{w}} \partial_{ij}(\mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p)], \end{aligned}$$

and consequently, recalling (3.11), we have

$$\mathbb{E}[\mathcal{K}_{\mathbf{vw}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p] + X_1 = -\mathbb{E}[\mathcal{J}_{\mathbf{vw}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p] + \sum_{i,j} v_i \mathbb{E}[H_{ij}^2] \mathbb{E}[G_{j\mathbf{w}} \partial_{ij}(\mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p)]. \quad (3.27)$$

Since (3.13) implies $\mathbb{E}[\mathcal{J}_{\mathbf{vw}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p] = O_{\prec}(\tilde{\zeta}) \cdot \mathbb{E}|\mathcal{D}_{\mathbf{vw}}|^{2p-1}$, it remains to estimate the second term on the right-hand side of (3.27).

By (3.18) and (3.19) we have

$$\begin{aligned} & \sum_{i,j} v_i \mathbb{E}[H_{ij}^2] \mathbb{E}[G_{j\mathbf{w}} \partial_{ij}(\mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p)] \\ &= (p-1) \sum_{i,j} v_i \mathbb{E}[H_{ij}^2] \mathbb{E}[G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-2} \overline{\mathcal{D}}_{\mathbf{vw}}^p \partial_{ij} \mathcal{D}_{\mathbf{vw}}] + p \sum_{i,j} v_i \mathbb{E}[H_{ij}^2] \mathbb{E}[G_{j\mathbf{w}} |\mathcal{D}_{\mathbf{vw}}|^{2p-2} \partial_{ij} \overline{\mathcal{D}}_{\mathbf{vw}}] \\ &= \frac{p-1}{N} \sum_{i,j} \mathbb{E} \left[v_i s_{ij} G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-2} \overline{\mathcal{D}}_{\mathbf{vw}}^p \left(-\mathcal{D}_{\mathbf{vi}} G_{j\mathbf{w}} - \mathcal{D}_{\mathbf{vj}} G_{i\mathbf{w}} + v_i G_{j\mathbf{w}} + v_j G_{i\mathbf{w}} \right. \right. \\ & \quad \left. \left. - \frac{1}{N} \sum_r G_{ri} G_{jr} G_{\mathbf{v}^r \mathbf{w}} - \frac{1}{N} \sum_r G_{rj} G_{ir} G_{\mathbf{v}^r \mathbf{w}} \right) \right] + \frac{p}{N} \sum_{i,j} \mathbb{E} \left[v_i s_{ij} G_{j\mathbf{w}} |\mathcal{D}_{\mathbf{vw}}|^{2p-2} \right. \\ & \quad \left. \left(-\overline{\mathcal{D}}_{\mathbf{vi}} \overline{G}_{j\mathbf{w}} - \overline{\mathcal{D}}_{\mathbf{vj}} \overline{G}_{i\mathbf{w}} + v_i \overline{G}_{j\mathbf{w}} + v_j \overline{G}_{i\mathbf{w}} - \frac{1}{N} \sum_r \overline{G}_{ri} \overline{G}_{jr} \overline{G}_{\mathbf{v}^r \mathbf{w}} - \frac{1}{N} \sum_r \overline{G}_{rj} \overline{G}_{ir} \overline{G}_{\mathbf{v}^r \mathbf{w}} \right) \right]. \end{aligned} \quad (3.28)$$

Next, we show that each of the terms on the right-hand side of (3.28) is bounded in absolute value by $O_{\prec}(\tilde{\zeta}\lambda) \cdot \mathbb{E}|\mathcal{D}_{\mathbf{vw}}|^{2p-2}$. We only discuss the first, third, and fifth terms on the right-hand side of (3.28); all the others are estimated analogously.

We start with the first term on the right-hand side of (3.28). By the assumption $\mathcal{D} = O_{\prec}(\lambda)$ we know

$$\sup_{j \in [N]} |\mathcal{D}_{\mathbf{v}^j \mathbf{w}}| \prec \lambda.$$

Thus, with the help of the estimate (3.6), we obtain

$$\begin{aligned} & \left| \frac{p-1}{N} \sum_{i,j} \mathbb{E} \left[-v_i s_{ij} G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-2} \overline{\mathcal{D}}_{\mathbf{vw}}^p \mathcal{D}_{\mathbf{vi}} G_{j\mathbf{w}} \right] \right| = \left| \frac{p-1}{N} \sum_j \mathbb{E} \left[G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-2} \overline{\mathcal{D}}_{\mathbf{vw}}^p \mathcal{D}_{\mathbf{v}^j \mathbf{w}} G_{j\mathbf{w}} \right] \right| \\ &= O_{\prec}(\lambda) \cdot \mathbb{E} \left[\frac{1}{N} \sum_j |G_{j\mathbf{w}}|^2 \cdot |\mathcal{D}_{\mathbf{vw}}|^{2p-2} \right] = O_{\prec}(\lambda \zeta^2) \cdot \mathbb{E}|\mathcal{D}_{\mathbf{vw}}|^{2p-2}, \end{aligned} \quad (3.29)$$

and the desired estimate follows from $\zeta \leq \tilde{\zeta} \leq 1$.

Next, we estimate the third term on the right-hand side of (3.28). Using the Ward identity we get

$$\begin{aligned} & \left| \frac{p-1}{N} \sum_{i,j} \mathbb{E} \left[-v_i s_{ij} G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-2} \overline{\mathcal{D}}_{\mathbf{vw}}^p v_i G_{j\mathbf{w}} \right] \right| \\ &= O(1) \cdot \sum_i |v_i|^2 \cdot \mathbb{E} \left[\frac{1}{N} \sum_j |G_{j\mathbf{w}}|^2 \cdot |\mathcal{D}_{\mathbf{vw}}|^{2p-2} \right] = O_{\prec}(\zeta^2) \cdot \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-2}, \end{aligned}$$

and the result follows from the inequality $\zeta \leq \tilde{\zeta} \leq \lambda$.

Finally, we estimate the fifth term on the right-hand side of (3.28). By the Cauchy-Schwarz inequality and the Ward identity we have

$$\begin{aligned} & \left| \frac{p-1}{N^2} \sum_{i,j,r} \mathbb{E} \left[-v_i s_{ij} G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-2} \overline{\mathcal{D}}_{\mathbf{vw}}^p G_{ri} G_{jr} G_{\mathbf{v}^r \mathbf{w}} \right] \right| = \left| \frac{p-1}{N^2} \sum_{j,r} \mathbb{E} \left[G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-2} \overline{\mathcal{D}}_{\mathbf{vw}}^p G_{r\mathbf{v}j} G_{jr} G_{\mathbf{v}^r \mathbf{w}} \right] \right| \\ &= O_{\prec}((1+\phi)^2) \cdot \frac{1}{N} \sum_r \mathbb{E} \left[\frac{1}{N} \sum_j |G_{j\mathbf{w}} G_{jr}| \cdot |\mathcal{D}_{\mathbf{vw}}|^{2p-2} \right] = O_{\prec}((1+\phi)^2 \zeta^2) \cdot \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-2}, \end{aligned}$$

and the desired estimate follows from $(1+\phi)\zeta \leq \tilde{\zeta} \leq \lambda$. One can check that the estimates of other terms on the right-hand side of (3.28) follow analogously. We conclude that the term on the right-hand side of (3.27) is bounded by

$$O_{\prec}(\tilde{\zeta}\lambda) \cdot \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-2},$$

as claimed. \square

Proof of Lemma 3.4 (ii). Fix $k \geq 2$. By Lemma 2.5 we have

$$\begin{aligned} X_k &= O(N^{-\frac{k+1}{2}}) \cdot \sum_{i,j} \mathbb{E} |v_i \partial_{ij}^k (G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p)| \\ &= O(N^{-\frac{k+1}{2}}) \cdot \sum_{\substack{r,s,t \geq 0 \\ r+s+t=k}} \sum_{i,j} \mathbb{E} |v_i (\partial_{ij}^r G_{j\mathbf{w}}) (\partial_{ij}^s \mathcal{D}_{\mathbf{vw}}^{p-1}) (\partial_{ij}^t \overline{\mathcal{D}}_{\mathbf{vw}}^p)|. \end{aligned}$$

As the sum over r, s, t is finite it suffices to deal with each term separately. To simplify notation, we drop the complex conjugates of Q (which play no role in the subsequent analysis), and estimate the quantity

$$O(N^{-\frac{k+1}{2}}) \cdot \sum_{i,j} \mathbb{E} |v_i (\partial_{ij}^r G_{j\mathbf{w}}) (\partial_{ij}^{k-r} \mathcal{D}_{\mathbf{vw}}^{2p-1})| \quad (3.30)$$

for $r = 0, \dots, k$. Computing the derivative ∂_{ij}^{k-r} , we find that (3.30) is bounded by a sum of terms of the form

$$O(N^{-\frac{k+1}{2}}) \cdot \sum_{i,j} \mathbb{E} \left| v_i (\partial_{ij}^r G_{j\mathbf{w}}) \left(\prod_{m=1}^q (\partial_{ij}^{l_m} \mathcal{D}_{\mathbf{vw}}) \right) \mathcal{D}_{\mathbf{vw}}^{2p-1-q} \right|, \quad (3.31)$$

where the sum ranges over integers $q = 0, \dots, (k-r) \wedge (2p-1)$ and $l_1, \dots, l_q \geq 1$ satisfying $l_1 + \dots + l_q = k-r$. Using Lemma 3.3, we find that (3.31) is bounded by

$$O_{\prec}(N^{-\frac{k+1}{2}}) \cdot \sum_{i,j} \mathbb{E} \left[|v_i (\partial_{ij}^r G_{j\mathbf{w}})| N^{\frac{q}{2}} (1+\phi)^{k-r+2q} \zeta^q |\mathcal{D}_{\mathbf{vw}}|^{2p-1-q} \right].$$

Note that by (3.16) the derivative $\partial_{ij}^r G_{j\mathbf{w}}$ can be written as a sum of terms, each of which is a product of $r+1$ entries of the matrices G , with one entry of the form $G_{i\mathbf{w}}$ or $G_{j\mathbf{w}}$. For definiteness, we suppose that this entry is $G_{i\mathbf{w}}$ in the product. Using (3.3), the Cauchy-Schwarz inequality, and the Ward identity, we get

$$\sum_{i,j} |v_i \partial_{ij}^r G_{i\mathbf{w}}| \prec (1+\phi)^r \cdot \sum_{i,j} |v_i G_{i\mathbf{w}}| \prec N^{\frac{3}{2}} (1+\phi)^r \zeta,$$

and therefore

$$\begin{aligned} (3.31) &\prec N^{\frac{q+2-k}{2}} \cdot (1+\phi)^{k+2q} \zeta^{q+1} \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-1-q} \\ &= (N^{-\frac{1}{2}} (1+\phi))^{k-2-q} \cdot (1+\phi)^{3q+2} \zeta^{q+1} \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-1-q} \\ &\leq (N^{-\frac{1}{2}} (1+\phi))^{k-2-q} \cdot \tilde{\zeta}^{q+1} \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-1-q}, \end{aligned}$$

where for the last inequality we used $(1+\phi)^3 \zeta = \tilde{\zeta}$. Clearly, if $q \leq k-2$, we conclude that (3.31) is bounded by $\tilde{\zeta}^{q+1} \mathbb{E} [|\mathcal{D}_{\mathbf{vw}}|^{2p-1-q}]$, and hence the desired estimate follows from the assumption $\tilde{\zeta} \leq \lambda$.

What remains, therefore, is to estimate (3.31) for $q \geq k-1$, which we assume from now on. Because $k \geq 2$ by assumption, we find that $q \geq 1$. Moreover, since $q \leq k-r$, we find that $r \leq 1$. Thus, it remains to consider the three cases $(r, q) = (0, k)$, $(r, q) = (1, k-1)$ and $(r, q) = (0, k-1)$. We deal with them separately.

The case $(r, q) = (0, k)$. In this case $l_1 = l_2 = \dots = l_q = 1$, so that (3.31) reads

$$\begin{aligned} &O(N^{-\frac{k+1}{2}}) \cdot \sum_{i,j} \mathbb{E} \left| v_i G_{j\mathbf{w}} (\partial_{ij} \mathcal{D}_{\mathbf{vw}})^k \mathcal{D}_{\mathbf{vw}}^{2p-1-k} \right| \\ &\prec N^{-\frac{k+1}{2}} \cdot (N^{\frac{1}{2}} (1+\phi)^3 \zeta)^{k-2} \cdot \sum_{i,j} \mathbb{E} \left| v_i G_{j\mathbf{w}} (\partial_{ij} \mathcal{D}_{\mathbf{vw}})^2 \mathcal{D}_{\mathbf{vw}}^{2p-1-k} \right| \\ &= \tilde{\zeta}^{k-2} \cdot N^{-\frac{3}{2}} \cdot \sum_{i,j} \mathbb{E} \left| v_i G_{j\mathbf{w}} (\partial_{ij} \mathcal{D}_{\mathbf{vw}})^2 \mathcal{D}_{\mathbf{vw}}^{2p-1-k} \right|, \end{aligned} \quad (3.32)$$

where we used Lemma 3.3 and the assumption $k \geq 2$. By (3.18), (3.3), the Ward identity, the bound $\lambda \geq \zeta$, and the assumption $D = O_{\prec}(\lambda)$ we have

$$(1 + \delta_{ij}) \partial_{ij} \mathcal{D}_{\mathbf{vw}} = O_{\prec}((1+\phi)\lambda) + v_i G_{j\mathbf{w}} + v_j G_{i\mathbf{w}}. \quad (3.33)$$

Thus,

$$\sum_{i,j} |v_i G_{j\mathbf{w}} (\partial_{ij} \mathcal{D}_{\mathbf{vw}})^2| \leq \sum_{i,j} \left| v_i G_{j\mathbf{w}} (O_{\prec}((1+\phi)\lambda) + v_i G_{j\mathbf{w}} + v_j G_{i\mathbf{w}})^2 \right|. \quad (3.34)$$

We expand the square on the right-hand side of (3.34) and estimate the result term by term, using (3.3), the Ward identity, and $|\mathbf{v}| = 1$. For example we have

$$\sum_{i,j} |v_i G_{j\mathbf{w}} O_{\prec}((1+\phi)\lambda)^2| \prec N^{\frac{3}{2}} (1+\phi)^2 \zeta \lambda^2,$$

and

$$\sum_{i,j} |v_i G_{j\mathbf{w}} v_j G_{i\mathbf{w}} v_i G_{j\mathbf{w}}| \prec (1+\phi) \cdot \sum_i v_i^2 \sum_j |G_{j\mathbf{w}}|^2 \prec N(1+\phi) \zeta^2 \leq N^{\frac{3}{2}} \tilde{\zeta} \lambda^2,$$

where in the last step we used the estimates $(1 + \phi)\zeta \leq \tilde{\zeta}$ and $N^{-1/2} \leq \zeta \leq \lambda$. By estimating other terms on the right-hand side of (3.34) in a similar fashion, we get the bound

$$\sum_{i,j} |v_i G_{j\mathbf{w}} (\partial_{ij} \mathcal{D}_{\mathbf{vw}})^2| \prec N^{\frac{3}{2}} \tilde{\zeta} \lambda^2.$$

Together with (3.32) we therefore conclude that

$$O(N^{-\frac{k+1}{2}}) \cdot \sum_{i,j} \mathbb{E} \left| v_i G_{j\mathbf{w}} (\partial_{ij} \mathcal{D}_{\mathbf{vw}})^k \mathcal{D}_{\mathbf{vw}}^{2p-1-k} \right| \prec \tilde{\zeta}^{k-1} \lambda^2 \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-k-1}.$$

The desired estimate then follows from the assumptions $\tilde{\zeta} \leq \lambda$ and $k \geq 2$.

The case $(r, q) = (1, k-1)$. We apply a similar argument to the one from the previous case. In this case $l_1 = l_2 = \dots = l_{k-1} = 1$, and (3.31) reads

$$\begin{aligned} O(N^{-\frac{k+1}{2}}) \cdot \sum_{i,j} \mathbb{E} |v_i (\partial_{ij} G_{j\mathbf{w}}) (\partial_{ij} \mathcal{D}_{\mathbf{vw}})^{k-1} \mathcal{D}_{\mathbf{vw}}^{2p-k}| \\ \prec N^{-\frac{k+1}{2}} \cdot (N^{\frac{1}{2}} (1 + \phi)^3 \zeta)^{k-2} \cdot \sum_{i,j} \mathbb{E} |v_i (\partial_{ij} G_{j\mathbf{w}}) (\partial_{ij} \mathcal{D}_{\mathbf{vw}}) \mathcal{D}_{\mathbf{vw}}^{2p-k}| \\ = \tilde{\zeta}^{k-2} \cdot N^{-\frac{3}{2}} \cdot \sum_{i,j} \mathbb{E} |v_i (\partial_{ij} G_{j\mathbf{w}}) (\partial_{ij} \mathcal{D}_{\mathbf{vw}}) \mathcal{D}_{\mathbf{vw}}^{2p-k}|, \end{aligned} \quad (3.35)$$

where in the second step we used Lemma 3.3 and the assumption $k \geq 2$. Using (3.16) and (3.33) to rewrite $\partial_{ij} G_{j\mathbf{w}}$ and one factor of $\partial_{ij} \mathcal{D}_{\mathbf{vw}}$ respectively, we have

$$\sum_{i,j} |v_i (\partial_{ij} G_{j\mathbf{w}}) (\partial_{ij} \mathcal{D}_{\mathbf{vw}})| \leq \sum_{i,j} |v_i (G_{ij} G_{j\mathbf{w}} + G_{ji} G_{j\mathbf{w}}) (O_{\prec}((1 + \phi)\lambda) + v_i G_{j\mathbf{w}} + v_j G_{i\mathbf{w}})|. \quad (3.36)$$

We expand the right-hand side of the above and estimate the result term by term, using (3.3), the Ward identity, and $|\mathbf{v}| = 1$. For example we have

$$\begin{aligned} \sum_{i,j} |v_i G_{ij} G_{j\mathbf{w}} \cdot v_j G_{i\mathbf{w}}| \prec (1 + \phi) \cdot \left(\sum_{i,j} v_i^2 |G_{j\mathbf{w}}|^2 \right)^{\frac{1}{2}} \cdot \left(\sum_{i,j} v_j^2 |G_{i\mathbf{w}}|^2 \right)^{\frac{1}{2}} \\ \prec (1 + \phi) N \zeta^2 \leq N^{\frac{3}{2}} \tilde{\zeta}^2, \end{aligned}$$

and

$$\sum_{i,j} |v_i G_{ij} G_{j\mathbf{w}} \cdot O_{\prec}((1 + \phi)\lambda)| \prec N^{\frac{3}{2}} (1 + \phi)^2 \zeta \lambda.$$

By estimating the other terms on the right-hand side of (3.36) in a similar fashion, we get the bound

$$\sum_{i,j} |v_i (\partial_{ij} G_{j\mathbf{w}} (\partial_{ij} \mathcal{D}_{\mathbf{vw}}))| \prec N^{\frac{3}{2}} \tilde{\zeta} \lambda.$$

Together with (3.35) we conclude that

$$O(N^{-\frac{k+1}{2}}) \cdot \sum_{i,j} \mathbb{E} |v_i (\partial_{ij} G_{j\mathbf{w}}) (\partial_{ij} \mathcal{D}_{\mathbf{vw}})^{k-1} \mathcal{D}_{\mathbf{vw}}^{2p-k}| \prec \tilde{\zeta}^{k-1} \lambda \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-k}.$$

The desired estimate then follows from the assumptions $\tilde{\zeta} \leq \lambda$ and $k \geq 2$.

The case $(r, q) = (0, k-1)$. Since $l_1 + \dots + l_{k-1} = k$ and $l_m \geq 1$ for every $m \in \{1, \dots, k-1\}$, there exists exactly one m such that $l_m = 2$ and the remaining l_m 's are 1. Hence, (3.31) reads

$$O(N^{-\frac{k+1}{2}}) \cdot \sum_{i,j} \mathbb{E} |v_i G_{j\mathbf{w}}(\partial_{ij}^2 \mathcal{D}_{\mathbf{vw}})(\partial_{ij} \mathcal{D}_{\mathbf{vw}})^{k-2} \mathcal{D}_{\mathbf{vw}}^{2p-k}| \\ \prec \tilde{\zeta}^{k-2} \cdot N^{-\frac{3}{2}} \cdot \sum_{i,j} \mathbb{E} |v_i G_{j\mathbf{w}}(\partial_{ij}^2 \mathcal{D}_{\mathbf{vw}}) \mathcal{D}_{\mathbf{vw}}^{2p-k}|, \quad (3.37)$$

where we used Lemma 3.3. By (3.18) we have

$$\sum_{i,j} |v_i G_{j\mathbf{w}}(\partial_{ij}^2 \mathcal{D}_{\mathbf{vw}})| \leq \sum_{i,j} |v_i G_{j\mathbf{w}} \partial_{ij} (-\mathcal{D}_{\mathbf{v}i} G_{j\mathbf{w}} - \mathcal{D}_{\mathbf{v}j} G_{i\mathbf{w}} + U_{\mathbf{vw}}^{ij})|,$$

where U is defined as in (3.19). Now we apply the differential ∂_{ij} on the right-hand side, and estimate the result term by term. By using the bounds (3.3) and $\mathcal{D} = O_{\prec}(\lambda)$, the Ward identity, and $|\mathbf{v}| = 1$ in a similar fashion as in the previous two cases, we get

$$\sum_{i,j} |v_i G_{j\mathbf{w}}(\partial_{ij}^2 \mathcal{D}_{\mathbf{vw}})| \prec N^{\frac{3}{2}} \tilde{\zeta} \lambda.$$

Together with (3.37) we conclude that

$$O(N^{-\frac{k+1}{2}}) \cdot \sum_{i,j} \mathbb{E} |v_i G_{j\mathbf{w}}(\partial_{ij}^2 \mathcal{D}_{\mathbf{vw}})(\partial_{ij} \mathcal{D}_{\mathbf{vw}})^{k-2} \mathcal{D}_{\mathbf{vw}}^{2p-k}| \prec \tilde{\zeta}^{k-1} \lambda \mathbb{E} |\mathcal{D}_{\mathbf{vw}}|^{2p-k}.$$

The desired estimate then follows from the assumptions $\tilde{\zeta} \leq \lambda$ and $k \geq 2$. This concludes the proof. \square

Proof of Lemma 3.4 (iii). The remainder term of the cumulation expansion was analysed in a slightly different context in [27], and we shall follow the same method.

Let $D > 0$ be given. Fix i, j , and choose $t := N^{\tau/5-1/2}$ in (3.25). Define $S := H_{ij} \Delta^{ij}$, where we recall the notation $\Delta_{kl}^{ij} := (\delta_{ik} \delta_{jl} + \delta_{jk} \delta_{il})(1 + \delta_{ij})^{-1}$. Then we have $H^{ij} = H - S$. Let $\hat{G} := (H^{ij} - E - i\eta)^{-1}$. We have the resolvent expansions

$$\hat{G} = G + (GS)G + (GS)^2 \hat{G} \quad (3.38)$$

and

$$G = \hat{G} - (\hat{G}S)\hat{G} + (\hat{G}S)^2 G. \quad (3.39)$$

Note that only two entries of S are nonzero, and they are stochastically dominated by $N^{-1/2}$. Then the bound

$$\sup_{\mathbf{x}, \mathbf{y} \in \mathbb{S}} |\hat{G}_{\mathbf{xy}}| = \|\hat{G}\| \leq \eta^{-1} \leq N^{1-\tau},$$

together with (3.3), (3.38), and the assumption $\phi \leq N^{\tau/10}$, show that

$$\max_{\mathbf{x}, \mathbf{y} \in \mathbb{X}} |\hat{G}_{\mathbf{xy}}| \prec 1 + \phi.$$

Combining with (3.39), the trivial bound $\sup_{\mathbf{x}, \mathbf{y} \in \mathbb{S}} |G_{\mathbf{xy}}| \leq N^{1-\tau}$, and the fact \hat{G} is independent of S , we have

$$\max_{\mathbf{x}, \mathbf{y} \in \mathbb{X}} \sup_{|H_{ij}| \leq t} |G_{\mathbf{xy}}| \prec 1 + \phi. \quad (3.40)$$

Then a simple induction using (3.16) implies

$$\max_{\mathbf{x}, \mathbf{y} \in \mathbb{X}} \sup_{|H_{ij}| \leq t} |\partial_{ij}^m G_{\mathbf{xy}}| \prec (1 + \phi)^{m+1} \quad (3.41)$$

for every fixed $m \in \mathbb{N}$. Also, by Assumption 1.1 we have a trivial bound

$$\max_{a, b \in \llbracket N \rrbracket} \sup_{|H_{ij}| \leq t} |H_{ab}| \prec 1. \quad (3.42)$$

Now let us estimate the second term on the right-hand side of (3.25). By our definition of f in (3.22), together with (3.16), (3.18), (3.41) and (3.42), one easily shows that

$$\sup_{|x| \leq t} \left| \partial_{ji}^{\ell+1} f(H^{ij} + x\Delta^{ij}) \right| \prec (1 + \phi)^{\ell+1} \cdot N^{100p}$$

for any fixed $\ell \in \mathbb{N}$. Note that $\mathbb{E}|H_{ij}|^{\ell+2} = O(N^{-(\ell+2)/2})$, and by the bound $\phi \leq N^{\tau/10}$ we can find $\ell \equiv \ell(D, p) \geq 1$ such that

$$\mathbb{E}|H_{ij}|^{\ell+2} \cdot \mathbb{E} \left[\sup_{|x| \leq t} \left| \partial_{ij}^{\ell+1} f(H^{ij} + x\Delta^{ij}) \right| \right] = O(N^{-D}).$$

Finally, we estimate the first term on the right-hand side of (3.25). For $a, b \in \llbracket N \rrbracket$, we look at H_{ab} as a function of H . Note that we have the bound

$$\sup_{|x| \leq |H_{ij}|} |H_{ab}(H^{ij} + x\Delta^{ij})| \leq |H_{ab}|$$

uniformly for all $a, b \in \llbracket N \rrbracket$. Together with Assumption 1.1, (3.16), (3.18), and the trivial bound $\sup_{\mathbf{x}, \mathbf{y} \in \mathbb{S}} |G_{\mathbf{xy}}| \leq N$, we have

$$\mathbb{E} \sup_{|x| \leq |H_{ij}|} \left| \partial_{ij}^{\ell+1} f(H^{ij} + x\Delta^{ij}) \right|^2 = O(N^{O_p, \ell(1)}). \quad (3.43)$$

From Assumption 1.1 we find $\max_{i, j} |H_{ij}| \prec N^{-1/2}$, thus by Cauchy-Schwarz inequality we have

$$\mathbb{E} |H_{ij}^{2\ell+4} \mathbf{1}_{|H_{ij}| > t}| = O_{\prec}(N^{-(\ell+2)}) \cdot \mathbb{P}(|H_{ij}| > t) = O(N^{-2D - O_p, \ell(1)}). \quad (3.44)$$

A combination of (3.43), and (3.44) shows that the first term on the right-hand side of (3.25) is bounded by $O(N^{-D})$.

This completes the proof of $\mathcal{R}_{\ell+1}^{ij} = O(N^{-D})$, and our steps also show that the bound is uniform for all $i, j \in \llbracket N \rrbracket$. \square

3.3. Proof of Proposition 3.2. For a fixed $p \in \mathbb{N}$ write

$$\begin{aligned}\mathbb{E}|\underline{\mathcal{D}B}|^{2p} &= \mathbb{E}[(\underline{\mathcal{K}} + \underline{HG})\underline{B} \cdot \underline{\mathcal{D}B}^{p-1} \overline{\underline{\mathcal{D}B}}^p] \\ &= \mathbb{E}[\underline{\mathcal{K}B} \cdot \underline{\mathcal{D}B}^{p-1} \overline{\underline{\mathcal{D}B}}^p] + \frac{1}{N} \sum_{i,j} \mathbb{E}[H_{ij}(GB)_{ji} \underline{\mathcal{D}B}^{p-1} \overline{\underline{\mathcal{D}B}}^p].\end{aligned}\quad (3.45)$$

Using the cumulant expansion (Lemma 2.4) for the second term in (3.45) with $h = H_{ij}$ and $f = f_{ij}(H) = (GB)_{ji} \underline{\mathcal{D}B}^{p-1} \overline{\underline{\mathcal{D}B}}^p$, we obtain

$$\mathbb{E}|\underline{\mathcal{D}B}|^{2p} = \mathbb{E}[\underline{\mathcal{K}B} \cdot \underline{\mathcal{D}B}^{p-1} \overline{\underline{\mathcal{D}B}}^p] + \sum_{k=1}^{\ell} Y_k + \sum_{i,j} \tilde{\mathcal{R}}_{\ell+1}^{ij}, \quad (3.46)$$

where we used the notation

$$Y_k := \frac{1}{N} \sum_{i,j} \frac{1}{k!} \mathcal{C}_{k+1}(H_{ij}) \mathbb{E} \left[\partial_{ij}^k ((GB)_{ji} \underline{\mathcal{D}B}^{p-1} \overline{\underline{\mathcal{D}B}}^p) \right]. \quad (3.47)$$

Note that the sum in (3.46) begins from $k = 1$ because $\mathcal{C}_1(H_{ij}) = 0$. Here ℓ is a fixed positive integer to be chosen later, and $\tilde{\mathcal{R}}_{\ell+1}^{ij}$ is a remainder term defined analogously to $\mathcal{R}_{\ell+1}^{ij}$ in (3.25).

As in the proof of Proposition 3.1, the proof can be broken down into the following lemma.

Lemma 3.5. *Fix $p \geq 2$. Under the assumptions of Theorem 1.5 (ii), we have the following estimates.*

- (i) $\mathbb{E}[\underline{\mathcal{K}B} \cdot \underline{\mathcal{D}B}^{p-1} \overline{\underline{\mathcal{D}B}}^p] + Y_1 = O_{\prec}(\zeta^2) \cdot \mathbb{E}|\underline{\mathcal{D}B}|^{2p-1} + O_{\prec}((1+\phi)^4 \zeta^4) \cdot \mathbb{E}|\underline{\mathcal{D}B}|^{2p-2}$.
- (ii) For $k \geq 2$, $Y_k = (1+\phi)^6 \sum_{n=1}^{2p} O_{\prec}(\zeta^{2n}) \cdot \mathbb{E}|\underline{\mathcal{D}B}|^{2p-n}$.
- (iii) For any $D > 0$, there exists $\ell \equiv \ell(D) \geq 1$ such that $\tilde{\mathcal{R}}_{\ell+1}^{ij} = O(N^{-D})$ uniformly for all $i, j \in \llbracket N \rrbracket$.

Indeed, combining Lemma 3.5 for $\ell = \ell(4p+2)$ together with (3.46) we obtain

$$\mathbb{E}|\underline{\mathcal{D}B}|^{2p} = (1+\phi)^6 \sum_{n=1}^{2p} O_{\prec}(\zeta^{2n}) \cdot \mathbb{E}|\underline{\mathcal{D}B}|^{2p-n} + O(N^{-4p}).$$

From Hölder's inequality we find

$$\mathbb{E}|\underline{\mathcal{D}B}|^{2p} \leq (1+\phi)^6 \sum_{n=1}^{2p} O_{\prec}(\zeta^{2n}) \cdot (\mathbb{E}|\underline{\mathcal{D}B}|^{2p})^{(2p-n)/2p} + O(N^{-4p}),$$

which by Young's inequality implies $\mathbb{E}|\underline{\mathcal{D}B}|^{2p} = O_{\prec}(((1+\phi)^6 \zeta^2)^{2p})$, because $\zeta \geq N^{-1/2}$. Since p was arbitrary, Proposition 3.2 follows by Markov's inequality.

What remains, therefore, is the proof of Lemma 3.5. A crucial ingredient in the proof of Lemma 3.5 is an upper bound on the absolute value of partial derivatives of $\underline{\mathcal{D}B}$. To this end, define the matrix

$$Q_{ij} := \frac{1}{N} (GBHG)_{ij} + \frac{1}{N^2} \sum_{a,b} G_{ia} G_{aj} (GB)_{e_b^a} + \frac{1}{N^2} \sum_{a,b} G_{aa} (GB)_{ib} G_{e_b^a} \quad (3.48)$$

(recall the notation (3.2)) and note that

$$(1 + \delta_{ij})\partial_{ij}\underline{\mathcal{D}B} = \frac{1}{N}(GB)_{ij} + \frac{1}{N}(GB)_{ji} - Q_{ij} - Q_{ji}. \quad (3.49)$$

The following estimate on Q and its derivatives is the key tool behind the proof of Lemma 3.5.

Lemma 3.6. *Suppose that the assumptions of Theorem 1.5 (ii) hold.*

(i) *For every fixed $l \geq 0$ we have $\partial_{ij}^l Q = O_{\prec}((1 + \phi)^{l+1}\zeta^2)$.*

(ii) *We have $Q = O_{\prec}((1 + \phi)^4\zeta^3)$.*

Note that the bound from (ii) is stronger than that from (i) for $l = 0$. Indeed, it relies on a special cancellation that is false for $l > 0$. This cancellation is an essential ingredient of our proof, and our ability to exploit it hinges on the identification of Q as a central object of our analysis. As it turns out, we are able to derive bounds on Q that are closed in the sense that they are self-improving; this allows us to iterate these bounds and hence obtain the optimal bounds on Q stated in Lemma 3.6.

We postpone the proof of Lemma 3.6 and turn to complete the proof of Lemma 3.5.

Proof of Lemma 3.5. Using (3.47) and Lemma 2.5, for every fixed k we have

$$|Y_k| = O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} |\partial_{ij}^k ((GB)_{ji} \underline{\mathcal{D}B}^{p-1} \overline{\mathcal{D}B}^p)|.$$

The last expression can be estimated by a sum, over $r, s, t \geq 0$ such that $r + s + t = k$, of terms of the form

$$O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} |(\partial_{ij}^r (GB)_{ji})(\partial_{ij}^s \underline{\mathcal{D}B}^{p-1})(\partial_{ij}^t \overline{\mathcal{D}B}^p)|. \quad (3.50)$$

To simplify notation, here we ignore the complex conjugate on $\underline{\mathcal{D}B}$, which plays no role in the following analysis and represents a trivial notational complication, and estimate

$$O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} |(\partial_{ij}^r (GB)_{ji})(\partial_{ij}^{k-r} \underline{\mathcal{D}B}^{2p-1})| \quad (3.51)$$

for $r = 0, \dots, k$.

Computing the derivative ∂_{ij}^{k-r} , we find that (3.51) is bounded by a sum of terms of the form

$$O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} \left| (\partial_{ij}^r (GB)_{ji}) \underline{\mathcal{D}B}^{2p-1-q} \prod_{m=1}^q (\partial_{ij}^{l_m} \underline{\mathcal{D}B}) \right|, \quad (3.52)$$

where the sum ranges of integers $q = 0, \dots, (k-r) \wedge (2p-1)$ and $l_1, \dots, l_q \geq 1$ satisfying $l_1 + \dots + l_q = k - r$. Using (3.49) to rewrite the derivative of $\underline{\mathcal{D}B}$ and noting that $\partial_{ij}^l ((GB)_{ji}) \prec (1 + \phi)^{l+1}$ for $l \geq 0$ by (3.3), we conclude that (3.52) is stochastically dominated by

$$O(N^{-\frac{k-1}{2}})(1 + \phi)^{r+1} \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} \left[\left| \prod_{m=1}^q [|\partial_{ij}^{l_m-1}(Q_{ij} + Q_{ji})| + N^{-1}(1 + \phi)^{l_m}] \right| \cdot |\underline{\mathcal{D}B}|^{2p-1-q} \right].$$

Using Lemma 3.6 (i) to estimate the derivatives of Q_{ij} and Q_{ji} , we conclude

$$\begin{aligned} |Y_k| &= \sum_{q=0}^{k \wedge (2p-1)} O(N^{-\frac{k-1}{2}}) \cdot O_{\prec}((1+\phi)^{k+1} \zeta^{2q}) \cdot \mathbb{E}|\underline{\mathcal{D}B}|^{2p-1-q} \\ &\leq \sum_{q=0}^{2p-1} O_{\prec}(N^{-\frac{k-3}{2}} (1+\phi)^{k+1} \zeta^{2q+2}) \cdot \mathbb{E}|\underline{\mathcal{D}B}|^{2p-1-q}. \end{aligned}$$

Noting that $N^{-\frac{k-3}{2}} (1+\phi)^{k+1} \leq (1+\phi)^4$ whenever $k \geq 3$, we conclude the proof for the case $k \geq 3$.

In the remaining cases for Lemma 3.5 (i) and (ii), that is $k \in \{1, 2\}$, this rough argument is not precise enough, and one needs to obtain an additional factor of ζ . As it turns out, those terms are the ones in which $r = 1$. This can be done by a more careful analysis, which we now perform.

(i) *The case $k = 1$.* This case makes use of an algebraic cancellation, and we therefore track the complex conjugates carefully. Using that $\mathcal{C}_2(H_{ij}) = \mathbb{E}[H_{ij}^2]$, the formula for the derivative of G (see (3.16)), and the one for the derivative of $\underline{\mathcal{D}B}$ (see (3.49)), one can verify that

$$\begin{aligned} &(1 + \delta_{ij})(\mathbb{E}[\underline{\mathcal{K}B} \cdot \underline{\mathcal{D}B}^{p-1} \overline{\underline{\mathcal{D}B}}^p] + Y_1) \\ &= -(1 + \delta_{ij})\mathbb{E}[\underline{\mathcal{J}B} \cdot \underline{\mathcal{D}B}^{p-1} \overline{\underline{\mathcal{D}B}}^p] \\ &+ \frac{(p-1)}{N^3} \sum_{i,j} s_{ij} \mathbb{E}[\left((GB)_{ji}(GB)_{ij} + (GB)_{ji}(GB)_{ji}\right) \underline{\mathcal{D}B}^{p-2} \overline{\underline{\mathcal{D}B}}^p] \\ &+ \frac{p}{N^3} \sum_{i,j} s_{ij} \mathbb{E}[\left((GB)_{ji}(\overline{GB})_{ij} + (GB)_{ji}(\overline{GB})_{ji}\right) \underline{\mathcal{D}B}^{p-1} \overline{\underline{\mathcal{D}B}}^{p-1}] \\ &- \frac{p-1}{N^2} \sum_{i,j} s_{ij} \mathbb{E}[(GB)_{ji}(Q_{ij} + Q_{ji}) \underline{\mathcal{D}B}^{p-2} \overline{\underline{\mathcal{D}B}}^p] \\ &- \frac{p}{N^2} \sum_{i,j} s_{ij} \mathbb{E}[(GB)_{ji}(\overline{Q}_{ij} + \overline{Q}_{ji}) \underline{\mathcal{D}B}^{p-1} \overline{\underline{\mathcal{D}B}}^{p-1}]. \end{aligned} \tag{3.53}$$

Note that on the right-hand side of this identity a crucial cancellation took place: the first line comes from the sum of the term $\underline{\mathcal{K}B}$ and the term $r = 1$ in Y_1 , which almost cancel, yielding the small term $\underline{\mathcal{J}B}$.

Taking an absolute value, using (3.14) for the first term, the Ward identity and the fact that $s_{ij} = O(1)$ for the second and third terms and Lemma 3.6 (ii) together with the Ward identity for the last two terms, the claim follows.

(ii) *The case $k = 2$.* From Lemma 2.5 we get $|\mathcal{C}_3(H_{ij})| = O(N^{-3/2})$, and therefore

$$|Y_2| = O(N^{-5/2}) \cdot \sum_{i,j} \mathbb{E}|\partial_{ij}^2((GB)_{ji} \underline{\mathcal{D}B}^{p-1} \overline{\underline{\mathcal{D}B}}^p)|. \tag{3.54}$$

As before, we ignore the complex conjugates to simplify notation, and estimate

$$|Y_2| = O(N^{-5/2}) \cdot \sum_{i,j} \mathbb{E}|\partial_{ij}^2((GB)_{ji} \underline{\mathcal{D}B}^{2p-1})|. \tag{3.55}$$

instead of (3.54). The proof is performed using an explicit computation of the second derivative

$$\begin{aligned}\partial_{ij}^2((GB)_{ji}\underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-1}) &= (\partial_{ij}^2(GB)_{ji})\underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-1} + 2(2p-1)(\partial_{ij}(GB)_{ji})(\partial_{ij}\underline{\mathcal{D}}\underline{\mathcal{B}})\underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-2} \\ &\quad + (2p-1)(GB)_{ji}(\partial_{ij}^2\underline{\mathcal{D}}\underline{\mathcal{B}})\underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-2} + (2p-1)(2p-2)(GB)_{ji}(\partial_{ij}\underline{\mathcal{D}}\underline{\mathcal{B}})^2\underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-3} \\ &=: (1+\delta_{ij})^{-2}(Z_{ij}^1\underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-1} + Z_{ij}^2\underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-2} + Z_{ij}^3\underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-3}),\end{aligned}$$

in self-explanatory notation.

Starting with Z_{ij}^1 , we have

$$(1+\delta_{ij})^2 Z_{ij}^1 = 2G_{ji}G_{ji}(GB)_{ji} + 2G_{jj}G_{ii}(GB)_{ji} + 2G_{ji}G_{jj}(GB)_{ii} + 2G_{jj}G_{ij}(GB)_{ii}.$$

Using the bound $N^{-1/2} \leq \zeta$, the Ward identity, and (3.3) on each of the four terms, we obtain

$$O(N^{-5/2}) \sum_{i,j} \mathbb{E}|Z_{ij}^1 \underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-1}| \prec O_{\prec}((1+\phi)^2 \zeta^2) \mathbb{E}|\underline{\mathcal{D}}\underline{\mathcal{B}}|^{2p-1},$$

as desired.

The first term of Z_{ij}^2 yields the contribution

$$\begin{aligned}O(N^{-5/2}) \sum_{i,j} \mathbb{E}|(\partial_{ij}(GB)_{ji})(\partial_{ij}\underline{\mathcal{D}}\underline{\mathcal{B}})\underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-2}| \\ = O(N^{-5/2}) \sum_{i,j} \mathbb{E}|(\partial_{ij}(GB)_{ji})(N^{-1}(GB)_{ij} + N^{-1}(GB)_{ji} - Q_{ij} - Q_{ji})\underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-2}|,\end{aligned}$$

where we used (3.49). Using (3.3), (3.16), the bound $N^{-1/2} \leq \zeta$, and the Ward identity, one can estimate the contribution of the first two terms by $O_{\prec}((1+\phi)^2 \zeta^4) \mathbb{E}|\underline{\mathcal{D}}\underline{\mathcal{B}}|^{2p-2}$. Using (3.3), Lemma 3.6 (ii), and the Ward identity, one can bound the contribution of the last two terms by $O_{\prec}((1+\phi)^6 \zeta^4) \mathbb{E}|\underline{\mathcal{D}}\underline{\mathcal{B}}|^{2p-2}$, as desired.

In the same spirit, using Lemma 3.6 (i), the second term of Z_{ij}^2 is of the form

$$\begin{aligned}O(N^{-5/2}) \sum_{i,j} \mathbb{E}|(GB)_{ji}(\partial_{ij}^2\underline{\mathcal{D}}\underline{\mathcal{B}})\underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-2}| \\ = O(N^{-5/2}) \sum_{i,j} \mathbb{E}|(GB)_{ji}(\partial_{ij}(N^{-1}((GB)_{ij} + (GB)_{ji}) - (Q_{ij} + Q_{ji})))\underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-2}| \\ = O(N^{-1/2}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E}|(GB)_{ji} O_{\prec}((1+\phi)^2 \zeta^2) \underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-2}| = O((1+\phi)^2 \zeta^4) \mathbb{E}|\underline{\mathcal{D}}\underline{\mathcal{B}}|^{2p-2}.\end{aligned}$$

Finally, the contribution of Z_{ij}^3 is of the form

$$O(N^{-5/2}) \sum_{i,j} \mathbb{E}|(GB)_{ji}(\partial_{ij}\underline{\mathcal{D}}\underline{\mathcal{B}})^2 \underline{\mathcal{D}}\underline{\mathcal{B}}^{2p-3}|,$$

using Lemma 3.6 (i) to bound $(\partial_{ij}\underline{\mathcal{D}}\underline{\mathcal{B}})^2 = O_{\prec}((1+\phi)^4 \zeta^4)$ and using the Ward identity for $(GB)_{ji}$, we conclude that the last term equals $O_{\prec}((1+\phi)^4 \zeta^6)$, as desired. This concludes the proof of Lemma 3.5 (i) and (ii).

(iii) *The remainder term.* The proof is similar to that of Lemma 3.4 (iii), and we omit the details. \square

3.4. Proof of Lemma 3.6. We start with several observations regarding the matrix Q . Define the random matrix

$$F_{ab} \equiv F(H)_{ab} := \frac{1}{N}(BH)_{ab} + \frac{\delta_{ab}}{N^2} \sum_c (GB)_{e_c^b} + \frac{1}{N^2} B_{ab} \sum_c G_{e_c^b}.$$

Thus we have

$$Q = GFG,$$

and therefore by (3.16) we have for all $\mathbf{v}, \mathbf{w} \in \mathbb{S}$

$$(1 + \delta_{ij})\partial_{ij}Q_{\mathbf{v}\mathbf{w}} = -(G_{\mathbf{v}i}Q_{j\mathbf{w}} + G_{\mathbf{v}j}Q_{i\mathbf{w}} + Q_{\mathbf{v}i}G_{j\mathbf{w}} + Q_{\mathbf{v}j}G_{i\mathbf{w}}) + E_{\mathbf{v}\mathbf{w}}^{ij}, \quad (3.56)$$

where $E^{ij} := (1 + \delta_{ij})G(\partial_{ij}F)G$.

Proof of Lemma 3.6 (i). We prove the claim by induction on l . We start with the case $l = 0$. Fix $\mathbf{v}, \mathbf{w} \in \mathbb{S}$. Using (3.48), (3.3), and the Ward identity, we obtain

$$\begin{aligned} |Q_{\mathbf{v}\mathbf{w}}| &\leq \frac{1}{N}|(GBHG)_{\mathbf{v}\mathbf{w}}| + \frac{1}{N^2} \sum_{a,b} |G_{\mathbf{v}a}G_{a\mathbf{w}}(GB)_{e_b^a}| + \frac{1}{N^2} \sum_{a,b} |G_{aa}(GB)_{\mathbf{v}b}G_{e_b^a\mathbf{w}}| \\ &\prec \frac{1}{N} \sqrt{|(GBB^*G^*)_{\mathbf{v}\mathbf{v}}|} \sqrt{|(G^*H^*HG)_{\mathbf{w}\mathbf{w}}|} + (1 + \phi)\zeta^2. \end{aligned}$$

Using the bound (3.3) the Ward identity, and Lemma 2.2, we therefore get

$$\frac{1}{N}|(G^*H^*HG)_{\mathbf{w}\mathbf{w}}| \leq \frac{1}{N}\|H\|^2(G^*G)_{\mathbf{w}\mathbf{w}} = \|H\|^2 \frac{\text{Im } G_{\mathbf{w}\mathbf{w}}}{N\eta} \leq \|H\|^2 \zeta^2 \prec \zeta^2.$$

Similarly, using the assumption $\|B\| = O(1)$ we find $\frac{1}{N}|(GBB^*G^*)_{\mathbf{v}\mathbf{v}}| \leq O(\zeta^2)$. We conclude that $|Q_{\mathbf{v}\mathbf{w}}| \prec (1 + \phi)\zeta^2$, which completes the proof of the case $l = 0$.

We now perform the induction step. Suppose that $l \geq 1$ and $\partial_{ij}^m Q = O_{\prec}((1 + \phi)^{m+1}\zeta^2)$ for $0 \leq m \leq l - 1$. By (3.56) we have

$$(1 + \delta_{ij})\partial_{ij}^l Q_{\mathbf{v}\mathbf{w}} = -\partial_{ij}^{l-1}(G_{\mathbf{v}i}Q_{j\mathbf{w}} + G_{\mathbf{v}j}Q_{i\mathbf{w}} + Q_{\mathbf{v}i}G_{j\mathbf{w}} + Q_{\mathbf{v}j}G_{i\mathbf{w}}) + \partial_{ij}^{l-1}E_{\mathbf{v}\mathbf{w}}^{ij}. \quad (3.57)$$

We deal with each of the terms on the right-hand side separately. The term $\partial_{ij}^{l-1}(G_{\mathbf{v}i}Q_{j\mathbf{w}})$ is estimated using (3.17) as

$$\begin{aligned} |\partial_{ij}^{l-1}(G_{\mathbf{v}i}Q_{j\mathbf{w}})| &\leq \sum_{m=0}^{l-1} \binom{l-1}{m} |(\partial_{ij}^m G_{\mathbf{v}i})(\partial_{ij}^{l-1-m} Q_{j\mathbf{w}})| \prec \sum_{m=0}^{l-1} (1 + \phi)^{m+1} |\partial_{ij}^{l-1-m} Q_{j\mathbf{w}}| \\ &\prec \sum_{m=0}^{l-1} (1 + \phi)^{m+1} (1 + \phi)^{l-m} \zeta^2 = O((1 + \phi)^{l+1} \zeta^2), \end{aligned}$$

where in the third step we used the induction assumption. The three subsequent terms of (3.57) are estimated analogously.

In order to estimate the last term of (3.57), we write

$$\begin{aligned}
E_{\mathbf{vw}}^{ij} &= (1 + \delta_{ij}) \sum_{a,b} G_{\mathbf{va}} (\partial_{ij} F_{ab}) G_{b\mathbf{w}} \\
&= \frac{1}{N} (GB)_{\mathbf{vi}} G_{j\mathbf{w}} + \frac{1}{N} (GB)_{\mathbf{vj}} G_{i\mathbf{w}} \\
&\quad - \frac{1}{N^2} \sum_{a,b} G_{\mathbf{va}} G_{a\mathbf{w}} G_{\mathbf{e}_b^a i} (GB)_{jb} - \frac{1}{N^2} \sum_{a,b} G_{\mathbf{va}} G_{a\mathbf{w}} G_{\mathbf{e}_b^a j} (GB)_{ib} \\
&\quad - \frac{1}{N^2} \sum_{a,b} (GB)_{\mathbf{va}} G_{a\mathbf{w}} G_{\mathbf{e}_b^a i} G_{jb} - \frac{1}{N^2} \sum_{a,b} (GB)_{\mathbf{va}} G_{a\mathbf{w}} G_{\mathbf{e}_b^a j} G_{ib}.
\end{aligned} \tag{3.58}$$

Using (3.16), the assumption $\|B\| = O(1)$ and the bound (3.3) it follows that

$$\partial_{ij}^{l-1} \left(\frac{1}{N} (GB)_{\mathbf{vi}} G_{j\mathbf{w}} + \frac{1}{N} (GB)_{\mathbf{vj}} G_{i\mathbf{w}} \right) \prec \frac{1}{N} (1 + \phi)^{l+1} \leq (1 + \phi)^{l+1} \zeta^2.$$

What remains, therefore, is to estimate ∂_{ij}^{l-1} applied to the four last terms of (3.58). They are all similar, and we consider $\partial_{ij}^{l-1} \left(\frac{1}{N^2} \sum_{a,b} G_{\mathbf{va}} G_{a\mathbf{w}} G_{\mathbf{e}_b^a i} (GB)_{jb} \right)$ for definiteness. By computing the derivatives and using (3.16), we obtain a sum of terms, each of which is a product of $l + 3$ factors of the form $G_{\mathbf{xy}}$ for some $\mathbf{x}, \mathbf{y} \in \{\mathbf{v}, \mathbf{w}, \mathbf{e}_a, \mathbf{e}_b, \mathbf{e}_b^a, \mathbf{e}_i, \mathbf{e}_j\}$ and one factor of the form $(GB)_{\mathbf{x}b}$ for some $\mathbf{x} \in \{i, j\}$. Furthermore, exactly two of those factors are of the form $G_{\mathbf{x}a}$ or $G_{a\mathbf{x}}$ with $\mathbf{x} \in \{\mathbf{v}, \mathbf{w}, \mathbf{e}_i, \mathbf{e}_j\}$. Applying the Ward identity with the sum over a to the two factors containing an index a , and estimating the remaining $l + 1$ factors using (3.3), we conclude that

$$\partial_{ij}^{l-1} \left(\frac{1}{N^2} \sum_{a,b} G_{\mathbf{va}} G_{a\mathbf{w}} G_{\mathbf{e}_b^a i} (GB)_{jb} \right) \prec (1 + \phi)^{l+1} \zeta^2,$$

as desired. This concludes the proof. \square

Proof of Lemma 3.6 (ii). We have to improve the naive bound $Q = O_{\prec}((1 + \phi)\zeta^2)$ from part (i) by an order ζ . We do this by deriving a stochastic self-improving estimate on Q , which makes use of a crucial cancellation analogous to, but more subtle than, the one in (3.53). We derive this self-improving bound using another high-moment estimate for the entries of Q , which is derived using the cumulant expansion from Lemma 2.4.

To this end, fix $\mathbf{v}, \mathbf{w} \in \mathbb{S}$ and $p \in \mathbb{N}$. From (3.48) we immediately get

$$Q_{\mathbf{vw}} = \frac{1}{N} (GBHG)_{\mathbf{vw}} + \frac{1}{N^2} \sum_{i,j} G_{\mathbf{vj}} G_{j\mathbf{w}} (GB)_{\mathbf{e}_i^j} + \frac{1}{N^2} \sum_{i,j} G_{jj} (GB)_{\mathbf{vi}} G_{\mathbf{e}_i^j \mathbf{w}}.$$

Thus,

$$\mathbb{E}|Q_{\mathbf{vw}}|^{2p} = \mathbb{E} \left[\mathcal{W}_{\mathbf{vw}} Q_{\mathbf{vw}}^{p-1} \overline{Q}_{\mathbf{vw}}^p \right] + \frac{1}{N} \sum_{i,j} \mathbb{E} \left[(GB)_{\mathbf{vi}} H_{ij} G_{j\mathbf{w}} Q_{\mathbf{vw}}^{p-1} \overline{Q}_{\mathbf{vw}}^p \right],$$

where we defined

$$\mathcal{W}_{\mathbf{vw}} := \frac{1}{N^2} \sum_{i,j} G_{\mathbf{vj}} G_{j\mathbf{w}} (GB)_{\mathbf{e}_i^j} + \frac{1}{N^2} \sum_{i,j} G_{jj} (GB)_{\mathbf{vi}} G_{\mathbf{e}_i^j \mathbf{w}}.$$

Using the cumulant expansion (Lemma 2.4) for the last term on the right-hand side with $h = H_{ij}$ and $f(H_{ij}) = (GB)_{vi}G_{jw}Q_{vw}^{p-1}\overline{Q}_{vw}^p$, we obtain

$$\mathbb{E}|Q_{vw}|^{2p} = \mathbb{E}\left[\mathcal{W}_{vw}Q_{vw}^{p-1}\overline{Q}_{vw}^p\right] + \sum_{k=1}^{\ell} Z_k + \sum_{i,j} \hat{\mathcal{R}}_{\ell+1}^{ij}, \quad (3.59)$$

where we used the notation

$$Z_k = \frac{1}{N} \sum_{i,j} \frac{1}{k!} \mathcal{C}_{k+1}(H_{ij}) \mathbb{E}\left[\partial_{ij}^k \left((GB)_{vi}G_{jw}Q_{vw}^{p-1}\overline{Q}_{vw}^p\right)\right]$$

Here ℓ is a fixed positive integer to be chosen later, and $\hat{\mathcal{R}}_{\ell+1}^{ij}$ is a remainder term defined analogously to $\mathcal{R}_{\ell+1}^{ij}$ in (3.25).

The proof of Lemma 3.6 can be broken into the following estimates.

Lemma 3.7. *Suppose that $Q = O_{\prec}(\lambda)$ for some $\lambda \geq (1 + \phi)^4 \zeta^3$.*

(i) *We have*

$$\mathbb{E}\left[\mathcal{W}_{vw}Q_{vw}^{p-1}\overline{Q}_{vw}^p\right] + Z_1 = O_{\prec}(\zeta^3) \mathbb{E}|Q_{vw}|^{2p-1} + O_{\prec}(\zeta^3 \lambda) \mathbb{E}|Q_{vw}|^{2p-2}$$

(ii) *For $k \geq 2$ we have*

$$Z_k = \sum_{n=1}^{2p} (1 + \phi)^{2n} O_{\prec}((\zeta^3 \lambda)^{n/2}) \mathbb{E}|Q_{vw}|^{2p-n}. \quad (3.60)$$

(iii) *For any $D > 0$, there exists $\ell \equiv \ell(D) \geq 1$ such that $\hat{\mathcal{R}}_{\ell+1}^{ij} = O(N^{-D})$ uniformly for all $i, j \in \llbracket N \rrbracket$.*

We now conclude the proof of Lemma 3.6 (ii) using Lemma 3.7. Suppose that $Q = O_{\prec}(\lambda)$ for some $\lambda \geq \zeta^3$. Combining Lemma 3.7 for $\ell = \ell(6p + 2)$ together with (3.59) we obtain

$$\mathbb{E}|Q_{vw}|^{2p} = \sum_{n=1}^{2p} O_{\prec}((1 + \phi)^{2n} (\zeta^3 \lambda)^{n/2}) \cdot \mathbb{E}|Q_{vw}|^{2p-n} + O(N^{-6p}).$$

Since $\zeta \geq N^{-1/2}$, it follows from the Hölder's inequality that

$$\mathbb{E}|Q_{vw}|^{2p} \leq \sum_{n=1}^{2p} O_{\prec}((1 + \phi)^{2n} (\zeta^3 \lambda)^{n/2}) \cdot (\mathbb{E}|Q_{vw}|^{2p})^{\frac{2p-n}{2p}},$$

and therefore by Young's inequality

$$\mathbb{E}|Q_{vw}|^{2p} = O_{\prec}(((1 + \phi)^4 \zeta^3 \lambda)^p).$$

Next, since p was arbitrary, we conclude from Markov's inequality that

$$Q = O_{\prec}(\lambda) \quad \implies \quad Q = O_{\prec}\left(\left((1 + \phi)^4 \zeta^3 \lambda\right)^{1/2}\right) \quad (3.61)$$

for any $\lambda \geq (1 + \phi)^4 \zeta^3$. Moreover, by Lemma 3.6 (i) we have the a priori bound $Q = O_{\prec}((1 + \phi)\zeta^2)$. An iteration of (3.61), analogous to Lemma 2.6, yields $Q = O_{\prec}((1 + \phi)^4 \zeta^3)$, as claimed. \square

What remains, therefore, is the proof of Lemma 3.7.

Proof of Lemma 3.7. We begin with (i). A straightforward computation yields

$$\begin{aligned}
& \mathbb{E} \left[\mathcal{W}_{\mathbf{vw}} Q_{\mathbf{vw}}^{p-1} \overline{Q}_{\mathbf{vw}}^p \right] + Z_1 \\
&= -\frac{1}{N^2} \sum_{i,j} \mathbb{E} \left[G_{\mathbf{ve}_i^j} (GB)_{ji} G_{j\mathbf{w}} Q_{\mathbf{vw}}^{p-1} \overline{Q}_{\mathbf{vw}}^p \right] - \frac{1}{N^2} \sum_{i,j} \mathbb{E} \left[(GB)_{\mathbf{ve}_i^j} G_{ji} G_{j\mathbf{w}} Q_{\mathbf{vw}}^{p-1} \overline{Q}_{\mathbf{vw}}^p \right] \\
&+ \frac{p-1}{N^2} \sum_{i,j} \mathbb{E} \left[(GB)_{\mathbf{ve}_i^j} G_{j\mathbf{w}} (\partial_{ij} Q_{\mathbf{vw}}) Q_{\mathbf{vw}}^{p-2} \overline{Q}_{\mathbf{vw}}^p \right] + \frac{p}{N^2} \sum_{i,j} \mathbb{E} \left[(GB)_{\mathbf{ve}_i^j} G_{j\mathbf{w}} (\partial_{ij} \overline{Q}_{\mathbf{vw}}) Q_{\mathbf{vw}}^{p-1} \overline{Q}_{\mathbf{vw}}^{p-1} \right].
\end{aligned} \tag{3.62}$$

We emphasize that here an important cancellation takes place between the two terms on the left-hand side, whereby two large terms arising from the computation of Z_1 precisely cancel out the two terms of $\mathcal{W}_{\mathbf{vw}}$. Using the Ward identity and $s_{ij} = O(1)$, one can verify that the absolute value of the first two terms on the right-hand side of (3.62) is $O_{\prec}(\zeta^3) \mathbb{E} |Q_{\mathbf{vw}}|^{2p-1}$. Hence, it suffices to show that the remaining two terms are $O_{\prec}(\zeta^3 \lambda) \mathbb{E} |Q_{\mathbf{vw}}|^{2p-2}$. Taking for example the third term on the right-hand side of (3.62), by (3.56) we find

$$\begin{aligned}
& \left| \frac{1}{N^2} \sum_{i,j} \mathbb{E} \left[(GB)_{\mathbf{ve}_i^j} G_{j\mathbf{w}} (\partial_{ij} Q_{\mathbf{vw}}) Q_{\mathbf{vw}}^{p-2} \overline{Q}_{\mathbf{vw}}^p \right] \right| \\
& \leq \frac{1}{N^2} \sum_{i,j} \mathbb{E} \left[\left| (GB)_{\mathbf{ve}_i^j} G_{j\mathbf{w}} (G_{\mathbf{vi}} Q_{j\mathbf{w}} + G_{\mathbf{vj}} Q_{i\mathbf{w}} + Q_{\mathbf{vi}} G_{j\mathbf{w}} + Q_{\mathbf{vj}} G_{i\mathbf{w}}) \right| \cdot |Q_{\mathbf{vw}}|^{2p-2} \right] \\
& \quad + \frac{1}{N^2} \sum_{i,j} \mathbb{E} \left[\left| (GB)_{\mathbf{ve}_i^j} G_{j\mathbf{w}} E_{\mathbf{vw}}^{ij} \right| \cdot |Q_{\mathbf{vw}}|^{2p-2} \right].
\end{aligned}$$

Due to the assumption $G = O_{\prec}(\lambda)$ and using the Ward identity, one can bound the first term on the right-hand side by $O_{\prec}(\zeta^3 \lambda) \mathbb{E} |Q_{\mathbf{vw}}|^{2p-2}$. As for the last term, using (3.58) and the Ward identity on the indices i, j and when possible also a, b , we find that it is $O_{\prec}(\zeta^6) \mathbb{E} |Q_{\mathbf{vw}}|^{2p-2}$. This concludes the proof of part (i).

Next, we prove (ii). Fix $k \geq 2$ and estimate

$$\begin{aligned}
|Z_k| &= O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} \left| \partial_{ij}^k \left((GB)_{\mathbf{vi}} G_{j\mathbf{w}} Q_{\mathbf{vw}}^{p-1} \overline{Q}_{\mathbf{vw}}^p \right) \right| \\
&= O(N^{-\frac{k-1}{2}}) \sum_{\substack{r,s,t \geq 0 \\ r+s+t=k}} \frac{1}{N^2} \sum_{i,j} \mathbb{E} \left| (\partial_{ij}^r \left((GB)_{\mathbf{vi}} G_{j\mathbf{w}} \right)) (\partial_{ij}^s Q_{\mathbf{vw}}^{p-1}) (\partial_{ij}^t \overline{Q}_{\mathbf{vw}}^p) \right|.
\end{aligned}$$

As the sum over r, s, t is finite it suffices to deal with each term separately. To simplify notation, we drop the complex conjugates of Q (which play no role in the subsequent analysis), and estimate the quantity

$$O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} \left| (\partial_{ij}^r \left((GB)_{\mathbf{vi}} G_{j\mathbf{w}} \right)) (\partial_{ij}^{k-r} Q_{\mathbf{vw}}^{2p-1}) \right| \tag{3.63}$$

for $r = 0, \dots, k$. Computing the derivative ∂_{ij}^{k-r} , we find that (3.63) is bounded by a sum of terms of the form

$$O(N^{-\frac{k-1}{2}}) \frac{1}{N^2} \sum_{i,j} \mathbb{E} \left| (\partial_{ij}^r \left((GB)_{\mathbf{vi}} G_{j\mathbf{w}} \right)) \left(\prod_{m=1}^q (\partial_{ij}^{l_m} Q_{\mathbf{vw}}) \right) Q_{\mathbf{vw}}^{2p-1-q} \right|, \tag{3.64}$$

where the sum ranges over integers $q = 0, \dots, (k-r) \wedge (2p-1)$ and $l_1, \dots, l_q \geq 1$ satisfying $l_1 + \dots + l_q = k-r$. Using Lemma 3.6 (i), we find that (3.64) is bounded by

$$O_{\prec}(N^{-\frac{k-1}{2}}) \frac{1}{N^2} \sum_{i,j} \mathbb{E} [|\partial_{ij}^r((GB)_{\mathbf{v}i} G_{j\mathbf{w}})| (1+\phi)^{k-r+q} \zeta^{2q} |Q_{\mathbf{v}\mathbf{w}}|^{2p-1-q}].$$

Note that by (3.16) the derivative $\partial_{ij}^r((GB)_{\mathbf{v}i} G_{j\mathbf{w}})$ can be written as a sum of terms, each of which is a product of $r+2$ entries of the matrices GB or G , with one entry of the form $(GB)_{\mathbf{v}a}$ or $G_{\mathbf{v}a}$ with $a \in \{i, j\}$ and one of the form $G_{a\mathbf{w}}$ with $a \in \{i, j\}$. Using the Ward identity for the two specified terms in the product and using (3.3) to bound the remaining terms in the product, we conclude that $\frac{1}{N^2} \sum_{i,j} |(\partial_{ij}^r((GB)_{\mathbf{v}i} G_{j\mathbf{w}}))| \prec (1+\phi)^r \zeta^2$, and therefore

$$\begin{aligned} (3.64) &\prec O(N^{-\frac{k-1}{2}}) (1+\phi)^{k+q} \zeta^{2q+2} \mathbb{E} |Q_{\mathbf{v}\mathbf{w}}|^{2p-1-q} \\ &= O(N^{-\frac{k-1}{2}} \zeta^{-q-1} (1+\phi)^{k+q}) \zeta^{3q+3} \mathbb{E} |Q_{\mathbf{v}\mathbf{w}}|^{2p-1-q} \\ &\leq O(\zeta^{k-2-q} (1+\phi)^{q+1}) \zeta^{3q+3} \mathbb{E} |Q_{\mathbf{v}\mathbf{w}}|^{2p-1-q}, \end{aligned} \quad (3.65)$$

where for the last inequality we used $N^{-1/2}(1+\phi) \leq \zeta$. Clearly, if $q \leq k-2$ then (3.65) is bounded by $O_{\prec}((1+\phi)^{q+1} \zeta^{3q+3}) \cdot \mathbb{E} |Q_{\mathbf{v}\mathbf{w}}|^{2p-1-q}$, as desired.

What remains, therefore, is to estimate (3.64) for $q \geq k-1$, which we assume from now on. Because $k \geq 2$ by assumption, we find that $q \geq 1$. Moreover, since $q = 0, \dots, (k-r) \wedge (2p-1)$, we find that $r \leq 1$. Thus, it remains to consider the three cases $(r, q) = (0, k)$, $(r, q) = (1, k-1)$ and $(r, q) = (0, k-1)$. We deal with them separately.

The case $(r, q) = (0, k)$. In this case $l_1 = l_2 = \dots = l_q = 1$, so that (3.64) reads

$$\begin{aligned} &O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} |(GB)_{\mathbf{v}i} G_{j\mathbf{w}} (\partial_{ij} Q_{\mathbf{v}\mathbf{w}})^k Q_{\mathbf{v}\mathbf{w}}^{2p-1-k}| \\ &\leq O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} |(GB)_{\mathbf{v}i} G_{j\mathbf{w}} (G_{\mathbf{v}i} Q_{j\mathbf{w}} + G_{\mathbf{v}j} Q_{i\mathbf{w}} + Q_{\mathbf{v}i} G_{j\mathbf{w}} + Q_{\mathbf{v}j} G_{i\mathbf{w}}) (\partial_{ij} Q_{\mathbf{v}\mathbf{w}})^{k-1} Q_{\mathbf{v}\mathbf{w}}^{2p-1-k}| \\ &\quad + O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} |(GB)_{\mathbf{v}i} G_{j\mathbf{w}} E_{\mathbf{v}\mathbf{w}}^{ij} (\partial_{ij} Q_{\mathbf{v}\mathbf{w}})^{k-1} Q_{\mathbf{v}\mathbf{w}}^{2p-1-k}|, \end{aligned} \quad (3.66)$$

where for the inequality we used (3.56).

We estimate the second line of (3.66) using the Ward identity on the summations over i and j , using the bound $Q = O_{\prec}(\lambda)$, and the bound $\partial_{ij} Q_{\mathbf{v}\mathbf{w}} \prec (1+\phi)^2 \zeta^2$ from Lemma 3.6 (i). The result is

$$O(N^{-\frac{k-1}{2}}) \zeta^3 \lambda (1+\phi)^{2k} \zeta^{2(k-1)} \mathbb{E} |Q_{\mathbf{v}\mathbf{w}}|^{2p-1-k} \leq (1+\phi)^{2k} \zeta^{3k} \lambda \mathbb{E} |Q_{\mathbf{v}\mathbf{w}}|^{2p-1-k},$$

as desired, where we used $N^{-1/2} \leq \zeta$. For the third line of (3.66), we use (3.58) and the Ward identity to obtain

$$\frac{1}{N^2} \sum_{i,j} |(GB)_{\mathbf{v}i} G_{j\mathbf{w}} E_{\mathbf{v}\mathbf{w}}^{ij}| \prec \zeta^6,$$

which, together with the bound $|\partial_{ij} Q_{\mathbf{v}\mathbf{w}}| \prec (1+\phi)^2 \zeta^2$ obtained in Lemma 3.6 (i), gives the bound

$$(1+\phi)^{2k} \zeta^{3k+3} \mathbb{E} |Q_{\mathbf{v}\mathbf{w}}|^{2p-1-k}$$

for the third line of (3.66). Since $\zeta^3 \leq \lambda$, this bound is good enough.

The case $(r, q) = (1, k - 1)$. In this case $l_1 = l_2 = \dots = l_{k-1} = 1$, and (3.64) reads

$$O(N^{-\frac{k-1}{2}}) \frac{1}{N^2} \sum_{i,j} \mathbb{E} |(\partial_{ij}((GB)_{vi}G_{jw}))(\partial_{ij}Q_{vw})^{k-1} Q_{vw}^{2p-k}|. \quad (3.67)$$

Using (3.56) to rewrite one factor $\partial_{ij}Q_{vw}$, we conclude that (3.67) equals

$$\begin{aligned} O(N^{-\frac{k-1}{2}}) \frac{1}{N^2} \sum_{i,j} \mathbb{E} |(\partial_{ij}((GB)_{vi}G_{jw}))(Q_{vi}G_{jw} + Q_{vj}G_{iw} + G_{vi}Q_{jw} + G_{vj}Q_{iw})(\partial_{ij}Q_{vw})^{k-2} Q_{vw}^{2p-k}| \\ + O(N^{-\frac{k-1}{2}}) \frac{1}{N^2} \sum_{i,j} \mathbb{E} |(\partial_{ij}((GB)_{vi}G_{jw}))E_{vw}^{ij}(\partial_{ij}Q_{vw})^{k-2} Q_{vw}^{2p-k}|. \end{aligned}$$

We now apply a similar argument to the one used in the previous case $(r, q) = (0, k)$. We use (3.58), the assumption $Q = O_{\prec}(\lambda)$ to bound Q , and the Ward identity to bound the product of entries of G and (GB) . This gives

$$\frac{1}{N^2} \sum_{i,j} |(\partial_{ij}((GB)_{vi}G_{jw}))(\partial_{ij}Q_{vw})| \prec (1 + \phi)(\lambda\zeta^2 + \zeta^5).$$

Using $\zeta^3 \leq \lambda$ and the bound $|\partial_{ij}Q_{vw}|^{k-2} \prec (1 + \phi)^{2k-4} \zeta^{2k-4}$ from Lemma 3.6 (i), we therefore get

$$\begin{aligned} (3.67) &= O(N^{-\frac{k-1}{2}})(1 + \phi)\lambda\zeta^2(1 + \phi)^{2k-4}\zeta^{2k-4}\mathbb{E}|Q_{vw}|^{2p-k} \\ &\prec (1 + \phi)^{2k}\zeta^{3k-3}\lambda\mathbb{E}|Q_{vw}|^{2p-k}, \end{aligned}$$

as desired

The case $(r, q) = (0, k - 1)$. Since $l_1 + \dots + l_{k-1} = k$ and $l_m \geq 1$ for every $m \in \{1, \dots, k - 1\}$, there exists exactly one m such that $l_m = 2$ and the remaining l_m 's are 1. Hence, (3.64) reads

$$\begin{aligned} O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} |(GB)_{vi}G_{jw}(\partial_{ij}^2 Q_{vw})(\partial_{ij}Q_{vw})^{k-2} Q_{vw}^{2p-k}| \\ \leq O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} [|(GB)_{vi}G_{jw}(\partial_{ij}(G_{vi}Q_{jw} + G_{vj}Q_{iw} + Q_{vi}G_{jw} + Q_{vj}G_{iw}))(\partial_{ij}Q_{vw})^{k-2} Q_{vw}^{2p-k}|] \\ + O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} |(GB)_{vi}G_{jw}(\partial_{ij}E_{vw}^{ij})(\partial_{ij}Q_{vw})^{k-2} Q_{vw}^{2p-k}|, \quad (3.68) \end{aligned}$$

where we used (3.56). We deal with each of the terms on the right-hand side separately. For the third line of (3.68), using the derivative formulas (3.16) and (3.58) together with the Ward identity, we find

$$\frac{1}{N^2} \sum_{i,j} |(GB)_{vi}G_{jw}\partial_{ij}E_{vw}^{ij}| \prec (1 + \phi)^3\zeta^5 \leq (1 + \phi)^3\lambda\zeta^2. \quad (3.69)$$

Using (3.69) which we can estimate the third line of (3.68) by the right-hand side of (3.60), as in the previous case.

All four terms in the second line of (3.68) are similar, and we estimate

$$\begin{aligned}
& O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} |(GB)_{\mathbf{v}i} G_{j\mathbf{w}} (\partial_{ij} G_{\mathbf{v}i} Q_{j\mathbf{w}}) (\partial_{ij} Q_{\mathbf{v}\mathbf{w}})^{k-2} Q_{\mathbf{v}\mathbf{w}}^{2p-k}| \\
&= O(N^{-\frac{k-1}{2}}) \cdot \frac{1}{N^2} \sum_{i,j} \mathbb{E} |(GB)_{\mathbf{v}i} G_{j\mathbf{w}} (-G_{\mathbf{v}i} G_{ji} Q_{j\mathbf{w}} - G_{\mathbf{v}j} G_{ii} Q_{j\mathbf{w}} + G_{\mathbf{v}i} \partial_{ij} Q_{j\mathbf{w}}) (\partial_{ij} Q_{\mathbf{v}\mathbf{w}})^{k-2} Q_{\mathbf{v}\mathbf{w}}^{2p-k}|.
\end{aligned} \tag{3.70}$$

Using the bound $Q = O_{\prec}(\lambda)$ to bound $Q_{j\mathbf{w}}$, the bound $|\partial_{ij} Q_{\mathbf{v}\mathbf{w}}| \prec (1+\phi)^2 \zeta^2$ from Lemma 3.6 (i) and the Ward identity (which can be applied to obtain at least ζ^3 in the last term and ζ^2 in the first two terms), we conclude that (3.70) is bounded by

$$O(N^{-\frac{k-1}{2}}) (1+\phi)^2 (\zeta^2 \lambda + \zeta^5) (1+\phi)^{2k-4} \zeta^{2k-4} \mathbb{E} |Q_{\mathbf{v}\mathbf{w}}|^{2p-k} = (1+\phi)^{2k-2} \zeta^{3k-3} \lambda \mathbb{E} |Q_{\mathbf{v}\mathbf{w}}|^{2p-k},$$

which is bounded by the right-hand side of (3.60). This concludes the proof of part (ii).

Finally, the proof of (iii) is similar to that of Lemma 3.4 (iii), and we omit the details. \square

4. Proof of Theorem 1.9

4.1. Preliminaries. We start with a few simple deterministic estimates.

Lemma 4.1. *Under the assumptions of Theorem 1.9 we have $\eta = O(\text{Im } m)$.*

Proof. From $\|A\| = O(1)$, (1.7), and (1.9), it is not hard to deduce that the support of ρ lies in a ball of radius $O(1)$ around the origin. The claim then follows from the fact that $\text{Im } m(z) = \int \frac{\eta}{(E-x)^2 + \eta^2} \nu(dx) \geq c\eta$, since $(E-x) = O(1)$ on the support of ν . \square

Lemma 4.2. *Under the assumptions of Theorem 1.9 we have $\|\text{Im } M\| = O(\text{Im } m)$.*

Proof. By the resolvent identity we have

$$\|\text{Im } M\| = \|M - M^*\|/2 = \|MM^*\| \text{Im}(z + m) = \|MM^*\|(\eta + \text{Im } m),$$

and the claim follows by Lemma 4.1. \square

For the remainder of this section we abbreviate $\delta := \tau/10$.

Definition 4.3. A *control parameter* is a deterministic continuous function $\phi : \mathbf{S} \rightarrow [N^{-1}, N^\delta]$ such that $\eta \mapsto \phi(E + i\eta)$ is decreasing for each E .

For a control parameter ϕ we define

$$h(\phi) := \sqrt{\frac{\text{Im } m + \phi}{N\eta}}. \tag{4.1}$$

The following result is a simple consequence of Theorem 1.5 applied to the special case of Wigner matrices.

Lemma 4.4. *Suppose that the assumptions of Theorem 1.9 hold. Let ϕ be a control parameter. Let $z \in \mathbf{S}$ and suppose that $G - M = O_{\prec}(\phi)$ at z .*

(i) *We have at z*

$$\Pi(G) = O_{\prec}((1 + \phi)^3 h(\phi)).$$

(ii) *For any deterministic $B \in \mathbb{C}^{N \times N}$ satisfying $\|B\| = O(1)$ we have at z*

$$\underline{B\Pi(G)} = O_{\prec}((1 + \phi)^6 h(\phi)^2).$$

Proof. To prove (i), by Lemmas 4.1 and 4.2, it suffices to prove

$$\mathcal{S}(G)G - gG = O_{\prec}((1 + \phi)h(\phi)), \quad \mathcal{S}(M)M - mM = O_{\prec}((1 + \phi)h(\phi)), \quad (4.2)$$

where \mathcal{S} denotes the map from (1.4). By (3.11) we have $\mathcal{S}(G)G = \mathcal{J} + \mathcal{K}$, and (3.13) shows $\mathcal{J} = O_{\prec}((1 + \phi)\zeta) = O_{\prec}((1 + \phi)h(\phi))$. Also, Assumption 1.7 shows $s_{ij} = 1 + O(\delta_{ij})$, and we obtain from (3.11) that

$$|\mathcal{K}_{\mathbf{v}\mathbf{w}} - gG_{\mathbf{v}\mathbf{w}}| = O(1) \cdot \frac{1}{N} \sum_i |G_{ii}G_{i\mathbf{w}}v_i| = O_{\prec}((1 + \phi)h(\phi))$$

for any fixed $\mathbf{v}, \mathbf{w} \in \mathbb{S}$, where in the last step we used (3.3), $|v_i| \leq 1$, and the Ward identity. This proves the first estimate of (4.2). The second estimate of (4.2) is proved similarly.

The proof of (ii) is analogous, using the estimates (3.14) and

$$\begin{aligned} |\underline{\mathcal{K}B} - g\underline{GB}| &= O(1) \cdot \frac{1}{N^2} \sum_{i,j} |G_{ii}G_{ij}B_{ji}| \\ &= O_{\prec}(1 + \phi) \cdot \left(\frac{1}{N^2} \sum_{i,j} |G_{ij}|^2 \right)^{1/2} \cdot \left(\frac{1}{N^2} \sum_{i,j} |B_{ij}|^2 \right)^{1/2} = O_{\prec}((1 + \phi)h(\phi)^2), \end{aligned}$$

where in the last step we used Lemma 2.3 and the estimate $\sum_{i,j} |B_{ij}|^2 = \text{Tr} BB^* = O(N)$, and Lemmas 4.1 and 4.2. This concludes the proof. \square

For $u \in \mathbb{C}$ define

$$R_u := (A - u - z)^{-1}.$$

With this notation, we have

$$M = R_m, \quad G = R_g - R_g \Pi(G). \quad (4.3)$$

Taking the difference yields

$$G - M = R_g - R_m - R_g \Pi(G). \quad (4.4)$$

Moreover, taking the trace of the second identity of (4.3) yields

$$\pi(g) = -\underline{R_g \Pi(G)}, \quad (4.5)$$

where we defined

$$\pi(u) \equiv \pi(u, z) := u - \underline{R_u} = u - \int \frac{\nu(\mathrm{d}a)}{a - u - z}.$$

Note that $\pi(m) = 0$, by (1.7). The formulas (4.4) and (4.5) are the main identities behind the following analysis.

In order to estimate the error terms on the right-hand side of (4.4) and (4.5) using Lemma 4.4, we need to deal with the fact that the matrix R_g is random.

Definition 4.5. An event Ω holds with high probability if $\mathbf{1}(\Omega^c) \prec 0$, i.e. if $\mathbb{P}(\Omega^c) \leq N^{-D}$ for all $D > 0$ and large enough N depending on D .

Lemma 4.6. Suppose that the assumptions of Lemma 4.4 hold. Suppose that $\|R_g\| = O(1)$ with high probability. Then the conclusions (i) and (ii) of Lemma 4.4 hold with $\Pi(G)$ replaced by $R_g\Pi(G)$.

Proof. Let Ω an event of high probability such that $\mathbf{1}(\Omega)\|R_g\| = O(1)$, and define the set $\mathcal{U} := \{g \equiv g(H) : H \in \Omega\} \subset \mathbb{C}$. Since $|g| \leq N$, we find that \mathcal{U} is contained in the ball of radius N around the origin. Let $\hat{\mathcal{U}}$ be an N^{-3} net of \mathcal{U} , i.e. a set $\hat{\mathcal{U}} \subset \mathcal{U}$ such that $|\hat{\mathcal{U}}| = O(N^8)$ and for each $u \in \mathcal{U}$ there exists a $\hat{u} \in \hat{\mathcal{U}}$ such that $|u - \hat{u}| \leq N^{-3}$.

By a union bound, from Lemma 4.4 (i) it follows that for any $\mathbf{v}, \mathbf{w} \in \mathbb{S}$

$$\sup_{\hat{u} \in \hat{\mathcal{U}}} |(R_{\hat{u}}\Pi(G))_{\mathbf{v}\mathbf{w}}| \prec (1 + \phi)^3 h(\phi). \quad (4.6)$$

Writing

$$|(R_g\Pi(G))_{\mathbf{v}\mathbf{w}}| \leq |(R_{\hat{g}}\Pi(G))_{\mathbf{v}\mathbf{w}}| + |((R_{\hat{g}} - R_g)\Pi(G))_{\mathbf{v}\mathbf{w}}|$$

and estimating the first term by $O_{\prec}((1 + \phi)^3 h(\phi))$ using (4.6) and the second term by $\|R_{\hat{g}} - R_g\| \|\Pi(G)\| = O(|\hat{g} - g| \|\Pi(G)\|) = O(N^{-1})$ (since $\|\Pi(G)\| = O(N^2)$ trivially), we conclude that $R_g\Pi(G) = O_{\prec}((1 + \phi)^3 h(\phi))$. Here we used that $N^{-1/2} = O(h(\phi))$ by Lemma 4.1. The quantity $\underline{BR_g\Pi(G)}$ is estimated analogously. This concludes the proof. \square

4.2. Proof of the local law. We begin by estimating $g - m$ for large η .

Lemma 4.7. Under the assumptions of Theorem 1.9, we have $g - m = O_{\prec}((N\eta)^{-1})$ for $\eta \geq 2$.

Proof. For $\eta \geq 1$ we have $\|G\| \leq 1$ and therefore $G - M = O_{\prec}(1)$. Moreover, we have $\|R_g\| \leq 1$ and $\text{Im } m \leq 1$. From (4.5), $\pi(m) = 0$, and Lemma 4.6 we therefore find

$$g - m = (g - m) \int \frac{\nu(\mathrm{d}a)}{(a - g - z)(a - m - z)} + O_{\prec}\left(\frac{1}{N\eta}\right).$$

For $\eta \geq 2$ we have $|a - g - z||a - m - z| \geq \text{Im}(a - g - z)\text{Im}(a - m - z) \geq 4$, and the claim follows since ν is a probability measure. \square

Lemma 4.8. Suppose that at $z \in \mathbf{D}$ we have $\|M\| = O(1)$, $G - M = O_{\prec}(N^\delta)$, and $g - m = O_{\prec}(\theta)$ for some control parameter $\theta \leq N^{-\delta}$. Then we have at z

$$G - M = O_{\prec}(\theta + \psi), \quad \psi := \sqrt{\frac{\text{Im } m}{N\eta}} + \frac{1}{N\eta}.$$

Proof. From the resolvent identity $R_g = R_m + R_g R_m (g - m)$ and the assumptions $|g - m| \prec N^{-\delta}$ and $\|R_m\| = \|M\| = O(1)$, we conclude that $\|R_g\| = O(1)$ with high probability.

Now suppose that $G - M = O_{\prec}(\phi)$ for some control parameter ϕ . Then from (4.4) and Lemma 4.6 we find

$$G - M = R_g R_m (g - m) + O_{\prec}((1 + \phi)^3 h(\phi)).$$

Using that $\|R_g\| = O(1)$ with high probability, we deduce the implication

$$G - M = O_{\prec}(\phi) \quad \implies \quad G - M = O_{\prec}(\theta + (1 + \phi)^3 h(\phi)). \quad (4.7)$$

The implication (4.7) is a self-improving bound, which we iterate to obtain successively better bounds on $G - M$. The iteration gives rise to a sequence ϕ_0, ϕ_1, \dots , where $\phi_0 := N^\delta$ and $\phi_{k+1} := \theta + (1 + \phi_k)^3 h(\phi_k)$. By the definition of \prec , we easily conclude the claim after a bounded number of iterations. Note that here we used the fact that $\delta = \tau/10$ in order to make sure that $\phi_1 = O(1)$, and for Theorem 1.5, and thus Lemma 4.4, to hold. See also Lemma 2.6 and its proof. \square

Lemma 4.8 provides a bound on $G - M$ starting from a bound on $g - m$ and a rough bound on $G - M$. The needed bound on $g - m$ is obtained from a stochastic continuity argument, which propagates the bound on $g - m$ from large values of η to small values of η . It makes use of the following notion of stability of the equation (1.7).

Definition 4.9 (Stability). Let $\mathbf{S} \subset \mathbf{D}$ be a spectral domain. We say that (1.7) is *stable on \mathbf{S}* if the following holds for some constant $C > 0$. Suppose that the line $\mathcal{L} := \{E\} \times [a, b]$ is a subset of \mathbf{S} . Let $\xi : \mathcal{L} \rightarrow [N^{-1}, N^{-\delta/2}]$ be continuous such that $\eta \mapsto \xi(E + i\eta)$ is decreasing on $[a, b]$. Suppose that $|\pi(g)| \leq \xi$ for all $z \in \mathcal{L}$, and

$$|g - m| \leq \frac{C\xi}{\operatorname{Im} m + \sqrt{\xi}} \quad (4.8)$$

for $z = E + ib$. Then (4.8) holds for all $z \in \mathcal{L}$.

For the following we fix $E \in \mathbb{R}$, and establish (1.10) and (1.11) for $z \in (\{E\} \times \mathbb{R}) \cap \mathbf{S} = \bigcup_{k=0}^K \mathcal{L}_k$, where $\mathcal{L}_0 := (\{E\} \times [2, \infty)) \cap \mathbf{S}$ and $\mathcal{L}_k := (\{E\} \times [2N^{-\delta k}, 2N^{-\delta(k-1)}]) \cap \mathbf{S}$ for $k = 1, \dots, K$. Here $K \leq 1/\delta$.

The stochastic continuity argument is an induction on k . We start with $k = 0$. By Lemmas 4.7 and 4.8 with $\theta = (N\eta)^{-1}$ we have (1.10) and (1.11) for $z \in \mathcal{L}_0$, since $G - M = O_{\prec}(1)$ trivially.

The induction step follows from the following result.

Lemma 4.10. *Fix $k = 1, \dots, K$ and suppose that (1.10) and (1.11) hold for all $z \in \mathcal{L}_{k-1}$. Then (1.10) and (1.11) hold for all $z \in \mathcal{L}_k$.*

Proof. Denote by $b_k := E + i2N^{-\delta(k-1)}$ the upper edge of the line \mathcal{L}_k . First we note that by a simple monotonicity of the resolvent (see e.g. [10, Lemma 10.2]), the estimate $G = O_{\prec}(1)$ for $z = b_k$ implies $G = O_{\prec}(N^\delta)$ for $z \in \mathcal{L}_k$. Setting $\phi := N^\delta$ in Lemma 4.6, we find from (4.5) that $|\pi(g)| \prec N^{7\delta}(N\eta)^{-1}$ for all $z \in \mathcal{L}_k$. Using that $\pi(g)$ is N^2 -Lipschitz in \mathbf{D} , we find using a simple N^{-3} -net argument that with high probability we have $|\pi(g)| \leq (N\eta)^{-1/2}$ for all $z \in \mathcal{L}_k$; here we also used that $N^{8\delta}(N\eta)^{-1} \leq (N\eta)^{-1/2}$ by definition of δ . Moreover, from (1.11), we find that with high probability we have $|g - m| \leq (N\eta)^{-1/2}$ for $z = b_k$. From Definition 4.9 we therefore deduce that

$$|g - m| \prec (N\eta)^{-1/4} \quad \text{for all } z \in \mathcal{L}_k \quad (4.9)$$

Next, suppose that $\theta \leq N^{-\delta}$ is a control parameter and $|g - m| \prec \theta$ for $z \in \mathcal{L}_k$. In particular, from $R_g = R_m + R_g R_m (g - m)$ we deduce that $\|R_g\| = O(1)$ with high probability for all $z \in \mathcal{L}_k$.

From Lemma 4.8 we deduce that $G - M = O_{\prec}(\theta + \psi)$. By Lemma 4.6 with $\phi = \theta + \psi$ and (4.5) we therefore have

$$|\pi(g)| \prec \xi := \frac{\operatorname{Im} m}{N\eta} + \frac{1}{(N\eta)^2} + \frac{\theta}{N\eta}. \quad (4.10)$$

It is easy to verify that ξ is a control parameter. Moreover, by the induction assumption we have

$$|g - m| \prec \frac{1}{N\eta} \leq \frac{C\xi}{\operatorname{Im} m + \sqrt{\xi}}.$$

From an N^{-3} -net argument on \mathcal{L}_k analogous to the one given in the previous paragraph, we conclude using Definition 4.9 that for all $z \in \mathcal{L}_k$ we have

$$|g - m| \prec \frac{\xi}{\operatorname{Im} m + \sqrt{\xi}} \leq \frac{C}{N\eta} + \sqrt{\frac{\theta}{N\eta}}.$$

In conclusion, for every $z \in \mathcal{L}_k$ we have the implication

$$|g - m| \prec \theta \quad \implies \quad |g - m| \prec \frac{1}{N\eta} + \sqrt{\frac{\theta}{N\eta}}.$$

Starting from (4.9) and iterating this implication a bounded number of times and using the definition of \prec , we obtain (1.11). See Lemma 2.6 and its proof. Now (1.10) follows from Lemma 4.8. This concludes the proof. \square

We have established the claim of Theorem 1.9 on the whole line $(\{E\} \times \mathbb{R}) \cap \mathbf{S}$. Since $E \in \mathbb{R}$ was arbitrary, this concludes the proof of Theorem 1.9.

A. Proof of Lemma 2.4

In this section we prove the cumulant expansion formula with the remainder term (2.3). We start with an elementary inequality.

Lemma A.1. *Let X be a nonnegative random variable with finite moments. Then for any $a, b, t \geq 0$, we have*

$$\mathbb{E}X^a \mathbb{E}[X^b \mathbf{1}_{X>t}] \leq \mathbb{E}[X^{a+b} \mathbf{1}_{X>t}].$$

Proof. It suffices to assume $a > 0$. Let us abbreviate $\|X\|_a := (\mathbb{E}X^a)^{1/a}$. For $t \geq \|X\|_a$, we have

$$\mathbb{E}X^a \mathbb{E}[X^b \mathbf{1}_{X>t}] \leq \mathbb{E}[t^a X^b \mathbf{1}_{X>t}] \leq \mathbb{E}[X^{a+b} \mathbf{1}_{X>t}], \quad (\text{A.1})$$

which is the desired result. For $t < \|X\|_a$, we have

$$\mathbb{E}X^a \mathbb{E}[X^b \mathbf{1}_{X \leq t}] > \mathbb{E}[t^a X^b \mathbf{1}_{X \leq t}] \geq \mathbb{E}[X^{a+b} \mathbf{1}_{X \leq t}]. \quad (\text{A.2})$$

Jensen's (or Hölder's) inequality yields

$$\mathbb{E}X^a \mathbb{E}X^b \leq \mathbb{E}X^{a+b}, \quad (\text{A.3})$$

and the claim follows from (A.2) and (A.3), using $1 = \mathbf{1}_{X \leq t} + \mathbf{1}_{X > t}$. \square

Let $\chi(t) := \log \mathbb{E} e^{ith}$. For $n \geq 1$, we have

$$\partial_t^n (e^{\chi(t)}) = \partial_t^{n-1} (\chi'(t) e^{\chi(t)}) = \sum_{k=1}^n \binom{n-1}{k-1} \partial_t^k (\chi(t)) \partial_t^{n-k} (e^{\chi(t)}),$$

hence

$$\mathbb{E} h^n = (-i)^n \partial_t^n e^{\chi(t)} \Big|_{t=0} = \sum_{k=1}^n \binom{n-1}{k-1} \mathcal{C}_k(h) \mathbb{E} h^{n-k}.$$

For $g(h) = h^\ell$, we have

$$\mathbb{E}[h \cdot g(h)] = \mathbb{E} h^{\ell+1} = \sum_{k=1}^{\ell+1} \binom{\ell}{k-1} \mathcal{C}_k(h) \mathbb{E} h^{\ell+1-k} = \sum_{k=0}^{\ell} \frac{1}{k!} \mathcal{C}_{k+1}(h) \mathbb{E}[g^{(k)}(h)],$$

and by linearity the same relation holds for any polynomial P of degree $\leq \ell$:

$$\mathbb{E}[h \cdot P(h)] = \sum_{k=0}^{\ell} \frac{1}{k!} \mathcal{C}_{k+1}(h) \mathbb{E}[P^{(k)}(h)]. \quad (\text{A.4})$$

Next, let f be as in the statement of Lemma 2.4, and fix $\ell \in \mathbb{N}$. By Taylor expansion we can find a polynomial P of degree at most ℓ , such that for any $0 \leq k \leq \ell$,

$$f^{(k)}(h) = P^{(k)}(h) + \frac{1}{(\ell+1-k)!} f^{(\ell+1)}(\xi_k) h^{\ell+1-k}, \quad (\text{A.5})$$

where $\xi_k \equiv \xi_k(h)$ is a random variable taking values between 0 and h .

By (A.4), (A.5), homogeneity of the cumulants, and Jensen's inequality we find that the error term in (2.2) satisfies

$$\begin{aligned} \mathcal{R}_{\ell+1} &= \mathbb{E}[h \cdot f(h)] - \sum_{k=0}^{\ell} \frac{1}{k!} \mathcal{C}_{k+1}(h) \mathbb{E}[f^{(k)}(h)] \\ &= \mathbb{E}[h \cdot (f(h) - P(h))] - \sum_{k=0}^{\ell} \frac{1}{k!} \mathcal{C}_{k+1}(h) \mathbb{E}[f^{(k)}(h) - P^{(k)}(h)] \\ &= \frac{1}{(\ell+1)!} \mathbb{E}[f^{(\ell+1)}(\xi_0) \cdot h^{\ell+2}] - \sum_{k=0}^{\ell} \frac{1}{k!(\ell+1-k)!} \mathcal{C}_{k+1}(h) \mathbb{E}[f^{(\ell+1)}(\xi_k) \cdot h^{\ell+1-k}] \\ &\leq O(1) \cdot \sum_{k=0}^{\ell+1} \mathbb{E}|h|^k \cdot \mathbb{E} \left[\sup_{|x| \leq |h|} |f^{(\ell+1)}(x)| \cdot h^{\ell+2-k} \right] \\ &\leq O(1) \cdot \mathbb{E}|h|^{\ell+2} \cdot \sup_{|x| \leq t} |f^{(\ell+1)}(x)| + O(1) \cdot \sum_{k=0}^{\ell+1} \mathbb{E}|h|^k \cdot \mathbb{E} \left[\sup_{|x| \leq |h|} |f^{(\ell+1)}(x)| \cdot h^{\ell+2-k} \cdot \mathbf{1}_{|h| > t} \right]. \end{aligned} \quad (\text{A.6})$$

The desired result then follows from estimating the last term of (A.6) by Cauchy-Schwarz inequality and Lemma A.1.

B. The complex case

In this appendix we explain how to generalize our results to the complex case. We briefly explain how to conclude the proof of Theorem 1.5 (i); Theorem 1.5 (ii) follows in a similar fashion.

If H is complex Hermitian, in general $\mathbb{E}H_{ij}^2$ and $\mathbb{E}|H_{ij}|^2$ are different. We define, in addition to (3.1),

$$t_{ij} := (1 + \delta_{ij})^{-1} N \mathbb{E}H_{ij}^2, \quad (\text{B.1})$$

and, in addition to (3.2), for a vector $\mathbf{x} = (x_i)_{i \in \mathbb{N}} \in \mathbb{C}^N$,

$$\underline{\mathbf{x}}^j = (\underline{x}_i^j)_{i \in [N]}, \quad \text{where} \quad \underline{x}_i^j := x_i t_{ij}. \quad (\text{B.2})$$

Now (3.11) generalizes to

$$(\mathcal{S}(G)G)_{\mathbf{vw}} = \mathcal{J}_{\mathbf{vw}} + \mathcal{K}_{\mathbf{vw}}, \quad \mathcal{J}_{\mathbf{vw}} := \frac{1}{N} \sum_j G_{j\bar{v}j} G_{j\mathbf{w}}, \quad \mathcal{K}_{\mathbf{vw}} := \frac{1}{N} \sum_j G_{jj} G_{\mathbf{v}j\mathbf{w}}, \quad (\text{B.3})$$

and $\mathcal{D}_{\mathbf{vw}}$ remains the same as in (3.12).

Following a similar argument as in Section 3.1, we see that Proposition 3.1 is enough to conclude Theorem 1.5 (i). As in (3.21), we fix a constant $p \in \mathbb{N}$ and write

$$\mathbb{E}[|\mathcal{D}_{\mathbf{vw}}|^{2p}] = \mathbb{E}[\mathcal{K}_{\mathbf{vw}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p] + \sum_{i,j} v_i \mathbb{E}[H_{ij} G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p]. \quad (\text{B.4})$$

We compute the second term on the right-hand side of (B.4) using the complex cumulant expansion, given in [27, Lemma 7.1], which replaces the real cumulant expansion from Lemma 2.4. The result is

$$\mathbb{E}|\mathcal{D}_{\mathbf{vw}}|^{2p} = \mathbb{E}[\mathcal{K}_{\mathbf{vw}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p] + \sum_{k=1}^{\ell} \tilde{X}_k + \sum_{i,j} v_i \tilde{\mathcal{R}}_{\ell+1}^{ij}, \quad (\text{B.5})$$

where \tilde{X}_k and $\tilde{\mathcal{R}}_{\ell+1}^{ij}$ are defined analogously to (3.24) and (3.25), respectively. Using the same proof, one can easily extend Lemma 3.4 (ii)–(iii) with X_k and $\mathcal{R}_{\ell+1}^{ij}$ replaced by \tilde{X}_k and $\tilde{\mathcal{R}}_{\ell+1}^{ij}$, so that it remains to estimate $\mathbb{E}[\mathcal{K}_{\mathbf{vw}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p] + \tilde{X}_1$. Using the complex cumulant expansion, we find

$$\begin{aligned} \tilde{X}_1 &= \sum_{i,j} \left(v_i \mathbb{E}[H_{ij}^2] \mathbb{E}[\partial_{ji}(G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p)] + v_i \mathbb{E}[H_{ij}^2] \mathbb{E}[\partial_{ij}(G_{j\mathbf{w}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p)] \right) (1 + \delta_{ij})^{-1} \\ &= \sum_{i,j} \left(v_i \mathbb{E}[H_{ij}^2] \mathbb{E}[G_{j\mathbf{w}} \partial_{ji}(\mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p)] + v_i \mathbb{E}[H_{ij}^2] \mathbb{E}[G_{j\mathbf{w}} \partial_{ij}(\mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p)] \right) (1 + \delta_{ij})^{-1} \\ &\quad - \mathbb{E}[(\mathcal{S}(G)G)_{\mathbf{vw}} \mathcal{D}_{\mathbf{vw}}^{p-1} \overline{\mathcal{D}}_{\mathbf{vw}}^p], \end{aligned}$$

where in the second step we used (B.3) and

$$\frac{\partial G_{ij}}{\partial H_{kl}} = -G_{ik} G_{lj}. \quad (\text{B.6})$$

(Here we take the conventions of [27, Section 7] for the derivatives in the complex entries of H .) Thus we have

$$\begin{aligned} \mathbb{E}[\mathcal{K}_{\mathbf{vw}}\mathcal{D}_{\mathbf{vw}}^{p-1}\overline{\mathcal{D}}_{\mathbf{vw}}^p] + \tilde{X}_1 &= -\mathbb{E}[\mathcal{J}_{\mathbf{vw}}\mathcal{D}_{\mathbf{vw}}^{p-1}\overline{\mathcal{D}}_{\mathbf{vw}}^p] + \sum_{i,j} \left(v_i \mathbb{E}[H_{ij}^2] \mathbb{E}[G_{jw}\partial_{ji}(\mathcal{D}_{\mathbf{vw}}^{p-1}\overline{\mathcal{D}}_{\mathbf{vw}}^p)] \right. \\ &\quad \left. + v_i \mathbb{E}[H_{ij}^2] \mathbb{E}[G_{jw}\partial_{ij}(\mathcal{D}_{\mathbf{vw}}^{p-1}\overline{\mathcal{D}}_{\mathbf{vw}}^p)] \right) (1 + \delta_{ij})^{-1}. \end{aligned} \tag{B.7}$$

By estimating the right-hand side of (B.7) in a similar way as we dealt with the right-hand side of (3.27), we obtain

$$\mathbb{E}[\mathcal{K}_{\mathbf{vw}}\mathcal{D}_{\mathbf{vw}}^{p-1}\overline{\mathcal{D}}_{\mathbf{vw}}^p] + \tilde{X}_1 = O_{\prec}(\tilde{\zeta}) \cdot \mathbb{E}|\mathcal{D}_{\mathbf{vw}}|^{2p-1} + O_{\prec}(\tilde{\zeta}\lambda) \cdot \mathbb{E}|\mathcal{D}_{\mathbf{vw}}|^{2p-2},$$

which completes the proof.

References

- [1] O. Ajanki, L. Erdős, and T. Krüger, *Local eigenvalue statistics for random matrices with general short range correlations*, Preprint arXiv:1604.08188.
- [2] ———, *Quadratic vector equations on complex upper half-plane*, Preprint arXiv:1506.05095.
- [3] ———, *Singularities of solutions to quadratic vector equations on complex upper half-plane*, Preprint arXiv:1512.03703.
- [4] ———, *Universality for general Wigner-type matrices*, Preprint arXiv:1506.05098.
- [5] ———, *Local spectral statistics of Gaussian matrices with correlated entries*, J. Stat. Phys. **163** (2016), 280–302.
- [6] G.W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, vol. 118, Cambridge university press, 2010.
- [7] Z. Bao, L. Erdős, and K. Schnelli, *Local law of addition of random matrices on optimal scale*, Preprint arXiv:1509.07080.
- [8] ———, *Convergence rate for spectral distribution of addition of random matrices*, Preprint arXiv:1606.03076 (2016).
- [9] R. Bauerschmidt, A. Knowles, and H.-T. Yau, *Local semicircle law for random regular graphs*, Preprint arXiv:1503.08702.
- [10] F. Benaych-Georges and A. Knowles, *Lectures on the local semicircle law for Wigner matrices*, Preprint arXiv:1601.04055.
- [11] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Isotropic local laws for sample covariance and generalized Wigner matrices*, Electron. J. Probab **19** (2014), 1–53.
- [12] A. Bloemendal, A. Knowles, H.-T. Yau, and J. Yin, *On the principal components of sample covariance matrices*, Prob. Theor. Rel. Fields **164** (2016), 459–552.

- [13] P. Bourgade, J. Huang, and H.-T. Yau, *Eigenvector statistics of sparse random matrices*, Preprint arXiv:1609.09022.
- [14] P. Bourgade and H.-T. Yau, *The eigenvector moment flow and local quantum unique ergodicity*, Comm. in Math. Phys. **350** (2017), 231–278.
- [15] Z. Che, *Universality of random matrices with correlated entries*, Preprint arXiv:1604.05709.
- [16] A. Boutet de Monvel and A. Khorunzhy, *Asymptotic distribution of smoothed eigenvalue density. II. Wigner random matrices*, Random Oper. and Stoch. Equ. **7** (1999), 149–168.
- [17] L. Erdős and T. Krüger, In preparation.
- [18] L. Erdős, A. Knowles, and H.-T. Yau, *Averaging fluctuations in resolvents of random band matrices*, Ann. H. Poincaré **14** (2013), 1837–1926.
- [19] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Delocalization and diffusion profile for random band matrices*, Comm. Math. Phys. **323** (2013), 367–416.
- [20] ———, *The local semicircle law for a general class of random matrices*, Electron. J. Probab **18** (2013), 1–58.
- [21] ———, *Spectral statistics of Erdős-Rényi graphs I: Local semicircle law*, Ann. Prob. **41** (2013), 2279–2375.
- [22] L. Erdős, B. Schlein, and H.-T. Yau, *Local semicircle law and complete delocalization for Wigner random matrices*, Comm. Math. Phys. **287** (2009), 641–655.
- [23] L. Erdős, H.-T. Yau, and J. Yin, *Universality for generalized Wigner matrices with Bernoulli distribution*, J. Combinatorics **1** (2011), no. 2, 15–85.
- [24] ———, *Bulk universality for generalized Wigner matrices*, Prob. Theor. Rel. Fields **154** (2012), 341–407.
- [25] ———, *Rigidity of eigenvalues of generalized Wigner matrices*, Adv. Math **229** (2012), 1435–1515.
- [26] V.L. Girko, *Theory of stochastic canonical equations*, vol. 2, Springer, 2001.
- [27] Y. He and A. Knowles, *Mesosopic eigenvalue statistics of Wigner matrices*, Preprint arXiv:1603.01499.
- [28] J.W. Helton, R.R. Far, and R. Speicher, *Operator-valued semicircular elements: solving a quadratic matrix equation with positivity constraints*, Int. Math. Res. Not. **2007** (2007).
- [29] A.M. Khorunzhy, B.A. Khoruzhenko, and L.A. Pastur, *Asymptotic properties of large random matrices with independent entries*, J. Math. Phys. **37** (1996), 5033–5060.
- [30] A. Knowles and J. Yin, *Anisotropic local laws for random matrices*, Preprint arXiv:1410.3516.
- [31] A. Knowles and J. Yin, *The isotropic semicircle law and deformation of Wigner matrices*, Comm. Pure Appl. Math. **66** (2013), 1663–1749.

- [32] J. O. Lee and K. Schnelli, *Edge universality for deformed Wigner matrices*, Rev. Math. Phys. **27** (2015), 1550018.
- [33] J.O. Lee and K. Schnelli, *Local law and Tracy-Widom limit for sparse random matrices*, Preprint arXiv:1605.08767.
- [34] ———, *Local deformed semicircle law and complete delocalization for Wigner matrices with random potential*, J. of Math. Phys. **54** (2013), 103504.
- [35] J.O. Lee, K. Schnelli, B. Stetler, and H.-T. Yau, *Bulk universality for deformed Wigner matrices*, Ann. Prob. **44** (2016), 2349–2425.
- [36] M. L. Mehta, *Random matrices*, Academic press, 2004.
- [37] L.A. Pastur and M. Shcherbina, *Eigenvalue distribution of large random matrices*, vol. 171, Amer. Math. Soc., 2011.
- [38] N. S. Pillai and J. Yin, *Universality of covariance matrices*, Ann. Appl. Probab. **24** (2014), 935–1001.
- [39] D. V. Voiculescu, K. J. Dykema, and A. Nica, *Free random variables: A noncommutative probability approach to free products with applications to random matrices, operator algebras and harmonic analysis on free groups*, vol. 1, American Mathematical Society, 1992.
- [40] E.P. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, Ann. Math. **62** (1955), 548–564.