The Annals of Applied Probability
2020, Vol. 30, No. 6, 2846–2879
https://doi.org/10.1214/20-AAP1575

BULK EIGENVALUE FLUCTUATIONS OF SPARSE RANDOM MATRICES

BY YUKUN HE

Institute of Mathematics, University of Zürich, yukun.he@math.uzh.ch

We consider a class of sparse random matrices, which includes the adjacency matrix of Erdős–Rényi graphs $\mathcal{G}(N, p)$ for $p \in [N^{\varepsilon-1}, N^{-\varepsilon}]$. We identify the joint limiting distributions of the eigenvalues away from 0 and the spectral edges. Our result indicates that unlike Wigner matrices, the eigenvalues of sparse matrices satisfy central limit theorems with normalization $N\sqrt{p}$. In addition, the eigenvalues fluctuate simultaneously: the correlation of two eigenvalues of the same/different sign is asymptotically 1/-1. We also prove CLTs for the eigenvalue counting function and trace of the resolvent at mesoscopic scales.

1. Introduction and statements of results. Let $\mathcal{A}$ be the adjacency matrix of a sparse Erdős–Rényi graph $\mathcal{G}(N, p)$. That is, $\mathcal{A}$ is a symmetric $N \times N$ matrix with independent upper triangular entries satisfying

$$\mathcal{A}_{ij} = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases}$$

Note that each row and column of $\mathcal{A}$ has typically Np nonzero entries, and we are interested in the case when $\mathcal{A}$ is sparse; more precisely, we set $p \in [N^{-1+\varepsilon}, N^{-\varepsilon}]$ for some fixed $\varepsilon > 0$. It is convenient to introduce the normalized matrix

$$A := \sqrt{\frac{1}{p(1-p)N}}\mathcal{A} \tag{1.1}$$

so that the typical eigenvalue spacing of A is of order N^{-1}. We also introduce the new variable

$$q := \sqrt{Np}.$$

In this paper, we consider random matrices of the following class; it is an easy exercise to check that A defined in (1.1) in terms of $\mathcal{G}(N, p)$ satisfies the following conditions.

DEFINITION 1.1 (Sparse matrix). Fix $\beta \in (0, 1/2)$ and set $q := N^{\beta}$. A sparse matrix is a real symmetric $N \times N$ matrix $H = H^* \in \mathbb{R}^{N\times N}$ whose entries H_{ij} satisfy the following conditions.

(i) The upper-triangular entries $(H_{ij} : 1 \le i \le j \le N)$ are independent.
(ii) The off-diagonal entries $(H_{ij} : i \ne j)$ are identically distributed.
(iii) We have $\mathbb{E}H_{ij} = 0$ and $\mathbb{E}H_{ij}^2 = (1 + O(\delta_{ij}))/N$ for all i, j.
(iv) For any $k \ge 3$, we have $\mathbb{E}|H_{ij}|^k \le C_k/(Nq^{k-2})$ for all i, j.

We define the adjacency matrix A by

$$A = H + f\mathbf{e}\mathbf{e}^*,$$

where $\mathbf{e} := N^{-1/2}(1, 1, \ldots, 1)^*$, and $f \ge 0$.

Received May 2019; revised November 2019.
MSC2020 subject classifications. 05C80, 15B52, 60B20, 05C50.
Key words and phrases. Random matrices, sparse Erdős–Rényi graphs, CLT.

A special case of the above model is the Wigner matrix. Recall that Wigner matrix is an $N \times N$ real symmetric matrix W satisfying the assumptions (i)–(iii) in Definition 1.1, and $\|W_{ij}\|_k \asymp \|W_{ij}\|_2$ for all $k \geq 3$. W is the Gaussian orthogonal ensemble (GOE) if we further assume that W_{ij} have Gaussian distributions.

The celebrated Wigner–Dyson–Mehta (WDM) universality conjecture asserts that the local spectral properties of a random matrix do not depend on the explicit distribution of the matrix entries, and they are only determined by the symmetry class of the matrix. During the past decade, the universality conjecture for Wigner matrices has been established in a series of papers [8–11, 24, 25] in great generality. In particular, it has been shown that for a symmetric Wigner matrix, the averaged n-point correlation functions and distribution of a single eigenvalue gap coincide with those of the GOE.

The study of universality for sparse matrices was initiated in [6, 7], where the authors proved local semicircle law on optimal scales, and established bulk universality for $q \geq N^{1/3}$. Later in [17], the result was extended to all $q \geq N^{\varepsilon}$. In particular, it was proved that for the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$ of A and the eigenvalues $\mu_1 \leq \cdots \leq \mu_N$ of GOE, one has

$$\lim_{N\to\infty} \mathbb{E}\big[f\big(N\varrho(\gamma_i)(\lambda_i - \lambda_{i+1})\big) - f\big(N\varrho(\gamma_i)(\mu_i - \mu_{i+1})\big)\big] = 0 \tag{1.2}$$

whenever $i \in [\varepsilon N, (1-\varepsilon)N]$. Here $f \in C_c^{\infty}(\mathbb{R})$, ϱ and γ_i are the semicircle density and its ith N-quantile γ_i respectively, that is,

$$\varrho(x) := \frac{1}{2\pi}\sqrt{(4-x^2)_+} \text{and} N\int_{-2}^{\gamma_i} \varrho(x)\,\mathrm{d}x = i - 1/2.$$

Unlike the averaged n-point correlation functions and single eigenvalue gaps, the fluctuations of single eigenvalues are understood much later. The single eigenvalue fluctuation was first considered in [12] for Gaussian unitary ensembles (GUE), where the author proved that

$$\frac{\mu_i - \gamma_i}{\sqrt{\frac{2\log N}{(4-\gamma_i^2)N^2}}} \xrightarrow{d} \mathcal{N}(0,1) \tag{1.3}$$

as $N \to \infty$, for all bulk eigenvalues μ_i of GUE. In [22], the result was extended to GOE and a special class of Wigner matrices. Recently in [3, 19], it was showed that (1.3) remains valid for all Wigner matrices.

In this paper, we study the single eigenvalue fluctuation of the sparse matrices. For the remaining of this paper we replace the assumption (iv) in Definition 1.1 by

(iv)′. For any $k \geq 3$, we have $\mathbb{E}|H_{ij}|^k \asymp 1/(Nq^{k-2})$ for all i, j,

so that H and A are strictly sparse. Let us denote

$$\zeta := \min\left\{\frac{1}{2} - \beta, \beta\right\} > 0. \tag{1.4}$$

We may now state our main result.

THEOREM 1.2 (Main result). *Fix $\tau > 0$. Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$ be the eigenvalues of A. Set*

$$X_i := \frac{\lambda_i - \mathbb{E}\lambda_i}{\gamma_i\sqrt{\frac{1}{2}\mathbb{E}H_{12}^4}}$$

for all $i \in \{1, 2, \ldots, N\}$. We denote the index set $\mathcal{I} := ([\tau N, N/2 - N^{1-\zeta/17}] \cup [N/2 + N^{1-\zeta/17}, (1-\tau N)]) \cap \mathbb{N}$. Then for any fixed k and $i_1, \ldots, i_k \in \mathcal{I}$,

$$(X_{i_1}, \ldots, X_{i_k}) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \mathcal{J}), \tag{1.5}$$

where $\mathcal{J} \in \mathbb{R}^{k\times k}$ is the matrix of ones, that is, $\mathcal{J}_{ij} = 1$ for all $i, j \in \{1, 2, \ldots, k\}$.

For $\lim_{N\to\infty} i/N = 1/2$, *we have*

$$\frac{\lambda_i - \mathbb{E}\lambda_i}{\sqrt{\mathbb{E}H_{12}^4}} \xrightarrow{d} 0.$$

By assumption (iv)′, we have $\gamma_i\sqrt{\frac{1}{2}\mathbb{E}H_{12}^4} \asymp |\gamma_i| N^{-1/2-\beta}$. Thus (1.5) implies that, when γ_i is away from 0, the corresponding λ_i fluctuates on a much larger scale than the bulk eigenvalues of a Wigner matrix. Since the limiting covariance matrix $\mathcal{J}$ is the matrix of ones, we see that all the eigenvalues fluctuate simultaneously. Note the phenomenon of the co-existence of Theorem 1.2 and the gap universality (1.2): although the eigenvalues of A fluctuate on large scales, the fluctuations of consecutive eigenvalues are almost identical, and hence the fluctuations make little impact on the gap distribution.

We also remark on the fluctuation near the edge. For $q \gg N^{1/6}$, the extreme eigenvalues of A are known to exhibit Tracy–Widom fluctuations [6, 20]. When $N^{1/9} \ll q \ll N^{1/6}$, it was proved in [18] that

$$\frac{\lambda_{N-1} - (2 + 1/q^2 - 5/(4q^4))}{\sqrt{2\mathbb{E}H_{12}^4}} \xrightarrow{d} \mathcal{N}(0,1). \tag{1.6}$$

Note that

$$X_{N-1} = \frac{\lambda_{N-1} - \mathbb{E}\lambda_{N-1}}{\sqrt{2\mathbb{E}H_{12}^4}}\big(1 + O(N^{-2/3})\big),$$

thus for $N^{1/9} \ll q \ll N^{1/6}$, the bulk fluctuation (1.5) exhibits exactly the same behavior as the edge fluctuation (1.6). In fact, in both cases the fluctuations come from the sparsity of A. We believe that the source of the edge fluctuation remains the same for small q, and Theorem 1.2 can be extended to the edge for all $N^{\varepsilon} \le q \ll N^{1/6}$.

We also have the following central limit theorem for the eigenvalue counting function of A.

THEOREM 1.3. *Fix* $\tau > 0$. *Let* $\Sigma(E) := |\{i : \lambda_i \le E\}|$ *denote the eigenvalue counting function of* A. *For* $E \in [-2+\tau, -N^{-\zeta/17}] \cup [N^{-\zeta/17}, 2-\tau]$, *we have*

$$\frac{\Sigma(E) - \mathbb{E}\Sigma(E)}{\sigma(E)} \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\sigma(E) := E\sqrt{4-E^2}\left(\frac{\mathbb{E}H_{12}^4}{8\pi^2}\right)^{1/2} N.$$

For $\lim_{N\to\infty} E = 0$, *we have*

$$\frac{\sqrt{N}}{q}\big(\Sigma(E) - \mathbb{E}\Sigma(E)\big) \xrightarrow{d} 0.$$

Let F be a smooth test function independent of N. Recall that for a Wigner matrix W, the macroscopic linear statistic $\operatorname{Tr} F(W)$ fluctuates on the scale 1 (see [21]), while [19] shows $\Sigma_W(E)$ fluctuates on the scale $\sqrt{\log N} \gg 1$. This is due to the fact that as the derivative of the test function becomes more singular, the leading contribution of the fluctuation will start to come from the fluctuations of individual eigenvalues, which are much larger than the averaging fluctuation from linear statistics.

Our observation is that for a sparse matrix, $\Sigma(E)$ should fluctuate on the same scale as $\operatorname{Tr} F(H)$. From [23] we know that $\operatorname{Tr} F(H)$ fluctuates on the scale $\sqrt{N}/q \gg \sqrt{\log N}$. The source of this is the fourth moment assumption $\mathbb{E}(\sqrt{N}H_{12})^4 \asymp N/q^2 \gg 1$, which gives rise to large, but simultaneous fluctuations for all eigenvalues. When we switch from continuous to a jump test function, the result remains the same, as the source of the fluctuation is unchanged.

To study λ_i and $\Sigma(E)$, the main step is obtaining good estimates for linear statistics of Green functions at small scales. Let us define the spectral domain

$$\mathbf{D}_\tau := \{E + \mathrm{i}\eta : |4 - E^2| + \eta \geq \tau, |E| \leq 4, N^{-1+\tau} \leq \eta \leq 4\}.$$

We denote the resolvent of H by $G(z) := (H - z)^{-1}$, where $\operatorname{Im} z \neq 0$. The key step of our proof is a result on centered moments of mesoscopic linear statistics of the Green functions (see Proposition 3.1 below), which in particular implies the optimal estimate

$$\frac{1}{N}\operatorname{Tr} G(z) - \frac{1}{N}\mathbb{E}\operatorname{Tr} G(z) \prec \frac{1}{N\eta} + \frac{1}{\sqrt{N}q}$$

for all $z = E + \mathrm{i}\eta \in \mathbf{D}_\tau$. Here "$\prec$" is the notion of stochastic domination given in Definition 2.5 below.

By computing the high moments of $N^{-1}\operatorname{Tr} G - N^{-1}\mathbb{E}\operatorname{Tr} G$ using cumulant expansion/Schur complement formula, it can be proven, as previously in [7] that

$$\frac{1}{N}\operatorname{Tr} G(z) - \frac{1}{N}\mathbb{E}\operatorname{Tr} G(z) \prec \frac{1}{N\eta} + \frac{1}{q^2}.$$

In order to improve the second term $1/q^2$ to the optimal scale $1/(\sqrt{N}q)$, we need more expansions. However, each additional expansion, in the worst case, only results in an improvement of factor $1/q^2$. When $q = N^\varepsilon$, it is impossible to write down each expansion explicitly, and one has to introduce general formulas that allows recursive expansions. In order to do so, we implement the ideas in [14], to construct a hierarchy of Schwinger–Dyson equations for a sufficiently large class of polynomials in the entries of the Green function. As [14] deals with the covariance of two Green functions of Wigner matrices, we also need to adapt the method to our current setting, which deals with high-moment estimates of Green functions of sparse matrices. See Section 4.2 for more details.

We also apply Proposition 3.1 to prove the following CLT for mesoscopic linear statistics of Green functions.

THEOREM 1.4. *Let $z = E + \mathrm{i}\eta \in \mathbf{D}_\tau$.*

(i) *When $\eta \gg q/\sqrt{N}$,*

$$\frac{1}{m(z)m'(z)\sqrt{2\mathbb{E}H_{12}^4 N}}\big(\operatorname{Tr} G(z) - \mathbb{E}\operatorname{Tr} G(z)\big) \xrightarrow{d} \mathcal{N}(0,1), \tag{1.7}$$

where m is the Stieltjes transform of the Wigner semicircle law.

(ii) *When $\eta \ll q/\sqrt{N}$,*

$$\sqrt{2}\eta\big(\operatorname{Tr} G(z) - \mathbb{E}\operatorname{Tr} G(z)\big) \xrightarrow{d} \mathcal{N}_{\mathbb{C}}(0,1). \tag{1.8}$$

Here $\mathcal{N}_{\mathbb{C}}(0,1)$ denotes the distribution of the standard complex Gaussian random variable.

Note that (1.8) coincides with the mesoscopic linear statistics for GOE [4], whose source is the extrapolation of WDM (or sine-kernel) statistics to mesoscopic scales. On the other hand, (1.7) comes from the sparsity of H. Thus our result shows that, although the eigenvalue statistics for sparse matrices are different from WDM statistics on large scales, WDM

statistics remain valid on small enough mesoscopic scales. This bridges the results on microscopic [6, 17] and macroscopic [1, 23] statistics of H.

The rest of the paper is organized as follows. In Section 2 we introduce the notations and previous results that we use in this paper. In Section 3 we prove our main results, Theorems 1.2– 1.4, assuming a key result on centered moments of mesoscopic linear statistics, Proposition 3.1. In Section 4 we introduce a class of polynomials in the entries of the Green function, and construct a hierarchy of its Schwinger–Dyson equations. We then use this construction to prove Proposition 3.1. Finally in Section 5 we prove the general estimates for the class of polynomials of Green function that we used in Section 4.

Conventions. Throughout this paper, we regard N as our fundamental large parameter. Any quantities that are not explicitly constant or fixed may depend on N; we almost always omit the argument N from our notation. We use τ to denote some generic (small) positive constant, whose value may change from one expression to the next. Similarly, we use C to denote some generic (large) positive constant. For $A, B > 0$, we use $A = O(B)$ to denote $A \le CB$ and $A \asymp B$ to denote $C^{-1}B \le A \le CB$. When we write $A \ll B$ and $A \gg B$, we mean $A \le CN^{-\tau}B$ and $A \ge C^{-1}N^{\tau}B$ for some constants $C, \tau > 0$ respectively.

2. Preliminaries. In this section we collect notations and tools that are used in the paper.

Let M be an $N \times N$ matrix. We denote $M^{*n} := (M^*)^n$, $M^*_{ij} := (M^*)_{ij} = \overline{M}_{ji}$, $M^n_{ij} := (M_{ij})^n$, and the normalized trace of M by $\underline{M} := \frac{1}{N}\operatorname{Tr} M$. We abbreviate $\langle X\rangle := X - \mathbb{E}X$ for any random variable X with finite expectation. For the Green function G, we have the differential rule

$$\frac{\partial G_{ij}}{\partial H_{kl}} = -(G_{ik}G_{lj} + G_{il}G_{kj})(1+\delta_{kl})^{-1}. \tag{2.1}$$

Let μ be the empirical spectral measure of H. Its Stieltjes transform is denoted by

$$\underline{G}(z) := \frac{1}{N}\operatorname{Tr} G(z) = \int \frac{\mu(x)}{x-z}\,\mathrm{d}x.$$

We also have

$$\frac{\partial(\underline{H^2}-1)}{\partial H_{ij}} = \frac{4}{N}H_{ij}(1+\delta_{ij})^{-1} \prec \frac{1}{Nq} \quad \text{and} \quad \frac{\partial^2(\underline{H^2}-1)}{\partial H_{ij}^2} = \frac{4}{N}(1+\delta_{ij})^{-1}. \tag{2.2}$$

For $z \in \mathbb{C}$ with $\operatorname{Im} z \ne 0$, the Stieltjes transform of the Wigner semicircle law is defined by

$$m(z) := \int \frac{\varrho(x)}{x-z}\,\mathrm{d}x.$$

One elementary fact is that m is the unique solution of

$$1 + zm(z) + m(z)^2 = 0 \tag{2.3}$$

satisfying $\operatorname{Im} m(z) \operatorname{Im} z > 0$. Let us define the spectral domains

$$\mathbf{S} = \{E + \mathrm{i}\eta : |E| \le 4, 0 < \eta \le 4\} \quad \text{and} \quad \widetilde{\mathbf{S}}_\tau = \{E + \mathrm{i}\eta : |E| \le 4, N^{-1+\tau} \le \eta \le 4\}.$$

We denote the distance to spectral edge by

$$\kappa \equiv \kappa_E := \min\{|2-E|, |2+E|\}.$$

LEMMA 2.1 (Basic properties of m). *We have*

$$|m(z)| \asymp 1 \quad \textit{and} \quad |m'(z)| \asymp \frac{1}{\sqrt{\kappa+\eta}}$$

for all $z \in \mathbf{S}$. *In particular,* $|m'(z)| \asymp 1$ *for all* $z \in \mathbf{D}_\tau$.

PROOF. The proof is an elementary exercise using (2.3). □

If h is a real-valued random variable with finite moments of all order, we denote by $\mathcal{C}_k(h)$ the kth cumulant of h, that is,

$$\mathcal{C}_k(h) := (-\mathrm{i})^k \cdot \big(\partial_\lambda^k \log \mathbb{E}\mathrm{e}^{\mathrm{i}\lambda h}\big)|_{\lambda=0}.$$

We state the cumulant expansion formula, whose proof is given in, for example, [16], Appendix A.

LEMMA 2.2 (Cumulant expansion). *Let $f : \mathbb{R} \to \mathbb{C}$ be a smooth function, and denote by $f^{(k)}$ its kth derivative. Then, for every fixed $\ell \in \mathbb{N}$, we have*

$$\mathbb{E}\big[h \cdot f(h)\big] = \sum_{k=0}^{\ell} \frac{1}{k!} \mathcal{C}_{k+1}(h) \mathbb{E}\big[f^{(k)}(h)\big] + \mathcal{R}_{\ell+1}, \tag{2.4}$$

assuming that all expectations in (2.4) exist, where $\mathcal{R}_{\ell+1}$ is a remainder term (depending on f and h), such that for any $t > 0$,

$$\mathcal{R}_{\ell+1} = O(1) \cdot \Big(\mathbb{E} \sup_{|x|\le|h|} \big|f^{(\ell+1)}(x)\big|^2 \cdot \mathbb{E}\big|h^{2\ell+4}\mathbf{1}_{|h|>t}\big|\Big)^{1/2} + O(1) \cdot \mathbb{E}|h|^{\ell+2} \cdot \sup_{|x|\le t} \big|f^{(\ell+1)}(x)\big|.$$

The following result gives bounds on the cumulants of the entries of H, whose proof follows by the homogeneity of the cumulants.

LEMMA 2.3. *For every $k \in \mathbb{N}$ we have*

$$\mathcal{C}_k(H_{ij}) = O_k\big(1/(Nq^{k-2})\big)$$

uniformly for all i, j.

The following is a standard complex analysis result from [5].

LEMMA 2.4 (Helffer–Sjöstrand formula). *Let $f \in C^2(\mathbb{R})$, and let $\tilde{f}$ be the almost analytic extension of f defined by*

$$\tilde{f}(x + \mathrm{i}y) := f(x) + \mathrm{i}y f'(x).$$

Let $\chi \in \mathcal{C}_c^\infty(\mathbb{R})$ be a cutoff function satisfying $\chi(0) = 1$, and by a slight abuse of notation write $\chi(z) \equiv \chi(\operatorname{Im} z)$. Then for any $\lambda \in \mathbb{R}$ we have

$$f(\lambda) = \frac{1}{\pi} \int_{\mathbb{C}} \frac{\partial_{\bar{z}}(\tilde{f}(z)\chi(z))}{\lambda - z} \,\mathrm{d}^2 z,$$

where $\partial_{\bar{z}} := \frac{1}{2}(\partial_x + \mathrm{i}\partial_y)$ is the antiholomorphic derivative and $\mathrm{d}^2 z$ the Lebesgue measure on $\mathbb{C}$.

The following definition introduces a (conventional) notion of a high-probability bound that is used commonly in random matrix theory.

DEFINITION 2.5 (Stochastic domination). Let

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

be two families of random variables, where $Y^{(N)}(u)$ are nonnegative and $U^{(N)}$ 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

$$\sup_{u \in U^{(N)}} \mathbb{P}\big[|X^{(N)}(u)| > N^{\varepsilon} Y^{(N)}(u)\big] \le N^{-D}$$

for large enough $N \ge N_0(\varepsilon, D)$. If X is stochastically dominated by Y, uniformly in u, we use the notation $X \prec Y$, or equivalently $X = O_{\prec}(Y)$. (Note that for deterministic X and Y, $X = O_{\prec}(Y)$ means $X = O_{\varepsilon}(N^{\varepsilon} Y)$ for any $\varepsilon > 0$.)

Next we recall the local semicircle law for Erdős–Rényi graphs in [7].

THEOREM 2.6 (Theorem 2.8, [7]). *Let H be a sparse matrix defined as in Definition 1.1. We have*

$$\max_{i,j} |G_{ij}(z) - \delta_{ij} m(z)| \prec \frac{1}{q} + \sqrt{\frac{\operatorname{Im} m(z)}{N\eta}} + \frac{1}{N\eta}$$

and

$$|\underline{G} - m| \prec \frac{1}{q} \wedge \frac{1}{q^2(\eta + \kappa_E)} + \frac{1}{N\eta}$$

uniformly in $z = E + \mathrm{i}\eta \in \mathbf{S}$.

REMARK 2.7. Theorem 2.6 was proved in [7] under the additional assumption $\mathbb{E}H_{ii}^2 = 1/N$ for all i. However, the proof is insensitive to the variance of the diagonal entries, and one can easily repeat the steps in [7] under the general assumption $\mathbb{E}H_{ij}^2 = C_i/N$. A weak local law for H with general variances on the diagonal can also be found in [15].

We also need the following result from [7] concerning the density of states of A.

LEMMA 2.8 (Theorem 2.10, [7]). *Let $\widetilde{\mu}$ be the empirical eigenvalue density of A. For any interval $I \subset \mathbb{R}$, we have*

$$|\widetilde{\mu}(I) - \varrho(I)| \prec \frac{1}{N} + \frac{|I|}{q}.$$

We recall the magical Ward identity.

LEMMA 2.9 (Ward identity). *We have*

$$\sum_j |G_{ij}|^2 = \frac{\operatorname{Im} G_{ii}}{\eta}$$

for all $z = E + \mathrm{i}\eta \in \mathbf{S}$.

Finally, we collect some estimates in the following lemma, whose proof is postponed to Appendix 5.2.

LEMMA 2.10. (*i*) *For any fixed $m, n \in \mathbb{N}$ such that $m + n \ge 1$, we have*

$$\langle \underline{G^m G^{*n}} \rangle \prec \eta^{1-(m+n)} \left(\frac{1}{q} + \frac{1}{N\eta} \right) \tag{2.5}$$

as well as

$$(G^m G^{*n})_{ij} \prec \begin{cases} \eta^{1-(m+n)} & \text{if } i = j, \\ \eta^{1-(m+n)}\Big(\dfrac{1}{q} + \dfrac{1}{\sqrt{N\eta}}\Big) & \text{if } i \neq j, \end{cases} \tag{2.6}$$

uniformly in i, j *and* $z = E + \mathrm{i}\eta \in \widetilde{\mathbf{S}}_\tau$.

(*ii*) *For any fixed* $m, n \in \mathbb{N}$ *such that* $m + n \geq 1$, *we have*

$$\sum_i |(G^m G^{*n})_{ij}|^2 \prec \eta^{1-2(m+n)}$$

uniformly in j *and* $z = E + \mathrm{i}\eta \in \widetilde{\mathbf{S}}_\tau$.

(*iii*) *For* $k = 2, 3$, *we have*

$$\mathbb{E}\underline{G}^k \prec \Big(\frac{1}{q} + \frac{1}{N\eta} + \eta\Big)\eta^{1-k}$$

uniformly for all $z = E + \mathrm{i}\eta \in \mathbf{D}_\tau$.

(*iv*) *For any fixed* $n \in \mathbb{N}_+$,

$$\mathbb{E}(\underline{H^2} - 1)^n = \begin{cases} (n-1)!!(2\mathbb{E}H_{12}^4)^{n/2} + O\Big(\dfrac{1}{(\sqrt{N}q)^n} \cdot \dfrac{q}{\sqrt{N}}\Big) & \text{if } n \text{ is even}, \\ O\Big(\dfrac{1}{(\sqrt{N}q)^n} \cdot \dfrac{q}{\sqrt{N}}\Big) & \text{if } n \text{ is odd}. \end{cases} \tag{2.7}$$

3. Proof of main results. For $z = E + \mathrm{i}\eta \in \mathbf{D}_\tau$, we write

$$\alpha := -\log_N \eta$$

so that $\eta = N^{-\alpha}$, $\alpha \in [0, 1 - \tau]$. We define

$$\delta \equiv \delta(z) := \min\Big\{\beta, \frac{1}{2} - \beta, \frac{1-\alpha}{2}\Big\} > 0, \tag{3.1}$$

and

$$\xi \equiv \xi(z) = \frac{1}{8}\min\Big\{\frac{\alpha}{2}, \delta\Big\} \geq 0. \tag{3.2}$$

We define the linear statistics with a random shift

$$[G] \equiv [G(z)] := \frac{1}{N}\operatorname{Tr} G(z) - \frac{1}{N}\mathbb{E}\operatorname{Tr} G(z) - (\underline{H^2} - 1)m(z)m'(z).$$

Note that (2.7) implies

$$\frac{1}{\sqrt{2\mathbb{E}H_{12}^4}}(\underline{H^2} - 1) \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{and} \quad \underline{H^2} - 1 \prec N^{-1/2-\beta}. \tag{3.3}$$

The term $\underline{H^2} - 1$ was introduced in [18] to study the eigenvalue fluctuations of A near the edge.

In this section we shall prove Theorems 1.2–1.4 assuming the following proposition, whose proof is postponed to Section 4.

PROPOSITION 3.1. *Let* $m, n \in \mathbb{N}_+$. *We have*

$$\mathbb{E}|[G]|^{2n} = \frac{n!}{2^n}\Big(\frac{1}{N\eta}\Big)^{2n} + O_\prec\Big(\frac{N^{-\xi}}{(N\eta)^{2n}} + \Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^{2n}\Big) \tag{3.4}$$

and

$$\mathbb{E}[G^*]^n[G]^m = O_\prec\Big(\frac{N^{-\xi}}{(N\eta)^{m+n}} + \Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^{m+n}\Big) \tag{3.5}$$

for $m \neq n$, uniformly for all $z \in \mathbf{D}_\tau$.

We observe that Theorem 1.4 is an immediate consequence of Lemma 2.1, Lemma 2.10(iv) and Proposition 3.1. One can follow, for example, the steps in [13], to show Theorem 1.4 for general test functions. We do not pursue it here.

3.1. *Proof of Theorem 1.3.* In this section, we prove the following result, which trivially implies Theorem 1.3 by (3.3).

PROPOSITION 3.2. *Let $E \in [-2+\tau, 2-\tau]$. We have*

$$\Sigma(E) - \mathbb{E}\Sigma(E) - \frac{E\sqrt{4-E^2}}{4\pi}(\underline{H^2-1})N \prec N^{1/2-\beta-\zeta/16}.$$

PROOF. Let $f \in C^\infty(\mathbb{R})$ such that $f = 1$ in $(-3+\frac{1}{N}, E-\frac{1}{N}]$ and $f = 0$ in $(-\infty, -3-\frac{1}{N}] \cup [E+\frac{1}{N}, +\infty)$. We further assume $|f'| = O(N)$ and $|f''| = O(N^2)$. Let us write $z = x + \mathrm{i}y$ and choose $\chi \equiv \chi(y)$ such that $\chi(y) = 1$ for $|y| \le 1$ and $\chi(y) = 0$ for $|y| \ge 2$. Note that by Green's theorem we have

$$\int_{\mathbb{C}} \partial_{\bar z}\big(\tilde f(z)\chi(z)\big)m(z)m'(z)\,\mathrm{d}^2 z = \frac{1}{2}\int_{-2}^{2} f(x)\frac{2-x^2}{\sqrt{4-x^2}} = \frac{E\sqrt{4-E^2}}{4} + O(N^{-1}),$$

and Lemma 2.4 implies

$$\operatorname{Tr} f(H) - \mathbb{E}\operatorname{Tr} f(H) = \frac{N}{\pi}\int_{\mathbb{C}} \partial_{\bar z}\big(\tilde f(z)\chi(z)\big)\langle \underline{G}\rangle\,\mathrm{d}^2 z.$$

Combining the above two relations, and together with (3.3), we have

$$\begin{aligned}&\operatorname{Tr} f(H) - \mathbb{E}\operatorname{Tr} f(H) - \frac{E\sqrt{4-E^2}}{4\pi}(\underline{H^2-1})N \\ &\quad= \frac{N}{\pi}\int_{\mathbb{C}} \partial_{\bar z}\big(\tilde f(z)\chi(z)\big)\big[G(z)\big]\,\mathrm{d}^2 z + O_\prec(N^{-1/2-\beta}).\end{aligned} \tag{3.6}$$

Recall the definition of $\zeta > 0$ from (1.4). By (3.4) we have

$$\big[G(z)\big] \prec \frac{1}{Ny} + \frac{N^{-\zeta/16}}{\sqrt{N}q}$$

uniformly for $z = x + \mathrm{i}y \in \mathbf{D}_{\zeta/2}$, and an N^{-3}-net argument [2], Remark 2.7, shows

$$\sup_{z\in\mathbf{D}_{\zeta/2}} \big|\big[G(z)\big]\big|\Big(\frac{1}{Ny} + \frac{N^{-\zeta/16}}{\sqrt{N}q}\Big)^{-1} \prec 1. \tag{3.7}$$

Fix $\varepsilon > 0$. By Theorem 2.6, (3.3) and an N^{-3}-net argument, we see that

$$\sup_{z\in\mathbf{S}, N^{-1+\varepsilon}\le y\le N^{-1+\zeta/2}} \big|Ny\big[G(z)\big]\big| \prec \sum_{N^{-1+\varepsilon}\le y\le N^{-1+\zeta/2}} Ny\Big(\frac{1}{q} + \frac{1}{Ny} + \frac{1}{\sqrt{N}q}\Big) \prec 1. \tag{3.8}$$

From Theorem 2.6, $|m(z)| \le C$, and an N^{-3}-net argument, we have

$$\sup_{|x|\le 4}\max_{i,j}\big|G_{ij}(x + \mathrm{i}N^{-1+\varepsilon})\big| \prec 1,$$

and a deterministic monotonicity result [2], Lemma 10.2, shows

$$\sup_{z=x+\mathrm{i}y\in\mathbf{S},\,y\le N^{-1+\varepsilon}}\max_{i,j}\bigl|NyG_{ij}(z)\bigr|\prec N^{\varepsilon}.$$

Using the above relation, together with (3.3), (3.8) and the arbitrarity of ε, we have

$$\sup_{z\in\mathbf{S},\,y\le N^{-1+\zeta/2}}\bigl|Ny\bigl[G(z)\bigr]\bigr|\prec 1. \tag{3.9}$$

We split

$$\begin{aligned}&\biggl|\frac{N}{\pi}\int_{\mathbb{C}}\partial_{\bar z}\bigl(\tilde f(z)\chi(z)\bigr)\bigl[G(z)\bigr]\,\mathrm{d}^2z\biggr|\\&\quad\le C\biggl|N\int_{1\le y\le 2}\bigl(f(x)+\mathrm{i}yf'(x)\bigr)\chi'(y)\bigl[G(x+\mathrm{i}y)\bigr]\,\mathrm{d}^2z\biggr|\\&\qquad+C\biggl|N\int_{0<y\le N^{-1+\zeta/2}}f''(x)y\chi(y)\bigl[G(z)\bigr]\,\mathrm{d}^2z\biggr|\\&\qquad+C\biggl|N\int_{N^{-1+\zeta/2}\le y\le 2}f''(x)y\chi(y)\bigl[G(z)\bigr]\,\mathrm{d}^2z\biggr|.\end{aligned} \tag{3.10}$$

By (3.7) and $\|f\|_1+\|f'\|_1\le C$ we have

$$\begin{aligned}&\biggl|N\int_{1\le y\le 2}\bigl(f(x)+\mathrm{i}yf'(x)\bigr)\chi'(y)\bigl[G(x+\mathrm{i}y)\bigr]\,\mathrm{d}^2z\biggr|\\&\quad\prec N\biggl(\frac{1}{N}+\frac{N^{-\zeta/16}}{\sqrt{Nq}}\biggr)\prec N^{1/2-\beta-\zeta/16}.\end{aligned} \tag{3.11}$$

By (3.9) and $\|f''\|_1=O(N)$ we have

$$\begin{aligned}&\biggl|N\int_{0<y\le N^{-1+\zeta/2}}f''(x)y\chi(y)\bigl[G(z)\bigr]\,\mathrm{d}^2z\biggr|\\&\quad\prec N\cdot\int_0^{N^{-1+\zeta/2}}1\,\mathrm{d}y=N^{\zeta/2}\le N^{1/2-\beta-\zeta/2}.\end{aligned} \tag{3.12}$$

For the last term on RHS of (3.10), we do integration by parts, first in x and then in y, and get

$$\begin{aligned}&\biggl|N\int_{N^{-1+\zeta/2}\le y\le 2}f''(x)y\chi(y)\bigl[G(z)\bigr]\,\mathrm{d}^2z\biggr|\\&\quad\le\biggl|N\int_{N^{-1+\zeta/2}\le y\le 2}f'(x)\chi(y)\bigl[G(z)\bigr]\,\mathrm{d}^2z\biggr|\\&\qquad+\biggl|N\int_{1\le y\le 2}f'(x)y\chi'(y)\bigl[G(z)\bigr]\,\mathrm{d}^2z\biggr|\\&\qquad+\biggl|N\int f'(x)N^{-1+\zeta/2}\chi\bigl(N^{-1+\zeta/2}\bigr)\bigl[G\bigl(x+\mathrm{i}N^{-1+\zeta/2}\bigr)\bigr]\,\mathrm{d}x\biggr|,\end{aligned}$$

and again by (3.7) we have

$$\biggl|N\int_{N^{-1+\zeta/2}\le y\le 2}f''(x)y\chi(y)\bigl[G(z)\bigr]\,\mathrm{d}^2z\biggr|\prec N^{1/2-\beta-\zeta/16}. \tag{3.13}$$

From (3.6), (3.10)–(3.13) we have

$$\operatorname{Tr}f(H)-\mathbb{E}\operatorname{Tr}f(H)-\frac{E\sqrt{4-E^2}}{4\pi}\underline{(H^2-1)}N=O_{\prec}\bigl(N^{1/2-\beta-\zeta/16}\bigr). \tag{3.14}$$

Let $\widetilde{\Sigma}$ be the eigenvalue counting function of H. Note that $\|H\| \le 5/2$ with overwhelming probability. Thus

$$\begin{aligned}|\operatorname{Tr} f(H) - \widetilde{\Sigma}(E)| &\le C\big(\widetilde{\Sigma}(E+N^{-1}) - \widetilde{\Sigma}(E-N^{-1}) + \widetilde{\Sigma}(-3+N^{-1})\big)\\ &= C\big(\widetilde{\Sigma}(E+N^{-1}) - \widetilde{\Sigma}(E-N^{-1})\big) + O_{\prec}(1).\end{aligned}$$

From Theorem 2.6 we know

$$\widetilde{\Sigma}(E+N^{-1}) - \widetilde{\Sigma}(E-N^{-1}) \le \sum_i \frac{2}{N^2(\lambda_i - E)^2 + 1} = 2\operatorname{Im} \underline{G}(E + \mathrm{i}N^{-1}) \prec 1,$$

and thus

$$\operatorname{Tr} f(H) - \widetilde{\Sigma}(E) \prec 1. \tag{3.15}$$

Note that (3.15) also implies

$$\mathbb{E}\operatorname{Tr} f(H) - \mathbb{E}\widetilde{\Sigma}(E) \prec 1. \tag{3.16}$$

By Cauchy interlacing theorem (e.g., [6], Lemma 6.1) we have

$$\widetilde{\Sigma}(E) - 1 \le \Sigma(E) \le \widetilde{\Sigma}(E). \tag{3.17}$$

By (3.15)–(3.17) we have

$$\operatorname{Tr} f(H) - \mathbb{E}\operatorname{Tr} f(H) - \big(\Sigma(E) - \mathbb{E}\Sigma(E)\big) \prec 1. \tag{3.18}$$

Combining (3.14) and (3.18) completes the proof. □

3.2. *Proof of Theorem* 1.2. We shall prove the following result, and Theorem 1.2 then follows by Lemma 2.10(iv).

PROPOSITION 3.3. *Fix $\tau > 0$. We denote the eigenvalues of A by $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$. For all $i \in [\tau N, (1-\tau)N]$, we have*

$$\lambda_i - \mathbb{E}\lambda_i - \frac{\gamma_i}{2}\big(\underline{H^2} - 1\big) \prec N^{-1/2-\beta-\zeta/16}.$$

PROOF. Let $\widetilde{\mu}$ be the empirical eigenvalue density of A. Let us define the function $g : \mathbb{R} \to \mathbb{R}$ by

$$g(a) := \mathbb{E}\int_{-\infty}^{a} \widetilde{\mu}(x)\,\mathrm{d}x.$$

We claim that for any fixed (small) $\varepsilon > 0$, g has no jumps of size larger than $N^{-1+\varepsilon}$. In fact, by Lemma 2.8 we have

$$\begin{aligned}g(a + N^{-1+\varepsilon/2}) - g(a) &= \mathbb{E}\widetilde{\mu}((a, a+N^{-1+\varepsilon/2}])\\ &= \varrho((a, a+N^{-1+\varepsilon/2}]) + O_{\prec}(N^{-1}) \le CN^{-1+\varepsilon/2}.\end{aligned}$$

Pick $i \in [\tau N, (1-\tau)N]$. We can then choose deterministic $\theta_i \in \mathbb{R}$ satisfying

$$\left|\mathbb{E}\int_{-\infty}^{\theta_i} \widetilde{\mu}(x)\,\mathrm{d}x - \frac{i}{N}\right| \le N^{-1/2-\beta-\zeta/2},$$

so that $|\mathbb{E}\Sigma(\theta_i) - i| \le N^{1/2-\beta-\zeta/2}$. Fix $\varepsilon \in (0, \zeta/16)$. Let us abbreviate $\omega_i = \theta_i - N^{-1/2-\beta-\zeta/16+\varepsilon}$. We have

$$\begin{aligned}&\mathbb{P}\big(\lambda_i - \theta_i - (\underline{H^2} - 1)\gamma_i/2 \le -N^{-1/2-\beta-\zeta/16+\varepsilon}\big)\\ &= \mathbb{P}\big(\Sigma\big(\omega_i + (\underline{H^2} - 1)\gamma_i/2\big) \ge i\big)\\ &= \mathbb{P}\big(\Sigma\big(\omega_i + (\underline{H^2} - 1)\gamma_i/2\big) - \mathbb{E}\Sigma(\omega_i) \ge \mathbb{E}\Sigma(\theta_i) - \mathbb{E}\Sigma(\omega_i) + O\big(N^{1/2-\beta-\zeta/2}\big)\big).\end{aligned} \tag{3.19}$$

By Lemma 2.8 we know that

$$\widetilde{\mu}(I) - \varrho(I) \prec N^{-1/2-\beta-\zeta} \tag{3.20}$$

for any I satisfying $|I| \prec N^{-1/2-\beta}$. Together with (3.3) we have

$$\begin{aligned}
&\Sigma\big(\omega_i + (\underline{H^2} - 1)\gamma_i/2\big) - \Sigma(\omega_i) \\
&\quad = N\varrho\big([\omega_i, \omega_i + (\underline{H^2} - 1)\gamma_i/2]\big) + O_{\prec}(N^{1/2-\beta-\zeta}) \\
&\quad = \frac{\sqrt{4-\omega_i^2}}{4\pi} N(\underline{H^2} - 1)\gamma_i + O_{\prec}(N^{1/2-\beta-\zeta}) \\
&\quad = \frac{\sqrt{4-\omega_i^2}}{4\pi} N(\underline{H^2} - 1)\omega_i + O_{\prec}(N^{1/2-\beta-\zeta}).
\end{aligned} \tag{3.21}$$

In the last step of (3.21) we used $|\theta_i - \gamma_i| \prec N^{-\zeta}$, which also can be deduced from Lemma 2.8. By (3.20),

$$\begin{aligned}
\mathbb{E}\Sigma(\theta_i) - \mathbb{E}\Sigma(\omega_i) &= N\varrho\big([\omega_i, \theta_i]\big) + O_{\prec}(N^{1/2-\beta-\zeta}) \\
&= \frac{\sqrt{4-\omega_i^2}}{2\pi} N^{1/2-\beta-\zeta/16+\varepsilon} + O_{\prec}(N^{1/2-\beta-\zeta}).
\end{aligned} \tag{3.22}$$

A combination of (3.19), (3.21), and (3.22) shows that

$$\begin{aligned}
&\mathbb{P}\big(\lambda_i - \theta_i - (\underline{H^2} - 1)\gamma_i/2 \le -N^{-1/2-\beta-\zeta/16+\varepsilon}\big) \\
&\quad = \mathbb{P}\bigg(\Sigma(\omega_i) - \mathbb{E}\Sigma(\omega_i) - \frac{\sqrt{4-\omega_i^2}}{4\pi} N(\underline{H^2} - 1)\omega_i \\
&\qquad \ge \frac{\sqrt{4-\omega_i^2}}{2\pi} N^{1/2-\beta-\zeta/16+\varepsilon} + O_{\prec}(N^{1/2-\beta-\zeta/2})\bigg).
\end{aligned}$$

Since ε is arbitrary, by Proposition 2.10 we see that

$$(\lambda_i - \theta_i - (\underline{H^2} - 1)\gamma_i/2)_- \prec N^{-1/2-\beta-\zeta/16}.$$

Repeating the above process for

$$\mathbb{P}\big(\lambda_i - \theta_i - (\underline{H^2} - 1)\gamma_i/2 \ge N^{-1/2-\beta-\zeta/16+\varepsilon}\big)$$

we can also show that

$$(\lambda_i - \theta_i - (\underline{H^2} - 1)\gamma_i/2)_+ \prec N^{-1/2-\beta-\zeta/16}.$$

Thus

$$\lambda_i - \theta_i - (\underline{H^2} - 1)\gamma_i/2 \prec N^{-1/2-\beta-\zeta/16}, \tag{3.23}$$

which also implies

$$\mathbb{E}\lambda_i - \theta_i \prec N^{-1/2-\beta-\zeta/16}. \tag{3.24}$$

The proof then follows from (3.23) and (3.24). □

4. Proof of Proposition 3.1. In this section we prove (3.4); the proof of (3.5) is similar, and we omit the details. Throughout this section let us pick $n \in \mathbb{N}_+$ and

$$z = E + \mathrm{i}\eta \in \mathbf{D}_\tau. \tag{4.1}$$

Let us define

$$\mathcal{M} := \|[G]\|_{2n} = \big(\mathbb{E}|[G]|^{2n}\big)^{\frac{1}{2n}},$$

and we split

$$\mathcal{M}^{2n} = \mathbb{E}[G^*]^n[G]^{n-1}\langle \underline{G}\rangle - \mathbb{E}[G^*]^n[G]^{n-1}(\underline{H^2} - 1)mm'.$$

The proof of (3.4) is immediate from the next lemma.

LEMMA 4.1. *We have*

$$\begin{aligned}&\mathbb{E}[G^*]^n[G]^{n-1}\langle \underline{G}\rangle \\ &\quad = \frac{n}{2N^2\eta^2}\mathbb{E}|[G]|^{2n-2} + O_\prec(N^{-\delta})\mathcal{M}^{2n} \\ &\qquad + \sum_{r=1}^{2n} O_\prec\Big(\frac{N^{-\xi}}{(N\eta)^r} + \Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^r\Big)\mathcal{M}^{2n-r},\end{aligned} \tag{4.2}$$

and

$$\mathbb{E}[G^*]^n[G]^{n-1}(\underline{H^2} - 1)mm' = \sum_{r=1}^{2n} O_\prec\Big(\frac{N^{-\xi}}{(N\eta)^r} + \Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^r\Big)\mathcal{M}^{2n-r}. \tag{4.3}$$

In fact, Lemma 4.1 shows

$$\begin{aligned}\mathcal{M}^{2n} &= \frac{n}{2N^2\eta^2}\mathbb{E}|[G]|^{2n-2} + O_\prec(N^{-\delta})\mathcal{M}^{2n} \\ &\quad + \sum_{r=1}^{2n} O_\prec\Big(\frac{N^{-\xi}}{(N\eta)^r} + \Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^r\Big)\mathcal{M}^{2n-r},\end{aligned} \tag{4.4}$$

and together with $\mathbb{E}|[G]|^{2n-2} \le \mathcal{M}^{2n-2}$ we have

$$\mathcal{M}^{2n} \prec \sum_{r=1}^{2n}\Big(\frac{1}{(N\eta)^r} + \Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^r\Big)\mathcal{M}^{2n-r},$$

which implies

$$\mathcal{M}^{2n} \prec \frac{1}{(N\eta)^{2n}} + \Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^{2n}.$$

Since n is arbitrary, we have

$$[G] \prec \frac{1}{N\eta} + \frac{N^{-\delta/4}}{\sqrt{N}q}. \tag{4.5}$$

Inserting (4.5) back into (4.4), we have

$$\mathbb{E}|[G]|^{2n} = \mathcal{M}^{2n} = \frac{n}{2N^2\eta^2}\mathbb{E}|[G]|^{2n-2} + O_\prec\Big(\frac{N^{-\xi}}{(N\eta)^{2n}} + \Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^{2n}\Big),$$

and (3.4) follows by iteration.

In Sections 4.1–4.5 we shall prove (4.2), and in Section 4.6 we prove (4.3).

4.1. *First estimates.* By the resolvent identity $zG = GH - I$ we have

$$z\mathbb{E}\big[G^*\big]^n[G]^{n-1}\langle \underline{G}\rangle = \mathbb{E}\big\langle\big[G^*\big]^n[G]^{n-1}\big\rangle \underline{GH} = \frac{1}{N}\sum_{i,j}\mathbb{E}\big\langle\big[G^*\big]^n[G]^{n-1}\big\rangle G_{ij}H_{ji}. \tag{4.6}$$

We calculate the RHS of (4.6) using the cumulant formula (2.4) with $f = f_{ij}(H) := \langle[G^*]^n[G]^{n-1}\rangle G_{ij}$ and $h = H_{ji}$, and get

$$\begin{aligned}
&z\mathbb{E}\big[G^*\big]^n[G]^{n-1}\langle \underline{G}\rangle \\
&\quad= \frac{1}{N^2}\sum_{i,j}\mathbb{E}\big\langle\big[G^*\big]^n[G]^{n-1}\big\rangle\frac{\partial G_{ij}}{\partial H_{ji}}(1+\delta_{ji}) \\
&\quad\quad+ \frac{1}{N^2}\sum_{i,j}\mathbb{E}\frac{\partial(\langle[G^*]^n[G]^{n-1}\rangle)}{\partial H_{ji}}G_{ij}(1+\delta_{ji}) + \mathbb{E}K \\
&\quad\quad+ \sum_{k=2}^{l}\mathbb{E}L_k + \frac{1}{N}\sum_{i,j}\mathbb{E}\mathcal{R}_{l+1}^{(ji)} \\
&\quad=: (a)+(b)+K+\sum_{k=2}^{l}L_k + \frac{1}{N}\sum_{i,j}\mathbb{E}\mathcal{R}_{l+1}^{(ji)},
\end{aligned} \tag{4.7}$$

where

$$K = N^{-2}\sum_i \frac{\partial(\langle[G^*]^n[G]^{n-1}\rangle G_{ii})}{\partial H_{ii}}\big(N\mathcal{C}_2(H_{ii}) - 2\big), \tag{4.8}$$

and

$$L_k = N^{-1}\cdot\sum_{i,j}\bigg(\frac{1}{k!}\mathcal{C}_{k+1}(H_{ji})\frac{\partial^k(\langle[G^*]^n[G]^{n-1}G_{ij}\rangle)}{\partial H_{ji}^k}\bigg). \tag{4.9}$$

Here l is a fixed positive integer to be chosen later, and $\mathcal{R}_{l+1}^{(ji)}$ is a remainder term defined analogously to $\mathcal{R}_{l+1}$ in (2.4). Using the differential rule (2.1) we get

$$\begin{aligned}
(a) &= N^{-2}\sum_{i,j}\mathbb{E}\big\langle\big[G^*\big]^n[G]^{n-1}\big\rangle(-G_{ij}G_{ij} - G_{ii}G_{jj}) \\
&= -N^{-1}\mathbb{E}\big[G^*\big]^n[G]^{n-1}\langle\underline{G^2}\rangle - \mathbb{E}\big[G^*\big]^n[G]^{n-1}\langle\underline{G}\rangle^2 \\
&\quad- 2\mathbb{E}\big[G^*\big]^n[G]^{n-1}\langle\underline{G}\rangle\mathbb{E}\underline{G} + \mathbb{E}\big[G^*\big]^n[G]^{n-1}\mathbb{E}\langle\underline{G}\rangle^2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
(b) = -\frac{2}{N^2}\big(&n\mathbb{E}\big[G^*\big]^{n-1}[G]^{n-1}\big(\underline{GG^{*2}} + 2\underline{HG^*}\bar{m}\bar{m}'\big) \\
&+ (n-1)\mathbb{E}\big[G^*\big]^n[G]^{n-2}\big(\underline{G^3} + 2\underline{HG}mm'\big)\big).
\end{aligned}$$

Altogether we obtain

$$\begin{aligned}
&\mathbb{E}\big[G^*\big]^n[G]^{n-1}\langle\underline{G}\rangle \\
&\quad= \frac{1}{T}\bigg(\mathbb{E}\big[G^*\big]^n[G]^{n-1}\langle\underline{G}\rangle^2 - \mathbb{E}\big[G^*\big]^n[G]^{n-1}\mathbb{E}\langle\underline{G}\rangle^2 + \frac{1}{N}\mathbb{E}\big[G^*\big]^n[G]^{n-1}\langle\underline{G^2}\rangle
\end{aligned} \tag{4.10}$$

$$+\frac{2n-2}{N^2}\mathbb{E}[G^*]^n[G]^{n-2}(\underline{G^3}+2\underline{HG}mm')-\mathbb{E}K-\sum_{k=2}^{l}\mathbb{E}L_k$$
$$+\frac{2n}{N^2}\mathbb{E}[G^*]^{n-1}[G]^{n-1}(\underline{GG^{*2}}+2\underline{HG^*}\bar{m}\bar{m}')-\frac{1}{N}\sum_{i,j}\mathbb{E}\mathcal{R}_{l+1}^{(ji)}\Big),$$

where $T:=-z-2\mathbb{E}\underline{G}$. Note that by (2.3) and Theorem 2.6 we have

$$\frac{1}{T}=\frac{1}{-z-2m}+O_{\prec}\Big(\frac{1}{q}+\frac{1}{N\eta}\Big)=O_\tau(1) \tag{4.11}$$

uniformly for all $z\in\mathbf{D}_\tau$. Let us look at the terms in (4.10). By (3.3) we have

$$\langle\underline{G}\rangle=[G]+O_{\prec}(N^{-1/2-\beta}). \tag{4.12}$$

Together with Lemma 2.10(i) and Hölder's inequality we get

$$\mathbb{E}[G^*]^n[G]^{n-1}\langle\underline{G}\rangle^2\prec\Big(\frac{1}{q}+\frac{1}{N\eta}\Big)\mathcal{M}^{2n}+\Big(\frac{1}{q}+\frac{1}{N\eta}\Big)\frac{1}{\sqrt{N}q}\mathcal{M}^{2n-1} \tag{4.13}$$

and

$$\mathbb{E}[G^*]^n[G]^{n-1}\mathbb{E}\langle\underline{G}\rangle^2\prec\Big(\frac{1}{q}+\frac{1}{N\eta}\Big)\mathcal{M}^{2n}+\Big(\frac{1}{q}+\frac{1}{N\eta}\Big)\frac{1}{\sqrt{N}q}\mathcal{M}^{2n-1}. \tag{4.14}$$

Similarly, by Lemma 2.10(i), (iii) and Hölder's inequality we have

$$\frac{1}{N}\mathbb{E}[G^*]^n[G]^{n-1}\langle\underline{G^2}\rangle\prec\Big(\frac{1}{q}+\frac{1}{N\eta}\Big)\frac{1}{N\eta}\mathcal{M}^{2n-1} \tag{4.15}$$

and

$$\begin{aligned}&\frac{2n-2}{N^2}\mathbb{E}[G^*]^n[G]^{n-2}\underline{G^3}\\&=\frac{2n-2}{N^2}\mathbb{E}[G^*]^n[G]^{n-2}(\langle\underline{G^3}\rangle+\mathbb{E}\underline{G^3})\prec\Big(\frac{1}{N\eta}+\frac{1}{q}+\eta\Big)\Big(\frac{1}{N\eta}\Big)^2\mathcal{M}^{2n-2}.\end{aligned} \tag{4.16}$$

Note that

$$HG=I+zG\quad\text{and}\quad|z|\le 6, \tag{4.17}$$

hence

$$\begin{aligned}&\frac{4n-4}{N^2}\mathbb{E}[G^*]^n[G]^{n-2}\underline{HG}mm'\\&=\frac{4n-4}{N^2}\mathbb{E}[G^*]^n[G]^{n-2}(1+z\underline{G})mm'\prec\frac{1}{N^2}\mathcal{M}^{2n-2}.\end{aligned} \tag{4.18}$$

From resolvent identity, Theorem 2.6 and Lemma 2.10(iii) we have

$$\mathbb{E}\underline{GG^{*2}}=\frac{\mathbb{E}\underline{G}-\mathbb{E}\underline{G^*}}{(2\mathrm{i}\eta)^2}-\frac{\mathbb{E}\underline{G^{*2}}}{2\mathrm{i}\eta}=-\frac{m-\bar{m}}{4\eta^2}+O_{\prec}\Big(\Big(\frac{1}{q}+\frac{1}{N\eta}+\eta\Big)\eta^{-2}\Big).$$

Thus

$$\begin{aligned}&\frac{2n}{N^2}\mathbb{E}[G^*]^{n-1}[G]^{n-1}\underline{GG^{*2}}\\&=\frac{2n}{N^2}\mathbb{E}[G^*]^{n-1}[G]^{n-1}(\langle\underline{GG^{*2}}\rangle+\mathbb{E}\underline{GG^{*2}})\\&=-\frac{n(m-m^*)}{2N^2\eta^2}\mathbb{E}|[G]|^{2n-2}+O_{\prec}\Big(\frac{1}{q}+\frac{1}{N\eta}+\eta\Big)\Big(\frac{1}{N\eta}\Big)^2\mathcal{M}^{2n-2}.\end{aligned} \tag{4.19}$$

Similarly, by $N\mathcal{C}_2(H_{ii}) \asymp 1$ and the differential rule (2.1) one can easily check that

$$\mathbb{E}K \prec \frac{1}{N}\mathcal{M}^{2n-1} + \frac{1}{N^2\eta}\mathcal{M}^{2n-2}. \tag{4.20}$$

The estimate for the remainder term can be done routinely. One can follow, for example, the proof of Lemma 3.4 (iii) in [16], and readily check that

$$\frac{1}{N}\sum_{i,j}\mathbb{E}\mathcal{R}_{l+1}^{(ji)} \prec \frac{1}{N^{2n}} \tag{4.21}$$

for l large enough. From now on, we shall always assume the remainder term in cumulant expansion is negligible. Inserting the above estimates (4.11), (4.13)–(4.16), (4.18)–(4.21) into (4.10), we have

$$\begin{aligned}\mathbb{E}\big[G^*\big]^n[G]^{n-1}\langle \underline{G}\rangle = {} & \frac{n}{2N^2\eta^2}\mathbb{E}\big|[G]\big|^{2n-2} - \frac{1}{T}\sum_{k=2}^{l}\mathbb{E}L_k \\ & + O_{\prec}(N^{-\delta})\mathcal{M}^{2n} + \sum_{r=1,2} O_{\prec}\Big(\frac{N^{-\xi}}{(N\eta)^r} + \Big(\frac{N^{-\delta}}{\sqrt{N}q}\Big)^r\Big)\mathcal{M}^{2n-r},\end{aligned} \tag{4.22}$$

where we recall the definitions of δ, ξ from (3.1), (3.2). What is left, therefore, is the analysis of $\mathbb{E}L_k$, $k \geq 2$.

4.2. *Abstract polynomials and the recursive estimates.* We now introduce some additional notations that will be used frequently in the analysis of L_k. To motivate them, we note that the proof relies on a calculus of products of expectations of random variables of the type $(G^m)_{ij}$, $(G^{*n})_{ij}$, $[G]$, $[G^*]$ evaluated at $z \in \mathbf{D}_\tau$, for example,

$$a(z,z^*)N^{3/2}\mathbb{E}[G]\big[G^*\big], \qquad a_{i_1i_2i_3i_4}(z,z^*)N^{-1/2}\mathbb{E}\big(G^3\big)_{i_1i_2}G_{i_3i_4}\mathbb{E}\big[G^*\big]^2,$$

where $a(z,z^*)$ and $a_{i_1i_2i_3i_4}(z,z^*)$ are uniformly bounded functions that may depend on N. It is convenient to classify such expressions depending on the exponent $\nu_1 \in \mathbb{R}$ and on the number ν_0 of indices i_k. Below, we introduce the notations $\mathcal{U}^{(\nu_0,\nu_1)}(\mathcal{Y})$, $\mathcal{V}^{(\nu_0,\nu_1)}(\mathcal{Y})$ for the set of such expressions, where $\mathcal{Y}$ is the set of matrices appearing in them, in the above examples $\mathcal{Y} = \{G, G^*\}$.

To that end, we define a set of formal monomials in a set of formal variables. Here the word *formal* refers to the fact that these definitions are purely algebraic and we do not assign any values to variables or monomials. The formal variables are constructed from a finite set of formal matrices $\mathcal{Y}$ and the infinite set of formal indices $\{i_1, i_2, \ldots\}$.

- For $\nu_0 \in \mathbb{N}$, $\nu_1 \in \mathbb{R}$, denote by $\mathcal{U}^{(\nu_0,\nu_1)}(\mathcal{Y})$ the set of monomials with coefficient $a_{i_1,\ldots,i_{\nu_0}}N^{-\nu_1}$ in the variables $(Y^m)_{xy}$ and $[Y]$. Here $Y \in \mathcal{Y}$, $m \in \mathbb{N}_+$, $x, y \in \{i_1, \ldots, i_{\nu_0}\}$, and $(a_{i_1,\ldots,i_{\nu_0}})_{1\leq i_1,\ldots,i_{\nu_0}\leq N}$ is some family of complex numbers that is uniformly bounded in $i_1, \ldots, i_{\nu_0}$.
- Set $\mathcal{U}(\mathcal{Y}) = \bigcup_{\nu_0,\nu_1}\mathcal{U}^{(\nu_0,\nu_1)}(\mathcal{Y})$.

We also define the following subset of $\mathcal{U}(\mathcal{Y})$.

- We denote by $\mathcal{V}^{(\nu_0,\nu_1)}(\mathcal{Y})$ the subset of $\mathcal{U}^{(\nu_0,\nu_1)}(\mathcal{Y})$, where we further require $m \in \{1,2\}$ for all variables $(Y^m)_{xy}$.
- Set $\mathcal{V}(\mathcal{Y}) = \bigcup_{\nu_0,\nu_1}\mathcal{V}^{(\nu_0,\nu_1)}(\mathcal{Y}) \subset \mathcal{U}(\mathcal{Y})$.

Next, we define the following maps $\nu_0, \nu_1, \nu_2, \nu_3, \widetilde{\nu}_3, \nu_4 \colon \mathcal{U}(\mathcal{Y}) \to \mathbb{N}$.

(i) For $U \in \mathcal{U}^{(\nu_0,\nu_1)}(\mathcal{Y})$, $(\nu_0(U), \nu_1(U)) = (\nu_0, \nu_1)$.

(ii) $\nu_2(U) =$ sum of $m-1$ of all $(Y^m)_{xy}$ in U with $Y \in \mathcal{Y}$.
(iii) $\nu_3(U) = 2\wedge$ (number of $(Y^m)_{xy}$ in U with $x \neq y$ and $Y \in \mathcal{Y}$). Set $\widetilde{\nu}_3(U) = 2 - \nu_3(U)$.
(iv) $\nu_4(U) =$ number of $[Y]$ in U with $Y \in \mathcal{Y}$.

Next, we assign to each monomial $U \in \mathcal{U}^{(\nu_0,\nu_1)}(\mathcal{Y})$ a value $U_{i_1,\dots,i_{\nu_0}}$ as follows. Suppose that the set $\mathcal{Y}$ consists of $N \times N$ random matrices. Then for any ν_0-tuple $(i_1,\dots,i_{\nu_0}) \in \{1,2,\dots,N\}^{\nu_0}$ we define the number $U_{i_1,\dots,i_{\nu_0}}$ as the one obtained by taking the formal expression U and evaluating it with the laws of the matrices in $\mathcal{Y}$ and the numerical values of $i_1,\dots,i_{\nu_0}$. In the following arguments, the set $\mathcal{Y}$ will consist of Green functions of H for the spectral parameter z defined in (4.1), and the indices $i_1,\dots,i_{\nu_0}$ will be summed over.

The next result is a straightforward consequence of Lemma 2.10(i), (iii) and Hölder's inequality whose proof we omit.

LEMMA 4.2. *Let $\mathcal{Y} = \{G, G^*\}$, and fix $U \in \mathcal{U}^{(\nu_0,\nu_1)}(\mathcal{Y})$. Then*

$$\sum_{i_1,\dots,i_{\nu_0}} \mathbb{E}U_{i_1,\dots,i_{\nu_0}} = O_{\prec}\big(N^{b_0(U)}\big)\cdot \mathcal{M}^{\nu_4(U)}, \tag{4.23}$$

where $b_0(U) = \nu_0(U) - \nu_1(U) + \alpha\nu_2(U) + (\alpha/2 - 1/2)\nu_3(U)$.

Our first estimate is the following improved bound for the LHS of (4.23), whose proof is postponed to Section 5.1. The necessity of this result is explained in Remark 4.10 below.

LEMMA 4.3. *Let us adopt the assumptions in Lemma 4.2. Let $V \in \mathcal{V}^{(\nu_0,\nu_1)}(\mathcal{Y})$ satisfying $\nu_2(V) \geq 1$, we have*

$$\sum_{i_1,\dots,i_{\nu_0}} \mathbb{E}V_{i_1,\dots,i_{\nu_0}} \prec \mathcal{B}(V),$$

where

$$\mathcal{B}(V) = N^{b_0(V)}\mathcal{M}^{\nu_4(V)+1} + \sum_{k=0}^{\nu_4(V)} N^{b_0(V)}\frac{1}{(N\eta)^k}\bigg(\eta + \frac{1}{(N\eta)^{\widetilde{\nu}_3(V)/2}}\bigg)\cdot \mathcal{M}^{\nu_4(V)-k}.$$

EXAMPLE 4.4. Let $\mathcal{Y} = \{G, G^*\}$ and set

$$U \equiv U_{ij} := \frac{1}{N^2}\mathcal{C}_4(H_{ij})\mathbb{E}[G]^{n-1}[G^*]^{n-1}(G^{*2})_{ii}G^*_{jj}G_{ii}G_{jj}.$$

Note that $Nq^2\mathcal{C}_4(H_{ij}) \asymp 1$, and thus $U \in \mathcal{V}^{(\nu_0,\nu_1)}(\mathcal{Y}) \subset \mathcal{U}^{(\nu_0,\nu_1)}(\mathcal{Y})$, with $\nu_0 = 2$ and $\nu_1 = 2 + (1 + 2\log_N q) = 3 + 2\beta$. We also have $\nu_2(U) = 2 - 1 = 1$, $\nu_3(U) = 2 \wedge 0 = 0$, $\widetilde{\nu}_3(U) = 2 - 0 = 2$, $\nu_4(U) = 2n - 2$, and $b_0(U) = -1 - 2\beta + \alpha$.

By Lemma 2.10, we have

$$\sum_{i,j} U_{ij} \prec \frac{1}{N^3q^2}\cdot N^2\cdot \mathcal{M}^{2n-2}\cdot\frac{1}{\eta} = \frac{1}{Nq^2\eta}\cdot\mathcal{M}^{2n-2} = N^{b_0(U)}\cdot\mathcal{M}^{2n-2},$$

which agrees with Lemma 4.2. On the other hand, Lemma 4.3 implies the improved estimate

$$\sum_{i,j} U_{ij} \prec \frac{1}{Nq^2\eta}\mathcal{M}^{2n-1} + \sum_{k=0}^{2n-2}\frac{1}{Nq^2\eta}\frac{1}{(N\eta)^k}\bigg(\eta + \frac{1}{N\eta}\bigg)\cdot\mathcal{M}^{2n-2-k}.$$

In order to handle all terms in L_k, we also need the following formal polynomials.

- For $\nu_0 \in \mathbb{N}, \nu_1 \in \mathbb{R}$, denote by $\mathcal{W}^{(\nu_0,\nu_1)}(\mathcal{Y})$ the set of monomials with coefficient $a_{i_1,\dots,i_{\nu_0}} N^{-\nu_1}$ in the variable $[Y]$ and also contain exactly one factor of $\langle Y^{(1)}_{x_1 y_1} \cdots Y^{(k)}_{x_k y_k} \rangle$. Here $Y, Y^{(1)}, \dots, Y^{(k)} \in \mathcal{Y}$, $k \in \mathbb{N}_+$, $x, y \in \{i_1, \dots, i_{\nu_0}\}$, and $(a_{i_1,\dots,i_{\nu_0}})_{1 \le i_1,\dots,i_{\nu_0} \le N}$ is some family of complex numbers that is uniformly bounded in $i_1, \dots, i_{\nu_0}$.
- Set $\mathcal{W}(\mathcal{Y}) = \bigcup_{\nu_0,\nu_1} \mathcal{W}^{(\nu_0,\nu_1)}(\mathcal{Y})$.

Next, we define the following maps $\nu_0, \nu_1, \nu_3, \widetilde{\nu}_3, \nu_4 \colon \mathcal{W}(\mathcal{Y}) \to \mathbb{N}$.

(i) For $W \in \mathcal{W}^{(\nu_0,\nu_1)}(\mathcal{Y})$, $(\nu_0(W), \nu_1(W)) = (\nu_0, \nu_1)$.
(ii) $\nu_3(W) = 2\wedge$ (number of Y_{xy} in W with $x \neq y$ and $Y \in \mathcal{Y}$). Set $\widetilde{\nu}_3(W) = 2 - \nu_3(W)$.
(iii) $\nu_4(W) =$ number of $[Y]$ in W with $Y \in \mathcal{Y}$.

The following is a trivial result from Lemma 2.9.

LEMMA 4.5. *Let $\mathcal{Y} = \{G, G^*\}$, and fix $W \in \mathcal{W}^{(\nu_0,\nu_1)}(\mathcal{Y})$. Then*

$$\sum_{i_1,\dots,i_{\nu_0}} \mathbb{E} W_{i_1,\dots,i_{\nu_0}} = O_{\prec}\big(N^{b_1(W)}\big) \cdot \mathcal{M}^{\nu_4(W)},$$

where $b_1(W) = \nu_0(W) - \nu_1(W) + (\alpha/2 - 1/2)\nu_3(W)$.

We have the following improved estimate for Lemma 4.5, whose proof is postponed to Section 5.2. The necessity of this result is explained in Remark 4.10 below.

LEMMA 4.6. *Let us adopt the assumptions in Lemma 4.5. We have*

$$\sum_{i_1,\dots,i_{\nu_0}} \mathbb{E} W_{i_1,\dots,i_{\nu_0}} \prec \mathcal{B}^{(1)}(W),$$

where

$$\mathcal{B}^{(1)}(W) = N^{b_1(W)} \bigg(\frac{1}{(N\eta)^{\widetilde{\nu}_3(W)/2}} + \frac{1}{\sqrt{N}q} + \mathcal{M} \bigg) \cdot \mathcal{M}^{\nu_4(W)} + \sum_{k=1}^{\nu_4(V)} N^{b_1(W)} \frac{1}{(N\eta)^k} \bigg(\frac{\eta}{q} + \frac{1}{(N\eta)^{\widetilde{\nu}_3(W)/2}} \bigg) \cdot \mathcal{M}^{\nu_4(W)-k}.$$

We close this section with the following estimate.

LEMMA 4.7. *Let $\mathcal{Y} = \{G, G^*\}$, and fix $U \in \mathcal{U}^{(\nu_0,\nu_1)}(\mathcal{Y})$. For $i, j \in \{i_1, \dots i_{\nu_0}\}$, we have*

$$\sum_{i_1,\dots,i_{\nu_0}} \mathbb{E} H_{ij} U_{i_1,\dots,i_{\nu_0}} = O_{\prec}\big(N^{b_2(U)}\big) \cdot \sum_{k=0}^{\nu_4(U)} \frac{1}{(N\eta)^k} \mathcal{M}^{\nu_4(U)-k}, \tag{4.24}$$

where $b_2(U) = \nu_0(U) - \nu_1(U) + \alpha\nu_2(U) - 1$.

PROOF. The proof follows by applying Lemma 2.2 on LHS of (4.24) with $h = H_{ij}$, and then estimating the result by Lemma 4.2. We omit the details here. □

In the next two sections we shall estimate $\mathbb{E}L_2$ and $\mathbb{E}L_3$ using the above lemmas.

4.3. *The estimate of* $\mathbb{E}L_2$. In this section we prove the following result.

LEMMA 4.8. *Let* L_2 *be as in* (4.22). *Let* δ, ξ *be as in* (3.1), (3.2). *We have*

$$\mathbb{E}L_2 \prec N^{-\delta}\mathcal{M}^{2n} + \sum_{r=1}^{2n}\Big(\frac{N^{-\xi}}{(N\eta)^r} + \Big(\frac{N^{-\delta/4}}{\sqrt{Nq}}\Big)^r\Big)\mathcal{M}^{2n-r}. \tag{4.25}$$

PROOF. The differential ∂H_{ij}^2 gives rise to terms of three types depending on how many derivatives act on G_{ij}. We deal with each type separately.

Step 1. Let us look at the case when both derivatives in L_2 act on G_{ij}, namely the term

$$\mathbb{E}L_{2,1} := N^{-1}\cdot\sum_{i,j}\Big[\frac{1}{2!}\mathcal{C}_3(H_{ji})\mathbb{E}\langle[G^*]^n[G]^{n-1}\rangle\frac{\partial^2 G_{ij}}{\partial H_{ji}^2}\Big].$$

By Lemma 2.3 and the identity $\mathbb{E}\langle X\rangle Y = \mathbb{E}X\langle Y\rangle$, we see that the worst term in $L_{2,1}$ is

$$\frac{1}{N^2 q}\sum_{i,j}C_{ij}\mathbb{E}[G^*]^n[G]^{n-1}\langle G_{ii}G_{jj}G_{ij}\rangle =: \sum_{i,j}\mathbb{E}W_{i,j}, \tag{4.26}$$

where C_{ij} are constants uniformly bounded in i, j, and $W \in \mathcal{W}(\{G, G^*\})$. Note that $\nu_0(W) = 2$, $\nu_1(W) = 2+\beta$, $\nu_3(W) = 1$, $\nu_4(W) = 2n-1$, and $b_1(W) = -\beta - (1/2 - \alpha/2)$. Thus Lemma 4.6 shows

$$\begin{aligned}\sum_{i,j}\mathbb{E}W_{ij} &\prec \frac{1}{\sqrt{N\eta}q}\mathcal{M}^{2n} + \frac{1}{\sqrt{N\eta}q}\Big(\frac{1}{\sqrt{N\eta}} + \frac{1}{\sqrt{Nq}}\Big)\cdot\mathcal{M}^{2n-1}\\ &\quad + \sum_{k=1}^{2n-1}\frac{1}{\sqrt{N\eta}q}\frac{1}{(N\eta)^k}\Big(\frac{\eta}{q} + \frac{1}{\sqrt{N\eta}}\Big)\cdot\mathcal{M}^{2n-1-k}\\ &\prec N^{-\beta}\mathcal{M}^{2n} + N^{-\beta}\Big(\frac{1}{N\eta} + \frac{1}{\sqrt{Nq}}\Big)\cdot\mathcal{M}^{2n-1}\\ &\quad + \sum_{r=2}^{2n}\Big(\frac{N^{-\beta}}{\sqrt{Nq}}\frac{N^{-\alpha/2}}{(N\eta)^{r-1}} + \frac{N^{-\beta}}{(N\eta)^r}\Big)\mathcal{M}^{2n-r},\end{aligned}$$

which is bounded by the RHS of (4.25). Similarly, one can show that the other terms in $L_{2,1}$ satisfy the same bound.

Step 2. Let us look at the case when only one derivative in L_2 acts on G_{ij}, namely the term

$$\mathbb{E}L_{2,2} := N^{-1}\cdot\sum_{i,j}\Big[\mathcal{C}_3(H_{ji})\mathbb{E}\frac{\partial\langle[G^*]^n[G]^{n-1}\rangle}{\partial H_{ji}}\frac{\partial G_{ij}}{\partial H_{ji}}\Big]. \tag{4.27}$$

By (2.1), we see that the worst terms above will contain no off-diagonal terms of G from the second differential. Let us pick a representative of these, which is

$$\begin{aligned}&\frac{1}{N^3 q}\sum_{i,j}C_{ij}\mathbb{E}[G^*]^{n-1}[G]^{n-1}((G^{*2})_{ij} + 2H_{ij})G_{ii}G_{jj}\\ &= \frac{1}{N^3 q}\sum_{i,j}C_{ij}\mathbb{E}[G^*]^{n-1}[G]^{n-1}(G^{*2})_{ij}G_{ii}G_{jj}\\ &\quad + O_{\prec}\Big(\frac{1}{N^2 q}\Big)\sum_{k=0}^{2n-2}\frac{1}{(N\eta)^k}\mathcal{M}^{2n-2-k}\end{aligned} \tag{4.28}$$

$$=: \frac{1}{N^3 q}\sum_{i,j}\mathbb{E}V_{ij} + O_{\prec}\Big(\frac{1}{N^2 q}\Big)\sum_{k=0}^{2n-2}\frac{1}{(N\eta)^k}\mathcal{M}^{2n-2-k}.$$

Here C_{ij} are constants uniformly bounded in i, j, and in the first step of (4.28) we used Lemma 4.7. Note that $V \in \mathcal{V}$ satisfies $\nu_0(V) = 2$, $\nu_1(V) = 3 + \beta$, $\nu_2(V) = 1$, $\nu_3(V) = 1$, $\nu_4(V) = 2n - 2$, and $b_0(V) = 3(\alpha - 1)/2 - \beta$. Thus Lemma 4.3 shows

$$\begin{aligned}&\frac{1}{N^3 q}\sum_{i,j}\mathbb{E}V_{ij}\\ &\prec \frac{1}{(N\eta)^{3/2}q}\mathcal{M}^{2n-1} + \sum_{k=0}^{2n-2}\frac{1}{(N\eta)^{3/2}q}\Big(\frac{1}{N\eta}\Big)^k\Big(\eta + \frac{1}{\sqrt{N\eta}}\Big)\cdot\mathcal{M}^{2n-2-k} \qquad (4.29)\\ &\prec \frac{N^{-\beta}}{N\eta}\mathcal{M}^{2n-1} + \sum_{r=2}^{2n}\Big(\frac{1}{(N\eta)^{r-3/2}Nq} + \frac{N^{-\beta}}{(N\eta)^r}\Big)\mathcal{M}^{2n-r}.\end{aligned}$$

By (4.29) and

$$\frac{1}{(N\eta)^{r-3/2}Nq} = \frac{1}{(N\eta)^{r-3/2}(\sqrt{N}q)^{3/2}}\Big(\frac{q^2}{N}\Big)^{1/4} \prec \Big(\frac{N^{-\delta/8}}{(N\eta)^r} + \frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^r,$$

we see that (4.28) is bounded by the RHS of (4.25). Similarly, the other terms in $L_{2,1}$ can be shown to satisfy the same bound.

Step 3. Let us look at the case when no derivatives in L_2 act on G_{ij}, namely the term

$$\mathbb{E}L_{2,3} := N^{-1}\cdot\sum_{i,j}\Big[\frac{1}{2!}\mathcal{C}_3(H_{ji})\mathbb{E}\frac{\partial^2\langle[G^*]^n[G]^{n-1}\rangle}{\partial H_{ji}^2}G_{ij}\Big].$$

Similarly as in Step 2, one can use Lemma 4.3 to show that

$$\mathbb{E}L_{2,3} \prec \sum_{r=1}^{2n}\Big(\frac{N^{-\xi}}{(N\eta)^r} + \Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^r\Big)\mathcal{M}^{2n-r}.$$

From Steps 1–3 we conclude the proof. □

4.4. *The estimate of* $\mathbb{E}L_3$. Now let us look at the case $k = 3$. This is the crucial case where we see the cancellation between $\langle\underline{G}\rangle$ and $(\underline{H^2} - 1)mm'$. We shall prove the following lemma.

LEMMA 4.9. *Let* L_3 *be as in* (4.22). *Let* δ, ξ *be as in* (3.1), (3.2). *We have*

$$\mathbb{E}L_3 \prec N^{-\delta}\mathcal{M}^{2n} + \sum_{r=1}^{2n}\Big(\frac{N^{-\xi}}{(N\eta)^r} + \Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^r\Big)\mathcal{M}^{2n-r}. \qquad (4.30)$$

PROOF. We still split the estimates basing on how many derivatives hit G_{ij}.

Step 1. We investigate the case when all derivatives in L_3 act on G_{ij}, namely the term

$$\mathbb{E}L_{3,1} := N^{-1}\cdot\sum_{i,j}\Big[\frac{1}{3!}\mathcal{C}_4(H_{ji})\mathbb{E}\langle[G^*]^n[G]^{n-1}\rangle\frac{\partial^3 G_{ij}}{\partial H_{ji}^3}\Big].$$

From Lemma 2.3 we see that the worst term in $L_{3,1}$ is

$$\frac{1}{N^2q^2}\sum_{i,j}C_{ij}\mathbb{E}[G^*]^n[G]^{n-1}\langle G_{ii}^2G_{jj}^2\rangle =: \frac{1}{N^2q^2}\sum_{i,j}\mathbb{E}W_{ij},$$

where $W \in \mathcal{W}$. Note that $\nu_0(W) = 2$, $\nu_1(W) = 2 + 2\beta$, $\nu_3(W) = 0$, $\nu_4(W) = 2n - 1$, and $b_1(W) = -2\beta$. Thus Lemma 4.6 shows

$$\begin{aligned}\frac{1}{N^2q^2}\sum_{i,j}\mathbb{E}W_{ij} &\prec \frac{1}{q^2}\mathcal{M}^{2n} + \frac{1}{q^2}\left(\frac{1}{N\eta} + \frac{1}{\sqrt{N}q}\right)\cdot\mathcal{M}^{2n-1} \\ &\quad + \sum_{k=1}^{2n-1}\frac{1}{q^2}\frac{1}{(N\eta)^k}\left(\frac{\eta}{q} + \frac{1}{N\eta}\right)\cdot\mathcal{M}^{2n-1-k} \\ &\prec N^{-2\beta}\cdot\mathcal{M}^{2n} + N^{-2\beta}\left(\frac{1}{N\eta} + \frac{1}{\sqrt{N}q}\right)\cdot\mathcal{M}^{2n-1} \\ &\quad + \sum_{r=2}^{2n} O_{\prec}\left(\frac{N^{-\beta/2}}{Nq^2}\frac{N^{-\beta/2}}{(N\eta)^{r-2}} + \frac{N^{-2\beta}}{(N\eta)^r}\right)\cdot\mathcal{M}^{2n-r},\end{aligned}$$

which is bounded by RHS of (4.30).

Step 2. Let us look at the case when only one derivative in L_3 acts on G_{ij}, namely the term

$$\mathbb{E}L_{3,2} := \frac{1}{2N}\cdot\sum_{i,j}\left[\mathcal{C}_4(H_{ji})\mathbb{E}\frac{\partial^2\langle[G^*]^n[G]^{n-1}\rangle}{\partial H_{ji}^2}\frac{\partial G_{ij}}{\partial H_{ji}}\right].$$

We see that one of the worst terms is

$$\begin{aligned}&-\frac{n}{2N^3q^2}\sum_{i,j}s_4(1 + C_i\delta_{ij})\mathbb{E}[G^*]^{n-1}[G]^{n-1}\big(4(G^{*2})_{ii}G^*_{jj} - 4\bar{m}\bar{m}'\big)G_{ii}G_{jj} \\ &= -\frac{n}{2N^3q^2}\sum_{i,j}s_4\mathbb{E}[G^*]^{n-1}[G]^{n-1}\big(4(G^{*2})_{ii}G^*_{jj} - 4\bar{m}\bar{m}'\big)G_{ii}G_{jj} \\ &\quad + O_{\prec}\left(\frac{1}{N\eta}\frac{1}{Nq^2}\right)\mathcal{M}^{2n-2},\end{aligned} \tag{4.31}$$

where $s_4 = Nq^2\mathcal{C}_4(H_{12}) \asymp 1$, and C_i are constants uniformly bounded in i. Now let us look at the first term on RHS of (4.31), which is

$$-\frac{2n}{N^3q^2}\sum_{i,j}s_4\mathbb{E}[G^*]^{n-1}[G]^{n-1}(G^{*2})_{ii}G^*_{jj}G_{ii}G_{jj} =: \sum_{i,j}\mathbb{E}V_{ij},$$

where $V \in \mathcal{V}(\{G, G^*\})$. By the resolvent identity $\bar{z}G^* = HG^* - I$ and Lemma 2.2, we have

$$\begin{aligned}\sum_{i,j}\mathbb{E}V_{ij} = \frac{1}{\bar{z} + \mathbb{E}\underline{G^*}}\frac{2s_4n}{N^3q^2}\sum_{i,j}\mathbb{E}\Big[&[G^*]^{n-1}[G]^{n-1}G^*_{jj}G_{ii}G_{jj} \\ &\times\big(G^*_{ii}\underline{G^{*2}} + (G^{*2})_{ii}\langle\underline{G^*}\rangle + 2N^{-1}(G^{*3})_{ii}\big) \\ &+ \frac{1}{N}[G^*]^{n-1}[G]^{n-1}\cdot\big(2(G^{*3})_{ij}G^*_{ij}G_{ii}G_{jj} + 2(G^{*2}G)_{ii}G^*_{jj}G_{ii}G_{jj} \\ &+ 2(G^{*2}G)_{ij}G^*_{jj}G_{ii}G_{ij}\big) \\ &+ [G^*]^{n-1}[G]^{n-1}G^*_{ii}G^*_{jj}G_{ii}G_{jj} - \widetilde{K} - \sum_i\sum_{k=2}^{l}\widetilde{L}_k - \sum_{a=1}^{N}\widetilde{\mathcal{R}}^{(a)}_{l+1}\Big],\end{aligned}$$

where $\widetilde{K}$, $\widetilde{L}_k$, and $\widetilde{\mathcal{R}}^{(a)}_{l+1}$ are defined similarly as K, L_k, $\mathcal{R}^{(ij)}_{l+1}$ in (4.7). By (2.2) and Lemmas 4.2,4.3, one can check that

$$\sum_{i,j}\mathbb{E}V_{ij} = \frac{1}{\bar{z} + \mathbb{E}\underline{G^*}}\frac{2s_4n}{N^3q^2}\sum_{i,j}\mathbb{E}\big([G^*]^{n-1}[G]^{n-1}G^*_{jj}G_{ii}G_{jj}G^*_{ii}\underline{G^{*2}}$$

$$+[G^*]^{n-1}[G]^{n-1}G^*_{ii}G^*_{jj}G_{ii}G_{jj}) \tag{4.32}$$
$$+\sum_{r=1}^{2n} O_{\prec}\Big(\frac{N^{-\xi}}{(N\eta)^r}+\Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^r\Big)\mathcal{M}^{2n-r}.$$

Similarly, for the first term on RHS of (4.32), we have

$$\frac{1}{\bar z+\mathbb{E}\underline{G^*}}\frac{2s_4n}{N^3q^2}\sum_{i,j}\mathbb{E}[G^*]^{n-1}[G]^{n-1}G^*_{jj}G_{ii}G_{jj}G^*_{ii}\underline{G^{*2}}$$
$$=\frac{1}{-\bar z-2\mathbb{E}\underline{G^*}}\frac{1}{\bar z+\mathbb{E}\underline{G^*}}\frac{2s_4n}{N^3q^2}\sum_{i,j}\mathbb{E}[G^*]^{n-1}[G]^{n-1}G^*_{jj}G_{ii}G_{jj}G^*_{ii}\underline{G^*} \tag{4.33}$$
$$+\sum_{r=1}^{2n} O_{\prec}\Big(\frac{N^{-\xi}}{(N\eta)^r}+\Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^r\Big)\mathcal{M}^{2n-r}.$$

By (4.32), (4.33) and Theorem 2.6, we have

$$\sum_{i,j}\mathbb{E}V_{ij}=\frac{2s_4n}{N^3q^2}\frac{1}{\bar z+\bar m}\Big(\frac{\bar m^3}{-\bar z-2\bar m}+\bar m^2\Big)\sum_{i,j}\mathbb{E}[G^*]^{n-1}[G]^{n-1}G_{ii}G_{jj}$$
$$+\sum_{r=1}^{2n} O_{\prec}\Big(\frac{N^{-\xi}}{(N\eta)^r}+\Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^r\Big)\mathcal{M}^{2n-r}. \tag{4.34}$$

By $\bar m/(-\bar z-2\bar m)=\bar m'$ we have

$$\frac{1}{\bar z+\bar m}\Big(\frac{\bar m^3}{-\bar z-2\bar m}+\bar m^2\Big)=-\bar m\bar m'. \tag{4.35}$$

Combining (4.31), (4.34) and (4.35), we see the crucial cancellation of the first two terms on RHS of (4.31). As a result, we obtain

$$(4.31)\prec\sum_{r=1}^{2n}\Big(\frac{N^{-\xi}}{(N\eta)^r}+\Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^r\Big)\mathcal{M}^{2n-r}$$

as desired. The other terms in $\mathbb{E}L_{3,2}$ can be directly estimated by (2.1) and Lemma 4.3, and one readily checks that they satisfy the bound on RHS of (4.30).

Step 3. The remaining two cases, that is, when two derivatives or no derivative act on G_{ij}, can be analyzed similarly using (2.1) and Lemma 4.3. Note that the estimate is easier than those in Steps 1 and 2: by (2.1), every term now contains either at least two off diagonal entries of the Green function or derivatives of $\underline{H^2}-1$. We omit the details.

From Steps 1–3 we conclude the proof of Lemma 4.9. □

4.5. *Put things together.* Up to now, what is left, is the estimate of L_k for $k\geq 4$. This is similar but easier than the cases when $k=2,3$. In fact, by Lemma 2.3 we see that there will be additional factors of $1/q$ in L_k when $k\geq 4$. By a direct estimate using (2.1) and Lemma 4.3, we have

$$\sum_{k=4}^{l}\mathbb{E}L_k\prec N^{-\delta}\mathcal{M}^{2n}+\sum_{r=1}^{2n}\Big(\frac{N^{-\xi}}{(N\eta)^r}+\Big(\frac{N^{-\delta/4}}{\sqrt{N}q}\Big)^r\Big)\mathcal{M}^{2n-r} \tag{4.36}$$

for any fixed $l\in\mathbb{N}_+$.

By (4.22), (4.36), Lemmas 4.8 and 4.9, we conclude the proof of (4.2).

REMARK 4.10. The input of Lemmas 4.3 and 4.6 are essential in finishing the proof. For example, in the estimate of (4.26), the bound from Lemma 4.2 only implies

$$(4.26) \prec \frac{1}{N^2 q} \cdot N^2 \cdot \frac{1}{\sqrt{N\eta}} \cdot \mathcal{M}^{2n-1} = \frac{1}{\sqrt{N\eta} q} \mathcal{M}^{2n-1},$$

which is not enough to deduce (4.25). Also, in the estimate of (4.28), the trivial bound from Lemma 4.5 only implies

$$\frac{1}{N^3 q} \sum_{i,j} \mathbb{E} V_{ij} \prec \frac{1}{(N\eta)^{3/2} q} \mathcal{M}^{2n-2},$$

which is not enough to deduce (4.25).

4.6. *Proof of* (4.3). The proof of (4.3) is similar to that of (4.2). We only sketch the main steps.

Step 1. By Lemma 2.2, we have

$$\begin{aligned}
\mathbb{E}[G^*]^n [G]^{n-1} (\underline{H^2} - 1) &= \frac{1}{N} \sum_{i,j} \mathbb{E} H_{ij}^2 [G^*]^n [G]^{n-1} - \mathbb{E}[G^*]^n [G]^{n-1} \\
&= \mathbb{E}[G^*]^n [G]^{n-1} - \frac{n}{N} \mathbb{E}[G^*]^{n-1} [G]^{n-1} (2\underline{HG^2} + 4\underline{H^2} mm') \\
&\quad - \frac{n-1}{N} \mathbb{E}[G^*]^n [G]^{n-2} (2\underline{HG^{*2}} + 4\underline{H^2} \bar{m}\bar{m}') \\
&\quad + \mathbb{E}\widehat{K} + \sum_{k=2}^{l} \mathbb{E}\widehat{L}_{k,1} + \sum_{k=2}^{l} \mathbb{E}\widehat{L}_{k,2} + \sum_{i,j} \widehat{\mathcal{R}}_{l+1}^{(ij)} - \mathbb{E}[G^*]^n [G]^{n-1},
\end{aligned} \tag{4.37}$$

where

$$\widehat{K} = N^{-2} \sum_i \frac{\partial (H_{ii} [G^*]^n [G]^{n-1})}{\partial H_{ii}} (N\mathcal{C}_2(H_{ii}) - 2),$$

$$\widehat{L}_{k,1} = N^{-1} \cdot \sum_{i,j} \Big(\frac{1}{k!} \mathcal{C}_{k+1}(H_{ji}) H_{ij} \frac{\partial^k ([G^*]^n [G]^{n-1})}{\partial H_{ji}^k} \Big)$$

and

$$\widehat{L}_{k,2} = N^{-1} \cdot \sum_{i,j} \Big(\frac{1}{(k-1)!} \mathcal{C}_{k+1}(H_{ji}) \frac{\partial^{k-1} ([G^*]^n [G]^{n-1})}{\partial H_{ji}^{k-1}} \Big).$$

Here l is a fixed positive integer to be chosen later, and $\widehat{\mathcal{R}}_{l+1}^{(ji)}$ is a remainder term defined analogously to $\mathcal{R}_{l+1}$ in (2.4). Notice the cancellation between the first and last terms on RHS of (4.37). By (3.3) and (4.17) we have

$$\begin{aligned}
&\mathbb{E}[G^*]^n [G]^{n-1} (\underline{H^2} - 1) \\
&= O_\prec\Big(\frac{1}{N}\Big) \mathcal{M}^{2n-1} + O_\prec\Big(\frac{1}{N^2 \eta}\Big) \mathcal{M}^{2n-2} + \sum_{k=2}^{l} \mathbb{E}\widehat{L}_{k,1} + \sum_{k=2}^{l} \mathbb{E}\widehat{L}_{k,2}.
\end{aligned} \tag{4.38}$$

Step 2. Let us estimate $\sum_{k=2}^{l} \mathbb{E}\widehat{L}_{k,1}$. By Lemma 2.2 we have

$$\sum_{k=2}^{l} \mathbb{E}\widehat{L}_{k,1} = \sum_{k=2}^{l} \sum_{s=1}^{l'} N^{-1} \cdot \sum_{i,j} \Big(\frac{1}{k!s!} \mathcal{C}_{k+1}(H_{ji}) \mathcal{C}_{s+1}(H_{ij}) \mathbb{E} \frac{\partial^{k+s} ([G^*]^n [G]^{n-1})}{\partial H_{ji}^{k+s}} \Big),$$

assuming the remainder term is small enough for large l'. By Lemma 2.3 and

$$\frac{\partial^m [G]}{\partial H_{ij}^m} \prec \frac{1}{N\eta},$$

we have

$$\sum_{k=2}^{l} \mathbb{E}\widehat{L}_{k,1} \prec \sum_{r=2}^{2n} \frac{1}{Nq} \frac{1}{(N\eta)^{r-1}} \mathcal{M}^{2n-r} \prec \sum_{r=2}^{2n-r} \left(\frac{N^{-1/4}}{(N\eta)^r} + \left(\frac{N^{-1/4}}{\sqrt{Nq}} \right)^r \right) \mathcal{M}^{2n-r}.$$

Step 3. Now let us estimate at $\sum_{k=2}^{l} \mathbb{E}\widehat{L}_{k,2}$. Note that in Section 4.3, we have estimated $\mathbb{E}L_{2,2}$ defined in (4.27). In particular, we have estimated the term

$$N^{-1} \sum_{i,j} \left[\mathcal{C}_3(H_{ji}) \mathbb{E} \frac{\partial \langle [G^*]^n [G]^{n-1} \rangle}{\partial H_{ji}} G_{ii} G_{jj} \right],$$

where the method used can be applied almost exactly in estimating $\mathbb{E}\widehat{L}_{2,2}$. Similarly, we have estimated

$$\frac{1}{2N} \sum_{i,j} \left[\mathcal{C}_4(H_{ji}) \mathbb{E} \frac{\partial^2 \langle [G^*]^n [G]^{n-1} \rangle}{\partial H_{ji}^2} G_{ii} G_{jj} \right]$$

in Section 4.4, and this method can be applied in estimating $\mathbb{E}\widehat{L}_{3,2}$. Additionally, we can also estimate $\mathbb{E}\widehat{L}_{k,2}$, $k \geq 4$ using Lemmas 4.3 and 4.7. One can check that

$$\sum_{k=2}^{l} \mathbb{E}\widehat{L}_{k,2} \prec \sum_{r=1}^{2n} \left(\frac{N^{-\xi}}{(N\eta)^r} + \left(\frac{N^{-\delta/4}}{\sqrt{Nq}} \right)^r \right) \mathcal{M}^{2n-r}. \tag{4.39}$$

Step 4. Combining (4.38)–(4.39) we conclude the proof of (4.3).

5. Estimates of general polynomials of Green functions. In this section we prove Lemmas 4.3 and 4.6.

5.1. *Proof of Lemma 4.3.* To simplify notations, we shall prove the lemma for $\mathcal{Y} = \{G\}$, and one easily checks that the proof is the same for $\mathcal{Y} = \{G, G^*\}$. Let us take a general term $V \in \mathcal{V}^{(\nu_0,\nu_1)}(\{G\})$, and consider

$$\begin{aligned} &\sum_{i_1,\dots,i_{\nu_0}} \mathbb{E} V_{i_1,\dots,i_{\nu_0}} \\ &\quad = \frac{1}{N^{\nu_1}} \sum_{i_1,\dots,i_{\nu_0}} a_{i_1,\dots,i_{\nu_0}} \mathbb{E} G_{x_1 y_1} \cdots G_{x_k y_k} (G^2)_{z_1 w_1} \cdots (G^2)_{z_{\nu_2} w_{\nu_2}} [G]^{\nu_4}, \end{aligned} \tag{5.1}$$

where $x_1, y_1, \dots, x_k, y_k, z_1, w_1, \dots, z_{\nu_2}, w_{\nu_2} \in \{i_1, \dots, i_{\nu_0}\}$, and $a_{i_1,\dots,i_{\nu_0}}$ are complex numbers uniformly bounded in $i_1, \dots, i_{\nu_0}$. We break the proof into four steps.

Step 1. Since $\nu_2 \geq 1$, we can use resolvent identity $z(G^2)_{z_1 w_1} = (HG^2)_{z_1 w_1} - G_{z_1 w_1}$ and Lemma 2.2 to get

$$\begin{aligned} \mathbb{E} V_{i_1,\dots,i_{\nu_0}} &= \frac{1}{N^{\nu_1}} a_{i_1,\dots,i_{\nu_0}} \mathbb{E} G_{x_1 y_1} \cdots G_{x_k y_k} (G^2)_{z_1 w_1} \cdots (G^2)_{z_{\nu_2} w_{\nu_2}} [G]^{\nu_4} \\ &= \frac{1}{-z - \mathbb{E}\underline{G}} \mathbb{E} \Big[V_{i_1,\dots,i_{\nu_0}} / (G^2)_{z_1 w_1} \cdot \big(G_{z_1 w_1} \underline{G^2} + (G^2)_{z_1 w_1} \langle \underline{G} \rangle + 2N^{-1} (G^3)_{z_1 w_1} \big) \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{N}\sum_{m=1}^{k} V_{i_1,\dots,i_{\nu_0}}/\big((G^2)_{z_1w_1}G_{x_my_m}\big)\cdot\big((G^3)_{x_mw_1}G_{z_1y_m}+G_{x_mz_1}(G^3)_{w_1y_m}\big)\\
&+\frac{1}{N}\sum_{m=2}^{\nu_2} V_{i_1,\dots,i_{\nu_0}}/\big((G^2)_{z_1w_1}(G^2)_{z_mw_m}\big)\\
&\times\cdot\big((G^4)_{z_mw_1}G_{z_1w_m}+(G^2)_{z_mz_1}(G^3)_{w_1w_m}\\
&+(G^4)_{w_1w_m}G_{z_1z_m}+(G^2)_{z_1w_m}(G^3)_{z_mw_1}\big)\\
&+\frac{2\nu_4}{N^2}V_{i_1,\dots,i_{\nu_0}}/\big((G^2)_{z_1w_1}[G]\big)\cdot\big((G^4)_{z_1w_1}+2(HG^2)_{z_1w_1}mm'\big)\\
&+V_{i_1,\dots,i_{\nu_0}}/(G^2)_{z_1w_1}\cdot G_{z_1w_1}-K^{(1)}-\sum_i\sum_{k=2}^{l}L_k^{(1,i)}-\sum_i\mathcal{R}_{l+1}^{(1,z_1i)}\Bigg],
\end{aligned}
\tag{5.2}
$$

where

$$
K^{(1)}=\frac{a_{i_1,\dots,i_{\nu_0}}}{N^{1+\nu_1}}\frac{\partial(G_{x_1y_1}\cdots G_{x_ky_k}(G^2)_{z_1w_1}\cdots(G^2)_{z_{\nu_2}w_{\nu_2}}[G]^{\nu_4})}{\partial H_{z_1z_1}}\big(N\mathcal{C}_2(H_{z_1z_1})-2\big),
$$

and

$$
L_k^{(1,i)}=\frac{a_{i_1,\dots,i_{\nu_0}}}{N^{\nu_1}}\frac{1}{k!}\mathcal{C}_{k+1}(H_{iz_1})\frac{\partial^k(G_{x_1y_1}\cdots G_{x_ky_k}(G^2)_{iw_1}(G^2)_{z_2w_2}\cdots(G^2)_{z_{\nu_2}w_{\nu_2}}[G]^{\nu_4})}{\partial H_{z_1i}^k}.
$$

Here l is a fixed positive integer to be chosen later, and $\mathcal{R}_{l+1}^{(1,z_1i)}$ is a remainder term defined analogously to $\mathcal{R}_{l+1}$ in (2.4). Again by a routine verification, the remainder term is negligible for large l. Note that by Theorem 2.6 we have

$$
\frac{1}{-z-\mathbb{E}\underline{G}}=O(1)
$$

uniformly for $z\in\mathbf{D}_\tau$. Also note that

$$
\langle\underline{G}\rangle=[G]+O_\prec\Big(\frac{1}{\sqrt{Nq}}\Big)=[G]+O_\prec\Big(\eta+\frac{1}{N\eta}\Big).
\tag{5.3}
$$

Inserting (5.2) into (5.1), and by using (4.17), (5.3) and Lemma 4.2, we have

$$
\begin{aligned}
&\sum_{i_1,\dots,i_{\nu_0}}\mathbb{E}V_{i_1,\dots,i_{\nu_0}}\\
&=\sum_{i_1,\dots,i_{\nu_0}}\frac{1}{-z-\mathbb{E}\underline{G}}\mathbb{E}\Bigg[V_{i_1,\dots,i_{\nu_0}}/(G^2)_{z_1w_1}\cdot G_{z_1w_1}\underline{G^2}-\sum_i\sum_{k=2}^{l}L_k^{(1,i)}\Bigg]\\
&\quad+O_\prec(\mathcal{B}(V)).
\end{aligned}
\tag{5.4}
$$

Step 2. Now let us look closely at $L_k^{(1,i)}$. When none of the derivatives $\partial H_{z_1i}^k$ hit $[G]^{\nu_4}$, (2.1) shows that all the resulting terms are still in $\mathcal{V}$. When at least one derivative hits $[G]^{\nu_4}$, we expand the factors $[G]$ that were differentiated, and split the terms according to whether $(\underline{H^2}-1)$ is differentiated or not. For example, when $k=2$, let us take the term

$$
\frac{a_{i_1,\dots,i_{\nu_0}}}{N^{\nu_1}}\frac{1}{2!}\mathcal{C}_{2+1}(H_{iz_1})G_{x_1y_1}\cdots G_{x_ky_k}(G^2)_{iw_1}(G^2)_{z_2w_2}\cdots
$$

$$\times (G^2)_{z_{\nu_2} w_{\nu_2}} [G]^{\nu_4 - 2} \nu_4 (\nu_4 - 1) \Big(\frac{\partial [G]}{\partial H_{z_1 i}} \Big)^2 \tag{5.5}$$

$$=: X \Big(\frac{\partial [G]}{\partial H_{z_1 i}} \Big)^2$$

from $L_2^{(1,i)}$. Since

$$\frac{\partial [G]}{\partial H_{z_1 i}} = -2N^{-1} (G^2)_{z_1 i} - 4N^{-1} m m' H_{z_1 i},$$

we can split (5.5) into

$$X(-2N^{-1}(G^2)_{z_1 i})^2 + X(8N^{-2} m m' H_{z_1 i} (G^2)_{z_1 i} + 16 N^{-2} (m m' H_{z_1 i})^2), \tag{5.6}$$

and note that the first term in (5.6) is in $\mathcal{V}(\{G\})$, and the second term in (5.6) contains at least one derivative of $\underline{H^2} - 1$. In this way, we split

$$L_k^{(1,i)} = L_k^{(1,i,1)} + L_k^{(1,i,2)}, \tag{5.7}$$

where $L_k^{(1,i,1)}$ are all the terms in $L_k^{(1,i)}$ that do not contain the derivatives of $(\underline{H^2} - 1)$. By the above reasoning, we see that $L_k^{(1,i,1)}$ is a finite linear combination of elements in $\mathcal{V}$. Also observe from (2.2) that when the derivatives hit $(\underline{H^2} - 1)$, it gives us something small.

By Lemma 4.2, (2.2) and (5.4), one readily checks that

$$\sum_{i_1, \dots, i_{\nu_0}} \mathbb{E} V_{i_1, \dots, i_{\nu_0}}$$

$$= \sum_{i_1, \dots, i_{\nu_0}} \frac{1}{-z - \mathbb{E}\underline{G}} \mathbb{E} \Big[V_{i_1, \dots, i_{\nu_0}} / (G^2)_{z_1 w_1} \cdot G_{z_1 w_1} \underline{G^2} - \sum_i \sum_{k=2}^{l} L_k^{(1,i,1)} \Big] \tag{5.8}$$

$$+ O_{\prec}(\Psi(V)).$$

Step 3. Now let us handle the first term on RHS of (5.8). Define

$$V^{(1)}_{i_1, \dots i_n} := V_{i_1, \dots, i_{\nu_0}} / (G^2)_{z_1 w_1} \cdot G_{z_1 w_1}$$

$$= \frac{a_{i_1, \dots, i_{\nu_0}}}{N^{\nu_1}} G_{x_1 y_1} \cdots G_{x_k y_k} (G^2)_{z_2 w_2} \cdots (G^2)_{z_{\nu_2} w_{\nu_2}} [G]^{\nu_4} G_{z_1 w_1},$$

and we look at

$$\mathbb{E} V^{(1)}_{i_1, \dots i_n} \underline{G^2} = \frac{a_{i_1, \dots, i_{\nu_0}}}{N^{\nu_1}} \mathbb{E} G_{x_1 y_1} \cdots G_{x_k y_k} (G^2)_{z_2 w_2} \cdots (G^2)_{z_{\nu_2} w_{\nu_2}} [G]^{\nu_4} G_{z_1 w_1} \underline{G^2}. \tag{5.9}$$

Similarly as in (5.2), we use $z\underline{G^2} = \underline{HG^2} - \underline{G}$ and Lemma 2.2 to expand (5.9). We get

$$\mathbb{E} V^{(1)}_{i_1, \dots i_n} \underline{G^2} = \frac{1}{T} \mathbb{E} \Big[V^{(1)}_{i_1, \dots, i_{\nu_0}} \cdot (2\underline{G^2} \langle \underline{G} \rangle + 2N^{-1} \underline{G^3}) + \frac{2}{N^2} \sum_{m=1}^{k} V^{(1)}_{i_1, \dots, i_{\nu_0}} / G_{x_m y_m} \cdot (G^4)_{x_m y_m}$$

$$+ \frac{4}{N^2} \sum_{m=2}^{\nu_2} V^{(1)}_{i_1, \dots, i_{\nu_0}} / (G^2)_{z_m w_m} \cdot (G^5)_{z_m w_m}$$

$$+ \frac{2\nu_4}{N^2} V^{(1)}_{i_1, \dots, i_{\nu_0}} / [G] \cdot (\underline{G^4} + 2\underline{HG^2} m m') \tag{5.10}$$

$$+ V^{(1)}_{i_1, \dots, i_{\nu_0}} \cdot \underline{G} + \frac{2}{N^2} V^{(1)}_{i_1, \dots, i_{\nu_0}} / G_{z_1 m_1} (G^4)_{z_m y_m} - K^{(2)} - \sum_{i,j} \sum_{k=2}^{l} L_k^{(2,ji)}$$

$$-\frac{1}{N}\sum_{i,j}\mathcal{R}_{l+1}^{(2,ji)}\Bigg],$$

where $\mathcal{R}_{l+1}^{(2,ji)}$ is the remainder term,

$$K^{(2)}=\frac{a_{i_1,\dots,i_{\nu_0}}}{N^{2+\nu_1}}\sum_i\frac{\partial(G_{x_1y_1}\cdots G_{x_ky_k}(G^2)_{z_2w_2}\cdots(G^2)_{z_{\nu_2}w_{\nu_2}}[G]^{\nu_4}G_{z_1w_1}(G^2)_{ii})}{\partial H_{ii}}$$
$$\times\big(N\mathcal{C}_2(H_{ii})-2\big),$$

and

$$L_k^{(2,ji)}$$
$$=\frac{a_{i_1,\dots,i_{\nu_0}}}{N^{1+\nu_1}}\frac{1}{k!}\mathcal{C}_{k+1}(H_{iz_1})\frac{\partial^k(G_{x_1y_1}\cdots G_{x_ky_k}(G^2)_{z_2w_2}\cdots(G^2)_{z_{\nu_2}w_{\nu_2}}[G]^{\nu_4}G_{z_1w_1}(G^2)_{ji})}{\partial H_{ji}^k}.$$

Recall that $T=-z-2\mathbb{E}\underline{G}$ satisfies (4.11). Similarly as in (5.7), we can split

$$L_k^{(2,ji)}=L_k^{(2,ji,1)}+L_k^{(2,ji,2)},$$

where $L_k^{(2,ji,1)}$ is a finite linear combination of elements in $\mathcal{V}$. By inserting (5.10) into (5.8), applying (2.2) and Lemma 4.2, we have

$$\begin{aligned}\sum_{i_1,\dots,i_{\nu_0}}\mathbb{E}V_{i_1,\dots,i_{\nu_0}}&=\sum_{k=2}^{l}\sum_{i_1,\dots,i_{\nu_0},i,j}\frac{1}{(z+\mathbb{E}\underline{G})T}\mathbb{E}L_k^{(2,ji,1)}\\&\quad+\sum_{k=2}^{l}\sum_{i_1,\dots,i_{\nu_0},i}\frac{1}{z+\mathbb{E}\underline{G}}\mathbb{E}L_k^{(1,i,1)}+O_{\prec}\big(\mathcal{B}(V)\big).\end{aligned}\tag{5.11}$$

Step 4. Now let us see how to further (recursively) expand (5.11) and why the expansion ends in finitely many steps. Let V be as in (5.1). For any $V_*\in\mathcal{V}$ satisfying $\nu_4(V_*)\le\nu_4(V)$, let us define the ratio

$$I(V_*):=\frac{N^{b_0(V_*)}}{(N\eta)^{-\nu_4(V)+\nu_4(V_*)}N^{b_0(V)-1}}.$$

From Lemma 4.2 we know that

$$\sum_{i_1,\dots,i_{\nu_0(V_*)}}\mathbb{E}V_*\prec N^{b_0(V_*)}\mathcal{M}^{\nu_4(V_*)}\le I(V_*)\mathcal{B}(V).\tag{5.12}$$

By construction, $L_k^{(1,i,1)}$ and $L_k^{(2,ji,1)}$ are finite linear combinations of the elements in $\mathcal{V}$. Let us collect these elements in the set $\mathcal{V}^{(1)}$. Pick arbitrary $V^{(1)}\in\mathcal{V}^{(1)}$. We see that $\nu_2(V^{(1)})\ge1$, $\nu_3(V^{(1)})\ge\nu_3(V)$, and $\nu_4(V^{(1)})\le\nu_4(V)$. By Lemmas 2.3 and 4.2 one readily check that

$$I\big(V^{(1)}\big)\le I(V)\cdot(N\eta)^{(\nu_3(V)-\nu_3(V^{(1)}))/2}N^{-\beta}=I(V)\cdot(N\eta)^{(\widetilde{\nu}_3(V^{(1)})-\widetilde{\nu}_3(V))/2}N^{-\beta},\tag{5.13}$$

which together with (5.12) implies

$$\sum_{i_1,\dots,i_{\nu_{0,1}}}\mathbb{E}V^{(1)}\prec I(V)(N\eta)^{(\widetilde{\nu}_3(V^{(1)})-\widetilde{\nu}_3(V))/2}N^{-\beta}\cdot\mathcal{B}(V).$$

Here we abbreviate $\nu_{0,1} := \nu_0(V^{(1)})$. Repeat (5.11) we have

$$\sum_{i_1,\dots,i_{\nu_{0,1}}} \mathbb{E}V^{(1)}_{i_1,\dots,i_{\nu_{0,1}}} = \sum_{k=2}^{l} \sum_{i_1,\dots,i_{\nu_{0,1}},i,j} \frac{1}{(z+\mathbb{E}\underline{G})T} \mathbb{E}\widetilde{L}_k^{(2,ji,1)} + \sum_{k=2}^{l} \sum_{i_1,\dots,i_{\nu_{0,1}},i} \frac{1}{z+\mathbb{E}\underline{G}} \mathbb{E}\widetilde{L}_k^{(1,i,1)} + O_{\prec}\big(\mathcal{B}(V^{(1)})\big).$$

Note that (5.13) implies

$$\begin{aligned}\mathcal{B}\big(V^{(1)}\big) &:= N^{b_0(V^{(1)})}\mathcal{M}^{\nu_4(V^{(1)})+1} \\ &\quad + \sum_{k=0}^{\nu_4(V^{(1)})} N^{b_0(V^{(1)})} \frac{1}{(N\eta)^k}\bigg(\eta + \frac{1}{(N\eta)^{\widetilde{\nu}_3(V^{(1)})/2}}\bigg)\cdot \mathcal{M}^{\nu_4(V^{(1)})-k} \\ &\le N^{-\beta}(N\eta)^{(\widetilde{\nu}_3(V^{(1)})-\widetilde{\nu}_3(V))/2-\nu_4(V)+\nu_4(V^{(1)})}\bigg(N^{b_0(V)}\mathcal{M}^{\nu_4(V^{(1)})+1} \\ &\quad + \sum_{k=0}^{\nu_4(V^{(1)})} N^{b_0(V)} \frac{1}{(N\eta)^k}\bigg(\eta + \frac{1}{(N\eta)^{\widetilde{\nu}_3(V^{(1)})/2}}\bigg)\cdot \mathcal{M}^{\nu_4(V^{(1)})-k}\bigg) \le N^{-\beta}\mathcal{B}(V).\end{aligned}$$

By construction, $\widetilde{L}_k^{(1,i,1)}$ and $\widetilde{L}_k^{(2,ji,1)}$ are finite linear combinations of the elements in $\mathcal{V}$. Let us collect these elements in the set $\mathcal{V}^{(2)}$. Pick arbitrary $V^{(2)} \in \mathcal{V}^{(2)}$. We see that $\nu_2(V^{(2)}) \ge 1$, $\nu_3(V^{(2)}) \ge \nu_3(V^{(1)}) \ge \nu_3(V)$ and $\nu_4(V^{(2)}) \le \nu_4(V^{(1)}) \le \nu_4(V)$. By Lemmas 2.3, 4.2 and (5.13) we have

$$I\big(V^{(2)}\big) \le I\big(V^{(1)}\big)\cdot(N\eta)^{(\widetilde{\nu}_3(V^{(2)})-\widetilde{\nu}_3(V^{(1)}))/2}N^{-\beta} \le I(V)\cdot(N\eta)^{(\widetilde{\nu}_3(V^{(2)})-\widetilde{\nu}_3(V))/2}N^{-2\beta},$$

which together with (5.12) implies

$$\sum_{i_1,\dots,i_{\nu_{0,2}}} \mathbb{E}V^{(2)} \prec I(V)(N\eta)^{(\widetilde{\nu}_3(V^{(2)})-\widetilde{\nu}_3(V))/2}N^{-2\beta}\cdot\mathcal{B}(V).$$

Here we abbreviate $\nu_{0,2} := \nu_0(V^{(2)})$. Repeating the above steps we get

$$\sum_{i_1,\dots,i_{\nu_0}} \mathbb{E}V_{i_1,\dots,i_{\nu_0}} = \sum_{V^{(n)}\in\mathcal{V}^{(n)}} \sum_{i_1,\dots,i_{\nu_0(V^{(n)})}} \mathbb{E}V^{(n)} + O_{\prec}\big(\mathcal{B}(V)\big)$$

for any $n \in \mathbb{N}$. Here $|\mathcal{V}^{(n)}| \le C_n$, and each $V^{(n)} \in \mathcal{V}^{(n)}$ satisfies

$$\sum_{i_1,\dots,i_{\nu_{0,n}}} \mathbb{E}V^{(n)} \prec I(V)(N\eta)^{(\widetilde{\nu}_3(V^{(n)})-\widetilde{\nu}_3(V))/2}N^{-n\beta}\cdot\mathcal{B}(V).$$

Since $\beta > 0$, $\widetilde{\nu}_3 \in \{0,1,2\}$ and $I(V) = N$, setting n large enough we get

$$\sum_{i_1,\dots,i_{\nu_0}} \mathbb{E}V_{i_1,\dots,i_{\nu_0}} = O_{\prec}\big(\mathcal{B}(V)\big)$$

as desired.

5.2. *Proof of Lemma 4.6.* Again, to simplify notations, we shall prove the lemma for $\mathcal{Y}=\{G\}$, and one easily checks that the proof is the same for $\mathcal{Y}=\{G,G^*\}$. Let us take a general term $W\in\mathcal{W}^{(\nu_0,\nu_1)}(\{G\})$, and consider

$$\sum_{i_1,\dots,i_{\nu_0}}\mathbb{E}W_{i_1,\dots,i_{\nu_0}}=\frac{1}{N^{\nu_1}}\sum_{i_1,\dots,i_{\nu_0}}a_{i_1,\dots,i_{\nu_0}}\mathbb{E}\langle G_{x_1y_1}\cdots G_{x_ky_k}\rangle[G]^{\nu_4}, \tag{5.14}$$

where $x_1,y_1,\dots,x_k,y_k\in\{i_1,\dots,i_{\nu_0}\}$, and $a_{i_1,\dots,i_{\nu_0}}$ are complex numbers uniformly bounded in $i_1,\dots,i_{\nu_0}$. Resolvent identity $zG_{x_1y_1}=(HG)_{x_1y_1}-I_{x_1y_1}$ and $\mathbb{E}\langle X\rangle Y=\mathbb{E}X\langle Y\rangle$ gives

$$\begin{aligned}&z\mathbb{E}\langle G_{x_1y_1}\cdots G_{x_ky_k}\rangle[G]^{\nu_4}\\&\quad=\mathbb{E}(HG)_{x_1y_1}G_{x_2y_2}\cdots G_{x_ky_k}\langle[G]^{\nu_4}\rangle-\mathbb{E}\delta_{x_1y_1}G_{x_2y_2}\cdots G_{x_ky_k}\langle[G]^{\nu_4}\rangle.\end{aligned}$$

Similarly as in the proof of Lemma 4.3, we apply Lemma 2.2 to the above and get

$$\begin{aligned}\mathbb{E}W_{i_1,\dots,i_{\nu_0}}&=\frac{a_{i_1,\dots,i_{\nu_0}}}{N^{\nu_1}}\mathbb{E}\langle G_{x_1y_1}\cdots G_{x_ky_k}\rangle[G]^{\nu_4}\\&=\frac{a_{i_1,\dots,i_{\nu_0}}}{-z-\mathbb{E}\underline{G}}\mathbb{E}\bigg[\frac{1}{N^{\nu_1}}\langle\delta_{x_1y_1}G_{x_2y_2}\cdots G_{x_ky_k}\rangle[G]^{\nu_4}\\&\quad+\frac{1}{N^{\nu_1}}\big\langle\big(\langle\underline{G}\rangle G_{x_1y_1}+N^{-1}\big(G^2\big)_{x_1y_1}\big)G_{x_2y_2}\cdots G_{x_ky_k}\big\rangle[G]^{\nu_4}\\&\quad+\frac{1}{N^{1+\nu_1}}\sum_{m=2}^{k}\big\langle\big(\big(G^2\big)_{x_1y_m}G_{y_1x_m}+\big(G^2\big)_{x_1x_m}G_{y_1y_m}\big)\\&\quad\times G_{x_2y_2}\cdots G_{x_{m-1}y_{m-1}}G_{x_{m+1}y_{m+1}}\cdots G_{x_ky_k}\big\rangle[G]^{\nu_4}\\&\quad+\frac{2\nu_4}{N^{2+\nu_1}}G_{x_2y_2}\cdots G_{x_ky_k}\big(\big(G^3\big)_{y_1x_1}+2(HG)_{y_1x_1}mm'\big)[G]^{\nu_4-1}-K^{(3)}\\&\quad-\sum_i\sum_{k=2}^{l}L_k^{(3,i)}-\sum_i\mathcal{R}_{l+1}^{(3,ix_1)}\bigg],\end{aligned} \tag{5.15}$$

where

$$K^{(3)}=\frac{1}{N^{1+\nu_1}}\mathbb{E}\frac{\partial(G_{x_1y_1}\cdots G_{x_ky_k}\langle[G]^{\nu_4}\rangle)}{\partial H_{x_1x_1}}(N\mathcal{C}_2(H_{x_1x_1})-2),$$

and

$$L_k^{(3,i)}=\frac{1}{N^{\nu_1}}\frac{1}{k!}\mathcal{C}_{k+1}(H_{ix_1})\mathbb{E}\frac{\partial^k((G_{iy_1}\cdots G_{x_ky_k}\langle[G]^{\nu_4}\rangle)}{\partial H_{x_1i}^k}.$$

Now we insert (5.15) into (5.14), and by using (4.17), (5.3) and Lemma 4.2, we have

$$\begin{aligned}&\sum_{i_1,\dots,i_{\nu_0}}\mathbb{E}W_{i_1,\dots,i_{\nu_0}}\\&\quad=\sum_{i_1,\dots,i_{\nu_0}}a_{i_1,\dots,i_{\nu_0}}\frac{1}{-z-\mathbb{E}\underline{G}}\mathbb{E}\bigg[\frac{1}{N^{\nu_1}}\langle\delta_{x_1y_1}G_{x_2y_2}\cdots G_{x_ky_k}\rangle[G]^{\nu_4}-\sum_i\sum_{k=2}^{l}L_k^{(3,i)}\bigg]\\&\qquad+O_{\prec}(\mathcal{B}^{(1)}(W)).\end{aligned}$$

Similarly as in (5.7), we split

$$L_k^{(3,i)}=L_k^{(3,i,1)}+L_k^{(3,i,2)},$$

where $L_k^{(3,i,1)}$ are all the terms in $L_k^{(3,i)}$ that do not contain the derivatives of $(\underline{H^2}-1)$. By (2.2) and Lemma 4.2, we see that the terms associated with $L_k^{(3,i,2)}$ are negligible, thus

$$\begin{aligned}&\sum_{i_1,\dots,i_{\nu_0}} \mathbb{E}W_{i_1,\dots,i_{\nu_0}}\\ &= \sum_{i_1,\dots,i_{\nu_0}} a_{i_1,\dots,i_{\nu_0}} \frac{1}{-z-\mathbb{E}\underline{G}}\mathbb{E}\left[\frac{1}{N^{\nu_1}}\langle \delta_{x_1y_1}G_{x_2y_2}\cdots G_{x_ky_k}\rangle [G]^{\nu_4} - \sum_i \sum_{k=2}^{l} L_k^{(3,i,1)}\right]\\ &\quad + O_{\prec}\big(\mathcal{B}^{(1)}(W)\big).\end{aligned} \tag{5.16}$$

Now let us look at the terms in (5.16) carefully. For any $W_* \in \mathcal{W}$ satisfying $\nu_4(W_*) = \nu_4(W)$, we define the ratio

$$I^{(1)}(W_*) := \frac{N^{b_1(W_*)}}{N^{b_1(W)-1}},$$

where W is defined as in (5.14). By Lemma 4.5, $\sum \mathbb{E}W_* \prec I^{(1)}(W_*)\mathcal{B}^{(1)}(W)$. For $L_k^{(3,i,1)}$, we can apply all the differentials $\partial H^k_{x_1 i}$ and write $L_k^{(3,i,1)}$ in the form of linear combinations

$$L_k^{(3,i,1)} = \sum_{s=1}^{n} c_s V_s + \sum_{t=1}^{n'} c'_t W_t,$$

where $V_s \in \mathcal{V}$, $W_t \in \mathcal{W}$, and c_s, c'_t are constants. Each V_s is formed by requiring at least one differential hit $[G]^{\nu_4}$, and we can use Lemma 4.3 to show that

$$\sum_{i_1,\dots,i_{\nu_0},i} a_{i_1,\dots,i_{\nu_0}} \frac{1}{-z-\mathbb{E}\underline{G}}\mathbb{E}(-c_sV_s) = O_{\prec}\big(\mathcal{B}^{(1)}(W)\big).$$

Each W_t is formed by requiring none of the differentials hit $[G]^{\nu_4}$, thus $\nu_4(W_t) = \nu_4(W)$. By $\mathcal{C}_{k+1}(H_{x_1i}) \le C/(Nq)$ we find that

$$I^{(1)}(W_t) \le I^{(1)}(W)\cdot (N\eta)^{(\tilde{\nu}_3(W_t))-\tilde{\nu}_3(W))/2} N^{-\beta},$$

which implies

$$\sum_{i_1,\dots,i_{\nu_0},i} a_{i_1,\dots,i_{\nu_0}} \frac{1}{-z-\mathbb{E}\underline{G}}\mathbb{E}\big(-c'_tW_t\big) \prec I^{(1)}(W)(N\eta)^{(\tilde{\nu}_3(W_t))-\tilde{\nu}_3(W))/2} N^{-\beta}\mathcal{B}^{(1)}(W). \tag{5.17}$$

As for the first term on RHS of (5.16), one easily sees that it is small enough when $x_1 \not\equiv y_1$. When $x_i \equiv y_i$ for all $i = 1,\dots,k$, we rewrite (5.16) into

$$\begin{aligned}&\sum_{i_1,\dots,i_{\nu_0}} \mathbb{E}W_{i_1,\dots,i_{\nu_0}}\\ &= \sum_{i_1,\dots,i_{\nu_0}} \mathbb{E}W' + \sum_{i_1,\dots,i_{\nu_0}} a_{i_1,\dots,i_{\nu_0}} \frac{1}{-z-\mathbb{E}\underline{G}}\mathbb{E}\left[-\sum_i \sum_{t=1}^{n'} c'_tW_t\right] + O_{\prec}\big(\mathcal{B}^{(1)}(W)\big),\end{aligned} \tag{5.18}$$

where

$$W' = a_{i_1,\dots,i_{\nu_0}} \frac{1}{-z-\mathbb{E}\underline{G}}\frac{1}{N^{\nu_1}}\langle G_{x_2y_2}\cdots G_{x_ky_k}\rangle [G]^{\nu_4},$$

and each W_t in (5.18) satisfies (5.17). Note that $I^{(1)}(W) = I^{(1)}(W')$, and $\nu_i(W) = \nu_i(W')$ for $i = 0, 1, 3, 4$, thus we can repeat (5.18) with W replaced by W'. By doing this k times, we have

$$\begin{aligned}
&\sum_{i_1,\dots,i_{\nu_0}} \mathbb{E}W_{i_1,\dots,i_{\nu_0}} \\
&\quad = \sum_{i_1,\dots,i_{\nu_0}} \mathbb{E}W^{(k)} + \sum_{i_1,\dots,i_{\nu_0}} a_{i_1,\dots,i_{\nu_0}} \frac{1}{-z - \mathbb{E}\underline{G}} \mathbb{E}\bigg[-\sum_i \sum_{t=1}^{n^{(k)}} c_t' W_t \bigg] + O_{\prec}\big(\mathcal{B}^{(1)}(W)\big) \\
&\quad = \sum_{i_1,\dots,i_{\nu_0}} a_{i_1,\dots,i_{\nu_0}} \frac{1}{-z - \mathbb{E}\underline{G}} \mathbb{E}\bigg[-\sum_i \sum_{t=1}^{n^{(k)}} c_t' W_t \bigg] + O_{\prec}\big(\mathcal{B}^{(1)}(W)\big),
\end{aligned}$$

where we used

$$W^{(k)} = a_{i_1,\dots,i_{\nu_0}} \frac{1}{-z - \mathbb{E}\underline{G}} \frac{1}{N^{\nu_1}} \langle 1 \rangle [G]^{\nu_4} = 0.$$

To sum up, we have

$$\sum_{i_1,\dots,i_{\nu_0}} \mathbb{E}W_{i_1,\dots,i_{\nu_0}} = \sum_{i_1,\dots,i_{\nu_0}} a_{i_1,\dots,i_{\nu_0}} \frac{1}{-z - \mathbb{E}\underline{G}} \mathbb{E}\bigg[-\sum_i \sum_{t=1}^{n'} c_t' W_t \bigg] + O_{\prec}\big(\mathcal{B}^{(1)}(W)\big) \tag{5.19}$$

for some $n' \in \mathbb{N}$, where each W_t in (5.19) satisfies (5.17). In addition, note that $I^{(1)}(W) = N$, $\widetilde{\nu}_3 \in \{0, 1, 2\}$, and $\beta > 0$.

The above argument shows that, similarly as in Section 5.1, we can repeatably use (5.19) finitely many times, and eventually get

$$\sum_{i_1,\dots,i_{\nu_0}} a_{i_1,\dots,i_{\nu_0}} \mathbb{E}W_{i_1,\dots,i_{\nu_0}} = O_{\prec}\big(\mathcal{B}^{(1)}(W)\big)$$

as desired.

APPENDIX: PROOF OF LEMMA 2.10

(i) From Theorem 2.6 we have

$$\big|\langle \underline{G} \rangle\big| \le |\underline{G} - m| + |\mathbb{E}\underline{G} - m| \prec \frac{1}{q} + \frac{1}{N\eta}$$

uniformly for $z = E + \mathrm{i}\eta \in \mathbf{S}$, and this proves (2.5) for the case $m + n = 1$. In addition, by an N^{-3}-net argument and a deterministic monotonivity result [2], Remark 2.7, Lemma 10.2, we have

$$\sup_{z \in \mathbf{S}} \big|\langle \underline{G}(z) \rangle\big| \bigg(\frac{1}{q} + \frac{1}{N\eta} \bigg)^{-1} \prec 1. \tag{A.1}$$

For $m + n \ge 2$, note that

$$G^m G^{*n} = \eta^{-(m+n)} f\bigg(\frac{H - E}{\eta} \bigg),$$

where

$$f(x) = \bigg(\frac{x + \mathrm{i}}{x^2 + 1} \bigg)^m \bigg(\frac{x - \mathrm{i}}{x^2 + 1} \bigg)^n.$$

Writing $f_\eta(x) = f(\frac{x-E}{\eta})$ and applying Lemma 2.4, we have

$$\begin{aligned}\langle \underline{G^m G^{*n}} \rangle &= \frac{1}{\pi \eta^{m+n}} \int_{\mathbb{C}} \partial_{\bar{z}} \big(\tilde{f}_\eta(z)\chi(y/\eta)\big) \langle \underline{G}(z) \rangle \, \mathrm{d}^2 z \\ &= \frac{1}{2\pi \eta^{m+n}} \int_{\mathbb{R}^2} \bigg(\mathrm{i} y f_\eta''(x)\chi(y/\eta) + \frac{\mathrm{i}}{\eta} f_\eta(x)\chi'(y/\eta) - \frac{y}{\eta} f_\eta'(x)\chi'(y/\eta) \bigg) \\ &\quad \times \langle \underline{G}(x+\mathrm{i}y) \rangle \, \mathrm{d}x \, \mathrm{d}y, \end{aligned} \tag{A.2}$$

where we set $\chi(y) = 1$ for $|y| \le 1$ and $\chi(y) = 0$ for $|y| \ge 2$. Note that for $m + n \ge 2$, f and its derivatives are in $L^1(\mathbb{R})$. The proof of (2.5) then finishes by inserting (A.1) into (A.2).

The proof of (2.6) is similar. We omit the details.

(ii) The proof follows by

$$\sum_i \big| (G^m G^{*n})_{ij} \big|^2 = (G^n G^{*m} G^m G^{*n})_{jj}$$

and (2.6).

(iii) Let us first look at the case $k = 2$. By the resolvent identity $zG = HG - I$ and Lemma 2.2, we have

$$\mathbb{E}\underline{G^2} = \frac{1}{T} \bigg(\mathbb{E}\underline{G} + 2\mathbb{E}\langle \underline{G} \rangle \langle \underline{G^2} \rangle + \frac{1}{N} \mathbb{E}\underline{G^3} - \mathbb{E}K^{(4)} - \mathbb{E}L^{(4)} - \mathcal{R}_{l+1}^{(4)} \bigg), \tag{A.3}$$

where $T = -z - 2\mathbb{E}\underline{G}$, $\mathcal{R}_{l+1}^{(4)}$ is the remainder term,

$$K^{(4)} = N^{-2} \sum_i \frac{\partial (G^2)_{ii}}{\partial H_{ii}} (\mathbb{E}H_{ii}^2 - 2),$$

and

$$L^{(4)} = \frac{1}{N} \sum_{i,j} \bigg(\sum_{k=2}^{l} \frac{1}{k!} \mathcal{C}_{k+1}(H_{ji}) \frac{\partial^k (G^2)_{ij}}{\partial H_{ji}^K} \bigg).$$

The proof then follows by estimating the RHS of (A.3) by parts (i) and (ii).

The proof of the case $k = 3$ is similar, and we omit the details.

(iv) The proof is an elementary computation. One possible way is to write

$$\mathbb{E}(\underline{H^2} - 1)^n = \frac{1}{N} \sum_{i,j} \mathbb{E} \bigg(H_{ij}^2 - \frac{1}{N} \bigg) (\underline{H^2} - 1)^{n-1}$$

and apply Lemma 2.2 with $h = H_{ij}^2$. We omit the details.

Acknowledgments. The author would like to thank Gaultier Lambert for helpful discussions, and thank Antti Knowles and Benjamin Schlein for many useful comments on the preliminary draft.

This project received funding from NCCR Swissmap, the Swiss National Science Foundation (SNF) Grant No. 20020_172623 and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 715539_RandMat).

REFERENCES

[1] BENAYCH-GEORGES, F., GUIONNET, A. and MALE, C. (2014). Central limit theorems for linear statistics of heavy tailed random matrices. *Comm. Math. Phys.* **329** 641–686. MR3210147 https://doi.org/10.1007/s00220-014-1975-3

[2] BENAYCH-GEORGES, F. and KNOWLES, A. (2016). *Lectures on the local semicircle law for Wigner matrices*. Panoramas et Synthèses **53**.

[3] BOURGADE, P. and MODY, K. (2019). Gaussian fluctuations of the determinant of Wigner matrices. *Electron. J. Probab.* **24** Paper No. 96, 28. MR4017114 https://doi.org/10.1214/19-ejp356

[4] BOUTET DE MONVEL, A. and KHORUNZHY, A. (1999). Asymptotic distribution of smoothed eigenvalue density. I. Gaussian random matrices. *Random Oper. Stoch. Equ.* **7** 1–22. MR1678012 https://doi.org/10.1515/rose.1999.7.1.1

[5] DAVIES, E. B. (1995). The functional calculus. *J. Lond. Math. Soc.* (2) **52** 166–176. MR1345723 https://doi.org/10.1112/jlms/52.1.166

[6] ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2012). Spectral statistics of Erdős–Rényi Graphs II: Eigenvalue spacing and the extreme eigenvalues. *Comm. Math. Phys.* **314** 587–640. MR2964770 https://doi.org/10.1007/s00220-012-1527-7

[7] ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2013). Spectral statistics of Erdős–Rényi graphs I: Local semicircle law. *Ann. Probab.* **41** 2279–2375. MR3098073 https://doi.org/10.1214/11-AOP734

[8] ERDŐS, L., PÉCHÉ, S., RAMÍREZ, J. A., SCHLEIN, B. and YAU, H.-T. (2010). Bulk universality for Wigner matrices. *Comm. Pure Appl. Math.* **63** 895–925. MR2662426 https://doi.org/10.1002/cpa.20317

[9] ERDŐS, L., SCHLEIN, B. and YAU, H.-T. (2009). Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices. *Ann. Probab.* **37** 815–852. MR2537522 https://doi.org/10.1214/08-AOP421

[10] ERDŐS, L., SCHLEIN, B. and YAU, H.-T. (2010). Wegner estimate and level repulsion for Wigner random matrices. *Int. Math. Res. Not. IMRN* **3** 436–479. MR2587574 https://doi.org/10.1093/imrn/rnp136

[11] ERDŐS, L., SCHLEIN, B. and YAU, H.-T. (2011). Universality of random matrices and local relaxation flow. *Invent. Math.* **185** 75–119. MR2810797 https://doi.org/10.1007/s00222-010-0302-7

[12] GUSTAVSSON, J. (2005). Gaussian fluctuations of eigenvalues in the GUE. *Ann. Inst. Henri Poincaré Probab. Stat.* **41** 151–178. MR2124079 https://doi.org/10.1016/j.anihpb.2004.04.002

[13] HE, Y. and KNOWLES, A. (2017). Mesoscopic eigenvalue statistics of Wigner matrices. *Ann. Appl. Probab.* **27** 1510–1550. MR3678478 https://doi.org/10.1214/16-AAP1237

[14] HE, Y. and KNOWLES, A. (2020). Mesoscopic eigenvalue density correlations of Wigner matrices. *Probab. Theory Related Fields* **177** 147–216. MR4095015 https://doi.org/10.1007/s00440-019-00946-w

[15] HE, Y., KNOWLES, A. and MARCOZZI, M. (2019). Local law and complete eigenvector delocalization for supercritical Erdős–Rényi graphs. *Ann. Probab.* **47** 3278–3302. MR4021251 https://doi.org/10.1214/19-AOP1339

[16] HE, Y., KNOWLES, A. and ROSENTHAL, R. (2018). Isotropic self-consistent equations for mean-field random matrices. *Probab. Theory Related Fields* **171** 203–249. MR3800833 https://doi.org/10.1007/s00440-017-0776-y

[17] HUANG, J., LANDON, B. and YAU, H.-T. (2015). Bulk universality of sparse random matrices. *J. Math. Phys.* **56** 123301, 19. MR3429490 https://doi.org/10.1063/1.4936139

[18] HUANG, J., LANDON, B. and YAU, H.-T. (2020). Transition from Tracy–Widom to Gaussian fluctuations of extremal eigenvalues of sparse Erdős–Rényi graphs. *Ann. Probab.* **48** 916–962. MR4089498 https://doi.org/10.1214/19-AOP1378

[19] LANDON, B. and SOSOE, P. (2018). Applications of mesoscopic clts in random matrix theory. Preprint. Available at arXiv:1811.05915.

[20] LEE, J. O. and SCHNELLI, K. (2018). Local law and Tracy–Widom limit for sparse random matrices. *Probab. Theory Related Fields* **171** 543–616. MR3800840 https://doi.org/10.1007/s00440-017-0787-8

[21] LYTOVA, A. and PASTUR, L. (2009). Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *Ann. Probab.* **37** 1778–1840. MR2561434 https://doi.org/10.1214/09-AOP452

[22] O'ROURKE, S. (2010). Gaussian fluctuations of eigenvalues in Wigner random matrices. *J. Stat. Phys.* **138** 1045–1066. MR2601422 https://doi.org/10.1007/s10955-009-9906-y

[23] SHCHERBINA, M. and TIROZZI, B. (2012). Central limit theorem for fluctuations of linear eigenvalue statistics of large random graphs: Diluted regime. *J. Math. Phys.* **53** 043501, 18. MR2953145 https://doi.org/10.1063/1.3698291

[24] TAO, T. and VU, V. (2010). Random matrices: Universality of local eigenvalue statistics up to the edge. *Comm. Math. Phys.* **298** 549–572. MR2669449 https://doi.org/10.1007/s00220-010-1044-5

[25] TAO, T. and VU, V. (2011). Random matrices: Universality of local eigenvalue statistics. *Acta Math.* **206** 127–204. MR2784665 https://doi.org/10.1007/s11511-011-0061-3