Eigenstate thermalization hypothesis for generalized Wigner matrices

Arka Adhikari[*]   Sofiia Dubova[†]   Changji Xu [‡]   Jun Yin [§]

Abstract

In this paper, we extend results of Eigenvector Thermalization to the case of generalized Wigner matrices. Analytically, the central quantity of interest here are multiresolvent traces, such as $\Lambda_A := \frac{1}{N}\text{Tr}\, GAGA$. In the case of Wigner matrices, as in [14], one can form a self-consistent equation for a single Λ_A. There are multiple difficulties extending this logic to the case of general covariances. The correlation structure does not naturally lead to deriving a closed equation for Λ_A; this is due to the introduction of new terms that are quite distinct from the form of Λ_A. We find a way around this by carefully splitting these new terms and writing them as sums of Λ_B, for matrices B obtained by modifying A using the covariance matrix. The result is a system of inequalities relating families of deterministic matrices. Our main effort in this work is to derive this system of inequalities.

Keywords: random matrices; Eigenstate thermalization; generalized Wigner .
MSC2020 subject classifications: 60B20; 58J51.
Submitted to EJP on August 12, 2023, final version accepted on August 17, 2024.

1 Introduction

1.1 Background and history

Ever since Wigner proposed the study of random matrices in 1960 [36] in order to understand the energy spectra of heavy atoms, there has been significant effort in trying to understand the behavior of the eigenvalue fluctuations of random matrices. Wigner's celebrated conjecture states that the statistics of the eigenvalue differences should only depend on the symmetry class of the model, not on the details of the randomness that generated the model. There have been multiple works in recent years that shed light on this phenomenon, [23, 35].

[*]Stanford University, United States of America. E-mail: arkaa@math.stanford.edu
[†]Harvard University, United States of America. E-mail: sdubova@math.harvard.edu
[‡]Center for Mathematical Sciences and Application, United States of America.
 E-mail: cxu@cmsa.fas.harvard.edu
[§]University of California, Los Angeles, United States of America. E-mail: jyin@math.ucla.edu

Even though the eigenvalues of random matrices are relatively well understood, the eigenvector statistics of random matrices remain largely mysterious. In contrast to the statistics of eigenvalue distributions where there are many powerful tools such as the Dyson-Brownian motion [23, 26, 28, 22, 25, 6, 7, 27, 30, 21, 20], the four moment method [35], and direct computation via the Brezin-Hikami formula [10, 11], the equations determining the behavior of the eigenvector are less amenable to analysis.

There are multiple conjectures regarding the behavior of the eigenvectors of random matrices inspired by conjectures derived from studying the quantum analogues of dynamical systems. The BGS conjecture [5] proposed that the eigenvalue behavior of the quantum analogues of classically chaotic dynamical systems should follow appropriate random matrix statistics; this conjecture, and various others, suggested a deep link between dynamical systems and random matrix theory. The study of eigenstates of these quantum dynamical operators led to very rich behavior; such as the celebrated Quantum Unique Ergodicity conjecture by Rudnik and Sarnak [33]. This suggests that, as $i \to \infty$, all eigenstates $\phi_i(x)$ of the Laplace-Beltrami operator on a surface with ergodic geodesic flow have an associated measure $|\phi_i(x)|^2 dx$ that becomes flat as $i \to \infty$, except for an exceptional sequence. The Eigenstate Thermalization Hypothesis [18, 19, 34] is implied by similar claims regarding the value of the related observable $\langle \phi_i, A\phi_j \rangle$ $i, j \to \infty$, for appropriate operators A. For further discussion of these results and references, refer to [14].

There has recently been significant work in the random matrix community, to try to find analogs of these eigenvector behaviors in random matrix theory. Estimates from Green's functions [4, 24, 28, 31] showed delocalization of eigenvectors; namely, that the maximum entry of the eigenvector is of order close to $\frac{1}{N}$. These results have been strengthened in [8] to show Gaussian fluctuation for individual eigenvector entries; this is the appropriate analog of QUE for random matrices. The paper [8] shows

$$\sqrt{N}\langle u_i, q \rangle \to \mathcal{N}(0, 1), \tag{1.1}$$

i.e., the inner product of an eigenvector with a fixed vector approaches a standard normal random variable. Other results regarding proving QUE results include [3, 1, 8, 9, 37]. The paper [32] studied the correlation of a small number of eigenvector entries ($\mathcal{O}(N^\epsilon)$), and showed joint Gaussian behavior on these small windows. These works used was the eigenvector moment flow equations derived from Dyson Brownian motion. However, these equations are very difficult to analyze and do not yet give a complete description of the eigenvector statistics.

The random matrix analog of eigenstate thermalization was studied in [14] by G. Cipolloni, L. Erdős, and D. Schröder. In this paper, the authors tried to establish more global results on the distribution of the eigenvector. Namely, they were able to show, for a Wigner matrix, that, with overwhelming probability,

$$\max_{i,j} |\langle u_i, Au_j \rangle - \delta_{ij}\langle A \rangle| \lesssim \frac{N^\epsilon}{\sqrt{N}}, \tag{1.2}$$

where the error in the right hand side is optimal. In what follows, we use the notation $\langle \cdot, \cdot \rangle$ to denote inner product in vector computations or the normalized trace $\langle A \rangle := \frac{1}{N}\text{Tr}[A]$ as appropriate in context. Some of these results were extended to prove the normality of the terms $\langle u_i, Au_j \rangle$ in [17, 29] and multi-resolvent local laws in [15].

Rather than using the eigenvector moment flow, they directly studied more global quantities like $\Lambda := \langle GAGB \rangle$, where $G = (H - z)^{-1}$ is the Green's function of the Wigner random matrix H, while A and B are arbitrary matrices. These quantities reveal more information about the correlation of eigenvectors on larger scales and, furthermore, are easier to manipulate analytically. The method of this work involved using the cumulant

expansion to form a self-consistent equation for Λ. The details of the cumulant expansion procedure meant that the results of [14] were restricted to random matrices of Wigner type.

In the paper [15], multi-resolvent local laws for Wigner matrices were considered. They derived a hierarchy of equations to get more detailed estimates for traces of high powers of the form $(GA)^k$ which can also accommodate different traceless observables and handle them uniformly in all choices of observables. These results were expanded in the recent works [16, 13, 12, 2]: In [16], general local law for Wigner matrices which optimally handles observables of arbitrary rank were shown; thus, the paper unifies the averaged and isotropic local laws. [12] establishes the Eigenstate Thermalisation Hypothesis and Gaussian fluctuations for Wigner matrices with an arbitrary deformation. In [13], the authors prove an optimal lower bound on the diagonal overlaps of the corresponding non-Hermitian eigenvectors. [2] derives Gaussian fluctuations and gives a analog of the Berry conjecture for random matrices.

1.2 Difficulties in the case of generalized Wigner matrices

A Wigner matrix is a generalization of the the GUE or appropriate Gaussian ensemble. All of the entries are independent and identically distributed (i.i.d.) up to the appropriate symmetry constraints. Due to this nice symmetric structure, one might believe on an intuitive level, that all of the relevant eigenvalue and eigenvector statistics would match that of the corresponding Gaussian ensemble. Namely, the eigenvalues are distributed according to the sine kernel and, more relevant to our case, the eigenvectors are Haar distributed.

In the context of Eigenstate Thermalization, one can prove the following claim,

$$\langle GA_1 GA_2 \rangle \approx m^2 \langle A_1 A_2 \rangle. \tag{1.3}$$

Here, m is the solution to the semicircle equation

$$m^2 - zm + 1 = 0,$$

and one has the approximation $G_{ii} \approx m$, for all diagonal entries G_{ii} of the resolvent matrix. Thus, to leading order, one can derive the approximation in Eigenstate Thermalization by replacing the resolvent matrices G by the approximation mI. In this way, there is seemingly little contribution from the off-diagonal entries of G. As such, this statistic is further evidence for the approximate Haar distribution of the eigenvector entries in a Wigner matrix.

A generalized Wigner matrix is an ensemble of random matrices where every entry has an independent entry, up to symmetry conditions, but each entry has a different value of the variance; thus, the entries are not i.i.d. If W is our generalized Wigner matrix, then $\mathbb{E}[|W_{ij}|^2] = S_{ij}$, for some number S_{ij}. To ensure that the empirical spectral distribution scales to the semicircle law, we impose the following normalization constraint

$$\sum_j S_{ij} = 1, \forall i.$$

Even with this constraint in place, one still has the following leading order behavior of the entries of the resolvent,

$$G_{ii} \approx m, |G_{ij}| \approx \frac{1}{N \mathrm{Im}[z]} = o(1).$$

However, in the context of generalized Wigner matrices, we obtain an entirely different result. We have instead,

$$
\langle G(z_1)A_1 G(z_2)A_2 \rangle \approx m(z_1)m(z_2) \langle A_1 A_2 \rangle
$$
$$
+ m(z_1)m(z_2)\frac{1}{N}\sum_{\alpha,\beta}(A_1)_{\alpha\alpha}\left[S(I - m(z_1)m(z_2)\mathcal{C})^{-1}\right]_{\alpha\beta}(A_2)_{\beta\beta},
$$

(1.4)

where $\mathcal{C}_{\mu\nu} = S_{\mu\nu} - \frac{1}{N}$ for any $\mu,\nu \in [N]$.

To get the above expression, it is no longer possible to replace $G(z)$ by the most obvious approximation $G(z) \approx m(z)I$, even though the leading behavior entry of each of the entries in the resolvent G is the same as that of the Wigner random matrices. The implication of this fact is that there are detailed correlations present in the distribution of the eigenvectors of the generalized Wigner ensemble that are not present in the Wigner ensemble. In particular, the distribution of the eigenvectors of the generalized Wigner ensemble is far from Haar-distributed. Furthermore, the covariances of the terms $\langle u_i, A u_j \rangle$ would depend on the eigenvector indices i and j, while for the pure Wigner matrix, the covariance structure would be homogeneous in i and j. We also remark here that this is only an effect you see in full rank matrices A; in the context of QUE with finite rank matrices (or even N^ϵ rank matrices for $\epsilon < 1$), there is no difference in the covariance structure of eigenvectors for pure Wigner matrices and generalized Wigner matrices.

When coming to the proof of equation (1.4), the main difficult is a presence of a more complicated term during the derivation of the self-consistent equation for the quantity Λ. Namely, if we consider the case of computing $\langle GA_1 GA_2 \rangle$ and A_1, A_2 are both traceless matrices, we have to deal with a term of the following form,

$$
\frac{1}{N}\sum_{i,j}S_{ij}(G(z_1)A_1 G(z_2))_{jj}(G(z_2)A_2)_{ii}.
$$

(1.5)

In the Wigner case, we have that $S_{ij} = \frac{1}{N}$ for all i and j. Thus, the above quantity can be simplified as,

$$
\frac{1}{N}\sum_{i,j}S_{ij}(G(z_1)A_1 G(z_2))_{jj}(G(z_2)A_2)_{ii} = \langle G(z_1)A_1 G(z_2) \rangle \langle G(z_2)A_2 \rangle.
$$

(1.6)

Now, since we have the heuristic that $\langle G(z_2)A_2 \rangle \approx m(z_2)\langle A_2 \rangle = 0$ and m is the Stieltjes transform for the semicircle distribution, we can believe that the term above is merely a lower order term that should not complicate the analysis.

However, when $S_{ij} \neq \frac{1}{N}$ uniformly, there is no longer any way to write it as a product of traces. As such, it seems like using the fact that A_1 and A_2 are both traceless do not seem to give any cancellations. Indeed, if we take the approximation $G(z_i) \approx m(z_i)I$, we might guess that the term in (1.5) is at least as large as,

$$
\frac{1}{N}m^2\sum_{i,j}S_{ij}(A_1)_{jj}(A_2)_{ii}.
$$

(1.7)

We cannot hope for the quantity above to be of smaller order.

The fact that the contribution of the term (1.6) presents us with two problems. The first issue is to actually determine the value to leading order. The second is to actually present this term in such a way that we get a closed equation. As we have mentioned earlier, in the Wigner case, these terms can be presented as products of traces; this means that we can derive closed equations just involving these products of traces. Without a closed equation, we cannot hope to analyze the resulting self-consistent

equation; thus, it is of paramount importance to rewrite this term in a manner that is amenable to analysis. Our first main step is to write such terms as a product of traces by carefully decomposing the covariance matrix S_{ij}. By taking the square root, we have that,

$$S_{ij} = \sum_{\mu} \tilde{S}_{i\mu} \tilde{S}_{\mu j}.$$

With this decomposition in hand, we can rewrite,

$$
\begin{aligned}
&\frac{1}{N} \sum_{i,j} S_{ij} (G(z_1) A_1 G(z_2))_{jj} (G(z_2) A_2)_{ii} \\
&\qquad = \frac{1}{N} \sum_{i,j,\mu} \tilde{S}_{i\mu} \tilde{S}_{j\mu} (G(z_1) A_1 G(z_2))_{jj} (G(z_2) A_2)_{ii} \\
&\qquad = \frac{1}{N} \sum_{\mu} \left\langle G(z_2) A_2 N \mathrm{diag} \tilde{S}_{\mu} \right\rangle \left\langle G(z_1) A_1 G(z_2) N \mathrm{diag} \tilde{S}_{\mu} \right\rangle .
\end{aligned}
\tag{1.8}
$$

Here, $\mathrm{diag}\tilde{S}_{\mu}$ is the diagonal matrix whose ith entry is given by $\tilde{S}_{i\mu}$. The above expression looks like a more closed expression, due to the fact that we have written the above as a product of traces; however, we still need to consider traceless matrices if we actually want to consider eigenstate thermalization.

An immediate solution here is to consider the traceless parts of the matrices $A_2 N \mathrm{diag}\tilde{S}_{\mu}$ and $N \mathrm{diag}\tilde{S}_{\mu}$, but this is still not closed since we keep introducing new traceless matrices of the form $A_2 N \mathrm{diag}\tilde{S}_{\mu}$. The result of this procedure is to generate a chain of equations relating the Λ_A of certain matrices A to Λ_B of other matrices B. At each step of this procedure, the hierarchy of matrices considered grows rapidly, and it is not clear that this chain would lead to an effective bound. For instance, the matrices at level $k+1$ consists of any product of two matrices at level k. If one did not have precise control of appropriate prefactors when deriving the inequalities, then it would be impossible to derive useful information. For example, if one were to try to prove the case for non-diagonal matrices at the very beginning, one would have to deal with a cubic term that cannot be controlled via iteration. We circumvent this issue by first proving estimates for diagonal matrices, in which one can apply improved local law estimates, in order to have optimal estimates for the diagonal Λ_S. These estimates are key inputs for deriving bounds on Λ for the general case of non-diagonal matrices. The main achievement of Sections 3 and 4 of this manuscript is to derive this system of inequalities.

1.3 Conventions and notation

We use the notation $\prec$ to indicate stochastic domination (see also e.g. [14]) indicating a bound with very high probability up to a factor N^{ϵ} for any small $\epsilon > 0$. If

$$X = \left(X^{(N)}(u) \mid N \in \mathbb{N}, u \in U^{(N)} \right) \quad and \quad Y = \left(Y^{(N)}(u) \mid N \in \mathbb{N}, u \in U^{(N)} \right) \tag{1.9}$$

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

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

for large enough $N \geq N_0(\epsilon, D)$. In this case we use the notation $X \prec Y$ or $X = \mathcal{O}_{\prec}(Y)$.

For any $N \times N$ matrix M we use the following notation for the normalized trace:

$$\langle M \rangle = \frac{1}{N} \mathrm{tr} M. \tag{1.10}$$

2 Main results

In this paper, we consider strictly-complex generalized Wigner matrices. Namely, these are Hermitian matrices, where each entry is independent, but they are allowed to have different variances. Our normalization condition on the variances is that $\sum_j S_{ij} = 1, \forall i$, where S_{ij} is the variance of the (i, j)th entry. We let $S = [S_{ij}]$ denote the full covariance matrix of our generalized Wigner matrix. For a more formal definition, see Section 2 of [28]. We assume the following condition on S.

Assumption 2.1. *Let $\tilde{S}$ be the square root of S. We assume that there is a constant $C > 0$ such that for all i, j, we have,*

$$|\tilde{S}_{ij}| \le C\frac{1}{N}. \tag{2.1}$$

One can check that this condition holds if S were the matrix whose entries were all $\frac{1}{N}$. To simplify computations in this paper, we also assume that the second cumulant of each matrix entry w_{ab} (which will be denoted by $\kappa_{ab}^{2,0}$) will be equal to 0. Note that this implies that $\mathbb{E}[(\Re w_{ab})^2] = \mathbb{E}[(\Im w_{ab})^2]$.

We remark here that by assuming S has a square root, we can write many of the quantities that are obtained by performing a cumulant expansion by a normalized product of traces of matrices. By contrast, if we did not make such an assumption, we would have obtained a large hierarchy of terms of different forms that would be more complicated to control.

From the paper [28], we have the following a-priori estimate on the behavior of the Green's function of our generalized Wigner matrices. These will be used multiple times in the proof.

Theorem 2.2 ([28, Theorems 2.1, 2.2]). *Let W be a generalized Wigner matrix. We assume that the probability distributions of each entry of W have a uniform sub-exponential decay. Then the following estimates hold:*

$$\|G(z) - mI\|_{\max} \prec \sqrt{\frac{\Im m(z)}{N\eta}} + \frac{1}{N\eta} \quad \text{for all } |E| \le 5, \eta \ge N^{-1+\epsilon}. \tag{2.2}$$

$$|\lambda_i - \gamma_i| \prec \min(i, N - i + 1)^{-1/3} N^{-2/3} \text{ for all } 1 \le i \le N. \tag{2.3}$$

With these preliminaries in hand, we can state our main result. Given a Hermitian matrix H, we can order its eigenvalues as $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_N$ and write its corresponding eigenvectors as $u_1, \ldots, u_n$.

Theorem 2.3. *Let A be a Hermitian matrix with trace 0 and bounded norm $\|A\| \le 1$. Let W be our random generalized Wigner matrix as we have previously constructed. With overwhelming probability for any $\xi > 0$, we can derive the following estimates:*

$$\max_{i,j}|\langle u_i, Au_j\rangle| + \max_{i,j}|\langle u_i, A\overline{u}_j\rangle| \le \frac{N^\xi}{\sqrt{N}}. \tag{2.4}$$

We study the entrywise maximum through the following intermediate quantity Ξ_M, as in [14]. Ξ_M computes averaged versions of the quantity in interest in Theorem 2.3.

Definition 2.4. *Let A be a matrix and $J \in \mathbb{N}$. We define $\Xi_A, \overline{\Xi}_A$ as,*

$$\begin{aligned}
\Xi_A(J) &:= \frac{N}{(2J)^2} \max_{i_0, j_0} \sum_{|i-i_0| \le J} \sum_{|j-j_0| \le J} |\langle u_i, Au_j\rangle|^2, \\
\overline{\Xi}_A(J) &:= \frac{N}{(2J)^2} \max_{i_0, j_0} \sum_{|i-i_0| \le J} \sum_{|j-j_0| \le J} |\langle u_i, A\overline{u}_j\rangle|^2.
\end{aligned} \tag{2.5}$$

We will omit the dependence of Ξ_A on J when the context is clear.

In contrast to the paper [14], in which the authors could derive a self-consistent equation consisting of only one matrix, we have the relate the quantities Ξ_M of different families of matrices to each other. We now introduce the following classes of deterministic matrices of interest.

$$\mathbb{M}_0 := \{N \operatorname{diag} \tilde{S}_\mu\}_{1 \le \mu \le N}, \quad \mathbb{M}_1 := \{I, A\} \cup \mathbb{M}_0, \tag{2.6}$$

$$\mathbb{M}_k^\circ := \{B - \langle B \rangle : B \in \mathbb{M}_k\}, \tag{2.7}$$

$$\mathbb{M}_k := \{B_1 B_2 : B_1, B_2 \in \mathbb{M}_{k-1} \cup \mathbb{M}_{k-1}^\circ, 1 \le \mu \le N\} \quad \text{for } k \ge 2, \tag{2.8}$$

$$\Lambda_k := \max_{B \in \mathbb{M}_k^\circ} \Xi_B + \max_{B \in \mathbb{M}_k} \overline{\Xi}_B + 1. \tag{2.9}$$

We remark that our control on $\overline{\xi}_B$ comes from the fact that we are assuming that we are dealing with strictly complex generalized Wigner matrices with $\mathbb{E}|w_{ij}|^2 \ne \mathbb{E}W_{ij}^2$. Furthermore, since the identity matrix is always found in $\mathbb{M}_k$, one can check that $\mathbb{M}_{k-1} \subset \mathbb{M}_k$ and thus, $\Lambda_1 \le \Lambda_2 \ldots \le \Lambda_k$.

The bound in the following lemma is a simple consequence of our definitions.

Lemma 2.5. $\sup_{B \in \mathbb{M}_k \cup \mathbb{M}_k^\circ} \|B\|_{l^2 \to l^2} \le (\sup_{B \in \mathbb{M}_1 \cup \mathbb{M}_1^\circ} \|B\|_{l^2 \to l^2})^{2^k}$.

Our main result Theorem 2.3 is an easy corollary of the following result on the size of the control parameters Ξ_M and $\overline{\Xi}_M$.

Theorem 2.6. *Fix $1 > \epsilon > 0$ and $J \ge N^\epsilon$. Let W be our generalized Wigner matrix as before and let A be a trace-less Hermitian matrix with bounded norm $||A|| \le 1$. Then, we have the following estimates*

$$\Xi_A(J), \overline{\Xi}_A(J) \prec 1. \tag{2.10}$$

Proof of Theorem 2.3. From Theorem 2.6, we know that $\Xi_A \prec 1$. From this fact then necessarily, for any i, j, we must have that $|\langle u_i, A u_j \rangle|^2 \prec \frac{(2J)^2}{N}$. By taking square roots of both sides, we are done. $\square$

Furthermore, the function Ξ_A can be related to more standard functions of the resolvent of our Wigner matrix; $G(z) := (W - z)^{-1}$. In what appears later, recall from equation (1.10) the notion of a normalized trace.

For example, consider the following expression with $z_i = E_i + \mathrm{i}\eta_i$:

$$\langle \Im G(z_1) A \Im G(z_2) A^* \rangle = \frac{1}{N} \sum_{i,j} \frac{|\langle u_i, A u_j \rangle|^2 \eta_1 \eta_2}{((\lambda_i - E_1)^2 + \eta_1^2)((\lambda_j - E_2)^2 + \eta_2^2)}. \tag{2.11}$$

The following lemma explicitly writes out the relations between Ξ_A and the quantity presented in equation (2.11). Recall, the notion of a classical eigenvalue location γ_i defined implicitly as

$$\frac{i}{N} = \int_{-\infty}^{\gamma_i} \rho_{sc}(x) \mathrm{d}x.$$

Lemma 2.7. *Fix $E_1 = \gamma_{i_0}$, $E_2 = \gamma_{j_0}$ and $J \ge N^\epsilon$, where the $\gamma_i's$ represent the classical eigenvalue locations of the ith eigenvalue. Choose η_1 and η_2 so that the following equation holds $J = N \eta_i \rho_i$, where $\rho_i = \Im m(E_i + \mathrm{i}\eta_i)$. Then, we have the following claim,*

$$\frac{N}{(2J)^2} \sum_{\substack{|i-i_0| \le J \\ |j-j_0| \le J}} |\langle u_i, A u_j \rangle|^2 \prec \frac{\langle \Im G(z_1) A \Im G(z_2) A^* \rangle}{\rho_1 \rho_2} \prec \frac{N}{(2J)^2} \sum_{\substack{|i-i_0| \le J \\ |j-j_0| \le J}} |\langle u_i, A u_j \rangle|^2,$$

$$\frac{N}{(2J)^2} \sum_{\substack{|i-i_0| \le J \\ |j-j_0| \le J}} |\langle u_i, A \overline{u}_j \rangle|^2 \prec \frac{\langle \Im G(z_1) A \Im G^t(z_2) A^* \rangle}{\rho_1 \rho_2} \prec \frac{N}{(2J)^2} \sum_{\substack{|i-i_0| \le J \\ |j-j_0| \le J}} |\langle u_i, A \overline{u}_j \rangle|^2. \tag{2.12}$$

Proof. This is a consequence of eigenvalue rigidity (2.3) for generalized Wigner matrices. See [14, Lemma 3.2]. □

Our basic tool for deriving a self consistent equation for quantities of the form appearing in equation (2.11) is integration by parts. One of our main error terms produced by this integration by parts procedure is the following renormalized term.

Definition 2.8 (Renormalized Matrix Products)**.** *Given a matrix product of the from $f(W)Wg(W)$, we can define the renormalized matrix product $\underline{f(W)Wg(W)}$ as,*

$$\underline{f(W)Wg(W)} := f(W)Wg(W) - \mathbb{E}_{\tilde{W}}(\partial_{\tilde{W}}f)(W)\tilde{W}g(W) - \mathbb{E}_{\tilde{W}}f(W)\tilde{W}(\partial_{\tilde{W}}g)(W). \quad (2.13)$$

The derivative $\partial_{\tilde{W}}f = \sum_{i,j} \tilde{W}_{ij}\partial_{ij}f$, where $\partial_{ij}f$ is the standard partial derivative of f with respect to the ijth matrix entry and $\tilde{W}$ is an independent copy of W.

Remark 2.9. The terms subtracted in (2.13) are the first order terms in the integration by parts of $f(W)Wg(W)$ with respect to the middle W in the product.

Our final main lemma computes the size of the renormalized term for our relevant quantities of interest. The proof of this lemma can be found in section 5.

Lemma 2.10. *Let W be a generalized Wigner matrix satisfying the conditions lined out in Assumption 2.1. Suppose for $i \in \{1, 2\}$ $z_i \in \mathbb{C} \setminus \mathbb{R}$, $\eta_i = |\Im z_i|$, $\rho_i = \Im m_i$, $L = \min |N\eta_i\rho_i|$, $\eta_* = \min(\eta_1, \eta_2)$. Then, we have the following estimates.*
For $G_i \in \{G(z_i), G^*(z_i), G^t(z_i), \Im G(z_i)\}$ and $A \in \mathbb{M}_k^\circ$

$$\left| \langle \underline{WG_iA} \rangle \right| \prec \frac{\rho_i\Lambda_k}{\sqrt{NL}} . \quad (2.14)$$

For $G_i \in \{G(z_i), G^*(z_i), G^t(z_i)\}$ and $A \in \mathbb{M}_k^\circ$

$$\left| \langle \underline{WG_1G_2A} \rangle \right| \prec \frac{\Lambda_k}{L\sqrt{\eta_*}} , \quad \left| \langle \underline{WG_1\Im G_2A} \rangle \right| \prec \frac{\rho_2\Lambda_k}{L\sqrt{\eta_*}} ,$$
$$\left| \langle \underline{W\Im G_1G_2A} \rangle \right| \prec \frac{\rho_1\Lambda_k}{L\sqrt{\eta_*}} , \quad \left| \langle \underline{W\Im G_1\Im G_2A} \rangle \right| \prec \frac{\rho_1\rho_2\Lambda_k}{L\sqrt{\eta_*}} . \quad (2.15)$$

For $G_i \in \{G(z_i), G^*(z_i), G^t(z_i)\}$, $A_1 \in \mathbb{M}_k^\circ$ and $A_2 \in \mathbb{M}_l^\circ$

$$|\langle \underline{WG_1A_1G_2A_2} \rangle| \prec \frac{\Lambda_k\Lambda_l}{\sqrt{L}} , \quad |\langle \underline{WG_1A_1\Im G_2A_2} \rangle| \prec \frac{\rho_2\Lambda_k\Lambda_l}{\sqrt{L}} ,$$
$$|\langle \underline{W\Im G_1A_1G_2A_2} \rangle| \prec \frac{\rho_1\Lambda_k\Lambda_l}{\sqrt{L}} , \quad |\langle \underline{W\Im G_1A_1\Im G_2A_2} \rangle| \prec \frac{\rho_1\rho_2\Lambda_k\Lambda_l}{\sqrt{L}} . \quad (2.16)$$

3 Proof of Theorem 2.6 for M diagonal

To prove Theorem 2.6, we need the bounds on expressions of the form $\langle \Im GA\Im GA^* \rangle$ and $\langle \Im GA\Im G^tA^* \rangle$ in terms of Λ_k. To do this, we first have to study simpler expressions like $\langle GA \rangle$, $\langle GGA \rangle$, $\langle GAGA \rangle$, etc.

Throughout Sections 3 and 4 we use the following notation. Let $z_i \in \mathbb{C} \setminus \mathbb{R}$, $G_i \in \{G_i(z), G_i^*(z)\}$, $\eta_i = |\Im z_i|$, $\rho_i = \Im m_i$, $L = \min |N\eta_i\rho_i|$, $\eta_* = \min \eta_i$. In the case that the studied expression has a single resolvent, we omit the index i.

In this section we assume that A is diagonal and, thus, all matrices in the families $\mathbb{M}_k$ and $\mathbb{M}_k^\circ$ are diagonal.

3.1 Bounds on $\langle GA \rangle$

Lemma 3.1. *Let $A \in \mathbb{M}_k^\circ$. Then, we have that,*

$$|\langle GA \rangle| \prec \frac{\sqrt{\rho}\Lambda_k}{N\sqrt{\eta}} = \frac{\rho\Lambda_k}{\sqrt{NL}}, \tag{3.1}$$

and, therefore

$$|\langle \Im GA \rangle| \prec \frac{\sqrt{\rho}\Lambda_k}{N\sqrt{\eta}} = \frac{\rho\Lambda_k}{\sqrt{NL}}. \tag{3.2}$$

Proof. First, we start with the following identity,

$$G = mI - mWG - m^2 G. \tag{3.3}$$

Multiplying this by the matrix A, we get,

$$GA = mA - mWGA - m^2 GA. \tag{3.4}$$

We replace the term WGA by the renormalization from Definition 2.8 and derive,

$$(GA)_{ik} = mA_{ik} - m(\underline{WGA})_{ik} + m\sum_j S_{ij}(G_{jj} - m)(GA)_{ik}. \tag{3.5}$$

Taking the trace of this expression, we have, for traceless matrices A, that

$$\langle GA \rangle = -m\langle \underline{WGA} \rangle + m\frac{1}{N}\sum_{i,j} S_{ij}(G_{jj} - m)(GA)_{ii}. \tag{3.6}$$

We introduce the splitting $S_{ij} = \sum_\mu \tilde{S}_{i\mu}\tilde{S}_{\mu j}$ on the last term and we further introduce the traceless part $\tilde{S}_{i\mu} = \tilde{S}_{i\mu}^\circ + \frac{1}{N}$.

$$
\begin{aligned}
\frac{1}{N}\sum_{i,j} S_{ij}(G_{jj} - m)(GA)_{ii} &= \frac{1}{N}\sum_\mu \sum_{i,j} \tilde{S}_{i\mu}(GA)_{ii}\tilde{S}_{\mu j}(G_{jj} - m) \\
&= \frac{1}{N}\sum_\mu \sum_{i,j} \tilde{S}_{i\mu}(GA)_{ii}\tilde{S}_{\mu j}^\circ(G_{jj} - m) \\
&\quad + \frac{1}{N^2}\sum_\mu \sum_{i,j} \tilde{S}_{\mu i}(GA)_{ii}(G_{jj} - m) \\
&= \frac{1}{N}\sum_\mu \left\langle GAN\mathrm{diag}\tilde{S}_\mu \right\rangle \left\langle GN\mathrm{diag}\tilde{S}_\mu^\circ \right\rangle + \langle GA \rangle \langle G - mI \rangle \\
&= \frac{1}{N}m\sum_\mu \left\langle AN\mathrm{diag}\tilde{S}_\mu \right\rangle \left\langle GN\mathrm{diag}\tilde{S}_\mu^\circ \right\rangle \\
&\quad + \frac{1}{N}\sum_\mu \left\langle (G - mI)AN\mathrm{diag}\tilde{S}_\mu \right\rangle \left\langle GN\mathrm{diag}\tilde{S}_\mu^\circ \right\rangle + \langle GA \rangle \langle G - mI \rangle.
\end{aligned}
\tag{3.7}
$$

We will specialize $A = N\mathrm{diag}\tilde{S}_\nu^\circ$ to get a certain system of equations. First, we have

$$\frac{1}{N}\sum_{i,j} S_{ij}(G_{jj} - m)(GN\mathrm{diag}\tilde{S}_\nu^\circ)_{ii} = \sum_\mu C_{\nu\mu}\langle GN\mathrm{diag}\tilde{S}_\mu^\circ \rangle, \tag{3.8}$$

where the coefficients $C_{\nu\mu}$ are

$$
\begin{aligned}
C_{\nu\mu} &= \frac{m}{N}\left\langle N\mathrm{diag}\tilde{S}_\nu^\circ N\mathrm{diag}\tilde{S}_\mu \right\rangle + \frac{1}{N}\left\langle (G - mI)N\mathrm{diag}\tilde{S}_\nu^\circ N\mathrm{diag}\tilde{S}_\mu \right\rangle + \langle G - mI \rangle\, \delta_{\nu\mu} \\
&= m\left[S_{\nu\mu} - \frac{1}{N} \right] + \mathcal{O}_\prec\left(\frac{1}{N^{3/2}\eta^{1/2}} \right) + \delta_{\nu\mu}\mathcal{O}_\prec\left(\frac{1}{N\eta} \right).
\end{aligned}
\tag{3.9}
$$

In the line above, we used the local law to bound the diagonal entries of $G - mI$ by $\frac{1}{\sqrt{N\eta}}$. Furthermore, $N\mathrm{diag}\tilde{S}_\mu^\circ N\mathrm{diag}\tilde{S}_\nu$ is a diagonal matrix with entries being $\mathcal{O}(1)$. Thus, we see that,

$$
\frac{1}{N}\langle (G-mI)N\mathrm{diag}\tilde{S}_\nu^\circ N\mathrm{diag}\tilde{S}_\mu\rangle
$$
$$
= \frac{1}{N^2}\sum_{i=1}^N (G-mI)_{ii}[N\mathrm{diag}\tilde{S}_\nu^\circ N\mathrm{diag}\tilde{S}_\mu]_{ii} = \mathcal{O}_\prec\left(\frac{1}{N^{3/2}\eta^{1/2}}\right). \tag{3.10}
$$

Placing all of these estimates back into the equation (3.6) for specialized values of $A = N\mathrm{diag}\tilde{S}_\nu^\circ$, we have,

$$
(I - C)\left\langle GN\mathrm{diag}\tilde{S}^\circ\right\rangle = -m\left\langle \underline{WGN\mathrm{diag}\tilde{S}^\circ}\right\rangle. \tag{3.11}
$$

Here, $\langle G\mathrm{diag}\tilde{S}^\circ\rangle$ and $\langle \underline{WG\mathrm{diag}\tilde{S}^\circ}\rangle$ is a shorthand for the column vector constructed using these terms for the matrix $\mathrm{diag}\tilde{S}_\mu^\circ$ for each μ. We are abusing notation slightly here and treat $\left\langle BN\mathrm{diag}\tilde{S}^\circ\right\rangle$ (where the term $\tilde{S}^\circ$ does not have a subscript) as a vector whose entries are $\left\langle BN\mathrm{diag}\tilde{S}_\mu^\circ\right\rangle$

This finally gives us,

$$
\left\langle GN\mathrm{diag}\tilde{S}^\circ\right\rangle = -m(I-C)^{-1}\left\langle \underline{WGN\mathrm{diag}\tilde{S}^\circ}\right\rangle. \tag{3.12}
$$

Lemma 3.2. *Assume that we have $S_{ij} \geq \frac{c}{N}$ for all values i, j. The largest eigenvalue of $S - \frac{1}{N}\mathbf{1}^T\mathbf{1}$ in absolute value is bounded from above by $1 - c$. Here, $\mathbf{1}$ is the vector of all 1 s.*

Proof. The matrix $S - \frac{1}{N}\mathbf{1}^T\mathbf{1}$ can be decomposed as follows,

$$
S - \frac{1}{N}\mathbf{1}^T\mathbf{1} = S_2 + \frac{c-1}{N}\mathbf{1}^T\mathbf{1}, \tag{3.13}
$$

All of the entries of S_2 are positive; furthermore, the sum over each row and each column is bounded by $1 - c$. This shows that the $l_2 \to l_2$ operator norm of the matrix S_2 is less than $1 - c$. If we look at the orthogonal space to the vector $\mathbf{1}$, we see that $\sup_{\langle v,\mathbf{1}\rangle=0}|v\left[S - \frac{1}{N}\mathbf{1}^T\mathbf{1}\right]v^T| \leq 1 - c$.

Furthermore, the vector $\mathbf{1}$ is an eigenvector of the matrix $S - \frac{1}{N}\mathbf{1}^T\mathbf{1}$ with eigenvalue $c - 1$. Thus, the largest eigenvalue of S is bounded by,

$$
\max\left(|c-1|, \sup_{\langle v,\mathbf{1}\rangle=0} v\left[S - \frac{1}{N}\mathbf{1}^T\mathbf{1}\right]v^T\right) \leq 1 - c,
$$

as desired. $\qquad\square$

Because of the above lemma, along with the fact that $|m| < 1$ (as the Stieltjes transform of the semicircle distribution), we know that the largest eigenvalue of C is bounded from above in absolute value by $1 - c$. Thus, the inverse $(I - C)^{-1}\langle \underline{WNG\mathrm{diag}\tilde{S}^\circ}\rangle$ is well-defined and bounded in l_2 vector norm by $\sqrt{N}\frac{\sqrt{\rho}\Lambda_1}{\sqrt{NL}}$.

Placing this estimate back into the equation for an individual row in (3.11), we find that,

$$
\left\langle GN\mathrm{diag}\tilde{S}_\mu^\circ\right\rangle = \left\langle C_\mu, \left\langle GN\mathrm{diag}\tilde{S}^\circ\right\rangle\right\rangle - m\left\langle \underline{WGN\mathrm{diag}\tilde{S}_\mu}\right\rangle. \tag{3.14}
$$

C_μ is the μth row of the matrix C. Now, C_μ is bounded in l_2 norm by $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$. By the Cauchy-Schwartz inequality $\langle C_\mu, \langle GN\mathrm{diag}\tilde{S}^\circ\rangle\rangle$ can be bounded by $\prec \frac{\sqrt{\rho}\Lambda_1}{\sqrt{NL}}$.

This shows that the entries,

$$|\langle GN\mathrm{diag}\tilde{S}_\mu\rangle| \prec \frac{\sqrt{\rho}\Lambda_1}{\sqrt{NL}}. \tag{3.15}$$

At this point, we can return to an analysis for general matrices A. From the equation (3.7), we see that,

$$\langle GA\rangle[1 - \langle G - mI\rangle] = -m\langle \underline{WGA}\rangle + \frac{m^2}{N}\sum_\mu \langle A(N\mathrm{diag}(\tilde{S}^\mu))\rangle\langle GN\mathrm{diag}(\tilde{S}_\mu^\circ)\rangle$$

$$+ \frac{1}{N}\sum_\mu \langle (G - mI)A(N\mathrm{diag}\tilde{S}_\mu)\rangle\langle GN\mathrm{diag}(\tilde{S}_\mu^\circ)\rangle \tag{3.16}$$

$$\prec \frac{\sqrt{\rho}\Lambda_k}{\sqrt{NL}}$$

Our earlier estimates on $\langle GN\mathrm{diag}\tilde{S}_\mu^\circ\rangle$ as well as on $\langle \underline{WGA}\rangle$ ensure that the right hand side is $\prec \frac{\Lambda_k}{N\sqrt{\eta}}$. Here, we specifically used the fact that A was a diagonal matrix, so that we can write,

$$\left\langle (G - mI)AN\mathrm{diag}\tilde{S}_\mu\right\rangle = \frac{1}{N}\sum_{i=1}^N (G - mI)_{ii}A_{ii}[N\mathrm{diag}\tilde{S}_\mu]_{ii}.$$

All the terms A_{ii} and $(N\mathrm{diag}\tilde{S}_\mu)_{ii}$ are $\mathcal{O}(1)$. Furthermore, $|G - mI|_{ii} = \mathcal{O}_\prec\left(\frac{1}{\sqrt{N\eta}}\right)$. Thus, the normalized trace considered above is $\mathcal{O}(1)$.

We can easily divide by $1 - \langle G - mI\rangle = 1 - o(1)$ to derive the same error estimate for $\langle GA\rangle$. $\qquad\square$

3.2 Bounds on $\langle GGA\rangle$

Lemma 3.3. *Let $A \in \mathbb{M}_k^\circ$. Then*

$$|\langle G_1 G_2 A\rangle| \prec \frac{\Lambda_k}{L\sqrt{\eta_*}}, \tag{3.17}$$

$$|\langle G_1 \Im G_2 A\rangle| \prec \frac{\rho_2 \Lambda_k}{L\sqrt{\eta_*}}, \tag{3.18}$$

$$|\langle \Im G_1 \Im G_2 A\rangle| \prec \frac{\rho_1 \rho_2 \Lambda_k}{L\sqrt{\eta_*}}. \tag{3.19}$$

Proof. First, we use identity (3.3) on G_1 and replace WG_1 by its renormalization from Definition 2.8 as follows.

$$(G_1 G_2 A)_{ik} = \sum_{l=1}^N \left(m_1 \delta_{il} - m_1(\underline{WG_1})_{il} + m_1 \sum_{j=1}^N S_{ij}((G_1)_{jj} - m_1)(G_1)_{il}\right)(G_2 A)_{lk}$$

$$= m_1(G_2 A)_{ik} - m_1(\underline{WG_1}G_2 A)_{ik} + m_1 \sum_{j=1}^N S_{ij}((G_1)_{jj} - m_1)(G_1 G_2 A)_{ik}. \tag{3.20}$$

From Definition 2.8, we can see that

$$(\underline{WG_1 G_2 A})_{ik} = (\underline{WG_1}G_2 A)_{ik} + \sum_{j=1}^N S_{ij}(G_1 G_2)_{jj}(G_2 A)_{ik}. \tag{3.21}$$

Thus,

$$\langle G_1 G_2 A \rangle = m_1 \langle G_2 A \rangle - m_1 \langle \underline{W G_1 G_2 A} \rangle + \frac{1}{N} m_1 \sum_{i,j=1}^{N} S_{ij} (G_1 G_2)_{jj} (G_2 A)_{ii}$$

$$+ \frac{1}{N} m_1 \sum_{i,j=1}^{N} S_{ij} ((G_1)_{jj} - m_1)(G_1 G_2 A)_{ii}. \tag{3.22}$$

In the last two terms, we split S as follows.

$$\langle G_1 G_2 A \rangle = m_1 \langle G_2 A \rangle - m_1 \langle \underline{W G_1 G_2 A} \rangle$$

$$+ \frac{1}{N} m_1 \sum_{\mu=1}^{N} \left\langle G_1 G_2 N \mathrm{diag} \tilde{S}_\mu^\circ \right\rangle \left\langle G_2 A N \mathrm{diag} \tilde{S}_\mu \right\rangle + m_1 \langle G_1 G_2 \rangle \langle G_2 A \rangle$$

$$+ \frac{1}{N} m_1 \sum_{\mu=1}^{N} \left\langle G_1 N \mathrm{diag} \tilde{S}_\mu^\circ \right\rangle \left\langle G_1 G_2 A N \mathrm{diag} \tilde{S}_\mu \right\rangle + m_1 \langle G_1 - m_1 \rangle \langle G_1 G_2 A \rangle. \tag{3.23}$$

To bound the first term we use Lemma 3.1 and get

$$|m_1 \langle G_2 A \rangle| \prec \frac{\rho_2 \Lambda_k}{\sqrt{NL}}. \tag{3.24}$$

Suppose $B = A N \mathrm{diag} \tilde{S}_\mu$ or $B = I$. We apply Cauchy-Schwarz to bound $\langle G_1 G_2 B \rangle$.

$$|\langle G_1 G_2 B \rangle| \leq \langle G_1 G_1^* \rangle^{\frac{1}{2}} \langle G_2 B B^* G_2^* \rangle^{\frac{1}{2}} \prec \frac{\sqrt{\rho_1 \rho_2}}{\sqrt{\eta_1 \eta_2}} \prec \frac{N \rho_1 \rho_2}{L}. \tag{3.25}$$

This gives us

$$\left| \left\langle G_1 N \mathrm{diag} \tilde{S}_\mu^\circ \right\rangle \left\langle G_1 G_2 A N \mathrm{diag} \tilde{S}_\mu \right\rangle \right| \prec \frac{\Lambda_1}{N \sqrt{\eta_*}} \cdot \frac{N \rho_1 \rho_2}{L} \prec \frac{\rho_1 \rho_2 \Lambda_1}{L \sqrt{\eta_*}}. \tag{3.26}$$

and

$$|\langle G_1 G_2 \rangle \langle G_2 A \rangle| \prec \frac{\Lambda_k}{N \sqrt{\eta_*}} \cdot \frac{N \rho_1 \rho_2}{L} \prec \frac{\rho_1 \rho_2 \Lambda_k}{L \sqrt{\eta_*}} \tag{3.27}$$

Using this estimate, the estimate for $\langle G A \rangle$ from above and $\langle G_1 - m_1 \rangle \prec \frac{1}{N \eta_1}$, we get

$$\left(1 + \mathcal{O} \left(\frac{1}{N \eta_1} \right) \right) \langle G_1 G_2 A \rangle = -m_1 \langle \underline{W G_1 G_2 A} \rangle$$

$$+ \frac{1}{N} m_1 \sum_{\mu=1}^{N} \left\langle G_1 G_2 N \mathrm{diag} \tilde{S}_\mu^\circ \right\rangle \left\langle G_2 A N \mathrm{diag} \tilde{S}_\mu \right\rangle \tag{3.28}$$

$$+ \mathcal{O}_\prec \left(\frac{\Lambda_k \rho_1 \rho_2}{L \sqrt{\eta_*}} \right).$$

We can bound $\langle \underline{W G_1 G_2 A} \rangle \prec \frac{\Lambda_k}{L \sqrt{\eta_*}}$ via Lemma 2.10.

Now we plug in $A = N \mathrm{diag} \tilde{S}_\nu^\circ$ to get the system of equations

$$(I - C) \left\langle G_1 G_2 N \mathrm{diag} \tilde{S}^\circ \right\rangle = \mathcal{O}_\prec \left(\frac{\Lambda_1}{L \sqrt{\eta_*}} \right), \tag{3.29}$$

where C is a matrix with

$$
\begin{aligned}
C_{\nu\mu} &= \frac{1}{N}m_1 \left\langle G_2 N \mathrm{diag}\tilde{S}_\nu^\circ N \mathrm{diag}\tilde{S}_\mu \right\rangle + \mathcal{O}_\prec\left(\frac{1}{N\eta}\right)\delta_{\nu\mu} \\
&= \frac{1}{N}m_1 \left\langle (G_2 - m_2) N \mathrm{diag}\tilde{S}_\nu^\circ N \mathrm{diag}\tilde{S}_\mu \right\rangle + m_1 m_2 \left(S_{\nu\mu} - \frac{1}{N}\right) + \mathcal{O}_\prec\left(\frac{1}{N\eta}\right)\delta_{\nu\mu} \\
&= m_1 m_2 \left(S_{\nu\mu} - \frac{1}{N}\right) + \mathcal{O}_\prec\left(\frac{1}{N\eta}\right)\delta_{\nu\mu} + \mathcal{O}_\prec\left(\frac{1}{N^{3/2}\eta^{1/2}}\right).
\end{aligned}
\tag{3.30}
$$

To get the above estimates, we used the fact that $A = N\mathrm{diag}\tilde{S}_\nu^\circ$ has the better error bounds from (3.15).

Similarly to the proof of Lemma 3.1 we use Lemma 3.2 to invert matrix $I - C$ in (3.29) and get

$$
\left|\left\langle G_1 G_2 N \mathrm{diag}\tilde{S}^\circ \right\rangle\right| \prec \frac{\Lambda_1}{L\sqrt{\eta_*}}.
\tag{3.31}
$$

Now, we can plug these estimates into equation (3.28). In general, we see that we have,

$$
\begin{aligned}
\left[1 - \mathcal{O}_\prec\left(\frac{1}{N\eta}\right)\right]|\langle G_1 G_2 A \rangle| &\prec \left|\frac{1}{N}\sum_{\mu=1}^{N}\langle G_1 G_2 N \mathrm{diag}\tilde{S}_\mu \rangle\langle G_2 A N \mathrm{diag}\tilde{S}_\mu \rangle\right| + \frac{\Lambda_k}{L\sqrt{\eta_*}} \\
&\prec \left|\frac{1}{N}\sum_{\mu=1}^{N}\langle G_1 G_2 N \mathrm{diag}\tilde{S}_\mu \rangle\left[\langle (G_2 - m_2) A N \mathrm{diag}\tilde{S}_\mu \rangle + m_2\langle A N \mathrm{diag}\tilde{S}_\mu \rangle\right]\right| + \frac{\Lambda_k}{L\sqrt{\eta_*}} \\
&\prec \frac{\Lambda_k}{L\sqrt{\eta_*}}.
\end{aligned}
\tag{3.32}
$$

The fact that A is diagonal allows us to use the local law in order to bound

$$
|\langle (G_2 - m_2) A N \mathrm{diag}\tilde{S}_\mu \rangle| \prec \frac{1}{\sqrt{N\eta_2}}.
\tag{3.33}
$$

Additionally, $\langle A N \mathrm{diag}\tilde{S}_\mu \rangle$ is $\mathcal{O}(1)$, while $|\langle G_1 G_2 N \mathrm{diag}\tilde{S}_\mu \rangle| \prec \frac{\Lambda_k}{L\sqrt{\eta_*}}$. All these estimates together complete the proof of (3.17).

Other bounds in Lemma 3.3 are proved similarly. For example, to bound $\langle G_1 \Im G_2 A \rangle$ we use the identity

$$
\begin{aligned}
\langle G_1 \Im G_2 A \rangle &= m_1 \langle \Im G_2 A \rangle - m_1 \underline{\langle W G_1 \Im G_2 A \rangle} \\
&\quad + m_1 \frac{1}{N}\sum_{i,j=1}^{N} S_{ij}(G_1 \Im G_2)_{jj}(G_2^* A)_{ii} \\
&\quad + m_1 \frac{1}{N}\sum_{i,j=1}^{N} S_{ij}(G_1 G_2)_{jj}(\Im G_2 A)_{ii} \\
&\quad + m_1 \frac{1}{N}\sum_{i,j=1}^{N} S_{ij}((G_1)_{jj} - m_1)(G_1 \Im G_2 A)_{ii}
\end{aligned}
\tag{3.34}
$$

After splitting S, we get

$$
\begin{aligned}
\langle G_1 \Im G_2 A \rangle = m_1 \langle \Im G_2 A \rangle &- m_1 \underline{\langle W G_1 \Im G_2 A \rangle} \\
&+ m_1 \frac{1}{N} \sum_{\mu=1}^{N} \left\langle G_1 \Im G_2 N \mathrm{diag} \tilde{S}_\mu^\circ \right\rangle \left\langle G_2^* A N \mathrm{diag} \tilde{S}_\mu \right\rangle + m_1 \langle G_1 \Im G_2 \rangle \langle G_2^* A \rangle \\
&+ m_1 \frac{1}{N} \sum_{\mu=1}^{N} \left\langle G_1 G_2 N \mathrm{diag} \tilde{S}_\mu^\circ \right\rangle \left\langle \Im G_2 A N \mathrm{diag} \tilde{S}_\mu \right\rangle + m_1 \langle G_1 G_2 \rangle \langle \Im G_2 A \rangle \\
&+ m_1 \frac{1}{N} \sum_{\mu=1}^{N} \left\langle G_1 N \mathrm{diag} \tilde{S}_\mu^\circ \right\rangle \left\langle G_1 \Im G_2 A N \mathrm{diag} \tilde{S}_\mu \right\rangle + m_1 \langle G_1 - m_1 \rangle \langle G_1 \Im G_2 A \rangle
\end{aligned}
\tag{3.35}
$$

We use (3.17) to get

$$
\begin{aligned}
&\left| \left\langle G_1 G_2 N \mathrm{diag} \tilde{S}_\mu^\circ \right\rangle \left\langle \Im G_2 A N \mathrm{diag} \tilde{S}_\mu \right\rangle \right| \\
&\prec \frac{\Lambda_k}{L\sqrt{\eta_*}} \cdot \left| \left\langle (\Im G_2 - \Im m_2) A N \mathrm{diag} \tilde{S}_\mu \right\rangle + \Im m_2 \left\langle A N \mathrm{diag} \tilde{S}_\mu \right\rangle \right| \prec \frac{\Lambda_k \rho_2}{L\sqrt{\eta_*}}.
\end{aligned}
\tag{3.36}
$$

Using the bounds (3.24), (3.26), (3.27) and Lemma 2.10 we get the same self-consistent equation for $\langle G_1 \Im G_2 A \rangle$ as (3.29) with error term $\mathcal{O}\left(\frac{\rho_2 \Lambda_1}{L\sqrt{\eta_*}} \right)$ on the right. By inverting $I - C$ the same way we get

$$
\left| \left\langle G_1 \Im G_2 N \mathrm{diag} \tilde{S}_\mu^\circ \right\rangle \right| \prec \frac{\rho_2 \Lambda_1}{L\sqrt{\eta_*}}.
\tag{3.37}
$$

By plugging this into (3.35), we get (3.18).

The bound (3.19) is proved similarly. $\qquad\square$

Remark 3.4. As one can see from the proof above, the computation of the traces involving imaginary parts of one the Green's functions matrices involve more terms, but these terms can be analyzed in a manner that is very similar to those traces that do not involves the imaginary part. The most important point to realize is that the inclusion of the imaginary part causes the appearance of an extra factor of ρ. In most cases, this either uses the fact that $\langle \Im G \rangle = O(\rho)$ or that $\frac{1}{N\eta} = \frac{\rho}{L}$.

3.3 Bounds on $\langle GAGA \rangle$

Lemma 3.5. *For $A_1, A_2 \in \mathbb{M}_k^\circ$ we have*

$$
| \langle G_1 A_1 G_2 A_2 \rangle | \prec 1 + \frac{\Lambda_k^2}{\sqrt{L}} + \frac{\Lambda_1 \Lambda_k \Lambda_{k+1}}{\sqrt{NL}},
\tag{3.38}
$$

$$
| \langle G_1 A_1 \Im G_2 A_2 \rangle | \prec \rho_2 \left[1 + \frac{\Lambda_k^2}{\sqrt{L}} + \frac{\Lambda_1 \Lambda_k \Lambda_{k+1}}{\sqrt{NL}} \right],
\tag{3.39}
$$

and

$$
| \langle \Im G_1 A_1 \Im G_2 A_2 \rangle | \prec \rho_1 \rho_2 \left[1 + \frac{\Lambda_k^2}{\sqrt{L}} + \frac{\Lambda_1 \Lambda_k \Lambda_{k+1}}{\sqrt{NL}} \right].
\tag{3.40}
$$

Proof. We prove the first inequality here. The other two are proved similarly. See Remark 3.4 for details.

First, we use the identity

$$
\begin{aligned}
\langle G_1 A_1 G_2 A_2 \rangle = \; & m_1 m_2 \langle A_1 A_2 \rangle + m_1 \langle A_1 (G_2 - m_2 I) A_2 \rangle - m_1 \langle \underline{W G_1 A_1 G_2 A_2} \rangle \\
& + \frac{1}{N} m_1 \sum_{i,j} S_{ij} (G_1 - m_1 I)_{jj} (G_1 A_1 G_2 A_2)_{ii} \\
& + \frac{1}{N} m_1 \sum_{i,j} S_{ij} (G_1 A_1 G_2)_{jj} (G_2 A_2)_{ii}.
\end{aligned}
\tag{3.41}
$$

By splitting the terms on the last two lines, we get,

$$
\begin{aligned}
\langle G_1 A_1 G_2 A_2 \rangle = \; & m_1 m_2 \langle A_1 A_2 \rangle + m_1 \langle A_1 (G_2 - m_2 I) A_2 \rangle - m_1 \langle \underline{W G_1 A_1 G_2 A_2} \rangle \\
& + \frac{m_1}{N} \sum_{\mu} \left\langle N \mathrm{diag} \tilde{S}_{\mu}^{\circ} G_1 \right\rangle \left\langle G_1 A_1 G_2 A_2 N \mathrm{diag} \tilde{S}_{\mu} \right\rangle + m_1 \langle G_1 - m_1 \rangle \langle G_1 A_1 G_2 A_2 \rangle \\
& + \frac{m_1}{N} \sum_{\mu} \left\langle G_1 A_1 G_2 N \mathrm{diag} \tilde{S}_{\mu}^{\circ} \right\rangle \left\langle G_2 A_2 N \mathrm{diag} \tilde{S}_{\mu} \right\rangle + m_1 \langle G_1 A_1 G_2 \rangle \langle G_2 A_2 \rangle.
\end{aligned}
\tag{3.42}
$$

Now we plug in $A_2 = N \mathrm{diag} \tilde{S}_{\nu}^{\circ}$ into the identity above and get a system of equations.

$$
\begin{aligned}
\left\langle G_1 A_1 G_2 N \mathrm{diag} \tilde{S}_{\nu}^{\circ} \right\rangle = \; & m_1 m_2 \left\langle A_1 N \mathrm{diag} \tilde{S}_{\nu}^{\circ} \right\rangle + m_1 \left\langle A_1 (G_2 - m_2 I) N \mathrm{diag} \tilde{S}_{\nu}^{\circ} \right\rangle \\
& - m_1 \left\langle \underline{W G_1 A_1 G_2 N \mathrm{diag} \tilde{S}_{\nu}^{\circ}} \right\rangle + \frac{m_1}{N} \sum_{\mu} \left\langle N \mathrm{diag} \tilde{S}_{\mu}^{\circ} G_1 \right\rangle \left\langle G_1 A_1 G_2 N \mathrm{diag} \tilde{S}_{\nu}^{\circ} N \mathrm{diag} \tilde{S}_{\mu} \right\rangle \\
& + m_1 \langle G_1 A_1 G_2 \rangle \langle G_2 A_2 \rangle + \sum_{\mu} C_{\nu\mu} \left\langle G_1 A_1 G_2 N \mathrm{diag} \tilde{S}_{\mu}^{\circ} \right\rangle,
\end{aligned}
\tag{3.43}
$$

where

$$
\begin{aligned}
C_{\nu\mu} & = m_1 \langle G_1 - m_1 \rangle \delta_{\nu\mu} + m_1 \frac{1}{N} \left\langle G_2 N \mathrm{diag} \tilde{S}_{\nu}^{\circ} N \mathrm{diag} \tilde{S}_{\mu} \right\rangle \\
& = m_1 m_2 \left(S_{\nu\mu} - \frac{1}{N} \right) + \mathcal{O} \left(\frac{1}{N^{3/2} \eta_*^{1/2}} \right) + \mathcal{O} \left(\frac{1}{N \eta_*} \right) \delta_{\nu\mu}.
\end{aligned}
\tag{3.44}
$$

To bound the error terms in (3.43) we need the following lemma.

Lemma 3.6. *If $A \in \mathbb{M}_k^{\circ}$ and $B \in \mathbb{M}_l$, then*

$$
\begin{aligned}
|\langle G_1 A G_2 B \rangle| & \prec \Lambda_k \Lambda_l + \frac{\Lambda_k}{L \eta_*}, \\
|\langle \Im G_1 A G_2 B \rangle| & \prec \rho_1 \Lambda_k \Lambda_l + \frac{\rho_1 \Lambda_k}{L \eta_*}, \\
|\langle \Im G_1 A \Im G_2 B \rangle| & \prec \rho_1 \rho_2 \Lambda_k \Lambda_l + \frac{\rho_1 \rho_2 \Lambda_k}{L \eta_*}.
\end{aligned}
\tag{3.45}
$$

Proof. Let us divide the matrix $B = B^{\circ} + \langle B \rangle I$, where $B^{\circ} \in \mathbb{M}_l^{\circ}$ by the definition. Then, we have that

$$
\langle G_1 A G_2 B \rangle = \langle G_1 A G_2 B^{\circ} \rangle + \langle B \rangle \langle G_2 G_1 A \rangle.
\tag{3.46}
$$

The desired result follows from [14, (5.34)] and Lemma 3.3. $\qquad \square$

Then

$$
\left| \left\langle N \mathrm{diag} \tilde{S}_{\mu}^{\circ} G_1 \right\rangle \left\langle G_1 A_1 G_2 A_2 N \mathrm{diag} (\tilde{S}_{\mu}) \right\rangle \right| \prec \frac{\Lambda_1}{\sqrt{LN}} \left(\Lambda_k \Lambda_{k+1} + \frac{\Lambda_k}{L \sqrt{\eta_*}} \right),
\tag{3.47}
$$

and

$$
|\langle G_1 A_1 G_2 \rangle \langle G_2 A_2 \rangle| \prec \frac{\Lambda_k^2}{L^{3/2} \sqrt{N \eta_*}} \prec \frac{\Lambda_k^2}{L^2}.
\tag{3.48}
$$

From Lemma 2.10, we have

$$\left| \left\langle \underline{W G_1 A_1 G_2 N \mathrm{diag} \tilde{S}_\nu^\circ} \right\rangle \right| \prec \frac{\Lambda_k^2}{\sqrt{L}}. \tag{3.49}$$

Then by using the local law for $G_2 - m_2$, we have in general for diagonal A_1 and A_2,

$$|m_1 \left\langle A_1 (G_2 - m_2 I) A_2 \right\rangle| \prec \frac{1}{\sqrt{N \eta_2}} \prec \frac{1}{\sqrt{L}}. \tag{3.50}$$

This crucially used the fact that A_1 and A_2 are diagonal to get a simpler estimate. We specialize this estimate in the case that $A_2 = N \mathrm{diag}(\tilde{S}^\circ)$. Recall that we use $\mathrm{diag}(\tilde{S}^\circ)$ to denote the vector constructed by considering each $\mathrm{diag}(\tilde{S}_\mu^\circ)$.

Substituting all of these estimates in (3.43), we get

$$(I - C) \left\langle G_1 A_1 G_2 N \mathrm{diag} \tilde{S}^\circ \right\rangle = m_1 m_2 \left\langle A_1 N \mathrm{diag} \tilde{S}^\circ \right\rangle$$
$$+ \mathcal{O} \left(\frac{1}{\sqrt{L}} + \frac{\Lambda_k^2}{\sqrt{L}} + \frac{\Lambda_1 \Lambda_2 \Lambda_k}{\sqrt{NL}} + \frac{\Lambda_1 \Lambda_k}{L^2} \right). \tag{3.51}$$

By inverting matrix $I - C$ using a similar argument to the proof of Lemma 3.1, we get

$$\left| \left\langle G_1 A_1 G_2 N \mathrm{diag} \tilde{S}^\circ \right\rangle \right| \prec 1 + \frac{\Lambda_k^2}{\sqrt{L}} + \frac{\Lambda_1 \Lambda_2 \Lambda_k}{\sqrt{NL}}. \tag{3.52}$$

Now we bound $\langle G_1 A_1 G_2 A_2 \rangle$. To bound $\left\langle G_2 A_2 N \mathrm{diag} \tilde{S}_\mu \right\rangle$ we write $G_2 = (G_2 - m_2) + m_2$. Then since A_2 is diagonal, we have

$$\left| \left\langle G_2 A_2 N \mathrm{diag} \tilde{S}_\mu \right\rangle \right| \prec 1 + \frac{1}{\sqrt{N \eta_2}} \prec 1. \tag{3.53}$$

Then by using (3.52), we get

$$\left| \frac{m_1}{N} \sum_\mu \left\langle G_1 A_1 G_2 N \mathrm{diag} \tilde{S}_\mu \right\rangle \left\langle G_2 A_2 N \mathrm{diag} \tilde{S}_\mu \right\rangle \right| \prec 1 + \frac{\Lambda_k^2}{\sqrt{L}} + \frac{\Lambda_1 \Lambda_2 \Lambda_k}{\sqrt{NL}}. \tag{3.54}$$

Finally, by plugging in (3.54), (3.47), (3.48) and Lemma 2.10 into (3.42) and moving $\langle G_2 - m_2 \rangle \langle G_1 A_1 G_2 A_2 \rangle$ term to the left, we get

$$|\langle G_1 A_1 G_2 A_2 \rangle| \prec 1 + \frac{\Lambda_k^2}{\sqrt{L}} + \frac{\Lambda_1 \Lambda_k \Lambda_{k+1}}{\sqrt{NL}}. \tag{3.55}$$

$\square$

3.4 Bounds on $\langle \Im G A \Im G^t A \rangle$

Lemma 3.7. If $A_1, A_2 \in \mathbb{M}_k$, then

$$\left| \left\langle \Im G_1 A_1 \Im G_2^t A_2 \right\rangle \right| \prec \rho_1 \rho_2 \left(1 + \frac{\Lambda_1 \Lambda_{k+1}^2}{\sqrt{NL}} + \frac{\Lambda_{k+1}^2}{L} \right). \tag{3.56}$$

Proof. The imaginary part of the Green's function can be written as $\Im[G] = \frac{G - G^*}{2}$. Recall that we can write,

$$\begin{aligned} G &= mWG - m^2 G + mI \\ G^* &= \overline{m} W G^* - \overline{m}^2 G^* + \overline{m} I. \end{aligned} \tag{3.57}$$

By subtraction, we get,

$$\frac{G - G^*}{2} = mW \frac{G - G^*}{2} + \frac{m - \overline{m}}{2} W G^* + \frac{m - \overline{m}}{2} I - m^2 \frac{G - G^*}{2} - \frac{m^2 - \overline{m}^2}{2} G^*. \tag{3.58}$$

Thus, we have,

$$\Im G = mW\Im G + \Im mWG^* + \Im mI - m^2\Im G - \Im m(m + \overline{m})G^*. \qquad (3.59)$$

To derive an equation for $\langle \Im G_1 A_1 \Im G_2^t A_2 \rangle$, we substitute this expression for $\Im[G]_1$. First, we get,

$$\langle \Im G_1 A_1 \Im G_2^t A_2 \rangle = m_1 \langle W\Im G_1 A_2 \Im G_2^t A_2 \rangle + \Im m_1 \langle WG_1^* A_2 \Im G_2^t A_2 \rangle + \Im m_1 \langle A_1 \Im G_2^t A_2 \rangle$$
$$- m_1^2 \langle \Im G_1 A_1 \Im G_2^t A_2 \rangle - \Im m_1 (m_1 + \overline{m}_1) \langle G_1^* A_1 \Im G_2^t A_2 \rangle. \qquad (3.60)$$

From this expression, one does two things. The first is to write

$$\langle A_1 \Im G_2^t A_2 \rangle = \Im m_2 \langle A_1 A_2 \rangle + \langle A_1 (\Im G_2^t - m_2) A_2 \rangle.$$

The second is to perform the cumulant expansion for the two terms involving W. This will introduce the terms $m_1 \langle \underline{W\Im G_1 A_1 \Im G_2^t A_2} \rangle$ and $\Im m_1 \langle \underline{WG_1^* \Im G_2^t A_2} \rangle$. To perform the integration by parts, we write $\overline{\Im G = (GG^*)\Im z}$

For example, we have that,

$$\langle W\Im G_1 A_1 \Im G_2^t A_2 \rangle = \langle \underline{W\Im G_1 A_1 \Im G_2^t A_2} \rangle - \frac{1}{N}\sum_{i,j} S_{ij}(G_1)_{jj}(\Im z_1 G_1 G_1^* A_1 \Im G_2^t A_2)_{ii}$$

$$- \frac{1}{N}\sum_{i,j} S_{ij}(\Im z_1 G_1 G_1^*)_{jj}(G_1^* A_1 \Im G_2^t A_2)_{ii}$$

$$- \frac{1}{N}\sum_{i,j} S_{ij}(\Im G_1 A_1 (G_2^*)^t)_{ji}(\Im z_2 (G_2^*)^t G_2^t A_2)_{ji}$$

$$- \frac{1}{N}\sum_{i,j} S_{ij}(\Im G_1 A_1 \Im z_2 (G_2^*)^t G_2^t)_{ji}(G_2^t A_2)_{ji}. \qquad (3.61)$$

Here, now we split $S_{ij} = \sum_\mu \tilde{S}_{i\mu}\tilde{S}_{\mu j}$ for the last two terms. For the second term on the first line as well as the term on the second line, we instead use the decomposition, $S_{ij} = \sum_\mu \tilde{S}_{i\mu}\tilde{S}^\circ{}_{\mu j} + \frac{1}{N^2}$

Doing this, we see that,

$$\langle W\Im G_1 A_1 \Im G_2^t A_2 \rangle = \langle \underline{W\Im G_1 A_1 \Im G_2^t A_2} \rangle$$

$$- \frac{1}{N}\sum_\mu\sum_{ij} \tilde{S}^\circ_{j\mu}(G_1)_{jj}\tilde{S}_{i\mu}(\Im z_1 G_1 G_1^* A_1 \Im G_2^t A_2)_{ii}$$

$$- \frac{1}{N^2}\sum_{ij} (G_1)_{jj}(\Im z_1 G_1 G_1^* A_1 \Im G_2^t A_2)_{ii}$$

$$- \frac{1}{N}\sum_\mu\sum_{ij} S^\circ_{j\mu}(\Im z_1 G_1 G_1^*)_{jj} S_{i\mu}(G_1^* A_1 \Im G_2^t A_2)_{ii}$$

$$- \frac{1}{N^2}\sum_{i,j} (\Im z_1 G_1 G_1^*)_{jj}(G_1^* A_1 \Im G_2^t A_2)_{ii}$$

$$- \frac{1}{N}\sum_\mu\sum_{i,j} \tilde{S}_{i\mu}\tilde{S}_{j\mu}(\Im G_1 A_1 (G_2^*)^t)_{ji}(\Im z_2 (G_2^*)^t G_2^t A_2)_{ji} \qquad (3.62)$$

$$- \frac{1}{N}\sum_\mu\sum_{i,j} \tilde{S}_{i\mu}\tilde{S}_{j\mu}(\Im G_1 A_1 \Im z_2 (G_2^*)^t G_2^t)_{ji}(G_2^t A_2)_{ji}$$

$$= \langle \underline{W\Im G_1 A_1 \Im G_2^t A_2} \rangle - \frac{1}{N}\sum_\mu \left\langle N\mathrm{diag}\,\tilde{S}_\mu^\circ G_1 \right\rangle \left\langle \Im G_1 A_1 \Im G_2^t A_2 N\mathrm{diag}(\tilde{S}_\mu) \right\rangle$$

$$- \langle G_1 \rangle \langle \Im G_1 A_1 \Im G_2^t A_2 \rangle - \frac{1}{N} \sum_\mu \left\langle N \mathrm{diag} \tilde{S}_\mu^\circ \Im G_1 \right\rangle \left\langle G_1^* A_1 \Im G_2^t A_2 N \mathrm{diag}(\tilde{S}_\mu) \right\rangle$$

$$- \langle \Im G_1 \rangle \langle G_1^* A_1 \Im G_2^t A_2 \rangle - \frac{1}{N^2} \sum_\mu \left\langle \Im G_1 A_1 G_2^t N \mathrm{diag}(\tilde{S}_\mu) A_2^t \Im G_2 N \mathrm{diag}(\tilde{S}_\mu) \right\rangle$$

$$- \frac{1}{N^2} \sum_\mu \left\langle \Im G_1 A_1 \Im G_2^t N \mathrm{diag}(\tilde{S}_\mu) A_2^t G_2^* N \mathrm{diag}(\tilde{S}_\mu) \right\rangle$$

Combining all of these steps together we get the expression:

$$
\begin{aligned}
\left\langle \Im G_1 A_1 \Im G_2^t A_2 \right\rangle = {}& \Im m_1 \Im m_2 \left\langle A_1 A_2 \right\rangle + \Im m_1 \left\langle A_1 (\Im G_2^t - \Im m_2 I) A_2 \right\rangle \\
& - m_1 \left\langle W \Im G_1 A_1 \Im G_2^t A_2 \right\rangle - \Im m_1 \left\langle \overline{W G_1^* A_1 \Im G_2^t A_2} \right\rangle \\
& + \frac{m_1}{N} \sum_\mu \left\langle N \mathrm{diag} \tilde{S}_\mu^\circ G_1 \right\rangle \left\langle \Im G_1 A_1 \Im G_2^t A_2 N \mathrm{diag}(\tilde{S}_\mu) \right\rangle \\
& + \frac{m_1}{N} \sum_\mu \left\langle N \mathrm{diag} \tilde{S}_\mu^\circ \Im G_1 \right\rangle \left\langle G_1^* A_1 \Im G_2^t A_2 N \mathrm{diag}(\tilde{S}_\mu) \right\rangle \\
& + \frac{\Im m_1}{N} \sum_\mu \left\langle N \mathrm{diag} \tilde{S}_\mu^\circ G_1^* \right\rangle \left\langle G_1^* A_1 \Im G_2^t A_2 N \mathrm{diag}(\tilde{S}_\mu) \right\rangle \\
& + m_1 \left\langle G_1 - m_1 \right\rangle \left\langle \Im G_1 A_1 \Im G_2^t A_2 \right\rangle \\
& + m_1 \left\langle \Im G_1 - \Im m_1 \right\rangle \left\langle G_1^* A_1 \Im G_2^t A_2 \right\rangle \\
& + \Im m_1 \left\langle G_1^* - \overline{m_1} \right\rangle \left\langle G_1^* A_1 \Im G_2^t A_2 \right\rangle \\
& + \frac{m_1}{N^2} \sum_\mu \left\langle \Im G_1 A_1 G_2^t N \mathrm{diag}(\tilde{S}_\mu) A_2^t \Im G_2 N \mathrm{diag}(\tilde{S}_\mu) \right\rangle \\
& + \frac{m_1}{N^2} \sum_\mu \left\langle \Im G_1 A_1 \Im G_2^t N \mathrm{diag}(\tilde{S}_\mu) A_2^t G_2^* N \mathrm{diag}(\tilde{S}_\mu) \right\rangle \\
& + \frac{\Im m_1}{N^2} \sum_\mu \left\langle G_1^* A_1 G_2^t N \mathrm{diag}(\tilde{S}_\mu) A_2^t \Im G_2 N \mathrm{diag}(\tilde{S}_\mu) \right\rangle \\
& + \frac{\Im m_1}{N^2} \sum_\mu \left\langle G_1^* A_1 \Im G_2^t N \mathrm{diag}(\tilde{S}_\mu) A_2^t G_2^* N \mathrm{diag}(\tilde{S}_\mu) \right\rangle .
\end{aligned}
\tag{3.63}
$$

To get the right error terms, we can estimate each of these terms individually. Many of the bounds that we use, especially on terms involving the normalized traces of long sequences of products follow from estimates found in the paper [14] and one can find proofs and further estimates in said paper. Using Lemma 3.1, [14, (5.34), (5.35)], and (2.2), we get

$$\left| \left\langle N \mathrm{diag} \tilde{S}_\mu^\circ G_1 \right\rangle \left\langle \Im G_1 A_1 \Im G_2^t A_2 N \mathrm{diag}(\tilde{S}_\mu) \right\rangle \right| \prec \frac{\sqrt{\rho_1} \Lambda_1}{N \sqrt{\eta_1}} \cdot \rho_1 \rho_2 \Lambda_{k+1}^2 \prec \frac{\rho_1 \rho_2 \Lambda_1 \Lambda_{k+1}^2}{\sqrt{NL}},$$

$$\left| \left\langle \Im G_1 A_1 G_2^t N \mathrm{diag}(\tilde{S}_\mu) A_2^t \Im G_2 N \mathrm{diag}(\tilde{S}_\mu) \right\rangle \right| \prec \frac{\sqrt{\rho_1 \rho_2} \Lambda_k \Lambda_{k+1}}{\sqrt{\eta_1 \eta_2}} \prec \frac{\rho_1 \rho_2 \Lambda_{k+1}^2}{L},$$

$$\left| \Im m_1 \left\langle A_1 (\Im G_2^t - \Im m_2 I) A_2 \right\rangle \right| \prec \frac{\rho_1 \rho_2}{\sqrt{L}}.$$

Other terms are estimated similarly. Substituting these bounds into (3.63) gives the desired result. $\qquad\square$

3.5 Continuity argument for bounding Λ_1

For each value of E and J there is a unique value of η such that $N \eta \rho(E + i\eta) = J$. We let $F(E, J)$ be the unique η so that this is true.

We can now define the functions

$$
\begin{aligned}
G_A(J) &:= \max_{E_1, E_2 \in \{\gamma_a : 1 \leq a \leq N\}} \frac{\langle \Im G(E_1 + iF(E_1, J)) A \Im G(E_2 + iF(E_2, J)) A^* \rangle}{\rho_1(E_1 + iF(E_1, J))\rho_2(E_2 + iF(E_2, J))}, \\
G_A^t(J) &:= \max_{E_1, E_2 \in \{\gamma_a : 1 \leq a \leq N\}} \frac{\langle \Im G(E_1 + iF(E_1, J)) A \Im G(E_2 + iF(E_2, J))^t A^* \rangle}{\rho_1(E_1 + iF(E_1, J))\rho_2(E_2 + iF(E_2, J))}.
\end{aligned}
\tag{3.64}
$$

Lemma 3.8. *Fix some parameter $\epsilon > 0$ and let $J \geq N^\epsilon$. Uniformly in E_1, E_2, there exists a constant C such that,*

$$
\begin{aligned}
\left| \partial_J \frac{\langle \Im G(E_1 + iF(E_1, J)) A \Im G(E_2 + iF(E_2, J)) A^* \rangle}{\rho_1(E_1 + iF(E_1, J))\rho_2(E_2 + iF(E_2, J))} \right| &\leq N^C, \\
\left| \partial_J \frac{\langle \Im G(E_1 + iF(E_1, J)) A \Im G(E_2 + iF(E_2, J))^t A^* \rangle}{\rho_1(E_1 + iF(E_1, J))\rho_2(E_2 + iF(E_2, J))} \right| &\leq N^C.
\end{aligned}
\tag{3.65}
$$

Proof. First, let us find $\partial_J \eta$ at E. We have,

$$
\begin{aligned}
1 &= N\partial_J \eta (\rho + \eta \partial_\eta \rho), \\
\partial_J \eta &= \frac{1}{N(\rho + \eta \partial_\eta \rho)}.
\end{aligned}
\tag{3.66}
$$

Now, if $J \geq N^\epsilon$, using the fact that $\rho \leq 1$ implies $\eta \geq N^{-1+\epsilon}$. Furthermore, we would also know that $|\rho| \gtrsim \eta$ for $E \in [-2, 2]$.

We have the following integral expression for ρ.

$$
\begin{aligned}
\rho(E + i\eta) &= \int_{-2}^{2} \frac{\eta \rho_{sc}(x)}{(x - E)^2 + \eta^2} \mathrm{d}x, \\
\partial_\eta \rho(E + i\eta) &= \int_{-2}^{2} \frac{\rho_{sc}(x)((x - E)^2 - \eta^2)}{((x - E)^2 + \eta^2)^2} \mathrm{d}x, \\
\rho + \eta \partial_\eta \rho &= \int_{-2}^{2} \frac{2\eta \rho_{sc}(x)(x - E)^2}{((x - E)^2 + \eta^2)^2} \mathrm{d}x \gtrsim \eta.
\end{aligned}
\tag{3.67}
$$

Thus, $|\partial_J \eta| \leq N^{-\epsilon}$.

The above expression should also allow us to assert that $|\partial_\eta \rho| \leq \frac{1}{\eta^4} \leq N^4$.

$$
\max_{i,j} |\partial_\eta G_{ij}| = \max_{i,j} \left| \sum_\alpha \partial_\eta \left(\frac{1}{\lambda_\alpha - z} \right) u_\alpha(i) \overline{u}_\alpha(j) \right| \lesssim N^3.
\tag{3.68}
$$

By applying the product rule, this would imply equation (3.65). $\square$

Corollary 3.9. *For any matrix A with $\|A\| \leq 1$ and for any value of $N^\epsilon \leq J \leq N$, we have that*

$$
\begin{aligned}
G_A(J - N^{-C-2}) &\leq G_A(J) + N^{-2}, \\
G_A^t(J - N^{-C-2}) &\leq G_A^t(J) + N^{-2}.
\end{aligned}
\tag{3.69}
$$

Taking the union over $M \in \mathbb{M}_1$, which is a $O(N)$ family of matrices, and applying Lemma 2.7, we also have,

$$
\Lambda_1(J - N^{-C-2}) \prec \Lambda_1(J) + N^{-2}
\tag{3.70}
$$

Proof of Theorem 2.6 for Diagonal Matrices. By using Lemmas 2.7, 3.7, and 3.5, we would be able to derive the following relation:

$$
\Lambda_k^2(J) \prec 1 + \frac{\Lambda_1(J)\Lambda_{k+1}^2(J)}{\sqrt{J}}.
$$

That is, for any $\delta, D > 0$, there exists $N_0(\delta, D) > 0$ such that for all $N \geq N_0$,

$$\mathbb{P}\left(\Lambda_k^2(J) \leq N^\delta \left(1 + \frac{\Lambda_1(J)\Lambda_{k+1}^2(J)}{\sqrt{J}}\right)\right) \geq 1 - N^{-D}, \qquad (3.71)$$

for all $M \in \mathbb{M}_k$.

Let Ω_δ be the event that

$$\Lambda_k^2(J) \leq N^\delta \left(1 + \frac{\Lambda_1(J)\Lambda_{k+1}^2(J)}{\sqrt{J}}\right), \quad \Lambda_1(J - N^{-C-2}) \leq N^\delta \left(\Lambda_1(J) + N^{-2}\right)$$

holds for $k = 1, ..., \lceil 4/\epsilon \rceil$, $J = N - lN^{-C+2}$, $l \in \{0, \ldots, \lfloor N^{C-2} - N^{C-2+\epsilon} \rfloor\}$ and all $M \in \mathbb{M}_1$. Then there exists $N_0(\delta, D) > 0$ such that for all $N \geq N_0$, $\mathbb{P}(\Omega_\delta) \geq 1 - N^{-D}$.

This identity can be iterated to show that for $T = \lceil 4/\epsilon \rceil$,

$$\Lambda_1^2(J) \leq N^{\delta T}\left[1 + \sum_{l=1}^T \frac{(\Lambda_1^2(J))^l}{J^{l/2}} + \frac{(\Lambda_1^2(J))^{T+1}}{J^{(T+1)/2}}\Lambda_{T+1}^2(J)\right]. \qquad (3.72)$$

on Ω_δ.

Λ_{T+1} can be given the trivial bound η_*^{-1}, so this ultimately gives us,

$$\Lambda_1^2(J) \leq N^{\delta T}\left[1 + \sum_{l=1}^T \frac{(\Lambda_1^2(J))^l}{J^{l/2}} + \frac{(\Lambda_1^2(J))^{T+1}}{J^{(T+1)/2}}\frac{1}{\eta_*^2}\right]. \qquad (3.73)$$

Since $J \geq N^\epsilon$, if we choose $T = \lceil 8/\epsilon \rceil$, this implies that either $\Lambda_1^2(J) \leq (T+2)N^{\delta T}$ or $\Lambda_1^2(J) \geq N^{-\epsilon/4}\sqrt{J} \geq N^{\epsilon/4}$. Now we choose $\delta < \frac{\epsilon^2}{40}$. In more plain terms, the inequality established in equation (3.73) shows either that $\Lambda_1^2(J)$ is small or that it is relatively large.

Our ultimate goal is to show that $\Lambda_1^2(J)$ is small for J in some compact domain. On a high level, our argument is as follows. It is fairly straightforward to show that $\Lambda_1^2(J)$ is small when J is relatively small. Furthermore, on the event Ω_δ, we have a sort of continuity estimate on $\Lambda_1^2(J)$, which shows that if $\Lambda_1^2(J)$ is small, then $\Lambda_1^2(J - N^{-C-2})$ is not that much bigger. This piece of information is now combined with the dichotomy. The continuity estimate shows that $\Lambda_1^2(J - N^{-C-2})$ will not be bigger than the possible 'relatively large' lower bound in the dichotomy. Thus, $\Lambda_1^2(J - N^{-C-2})$ must be small as well. We write our argument more formally in what follows.

Our base case is to check that $\Lambda_1(N) \leq (T+2)N^{\delta T}$ with $T = \lceil 8/\epsilon \rceil$ and $\delta = \frac{\epsilon^2}{40}$. By direct computation from Definition 2.4, we see that $\Xi_M(N) = \frac{1}{2N}\text{tr}[MM^*]$, which will be $O(1)$ when $\|M\|$ is bounded. Thus, we see also that $\Lambda_1(N)$ is $O(1)$ and the base case is established.

Assume by induction for some l', we know that $\Lambda_1(N - l'N^{-C+2}) \leq (T+2)N^{\delta T}$.

On the event Ω_δ, we can assert that

$$\Lambda_1^2(N - (l'+1)N^{-C+2}) \leq N^\delta((T+2)N^{\delta T} + N^{-2}) \leq (T+3)N^{\delta(T+1)} < N^{\epsilon/4}.$$

Hence, we must have $\Lambda_1^2(N - (l'+1)N^{-C+2}) \leq (T+2)N^{\delta T}$ as well, due to the dichotomy that we have shown earlier. By induction, on Ω_δ we have

$$\Lambda_1(N^\epsilon) \leq (T+2)N^{\delta T}. \qquad \square$$

4 Proof of Theorem 2.6 for general M

4.1 Bounds on $\langle GA \rangle$

Lemma 4.1. *Let $A \in \mathbb{M}_k^\circ$. Then, we have that,*

$$|\langle GA \rangle| \prec \rho \left(\frac{\Lambda_k}{\sqrt{NL}} + \frac{\Lambda_{k+3}}{NL} \right). \tag{4.1}$$

Proof. At this point, we can return to an analysis for general matrices A. From the equation (3.7), we see that,

$$\langle GA \rangle[1 - \langle G - mI \rangle] = -m\langle \underline{WGA} \rangle + \frac{1}{N} \sum_\mu \langle GB_\mu \rangle \langle GN \mathrm{diag}(\tilde{S}_\mu^\circ) \rangle, \tag{4.2}$$

with $B_\mu := A(N\mathrm{diag}\tilde{S}_\mu)$. Using

$$\langle GB_\mu \rangle = m \langle B_\mu \rangle + \langle B_\mu \rangle \langle G - m \rangle + \langle GB_\mu^\circ \rangle,$$

and

$$|\langle GN\mathrm{diag}(\tilde{S}_\mu^\circ) \rangle| \prec \frac{\rho}{\sqrt{NL}}, \quad \langle B_\mu \rangle \prec 1, \quad |\langle G - m \rangle| \prec \frac{1}{N\eta}, \quad |\langle \underline{WGA} \rangle| \prec \frac{\rho\Lambda_k}{\sqrt{NL}},$$

we get

$$|\langle GA \rangle| \prec \frac{\rho}{\sqrt{NL}} \left(\Lambda_k + \max_\mu \langle GB_\mu^\circ \rangle \right). \tag{4.3}$$

By iterating this bound, we get

$$\sup_{A \in \mathbb{M}_k^\circ} |\langle GA \rangle| \prec \sum_{t=0}^{T} \left(\frac{\rho}{\sqrt{NL}} \right)^{t+1} \Lambda_{k+t} + \left(\frac{\rho}{\sqrt{NL}} \right)^{T+1} \Lambda_{k+T+1}. \tag{4.4}$$

If we take $T = 3$ and use the trivial bounds $\Lambda_{k+T+1} \prec \frac{1}{\eta}$, $\Lambda_k \geq 1$, we get

$$|\langle GA \rangle| \prec \sum_{t \geq 0}^{3} \frac{\Lambda_{k+t}\rho^{t+1}}{\sqrt{NL}^{t+1}} \prec \rho \frac{\Lambda_k}{\sqrt{NL}} + \rho \frac{\Lambda_{k+3}}{NL}. \tag{4.5}$$

$\square$

4.2 Bounds on $\langle GGA \rangle$

Lemma 4.2. *Let $A \in \mathbb{M}_k^\circ$. Then*

$$
\begin{aligned}
|\langle G_1 G_2 A \rangle| &\prec \frac{\Lambda_k}{L\sqrt{\eta_*}} + \frac{\Lambda_{k+4}}{L^2}, \\
|\langle \Im G_1 G_2 A \rangle| &\prec \rho_1 \frac{\Lambda_k}{L\sqrt{\eta_*}} + \rho_1 \frac{\Lambda_{k+4}}{L^2}, \\
|\langle \Im G_1 \Im G_2 A \rangle| &\prec \rho_1 \rho_2 \frac{\Lambda_k}{L\sqrt{\eta_*}} + \rho_1 \rho_2 \frac{\Lambda_{k+4}}{L^2}.
\end{aligned}
\tag{4.6}
$$

Proof. We prove the first inequality here. The other two are proved similarly. See Remark 3.4 for details.

From (3.23) we get

$$\langle G_1 G_2 A\rangle \left[1 - m_1 \langle G_1 - m_1\rangle\right] = m_1 \langle G_2 A\rangle - m_1 \underline{\langle W G_1 G_2 A\rangle}$$

$$+ \frac{1}{N} m_1 \sum_{\mu=1}^{N} \left\langle G_1 G_2 N \operatorname{diag}\tilde{S}_\mu^\circ \right\rangle \left\langle G_2 A N \operatorname{diag}\tilde{S}_\mu \right\rangle$$

$$+ m_1 \langle G_1 G_2\rangle \langle G_2 A\rangle$$

$$+ \frac{1}{N} m_1 \sum_{\mu=1}^{N} \left\langle G_1 N \operatorname{diag}\tilde{S}_\mu^\circ \right\rangle \left\langle G_1 G_2 A N \operatorname{diag}\tilde{S}_\mu \right\rangle .$$

Suppose $B = A N \operatorname{diag}\tilde{S}_\mu$ or $B = I$. We apply Cauchy-Schwarz to bound $\langle G_1 G_2 B\rangle$.

$$|\langle G_1 G_2 B\rangle| \leq \langle G_1 G_1^*\rangle^{\frac{1}{2}} \langle G_2 B B^* G_2^*\rangle^{\frac{1}{2}} \prec \frac{\sqrt{\rho_1 \rho_2}}{\sqrt{\eta_1 \eta_2}}. \tag{4.7}$$

Using this estimate, the estimate for $\langle GA\rangle$ from above and $\langle G_1 - m_1\rangle \prec \frac{1}{N\eta_1}$, we get

$$\left(1 + \mathcal{O}\left(\frac{1}{N\eta_1}\right)\right) \langle G_1 G_2 A\rangle = -m_1 \underline{\langle W G_1 G_2 A\rangle}$$

$$+ \frac{1}{N} m_1 \sum_{\mu=1}^{N} \left\langle G_1 G_2 N \operatorname{diag}\tilde{S}_\mu^\circ \right\rangle \left\langle G_2 A N \operatorname{diag}\tilde{S}_\mu \right\rangle \tag{4.8}$$

$$+ \mathcal{O}\left(\frac{\Lambda_k}{L\sqrt{\eta_*}} + \frac{\Lambda_{k+3}}{L^2}\right).$$

By Lemma 4.1,

$$\left|\left\langle G_2 A N \operatorname{diag}\tilde{S}_\mu \right\rangle\right| \prec 1 + \rho_2 \frac{\Lambda_{k+1}}{\sqrt{NL}} + \rho_2 \frac{\Lambda_{k+4}}{NL} \prec 1 + \rho_2 \frac{\Lambda_{k+4}}{\sqrt{NL}}. \tag{4.9}$$

Hence we get

$$|\langle G_1 G_2 A\rangle| \prec \frac{\Lambda_k}{L\sqrt{\eta_*}} + \frac{1}{L\sqrt{\eta_*}}\left(1 + \frac{\Lambda_{k+4}}{\sqrt{NL}}\right) + \frac{\Lambda_{k+3}}{L^2} \prec \frac{\Lambda_k}{L\sqrt{\eta_*}} + \frac{\Lambda_{k+4}}{L^2}. \qquad \square$$

4.3 Bounds on $\langle GAGA\rangle$

Lemma 4.3. *Suppose M is a traceless matrix with $\|M\| \prec 1$. For $A_1, A_2 \in \mathbb{M}_k^\circ$, we have*

$$|\langle G_1 A_1 G_2 A_2\rangle| \prec 1 + \frac{\Lambda_{k+4}^2}{\sqrt{L}},$$

$$|\langle \Im G_1 A_1 G_2 A_2\rangle| \prec \rho_1 \left(1 + \frac{\Lambda_{k+4}^2}{\sqrt{L}}\right), \tag{4.10}$$

$$|\langle \Im G_1 A_1 \Im G_2 A_2\rangle| \prec \rho_1 \rho_2 \left(1 + \frac{\Lambda_{k+4}^2}{\sqrt{L}}\right).$$

Proof. We prove the first inequality here. The other two are proved similarly. See Remark 3.4 for details.

The proof is similar to the proof of Lemma 3.5. The only difference is the size of the error terms.

Similarly to (3.42), we have

$$
\begin{aligned}
\langle G_1 A_1 G_2 A_2 \rangle = {} & m_1 m_2 \langle A_1 A_2 \rangle + m_1 \langle A_1 (G_2 - m_2 I) A_2 \rangle - m_1 \langle \underline{W G_1 A_1 G_2 A_2} \rangle \\
& + \frac{m_1}{N} \sum_\mu \left\langle N \mathrm{diag}\, \tilde{S}_\mu^\circ G_1 \right\rangle \left\langle G_1 A_1 G_2 A_2 N \mathrm{diag}\, \tilde{S}_\mu \right\rangle \\
& + m_1 \langle G_1 - m_1 \rangle \langle G_1 A_1 G_2 A_2 \rangle \\
& + \frac{m_1}{N} \sum_\mu \left\langle G_1 A_1 G_2 N \mathrm{diag}\, \tilde{S}_\mu^\circ \right\rangle \left\langle G_2 A_2 N \mathrm{diag}\, \tilde{S}_\mu \right\rangle \\
& + m_1 \langle G_1 A_1 G_2 \rangle \langle G_2 A_2 \rangle .
\end{aligned}
\tag{4.11}
$$

Now we plug in $A_2 = N \mathrm{diag}\, \tilde{S}_\nu^\circ$ into the identity above and get a system of equations.

$$
\begin{aligned}
\left\langle G_1 A_1 G_2 N \mathrm{diag}\, \tilde{S}_\nu^\circ \right\rangle = {} & m_1 m_2 \left\langle A_1 N \mathrm{diag}\, \tilde{S}_\nu^\circ \right\rangle + m_1 \left\langle A_1 (G_2 - m_2 I) N \mathrm{diag}\, \tilde{S}_\nu^\circ \right\rangle \\
& - m_1 \left\langle \underline{W G_1 A_1 G_2 N \mathrm{diag}\, \tilde{S}_\nu^\circ} \right\rangle + m_1 \langle G_1 A_1 G_2 \rangle \left\langle G_2 N \mathrm{diag}\, \tilde{S}_\nu^\circ \right\rangle \\
& + \frac{m_1}{N} \sum_\mu \left\langle N \mathrm{diag}\, \tilde{S}_\mu^\circ G_1 \right\rangle \left\langle G_1 A_1 G_2 N \mathrm{diag}\, \tilde{S}_\nu^\circ N \mathrm{diag}\, \tilde{S}_\mu \right\rangle \\
& + \sum_\mu C_{\nu\mu} \left\langle G_1 A_1 G_2 N \mathrm{diag}\, \tilde{S}_\mu^\circ \right\rangle ,
\end{aligned}
\tag{4.12}
$$

where

$$
\begin{aligned}
C_{\nu\mu} & = m_1 \langle G_1 - m_1 \rangle \delta_{\nu\mu} + m_1 \frac{1}{N} \left\langle G_2 N \mathrm{diag}\, \tilde{S}_\nu^\circ N \mathrm{diag}\, \tilde{S}_\mu \right\rangle \\
& = m_1 m_2 \left(S_{\nu\mu} - \frac{1}{N} \right) + \mathcal{O}\left(\frac{1}{N^{3/2} \eta_*^{1/2}} \right) + \mathcal{O}\left(\frac{1}{N \eta_1} \right) \delta_{\nu\mu}.
\end{aligned}
\tag{4.13}
$$

To bound the error terms in (4.12) we need the following lemma.

Lemma 4.4. *If $A \in \mathbb{M}_k^\circ$ and $B \in \mathbb{M}_l$, then*

$$
\begin{aligned}
|\langle G_1 A G_2 B \rangle| & \prec \Lambda_k \Lambda_l + \frac{\Lambda_k}{L \sqrt{\eta_*}} + \frac{\Lambda_{k+4}}{L^2}, \\
|\langle \Im G_1 A G_2 B \rangle| & \prec \rho_1 \Lambda_k \Lambda_l + \rho_1 \frac{\Lambda_k}{L \sqrt{\eta_*}} + \rho_1 \frac{\Lambda_{k+4}}{L^2}, \\
|\langle \Im G_1 A \Im G_2 B \rangle| & \prec \rho_1 \rho_2 \Lambda_k \Lambda_l + \rho_1 \rho_2 \frac{\Lambda_k}{L \sqrt{\eta^*}} + \rho_1 \rho_2 \frac{\Lambda_{k+4}}{L^2}.
\end{aligned}
\tag{4.14}
$$

Furthermore, for $B = N \mathrm{diag}\, \tilde{S}_\nu^\circ N \mathrm{diag}\, \tilde{S}_\mu$,

$$
\begin{aligned}
|\langle G_1 A G_2 B \rangle| & \prec \Lambda_k + \frac{\Lambda_k}{L \sqrt{\eta_*}} + \frac{\Lambda_{k+4}}{L^2}, \\
|\langle \Im G_1 A G_2 B \rangle| & \prec \rho_1 \Lambda_k + \rho_1 \frac{\Lambda_k}{L \sqrt{\eta_*}} + \frac{\Lambda_{k+4}}{L^2}, \\
|\langle \Im G_1 A \Im G_2 B \rangle| & \prec \rho_1 \rho_2 \Lambda_k + \rho_1 \rho_2 \frac{\Lambda_k}{L \sqrt{\eta_*}} + \rho_1 \rho_2 \frac{\Lambda_{k+4}}{L^2}.
\end{aligned}
\tag{4.15}
$$

Proof. Let us divide the matrix $B = B^\circ + \langle B \rangle I$, where $B^\circ \in \mathbb{M}_l^\circ$ by the definition. Then, we have that

$$
\langle G_1 A G_2 B \rangle = \langle G_1 A G_2 B^\circ \rangle + \langle B \rangle \langle G_2 G_1 A \rangle .
\tag{4.16}
$$

Now, (4.14) follows from [14, (5.34)], Lemma 4.2. Equation (4.15) follows from noting that $\Lambda_2 \prec 1$ for diagonal matrices like B, from the previous section. $\qquad\square$

Then

$$\left| \left\langle N \mathrm{diag} \tilde{S}_\mu^\circ G_1 \right\rangle \left\langle G_1 A_1 G_2 N \mathrm{diag} \tilde{S}_\nu^\circ N \mathrm{diag} \tilde{S}_\mu \right\rangle \right| \prec \frac{\rho_1}{\sqrt{NL}} \left(\Lambda_k + \frac{\Lambda_k}{L\sqrt{\eta_*}} + \frac{\Lambda_{k+4}}{L^2} \right). \quad (4.17)$$

and

$$\left| \left\langle G_1 A_1 G_2 \right\rangle \left\langle G_2 N \mathrm{diag} \tilde{S}_\nu^\circ \right\rangle \right| \prec \left(\frac{\Lambda_k}{L\sqrt{\eta_*}} + \frac{\Lambda_{k+4}}{L^2} \right) \frac{\rho_2}{\sqrt{NL}}. \quad (4.18)$$

Using Lemma 4.1, we get

$$\left| \left\langle A_1 (G_2 - m_2 I) N \mathrm{diag} \tilde{S}_\nu^\circ \right\rangle \right| = \left| \left\langle G_2 - m_2 \right\rangle \left\langle A_1 N \mathrm{diag} \tilde{S}_\nu^\circ \right\rangle + \left\langle (G_2 - m_2)(N \mathrm{diag} \tilde{S}_\nu^\circ A_1)^\circ \right\rangle \right|$$
$$\prec \frac{1}{N\eta_2} + \frac{\rho_2 \Lambda_{k+1}}{\sqrt{NL}} + \frac{\rho_2 \Lambda_{k+4}}{NL} \prec \frac{\rho_2}{L} + \frac{\rho_2 \Lambda_{k+4}}{\sqrt{NL}}. \quad (4.19)$$

From Lemma 2.10, we have

$$\left| \left\langle \underline{W G_1 A_1 G_2 N \mathrm{diag} \tilde{S}_\nu^\circ} \right\rangle \right| \prec \frac{\Lambda_k}{\sqrt{L}}. \quad (4.20)$$

Then from (4.12), we get

$$(I - C) \left\langle G_1 A_1 G_2 N \mathrm{diag} \tilde{S}^\circ \right\rangle = m_1 m_2 \left\langle A_1 N \mathrm{diag} \tilde{S}_\nu^\circ \right\rangle + \mathcal{O}_\prec \left(\frac{1}{L} + \frac{\Lambda_k}{\sqrt{L}} + \frac{\Lambda_{k+4}}{\sqrt{NL}} \right.$$
$$\left. + \frac{1}{\sqrt{NL}} \left(\Lambda_k + \frac{\Lambda_k}{L\sqrt{\eta_*}} + \frac{\Lambda_{k+4}}{L^2} \right) \right) \quad (4.21)$$
$$= m_1 m_2 \left\langle A_1 N \mathrm{diag} \tilde{S}_\nu^\circ \right\rangle + \mathcal{O}_\prec \left(\frac{\Lambda_{k+4}}{\sqrt{L}} \right).$$

By inverting matrix $I - C$ via a similar argument found in the proof of Lemma 3.1, we get

$$\left| \left\langle G_1 A_1 G_2 N \mathrm{diag} \tilde{S}^\circ \right\rangle \right| \prec 1 + \frac{\Lambda_{k+4}}{\sqrt{L}}. \quad (4.22)$$

We now return to our bound of $\langle G_1 A_1 G_2 A_2 \rangle$. At this point, we can substitute our bounds in equation (4.22) into equation (4.11).

We have some other error terms to deal with in (4.11). By using Lemma 4.4, we have,

$$\left| \left\langle N \mathrm{diag} \tilde{S}_\mu^\circ G_1 \right\rangle \left\langle G_1 A_1 G_2 A_2 N \mathrm{diag} \tilde{S}_\mu \right\rangle \right| \prec \frac{\rho_1}{\sqrt{NL}} \left(\Lambda_k \Lambda_{k+1} + \frac{\Lambda_k}{L\sqrt{\eta_*}} + \frac{\Lambda_{k+4}}{L^2} \right)$$
$$\prec \frac{\Lambda_{k+4}^2}{\sqrt{L}}. \quad (4.23)$$

We use (4.22) and Lemma 4.1 to get

$$\left| \left\langle G_1 A_1 G_2 N \mathrm{diag} \tilde{S}_\mu \right\rangle \left\langle G_2 A_2 N \mathrm{diag} \tilde{S}_\mu \right\rangle \right|$$
$$\prec \left| \left\langle G_1 A_1 G_2 N \mathrm{diag} \tilde{S}_\mu \right\rangle \right| \left(\langle G_2 \rangle \left| \left\langle A_2 N \mathrm{diag} \tilde{S}_\mu \right\rangle \right| + \left| \left\langle G_2 \left(A_2 N \mathrm{diag} \tilde{S}_\mu \right)^\circ \right\rangle \right| \right)$$
$$\prec \left(1 + \frac{\Lambda_{k+4}}{\sqrt{L}} \right) \left(1 + \frac{\Lambda_{k+1}}{\sqrt{NL}} + \frac{\Lambda_{k+4}}{NL} \right) \prec \left(1 + \frac{\Lambda_{k+4}}{\sqrt{L}} \right) \left(1 + \frac{\Lambda_{k+4}}{\sqrt{NL}} \right) \quad (4.24)$$
$$\prec 1 + \frac{\Lambda_{k+4}^2}{\sqrt{L}}.$$

Using Lemma 4.2 and Lemma 4.1, we get

$$\left| \langle G_1 A_1 G_2 \rangle \langle G_2 A_2 \rangle \right| \prec \left(\frac{\Lambda_k}{L\sqrt{\eta_*}} + \frac{\Lambda_{k+4}}{L^2} \right) \left(\frac{\Lambda_k}{\sqrt{NL}} + \frac{\Lambda_{k+3}}{NL} \right) \prec \frac{\Lambda_{k+4}^2}{L^2}. \quad (4.25)$$

Using Lemma 4.1, we get

$$|\langle A_1(G_2 - m_2 I)A_2\rangle| \leq |\langle G_2 - m_2\rangle \langle A_1 A_2\rangle| + |\langle (G_2 - m_2)(A_2 A_1)^\circ\rangle|$$
$$\prec \frac{1}{L} + \frac{\Lambda_{k+1}}{\sqrt{LN}} + \frac{\Lambda_{k+4}}{NL} \prec \frac{1}{L} + \frac{\Lambda_{k+4}^2}{\sqrt{NL}}. \tag{4.26}$$

From Lemma 2.10, we have

$$\left|\langle W G_1 A_1 G_2 A_2\rangle\right| \prec \frac{\Lambda_k^2}{\sqrt{L}}. \tag{4.27}$$

Now we use bounds (4.23), (4.24), (4.25), (4.26), (4.27) in (4.11), we have

$$\left[1 - \mathcal{O}_\prec\left(\frac{1}{N\eta_1}\right)\right]|\langle G_1 A_1 G_2 A_2\rangle| \prec 1 + \frac{\Lambda_{k+4}^2}{\sqrt{L}} \tag{4.28}$$

We can divide by $1 - \mathcal{O}_\prec\left(\frac{1}{N\eta_1}\right)$ on both sides to derive our result. □

4.4 Bounds on $\langle \Im G A \Im G^t A\rangle$

Lemma 4.5. *Suppose M is a diagonal traceless matrix with $\|M\| \prec 1$. If $A_1, A_2 \in \mathbb{M}_k$, then*

$$\left|\langle \Im G_1 A_1 \Im G_2^t A_2\rangle\right| \prec \rho_1\rho_2\left(1 + \frac{\Lambda_{k+3}^2}{\sqrt{L}}\right). \tag{4.29}$$

Proof. The proof is almost identical to the proof of Lemma 3.7. We start by using the identity (3.63) to express the normalized trace $\langle \Im G_1 A_1 \Im G_2^t A_2\rangle$ in terms of renormalized terms and traces of products of normalized traces involving a fairly small number of matrices. We then estimate each of the terms on the right hand side of (3.63) individually. As before, most of these estimates will be quoted from [14] and one could find full proofs of some of these technical estimates in said paper.

Like in the proof of Lemma 3.7 we use [14, (5.34), (5.35)] to get

$$\left|\left\langle N\text{diag}\tilde{S}_\mu^\circ \Im G_1\right\rangle \left\langle G_1^* A_1 \Im G_2^t A_2 N\text{diag}(\tilde{S}_\mu)\right\rangle\right| \prec \frac{\sqrt{\rho_1}}{N\sqrt{\eta_1}} \cdot \rho_2\Lambda_{k+1}^2 \prec \frac{\rho_1\rho_2\Lambda_{k+1}^2}{\sqrt{NL}},$$
$$\left|\left\langle \Im G_1 A_1 \Im G_2^t N\text{diag}(\tilde{S}_\mu) A_2^t G_2^* N\text{diag}(\tilde{S}_\mu)\right\rangle\right| \prec \frac{\rho_1\rho_2\Lambda_k\Lambda_{k+1}}{\sqrt{\eta_1\eta_2}} \prec \frac{\rho_1\rho_2\Lambda_{k+1}^2}{L},$$

The second term of (3.63) is estimated differently from the diagonal case:

$$|\Im m_1 \langle A_1(\Im G_2^t - \Im m_2 I)A_2\rangle| \leq \rho_1|\langle \Im G_2^t - \Im m_2\rangle \langle A_2 A_1\rangle| + \rho_1|\langle \Im G_2^t (A_2 A_1)^\circ\rangle|$$
$$\prec \frac{\rho_1\rho_2}{L} + \rho_1\rho_2\left(\frac{\Lambda_k}{\sqrt{NL}} + \frac{\Lambda_{k+3}}{NL}\right). \tag{4.30}$$

The bounds for the other terms in (3.63) can be obtained similarly. □

4.5 Proof of Main result

We now have enough results to prove our main Theorem 2.6.

Proof of Theorem 2.6. The result of Lemma 4.3 combined with Lemma 2.7 shows that,

$$\Lambda_k^2 \prec 1 + \frac{\Lambda_{k+4}^2}{\sqrt{J}}. \tag{4.31}$$

Starting from $k = 1$ and iterating this bound shows that,

$$\Lambda_1^2 \prec 1 + \frac{\Lambda_{1+4l}^2}{(\sqrt{J})^l}. \tag{4.32}$$

The important fact in this bound is that as l increases, the factor of $\sqrt{J}^l$ that appears in the denominator can become as large as one desires. Furthermore, we also know that Λ_{1+4l} can be trivially bounded by $\frac{1}{\eta}$. Thus, by setting l sufficiently large, we would be able to show $\Lambda_1 \prec 1$.

Our argument can be stated formally in the following way. We can choose $l = \lceil 8/\epsilon \rceil$ and apply the trivial bound $\Lambda_{1+4l} \prec \frac{1}{\eta}$ as well as $J \geq N^\epsilon$ to show that $\Lambda_1^2 \prec 1$. $\qquad \square$

5 Proof of Lemma 2.10

5.1 Cumulant expansion

In this section we use cumulant expansion to estimate the moments

$$\mathbb{E}|\langle \underline{WG_1B_1G_2B_2\ldots G_lB_l} \rangle|^{2p}. \tag{5.1}$$

Parts of this section were adapted from the paper [14].

For simplicity we assume that $B_i \in \mathbb{M}_k$ for all $i \in [N]$. It is easy to see from our proof that if B_i are from different families $\mathbb{M}_k$, each B provides the corresponding Λ in the bound.

For any $m, n \in \mathbb{Z}_+$ define a $N \times N$ matrix $\kappa^{m,n}$, such that its entries $\kappa_{ab}^{m,n}$ are the joint cumulants of m copies of w_{ab} and n copies of w_{ba}. Note that $\kappa^{1,1} = S$ and $\kappa^{2,0} = 0$.

We use the following cumulant expansion:

$$\mathbb{E}w_{ab}f(W) = \sum_{k=1}^{R} \sum_{m+n=k} \kappa_{ab}^{m+1,n} \mathbb{E}\partial_{ab}^m \partial_{ba}^n f(W) + \Omega_R, \tag{5.2}$$

where $\partial_{ab} = \partial_{w_{ab}}$.

Applying the expansion (5.2) to (5.1) $2p$ times with respect to each W allows us to express the moments (5.1) in terms of Feynman diagrams (Lemma 5.5). We understand that the following definition is quite long, but we will soon give an example that will make these concepts more concrete.

Definition 5.1. *Define the class of diagrams $\mathcal{G}$ as follows. Each diagram Γ is a graph with two types of vertices $V = V_\kappa \cup V_i$ that are called κ-vertices and internal vertices and two types of edges $E = E_\kappa \cup E_g$ called κ-edges and G-edges. For any vertex $v \in V$ its G-degree $d_g(v)$ is defined as its degree in the graph (V, E_g). Internal vertices $v \in V_i$ satisfy $d_g(v) = 2$ and κ-vertices can be partitioned $V_\kappa = \bigcup_{k \geq 2} V_\kappa^k$ according to their degree, i.e. $d_g(v) = k$ for $v \in V_\kappa^k$. κ-edges can be partitioned $e_\kappa = \bigcup_{k \geq 2} E_\kappa^k$ so that any $e \in E_\kappa^k$ connects two vertices from V_κ^k.*

Each κ-edge $e = (v, w)$ carries labels $r(e), s(e)$ and the value of $\kappa_{vw}^{r(e),s(e)}$. Each edge $e \in E_\kappa^2$ carries an additional label $h(e) \in \{mat, res\}$, which will record whether the edge comes from the derivative ∂_e hitting a matrix W or a resolvent G_k. Each G-edge e has labels $i(e), t(e), *(e) \in \{0,1\}$ recording the type of the resolvent e represents (imaginary part, transpose and adjoint respectively). Label $z(e)$ records the parameter of the resolvent. Labels $L(e)$ and $R(e)$ record deterministic matrices that resolvent is multiplied by.

Remark 5.2. In this paper $L(e)$ and $R(e)$ will be products of matrices B_k and $diag(\tilde{S}_\mu^\circ)$ defined in Assumption 2.1.
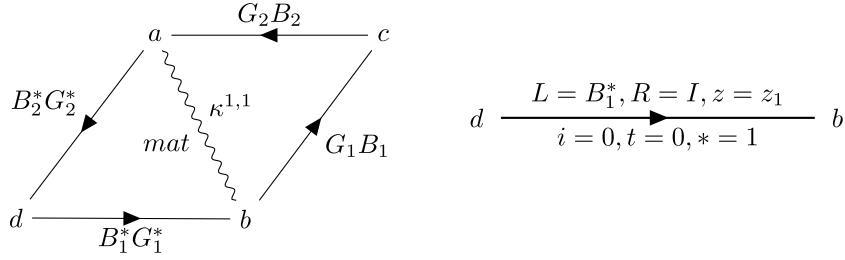
Figure 1: Diagram on the left corresponds to the first term of (5.4). On the right we show all labels of the edge $\{d, b\}$ with value $B_1^* G_1^*$.

In addition to the definition of diagrams, we also need to introduce the notion of values associated to each diagram. On an intuitive level, a diagram represents some product of matrix quantities organized in a particular way. The following definition formalizes the exact quantity associated to each diagram.

Definition 5.3. *For each* $\Gamma \in \mathcal{G}$ *and each* $e \in V_i$ *define the value* $\mathcal{G}^e$ *of the edge* e *as the resolvent* $L(e)G(z(e))R(e)$ *with imaginary part, transpose and* $*$ *applied according to the labels* $i(e), t(e), *(e)$. *For each* $e \in E_\kappa$ *define its value* $\mathcal{G}^e$ *as* $\kappa^{r(e), s(e)}$.

Define the value of the diagram Γ *as follows.*

$$\text{Val}(\Gamma) = \sum_{a_v \in [N], v \in V} \prod_{\{x,y\} \in E} \mathcal{G}_{a_x, a_y}^{\{x,y\}} \tag{5.3}$$

Here, we construct a few examples of the diagrams that appear after applying the cumulant expansion. Let us the consider the following terms,

$$
\begin{aligned}
\mathbb{E} \left| \langle \underline{W G_1 B_1 G_2 B_2} \rangle \right|^2 = {} & \mathbb{E} \sum_{a,b,c,d} \kappa_{a,b}^{1,1} (G_1 B_1)_{bc} (G_2 B_2)_{ca} (B_2^* G_2^*)_{ad} (B_1^* G_1^*)_{db} \\
& + \mathbb{E} \sum_{a,b,c,d} \kappa_{a,b}^{1,1} \kappa_{c,d}^{1,1} (G_1 B_1 G_2)_{bd} (G_2 B_2)_{ca} (B_2^* G_2^* B_1^* G_1^*)_{db} (G_1^*)_{ac} \\
& + \mathbb{E} \sum_{a,b,c,d} \kappa_{a,b}^{2,1} \kappa_{c,d}^{1,1} (G_1)_{bd} (G_1 B_1 G_2)_{ca} (G_2 B_2)_{ba} (B_2^* G_2^*)_{db} (G_2^* B_1^* G_1^*)_{ac} \\
& + \dots
\end{aligned}
\tag{5.4}
$$

Figure 1 shows the diagram corresponding to the first term of (5.4). Each edge has its value written next to it. On the right of Figure 1 we show the edge labels in more detail. Figure 2 shows the diagrams corresponding to the other two terms of (5.4).

The following definition is similar to the properties (P1)-(P8) in Proposition 5.3 of [14]. The main purpose of the definition is to encompass the properties of the graphs produced by cumulant expansion that are most important for our later counting bound. We remark that most of these properties are mechanical consequences of considering the algorithm of cumulant expansion.

Definition 5.4. *A diagram is said to be* $(l, p, i, \mathfrak{a}, \mathfrak{t})$-*regular if there exist a subset* V_o *of orthogonality vertices such that the following condition holds:*

1. *The graph* (V_κ, E_κ) *is a perfect matching.*

2. *The internal vertices satisfies* $|V_i| = 2(l-1)p$. *The edges satisfy* $1 \le |E_\kappa| \le 2p$, $\#\{e \in E_g : i(e) = 1\} = 2ip$, *and* $|E_g| = \sum_{e \in E_\kappa} d_g(e) + 2(l-1)p \ge 2p$.
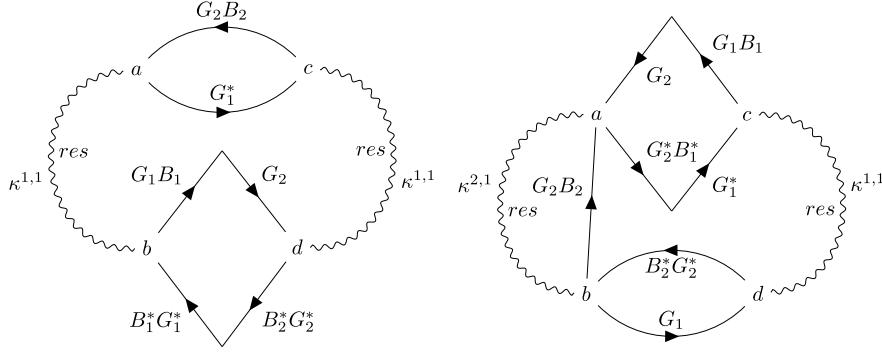
Figure 2: The diagrams corresponding to the second and third terms of (5.4).

3. *For any κ edge $(uv) \in E_\kappa^k$, the G-degrees of $u, v \in V_\kappa$ satisfy $d_g^{in}(u) = d_g^{out}(v)$, $d_g^{in}(v) = d_g^{out}(u)$, and $d_g(u) = d_g(v) \geq 2$. We can define the G-degree of (uv) as $d_g(uv) := d_g(u) = d_g(v) = k$.*

4. *Every E_g-cycle on $V_\kappa^2 \cup V_i$ contains at least two vertices in V_κ^2.*

5. *Denoting the number of isolated cycles in $(V_\kappa \cup V_i, E_g)$ with at most k vertices in V_o by $n_{cyc}^{o=k}$, we have $2n_{cyc}^{o=0} + n_{cyc}^{o=1} \leq 2|E_\kappa^2| - |V_o \cap V_\kappa^2|$.*

6. *$|V_i \cap V_o| = 2p(a + t - \mathbf{1}\{l \in \mathfrak{a} \cup \mathfrak{t}\})$.*

7. *If $l \in \mathfrak{a} \cup \mathfrak{t}$, then $2|E_\kappa^2| + |E_\kappa^{\geq 3}| - 2p \leq |V_o \cap V_\kappa^2| \leq 2p$, otherwise $V_o \cap V_\kappa^2$ is empty.*

8. *For any $\kappa^{1,1}$ edge e, the number of its endpoint in V_o is either 0 if $h(e) = mat$, or at most 1 if $h(e) = res$.*

With the notion of diagrams in hand, we can now describe the graph produced by cumulant expansion, which reduces the computations of moments of our renormalization terms to quantities defined on our $(l, p, i, \mathfrak{a}, \mathfrak{t})$ regular graphs.

Lemma 5.5 (Cumulant expansion). *For any $p \in \mathbb{N}$, there is a collection of graphs $\mathcal{G}_p^{av}$ such that*

$$\mathbb{E}| \langle \underline{W G_1 B_1 ... G_l B_l} \rangle |^{2p} = \sum_{\Gamma \in \mathcal{G}_p^{av}} \mathbb{E}\text{Val}(\Gamma) + O(N^{-2p}) . \tag{5.5}$$

Furthermore, for all diagrams in $\mathcal{G}_p^{av}$ are "$(l, p, i, \mathfrak{a}, \mathfrak{t})$-regular" in the sense of Definition 5.4.

Proof. This is the result of Proposition 5.3 in [14]. The only modification we make is the additional property 8 from Definition 5.4. This property holds because the vertices from V_κ^2 can only be selected as orthogonality vertices if they appear as a result of the derivative acting on a G (see (orth-2) in the proof of Proposition 5.3 in [14]). $\square$

In our proof we need to emphasize another property of the diagrams appearing in equation (5.5).

Lemma 5.6. *For any $\Gamma \in \mathcal{G}_p^{av}$ and any $\kappa^{1,1}$-edge e with $h(e) = res$, one of its endpoint has one outgoing G-edge e' with $L(e') = I$ and one incoming edge e'' with $R(e'') = I$.*

Proof. Suppose $e = (ab)$ comes from the cumulant expansion with respect to w_{ab} in $w_{ab}\langle \Delta^{ab}G_1 B_1 \dots G_l B_l \rangle$. Since $h(e) = $ res, ∂_{ba} hits a resolvent G_k, which becomes $G_k \Delta^{ba} G_k$. Then vertex b has an outgoing edge e' with resolvent G_1 and $L(e') = I$ and an incoming edge with resolvent G_k and $R(e) = I$. $\square$

Our final lemma computes the values of regular graphs along with the extra condition that $\kappa_{1,1} = \frac{1}{N}$. This lemma is from [14]. The reason we cannot apply this directly to our cumulant expansion is that the value of $\kappa^{1,1}$ we use is not uniform; our work in the next section is to modify the graphs so that we can reduce the computation of our graph values to those that appear in the following lemma.

Lemma 5.7. *If $\Gamma_{\kappa,G}$ is $(l, p, i, \mathfrak{a}, \mathfrak{t})$-regular, $\kappa_{xy}^{1,1} = \frac{1}{N}$, $|\kappa_{xy}^{p,q}| \leq CN^{-(p+q)/2}$, and $L(e), R(e) \in \mathbb{M}_k \cup \mathbb{M}_k^\circ$ for all G-edges e, then*

$$\mathrm{Val}(\Gamma_{\kappa,G}) \prec \begin{cases} \rho^{2(b+1)p} N^{2bp} L^{-2bp}, & b = l, \\ \Lambda_k^{2(a+t)p} \rho^{2ip \vee 2(b+1)p} N^{p(a+t+2b)} L^{-p(1+2b)}, & b < l, \end{cases} \tag{5.6}$$

where $b := l - a - t$.

5.2 The graph splitting procedure

We remark that there is only one difference between the case that we are considering here and the diagrams from [14]: they use the fact that the lowest order cumulants $\kappa^{1,1}$ are all uniformly $\frac{1}{N}$. This allows them to re-express some of the quantities related to the diagrams in terms of traces of matrix products. This is the key step that allows them to apply Proposition 5.6 to bound the value of the graph. In the absence of this important condition, what we must do is find a way to take our graphs into an expression that would be useful.

We find a procedure that takes any diagram Γ and re-expresses it as a sum of other diagrams. Namely,

$$\mathrm{Val}(\Gamma) = \sum_\mu \mathrm{Val}(\tilde{\Gamma}^\mu). \tag{5.7}$$

The details of our transformation will show that the diagrams $\tilde{\Gamma}^\mu$ are $(l, p, i, \mathfrak{a}, \mathfrak{t})$-regular diagrams along with the property that the new '$\kappa^{1,1}$' edges have value $\frac{1}{N}$. Formally, we find another way to write the covariance matrices $\kappa^{1,1}$ and incorporate these terms into one of the L or R matrices that multiply the G. The end result of this procedure is to formally treat the old $\kappa^{1,1}$ edges as having value $\frac{1}{N}$. We now begin to describe this procedure more formally.

For every $\kappa^{1,1}$ edge, we will decompose the diagram Γ into $N + 1$ further diagrams. Thus, if we let E_2 be the set of all $\kappa^{1,1}$ edges, we will decompose Γ into $(N + 1)^{|E_2|}$ edges. The graph splitting procedure essentially treats every edge independently, so to describe the construction, it is best to consider the case that there is only a single $\kappa^{1,1}$ edge.

First, consider a $\kappa^{1,1}$ edge $e = (x, y)$. In case $h(e) = res$, we know from Lemma 5.6 that one of x or y has the property that it has a single incoming edge with $R(e) = I$ and a single outgoing edge with $L(e) = I$. Assume that this vertex is x. When we split $\kappa^{1,1}$ later, then this property allow us to introduce a trace 0 matrix in a location that is useful for cancellations. We remark that by Definition 5.4 (1) if there were more $\kappa^{1,1}$ edges, this vertex x would not be shared with the other $\kappa^{1,1}$ edges. In case $h(e) = mat$, x can be chosen arbitrarily among the two vertices adjacent to the $\kappa^{1,1}$ edge.

We have the following matrix decomposition of $S_{xy} := \kappa_{x,y}^{1,1}$.

$$S_{xy} = \sum_\mu \tilde{S}_{x\mu} \tilde{S}_{\mu y}. \tag{5.8}$$

Though this procedure formally introduces the vertex μ, these vertices will not be introduced into our diagram. Instead, the first N of $N + 1$ diagrams will be indexed by this vertex μ; we call these diagrams $\tilde{\Gamma}^1, \ldots, \tilde{\Gamma}^N$. The remaining graph we will call $\tilde{\Gamma}^{\mathrm{ext}}$, and will be derived via a more complicated resummation procedure.

After fixing a value of μ, we still have to perform more manipulation in order to derive the diagrams $\tilde{\Gamma}^\mu$. This procedure is quite similar to those performed in Sections 3 and 4; namely, we further split $\tilde{S}_{x\mu}$ and interpret this as a trace 0 matrix. Recall the traceless part as follows:

$$\tilde{S}^\circ_{x\mu} = \tilde{S}_{x\mu} - \frac{1}{N}$$
$$\mathrm{diag}\tilde{S}^\circ_\mu = \mathrm{diag}\tilde{S}_\mu - \left\langle \mathrm{diag}\tilde{S}_\mu \right\rangle = \mathrm{diag}\tilde{S}_\mu - \frac{1}{N}. \tag{5.9}$$

We define the diagram $\tilde{\Gamma}^\mu$ as follows. We take the diagram Γ and redefine labels so that for the incoming edge to vertex x, we multiply R on the right by $N\mathrm{diag}\tilde{S}^\circ_\mu$. At y, we change the label so that at the incoming edge to y, we multiply R on the right by $N\mathrm{diag}\tilde{S}_\mu$. Finally, we change the value of $\kappa^{1,1}$ formally to $\frac{1}{N}$ on the edge $\{x,y\}$.

In case $h(e) = res$, the main benefit is that even though we remove the orthogonality at vertex y, vertex x becomes a trace zero orthogonality vertex. Thus, there is no net loss in the number of orthogonality vertices in all of the diagrams $\tilde{\Gamma}^\mu$ for μ between 1 and N. Moreover, the traceless matrix involved in this new orthogonality vertex is in $\mathbb{M}^\circ_0$.

In case $h(e) = mat$, neither x nor y are orthogonality vertices of Γ, so our procedure doesn't affect V_o.

Finally, we define our final diagram $\tilde{\Gamma}^{\mathrm{ext}}$ by taking the diagram Γ and changing the label $\kappa^{1,1}$ to $\frac{1}{N}$ at the edge $\{x,y\}$. It is easy to check that

$$\mathrm{Val}(\Gamma) = \frac{1}{N} \sum_{\mu=1}^{N} \mathrm{Val}(\tilde{\Gamma}^\mu) + \mathrm{Val}(\tilde{\Gamma}^{\mathrm{ext}}). \tag{5.10}$$

Here is an example of the graph splitting procedure. We only draw the vertices and edges that connect to the $\kappa^{1,1}$ edge.
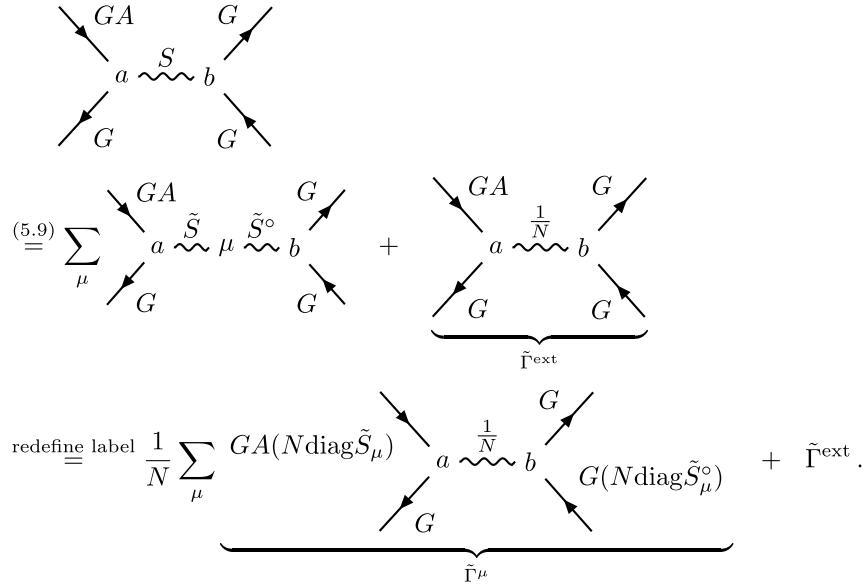
Lemma 5.8. *The diagrams $\tilde{\Gamma}^\mu$ or $\tilde{\Gamma}^{\mathrm{ext}}$ that have been constructed are $(l,p,i,\mathfrak{a},\mathfrak{t})$-regular diagrams.*

Proof. Let $e = (a,b)$ be the single $\kappa^{1,1}$ edge in the original diagram Γ.

The only modifications that we have to the diagrams $\tilde{\Gamma}^i$ from the original diagram Γ are in the labels of the edges which are adjacent to the vertices (a,b). Of these two

vertices, let b be the one that is multiplied by two pure G matrices on either side. In addition, vertices a and b are our only possible changes for V_o. From this property, it is clear to see that properties 1-4 of Definition 5.4 still hold after the modification. Furthermore, property 5 of Definition 5.4 is a consequence of property 4 regardless of the choice of V_o. This is a simple counting argument.

With regards to the sets of vertices V_o, we describe a bit more elaborately what changes are made. Due to property 8 of Definition 5.4, we know that at most one of x and y is in the set V_o. Recall that we have earlier chosen the vertex b to lie between the product of two pure G matrices without any intermittent matrix product. Once we multiply in between the two pure G matrices at b by the matrix $\text{diag}(\tilde{S}^\circ)_\mu$, we see that b can be an orthogonality matrix due to the multiplication of the additional trace 0 matrix. Regardless of whether $a \in V_o$ or $b \in V_o$ in the graph Γ, we put the vertex b into V_o for $\tilde{\Gamma}^i$. Thus, the cardinality of any the sets V_o and $V_o \cap V_\kappa^2$ do not change. Thus, properties 6-8 of Definition 5.4 still hold. $\qquad\square$

The previous discussion has established the following lemma.

Lemma 5.9. *Every $\Gamma \in \mathcal{G}_p$ can be split into a set of diagrams, $\widetilde{\Gamma}^{(\mu_1,\dots,\mu_m)}$ for $\mu_i \in [n] \cup \{0\}, 1 \le i \le m$ with m being the number of $\kappa^{1,1}$ edges in Γ, in the sense that*

$$\text{Val}(\Gamma_{\kappa,G}) = \sum_{\mu_i \in [n] \cup \{\text{ext}\}, 1 \le i \le m} \left(\frac{1}{N}\right)^{\#\{i:\mu_i \ne \text{ext}\}} \text{Val}(\widetilde{\Gamma}^{(\mu_1,\dots,\mu_m)}_{\kappa',G}), \qquad (5.11)$$

where $\kappa'^{1,1} = \frac{1}{N}$ and $\kappa'^{p,q} = \kappa^{p,q}$ for $(p,q) \ne (1,1)$. Furthermore, all of the diagrams $\widetilde{\Gamma}^{(\mu_1,\dots,\mu_m)}_{\kappa',G}$ are $(l,p,i,\mathfrak{a},\mathfrak{t})$-regular.

Proof of Lemma 2.10. The high probability estimates on $\langle \underline{WG_1A_1G_2A_2} \rangle$ and other similar quantities are readily derived by computing high moments and applying Markov's inequality. The cumulant expansion Lemma 5.5 gives an expression of the moments in terms of values of graphs. The values of these graphs are determined by the splitting procedure in Lemma 5.9 and the evaluation of graph values in appropriate conditions as in Lemma 5.7. The combinatorics of the sum over graphs is exactly the same as that of [14] and we can derive the exact same estimates. $\qquad\square$

References

[1] L. Benigni. Eigenvectors distribution and quantum unique ergodicity for deformed Wigner matrices. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(4):2822 – 2867, 2020. MR4164858

[2] L. Benigni and G. Cipolloni. Fluctuations of eigenvector overlaps and the berry conjecture for wigner matrices. *arXiv preprint arXiv:2212.10694*, 2022.

[3] L. Benigni and P. Lopatto. Fluctuations in local quantum unique ergodicity for generalized wigner matrices. *arXiv:2103.12013*, 2021. MR4397177

[4] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.*, 19(33):1–53, 2014. MR3183577

[5] O. Bohigas, M. Giannoni, and C. Schmid. Characterization of chaotic quantum spectra and universality of level fluctuation laws. *Phys. Rev. Lett.*, 52(1-4), 1984. MR0730191

[6] P. Bourgade, L. Erdős, H.-T. Yau, and J. Yin. Fixed energy universality for generalized Wigner matrices. *Communications on Pure and Applied Mathematics*, 69(10):1815–1881. MR3541852

[7] P. Bourgade, L. Erdős, H.-T. Yau, and J. Yin. Universality for a class of random band matrices. *Advances in Theoretical and Mathematical Physics*, 21(3):739–800, 2017. MR3695802

[8] P. Bourgade and H.-T. Yau. The eigenvector moment flow and local quantum unique ergodicity. *Communications in Mathematical Physics*, 350(1):231–278, 2017. MR3606475

[9] P. Bourgade, H.-T. Yau, and J. Yin. Random band matrices in the delocalized phase, I: Quantum unique ergodicity and universality. *Communications on Pure and Applied Mathematics*, 73(7):1526–1596, 2020. MR4156609

[10] E. Brézin and S. Hikami. Correlations of nearby levles induced by a random potential. *Nucl. Phys. B*, 476:697–706, 1996. MR1418841

[11] E. Brézin and S. Hikami. Spectral form factor in a random matrix theory. *Phys. Rev. E*, 55:4067–4083, 1997. MR1449379

[12] G. Cipolloni, L. Erdős, J. Henheik, and O. Kolupaiev. Gaussian fluctuations in the equipartition principle for wigner matrices. *arXiv preprint arXiv:2301.05181*, 2023. MR4634106

[13] G. Cipolloni, L. Erdős, J. Henheik, and D. Schröder. Optimal lower bound on eigenvector overlaps for non-hermitian random matrices. *arXiv preprint arXiv:2301.03549*, 2023. MR4748758

[14] G. Cipolloni, L. Erdős, and D. Schröder. Eigenstate thermalization hypothesis for wigner matrices. *Communications in Mathematical Physics*, 388(2):1005–1048, 2021. MR4334253

[15] G. Cipolloni, L. Erdős, and D. Schröder. Optimal multi-resolvent local laws for wigner matrices. *Electronic Journal of Probability*, 27:1–38, 2022. MR4479913

[16] G. Cipolloni, L. Erdős, and D. Schröder. Rank-uniform local law for wigner matrices. In *Forum of Mathematics, Sigma*, volume 10, page e96. Cambridge University Press, 2022. MR4502022

[17] G. Cipolloni, L. Erdős, and D. Schröder. Normal fluctuation in quantum ergodicity for Wigner matrices. *The Annals of Probability*, 50(3):984 – 1012, 2022. MR4413210

[18] J. Deutsch. Quantum statistical mechanics in a closed system. *Phys. Rev. A*, 43:2046–2049, 1991.

[19] J. Deutsch. Eigenstate thermalization hypothesis. *Rep. Prog. Phys.*, 81, 2018. MR3829412

[20] L. Erdős, T. Kruger, and D. Schroder. Random matrices with slow correlation decay. *Forum Math Sigma*, 7, 2019. MR3941370

[21] L. Erdős, T. Kruger, and D. Schroder. Cusp universality for random matrices i: local law and the complex hermitian case. *Comm. Math Phys*, 378, 2020. MR4134946

[22] L. Erdős, S. Péché, J. A. Ramírez, B. Schlein, and H.-T. Yau. Bulk universality for Wigner matrices. *Comm. Pure Appl. Math.*, 63(7):895–925. MR2662426

[23] L. Erdős, B. Schlein, and H.-T. Yau. Universality of random matrices and local relaxation flow. *Invent. Math.*, 185(1):75– 119. MR2810797

[24] L. Erdős, B. Schlein, and H.-T. Yau. Local semicircle law and complete delocalization for Wigner random matrices. *Commun. Math. Phys.*, 287(2):641–655, 2008. MR2481753

[25] L. Erdős and H.-T. Yau. Gap universality of generalized wigner and beta ensembles. *J. Eur. Math. Soc.*, 17:1927– 2036. MR3372074

[26] L. Erdős and H.-T. Yau. Universality of local spectral statistics of random matrices. *Bull. Amer. Math. Soc. (N.S.)*, 49(3):377– 414. MR2917064

[27] L. Erdős, H.-T. Yau, and J. Yin. Bulk universality for generalized Wigner matrices. *Probab. Theory Related Fields*, 154(1-2):341– 407, 2012. MR2981427

[28] L. Erdős, H.-T. Yau, and J. Yin. Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.*, 229(3):1435– 1515, 2012. MR2871147

[29] L. Erdős, G. Cipolloni, and D. Schröder. Functional central limit theorems for wigner matrices. 12 2020.

[30] A. Knowles and J. Yin. Eigenvector distribution of wigner matrices. *Probab. Theory Related Fields*, 155(3):543–582. MR3034787

[31] A. Knowles and J. Yin. The isotropic semicircle law and deformation of Wigner matrices. *Comm. Pure Appl. Math.*, 66:1663–1749. MR3103909

[32] J. Marcinek and H.-T. Yau. High dimensional normality of noisy eigenvectors. *arXiv:2005.08425*, 2020. MR4272266

[33] Z. Rudnick and P. Sarnak. The behaviour of eigenstates of arithmetic hyperbolic manifolds. *Comm. Math. Phys.*, 161(1):195–213. MR1266075

[34] M. Srednicki. Chaos and quantum thermalization. *Phys. Rev. E*, 50:888–901, 1994.

[35] T. Tao and V. Vu. Random matrices: universality of local eigenvalue statistics. *Acta Math.*, 206(1):127–204, 2011. MR2784665

[36] E. P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564, 1955. MR0077805

[37] C. Xu, F. Yang, H.-T. Yau, and J. Yin. Bulk universality and quantum unique ergodicity for random band matrices in high dimensions. *arXiv:2207.14533*.

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards

- Free for authors, free for readers

- Quick publication (no backlog)

- Secure publication (LOCKSS[1])

- Easy interface (EJMS[2])

Economical model of EJP-ECP

- Non profit, sponsored by IMS[3], BS[4], ProjectEuclid[5]

- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)

- Submit your best articles to EJP-ECP

- Choose EJP-ECP over for-profit journals

---

[1]LOCKSS: Lots of Copies Keep Stuff Safe http://www.lockss.org/
[2]EJMS: Electronic Journal Management System: https://vtex.lt/services/ejms-peer-review/
[3]IMS: Institute of Mathematical Statistics http://www.imstat.org/
[4]BS: Bernoulli Society http://www.bernoulli-society.org/
[5]Project Euclid: https://projecteuclid.org/
[6]IMS Open Access Fund: https://imstat.org/shop/donation/