

Multi-point functional central limit theorem for Wigner matrices*

Jana Reker^{†‡}

Abstract

Consider the random variable $\text{Tr}(f_1(W)A_1 \dots f_k(W)A_k)$ where W is an $N \times N$ Hermitian Wigner matrix, $k \in \mathbb{N}$, and choose (possibly N -dependent) regular functions f_1, \dots, f_k as well as bounded deterministic matrices A_1, \dots, A_k . We give a functional central limit theorem showing that the fluctuations around the expectation are Gaussian. Moreover, we determine the limiting covariance structure and give explicit error bounds in terms of the scaling of f_1, \dots, f_k and the number of traceless matrices among A_1, \dots, A_k , thus extending the results of [14] to products of arbitrary length $k \geq 2$. As an application, we consider the fluctuation of $\text{Tr}(e^{itW}A_1e^{-itW}A_2)/N$ around its thermal value $\text{Tr}(A_1)\text{Tr}(A_2)/N^2$ when t is large and give an explicit formula for the variance.

Keywords: Wigner matrix; central limit theorem; fluctuations; thermalization.

MSC2020 subject classifications: 60B20; 15B52.

Submitted to EJP on February 22, 2024, final version accepted on November 26, 2024.

Supersedes arXiv:2307.11028.

1 Introduction

The eigenvalues $\{\lambda_j\}_{j=1}^N$ of a large $N \times N$ Hermitian random matrix W constitute a strongly correlated system of random points on the real line. Due to the strong dependence, classical central limit theorems (CLTs) aimed at independent or weakly dependent random variables do not apply. However, the linear statistics $\text{Tr} f(W) = \sum_{j=1}^N f(\lambda_j)$ with a regular test function $f : \mathbb{R} \rightarrow \mathbb{R}$ have a variance of order one (see [31]) and, in fact, satisfy a central limit theorem with a Gaussian limit, as shown, e.g., in [32] for the Wigner case and in [30] for invariant ensembles, see also [47, 48]. Remarkably, the effect of the dependent random variables only manifests in the anomalous scaling, and removing the classical $N^{-1/2}$ prefactor fully compensates for the correlations. We

*Partially supported by ERC Advanced Grants “RMTBeyond” No. 101020331 and “LDRaM” No. 884584.

[†]IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria.

[‡]Current address: ENS de Lyon UMPA, 46 Allée d’Italie, 69007 Lyon, France.

E-mail: jana.reker@ens-lyon.fr

emphasize that this question is well-studied for Wigner matrices, see, e.g., [25, 3, 37, 43, 50, 4, 33] and that recent work by Diaz and Mingo [19] establishes a CLT for a large class of random matrix models and expresses the limiting covariance structure in terms of a Fréchet integral.

Note that the information obtained from a CLT is twofold: It characterizes the fluctuations of the linear statistics around its mean as Gaussian and simultaneously identifies the limiting variance or, more generally, the limiting covariance structure. To generalize the CLT for $\sum_j f(\lambda_j)$, the linear statistics can be modified in different ways. First, one may replace the N -independent function f by a function of the build

$$f(x) = g(N^\gamma(x - E)) \tag{1.1}$$

where g is a regular N -independent function, $E \in \mathbb{R}$ lies in the limiting spectrum of W , and $N^{-\gamma}$ is larger than the typical eigenvalue spacing around E . Considering the linear statistics for a function f that is concentrated around a value E on a mesoscopic scale allows us to zoom into the spectrum and thus study the problem locally. For Wigner matrices, this problem was studied by He and Knowles in [27, 26, 28], yielding a tracial CLT for the bulk spectrum that spans the entire mesoscopic regime. Similar questions have also been studied for other models, including deformed Wigner matrices [29, 34], generalized Wigner [35], and Wigner-type [41] matrices, sample covariance matrices [2, 34], Haar distributed random matrices on the classical compact groups [45, 46, 49], β -ensembles [8, 44, 7, 6, 24, 9], free sums [5], and non-Hermitian random matrices [10, 20]. See also [14] and references therein for a discussion of further examples and previous results.

The second generalization addresses that $\sum_{j=1}^N f(\lambda_j)$ is inherently *tracial*, i.e., the statistics only involve the eigenvalues of the random matrix, but not its eigenvectors. By testing $f(W)$ against a bounded *deterministic* matrix A with $\|A\| \leq 1$, i.e., by modifying the centered statistics to the form

$$\text{Tr}[f(W)A] - \mathbb{E} \text{Tr}[f(W)A] = \sum_{j=1}^N f(\lambda_j) \langle \mathbf{u}_j, A\mathbf{u}_j \rangle - \mathbb{E}[\dots], \tag{1.2}$$

the normalized eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_N$ of W enter into the problem. In the Wigner case, Lytova [36] obtained a CLT for (1.2) on macroscopic scales including an explicit formula for the limiting variance. We refer to the CLTs that also involve eigenvectors as *functional* in contrast to the tracial CLTs above. The recent paper [14] extended these results to all mesoscopic scales and further established that decomposing the matrix A in (1.2) according to

$$A = \langle A \rangle \text{Id} + \mathring{A}_d + \mathring{A}_{od}, \quad \langle A \rangle := \frac{1}{N} \text{Tr} A,$$

gives rise to three asymptotically independent fluctuation modes. Here, Id denotes the identity matrix, and \mathring{A}_d and \mathring{A}_{od} denote the diagonal and off-diagonal components of $\mathring{A} = A - \langle A \rangle \text{Id}$, the traceless part of A , respectively. Moreover, the results in [14] show that the modes corresponding to the tracial and traceless part of A fluctuate on different scales in the mesoscopic regime and two modes of the build (1.2) are asymptotically independent if the involved functions live on different scales.

In this work, we study a third generalization of the original linear statistics $\sum_{j=1}^N f(\lambda_j)$, which extends (1.2) from involving one (possibly N -dependent, mesoscopically scaled) function of W and one (possibly traceless) bounded deterministic matrix to alternating products involving $k \in \mathbb{N}$ functions and bounded deterministic matrices, respectively. More precisely, we consider the fluctuation of the statistics

$$Y := \langle f_1(W)A_1 f_2(W)A_2 \dots f_k(W)A_k \rangle - \mathbb{E} \langle f_1(W)A_1 f_2(W)A_2 \dots f_k(W)A_k \rangle, \tag{1.3}$$

show that Y satisfies a CLT with a Gaussian limit and give the limiting covariance structure as well as explicit error estimates. This generalizes [14, Thm. 2.4] to arbitrary $k \geq 1$. We refer to the result as a *multi-point functional CLT*. Similar to the results in [14], we further verify that two modes are asymptotically independent if the functions f_j are rescaled to different scales or around different numbers E_j via (1.1). However, while the $k = 1$ case only allows for two relevant classes of deterministic matrices (corresponding to the tracial and traceless modes, respectively), considering $k \geq 2$ further allows us to pinpoint the size of the limiting covariance explicitly in terms of the lengths of the matrix products and the number of traceless matrices involved. We further find that two modes of the build (1.3) are asymptotically independent whenever the total number of traceless matrices involved is odd.

A key ingredient for studying the fluctuation of (1.3) is information on the $1/N$ correction to $\mathbb{E}\langle f_1(W)A_1 \dots f_k(W)A_k \rangle$, which was included in the error terms of previous results (cf. [12, Cor. 2.7]). Before considering the CLT, we hence give an expansion of the expectation. Note that the leading term of this expansion was already identified in [13, 12]. As the corresponding local laws are obtained by induction, the limiting object naturally arises through a recursion. The explicit form of the expectation obtained in [13] from solving the recursion mirrors the combinatorics encountered in (*first-order*) free probability, e.g., for the alternating moments $\mathbb{E}\langle W_1 D_1 \dots W_k D_k \rangle$ of a finite family of independent Wigner matrices $(W_j)_j$ and a finite family of deterministic matrices $(D_j)_j$ (see, e.g., [39, Sect. 4.4]). Note, however, that free probability methods are typically restricted to (N -independent) polynomials and often require an independent family of Wigner matrices, while the resolvent approach presented in [13, 12] applies to a much wider class of functions including resolvents and mesoscopically rescaled Sobolev functions. In a similar spirit, the limiting covariance in our CLT also naturally arises through a recursion which can be solved to obtain an explicit formula. We carry out the necessary combinatorics in the companion paper [40] to show that the parallels to free probability identified in [13] for the expectation continue to hold for the fluctuations. More precisely, the structure of the limiting covariance in our CLT mirrors the combinatorics in *second-order* free probability theory (see [39, Ch. 5] and [15] for an introduction) and, in the special case $f_j(x) = x$, correctly reproduces the structure of the fluctuation moments of Wigner and deterministic matrices that was recently computed in [38]. To avoid introducing additional notation, we work with the recursive definitions in the present paper and only refer to the formulas in [40] for explicit computations and examples.

Lastly, as an application of the functional CLT, we consider the special case $f_j(x) = e^{it_j x}$ with $t_j \in \mathbb{R}$. Interpreting W as the Hamiltonian of a mean-field quantum system and the deterministic bounded matrix A as an observable, the quantity

$$A(t) := e^{itW} A e^{-itW}$$

describes the Heisenberg time evolution of A . In this context, applying the CLT for the linear statistics (1.3) yields information about the fluctuations around the equilibrium in certain thermalization problems. For $k = 1$, the main interest lies in a CLT for averages of diagonal eigenvector overlaps $\langle \mathbf{u}_j, A \mathbf{u}_j \rangle$ (see [14, Thm. 2.3]) due to their connection to the fluctuations in the eigenstate thermalization hypothesis (see [17]) which is referred to as quantum unique ergodicity in mathematics (see [42], further references can be found in [11]). For $k \geq 2$, the statistics in (1.3) translate to the simultaneous time evolution of different observables in the same quantum system. It is expected that two observables $A_1(t)$ and A_2 become *thermalized* for $t \gg 1$, i.e., that

$$\langle A_1(t) A_2 \rangle \approx \langle A_1 \rangle \langle A_2 \rangle$$

in the large t regime. More precisely, if both A_1 and A_2 are traceless we have

$$\langle A_1(t)A_2 \rangle = \langle A_1A_2 \rangle \frac{J_1(2t)^2}{t^2} + \frac{\xi(t)}{N} + \mathcal{O}\left(\frac{N^\varepsilon}{N^{3/2}}\right) \tag{1.4}$$

for any fixed $t \in \mathbb{R}$, where J_1 denotes a Bessel function of the first kind and $\xi(t)$ is a centered Gaussian random variable with a t -dependent variance. The first term of (1.4) was established in the recent paper [13] in the form of a *law of large numbers*-type result with an effective but non-optimal error bound. Applying our functional CLT for $k = 2$ shows that the fluctuations around the thermal value are Gaussian and thus gives the second term of the expansion. Considering asymptotics for $t \gg 1$ after letting $N \rightarrow \infty$ further yields an explicit expansion for the variance in the regime that is relevant for thermalization.

We conclude this section with a brief overview of the paper. After introducing some commonly used notations, we collect our assumptions on the Wigner matrix W in Assumption 1.1. We then briefly recall the optimal multi-resolvent local law [12, Thm. 2.5], which constitutes one of the key tools for the analysis. The main results of the paper are then given in Section 2. We start by giving a precise expansion of the expectation $\mathbb{E}\langle f_1(W)A_1 \dots f_k(W)A_k \rangle$ beyond the leading term (Theorem 2.4). Considering the fluctuations of the statistics in (1.3), we then establish a CLT and give an explicit formula for the limiting covariance (Theorem 2.7, Corollary 2.9). This is followed by a discussion of the result, including the asymptotics in the mesoscopic regime (Theorem 2.10), sufficient conditions for two modes to be asymptotically independent (Corollary 2.11) as well as the case of multiple independent Wigner matrices. We conclude Section 2 by applying the functional CLT to thermalization problems. In Section 3, we consider the special case of the resolvents $f_j(W) := G(z_j) = (W - z_j)^{-1}$ for some suitable spectral parameters $z_j \in \mathbb{C}$, which provides the key ingredient for the proof of our main results. Here, the first step is introducing a recursively defined set function $\mathcal{E}[\cdot]$ (Definition 3.1), which we then identify as the subleading $\frac{1}{N}$ term of the expectation $\mathbb{E}\langle G(z_1)A_1 \dots G(z_k)A_k \rangle$. This added resolution is the main tool in proving the CLT in the case that all functions f_j are resolvents (Theorem 3.8). The role of the limiting covariance in the theorem is played by a recursively defined set function $\mathfrak{m}_2[\cdot, \cdot]$ (Definition 3.5). Lastly, the proofs are given in Section 4. To keep the presentation concise, some routine calculations are deferred to the appendix.

1.1 Notation and conventions

We start by introducing some notation used throughout the paper. For two positive quantities f, g , we write $f \lesssim g$ and $f \sim g$ whenever there exist (deterministic, N -independent) constants $c, C > 0$ such that $f \leq Cg$ and $cg \leq f \leq Cg$, respectively. We denote the Hermitian conjugate of a matrix A by A^* and the complex conjugate of a scalar $z \in \mathbb{C}$ by \bar{z} . Moreover, $\|\cdot\|$ denotes the operator norm, $\text{Tr}(\cdot)$ is the usual trace and $\langle \cdot \rangle = N^{-1} \text{Tr}(\cdot)$. We further denote the covariance of two complex random variables Y_1, Y_2 by $\text{Cov}(Y_1, Y_2)$ and follow the convention

$$\text{Cov}(Y_1, Y_2) = \mathbb{E}(Y_1 - \mathbb{E}Y_1)\overline{(Y_2 - \mathbb{E}Y_2)},$$

i.e., the covariance is linear in the first and anti-linear in the second entry. For $k, a, b \in \mathbb{N}$ with $a \leq b$, we set $[k] = \{1, \dots, k\}$ and adopt the interval notation $[a, b] = \{a, a + 1, \dots, b\}$. We further write $\langle a, b \rangle$ or $[a, b)$ to indicate that a or b are excluded from the interval, respectively. Ordered sets are denoted by (\dots) instead of $\{\dots\}$.

Given a matrix $A \in \mathbb{C}^{N \times N}$, the traceless part of A is denoted by $\mathring{A} := A - \langle A \rangle \text{Id}$ where Id denotes the identity matrix. Further, $\mathfrak{a} := \text{diag}(A)$ denotes the diagonal matrix

obtained from extracting only the diagonal entries of A and $A_1 \odot A_2$ denotes the entry-wise (or Hadamard) product of two matrices A_1 and A_2 . For a Hermitian matrix W and $z_1, \dots, z_k \in \mathbb{C} \setminus \mathbb{R}$, we write the corresponding resolvents as $G_j = G(z_j) := (W - z_j)^{-1}$ and index products of resolvents using the interval notation

$$G_{[a,b]} := G_a G_{a+1} \dots G_b$$

for $a, b \in \mathbb{N}$ with $a \leq b$. Recalling that angled brackets indicate that an edge point of the interval is excluded, we write $G_{\langle a,b \rangle}$ and $G_{[a,b)}$ to exclude G_a or G_b from the product, respectively. Moreover, G_\emptyset is interpreted as zero. Note that this notation differs slightly from [13, 12]. As we often consider alternating products of resolvents with deterministic matrices A_1, \dots, A_k , define $T_j := G_j A_j$ and apply the same interval notation as above to write

$$T_{[k]} := T_1 \dots T_k = G_1 A_1 \dots G_k A_k, \quad T_{[a,b]} := T_a T_{a+1} \dots T_b. \tag{1.5}$$

Again, angled brackets are used to exclude T_a or T_b from the product, respectively, and T_\emptyset is interpreted as zero. We call a product of the type (1.5) *resolvent chain* of length k .

Throughout the paper, we assume W to be an $N \times N$ complex¹ Wigner matrix satisfying the following assumptions.

Assumption 1.1. *The matrix elements of W are independent up to Hermitian symmetry $W_{ij} = \overline{W_{ji}}$ and we assume identical distribution in the sense that there is a centered real random variable χ_d and a centered complex random variable χ_{od} such that $W_{ij} \stackrel{d}{=} N^{-1/2} \chi_{od}$ for $i < j$ and $W_{jj} \stackrel{d}{=} N^{-1/2} \chi_d$, respectively. We further assume that $\mathbb{E}|\chi_{od}|^2 = \mathbb{E}\chi_d^2 = 1$ as well as the existence of all moments of χ_d and χ_{od} , i.e., there exist constants $C_p > 0$ for any $p \in \mathbb{N}$ such that*

$$\mathbb{E}|\chi_d|^p + \mathbb{E}|\chi_{od}|^p \leq C_p.$$

Lastly, we assume that the pseudo-variance vanishes, i.e.,

$$\mathbb{E}\chi_{od}^2 = 0.$$

We further introduce the notation

$$\kappa_4 := \mathbb{E}|\chi_{od}|^4 - 2 \tag{1.6}$$

for the normalized fourth cumulant of the off-diagonal entries. Note that the notation matches [14], however, we restrict the model to complex matrices with vanishing pseudo-variance, i.e., $\mathbb{E}W_{ij}^2 = 0$ for $i \neq j$, for technical simplicity. The more general model from [14] is studied for macroscopic scales in the companion paper [40], and the necessary modifications for an extension to mesoscopic scales are sketched.

The eigenvalue density profile of W is described by the semicircle law (see, e.g., [1, Ch. 2] for some background), i.e., the probability measure with density

$$\rho_{sc}(x) := \frac{\sqrt{4 - x^2}}{2\pi} \mathbb{1}_{[-2,2]}(x) \tag{1.7}$$

which mainly enters our analysis in the form of its Stieltjes transform

$$m(z) := \int \frac{\rho_{sc}(x)}{x - z} dx, \quad z \in \mathbb{C} \setminus \mathbb{R}. \tag{1.8}$$

¹The same method applies to the real case with only small modifications. For simplicity of the presentation, we restrict the following analysis to the complex case only.

We remind the reader that $m(z)$ is the unique solution of the Dyson equation

$$-\frac{1}{m(z)} = m(z) + z, \quad \Im z \Im(m(z)) > 0 \tag{1.9}$$

and that its derivative satisfies

$$m'(z) = \frac{m(z)^2}{1 - m(z)^2}. \tag{1.10}$$

Given fixed $z_1, \dots, z_k \in \mathbb{C} \setminus \mathbb{R}$, set $m_j = m(z_j)$ and $m'_j = m'(z_j)$, respectively, and let

$$q_{i,j} = \frac{m_i m_j}{1 - m_i m_j}, \tag{1.11}$$

possibly setting $q_{j,j} = m'_j$ whenever $i = j$. Moreover, we define the *iterated divided difference* for finite multi-sets $\{z_1, \dots, z_k\} \subset \mathbb{C} \setminus \mathbb{R}$ recursively by

$$m[z_1, \dots, z_k] := \frac{m[z_2, \dots, z_k] - m[z_1, \dots, z_{k-1}]}{z_k - z_1} \tag{1.12}$$

whenever there are two distinct $z_1 \neq z_k$ among z_1, \dots, z_k and otherwise set

$$m[\underbrace{z, \dots, z}_k] := \frac{m^{(k-1)}(z)}{(k-1)!}$$

where $m^{(k-1)}$ is the $(k-1)$ -th derivative of the function m in (1.8). Note that this is well-defined in the sense that $m[z_1, \dots, z_k]$ is independent of the ordering of the multi-set $\{z_1, \dots, z_k\}$. We abbreviate $m[1, \dots, k] := m[z_1, \dots, z_k]$ and note that $q_{i,j}$ in (1.11) coincides with $m[i, j]$.

1.2 Preliminaries: multi-resolvent local laws

Before considering the fluctuations, we briefly recall the optimal multi-resolvent local law [12, Thm. 2.5], which characterizes the deterministic approximation of $\langle T_{[1,k]} \rangle$. We start by introducing the commonly used definition of stochastic domination.

Definition 1.2 (Stochastic domination). *Let*

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

be two families of non-negative random variables that are indexed by N and possibly some other parameter u . We say that X is stochastically dominated by Y , denoted by $X \prec Y$ or $X = \mathcal{O}_{\prec}(Y)$, if, for all $\varepsilon, C > 0$ we have

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

for large enough $N \geq N_0(\varepsilon, C)$.

Given $z_1, \dots, z_k \in \mathbb{C}$ and matrices A_1, \dots, A_k , we define the set function² $M_{[k]} = M_{[1,k]}$ for $k \geq 2$ through the recursion

$$M_{[k]} = m_1 \left(A_1 M_{[2,k]} + q_{1,k} \langle A_1 M_{[2,k]} \rangle + \sum_{j=2}^{k-1} \langle M_{[1,j]} \rangle (M_{[j,k]} + q_{1,k} \langle M_{[j,k]} \rangle) \right) \tag{1.13}$$

²Note that $M_{[k]}$ depends on $(z_j)_{j \in [k]}$ and $(A_j)_{j \in [k]}$, i.e., both the spectral parameters and the deterministic matrices are indexed by the same set. We hence interpret $M_{(\cdot)}$ as a function of the (ordered) index set to match the notation in the following sections.

with initial condition $M_{[1]} = m_1 \text{Id}$. This function was introduced in [13, Lem. 5.4] and an explicit (non-recursive) formula for $M_{[k]}$ was given in [13, Thm. 3.4]. However, we will only use the recursive definition in the present paper. Analogously to (1.13), we may define M_S for any (cyclically) ordered set $S = (s_1, \dots, s_k)$ instead of an interval. In this case, we write

$$M_S = M_{(s_1, \dots, s_k)}. \tag{1.14}$$

The set function $M_{[k]}$ plays the role of the deterministic approximation of $T_{[1,k]}G_k$ in the following multi-resolvent local laws.

Theorem 1.3 (Multi-resolvent local law, [12, Thm. 2.5]). *Fix $\zeta > 0$ and $k \in \mathbb{N}$. Let $z_1, \dots, z_k \in \mathbb{C} \setminus \mathbb{R}$ with $\max_j |z_j| \leq N^{100}$ and $d := \min_j \text{dist}(z_j, [-2, 2])$, deterministic matrices $A_1, \dots, A_k \in \mathbb{C}^{N \times N}$ with $\|A_i\| \lesssim 1$ such that a out of them are traceless. Set further $\eta_* := \min_j |\Im z_j| \geq N^{-1+\zeta}$. Recalling that $T_j = G_j A_j$, we have the averaged local law*

$$\langle T_{[1,k]} \rangle = \langle M_{[k]} A_k \rangle + \mathcal{O}_{\prec} \left(\frac{1}{N \eta_* \eta_*^{k-a/2-1}} \right), \tag{1.15}$$

and for $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$ with $\|\mathbf{x}\|, \|\mathbf{y}\| \lesssim 1$ we have the isotropic local law

$$\langle \mathbf{x}, T_{[1,k]} G_k \mathbf{y} \rangle = \langle \mathbf{x}, M_{[k]} \mathbf{y} \rangle + \mathcal{O}_{\prec} \left(\frac{1}{\sqrt{N} \eta_* \eta_*^{k-a/2-1}} \right). \tag{1.16}$$

Note that $A_1, \dots, A_k \in \mathbb{C}^{N \times N}$ are, in fact, N -dependent quantities, but we do not carry the N -dependence explicitly. As we frequently encounter $\langle M_{[k]} A_k \rangle$ in the following sections, we further introduce the notation³

$$\mathbf{m}_1[T_1, \dots, T_k] = \mathbf{m}_1[z_1, A_1, \dots, z_k, A_k] := \langle M_{[k]} A_k \rangle. \tag{1.17}$$

It is immediate from (1.13) that the function $\mathbf{m}_1[\cdot]$ satisfies the recursion

$$\begin{aligned} \mathbf{m}_1[T_1, \dots, T_k] &= m_1 \left(\mathbf{m}_1[T_2, \dots, T_k] + q_{1,k} \mathbf{m}_1[T_2, \dots, T_{k-1}, G_k] \langle A_k \rangle \right) \\ &\quad + \sum_{j=2}^{k-1} \mathbf{m}_1[T_1, \dots, T_{j-1}, G_j] \left(\mathbf{m}_1[T_j, \dots, T_k] + q_{1,k} \mathbf{m}_1[T_j, \dots, T_{k-1}, G_k] \langle A_k \rangle \right) \end{aligned} \tag{1.18}$$

with initial condition $\mathbf{m}_1[T_1] = \langle A_1 \rangle m_1$. The arguments in the notation $\mathbf{m}_1[T_1, \dots, T_k]$ indicate the deterministic approximation of $\langle T_1 \dots T_k \rangle$. Whenever $A_1 = \dots = A_k = \text{Id}$, it follows that the deterministic approximation is given by the iterated divided differences, i.e.,

$$\mathbf{m}_1[G_1, \dots, G_k] = m[1, \dots, k], \tag{1.19}$$

which can be seen from the resolvent identity

$$G_j G_{j-1} = \frac{G_j - G_{j-1}}{z_j - z_{j-1}} \tag{1.20}$$

and the averaged local law (1.15). Note that (1.15) and (1.16) may also be applied for any product $T_{s_1} \dots T_{s_{k-1}} G_{s_k}$ that is indexed by a (cyclically) ordered set $S = (s_1, \dots, s_k)$ instead of an interval. In this case, the deterministic approximation is given by (1.14).

We further note the following a priori bounds for $m[\cdot]$, $\mathbf{m}_1[\cdot]$, and $M_{[\cdot]}$ (cf. Lemma 2.4 and Appendix A of [12]). Recall that $\eta_j = \Im z_j$ is N -dependent in the mesoscopic regime.

³Note that $\mathbf{m}_1[T_1, \dots, T_n]$ is a deterministic quantity. We slightly abuse notation to use the symbol $T_j = G_j A_j$ and the tuple (z_j, A_j) interchangeably, as this allows writing some equations more compactly.

Lemma 1.4. *Let $k \in \mathbb{N}$, pick spectral parameters z_1, \dots, z_k and deterministic matrices such that a matrices among A_1, \dots, A_k are traceless. Further, set $\eta_* = \min_j |\Im z_j|$ and assume $d := \min_j \text{dist}(z_j, [-2, 2]) \leq 1$. Then,*

$$\begin{aligned} |m[1, \dots, k]| &\lesssim \frac{1}{\eta_*^{k-1}}, \\ |\mathfrak{m}_1[T_1, \dots, T_k]| &\lesssim \frac{1}{\eta_*^{k-1-\lceil a/2 \rceil}}, \\ |(M_{[k]})_{ij}| \leq \|M_{[k]}\| &\lesssim \frac{1}{\eta_*^{k-1-\lceil a'/2 \rceil}}, \end{aligned}$$

where a' denotes the number of traceless matrices among A_1, \dots, A_{k-1} and $\lceil x \rceil$ denotes the upper integer part of $x \in \mathbb{R}$. Generically, the above bounds are sharp⁴ when not all $\Im z_j$ have the same sign.

Theorem 1.3 together with the optimality of the bounds in Lemma 1.4 asserts that the deterministic $M_{[k]}$ is indeed the leading order approximation of $T_{[1,k]}G_k$. In particular, the error terms in (1.15) and (1.16) are smaller than the natural upper bound on their leading term by a factor of $(N\eta_*)^{-1}$ and $(N\eta_*)^{-1/2}$, respectively. We remark that the sharpness of the bounds in Lemma 1.4 was discussed in [12, App. A] and follows in essence from a suitable integral representation for $m[\cdot]$ as well as the non-recursive formulas for $\mathfrak{m}_1[\cdot]$ and $M_{[k]}$. If all $\Im z_j$ have the same sign, the iterated divided difference stays bounded due to the smoothness of (1.8) in the bulk, showing that the bound in Lemma 1.4 cannot be sharp in this case.

2 Main results

The main result of the present paper is a functional CLT for the centered statistics

$$Y_\alpha^{(k,a)} := \langle f_1(W)A_1 \dots f_k(W)A_k \rangle - \mathbb{E} \langle f_1(W)A_1 \dots f_k(W)A_k \rangle, \tag{2.1}$$

where α is a multi-index containing the deterministic matrices and test functions involved and a denotes the number of traceless matrices among A_1, \dots, A_k . Note that we omit the superscripts of $Y_\alpha^{(k,a)}$ whenever a or k are not used explicitly. The test functions f_1, \dots, f_k are chosen according to the following set of assumptions.

Assumption 2.1 (Test functions). *For $k, p \in \mathbb{N}$ let $g_1, \dots, g_k \in H_0^p(\mathbb{R})$ be (N -independent) real-valued compactly supported test functions with $\|g_j\|_{H_0^p} \lesssim 1$. Fixing $\delta, \gamma \geq 0$ as well as $\gamma_1, \dots, \gamma_k \geq 0$ either as*

- (1) [Macro] $\delta = \gamma = \gamma_1 = \dots = \gamma_k = 0$ or as
- (2) [Meso] $\delta > 0, \gamma \in (0, 1)$, and $0 < \gamma_j \leq \gamma$,

we pick (N -independent) reference energies $E_j \in [-2 + \delta, 2 - \delta]$ for $j = 1, \dots, k$. Lastly, we define the test function rescaled to a scale $N^{-\gamma_j}$ around E_j by

$$f_j(x) := g_j(N^{\gamma_j}(x - E_j)). \tag{2.2}$$

Note that Assumption 2.1 includes both the *macroscopic* scale (1) and the bulk regime for *mesoscopic* scales (2). While it is possible to mix both regimes by choosing part of the test functions to satisfy (1) and the remaining ones to satisfy (2), the error estimate obtained from the results below is of the same size as if all functions were chosen on the mesoscopic scale. Treating this mixed regime optimally would require a multi-resolvent

⁴The bounds are "sharp" in the sense that they are optimal in the class of bounds involving only η_* in the small η_* regime, see also [12].

local law that carries the number of spectral parameters with N -dependent imaginary part as a parameter. Since the present paper focuses on the proof of the multi-point functional central limit theorem, we use the already established Theorem 1.3 as input and choose either (1) or (2) for all test functions.

We further remark that the restriction to real-valued test functions is only for simplicity. The extension of the results in this section to complex-valued g_1, \dots, g_k follows by standard arguments. Moreover, as one can decompose any matrix A_j in Y_α as $A_j = \langle A_j \rangle \text{Id} + \check{A}_j$, by multi-linearity, it is sufficient to consider Y_α for deterministic matrices A_j that are either traceless or equal to the identity matrix.

Throughout the paper, we denote the multi-index α in the form

$$\alpha := ((g_1, \gamma_1, E_1, A_1), \dots, (g_k, \gamma_k, E_k, A_k))$$

with g_j, γ_j , and E_j chosen following Assumption 2.1. Moreover, we introduce $F_j := f_j(W)A_j$ and use the interval notation

$$F_{[i,j]} := f_i(W)A_i \dots f_j(W)A_j$$

for $i < j$ as well as $F_\emptyset = 0$. Note that $(g_j, \gamma_j, E_j, A_j)$ and F_j contain the same information. For this reason, we will occasionally abuse notation and use both quantities interchangeably.

We further introduce the random variables

$$X_\alpha^{(k,a)} := \langle T_{[1,k]} \rangle - \mathbb{E} \langle T_{[1,k]} \rangle = \langle G_1 A_1 \dots G_k A_k \rangle - \mathbb{E} \langle G_1 A_1 \dots G_k A_k \rangle, \quad (2.3)$$

as a special case of (2.1). By a suitable functional calculus (cf. [16]), information on (2.3) carries over to the general statistics (2.1), thus yielding a key tool for the proof of our main results. We, therefore, consider the analog of the results in Sections 2.1 and 2.2 for the resolvent case separately in Section 3. Throughout the paper, we write

$$\alpha = ((z_1, A_1), \dots, (z_k, A_k))$$

for the multi-index in (2.3) containing the spectral parameters z_1, \dots, z_k appearing in the resolvents as well as the deterministic matrices. Whenever we do not need the number k of resolvents (resp. deterministic matrices) in the product or the number a of traceless deterministic matrices among A_1, \dots, A_k explicitly, we again omit the superscript, and further occasionally abuse notation to use (z_j, A_j) and $T_j = G_j A_j$ interchangeably. In the context of (2.2), we may interpret the resolvent $G(z)$ as a function rescaled to scale $|\Im z|^{-1}$ around $\Re z$ (even though the corresponding function g is not compactly supported). The analog of Assumption 2.1 for the spectral parameters now reads as follows.

Assumption 2.2 (Spectral parameters). *Let $k \in \mathbb{N}$. Fixing $\delta, \zeta \geq 0$ either as*

- (1) [Macro] $\delta = \zeta = 0$ or as
- (2) [Meso] $\delta > 0$ and $\zeta \in (0, 1)$,

pick (N -independent) reference energies $E_j \in [-2 + \delta, 2 - \delta]$. We choose the spectral parameters $z_1, \dots, z_k \in \mathbb{C}$ such that $z_j = E_j + i\eta_j$ with $|\eta_j| \gtrsim N^{-\zeta}$ and $\max_j |z_j| \leq N^{100}$.

Note that we consider spectral parameters z_j for which $|\Im z_j|$ is either of order one (macroscopic scale) or only slightly above the typical eigenvalue spacing (mesoscopic scales). Whenever $|\eta_j|$ is small, we further restrict to the bulk regime, i.e., those z_j for which $\Re z_j$ is bounded away from the boundary of the support of the semicircle density at ± 2 .

2.1 The $\frac{1}{N}$ term of $\mathbb{E}\langle f_1(W)A_1 \dots f_k(W)A_k \rangle$

We start our analysis by considering an expansion of $\mathbb{E}\langle F_{[1,k]} \rangle$ which identifies the subleading $1/N$ term. To state the theorem, we introduce a set function $\mathcal{E}[\cdot]$ that plays the role of the $1/N$ term for the resolvent case $\mathbb{E}\langle T_{[1,k]} \rangle$. Note that $\mathcal{E}[\cdot]$ characterizes the error of order $1/N$ that is obtained from interchanging $\langle T_{[1,k]} \rangle - \mathbb{E}\langle T_{[1,k]} \rangle$ and $\langle T_{[1,k]} \rangle - m_1[T_1, \dots, T_k]$, i.e., it relates X_α in (2.3) to the bounds in the local law (1.15). The proof of Lemma 2.3 is carried out in Section 4.2 and a discussion of the properties of $\mathcal{E}[\cdot]$ is included in Section 3.1.

Lemma 2.3. *Let $k \in \mathbb{N}$, W be a Wigner matrix satisfying Assumption 1.1, and fix spectral parameters z_1, \dots, z_k satisfying Assumption 2.2 as well as deterministic matrices A_1, \dots, A_k with $\|A_j\| \lesssim 1$. Moreover, assume that a matrices among A_1, \dots, A_k are traceless. Then there exists a set function $\mathcal{E}[\cdot]$ (defined recursively in Definition 3.1 below) such that*

$$\mathbb{E}\langle T_1 \dots T_k \rangle = m_1[T_1, \dots, T_k] + \frac{\kappa_4}{N} \mathcal{E}[T_1, \dots, T_k] + \mathcal{O}\left(\frac{N^\varepsilon}{N \sqrt{N} \eta_* \eta_*^{k-a/2}}\right) \tag{2.4}$$

with $m_1[\cdot]$ as in (1.17), κ_4 as in (1.6), and $\eta_* := \min_j |\Im z_j|$.

We remark that (cf. Lemma 3.2 below)

$$\mathcal{E}[T_1, \dots, T_k] \lesssim \frac{1}{\eta_*^{k-1-\lfloor a/2 \rfloor}},$$

i.e., there is a clear distinction between an $\mathcal{O}(1)$ leading term, an $\mathcal{O}(N^{-1})$ subleading term, and an $\mathcal{O}(N^{\varepsilon-3/2})$ error term for (2.4) in the macroscopic regime (Part (1) of Assumption 2.2), but the hierarchy of the three summands is not as distinct in the mesoscopic regime (Part (2) of Assumption 2.2). However, the resolution given in Lemma 2.3 is sufficient for the proofs that follow, see also Remark 3.3 below. We now give the expansion of $\mathbb{E}\langle F_{[1,k]} \rangle$.

Theorem 2.4. *Let $k \in \mathbb{N}$ and pick deterministic matrices $A_1, \dots, A_k \in \mathbb{C}^{N \times N}$ with $\|A_j\| \lesssim 1$ such that a out of them are traceless. Let further W be a Wigner matrix satisfying Assumption 1.1 and let f_1, \dots, f_k be test functions satisfying Assumption 2.1 with $p = k - \lfloor a/2 \rfloor + 1$. Then, for any $\varepsilon > 0$, we have the expansion*

$$\begin{aligned} \mathbb{E}\langle F_{[1,k]} \rangle &= \frac{1}{\pi^k} \int_{\mathbb{R}^k} \int_{[0,10]^k} \left[\prod_{j=1}^k (\partial_{\bar{z}}(f_j)_{\mathbb{C},p})(z_j) \right] m_1[G(z_1)A_1, \dots, G(z_k)A_k] d\eta_{[k]} dx_{[k]} \\ &+ \frac{\kappa_4}{N \pi^k} \int_{\mathbb{R}^k} \int_{[0,10]^k} \left[\prod_{j=1}^k (\partial_{\bar{z}}(f_j)_{\mathbb{C},p})(z_j) \right] \mathcal{E}[G(z_1)A_1, \dots, G(z_k)A_k] d\eta_{[k]} dx_{[k]} \\ &+ \mathcal{O}\left(\frac{N^\varepsilon \max_j \|f_j\|_{H^p}}{N^{3/2}}\right) \end{aligned} \tag{2.5}$$

where we write $z_j = x_j + i\eta_j$, $dx_{[k]} = dx_1 dx_2 \dots dx_k$ as well as $d\eta_{[k]} = d\eta_1 d\eta_2 \dots d\eta_k$, and $(f_j)_{\mathbb{C},p}$ denotes the almost analytic extension of f_j of order p (see (4.37) for the detailed definition of $(f_j)_{\mathbb{C},p}$).

Theorem 2.4 follows from Lemma 2.3 and the Helffer-Sjöstrand formula (see [16]). As a similar argument will be used for the more involved proof of the multi-point functional CLT in Theorem 2.7, we omit the details here.

It follows from (93) in [14] that $\mathcal{E}[T_1]$ is given by

$$\mathcal{E}[T_1] = \langle A_1 \rangle \frac{m_1^5}{1 - m_1^2} = \langle A_1 \rangle m_1' m_1^3. \tag{2.6}$$

Hence, computing the second integral in (2.5) shows that the $1/N$ term $\mathcal{E}[f_1(W)A_1]$ of $\mathbb{E}\langle f_1(W)A_1 \rangle$ is

$$\mathcal{E}[f_1(W)A_1] = \langle A_1 \rangle \left(\int_{-2}^2 f_1(x)\rho_{sc}(x)dx + \frac{\kappa_4}{2\pi} \int_{-2}^2 \frac{(x^4 - 4x^2 + 2)f_1(x)}{\sqrt{4 - x^2}} dx - \frac{f_1(0)}{2} \right),$$

where ρ_{sc} denotes the density of the semicircle law in (1.7). We remark that this formula was already included in [14, Thm. 2.4]. Theorem 2.4 hence generalizes Equation (21) in [14] to arbitrary $k \geq 1$ in the setting of Assumptions 1.1 and 2.1.

2.2 Statement of the multi-point functional CLT

We now state our main result, the multi-point functional CLT for the statistics Y_α in (2.1). To give the limiting covariance structure explicitly, we introduce a set function $\mathfrak{m}_2[\cdot|\cdot]$ to play the role of the leading term of the (appropriately scaled) covariance⁵ of $\langle T_{[1,k]} \rangle$ and $\langle T_{[k+1,k+\ell]} \rangle$ in the same way that $M_{[k]}$ and $\mathfrak{m}_1[\cdot]$ do for the expectation of $T_{[1,k]}G_k$ (see Theorem 1.3 as well as (1.17)). Recall that we use (z_j, A_j) and T_j interchangeably. In particular, we may write

$$\mathfrak{m}_2[\alpha|\beta] = \mathfrak{m}_2[T_1, \dots, T_k | T_{k+1}, \dots, T_{k+\ell}]$$

where the two multi-indices α and β contain the spectral parameters and deterministic matrices in T_1, \dots, T_k and $T_{k+1}, \dots, T_{k+\ell}$, respectively.

Lemma 2.5. *Fix $k, \ell \in \mathbb{N}$, let α, β be two multi-indices of length k and ℓ , respectively, and let W be a Wigner matrix satisfying Assumption 1.1. Pick two sets of spectral parameters z_1, \dots, z_k and $z_{k+1}, \dots, z_{k+\ell}$ that either both satisfy Case 1 or both satisfy Case 2 of Assumption 2.2, and denote $\eta_* = \min_j |\Im z_j|$. Moreover, pick deterministic matrices $A_1, \dots, A_{k+\ell}$ with $\|A_j\| \lesssim 1$ such that a matrices among A_1, \dots, A_k and b matrices among $A_{k+1}, \dots, A_{k+\ell}$ are traceless. Then there exists a set function $\mathfrak{m}_2[\cdot|\cdot]$ (defined recursively in Definition 3.5 below) such that*

$$N^2 \mathbb{E} X_\alpha^{(k,a)} X_\beta^{(\ell,b)} = \mathfrak{m}_2[\alpha|\beta] + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N\eta_*} \eta_*^{k-a/2} \eta_*^{\ell-b/2}}\right) \tag{2.7}$$

for any $\varepsilon > 0$.

We remark that (cf. (3.14) below)

$$|\mathfrak{m}_2[T_1, \dots, T_k | T_{k+1}, \dots, T_{k+\ell}]| \lesssim \frac{1}{\eta_*^{k+\ell - \lceil (a+b)/2 \rceil}},$$

i.e., there is a clean distinction between an $\mathcal{O}(1)$ leading term and an $\mathcal{O}(N^{\varepsilon-1/2})$ error term for (2.7) in the macroscopic regime (Part (1) of Assumption 2.2), but the hierarchy of the two summands may not be as distinct in the mesoscopic regime (Part (2) of Assumption 2.2), see also Remark 3.9 below. The statistics X_α and the corresponding CLT, as well as the properties of the function $\mathfrak{m}_2[\cdot|\cdot]$ are discussed in more detail in Section 3.2. Moreover, explicit (non-recursive) formulas for $\mathfrak{m}_2[\cdot|\cdot]$ are derived in the companion paper [40]. We further recall the following definition.

Definition 2.6. *Consider two functions of the Wigner matrix W in Assumption 1.1, which we denote as N -dependent random variables $X^{(N)}$ and $Y^{(N)}$. We say that $X^{(N)} = Y^{(N)} + \mathcal{O}(N^{-c})$ with $c > 0$ in the sense of moments if for any polynomial \mathcal{P} it holds that*

$$\mathbb{E}\mathcal{P}(X^{(N)}) = \mathbb{E}\mathcal{P}(Y^{(N)}) + \mathcal{O}(N^{-c+\varepsilon}),$$

⁵Note the similarity between the notations $\mathfrak{m}_1[\cdot]$ and $\mathfrak{m}_2[\cdot|\cdot]$, which take one and two resolvent chains as arguments, respectively.

for any small $\varepsilon > 0$, where the implicit constant in $\mathcal{O}(\cdot)$ only depends on the polynomial \mathcal{P} and the constants in Assumption 1.1.

The main result of the paper can now be stated as follows.

Theorem 2.7 (Multi-point functional CLT). *Under the assumptions of Theorem 2.4 it holds that, for any $\varepsilon > 0$, the centered statistics (2.1) are approximately distributed (in the sense of moments) as*

$$NY_\alpha^{(k,a)} = \xi(\alpha) + \mathcal{O}\left(\frac{N^\varepsilon \max_j \|f_j\|_{H^p}}{\sqrt{N}}\right) \tag{2.8}$$

with a centered (N -dependent) Gaussian process $\xi(\alpha)$ satisfying

$$\begin{aligned} & \mathbb{E}[\xi(\alpha)\xi(\beta)] \tag{2.9} \\ &= \frac{1}{\pi^{k+\ell}} \int_{\mathbb{R}^k} dx_{[k]} \int_{[0,10]^k} d\eta_{[k]} \left[\prod_{i=1}^k (\partial_{\bar{z}}(f_i)_{\mathbb{C},p})(z_i) \right] \int_{\mathbb{R}^\ell} dx_{[k+1,k+\ell]} \int_{[0,10]^\ell} d\eta_{[k+1,k+\ell]} \\ & \times \left[\prod_{j=k+1}^\ell (\partial_{\bar{z}}(f_j)_{\mathbb{C},q})(z_j) \right] \mathfrak{m}_2[G(z_1)A_1, \dots, G(z_k)A_k | G(z_{k+1})A_{k+1}, \dots, G(z_{k+\ell})A_{k+\ell}]. \end{aligned}$$

Here, $z_j = x_j + i\eta_j$, $dx_{[i,j]} = dx_i dx_{i+1} \dots dx_j$ as well as $d\eta_{[i,j]} = d\eta_i d\eta_{i+1} \dots d\eta_j$ for $i < j$, and $(f_j)_{\mathbb{C},p}$ denotes the almost analytic extension of f_j of order p . Further, β denotes another multi-index of length ℓ containing the deterministic matrices $A_{k+1}, \dots, A_{k+\ell}$ with $\|A_j\| \lesssim 1$ out of which b are traceless, as well as the test functions $f_{k+1}, \dots, f_{k+\ell}$ satisfying Assumption 2.1 with $q = \ell - \lfloor b/2 \rfloor + 1$. Recall that $\mathfrak{m}_2[\cdot, \cdot]$ was introduced in Lemma 2.5.

The key ingredient for the proof of Theorem 2.7 is the case of all f_j being resolvents, which we discuss in detail in Section 3.2 below. The full multi-point functional CLT is then obtained from the resolvent CLT using the Helffer-Sjöstrand formula (cf. [16]). We carry out the argument in Section 4.5.

2.3 Discussion of the multi-point functional CLT for mesoscopic scales

In this section, we consider the mesoscopic regime of Theorem 2.7 (Case 2 of Assumption 2.1) in more detail. A key ingredient in this discussion is the decomposition

$$\mathfrak{m}_2[\cdot, \cdot] = \mathfrak{m}_{GUE}[\cdot, \cdot] + \kappa_4 \mathfrak{m}_\kappa[\cdot, \cdot] \tag{2.10}$$

for the set function $\mathfrak{m}_2[\cdot, \cdot]$ in Lemma 2.5 into two functions $\mathfrak{m}_{GUE}[\cdot, \cdot]$ and $\mathfrak{m}_\kappa[\cdot, \cdot]$ that do not depend on any parameters of the underlying Wigner matrix W . Equation (2.10) induces a similar decomposition for the limiting covariance in (2.9). We remark that the two contributions are not of comparable size in the mesoscopic regime and that the summand with prefactor κ_4 is of lower order. This reduces the leading term in (2.9) to the case $\kappa_4 = 0$, thus simplifying it considerably. We give a brief example in the resolvent case to illustrate this phenomenon. More general bounds are given in Lemma 3.7 below. Recall that we may interpret the resolvent $G(z)$ as a function rescaled to scale $|\Im z|^{-1}$ around $\Re z$ to match (2.2).

Example 2.8. For $k = \ell = 1$, we have by (92) of [14] (or Definition 3.5 below) that

$$\begin{aligned} \mathfrak{m}_2[T_1|T_2] &:= \langle A_1 A_2 \rangle \frac{m_1^2 m_2^2}{(1 - m_1 m_2)} + \langle \mathbf{a}_1 \mathbf{a}_2 \rangle \cdot \kappa_4 m_1^3 m_2^3 \tag{2.11} \\ &+ \langle A_1 \rangle \langle A_2 \rangle \left(\frac{m_1' m_2'}{(1 - m_1 m_2)^2} - \frac{m_1^2 m_2^2}{(1 - m_1 m_2)} + 2\kappa_4 m_1 m_1' m_2 m_2' - \kappa_4 m_1^3 m_2^3 \right), \end{aligned}$$

where \mathbf{a}_i denotes the diagonal part of the matrix A_i . Assuming that $\|A_1\|, \|A_2\| \lesssim 1$ and $|\Im z_1|, |\Im z_2| \geq \eta_*$, a brief computation using the explicit form

$$m(z) = \frac{-z + \sqrt{z^2 - 4}}{2}$$

of the solution to (1.9) shows that $|(1 - m_1 m_2)^{-1}| \lesssim \eta_*^{-1}$. Hence, the function $\mathbf{m}_2[T_1|T_2]$ in (2.11) satisfies the bounds

$$\mathbf{m}_2[T_1|T_2] \lesssim \begin{cases} \eta_*^{-2}, & \text{if } \langle A_1 \rangle \langle A_2 \rangle \neq 0, \\ \eta_*^{-1}, & \text{if } \langle A_1 \rangle \langle A_2 \rangle = 0. \end{cases}$$

Both inequalities are sharp if $\Im z_1$ and $\Im z_2$ have opposite signs. We further note that

$$\mathbf{m}_{GUE}[T_1|T_2] \lesssim \begin{cases} \eta_*^{-2}, & \text{if } \langle A_1 \rangle \langle A_2 \rangle \neq 0, \\ \eta_*^{-1}, & \text{if } \langle A_1 \rangle \langle A_2 \rangle = 0, \end{cases} \quad \mathbf{m}_\kappa[T_1|T_2] = \mathcal{O}(1),$$

i.e., the two parts of $\mathbf{m}_2[\cdot|\cdot]$ (cf. decomposition in (2.10)) do not contribute equally unless $\eta_* \gtrsim 1$ (macroscopic regime).

We hence restrict the following discussion to the case $\kappa_4 = 0$. Even with this simplification, the limiting covariance in Theorem 2.7 may be tedious to compute using the recursive definition of $\mathbf{m}_2[\cdot|\cdot]$ alone. In the companion paper [40], we consider the recursion defining $\mathbf{m}_2[\cdot|\cdot]$ in detail and derive explicit formulas. Combining [40, Thm. 2.4] with (2.9) then yields a more direct way of computing the limiting covariance which is fully explicit whenever $\kappa_4 = 0$. The result is given in Corollary 2.9 below. We emphasize that $\mathbb{E}\xi(\alpha)\xi(\beta)$ is a sum of terms that decompose into a product of a function of the deterministic matrices A_1, \dots, A_k and an expression in the test functions $f_1, \dots, f_{k+\ell}$, respectively. Moreover, the combinatorics underlying the summation mirror those encountered in second-order free probability theory. The proof of Corollary 2.9 is given in Section 4.6.

To state the result, we denote by $NCP(k)$ the set of non-crossing partitions of $[k]$, by $\overline{NCP}(k, \ell)$ the set of non-crossing permutations of the (k, ℓ) -annulus, and by $K(\pi)$ the Kreweras complement associated with an element $\pi \in NCP(k)$ or $\pi \in \overline{NCP}(k, \ell)$, respectively. Moreover, the first and second-order cumulants h_\circ and $h_{\circ\circ}$ associated with set functions $h[\cdot]$ and $h[\cdot|\cdot]$ are computed recursively from the moment-cumulant relations

$$h[S] = \sum_{\pi \in NCP(S)} \prod_{B \in \pi} h_\circ[B], \tag{2.12}$$

$$h[S_1|S_2] = \sum_{\pi \in \overline{NCP}(|S_1|, |S_2|)} \prod_{B \in \pi} h_\circ[B] + \sum_{\substack{\pi_1 \times \pi_2 \in NCP(|S_1|) \times NCP(|S_2|), \\ U_1 \in \pi_1, U_2 \in \pi_2 \text{ marked}}} h_{\circ\circ}[U_1|U_2] \prod_{\substack{B \in \pi_1 \setminus U_1 \\ \cup \pi_2 \setminus U_2}} h_\circ[B]. \tag{2.13}$$

The full definitions and some illustrative examples are, e.g., given in [40, Sect. 1] or [39].

Corollary 2.9. *Consider the setup of Theorem 2.7 for $\kappa_4 = 0$. Then,*

$$\begin{aligned} \mathbb{E}\xi(\alpha)\xi(\beta) &= \sum_{\pi \in \overline{NCP}(k, \ell)} \left(\prod_{B \in K(\pi)} \left\langle \prod_{j \in B} A_j \right\rangle \right) \prod_{B \in \pi} \Phi_{\pi, B}(f_j | j \in B) \\ &+ \sum_{\substack{\pi_1 \times \pi_2 \in NCP(k) \times NCP(\ell), \\ U_1 \in \pi_1, U_2 \in \pi_2 \text{ marked}}} \left(\prod_{\substack{B_1 \in K(\pi_1), \\ B_2 \in K(\pi_2)}} \left\langle \prod_{j \in B_1} A_j \right\rangle \left\langle \prod_{j \in B_2} A_j \right\rangle \right) \Phi_{\pi_1 \times \pi_2, U_1 \times U_2}(f_1, \dots, f_{k+\ell}). \end{aligned} \tag{2.14}$$

The functions $\Phi_{\pi,B}$ and $\Phi_{\pi_1 \times \pi_2, U_1 \times U_2}$ in (2.14) are given by

$$\Phi_{\pi,B}(f_j | j \in B) = \text{sc}_o[B], \tag{2.15}$$

where $\text{sc}_o[\cdot]$ denotes the first-order free cumulant function associated with

$$\text{sc}[i_1, \dots, i_n] := \int_{-2}^2 \left[\prod_{j=1}^n f_{i_j}(x) \right] \rho_{sc}(x) dx, \tag{2.16}$$

with ρ_{sc} as in (1.7), and

$$\Phi_{\pi_1 \times \pi_2, U_1 \times U_2}(f_1, \dots, f_{k+\ell}) = \text{sc}_{oo}[U_1 | U_2] \prod_{\substack{B_1 \in \pi_1 \setminus U_1, \\ B_2 \in \pi_2 \setminus U_2}} \text{sc}_o[B_1] \text{sc}_o[B_2]. \tag{2.17}$$

Here, $\text{sc}_{oo}[\cdot | \cdot]$ denotes the second-order free cumulants associated with $\text{sc}[\cdot]$ in (2.16) and

$$\text{sc}[i_1, \dots, i_n | i_{n+1}, \dots, i_{n+m}] := \frac{1}{2} \int_{-2}^2 \int_{-2}^2 \left(\prod_{j=1}^n f_{i_j}(x) \right)' \left(\prod_{j=1}^m f_{i_{n+j}}(y) \right)' u(x, y) dx dy, \tag{2.18}$$

where the integral kernel $u : [-2, 2] \times [-2, 2] \rightarrow \mathbb{R}$ is given by

$$u(x, y) := \frac{1}{4\pi^2} \ln \left[\frac{(\sqrt{4-x^2} + \sqrt{4-y^2})^2 (xy + 4 - \sqrt{4-x^2} \sqrt{4-y^2})}{(\sqrt{4-x^2} - \sqrt{4-y^2})^2 (xy + 4 + \sqrt{4-x^2} \sqrt{4-y^2})} \right]. \tag{2.19}$$

We remark that the structure of (2.14) resembles the formula in [38, Thm. 6] for the covariance of alternating products of GUE and deterministic matrices in second-order free probability. This connection is discussed further in the companion paper [40]. In particular, applying Theorem 2.7 and (2.14) for $f_1(x) = \dots = f_{k+\ell}(x) = x$ reproduces the corresponding formulas in [38] (cf. [40, Cor. 2.11]).

Theorem 2.7 and Corollary 2.9 identify the limiting process $\xi(\alpha)$ in terms of the test functions $f_1, \dots, f_{k+\ell}$. Similar to [14, Sect. 2.3] in the case $k = \ell = 1$, we can make use of the mesoscopic scaling (2.2) to give asymptotic formulas in terms of the functions $g_1, \dots, g_{k+\ell}$ whenever $\gamma_j > 0$ for all j . The key quantities $\text{sc}[\cdot]$ and $\text{sc}[\cdot | \cdot]$ characterizing the covariance structure of two modes $Y_\alpha^{(k,a)}$ and $Y_\beta^{(\ell,b)}$ can then be conveniently expressed in terms of the L^2 and $\dot{H}^{1/2}$ inner products

$$\langle f, g \rangle_{L^2} := \int_{\mathbb{R}} f(x)g(x)dx, \quad \langle f, g \rangle_{\dot{H}^{1/2}} := \int_{\mathbb{R}^2} \frac{f(x) - f(y)}{x - y} \frac{g(x) - g(y)}{x - y} dx dy.$$

Theorem 2.10 (Bulk scaling asymptotics). *Under the assumptions of Theorem 2.7, pick test functions $f_1, \dots, f_{k+\ell}$ that satisfy Assumption 2.1 with some $\delta, \gamma_j > 0$.*

- (i) *Whenever $\gamma_1 = \dots = \gamma_{n+m} = \gamma$ and the functions $f_{i_1}, \dots, f_{i_{n+m}}$ are rescaled around the same $E_0 \in [-2 + \delta, 2 - \delta]$, it holds that*

$$N^\gamma \text{sc}[i_1, \dots, i_{n+m}] = \rho_{sc}(E_0) \left\langle \prod_{j=1}^n g_{i_j}, \prod_{j=n+1}^{n+m} g_{i_j} \right\rangle_{L^2} + \mathcal{O}(N^{-\gamma}),$$

$$\text{sc}[i_1, \dots, i_n | i_{n+1}, \dots, i_{n+m}] = \frac{1}{4\pi^2} \left\langle \prod_{j=1}^n g_{i_j}, \prod_{j=n+1}^{n+m} g_{i_j} \right\rangle_{\dot{H}^{1/2}} + \mathcal{O}(N^{-\gamma}).$$

- (ii) *If $\gamma_{i_1} = \dots = \gamma_{i_{n+m}} = \gamma$, but the functions $f_{i_1}, \dots, f_{i_{n+m}}$ are not rescaled around the same energy, we have the bounds*

$$N^\gamma \text{sc}[i_1, \dots, i_{n+m}] = \mathcal{O}(N^{-\gamma}),$$

$$\text{sc}[i_1, \dots, i_n | i_{n+1}, \dots, i_{n+m}] = \mathcal{O}(N^{-\gamma}).$$

Recall that $E_{i_1}, \dots, E_{i_{n+m}}$ are fixed and N -independent.

(iii) If $E_{i_1} = \dots = E_{i_{n+m}}$, but the scales $\gamma_1, \dots, \gamma_{n+m}$ do not all coincide, we have the bounds

$$N^{\gamma_{\min}} \text{sc}[i_1, \dots, i_{n+m}] = \mathcal{O}\left(N^{-(\gamma_{\min,2} - \gamma_{\min})}\right),$$

$$\text{sc}[i_1, \dots, i_n | i_{n+1}, \dots, i_{n+m}] = \mathcal{O}\left(N^{-(\gamma_{\min,2} - \gamma_{\min})}\right).$$

with $\gamma_{\min} = \min_j \gamma_{i_j}$ and $\gamma_{\min,2} = \min\{\gamma_{i_j} | \gamma_{i_j} > \gamma_{\min}\}$.

The implicit constants in the error terms depend only on δ and the scaling exponents γ_j as well as the test functions $g_1, \dots, g_{k+\ell}$ through $\|g_j\|_{H_0^p}$ and $|\text{supp} g_j|$.

The proof of Theorem 2.10 follows from the definition of $\text{sc}[\cdot]$ in (2.16) and (careful) integration by parts of (2.18). We omit the details.

Note that the random variables $\xi(\alpha_1), \dots, \xi(\alpha_k)$ obtained for $k \geq 1$ multi-indices $\alpha_1, \dots, \alpha_k$ are jointly Gaussian as a consequence of Theorem 3.8 below. Next, we discuss conditions under which two modes Y_α and Y_β in Theorem 2.7 are asymptotically independent. The case $k = \ell = 1$ of Corollary 2.11 was already discussed in [14, Thm. 2.13].

Corollary 2.11 (Independent modes). *Under the assumptions of Theorem 2.7, let α and β be multi-indices such that the test functions f_1, \dots, f_k associated with α are all rescaled around a common reference energy E_α on the scale $\gamma_\alpha > 0$ and the test functions $f_{k+1}, \dots, f_{k+\ell}$ associated with β are all rescaled around E_β on the scale $\gamma_\beta > 0$. Further, assume that α and β contain a and b traceless matrices, respectively, and denote by $\xi(\alpha)$ and $\xi(\beta)$ the corresponding limiting Gaussian processes in Theorem 2.7.*

(i) *If $E_\alpha \neq E_\beta$, then the processes $\xi(\alpha)$ and $\xi(\beta)$ are asymptotically independent in the sense that*

$$|\mathbb{E}[\xi(\alpha)\xi(\beta)]| \lesssim N^{-\min\{\gamma_\alpha, \gamma_\beta\}}.$$

(ii) *If $\gamma_\alpha \neq \gamma_\beta$, then the processes $\xi(\alpha)$ and $\xi(\beta)$ are asymptotically independent in the sense that*

$$|\mathbb{E}[\xi(\alpha)\xi(\beta)]| \lesssim N^{-|\gamma_\alpha - \gamma_\beta|}.$$

(iii) *If $a + b$ is odd, then the processes $\xi(\alpha)$ and $\xi(\beta)$ are always asymptotically independent in the sense that*

$$\mathbb{E}[\xi(\alpha)\xi(\beta)] = \mathcal{O}\left(\frac{N^\varepsilon \max_{i \in [k]} \|f_i\|_{H^p} \max_{j \in [k+1, k+\ell]} \|f_j\|_{H^q}}{\sqrt{N}}\right),$$

where $p = k - \lfloor a/2 \rfloor + 1$ and $q = \ell - \lfloor b/2 \rfloor + 1$, i.e., the leading deterministic term in (2.9) has the same size as the error term in Theorem 2.7.

The proof of Corollary 2.11 is immediate from Theorem 2.10 and Remark 3.9 on independent modes in the resolvent CLT. As $m_\kappa[\cdot]$ is negligible in the mesoscopic regime (cf. bounds in Lemma 3.7), Theorem 2.10 is indeed enough to conclude.

Remark 2.12 (Multiple independent Wigner matrices). In both cases of Assumption 2.1, the high-probability sense of Theorem 2.7 allows us to generalize the result to multiple independent Wigner matrices by resolving the individual matrices iteratively while conditioning on all others. We remark that a similar mechanism has also been applied for computing the deterministic approximation of $\langle f_1(W_{i_1})A_1 \dots f_k(W_{i_k})A_k \rangle$ if W_{i_1}, \dots, W_{i_k} are taken, possibly with repetitions, from a family of independent Wigner matrices (see [13, Ext. 2.13]). We give an example in the case $k = \ell = 2$. Let W, W' denote two independent GUE matrices and pick bounded deterministic matrices A_1, \dots, A_4

with $\langle A_j \rangle \neq 0$ as well as test functions f_1, \dots, f_4 satisfying Case 2 of Assumption 2.1 with $p = 3$. Then,

$$\begin{aligned} & N^2 \mathbb{E} \left[\left(\langle f_1(W) A_1 f_2(W') A_2 \rangle - \mathbb{E} \langle f_1(W) A_2 f_2(W') A_2 \rangle \right) \right. \\ & \quad \times \left. \left(\langle f_3(W) A_3 f_4(W') A_4 \rangle - \mathbb{E} \langle f_3(W) A_3 f_4(W') A_4 \rangle \right) \right] \\ &= \text{sc}_{\text{oo}}[1|3] \langle A_1 f_2(W') A_2 \rangle \langle A_3 f_4(W') A_4 \rangle + \text{sc}_{\text{o}}[1, 3] \langle A_1 f_2(W') A_2 A_3 f_4(W') A_4 \rangle \\ & \quad + \text{sc}_{\text{o}}[1] \text{sc}_{\text{o}}[3] N^2 \mathbb{E} \left[\left(\langle A_1 f_2(W') A_2 \rangle - \mathbb{E} \langle A_2 f_2(W') A_2 \rangle \right) \right. \\ & \quad \times \left. \left(\langle A_3 f_4(W') A_4 \rangle - \mathbb{E} \langle A_3 f_4(W') A_4 \rangle \right) \right] + \mathcal{O} \left(\frac{N^\varepsilon \max\{\|f_1\|_{H^3}, \|f_3\|_{H^3}\}}{\sqrt{N}} \right) \\ &= \langle A_1 A_2 \rangle \langle A_3 A_4 \rangle (\text{sc}_{\text{oo}}[1|3] \text{sc}_{\text{o}}[2] \text{sc}_{\text{o}}[4] + \text{sc}_{\text{oo}}[2|4] \text{sc}_{\text{o}}[1] \text{sc}_{\text{o}}[3]) \\ & \quad + \langle A_1 A_4 \rangle \langle A_2 A_3 \rangle \text{sc}_{\text{o}}[1, 3] \text{sc}_{\text{o}}[2, 4] + \langle A_1 A_2 A_3 A_4 \rangle \text{sc}_{\text{o}}[1, 3] \text{sc}_{\text{o}}[2] \text{sc}_{\text{o}}[4] \\ & \quad + \langle A_2 A_1 A_4 A_3 \rangle \text{sc}_{\text{o}}[1] \text{sc}_{\text{o}}[2, 4] \text{sc}_{\text{o}}[3] + \mathcal{O} \left(\frac{N^\varepsilon \max_j \|f_j\|_{H^3}}{\sqrt{N}} \right). \end{aligned}$$

In the first step, we conditioned on W' and applied Theorem 2.7 as well as Corollary 2.9, treating W' , and hence $f_j(W')$, as deterministic. After computing the leading term, Corollary 2.9 is applied again for W' . Lastly, the remaining terms are identified using the local law [12, Cor. 2.7], which yields a total of five summands. In contrast, if $W = W'$, all terms on the right-hand side of (2.14) may contribute. For $k = \ell = 2$, this yields 27 terms in total (cf. [40, Ex. 1.18]). Analogous statements hold for an arbitrary number of independent Wigner matrices with possible repetitions. We remark that the underlying combinatorial structure for n independent GUE matrices is given by the so-called *non-mixing* annular non-crossing permutations and *non-mixing* marked partitions for n colors, which was also established in [38] for the special case $f_1(x) = \dots = f_k(x) = x$ and general Wigner matrices.

2.4 Application to thermalization problems

We now specialize Theorem 2.7 to the functions $f_j(x) = e^{it_j x}$ with (N -independent) numbers $t_j \in \mathbb{R}$. Recall that

$$A(t) := e^{itW} A e^{-itW} \tag{2.20}$$

describes the Heisenberg time evolution of an observable A and that it follows from [13, Cor. 2.9] that the observables $A_1(t)$ and A_2 become thermalized for $t \gg 1$, i.e., that

$$\langle A_1(t) A_2 \rangle \approx \langle A_1 \rangle \langle A_2 \rangle$$

in the large t regime. Fixing $t \sim 1$ and using Theorem 2.7, we readily conclude that the fluctuations around the thermal value are Gaussian and can give the leading terms of the variance explicitly by taking an $t \rightarrow \infty$ limit after letting $N \rightarrow \infty$. As the randomness in $\langle A_1(t) A_2 \rangle$ cancels out if either $A_1 = \text{Id}$ or $A_2 = \text{Id}$, we omit the deterministic term $\langle A_1 \rangle \langle A_2 \rangle$ from the following result and assume w.l.o.g. that $\langle A_1 \rangle = \langle A_2 \rangle = 0$.

Corollary 2.13. *Let $\kappa_4 = 0$, $\langle A_1 \rangle = \langle A_2 \rangle = 0$, and $\|A_1\|, \|A_2\| \lesssim 1$. Then,*

$$\langle A_1(t) A_2 \rangle = \langle A_1 A_2 \rangle \frac{J_1(2t)^2}{t^2} + \frac{\xi(t)}{N} + \mathcal{O} \left(\frac{N^\varepsilon}{N^{3/2}} \right),$$

in the sense of moments, where $A_1(t)$ is as defined in (2.20), J_1 is a Bessel function of the first kind, and $\xi(t)$ is a centered Gaussian random variable. In the $t \rightarrow \infty$ limit, the variance of $\xi(t)$ satisfies the asymptotics

$$\text{Var}[\xi(t)] = \langle |A_1|^2 \rangle \langle |A_2|^2 \rangle + \mathcal{O} \left(\frac{1}{t^2} \right) \tag{2.21}$$

and we further obtain

$$\mathbb{E}\xi(t_1)\overline{\xi(t_2)} = \langle |A_1|^2 \rangle \langle |A_2|^2 \rangle \left(\frac{J_1(2(t_1 - t_2))}{t_1 - t_2} \right)^2 + \mathcal{O}\left(\frac{1}{(\min\{t_1, t_2\})^2} \right). \tag{2.22}$$

In particular, $\xi(t_1)$ and $\xi(t_2)$ are asymptotically uncorrelated if we take an $|t_1 - t_2| \rightarrow \infty$ limit after letting $N \rightarrow \infty$.

Corollary 2.13 follows directly from Theorem 2.7 and Corollary 2.9 by specifying the test functions. We carry out the details in Section 4.7 and remark that the case $\kappa_4 \neq 0$ can be treated analogously. As the proof of Corollary 2.13 only uses the macroscopic regime of Theorem 2.7, the quantity $\langle A_1(t)A_2 \rangle$ is even accessible for more general Wigner matrices using [40, Thm. 2.9]. However, the resulting expressions for the variance are much more involved and may require some terms to be computed recursively. We hence restrict the detailed discussion to the GUE case.

Remark 2.14. The $\mathcal{O}(t^{-2})$ term in (2.21) can be given more explicitly, as the limiting variance can be read off from Corollary 2.9 or [37, Sect. 2] (since t is fixed). We obtain

$$\begin{aligned} \text{Var}[\xi(t)] &= \langle |A_1|^2 \rangle \langle |A_2|^2 \rangle \left(1 - \frac{J_1(2t)^2}{t^2} \right)^2 \\ &\quad + |\langle A_1 A_2^* \rangle|^2 \left[\left(\frac{J_1(4t)}{2t} - \frac{J_1(2t)^2}{t^2} \right)^2 + 2\text{sc}_{\circ\circ}[+, +] \frac{J_1(2t)^2}{t^2} - 2\text{sc}_{\circ\circ}[+, -] \frac{J_1(2t)^2}{t^2} \right] \\ &\quad + (\langle A_1 A_2 A_1^* A_2^* \rangle + \langle A_2 A_1 A_2^* A_1^* \rangle) \left(\frac{J_1(4t)}{2t} - \frac{J_1(2t)^2}{t^2} \right) \frac{J_1(2t)}{t^2} \\ &\quad + (\langle |A_1 A_2|^2 \rangle + \langle |A_2 A_1|^2 \rangle) \left(1 - \frac{J_1(2t)^2}{t^2} \right) \frac{J_1(2t)^2}{t^2} \end{aligned}$$

where

$$\begin{aligned} \text{sc}_{\circ\circ}[+, +] &= \frac{1}{2\pi^2} \int_{-2}^2 \int_{-2}^2 \left(\frac{e^{itx} - e^{ity}}{x - y} \right)^2 \frac{4 - xy}{\sqrt{4 - x^2}\sqrt{4 - y^2}} dx dy - \frac{J_1(4t)}{2t} + \frac{J_1(2t)^2}{t^2}, \\ \text{sc}_{\circ\circ}[+, -] &= \frac{1}{2\pi^2} \int_{-2}^2 \int_{-2}^2 \frac{1 - \cos(t(x - y))}{(x - y)^2} \frac{4 - xy}{\sqrt{4 - x^2}\sqrt{4 - y^2}} dx dy - 1 + \frac{J_1(2t)^2}{t^2}. \end{aligned}$$

Note that the $\langle |A_1|^2 \rangle \langle |A_2|^2 \rangle$ term is the only one having a $\mathcal{O}(1)$ contribution, in fact, its coefficient is $1 + \mathcal{O}(t^{-3})$. The first summand of the $|\langle A_1 A_2^* \rangle|^2$ term is $\mathcal{O}(t^{-3})$ while the double integrals behave like t for $t \rightarrow \infty$. Hence the entire $|\langle A_1 A_2^* \rangle|^2$ term is $\mathcal{O}(t^{-2})$. In the remaining contributions, the largest term is $\mathcal{O}(t^{-3})$ with the lower-order terms being as small as $\mathcal{O}(t^{-6})$.

3 Central limit theorem for resolvents

In this section, we supply the recursive definitions of the set functions $\mathcal{E}[\cdot]$ and $\mathfrak{m}_2[\cdot]$ introduced in Lemmas 2.3 and 2.5, respectively, and study their properties. We further state the analog of Theorem 2.7 for the resolvent case, which constitutes the main ingredient for the proof of the multi-point functional CLT.

3.1 The $\frac{1}{N}$ term of $\mathbb{E}\langle T_{[1,k]} \rangle$

As a first step, we revisit the function $\mathcal{E}[\cdot]$, starting with its recursive definition.

Definition 3.1. Let (T_1, \dots, T_k) be an ordered set of $\mathbb{C}^{N \times N}$ matrices of the form $T_j = G_j A_j$. We define $\mathcal{E}[\cdot]$ to be the set function taking values in \mathbb{C} that satisfies the linear recursion with a source term (in the last two lines)

$$\mathcal{E}[T_1, \dots, T_k]$$

$$\begin{aligned}
 &= m_1 \left(\mathcal{E}[T_2, \dots, T_{k-1}, T_k A_1] + q_{1,k} \mathcal{E}[T_2, \dots, T_{k-1}, G_k A_1] \langle A_k \rangle \right) \\
 &\quad + \sum_{j=1}^{k-1} \mathcal{E}[T_1, \dots, T_{j-1}, G_j] \left(m_1 [T_j, \dots, T_k] + q_{1,k} m_1 [T_j, \dots, T_{k-1}, G_k] \langle A_k \rangle \right) \\
 &\quad + \sum_{j=2}^k m_1 [T_1, \dots, T_{j-1}, G_j] \left(\mathcal{E}[T_j, \dots, T_k] + q_{1,k} \mathcal{E}[T_j, \dots, T_{k-1}, G_k] \langle A_k \rangle \right) \quad (3.1) \\
 &\quad + \sum_{1 \leq r \leq s \leq t \leq k} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle M_{[r,s]} \odot (M_{[t,k]} A_k) \rangle \\
 &\quad + q_{1,k} \sum_{1 \leq r \leq s \leq t \leq k} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle M_{[r,s]} \odot M_{[t,k]} \rangle \langle A_k \rangle
 \end{aligned}$$

and the initial condition $\mathcal{E}[\emptyset] = 0$. Note that $\mathcal{E}[T_1, \dots, T_{j-1}, G_j] = \mathcal{E}[G_1]$ for $j = 1$ and $\mathcal{E}[T_j, \dots, T_{k-1}, G_k] = \mathcal{E}[G_k]$ for $j = k$. Moreover, recall that \odot denotes the Hadamard product, $M_{[r]}$ was defined through the recursion in (1.13), $m_1[\cdot]$ was defined in (1.17), and $q_{1,k} = \frac{m_1 m_k}{1 - m_1 m_k}$.

We remark that $M_{[k]}$ is diagonal whenever $A_1 = \dots = A_k = \text{Id}$. In this case, the last two lines of (3.1) are readily evaluated and the recursion simplifies to

$$\begin{aligned}
 \mathcal{E}[G_1, \dots, G_k] &= \frac{m_1}{1 - m_1 m_k} \left(\mathcal{E}[G_2, \dots, G_k] + \sum_{j=1}^{k-1} \mathcal{E}[G_1, \dots, G_j] m[j, \dots, k] \right. \\
 &\quad \left. + \sum_{j=2}^k m[1, \dots, j] \mathcal{E}[G_j, \dots, G_k] \right. \\
 &\quad \left. + \sum_{1 \leq r \leq s \leq t \leq k} m[1, \dots, r] m[r, \dots, s] m[s, \dots, t] m[t, \dots, k] \right),
 \end{aligned}$$

where $m[\cdot]$ denotes the iterated divided differences in (1.12). Note that $\mathcal{E}[T_1, \dots, T_k]$ is generally not of order one, but its size is given in terms of η_* and the number of traceless matrices among A_1, \dots, A_k . We give a simple bound in the following lemma. The proof is carried out in Section 4.1.

Lemma 3.2. *Under the assumptions of Lemma 2.3, let a among the matrices A_1, \dots, A_k be traceless. Then,*

$$|\mathcal{E}[T_1, \dots, T_k]| \lesssim \frac{1}{\eta_*^{k-1 - \lceil a/2 \rceil}}. \quad (3.2)$$

Remark 3.3. Lemma 3.2 shows that the sub-leading term in (2.4) involving $\mathcal{E}[\cdot]$ may be smaller than the $\mathcal{O}(N^\varepsilon / (N \sqrt{N} \eta_* \eta_*^{k-a/2}))$ error term in some regimes. However, as we only apply (2.4) in the form

$$\mathbb{E} \langle T_{[1,k]} \rangle = m_1 [T_1, \dots, T_k] + \mathcal{O} \left(\frac{1}{N \eta_*^{k-1-a/2}} + \frac{N^\varepsilon}{N \sqrt{N} \eta_* \eta_*^{k-a/2}} \right)$$

for the proof of the CLT in the resolvent case in Section 4.4, a more careful resolution of the error is not needed.

Note that applying (3.1) once yields the formula

$$\mathcal{E}[T_1] = \langle A_1 \rangle \frac{m_1^5}{1 - m_1^2} = \langle A_1 \rangle m_1' m_1^3,$$

and we readily reobtain (93) of [14] by Lemma 2.3.

Having identified the function $\mathcal{E}[\cdot]$ as the $1/N$ term of $\mathbb{E}\langle T_{[1,k]} \rangle$, we may use identities that are valid on the random matrix side, i.e., the left-hand side of (2.4), to derive further identities for $\mathcal{E}[\cdot]$ (cf. the “meta argument” below [12, Lem. 4.1]). They are listed in Corollary 3.4 below and the proof is given in Appendix A. In fact, these identities can also be proven from Definition 3.1 directly, however, using (2.4) allows for a shorter proof. Note that $\mathfrak{m}[\cdot]$ satisfies the same properties by [13, Lem. 5.4].

Corollary 3.4. *Let $k \in \mathbb{N}$ and (T_1, \dots, T_k) be an ordered set of $\mathbb{C}^{N \times N}$ matrices of the form $T_j = G_j A_j$. Then*

- (i) $\mathcal{E}[\cdot]$ is cyclic in the sense that $\mathcal{E}[T_1, \dots, T_k] = \mathcal{E}[T_2, \dots, T_k, T_1]$.
- (ii) Whenever $z_1 \neq z_k$ and $A_k = \text{Id}$, we have

$$\mathcal{E}[T_1, \dots, T_{k-1}, G_k] = \frac{\mathcal{E}[T_2, \dots, T_{k-1}, G_k A_1] - \mathcal{E}[T_1, \dots, T_{k-1}]}{z_k - z_1}. \tag{3.3}$$

- (iii) Whenever $A_1 = \dots = A_k = \text{Id}$ and the spectral parameters z_1, \dots, z_k are distinct, $\mathcal{E}[\cdot]$ has a divided difference structure, i.e.,

$$\mathcal{E}[G_1, \dots, G_k] = \frac{\mathcal{E}[G_2, \dots, G_k] - \mathcal{E}[G_1, \dots, G_{k-1}]}{z_k - z_1} \tag{3.4}$$

and we have the closed formula

$$\mathcal{E}[G_1, \dots, G_k] = \sum_{j=1}^k \prod_{i \neq j} \frac{1}{z_i - z_j} \mathcal{E}[G_j] = \sum_{j=1}^k \prod_{i \neq j} \frac{m[i, j]}{m_i - m_j} \mathcal{E}[G_j] \tag{3.5}$$

with $m[\cdot]$ as in (1.19) and $\mathcal{E}[G_j] = m'_j m_j^3$. Moreover, $\mathcal{E}[\cdot]$ is invariant under any permutation of z_1, \dots, z_k in this case.

3.2 Statement of the resolvent central limit theorem

The main result of this section is a CLT for resolvents that identifies the joint distribution of multiple modes of the type (2.3), i.e., $X_{\alpha_i}^{(k_i, a_i)}$ with different k_i, a_i , and α_i , as asymptotically Gaussian in the sense of moments. We start by defining the set function $\mathfrak{m}_2[\cdot, \cdot]$, which characterizes the limiting covariance of the random variables X_α and X_β involving two distinct multi-indices α and β . Note that we only use the following recursive definition in the present work. However, closed formulas are obtained in the companion paper [40].

Definition 3.5. *Let $S_1 = (T_1, \dots, T_{k'})$ and $S_2 = (T_{k'+1}, \dots, T_{k'+\ell'})$ be two (ordered) finite sets of complex $N \times N$ -matrices of the form $T_j = G_j A_j$. We define $\mathfrak{m}_2[\cdot, \cdot]$ as the (deterministic) function of pairs of sets S_1, S_2 with values in \mathbb{C} and the following properties:*

- (i) *Symmetry: $\mathfrak{m}_2[\cdot, \cdot]$ is symmetric under the interchanging of its arguments, i.e., for any sets $B_1 \subseteq S_1, B_2 \subseteq S_2$ we have*

$$\mathfrak{m}_2[(T_i, i \in B_1) | (T_j, j \in B_2)] = \mathfrak{m}_2[(T_j, j \in B_2) | (T_i, i \in B_1)].$$

- (ii) *Initial condition: For any sets $B_1 \subseteq S_1, B_2 \subseteq S_2$ we have*

$$\mathfrak{m}_2[(T_i, i \in B_1) | \emptyset] = \mathfrak{m}_2[\emptyset | (T_j, j \in B_2)] = 0. \tag{3.6}$$

- (iii) *Recursion: Let $B_1 \subseteq S_1$ and $B_2 \subseteq S_2$ be ordered subsets with $|B_1| = k \leq k'$ and $|B_2| = \ell \leq \ell'$ elements, respectively. We index the matrices in B_1 by $[k]$ and the*

matrices in B_2 by $[k + 1, k + \ell]$. The function $\mathfrak{m}_2[\cdot|\cdot]$ satisfies the following linear recursion

$$\begin{aligned} & \mathfrak{m}_2[T_1, \dots, T_k | T_{k+1}, \dots, T_{k+\ell}] \\ &= m_1 \left(\mathfrak{m}_2[T_2, \dots, T_{k-1}, G_k A_k A_1 | T_{k+1}, \dots, T_{k+\ell}] \right. \\ & \quad + q_{1,k} \mathfrak{m}_2[T_2, \dots, T_{k-1}, G_k A_1 | T_{k+1}, \dots, T_{k+\ell}] \langle A_k \rangle \\ & \quad + \sum_{j=1}^{k-1} \mathfrak{m}_2[T_1, \dots, T_{j-1}, G_j | T_{k+1}, \dots, T_{k+\ell}] \\ & \quad \quad \times \left(\mathfrak{m}_1[T_j, \dots, T_k] + q_{1,k} \mathfrak{m}_1[T_j, \dots, T_{k-1}, G_k] \langle A_k \rangle \right) \\ & \quad + \sum_{j=2}^k \mathfrak{m}_1[T_1, \dots, T_{j-1}, G_j] \left(\mathfrak{m}_2[T_j, \dots, T_k | T_{k+1}, \dots, T_{k+\ell}] \right. \\ & \quad \quad \left. + q_{1,k} \mathfrak{m}_2[T_j, \dots, T_{k-1}, G_k | T_{k+1}, \dots, T_{k+\ell}] \langle A_k \rangle \right) + \mathfrak{s}_{GUE} + \mathfrak{s}_\kappa \Big) \end{aligned} \tag{3.7}$$

where the source terms \mathfrak{s}_{GUE} and \mathfrak{s}_κ are given by

$$\begin{aligned} \mathfrak{s}_{GUE} &:= \sum_{j=1}^{\ell} \left(\mathfrak{m}_1[T_1, \dots, T_k, T_{k+j}, \dots, T_{k+j-1}, G_{k+j}] \right. \\ & \quad \left. + q_{1,k} \mathfrak{m}_1[T_1, \dots, T_{k-1}, G_k, T_{k+j}, \dots, T_{k+j-1}, G_{k+j}] \langle A_k \rangle \right) \tag{3.8} \\ \mathfrak{s}_\kappa &:= \kappa_4 \sum_{r=1}^k \sum_{s=k+1}^{k+\ell} \left(\sum_{t=k+1}^s \langle M_{[r]} \odot M_{(s, \dots, k+\ell, k+1, \dots, t)} \rangle \langle (M_{[r,k]} A_k) \odot M_{[t,s]} \rangle \right. \\ & \quad \left. + \sum_{t=s}^{k+\ell} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle (M_{[r,k]} A_k) \odot M_{(t, \dots, k+\ell, k+1, \dots, s)} \rangle \right) \\ & \quad + \kappa_4 q_{1,k} \sum_{r=1}^k \sum_{s=k+1}^{k+\ell} \left(\sum_{t=k+1}^s \langle M_{[r]} \odot M_{(s, \dots, k+\ell, k+1, \dots, t)} \rangle \langle M_{[r,k]} \odot M_{[t,s]} \rangle \right. \\ & \quad \left. + \sum_{t=s}^{k+\ell} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle M_{[r,k]} \odot M_{(t, \dots, k+\ell, k+1, \dots, s)} \rangle \right) \langle A_k \rangle. \end{aligned} \tag{3.9}$$

Recall that \odot denotes the Hadamard product, $q_{1,k} = \frac{m_1 m_k}{1 - m_1 m_k}$ with m_1, m_k as in (1.8), $M_{(\dots)}$ was introduced in (1.14), and $\mathfrak{m}_1[\cdot]$ was defined in (1.17).

Remark 3.6. The special role of m_1 in (3.7) is a result of the identity (4.12) used for the proof of Lemma 2.5 in Section 4.3 below. Similar to the recursion for $\mathfrak{m}_1[\cdot]$ in [12, Lem. 4.1], it is possible to derive a version of (3.7) for every $j = 2, \dots, k$ that singles out the factor m_j instead of m_1 on the right-hand side, i.e., (3.7) is only one element in a family of equivalent recursions for $\mathfrak{m}_2[\cdot|\cdot]$.

The linearity of the recursion and the two different types of source terms induce the decomposition (2.10), where $\mathfrak{m}_{GUE}[\cdot|\cdot]$ satisfies (3.7) for $\kappa_4 = 0$, and $\kappa_4 \mathfrak{m}_\kappa[\cdot|\cdot]$ satisfies (3.7) without \mathfrak{s}_{GUE} . We remark that by [13, Thm. 3.4], both $\mathfrak{m}_1[\cdot]$ and $M_{(\cdot)}$ are fully expressible as functions of $A_1, \dots, A_{k+\ell}$ and $m_1, \dots, m_{k+\ell}$. Hence, the same holds for the source term $\mathfrak{s}_{GUE} + \mathfrak{s}_\kappa$ in (3.7), eventually making $\mathfrak{m}_2[\cdot|\cdot]$ a function of the same quantities. Similarly, we have the decomposition

$$\tilde{\mathfrak{m}}[\cdot|\cdot] = \tilde{\mathfrak{m}}_{GUE}[\cdot|\cdot] + \kappa_4 \tilde{\mathfrak{m}}_\kappa[\cdot|\cdot], \tag{3.10}$$

for the function $\tilde{m}[\cdot|\cdot]$ defined by the relation

$$\tilde{m}[1, \dots, k|k+1, \dots, k+\ell] := m_2[G_1, \dots, G_k|G_{k+1}, \dots, G_{k+\ell}] \tag{3.11}$$

in the special case $A_1 = \dots = A_k = \text{Id}$.

Next, we consider the size of $m_2[\cdot|\cdot]$. We have the following bounds, which we prove in Section 4.1.

Lemma 3.7. *Under the assumptions of Lemma 2.5, we have the estimates*

$$|m_{GUE}[T_1, \dots, T_k|T_{k+1}, \dots, T_{k+\ell}]| \lesssim \frac{1}{\eta_*^{k+\ell - \lceil (a+b)/2 \rceil}}, \tag{3.12}$$

$$|m_\kappa[T_1, \dots, T_k|T_{k+1}, \dots, T_{k+\ell}]| \lesssim \frac{1}{\eta_*^{k+\ell - 1 - \lceil (a+b)/2 \rceil}}. \tag{3.13}$$

In particular, $m_\kappa[\cdot|\cdot]$ is dominated by $m_{GUE}[\cdot|\cdot]$ on all mesoscopic scales and it holds that

$$|m_2[T_1, \dots, T_k|T_{k+1}, \dots, T_{k+\ell}]| \lesssim \frac{1}{\eta_*^{k+\ell - \lceil (a+b)/2 \rceil}}. \tag{3.14}$$

After this preparation, we state a CLT for resolvents which generalizes [14, Thm. 4.1] to handle resolvent chains of arbitrary length in the setting considered. The proof follows by induction on the number of factors $X_{\alpha_j}^{(k_j, a_j)}$ using the bounds from Lemma 2.5 and its proof as input. We give it in Section 4.4.

Theorem 3.8 (CLT for resolvents). *Fix $p \in \mathbb{N}$, let $\alpha_1, \dots, \alpha_p$ be multi-indices, and let W be a Wigner matrix satisfying Assumption 1.1. Moreover, for every $j = 1, \dots, p$ pick a set of spectral parameters $z_1^{(j)}, \dots, z_{k_j}^{(j)}$ such that either all sets satisfy Case 1 or all sets satisfy Case 2 of Assumption 2.2, and denote $\eta_* = \min_{i,j} |\Im z_i^{(j)}|$. Moreover, for every $j = 1, \dots, p$, pick deterministic matrices $A_1^{(j)}, \dots, A_{k_j}^{(j)}$ with $\|A_i^{(j)}\| \lesssim 1$ such that a_j of them are traceless. Then,*

$$N^p \mathbb{E} \left(\prod_{j=1}^p X_{\alpha_j}^{(k_j, a_j)} \right) = \sum_{Q \in \text{Pair}([p])} \prod_{\{i,j\} \in Q} m_2[\alpha_i|\alpha_j] + \mathcal{O} \left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \prod_{l=1}^p \eta_*^{k_l - a_l/2}} \right) \tag{3.15}$$

for any $\varepsilon > 0$. Here, $m_2[\cdot|\cdot]$ is as in Definition 3.5 and $\text{Pair}(S)$ denotes the pairings of a set S . Equation (3.15) establishes an asymptotic version of Wick’s rule and hence identifies the joint limiting distribution of the random variables $(X_{\alpha_j}^{(k_j, a_j)})_j$ as asymptotically complex Gaussian in the sense of moments in the limit $N\eta_* \rightarrow \infty$.

Remark 3.9 (Independent modes). Note that (3.14) implies that two modes $X_{\alpha_1}^{(k_1, a_1)}$ and $X_{\alpha_2}^{(k_2, a_2)}$ in Theorem 3.8 are asymptotically uncorrelated whenever $a_1 + a_2$ is odd and $\eta_* \ll 1$. This feature is exclusive to the mesoscopic regime, as all modes contribute equally in the macroscopic regime, i.e., if $\eta_* \gtrsim 1$. In the case $k = 1$, we may also write the deterministic matrix as

$$A = \langle A \rangle \text{Id} + \mathring{A}_d + \mathring{A}_{od} \tag{3.16}$$

with A_d and A_{od} denoting the diagonal and off-diagonal part of $\mathring{A} = A - \langle A \rangle \text{Id}$, respectively. The three resulting modes $\langle A \rangle \text{Tr} f(W)$, $\text{Tr} f(W) \mathring{A}_d$, and $\text{Tr} f(W) \mathring{A}_{od}$ are asymptotically uncorrelated as $N\eta_* \rightarrow \infty$ (cf. [14, Thm. 2.4]). Since this is a consequence of $\langle B_1 B_2 \rangle$ and $\langle B_1 \rangle \langle B_2 \rangle$ vanishing whenever $B_1 \neq B_2$ and $B_1, B_2 \in \{\text{Id}, \mathring{A}_d, \mathring{A}_{od}\}$, this phenomenon is exclusive to the $k = 1$ case and decomposing A_1, \dots, A_k for $k \geq 2$ according to (3.16) does not yield 3^k uncorrelated modes in general. As an example, consider the case $k = 2$ with two deterministic matrices A and B , where applying (3.16) gives rise to 9 distinct modes. Here, computing the covariance, e.g., of $\text{Tr}[G_1 \mathring{A}_d G_2 \mathring{B}_d]$ and $\text{Tr}[G_1 \mathring{A}_{od} G_2 \mathring{B}_{od}]$, yields terms that involve longer products, e.g., $\langle \mathring{A}_d \mathring{A}_{od} \mathring{B}_d \mathring{B}_{od} \rangle$, which do not vanish in general.

Similar to Corollary 3.4, we may use identities that are valid on the random matrix side, i.e., the left-hand side of (2.7), to derive further identities among the recursively defined quantities $\tilde{m}[\cdot|\cdot]$ and $m_2[\cdot|\cdot]$. The proof is analogous to the proof of Corollary 3.4 and hence omitted. We refer to Appendix A for the setup of the necessary “meta argument”.

Corollary 3.10. *Let $S_1, S_2 \neq \emptyset$ be two ordered multi-sets. Then*

- (i) $\tilde{m}[S_1|S_2]$ is invariant under any permutation of the elements of S_1 as well as S_2 .
- (ii) $m_2[\cdot|\cdot]$ is cyclic in the sense that

$$m_2[(T_j, j \in S_1)|T_1, \dots, T_k] = m_2[(T_j, j \in S_1)|T_2, \dots, T_k, T_1].$$

- (iii) Whenever the spectral parameters indexed by S_1 and S_2 are distinct, $\tilde{m}[\cdot|\cdot]$ has an entry-wise divided difference structure, i.e.,

$$\tilde{m}[S_1|1, \dots, k] = \frac{\tilde{m}[S_1|2, \dots, k] - \tilde{m}[S_1|1, \dots, k-1]}{z_k - z_1}, \tag{3.17}$$

and we have the closed formula

$$\begin{aligned} \tilde{m}[S_1|S_2] &= \sum_{(s,t) \in S_1 \times S_2} \left(\prod_{\substack{i \in S_1, \\ i \neq s}} \frac{1}{z_i - z_s} \prod_{\substack{j \in S_2, \\ j \neq t}} \frac{1}{z_j - z_t} \right) \tilde{m}[s|t] \\ &= \sum_{(s,t) \in S_1 \times S_2} \left(\prod_{\substack{i \in S_1, \\ i \neq s}} \frac{m[i, s]}{m_i - m_s} \prod_{\substack{j \in S_2, \\ j \neq t}} \frac{m[j, t]}{m_j - m_t} \right) \tilde{m}[s|t] \end{aligned} \tag{3.18}$$

with $\tilde{m}[s|t] = \frac{m'_s m'_t}{(1 - m_s m_t)^2}$. Recall that $\tilde{m}[\cdot|\cdot]$ was defined in (3.11).

- (iv) Whenever $z_1 \neq z_k$ and $A_k = \text{Id}$, we further have

$$\begin{aligned} &m_2[(T_j, j \in S_1)|T_1, \dots, T_{k-1}, G_k] \\ &= \frac{m_2[(T_j, j \in S_1)|T_2, \dots, T_{k-1}, G_k A_1] - m_2[(T_j, j \in S_1)|T_1, \dots, T_{k-1}]}{z_k - z_1}. \end{aligned} \tag{3.19}$$

Moreover, we have the following alternative integral representation for $\tilde{m}_{GUE}[\cdot|\cdot]$ (cf. decomposition in (3.10)). The proof of Corollary 3.11 is carried out in Section 4.3 below.

Corollary 3.11. *Let $k, \ell \in \mathbb{N}$. Then,*

$$\begin{aligned} &\tilde{m}_{GUE}[1, \dots, k|k+1, \dots, k+\ell] \\ &= \frac{1}{2} \iint \left(\sum_{i=1}^k \frac{1}{(x - z_i)^2} \cdot \prod_{j \neq i} \frac{1}{x - z_j} \right) \left(\sum_{i=k+1}^{k+\ell} \frac{1}{(y - z_i)^2} \cdot \prod_{j \neq i} \frac{1}{y - z_j} \right) u(x, y) dx dy \end{aligned} \tag{3.20}$$

with the kernel $u : [-2, 2] \times [-2, 2] \rightarrow \mathbb{R}$ in (2.19).

Remark 3.12. It is readily checked that the kernel u is non-negative and has a logarithmic singularity at $x = y$. Using that the two-body stability operator of the underlying Dyson equation (1.9) is given by $\mathcal{B}(z_1, z_2) = 1 - m(z_1)m(z_2)$, we can also express (2.19) in terms of 1×1 determinants as

$$u(x, y) = -\frac{1}{2\pi^2} \Re(\ln(\det[\mathcal{B}(x + i0, y + i0)]) - \ln(\det[\mathcal{B}(x + i0, y + i0)]))$$

to match the formulas in [41, Sect. 7] for $k = \ell = 1$ and $A_1 = A_2 = \text{Id}$. Note that W being a GUE matrix corresponds to the choice $\beta = 2$ and $\mathcal{C}^{(4)} = 0$ in the notation of [41].

4 Proofs

4.1 Proof of Lemmas 3.2 and 3.7 (size of $\mathcal{E}[\cdot]$ and $m_2[\cdot|\cdot]$)

In this section, we prove the estimates identifying the size of $\mathcal{E}[\cdot]$ and $m_2[\cdot|\cdot]$. We start by noting two bounds for $q_{1,2}$ that are used for both proofs.

Lemma 4.1. *Let $z_1, z_2 \in \mathbb{C}$ and define the constants $\eta_* := \min\{|\Im z_1|, |\Im z_2|\}$ as well as $\zeta := \pi^{-1} \min\{|\Im m_1|, |\Im m_2|\}$. Then,*

$$|q_{1,2}| = |m[1, 2]| = \left| \frac{m_1 m_2}{1 - m_1 m_2} \right| \lesssim \begin{cases} \zeta^{-1}, & \text{if } \Im z_1, \Im z_2 \text{ have the same sign,} \\ \eta_*^{-1}, & \text{if } \Im z_1, \Im z_2 \text{ have opposite signs.} \end{cases}$$

The estimates in Lemma 4.1 are immediate from the explicit form $m(z) = (-z + \sqrt{z^2 - 4})/2$ of the solution to (1.9). Next, we establish the estimate for $\mathcal{E}[\cdot]$.

Proof of Lemma 3.2. We show (3.2) by induction. The base case $k = 1$ readily follows using (2.6) and the explicit form $m(z) = (-z + \sqrt{z^2 - 4})/2$. Note that $\Im m(z)$ can be bounded from below independently of η_* for both the macroscopic scale and the bulk regime of the mesoscopic scales (cf. Assumption 2.2). Next, assume that the bound in (3.2) holds for up to $k - 1$ matrices T_1, \dots, T_{k-1} . W.l.o.g. assume further that A_k is either traceless or equal to the identity matrix.

We start by considering the case where $A_k = \text{Id}$ and $\Im z_1$ and $\Im z_k$ have opposite signs. Here, (3.2) follows immediately by using Corollary 3.4(ii) and applying the induction hypothesis for $\mathcal{E}[T_2, \dots, T_{k-1}, G_k A_1]$ and $\mathcal{E}[T_1, \dots, T_{k-1}]$, respectively. Note that $|z_1 - z_k| \geq 2\eta_*$ by assumption, which completes the bound for the right-hand side of (3.3).

In the remaining cases, (3.2) follows from the recursion (3.1), which allows rewriting $\mathcal{E}[T_1, \dots, T_k]$ in terms of $m_1[\cdot]$ and values of $\mathcal{E}[\cdot]$ for which the induction hypothesis applies. Whenever $\Im z_1$ and $\Im z_k$ have the same sign, Lemma 4.1 yields $q_{1,k} \lesssim \pi / \min\{|\Im m_1|, |\Im m_k|\}$. Thus, estimating the right-hand side of (3.1) using Lemma 1.4 and the induction hypothesis yields the claim. In particular, we obtain η_* with the exponent $-(k - 1 - \lceil \frac{\alpha}{2} \rceil)$ by using the inequality $\lceil \frac{x}{2} \rceil + \lceil \frac{y}{2} \rceil \geq \lceil \frac{x+y}{2} \rceil$ for $x, y \in \mathbb{N}$ to combine the powers of η_* when products with $M_{[\cdot]}$ are considered.

Whenever $\Im z_1$ and $\Im z_k$ have opposite signs, the prefactor $q_{1,k}$ is of size η_*^{-1} . However, it only remains to consider the case $\langle A_k \rangle = 0$ for this setting, in which the terms involving $q_{1,k}$ on the right-hand side of (3.1) do not contribute. Hence, (3.2) again follows from Lemma 1.4 and the induction hypothesis. \square

Next, we show the estimates for $m_{GUE}[\cdot|\cdot]$, $m_\kappa[\cdot|\cdot]$, and $m_2[\cdot|\cdot]$. To illustrate the tools at hand, the bound for $m_{GUE}[\cdot|\cdot]$ is obtained from the explicit formula in [40, Thm. 2.4] while the bound for $m_\kappa[\cdot|\cdot]$ is proved using the recursion (3.7). We start with a lemma.

Lemma 4.2. *Under the assumptions of Lemma 2.5, let $A_1 = \dots = A_{k+\ell} = \text{Id}$ and $\kappa_4 = 0$. Then,*

$$|\tilde{m}_{GUE}[1, \dots, k|k+1, \dots, k+\ell]| \lesssim \frac{1}{\eta_*^{k+\ell}} \tag{4.1}$$

$$|m_{\circ\circ}[1, \dots, k|k+1, \dots, k+\ell]| \lesssim \frac{1}{\eta_*^{k+\ell}} \tag{4.2}$$

where $m_{\circ\circ}[\cdot|\cdot]$ denotes the second-order free cumulant function associated with the iterated divided differences $m[\cdot]$ and $\tilde{m}_{GUE}[\cdot|\cdot]$.

If all $\Im z_j$ have the same sign, $\tilde{m}_{GUE}[\cdot|\cdot]$ stays bounded due to the divided difference structure (3.18) and smoothness of $\tilde{m}_{GUE}[s|t]$. Hence, the bound is not sharp in this

case. We expect, however, that (4.1) is generically sharp if not all $\Im z_j$ have the same sign. Note that a sharp bound in (4.1) immediately implies sharpness of (4.2) and (3.12) using the same argument as in [12, App. A].

Proof of Lemma 4.2. The bound (4.1) follows by induction on k and ℓ . As the base case $k = \ell = 1$ is covered by Example 2.8, assume that the bound for $\tilde{m}_{GUE}[\cdot|\cdot]$ holds for up to $k - 1$ indices in the first argument and a fixed number ℓ indices in the second argument. Recall that $\tilde{m}_{GUE}[\cdot|\cdot]$ is symmetric under the interchanging of its arguments by Definition 3.5(i) such that it is sufficient to carry out the induction step for one of the arguments only. We distinguish two cases for z_k depending on the sign of its imaginary part.

Case 1: $\Im z_1$ and $\Im z_k$ have the same sign: We note that $|q_{1,k}| \leq \pi / \min\{|\Im m_1|, |\Im m_k|\}$ by Lemma 4.1, where the right-hand side can be bounded from above independently of η_* under Assumption 2.2. Rewriting $\tilde{m}_{GUE}[1, \dots, k|k + 1, \dots, k + \ell]$ using the recursion (3.7), the bound (4.1) follows directly from the induction hypothesis and the estimate for $m[\cdot]$ from Lemma 1.4.

Case 2: $\Im z_1$ and $\Im z_k$ have opposite signs: Recalling that $\tilde{m}_{GUE}[\cdot|\cdot]$ has a divided difference structure (cf. Corollary 3.10), it follows from the induction hypothesis that

$$\begin{aligned} & |\tilde{m}_{GUE}[1, \dots, k|k + 1, \dots, k + \ell]| \\ &= \left| \frac{\tilde{m}_{GUE}[2, \dots, k|k + 1, \dots, k + \ell] - \tilde{m}_{GUE}[1, \dots, k - 1|k + 1, \dots, k + \ell]}{z_1 - z_k} \right| \\ &\lesssim \frac{1}{\eta_*^{k-1+\ell} |z_1 - z_k|}. \end{aligned}$$

As $\Im z_1$ and $\Im z_k$ are assumed to have opposite signs, we have $|z_1 - z_k| \geq 2\eta_*$, which gives (4.1). This concludes the induction step.

The bound (4.2) for $m_{\circ\circ}[\cdot|\cdot]$ is an immediate consequence of the second-order moment-cumulant relation (2.13) as well as the estimates in (4.1) and Lemma 1.4. Note that resolving the relation for $m_{\circ\circ}[1, \dots, k|k + 1, \dots, k + \ell]$ is straightforward, as its coefficient in the representation of $\tilde{m}_{GUE}[1, \dots, k|k + 1, \dots, k + \ell]$ is equal to one. \square

Proof of Lemma 3.7. Given Lemma 4.2, (3.12) readily follows from [40, Thm. 2.4]. We omit the details but include a brief sketch of the argument for the convenience of the reader. First, consider the case $k = \ell = 2$. Here,

$$\begin{aligned} \mathfrak{m}_{GUE}[T_1, T_2|T_3, T_4] &= \langle A_1 \rangle \langle A_2 \rangle \langle A_3 \rangle \langle A_4 \rangle m_{\circ\circ}[1, 2|3, 4] + \langle A_1 A_2 \rangle \langle A_3 \rangle \langle A_4 \rangle m_{\circ\circ}[1|3, 4] m_{\circ}[2] \\ &\quad + \langle A_1 A_2 \rangle \langle A_3 A_4 \rangle m_{\circ\circ}[1|3] m_{\circ}[2] m_{\circ}[4] + \dots \\ &\quad + \langle A_1 \rangle \langle A_2 A_4 \rangle \langle A_3 \rangle m_{\circ}[1, 2, 3, 4] + \langle A_1 \rangle \langle A_2 A_3 A_4 \rangle m_{\circ}[1, 2, 3] m_{\circ}[4] \\ &\quad + \dots, \end{aligned}$$

where $m_{\circ}[\cdot]$ and $m_{\circ\circ}[\cdot|\cdot]$ denote the first and second-order free cumulant functions associated with the iterated divided differences $m[\cdot]$ and $\tilde{m}_{GUE}[\cdot|\cdot]$. Note that the Kreweras complement is defined differently for non-crossing partitions and annular non-crossing permutations, and that both types of terms appear in the explicit formula for $\mathfrak{m}_{GUE}[\cdot|\cdot]$ in [40, Thm. 2.4] (similarly to (2.14) above). By Lemmas 1.4 and 4.2, the largest contribution is obtained from $m_{\circ\circ}[1, 2|3, 4]$ and $m_{\circ}[1, 2, 3, 4]$, respectively, yielding a total bound of order η_*^{-4} . Note, however, that these leading terms may vanish if at least one of the deterministic matrices is traceless, resulting in a smaller bound overall. This effect gets stronger the more traceless matrices are included. In the case that all four matrices A_1, \dots, A_4 are traceless, only 10 out of the initial 29 terms contributing to $\mathfrak{m}_{GUE}[T_1, T_2|T_3, T_4]$ remain, yielding a bound of order η_*^{-2} .

For general k and ℓ , the estimate for $\mathfrak{m}_{GUE}[\cdot|\cdot]$ is obtained from a careful balancing of the sizes of the largest cycles (resp. blocks) in the non-crossing permutation (resp. partition) π which governs the contribution of $m_\circ[\cdot|\cdot]$ (resp. $m_{\circ\circ}[\cdot|\cdot]$), and its Kreweras complement $K(\pi)$ which governs the contribution of $A_1, \dots, A_{k+\ell}$. This yields

$$\left| \sum_{\pi \in NCP(k,\ell)} \left(\prod_{B \in K(\pi)} \left\langle \prod_{j \in B} A_j \right\rangle \right) \prod_{B \in \pi} m_\circ[B] \right| \lesssim \left(\frac{1}{\eta_*} \right)^{k+\ell - \lceil (a+b)/2 \rceil}$$

where the leading contribution is obtained for an annular non-crossing permutation π with $|K(\pi)| \leq k + \ell - \lceil (a+b)/2 \rceil$, and

$$\begin{aligned} & \left| \sum_{\substack{\pi_1 \times \pi_2 \in NCP(k) \times NCP(\ell), \\ U_1 \in \pi_1, U_2 \in \pi_2 \text{ marked}}} \left(\prod_{\substack{B \in K(\pi_1) \\ \cup K(\pi_2)}} \left\langle \prod_{j \in B} A_j \right\rangle \right) m_{\circ\circ}[U_1|U_2] \prod_{\substack{B_1 \in \pi_1 \setminus U_1, \\ B_2 \in \pi_2 \setminus U_2}} m_\circ[B_1] m_\circ[B_2] \right| \\ & \lesssim \left(\frac{1}{\eta_*} \right)^{k+\ell - \lceil a/2 \rceil - \lceil b/2 \rceil}. \end{aligned}$$

where the leading contribution is obtained for a marked partition $\pi = \pi_1 \times \pi_2$ satisfying $|K(\pi_1)| \leq k - \lceil a/2 \rceil$ and $|K(\pi_2)| \leq \ell - \lceil b/2 \rceil$. In particular, the two parts of $\mathfrak{m}_{GUE}[\cdot|\cdot]$ only contribute equally to (3.12) if $\lceil \frac{a}{2} \rceil + \lceil \frac{b}{2} \rceil = \lceil \frac{a+b}{2} \rceil$.

The bound (3.13) follows by induction on k and ℓ . As the base case $k = \ell = 1$ is again covered by Example 2.8, assume that the bound for $\mathfrak{m}_\kappa[\cdot|\cdot]$ holds for a multi-index of length at most $k - 1$ in the first argument and a multi-index of length ℓ in the second argument. Recalling that $\mathfrak{m}_\kappa[\cdot|\cdot]$ is symmetric under the interchanging of its arguments, it is sufficient to carry out the induction step for one of the arguments only. To simplify notation, set $\beta = \{(z_{k+1}, A_{k+1}), \dots, (z_{k+\ell}, A_{k+\ell})\}$. We further assume w.l.o.g. that each A_j is either traceless or equal to the identity matrix. Similar to the proof of Lemma 3.2, we distinguish two cases depending on the deterministic matrices A_1, \dots, A_k .

Case 1 ($\exists j$ such that $A_j = \text{Id}$): We start by noting that $\mathfrak{m}_{GUE}[\cdot|\cdot]$ satisfies (i)-(iv) of Corollary 3.10. This implies that the same holds for $\mathfrak{m}_\kappa[\cdot|\cdot] = \mathfrak{m}_2[\cdot|\cdot] - \mathfrak{m}_{GUE}[\cdot|\cdot]$. In the following, we denote $s := \text{sign}(\Im z_j) = \text{sign}(\Im z_{j+1})$ whenever $\Im z_j$ and $\Im z_{j+1}$ have the same sign. Using the divided difference structure, we rewrite $\mathfrak{m}_\kappa[T_1, \dots, T_k | T_{k+1}, \dots, T_{k+\ell}]$ as a contour integral

$$\begin{aligned} & \mathfrak{m}_\kappa[T_1, \dots, T_{j-1}, G_j, T_{j+1}, \dots, T_k | \beta] \tag{4.3} \\ & = \frac{1}{2\pi i} \int_{\mathbb{R}} \frac{\mathfrak{m}_\kappa[\dots, T_{j-1}, G(x+i\eta)A_{j+1}, \dots | \beta] - \mathfrak{m}_\kappa[\dots, T_{j-1}, G(x-i\eta)A_{j+1}, \dots | \beta]}{(x+i\eta - z_j)(x+i\eta - z_{j+1})} dx \end{aligned}$$

whenever $0 < \eta < \Im z_j, \Im z_{j+1}$ ($s = 1$) or $\Im z_j, \Im z_{j+1} < -\eta < 0$ ($s = -1$), or as

$$\begin{aligned} & \mathfrak{m}_\kappa[T_1, \dots, T_{j-1}, G_j, T_{j+1}, \dots, T_k | \beta] \tag{4.4} \\ & = \frac{\mathfrak{m}_\kappa[T_1, \dots, T_{j-1}, G_j A_{j+1}, \dots, T_k | \beta] - \mathfrak{m}_\kappa[T_1, \dots, T_{j-1}, T_{j+1}, T_{j+2}, \dots, T_k | \beta]}{z_j - z_{j+1}} \end{aligned}$$

if $\Im z_j$ and $\Im z_{j+1}$ have opposite signs. Estimating (4.3) and (4.4) using the induction hypothesis yields (3.13).

Case 2 (all A_1, \dots, A_k traceless): As $\mathfrak{m}_\kappa[\cdot|\cdot]$ solves the recursion (3.7) without the source term \mathfrak{s}_{GUE} we can rewrite $\mathfrak{m}_\kappa[T_1, \dots, T_k | T_{k+1}, \dots, T_{k+\ell}]$ in terms of $\mathfrak{m}_1[\cdot]$ and values of $\mathfrak{m}_\kappa[\cdot|\cdot]$ for which the induction hypothesis applies. Note that $\langle A_k \rangle = 0$ implies that all terms with prefactor $q_{1,k}$ on the right-hand side of (3.7) vanish. The desired estimate thus readily follows from the induction hypothesis and Lemma 1.4. For the source term \mathfrak{s}_κ in (3.9), we obtain, e.g.,

$$\left| \langle M_{[r]} \odot M_{(s, \dots, k+\ell, k+1, \dots, t)} \rangle \langle M_{[r,k]} \odot M_{[t,s]} \rangle \right|$$

$$\begin{aligned} &\leq \frac{1}{N^2} \sum_{x,y \in [N]} |(M_{[r]})_{xx}(M_{(s,\dots,k+\ell,k+1,\dots,t)})_{xx}(M_{[r,k]})_{yy}(M_{[t,s]})_{yy}| \\ &\lesssim \left(\frac{1}{\eta_*}\right)^{k+\ell-\lceil(a+b)/2\rceil} \end{aligned}$$

for any $1 \leq r \leq k$ and $k+1 \leq s \leq t \leq k+\ell$. Recall that $\lceil \frac{x}{2} \rceil + \lceil \frac{y}{2} \rceil \geq \lceil \frac{x+y}{2} \rceil$ for any $x, y \in \mathbb{N}$, which yields the desired exponent $-(k+\ell-\lceil \frac{a+b}{2} \rceil)$ for η_* when bounds are multiplied. This concludes the proof of (3.13).

The bound (3.14) for $m_2[\cdot]$ is immediate from the decomposition (2.10) using the estimates (3.12) and (3.13). □

4.2 Proof of Lemma 2.3 (expansion for $\mathbb{E}\langle T_{[1,k]} \rangle$)

We use proof by induction to establish (2.4). As the base case $\mathbb{E}\langle T_\emptyset \rangle$ is trivial, assume that the expansion in Lemma 2.3 holds for resolvent chains of length up to $k-1$. We further assume w.l.o.g. that each A_j is either traceless or equal to the identity matrix.

First, consider resolvent chains that contain at least one deterministic matrix $A_j = \text{Id}$, i.e., that are of the form $T_{[1,k]} = T_{[1,j]}G_jG_{j+1}A_{j+1}T_{[j+1,k]}$ with the indices $j, j+1$ being interpreted mod k due to the cyclicity of the trace and the function $\mathcal{E}[\cdot]$. In this case, rewriting the product G_jG_{j+1} allows us to obtain the claim directly from the induction hypothesis. We distinguish two cases depending on the imaginary parts of z_j and z_{j+1} .

Case 1 ($\Im z_j$ and $\Im z_{j+1}$ have the same sign): Let again $s := \text{sign}(\Im z_j) = \text{sign}(\Im z_{j+1})$. By the residue theorem, we can write the product G_jG_{j+1} as a contour integral (cf. Equation (3.14) in the proof of [12, Lem. 3.2])

$$G_jG_{j+1} = \frac{1}{\pi} \int_{\mathbb{R}} \frac{\Im G(x+i\eta)}{(x+i s\eta-z_j)(x+i s\eta-z_{j+1})} dx, \quad G_j = G(z_j), \tag{4.5}$$

whenever $0 < \eta < \Im z_j, \Im z_{j+1}$ ($s = 1$) or $\Im z_j, \Im z_{j+1} < -\eta < 0$ ($s = -1$). Note that both $m_1[\cdot]$ and $\mathcal{E}[\cdot]$ have a similar representation, as

$$\begin{aligned} &m_1[T_1, \dots, T_{j-1}, G_j, T_{j+1}, \dots, T_k] \\ &= \frac{1}{2\pi i} \int_{\mathbb{R}} \frac{m_1[\dots, T_{j-1}, G(x+i\eta)A_{j+1}, \dots] - m_1[\dots, T_{j-1}, G(x-i\eta)A_{j+1}, \dots]}{(x+i s\eta-z_j)(x+i s\eta-z_{j+1})} dx \end{aligned} \tag{4.6}$$

by the residue theorem and [13, Lem. 5.4] as well as

$$\begin{aligned} &\mathcal{E}[T_1, \dots, T_{j-1}, G_j, T_{j+1}, \dots, T_k] \\ &= \frac{1}{2\pi i} \int_{\mathbb{R}} \frac{\mathcal{E}[\dots, T_{j-1}, G(x+i\eta)A_{j+1}, \dots] - \mathcal{E}[\dots, T_{j-1}, G(x-i\eta)A_{j+1}, \dots]}{(x+i s\eta-z_j)(x+i s\eta-z_{j+1})} dx \end{aligned} \tag{4.7}$$

using (3.3). Note that that (3.3) can be verified directly from the recursion (3.1). This distinction is crucial, as we use (2.4) later to apply the “meta argument” for the proof of Corollary 3.4 (see Appendix A). Rewriting the left-hand side of (2.4) using (4.5) and (4.6), we obtain an integral involving a resolvent chain of length $k-1$. Note that the parameter η has to be chosen large enough, e.g., as $\eta \geq \eta_*$. By the induction hypothesis, it now follows that

$$\begin{aligned} &\mathbb{E}\langle \langle T_{[1,k]} \rangle \rangle - m_1[T_1, \dots, T_k] \\ &= \frac{\kappa_4}{2\pi i N} \int_{\mathbb{R}} \frac{\mathcal{E}[\dots, T_{j-1}, G(x+i\eta)A_{j+1}, \dots] - \mathcal{E}[\dots, T_{j-1}, G(x-i\eta)A_{j+1}, \dots]}{(x+i s\eta-z_j)(x+i s\eta-z_{j+1})} dx \\ &\quad + \mathcal{O}\left(\frac{N^\varepsilon}{N\sqrt{N}\eta_*\eta_*^{k-a/2}}\right). \end{aligned}$$

Evaluating the integral using (4.7) gives (2.4) as desired.

Case 2 ($\Im z_j$ and $\Im z_{j+1}$ have opposite signs): Applying the resolvent identity (1.20) and the divided difference structure of $\mathfrak{m}_1[\cdot]$ (cf. [13, Lem. 5.4]) yields

$$\begin{aligned} & \mathbb{E}(\langle T_{[1,k]} \rangle - \mathfrak{m}_1[T_1, \dots, T_k]) \\ &= \mathbb{E} \left(\frac{\langle T_{[1,j]} G_j A_{j+1} T_{[j+1,k]} \rangle - \langle T_{[1,j]} T_{[j+1,k]} \rangle}{z_j - z_{j+1}} \right. \\ & \quad \left. - \frac{\mathfrak{m}_1[T_1, \dots, T_{j-1}, G_j A_{j+1}, T_{j+2}, \dots, T_k] - \mathfrak{m}_1[T_1, \dots, T_{j-1}, T_{j+1}, T_{j+2}, \dots, T_k]}{z_j - z_{j+1}} \right) \\ &= \frac{\kappa_4}{N} \frac{\mathcal{E}[T_1, \dots, T_{j-1}, G_j A_{j+1}, T_{j+2}, \dots, T_k] - \mathcal{E}[T_1, \dots, T_{j-1}, T_{j+1}, T_{j+2}, \dots, T_k]}{z_j - z_{j+1}} \\ & \quad + \mathcal{O} \left(\frac{N^\varepsilon}{N \sqrt{N} \eta_* \eta_*^{k-1-a/2} |z_j - z_{j+1}|} \right) \end{aligned}$$

by the induction hypothesis. As $\Im z_j$ and $\Im z_{j+1}$ are assumed to have opposite signs, it follows that $|z_j - z_{j+1}| \geq 2\eta_*$. The claim is now immediate from (3.3).

It remains to consider $T_{[1,k]}$ for which all matrices A_1, \dots, A_k are traceless. Here, we start the induction step by introducing

$$\underline{Wf(W)} := Wf(W) - \widetilde{\mathbb{E}} \widetilde{W} (\partial_{\widetilde{W}} f)(W) \tag{4.8}$$

with $\partial_{\widetilde{W}}$ denoting the directional derivative in direction \widetilde{W} and \widetilde{W} denoting an independent GUE matrix with expectation $\widetilde{\mathbb{E}}$. Comparing with the cumulant expansion formula (see, e.g., Equation (94) in [14] or the more general version in [22, Prop. 3.2]), we find that the renormalization in (4.8) cancels out the second-order term in the cumulant expansion of $\mathbb{E} Wf(W)$. In particular, $\mathbb{E} \underline{Wf(W)} = 0$ whenever W itself is a GUE matrix. Applying (4.8) for the resolvent $f(W) = (W - z)^{-1}$ yields the formulas

$$\begin{aligned} \underline{WG_1} &= WG_1 + \langle G_1 \rangle G_1, \\ \underline{WT_1 \dots T_k} &= \underline{WG_1} A_1 T_{[2,k]} + \sum_{j=2}^k \langle T_{[1,j]} G_j \rangle T_{[j,k]}, \end{aligned} \tag{4.9}$$

and we further recall the identities

$$\langle G_1 - m_1 \rangle = \frac{1}{1 - m_1^2} (-m_1 \langle \underline{WG_1} \rangle + m_1 \langle G_1 - m_1 \rangle^2) \tag{4.10}$$

$$\langle T_1 \rangle - \mathfrak{m}_1[T_1] = -m_1 \langle \underline{WT_1} \rangle + m_1^2 \langle G_1 - m_1 \rangle \langle A_1 \rangle + m_1 \langle G_1 - m_1 \rangle \langle T_1 - \mathfrak{m}_1[T_1] \rangle \tag{4.11}$$

from (96) in [14]. To complete the induction step, we need the analog of (4.11) for general $k \geq 1$. This allows rewriting $N(\langle T_{[1,k]} \rangle - \mathfrak{m}_1[T_1, \dots, T_k])$ in terms of shorter chains, and the claim follows by showing that the expectation matches the right-hand side of (3.1) up to an $\mathcal{O}(N^\varepsilon / (\sqrt{N} \eta_* \eta_*^{k-a/2}))$ error.

A brief calculation (see Equation (5.21) in [13]) yields

$$\begin{aligned} \langle T_{[1,k]} \rangle &= m_1 \left(- \langle \underline{WT_{[1,k]}} \rangle + \langle T_{[2,k]} A_1 \rangle + \sum_{j=2}^{k-1} \langle T_{[1,j]} G_j \rangle \langle T_{[j,k]} \rangle + \langle G_1 - m_1 \rangle \langle T_{[1,k]} \rangle \right. \\ & \quad \left. + \langle T_{[1,k]} G_k \rangle \langle (G_k - m_k) A_k \rangle \right). \end{aligned} \tag{4.12}$$

Next, we rewrite the equation to the form

$$\left(1 + \mathcal{O}_{\prec} \left(\frac{1}{N \eta_*} \right) \right) (\langle T_{[1,k]} \rangle - \mathfrak{m}_1[T_1, \dots, T_k])$$

$$\begin{aligned}
 &= m_1 \left(- \langle \underline{WT}_{[1,k]} \rangle + (\langle T_{[2,k]} G_k A_k A_1 \rangle - m_1[T_2, \dots, T_{k-1}, G_k A_k A_1]) \right. \\
 &\quad + \sum_{j=1}^{k-1} (\langle T_{[1,j]} G_j \rangle - m_1[T_1, \dots, T_{j-1}, G_j]) m_1[T_j, \dots, T_k] \\
 &\quad + \sum_{j=2}^k m_1[T_1, \dots, T_{j-1}, G_j] (\langle T_{[j,k]} \rangle - m_1[T_j, \dots, T_k]) \\
 &\quad \left. + \sum_{j=2}^k (\langle T_{[1,j]} G_j \rangle - m_1[T_1, \dots, T_{j-1}, G_j]) (\langle T_{[j,k]} \rangle - m_1[T_j, \dots, T_k]) \right), \tag{4.13}
 \end{aligned}$$

where we applied (1.15) for $\langle G_1 - m_1 \rangle$ on the right-hand side and used (1.18) to get the matching recursive form for $m_1[T_1, \dots, T_k]$. Note that $m_1[T_1, \dots, T_{j-1}, G_j] = m_1[G_1]$ for $j = 1$ resp. $m_1[T_j, \dots, T_k] = m_1[T_k]$ for $j = k$ on the right-hand side of (4.13) and recall that we set $\langle A_k \rangle = 0$ and $a = k$ in the case considered. Moving the factor $(1 + \mathcal{O}_{\prec}((N\eta_*)^{-1}))$ to the right-hand side, multiplying (4.13) with N and taking the expectation yields

$$\begin{aligned}
 N\mathbb{E}(\langle T_{[1,k]} \rangle - m_1[T_1, \dots, T_k]) &= \left(m_1 + \mathcal{O}_{\prec} \left(\frac{1}{N\eta_*} \right) \right) \left(- N\mathbb{E} \langle \underline{WT}_{[1,k]} \rangle + \kappa_4 \mathcal{E}[T_1, \dots, T_k] \right. \\
 &\quad \left. - \kappa_4 \sum_{1 \leq r \leq s \leq t \leq k} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle M_{[r,s]} \odot (M_{[t,k]} A_k) \rangle \right) \\
 &\quad + \mathcal{O} \left(\frac{N^\varepsilon}{\sqrt{N\eta_*} \eta_*^{k/2}} \right)
 \end{aligned}$$

by the induction hypothesis, (3.1) and the expansion $\frac{1}{1+x} = 1 + \mathcal{O}(x)$. We further applied (1.15) twice for the last line of (4.13) to obtain

$$(\langle T_{[1,j]} G_j \rangle - m_1[T_1, \dots, T_{j-1}, G_j]) (\langle T_{[j,k]} \rangle - m_1[T_j, \dots, T_k]) = \mathcal{O}_{\prec} \left(\frac{1}{N^2 \eta_*^{k/2+1}} \right).$$

Note that the index j appears in both terms and that the last factor of $T_{[1,j]} G_j$ being a resolvent implies that the term contains only $j - 1$ traceless matrices. It follows that

$$N\mathbb{E}(\langle T_{[1,j]} G_j \rangle - m_1[T_1, \dots, T_{j-1}, G_j]) (\langle T_{[j,k]} \rangle - m_1[T_j, \dots, T_k]) = \mathcal{O} \left(\frac{N^\varepsilon}{(N\eta_*) \eta_*^{k/2}} \right),$$

i.e., the term is indeed part of the error. Hence, (2.4) is established if

$$N\mathbb{E} \langle \underline{WT}_{[1,k]} \rangle = -\kappa_4 \sum_{1 \leq r \leq s \leq t \leq k} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle M_{[r,s]} \odot (M_{[t,k]} A_k) \rangle + \mathcal{O} \left(\frac{N^\varepsilon}{\sqrt{N\eta_*} \eta_*^{k/2}} \right). \tag{4.14}$$

By cumulant expansion (cf. Equation (94) in [14]), the underlined term on the left-hand side of (4.14) is given by

$$N\mathbb{E} \langle \underline{WT}_{[1,k]} \rangle = \sum_{n \geq 2} \sum_{x, y \in [N]} \sum_{\nu \in \{xy, yx\}^n} \frac{\kappa(xy, \nu)}{n!} \mathbb{E} \partial_\nu (T_{[1,k]})_{yx}, \tag{4.15}$$

where ∂_{xy} denotes the directional derivative in the direction of the xy entry of W and $\kappa(xy, \nu)$ denotes the joint cumulant of $W_{xy}, W_{\nu_1}, \dots, W_{\nu_n}$ for any n -tuple of double indices $\nu = (\nu_1, \dots, \nu_n)$. Note that the $n = 1$ term of the expansion (4.15) is canceled out by the renormalization (4.8).

Recall that $\kappa(xy, \nu) \sim N^{-(|\nu|+1)/2}$ by the scaling of W . It is hence sufficient to estimate the terms for $n \geq 4$ trivially using the bounds from Lemma 1.4, as the factor N^2 obtained from the double summation is canceled by the bound for the cumulant. Note that since

$$\partial_{xy} (T_{[1,k]})_{vw} = \sum_{r=1}^k (T_{[1,r]} G_r)_{vx} (T_{[r,k]})_{yw},$$

every derivative yields an additional resolvent factor, which increases the size of the bound for the corresponding resolvent chain by η_*^{-1} . However, each derivative also “breaks” the chain it acts on and increases the total number of resolvent chain entries in the term by one. This compensates for the additional factor η_*^{-1} (cf. Lemma 1.4) such that the power of η_*^{-1} contained in the bound for $\partial_\nu(T_{[1,k]})_{yx}$ does not depend on the order of the derivative. We conclude that

$$\sum_{n \geq 4} \sum_{x,y} \sum_{\nu \in \{xy, yx\}^n} \frac{\kappa(xy, \nu)}{n!} \mathbb{E} \partial_\nu(T_{[1,k]})_{yx} = \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \eta_*^{k/2}}\right)$$

and identify the term as part of the error in (2.4). It remains to consider the $n = 2$ and $n = 3$ contribution to (4.15). Recall that we write $T_{[i,j]} = T_i \dots T_j$ for $i \leq j$ and $T_{[i,i]} = T_\emptyset = 0$.

Estimate for the $n = 2$ term of (4.15): Evaluating the derivative yields, e.g.,

$$\begin{aligned} (\partial_{xy})^2(T_{[1,k]})_{yx} &= -\partial_{xy} \sum_{r=1}^k (T_{[1,r]} G_r)_{yx} (T_{[r,k]})_{yx} \\ &= \sum_{r=1}^k \left(\sum_{s=1}^r (T_{[1,s]} G_s)_{yx} (T_{[s,r]} G_r)_{yx} (T_{[r,k]})_{yx} \right. \\ &\quad \left. + \sum_{s=r}^k (T_{[1,r]} G_r)_{yx} (T_{[r,s]} G_s)_{yx} (T_{[s,k]})_{yx} \right), \end{aligned}$$

which only involves xy or yx entries of a resolvent chain. Note that the contribution for the case $x = y$ consists of just one sum and is, therefore, of lower order by power counting. We thus refer to any xy or yx entries as off-diagonal below. Further, we obtain

$$\begin{aligned} \partial_{xy} \partial_{yx} (T_{[1,k]})_{yx} &= \sum_{r=1}^k \left(\sum_{s=1}^r (T_{[1,s]} G_s)_{yx} (T_{[s,r]} G_r)_{yy} (T_{[r,k]})_{xx} \right. \\ &\quad \left. + \sum_{s=r}^k (T_{[1,r]} G_r)_{yx} (T_{[r,s]} G_s)_{yy} (T_{[s,k]})_{xx} \right), \end{aligned}$$

which involves two diagonal and one off-diagonal entries. The other terms arising for $n = 2$ are of a similar structure, i.e., they involve either zero or two diagonal entries.

As every term involves at least one off-diagonal entry of a resolvent chain, we use a procedure called *isotropic resummation* to estimate the x and y summations. To illustrate the strategy, consider the sum

$$N^{-3/2} \sum_{x,y \in [N]} (T_{[1,r]} G_r)_{yx} (T_{[r,s]} G_s)_{yy} (T_{[s,k]})_{xx} \tag{4.16}$$

with fixed $1 \leq r \leq s \leq k$. The factor $N^{-3/2}$ in front of the term accounts for the size of the cumulant $\kappa(xy, xy, yx)$. First, insert the deterministic approximation of the diagonal terms to decompose the resolvent chain entries in (4.16) into a deterministic term satisfying the bounds in Lemma 1.4 and a fluctuation that is controlled by the local law (1.16). Recalling that

$$\|M_{[i,j]} A_j\| \lesssim \|M_{[i,j]}\| \lesssim \frac{1}{\eta_*^{(j-i)/2}} \tag{4.17}$$

for $i \leq j$ by Lemma 1.4, as $M_{[i,j]}$ involves $j - i$ traceless matrices and $\|A_j\| \lesssim 1$ by assumption, and that

$$\max_{x,y \in [N]} |((T_{[i,j]} G_j - M_{[i,j]}) A_j)_{xy}| = \mathcal{O}_\prec\left(\frac{1}{\sqrt{N} \eta_* \eta_*^{(j-i)/2}}\right) \tag{4.18}$$

by the isotropic local law (1.16), we obtain the bound

$$|(T_{[1,r]}G_r)_{yx}| \leq \|M_{[r]}\| + \max_{x,y \in [N]} |(T_{[1,r]}G_r - M_{[r]})_{yx}| = \mathcal{O}_{\prec} \left(\frac{1}{\eta_*^{(r-1)/2}} + \frac{1}{\sqrt{N}\eta_* \eta_*^{(r-1)/2}} \right).$$

Recall that we abbreviated $M_{[r]} = M_{[1,r]}$. Putting everything together, it follows that

$$N^{-3/2} \sum_{x,y \in [N]} (T_{[1,r]}G_r)_{yx} (T_{[r,s]}G_s - M_{[r,s]})_{yy} (T_{[s,k]} - M_{[s,k]}A_k)_{xx} = \mathcal{O}_{\prec} \left(\frac{1}{\sqrt{N}\eta_* \eta_*^{k/2}} \right)$$

and we can include this term in the error in (2.4). Recall that we consider $N(\mathbb{E}\langle T_{[1,k]} - \mathbf{m}_1[T_1, \dots, T_k] \rangle)$, i.e., the error terms obtained differ from (2.4) by a factor of N .

Next, let $\text{dvec}(A) = (A_{jj})_{j=1}^N$ denote the (column) vector consisting of the diagonal entries of a matrix $A \in \mathbb{C}^{N \times N}$. In this notation, we have

$$\begin{aligned} & N^{-3/2} \sum_{x,y \in [N]} (T_{[1,r]}G_r)_{yx} (M_{[r,s]})_{yy} (M_{[s,k]}A_k)_{xx} \\ &= N^{-3/2} \langle \text{dvec}(M_{[r,s]}), T_{[1,r]}G_r \text{dvec}(M_{[s,k]}A_k) \rangle \end{aligned}$$

with deterministic vectors $\text{dvec}(M_{[r,s]})$ and $\text{dvec}(M_{[s,k]}A_k)$ satisfying

$$\begin{aligned} \|\text{dvec}(M_{[r,s]})\| &\leq \sqrt{N} \|M_{[r,s]}\| = \mathcal{O} \left(\frac{\sqrt{N}}{\eta_*^{(s-r)/2}} \right), \\ \|\text{dvec}(M_{[s,k]}A_k)\| &\leq \sqrt{N} \|M_{[s,k]}A_k\| = \mathcal{O} \left(\frac{\sqrt{N}}{\eta_*^{(k-s)/2}} \right), \end{aligned}$$

by (4.17). Recalling that $T_{[1,r]}G_r$ contains $r - 1$ traceless matrices by assumptions, we obtain

$$N^{-3/2} \sum_{x,y \in [N]} (T_{[1,r]}G_r)_{yx} (M_{[r,s]})_{yy} (M_{[s,k]}A_k)_{xx} = \mathcal{O}_{\prec} \left(\frac{1}{\sqrt{N}\eta_* \eta_*^{k/2-1}} \right)$$

from (4.17) and (1.16). Hence, the term can be included in the error in (2.4).

It remains to estimate the two terms in (4.16) that involve one deterministic and one fluctuation term each. By the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} & N^{-3/2} \sum_{x,y \in [N]} (T_{[1,r]}G_r)_{yx} (T_{[r,s]}G_s - M_{[r,s]})_{yy} (M_{[s,k]}A_k)_{xx} \\ &\leq N^{-3/2} \|T_{[1,r]}G_r\| \cdot \|\text{dvec}(T_{[r,s]}G_s - M_{[r,s]})\| \cdot \|\text{dvec}(M_{[s,k]}A_k)\| \end{aligned}$$

with a similar bound holding for the term involving $M_{[r,s]}$ and $(T_{[s,k]} - M_{[s,k]}A_k)$. Note that

$$\begin{aligned} \|\text{dvec}(T_{[r,s]}G_s - M_{[r,s]})\| &\leq \sqrt{N} \max_{x \in [N]} |(T_{[r,s]}G_s - M_{[r,s]})_{xx}| = \mathcal{O}_{\prec} \left(\frac{1}{\eta_*^{(s-r+1)/2}} \right), \\ \|T_{[1,r]}G_r\| &\leq \|M_{[r]}\| + N \max_{x,y \in [N]} |(T_{[1,r]}G_r - M_{[r]})_{xy}| = \mathcal{O}_{\prec} \left(\frac{1}{\eta_*^{(r-1)/2}} + \frac{\sqrt{N}}{\eta_*^{r/2}} \right) \end{aligned}$$

by (4.17), (4.18), and the fact that $\|B\| \leq N \max_{x,y} |B_{xy}|$ for any matrix $B \in \mathbb{C}^{N \times N}$. Overall, we obtain

$$N^{-3/2} \sum_{x,y \in [N]} (T_{[1,s]}G_s)_{yx} (T_{[s,r]}G_r)_{yy} (T_{[r,k]})_{xx} = \mathcal{O}_{\prec} \left(\frac{1}{\sqrt{N}\eta_* \eta_*^{k/2}} \right),$$

i.e., the term is part of the error in (2.4). The other terms arising from $\partial_\nu(T_1 \dots T_k)_{yx}$ with $\nu \in \{xy, yx\}^2$ are treated similarly. Hence, the entire $n = 2$ contribution of (4.15) can be included in the error term in (2.4).

Computation of the $n = 3$ term of (4.15): By a similar computation as in the $n = 2$ case, the terms arising from $\partial_\nu(T_{[1,k]})_{yx}$ for $|\nu| = 3$ involve either zero, two, or four diagonal entries of a resolvent chain. Whenever a term contains off-diagonal entries, we can include it in the error in (2.4) by decomposing each resolvent chain entry into a deterministic and a fluctuation part. Note that $M_{(\cdot)}$ is not necessarily diagonal, i.e., applying (1.16) alone is not sufficient. However, as every fluctuation term contributes a factor $N^{-1/2}$ and the presence of off-diagonal terms allows us to apply the Cauchy-Schwarz inequality to estimate the remaining (deterministic) double sum, we gain a factor $N^{-1/2}$ over the trivial bound as needed.

It remains to evaluate the contributions that consist of four diagonal entries, which can only occur for $\nu \in \{(xy, yx, yx), (yx, xy, yx), (yx, yx, xy)\}$. Here,

$$\begin{aligned} \partial_\nu(T_{[1,k]})_{yx} = & - \sum_{1 \leq t \leq s \leq r \leq k} (T_{[1,t]}G_t)_{yy}(T_{[t,s]}G_s)_{xx}(T_{[s,r]}G_r)_{yy}(T_{[r,k]})_{xx} \\ & - \sum_{1 \leq r \leq s \leq t \leq k} (T_{[1,r]}G_r)_{yy}(T_{[r,s]}G_s)_{xx}(T_{[s,t]}G_t)_{yy}(T_{[t,k]})_{xx} - \dots \end{aligned} \quad (4.19)$$

where the terms that are not written out in the last line involve two off-diagonal entries and, therefore, can be included in the error term. Since $\kappa(xy, \nu) = \kappa_4/N^2$ for the cases considered, we obtain

$$\begin{aligned} & \sum_{x,y} \sum_{\nu \in \{xy, yx\}^3} \frac{\kappa(xy, \nu)}{6} \mathbb{E} \partial_\nu(T_{[1,k]})_{yx} \\ &= \frac{\kappa_4}{2N^2} \sum_{x,y} \left(- \sum_{1 \leq t \leq s \leq r \leq k} (M_{[t]})_{yy}(M_{[t,s]})_{xx}(M_{[s,r]})_{yy}(M_{[r,k]}A_k)_{xx} \right. \\ & \quad \left. - \sum_{1 \leq r \leq s \leq t \leq k} (M_{[r]})_{yy}(M_{[r,s]})_{xx}(M_{[s,t]})_{yy}(M_{[t,k]}A_k)_{xx} \right) + \mathcal{O}\left(\frac{N^\varepsilon}{N \sqrt{N} \eta_* \eta_*^{k-a/2}}\right) \\ &= -\kappa_4 \sum_{1 \leq r \leq s \leq t \leq k} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle M_{[r,s]} \odot (M_{[t,k]}A_k) \rangle + \mathcal{O}\left(\frac{N^\varepsilon}{N \sqrt{N} \eta_* \eta_*^{k-a/2}}\right) \end{aligned}$$

where the last equality follows from the definition of the Hadamard product and the fact that the two sums involving r, s, t are the same up to relabeling of the indices. Note that the term in (4.19) is counted for each $\nu \in \{(xy, yx, yx), (yx, xy, yx), (yx, yx, xy)\}$ when summing over $\nu \in \{xy, yx\}^3$ while the other choices for ν yield lower-order terms.

Adding the contributions to (4.15) together, the cumulant expansion evaluates to

$$\begin{aligned} -N \mathbb{E} \langle \underline{WT}_{[1,k]} \rangle = & \kappa_4 \sum_{1 \leq r \leq s \leq t \leq k} \left(\frac{1}{N^2} \sum_{x,y} (M_{[r]})_{yy}(M_{[r,s]})_{xx}(M_{[s,t]})_{yy}(M_{[t,k]}A_k)_{xx} \right) \\ & + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \eta_*^{k-a/2}}\right). \end{aligned}$$

We conclude that $N(\mathbb{E} \langle T_{[1,k]} \rangle - \mathbf{m}_1[T_1, \dots, T_k])$ coincides with the right-hand side of (3.1) up to an error term and an additional factor κ_4 . Applying the recursion thus yields

$$N(\mathbb{E} \langle T_{[1,k]} \rangle - \mathbf{m}_1[T_1, \dots, T_k]) = \kappa_4 \mathcal{E}[T_1, \dots, T_k] + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \eta_*^{k-a/2}}\right)$$

as claimed. □

4.3 Proof of Lemma 2.5 (leading term of $\mathbb{E}X_\alpha X_\beta$)

We start by noting some general estimates for $X_\alpha^{(k,a)}$ and its derivatives that are needed for the proof of Lemma 2.5.

Lemma 4.3 (A priori estimates). *For $X_\alpha^{(k,a)} = \langle T_{[1,k]} \rangle - \mathbb{E}\langle T_{[1,k]} \rangle$ and its derivatives $\partial_\nu X_\alpha^{(k,a)}$ for any multi-index ν , we have the estimate*

$$|\partial_\nu X_\alpha^{(k,a)}| = \mathcal{O}_\prec\left(\frac{1}{N\eta_*^{k-a/2}}\right). \tag{4.20}$$

Moreover, we have the more precise expansions for the first and second derivatives

$$\partial_{xy} X_\alpha^{(k,a)} = -\frac{1}{N} \left[\sum_{r=1}^k (M_{(r,\dots,k,1,\dots,r)})_{yx} + \mathcal{O}_\prec\left(\frac{1}{\sqrt{N}\eta_* \eta_*^{k-a/2}}\right) \right], \tag{4.21}$$

$$\begin{aligned} \partial_{vw} \partial_{xy} X_\alpha^{(k,a)} &= \frac{1}{N} \left[\sum_{r=1}^k \left(\sum_{s=1}^r (M_{(r,\dots,k,1,\dots,s)})_{wx} (M_{[s,r]})_{yv} \right. \right. \\ &\quad \left. \left. + \sum_{s=r}^k (M_{[r,s]})_{wx} (M_{(s,\dots,k,1,\dots,r)})_{yv} \right) + \mathcal{O}_\prec\left(\frac{1}{\sqrt{N}\eta_* \eta_*^{k-a/2}}\right) \right]. \end{aligned} \tag{4.22}$$

Proof. The bounds in (4.21) and (4.22) follow directly from (1.16), Lemma 1.4, and

$$\begin{aligned} \partial_{xy} \langle T_{[1,k]} \rangle &= -\frac{1}{N} \sum_{r=1}^k (T_{[r,k]} T_{[1,r]} G_r)_{yx}, \\ \partial_{vw} \partial_{xy} \langle T_{[1,k]} \rangle &= \frac{1}{N} \sum_{r=1}^k \left(\sum_{s=1}^r (T_{[r,k]} T_{[1,s]} G_s)_{wx} (T_{[s,r]} G_r)_{yv} \right. \\ &\quad \left. + \sum_{s=r}^k (T_{[r,s]} G_s)_{wx} (T_{[s,k]} T_{[1,r]} G_r)_{yv} \right). \end{aligned}$$

The claim (4.20) follows inductively by (1.16) and Lemma 1.4. Note that (4.20) is also true for $\nu = \emptyset$ since

$$X_\alpha^{(k)} = (\langle T_{[1,k]} \rangle - \mathfrak{m}_1[T_1, \dots, T_k]) - (\mathbb{E}\langle T_{[1,k]} \rangle - \mathfrak{m}_1[T_1, \dots, T_k]) = \mathcal{O}_\prec\left(\frac{1}{N\eta_*^{k-a/2}}\right)$$

as a consequence of (1.15) and Lemma 2.3. □

Proof of Lemma 2.5. We use proof by induction on the length of the multi-indices α and β . First, note that the left-hand side of (2.7) vanishes if either $\langle T_{[1,k]} \rangle$ or $\langle T_{[k+1,k+\ell]} \rangle$ is zero, i.e., if one of the terms is indexed by the empty set. Comparing with (3.6), the base case is established. Due to the symmetry of the expression, it is further sufficient to carry out the induction step for one of the arguments only. Assume that (2.7) holds for multi indices α of length $1, \dots, k-1$ and multi-indices β of a fixed length $\ell \geq 1$. We need to show that the leading term of $N^2 \mathbb{E}X_\alpha^{(k)} X_\beta^{(\ell)}$ is given by $\mathfrak{m}_2[T_1, \dots, T_k | T_{k+1} \dots T_{k+\ell}]$ and estimate the error term. W.l.o.g. assume that each A_j is either traceless or equal to the identity matrix.

As a first step, consider resolvent chains $T_{[1,k]}$ that contain at least one deterministic matrix $A_j = \text{Id}$, i.e., that are of the form $T_{[1,k]} = T_{[1,j]} G_j G_{j+1} T_{[j+1,k]}$ with the indices being interpreted mod k due to the cyclicity of the trace and the first entry of $\mathfrak{m}_2[\cdot, \cdot]$. In this case, rewriting the product $G_j G_{j+1}$ in terms of a single resolvent allows us to obtain the claim directly from the induction hypothesis. We distinguish two cases depending on the imaginary parts of z_j and z_{j+1} .

Case 1: $\Im z_j$ and $\Im z_{j+1}$ have the same sign: Let $s := \text{sign}(\Im z_j) = \text{sign}(\Im z_{j+1})$ and recall from (4.5) that the product $G_j G_{j+1}$ can be rewritten as a contour integral using the residue theorem. We obtain a similar contour integral formula for $\mathfrak{m}_2[\cdot|\cdot]$ using the divided difference structure in (3.19), namely

$$\begin{aligned} & \mathfrak{m}_2[T_1, \dots, T_{j-1}, G_j, T_{j+1}, \dots, T_k|\beta] \tag{4.23} \\ &= \frac{1}{2\pi i} \int_{\mathbb{R}} \frac{\mathfrak{m}_2[\dots, T_{j-1}, G(x+i\eta)A_{j+1}, \dots|\beta] - \mathfrak{m}_2[\dots, T_{j-1}, G(x-i\eta)A_{j+1}, \dots|\beta]}{(x+i s\eta - z_j)(x+i s\eta - z_{j+1})} dx. \end{aligned}$$

Note that (3.17) can be verified directly from the recursion (3.7), which is crucial if we want to use Lemma 2.5 in the “meta argument” to prove Corollary 3.10. Rewriting the left-hand side of (2.7) using (4.6) yields a contour integral with an integrand of the form $N^2 \mathbb{E} X_{\alpha'}^{(k)} X_{\beta}^{(\ell)}$ for a multi-index α' of length $k-1$ and without the identity matrix $A_j = \text{Id}$, i.e., the number of traceless matrices a is unchanged. By the induction hypothesis, we thus obtain

$$\begin{aligned} & N^2 \mathbb{E} X_{\alpha}^{(k)} X_{\beta}^{(\ell)} \\ &= \frac{1}{2\pi i} \int_{\mathbb{R}} \frac{\mathfrak{m}_2[\dots, T_{j-1}, G(x+i\eta)A_{j+1}, \dots|\beta] - \mathfrak{m}_2[\dots, T_{j-1}, G(x-i\eta)A_{j+1}, \dots|\beta]}{(x+i s\eta - z_j)(x+i s\eta - z_{j+1})} dx \\ &+ \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N\eta_*} \eta_*^{(k+\ell-1)-(a+b)/2}}\right) \end{aligned}$$

where the integral evaluates to $\mathfrak{m}_2[\alpha|\beta]$ by (4.23). Hence, (2.7) holds in the case considered.

Case 2: $\Im z_j$ and $\Im z_{j+1}$ have opposite signs: Applying (1.20) and the divided difference structure of $\mathfrak{m}_2[\cdot|\cdot]$, it follows that

$$\begin{aligned} & N^2 \mathbb{E} X_{\alpha}^{(k,a)} X_{\beta}^{(\ell,b)} \\ &= N^2 \mathbb{E} \left(\frac{\langle \langle T_{[1,j]} G_j A_{j+1} T_{[j+1,k]} \rangle \rangle - \mathbb{E} \langle T_{[1,j]} G_j A_{j+1} T_{[j+1,k]} \rangle}{z_j - z_{j+1}} \right. \\ &\quad \left. - \frac{\langle \langle T_{[1,j]} T_{[j+1,k]} \rangle \rangle - \mathbb{E} \langle T_{[1,j]} T_{[j+1,k]} \rangle}{z_j - z_{j+1}} \right) X_{\beta}^{(\ell,b)} \\ &= \frac{\mathfrak{m}_2[T_1, \dots, T_{j-1}, T_{j+1}, \dots, T_k|\beta] - \mathfrak{m}_2[T_1, \dots, T_{j-1}, T_j A_{j+1}, T_{j+2}, \dots, T_k|\beta]}{z_j - z_{j+1}} \\ &+ \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N\eta_*} \eta_*^{k+\ell-(a+b)/2-1} |z_j - z_{j+1}|}\right) \\ &= \mathfrak{m}_2[T_1, \dots, T_k|\beta] + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N\eta_*} \eta_*^{k+\ell-(a+b)/2}}\right), \end{aligned}$$

since $|z_j - z_{j+1}| \geq 2\eta_*$ in the case considered. This concludes the proof of (2.7) for resolvent chains that contain at least one deterministic matrix $A_j = \text{Id}$.

It remains to consider $T_{[1,k]}$ for which all matrices A_1, \dots, A_k are traceless, i.e., for which $a = k$. The argument is similar to the proof of Lemma 2.3, i.e., we rewrite $N^2 \mathbb{E} X_{\alpha}^{(k,a)} X_{\beta}^{(\ell,b)}$ in terms of covariances of smaller chains and show that it satisfies the recursion (3.7) up to an $N^\varepsilon / (\sqrt{N\eta_*} \eta_*^{k-a/2} \eta_*^{\ell-b/2})$ error.

Combining (4.13) and Lemma 2.3 yields the starting point

$$\begin{aligned} X_{\alpha}^{(k,a)} &= m_1 \left(- \langle \langle \underline{W T_{[1,k]}} \rangle \rangle - \mathbb{E} \langle \underline{W T_{[1,k]}} \rangle \rangle + \langle \langle T_{[2,k]} G_k A_k A_1 \rangle \rangle - \mathbb{E} \langle T_{[2,k]} G_k A_k A_1 \rangle \rangle \right) \\ &+ \sum_{j=1}^{k-1} \langle \langle T_{[1,j]} G_j \rangle \rangle - \mathbb{E} \langle T_{[1,j]} G_j \rangle \rangle \mathfrak{m}_1[T_j, \dots, T_k] \tag{4.24} \end{aligned}$$

$$+ \sum_{j=2}^k \mathbf{m}_1[T_1, \dots, T_{j-1}, G_j] (\langle T_{[j,k]} \rangle - \mathbb{E} \langle T_{[j,k]} \rangle) + \mathcal{O}_{\prec} \left(\frac{1}{N \sqrt{N \eta_*} \eta_*^{k/2}} \right).$$

Next, multiply (4.24) by $N^2 X_{\beta}^{(\ell)}$ and compute the expectation. Applying the induction hypothesis for any terms for which the first factor involves a resolvent chain of length at most $k - 1$ yields

$$\begin{aligned} N^2 \mathbb{E} \left(X_{\alpha}^{(k)} X_{\beta}^{(\ell)} \right) &= \mathbf{m}_2[\alpha|\beta] - \sum_{j=1}^{\ell} \left(\mathbf{m}_1[T_1, \dots, T_k, T_{k+j}, \dots, T_{k+j-1}, G_{k+j}] \right. \\ &\quad - \kappa_4 \sum_{r=1}^k \sum_{s=k+1}^{k+\ell} \left(\sum_{t=k+1}^s \langle M_{[r]} \odot M_{(s, \dots, k+\ell, k+1, \dots, t)} \rangle \langle (M_{[r,k]} A_k) \odot M_{[t,s]} \rangle \right. \\ &\quad \left. \left. + \sum_{t=s}^{k+\ell} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle (M_{[r,k]} A_k) \odot M_{(t, \dots, k+\ell, k+1, \dots, s)} \rangle \right) \right) \\ &\quad + N^2 \mathbb{E} \left((\langle \underline{WT}_{[1,k]} \rangle - \mathbb{E} \langle \underline{WT}_{[1,k]} \rangle) X_{\beta}^{(\ell)} \right) + \mathcal{O} \left(\frac{N^{\varepsilon}}{\sqrt{N \eta_*} \eta_*^{(k+\ell)/2}} \right), \end{aligned}$$

where we used the recursion (3.7) to introduce $\mathbf{m}_2[\alpha|\beta]$. It remains to compute

$$N^2 \mathbb{E} \left((\langle \underline{WT}_{[1,k]} \rangle - \mathbb{E} \langle \underline{WT}_{[1,k]} \rangle) X_{\beta}^{(\ell)} \right) = N^2 \mathbb{E} \left(\langle \underline{WT}_{[1,k]} \rangle X_{\beta}^{(\ell)} \right).$$

By cumulant expansion, we obtain

$$\begin{aligned} N^2 \mathbb{E} \left(\langle \underline{WT}_{[1,k]} \rangle X_{\beta}^{(\ell)} \right) &= - \sum_{j=1}^{\ell} \mathbb{E} \langle T_{[1,k]} T_{[k+j, k+\ell]} T_{[1, k+j]} G_{k+j} \rangle \\ &\quad + N \sum_{n=2}^3 \sum_{x, y \in [N]} \sum_{\nu \in \{xy, yx\}^n} \frac{\kappa(xy, \nu)}{n!} \mathbb{E} \partial_{\nu} \left((T_{[1,k]})_{yx} X_{\beta}^{(\ell)} \right) \quad (4.25) \\ &\quad + \mathcal{O} \left(\frac{N^{\varepsilon}}{\sqrt{N \eta_*} \eta_*^{(k+\ell)/2}} \right), \end{aligned}$$

where the $n = 1$ term was evaluated using (4.8) and the terms for $n \geq 4$ were again estimated trivially using that $\kappa(xy, \nu) \sim N^{-(|\nu|+1)/2}$ due to the scaling of W as well as the bounds from Lemmas 1.4 and 4.3. By the isotropic local law (1.16) we have

$$\begin{aligned} &\mathbb{E} \langle T_{[1,k]} T_{[k+j, k+\ell]} T_{[1, k+j]} G_{k+j} \rangle \\ &= \mathbf{m}_1[T_1, \dots, T_k, T_{k+j}, \dots, T_{k+j-1}, G_{k+j}] + \mathcal{O} \left(\frac{N^{\varepsilon}}{(N \eta_*) \eta_*^{(k+\ell)/2}} \right), \end{aligned}$$

which yields an error term that can be included in the error in (2.7). Hence, it only remains to consider the contributions for $n = 2$ and $n = 3$.

Estimate for the $n = 2$ term of (4.25): Evaluating the derivative $\partial_{\nu} \left((T_{[1,k]})_{yx} X_{\beta}^{(\ell)} \right)$ always yields at least one xy or yx entry of a resolvent chain, which we can use to apply isotropic resummation (cf. estimate for (4.16) above). Note that $X_{\beta}^{(\ell)}$ always contributes a factor N^{-1} due to Lemma 4.3, either from evaluating derivatives using (4.21) and (4.22), or from the trivial estimate (4.20). The additional factor N in front of the x, y summation in (4.25) is thus balanced out. Overall, we obtain

$$N \sum_{x, y} \sum_{\nu \in \{xy, yx\}^2} \frac{\kappa(xy, \nu)}{2} \mathbb{E} \partial_{\nu} \left((T_{[1,k]})_{yx} X_{\beta}^{(\ell)} \right) = \mathcal{O} \left(\frac{N^{\varepsilon}}{\sqrt{N \eta_*} \eta_*^{k-a/2} \eta_*^{\ell-b/2}} \right).$$

In particular, the $n = 2$ term can be included in the error in the last line of (4.25).

Computation of the $n = 3$ term of (4.25): By applying the Leibniz rule, we distribute the derivatives as

$$\mathbb{E}\partial_\nu\left((T_{[1,k]})_{yx}X_\beta^{(\ell)}\right) = \sum_{\nu_1\cup\nu_2=\nu} \mathbb{E}\left(\partial_{\nu_1}(T_{[1,k]})_{yx}\partial_{\nu_2}X_\beta^{(\ell)}\right)$$

and distinguish three cases for ν_1 :

Case 1 ($|\nu_1| = 3$): The derivative $\partial_{\nu_1}(T_{[1,k]})_{yx}$ arising in this case was already computed in the proof of Lemma 2.3. Using (4.19) and (1.16), it follows that

$$\begin{aligned} & N \sum_{x,y} \sum_{\nu \in \{xy, yx\}^3} \frac{\kappa(xy, \nu)}{6} \mathbb{E}\left((\partial_\nu(T_{[1,k]})_{yx})X_\beta^{(\ell)}\right) \\ &= -\frac{\kappa_4}{N} \sum_{x,y} \sum_{1 \leq r \leq s \leq t \leq k} (M_{[r]})_{yy}(M_{[r,s]})_{xx}(M_{[s,t]})_{yy}(M_{[t,k]}A_k)_{xx} \mathbb{E}X_\beta^{(\ell)} \\ &\quad + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N}\eta_* \eta_*^{k-a/2} \eta_*^{\ell-b/2}}\right) \\ &= \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N}\eta_* \eta_*^{k-a/2} \eta_*^{\ell-b/2}}\right), \end{aligned}$$

since $X_\beta^{(\ell)}$ is centered. We hence include the term in the error in (4.25).

Case 2 ($|\nu_1| = 1$): Evaluating the ∂_ν derivative yields either zero, two, or four diagonal entries of a resolvent chain. Whenever the term contains an xy or yx entry, we use isotropic resummation to identify the term as part of the error in (4.25). It remains to consider the case $\nu_1 = (yx)$ and $\nu_2 \in \{(xy, yx), (yx, xy)\}$. Here, we compute

$$\begin{aligned} \partial_{\nu_1}(T_{[1,k]})_{yx} &= -\sum_{r=1}^k (T_{[1,r]}G_r)_{yy}(T_{[r,k]})_{xx}, \\ \partial_{\nu_2}X_\beta^{(\ell)} &= \frac{1}{N} \sum_{s=k+1}^{k+\ell} \left(\sum_{t=k+1}^s (T_{[s,k+\ell]}T_{[k+1,t]}G_t)_{yy}(T_{[t,s]}G_s)_{xx} \right. \\ &\quad \left. + \sum_{t=s}^{k+\ell} (T_{[s,t]}G_t)_{yy}(T_{[t,k+\ell]}T_{[k+1,s]}G_s)_{xx} \right) + \dots, \end{aligned}$$

where the terms that are left out contain at least one off-diagonal entry of a resolvent chain and hence can be included in the error term in (4.25) by isotropic resummation. Applying (1.16) and recalling that $\kappa(xy, \nu) = \kappa_4/N^2$ in the case considered, it follows that

$$\begin{aligned} & N \sum_{x,y} \frac{\kappa_4}{N^2} \mathbb{E}\left((\partial_{\nu_1}(T_{[1,k]})_{yx})(\partial_{\nu_2}X_\beta^{(\ell)})\right) \\ &= -\frac{\kappa_4}{N^2} \sum_{x,y} \left(\sum_{r=1}^k \sum_{s=k+1}^{k+\ell} (M_{[r]})_{yy}(M_{[r,k]}A_k)_{xx} \left(\sum_{t=k+1}^s (M_{(s,\dots,k+\ell,k+1,\dots,t)})_{yy}(M_{[t,s]})_{xx} \right. \right. \\ &\quad \left. \left. + \sum_{t=s}^{k+\ell} (M_{[s,t]})_{yy}(M_{(t,\dots,k+\ell,k+1,\dots,s)})_{xx} \right) \right) + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N}\eta_* \eta_*^{k-a/2} \eta_*^{\ell-b/2}}\right) \end{aligned} \tag{4.26}$$

which can again be rewritten using the definition of the Hadamard product. Note that we obtain six terms of the form (4.26) in total from applying the Leibniz rule.

Case 3 ($|\nu_1|$ even): For both $|\nu_1| = 0$ and $|\nu_1| = 2$, the derivative always contains at least one off-diagonal entry of a resolvent chain. Therefore, we can apply isotropic resummation and include the term in the error in (4.25).

Adding the contributions to (4.25) back together, the cumulant expansion evaluates to

$$\begin{aligned}
 & -N^2 \mathbb{E}(\langle \underline{WT}_{[1,k]} \rangle X_\beta^{(\ell)}) \\
 &= \sum_{j=1}^{\ell} \mathbf{m}[T_1, \dots, T_k, T_{k+j}, \dots, T_{k+\ell}, T_{k+1}, \dots, T_{k+j-1}, G_{k+j}] \\
 & \quad + \kappa_4 \sum_{r=1}^k \sum_{s=k+1}^{k+\ell} \left(\sum_{t=k+1}^s \langle M_{[r]} \odot M_{(s, \dots, k+\ell, k+1, \dots, t)} \rangle \langle (M_{[r,k]} A_k) \odot M_{[t,s]} \rangle \right. \\
 & \quad \left. + \sum_{t=s}^{k+\ell} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle (M_{[r,k]} A_k) \odot M_{(t, \dots, k+\ell, k+1, \dots, s)} \rangle \right) + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \eta_*^{k-a/2} \eta_*^{\ell-b/2}}\right).
 \end{aligned}$$

We conclude that, up to an $\mathcal{O}(N^\varepsilon/\sqrt{N})$ error, $\mathbb{E}(X_\alpha^{(k)} X_\beta^{(\ell)})$ equals the right-hand side of (3.7). Applying the recursion yields

$$\mathbb{E}\left(X_\alpha^{(k)} X_\beta^{(\ell)}\right) = \mathbf{m}_2[\alpha|\beta] + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \eta_*^{k-a/2} \eta_*^{\ell-b/2}}\right)$$

which completes the induction step. □

It remains to establish the alternative integral representation for $\tilde{m}_{GUE}[\cdot]$.

Proof of Corollary 3.11. We use proof by induction, starting with the base case $k = \ell = 1$. While we could establish (3.20) directly using a similar computation as in [14] or [41], it is much quicker to read off the kernel from the existing literature. To do so, note that the function $\tilde{m}_{GUE}[1|2]$ and its counterpart for W being a GOE matrix only differ by a constant factor. More precisely, evaluating (92) of [14] for GOE ($\sigma = 1, \kappa_4 = 0, \tilde{\omega}_2 = 0$) yields

$$\lim_{N \rightarrow \infty} N^2 \text{Cov}(\langle G_1 - \mathbb{E}G_1 \rangle, \overline{\langle G_2 - \mathbb{E}G_2 \rangle}) = 2 \cdot \frac{m'_1 m'_2}{(1 - m_1 m_2)^2} = 2 \cdot \tilde{m}_{GUE}[1|2]. \quad (4.27)$$

This allows us to obtain the desired integral representation for $\tilde{m}_{GUE}[1|2]$ from [18, Thm. 2.3], which only applies to real symmetric Gaussian matrices. Applying the theorem yields

$$\lim_{N \rightarrow \infty} N^2 \text{Cov}(\langle G_1 - \mathbb{E}G_1 \rangle, \overline{\langle G_2 - \mathbb{E}G_2 \rangle}) = \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{(x - z_1)^2} \frac{1}{(y - z_2)^2} u(x, y) dx dy, \quad (4.28)$$

where u is given by

$$u(x, y) = \frac{\sqrt{4 - x^2} \sqrt{4 - y^2}}{\pi^2} \cdot \int_0^1 \frac{1 - w^2}{w^2(x - y)^2 - wxy(1 - w)^2 + (1 - w^2)^2} dw. \quad (4.29)$$

In the notation of [18], Equation (4.28) corresponds to considering a matrix-valued Gaussian process for which the distribution at time 1 coincides with the law of a GOE matrix V , as well as the functions $f(x) = (x - z_1)^{-1}$ and $g(y) = (y - z_2)^{-1}$ for fixed $z_1, z_2 \in \mathbb{C}$ with $\Im z_1, \Im z_2 \gtrsim 1$. We remark that the functions f and g indeed satisfy the polynomial bound assumed in [18, Thm. 2.3]. As (4.28) is obtained for the fixed times $s = t = 1$, we omit the time dependence from the kernel. By substituting $v = (\frac{1-w}{1+w})^2$ and using partial fractions, the w -integration in (4.29) can be carried out explicitly, yielding the form in (2.19).

Comparing (4.27) and (4.28), we obtain (3.20) in the case $k = \ell = 1$.

The induction step uses the divided difference structure from Corollary 3.10(iii) and is similar to the proof of [13, Lem. 4.1]. Assume that the integral representation (3.20) holds for $\tilde{m}_{GUE}[S_1|S_2]$ with $|S_1| = k$ and $|S_2| \leq \ell$ for some fixed $k \geq 1$. We start by rewriting

$$\begin{aligned} & \tilde{m}_{GUE}[1, \dots, k|k+1, \dots, k+\ell+1] \\ &= \frac{\tilde{m}_{GUE}[1, \dots, k|k+2, \dots, k+\ell+1] - \tilde{m}_{GUE}[1, \dots, k|k+1, k+3, \dots, k+\ell+1]}{z_{k+2} - z_{k+1}}, \end{aligned}$$

where the induction hypothesis applies to both summands in the numerator, respectively. Noting that

$$\begin{aligned} & \sum_{i=k+2}^{k+\ell} \frac{1}{(y-z_i)^2} \cdot \prod_{j \neq i} \frac{1}{y-z_j} - \sum_{\substack{i=k+1 \\ i \neq k+2}}^{k+\ell} \frac{1}{(y-z_i)^2} \cdot \prod_{j \neq i} \frac{1}{y-z_j} \\ &= (z_{k+2} - z_{k+1}) \sum_{i=k+1}^{k+\ell} \frac{1}{(y-z_i)^2} \cdot \prod_{j \neq i} \frac{1}{y-z_j}, \end{aligned}$$

we obtain (3.20) as claimed. Since $\tilde{m}_{GUE}[S_1|S_2] = \tilde{m}_{GUE}[S_2|S_1]$ by definition, the same argument applies for the other entry of $\tilde{m}_{GUE}[\cdot, \cdot]$. \square

4.4 Proof of Theorem 3.8 (CLT for resolvents)

We use proof by induction over the number of factors on the left-hand side of (3.15). To keep the notation simple, we drop the superscripts (k_j, a_j) of $X_{\alpha_j}^{(k_j, a_j)}$ throughout this section. First, the base case $p = 1, 2$ readily follows from the definition and Lemma 2.5, which give $\mathbb{E}X_{\alpha_1} = 0$ and

$$N^2 \mathbb{E}(X_{\alpha_1} X_{\alpha_2}) = m_2[\alpha_1|\alpha_2] + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N}\eta_* \eta_*^{k_1-a_1/2} \eta_*^{k_2-a_2/2}}\right),$$

respectively. Next, fix $p \in \mathbb{N}$ and assume that

$$N^p \mathbb{E}\left(\prod_{j=1}^p X_{\alpha_j}\right) = \sum_{Q \in \text{Pair}(\{1, \dots, n\})} \prod_{(i,j) \in Q} m_2[\alpha_i|\alpha_j] + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N}\eta_* \prod_{l=1}^p \eta_*^{k_l-a_l/2}}\right) \quad (4.30)$$

holds for $n = 1, \dots, p$. It remains to consider $N^{p+1} \mathbb{E}(X_{\alpha_1} \dots X_{\alpha_{p+1}})$. When writing out any of the factors, we label the spectral parameters and matrices involved with a superscript $1, \dots, p+1$ according to the factor they appear in.

We start by considering the case $k_1 = 1$ and inductively extend the statement to larger products. For the first part of the induction step, we need to compute

$$N^{p+1} \mathbb{E}\left(\left(\langle T_1^{(1)} \rangle - \mathbb{E}\langle T_1^{(1)} \rangle\right) X_{\alpha_2} \dots X_{\alpha_{p+1}}\right). \quad (4.31)$$

Combining (4.11) and Lemma 2.3 (together with (4.10) and Theorem 1.3) yields

$$\begin{aligned} \langle T_1^{(1)} \rangle - \mathbb{E}\langle T_1^{(1)} \rangle &= -m(z_1^{(1)}) \left(\langle \underline{WT}_1^{(1)} \rangle + q_{1,1}^{(1)} \langle \underline{WG}_1^{(1)} \rangle \langle A_1^{(1)} \rangle \right) \\ &\quad - \frac{\kappa_4}{N} \mathcal{E}[T_1] + \mathcal{O}_{\prec}\left(\frac{1}{N \sqrt{N}\eta_* \eta_*^{k_1-a_1/2}}\right), \end{aligned} \quad (4.32)$$

where the superscript in $q^{(1)}$ indicates that the arguments are taken from α_1 . We use (4.32) to replace the first factor in (4.31). The underlined terms are again treated by cumulant expansion. This yields

$$-m(z_1^{(1)}) N^{p+1} \mathbb{E}\left(\langle \underline{WT}_1 \rangle X_{\alpha_2} \dots X_{\alpha_{p+1}}\right)$$

$$\begin{aligned}
 &= m(z_1^{(1)}) \sum_{i=2}^{p+1} \sum_{j=1}^{k_i} \mathbb{E} \left(\langle T_1^{(1)} T_j^{(i)} \dots T_{k_i}^{(i)} T_1^{(i)} \dots T_{j-1}^{(i)} G_j^{(i)} \rangle \cdot N^{p-1} \prod_{r \neq 1, i} X_{\alpha_r} \right) \\
 &\quad - m(z_1^{(1)}) N^p \sum_{n \geq 2} \sum_{x, y \in [N]} \sum_{\nu \in \{xy, yx\}^n} \frac{\kappa(xy, \nu)}{n!} \mathbb{E} \partial_\nu \left((T_1^{(1)})_{yx} X_{\alpha_2} \dots X_{\alpha_{p+1}} \right), \tag{4.33}
 \end{aligned}$$

where the $n = 1$ contribution was again evaluated using (4.8) and the computations in the proof of Lemma 4.3. Note that we obtain one term for each resolvent in the product $X_{\alpha_2} \dots X_{\alpha_{p+1}}$ from applying the Leibniz rule. By the local law (1.15), we have

$$\begin{aligned}
 &\mathbb{E} \langle T_1^{(1)} T_j^{(i)} \dots T_{k_i}^{(i)} T_1^{(i)} \dots T_{j-1}^{(i)} G_j^{(i)} \rangle \\
 &= m_1 [T_1^{(1)}, T_j^{(i)}, \dots, T_{j-1}^{(i)}, G_j^{(i)}] + \mathcal{O} \left(\frac{N^\varepsilon}{N \eta_* \eta_*^{k_i - (a_1 + a_i)/2}} \right).
 \end{aligned}$$

Further, note that every X_{α_j} contains a normalized trace, which yields a total of p factors N^{-1} when the derivative $\partial_\nu((T_1^{(1)})_{yx} X_{\alpha_2} \dots X_{\alpha_{p+1}})$ is evaluated. We can hence argue as in the proof of Lemma 2.5 to conclude that the contributions for $n \geq 4$ in (4.33) are sub-leading. Similarly, most of the terms arising for $n = 2$ and $n = 3$ in (4.33) were already computed in the proof of Lemma 2.5 and their treatment here is analogous. We, therefore, focus on the differences between (4.25) and (4.33) below. Recall that $|\nu_1|$ denotes the total number of derivatives acting on $(T_1^{(1)})_{yx}$ in a given term.

Estimate for the $n = 2$ term of (4.33): Compared to (4.25), new terms only arise if $|\nu_1| = 0$ and the two derivatives act on different factors of the product $X_{\alpha_2} \dots X_{\alpha_{p+1}}$. As the resulting derivative always includes the off-diagonal entry $(T_1^{(1)})_{yx}$, we obtain

$$N^p \sum_{x, y} \sum_{\nu \in \{xy, yx\}^2} \frac{\kappa(xy, \nu)}{2} \mathbb{E} \partial_\nu \left((T_1^{(1)})_{yx} X_{\alpha_2} \dots X_{\alpha_{p+1}} \right) = \mathcal{O} \left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \prod_{l=1}^{p+1} \eta_*^{k_l - a_l/2}} \right)$$

by isotropic resummation. Recall that every X_{α_j} contains a normalized trace such that the factor N^{-p} in front of the sum is balanced out.

Computation of the $n = 3$ term of (4.33): We distinguish three cases for $|\nu_1|$.

Case 1 ($|\nu_1| = 3$): After evaluating the derivative, we apply (1.16) to get

$$\begin{aligned}
 &N \sum_{x, y} \sum_{\nu \in \{xy, yx\}^3} \frac{\kappa(xy, \nu)}{6} \mathbb{E} \left((\partial_\nu (T_1)_{yx}) X_{\alpha_2} \dots X_{\alpha_{p+1}} \right) \\
 &= -\kappa_4 m(z_1^{(1)})^4 \langle A_1^{(1)} \rangle \mathbb{E} \left(X_{\alpha_2} \dots X_{\alpha_{p+1}} \right) + \mathcal{O} \left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \prod_{l=1}^{p+1} \eta_*^{k_l - a_l/2}} \right).
 \end{aligned}$$

In particular, adding the above terms for the cumulant expansions resulting from the two underlined terms in (4.32) yields

$$\begin{aligned}
 &-m(z_1^{(1)}) \left(-\kappa_4 m(z_1^{(1)})^4 \langle A_1^{(1)} \rangle - \kappa_4 q_{1,1}^{(1)} m(z_1^{(1)})^4 \langle A_1^{(1)} \rangle \right) \mathbb{E} \left(X_{\alpha_2} \dots X_{\alpha_{p+1}} \right) \\
 &= \kappa_4 \mathcal{E}[T_1] \mathbb{E} \left(X_{\alpha_2} \dots X_{\alpha_{p+1}} \right).
 \end{aligned}$$

Case 2 ($|\nu_1| = 1$): Whenever the two remaining derivatives act on different factors of the product $X_{\alpha_2} \dots X_{\alpha_{p+1}}$, (4.21) always yields an off-diagonal entry of a resolvent chain that can be used to apply isotropic resummation. We can thus include any terms of this build in the error in (4.33). Otherwise, we evaluate the term similar to (4.26). Let again $\nu_2 = \nu \setminus \nu_1$. For $i = 2, \dots, p + 1$, it follows that

$$N^p \sum_{x, y} \frac{\kappa_4}{N^2} \mathbb{E} \left((\partial_{\nu_1} (T_1)_{yx}) X_{\alpha_2} \dots X_{\alpha_{i-1}} (\partial_{\nu_2} X_{\alpha_i}) X_{\alpha_{i+1}} \dots X_{\alpha_{p+1}} \right)$$

$$\begin{aligned}
 &= -\kappa_4 \sum_{s=1}^{k_i} \left(\sum_{t=1}^s m(z_1^{(1)})^2 \langle A_1^{(1)} \odot M_{[s,t]}^{(i)} \rangle \langle M_{(t,\dots,k_i,1,\dots,s)}^{(i)} \rangle \right. \\
 &\quad \left. + \sum_{t=s}^{k_i} m(z_1^{(1)})^2 \langle A_1^{(1)} \odot M_{(s,\dots,k_i,1,\dots,t)}^{(i)} \rangle \langle M_{[t,s]}^{(i)} \rangle \right) \mathbb{E} \left(\prod_{r \neq 1,i} X_{\alpha_r} \right) \\
 &\quad + \mathcal{O} \left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \prod_{l=1}^{p+1} \eta_*^{k_l - a_l/2}} \right),
 \end{aligned}$$

where the superscript in $M^{(i)}$ indicates that the arguments are taken from α_i . Recall that \odot denotes the Hadamard product. We emphasize that this term is obtained six times in total when all mixed derivatives obtained from the Leibniz rule are summed up.

The last case for $n = 3$ ($|\nu_1|$ even) can be treated similarly to the corresponding cases of (4.25). We omit the details.

Repeating the expansion for the $\langle W G_1^{(1)} \rangle$ term in (4.32), we hence obtain

$$\begin{aligned}
 &N^{p+1} \mathbb{E} \left(-m(z_1^{(1)}) \langle W T_1^{(1)} \rangle + q_{1,1}^{(1)} \langle W G_1^{(1)} \rangle \langle A_1^{(1)} \rangle X_{\alpha_2} \dots X_{\alpha_{p+1}} \right) \\
 &= N^{p-1} \mathbb{E} \left(\prod_{r \neq 1,j} X_{\alpha_r} \right) \cdot m(z_1^{(1)}) \sum_{i=2}^{p+1} \left[\sum_{j=1}^{k_i} \mathbf{m}_1 [T_1^{(1)}, T_j^{(i)} \dots T_{j-1}^{(i)}, G_j^{(i)}] \right. \\
 &\quad \left. + q_{1,1}^{(1)} \sum_{j=1}^{k_i} \mathbf{m}_1 [G_1^{(1)}, T_j^{(i)} \dots T_{j-1}^{(i)}, G_j^{(i)}] \langle A_1^{(1)} \rangle \right. \\
 &\quad \left. + \kappa_4 \sum_{s=1}^{k_i} \left(\sum_{t=1}^s m(z_1^{(1)})^2 \langle A_1^{(1)} \odot M_{[s,t]}^{(i)} \rangle \langle M_{(t,\dots,k_i,1,\dots,s)}^{(i)} \rangle \right. \right. \\
 &\quad \left. \left. + \sum_{t=s}^{k_i} m(z_1^{(1)})^2 \langle A_1^{(1)} \odot M_{(t,\dots,k_i,1,\dots,s)}^{(i)} \rangle \langle M_{[s,t]}^{(i)} \rangle \right) \right. \\
 &\quad \left. + \kappa_4 q_{1,1}^{(1)} \sum_{s=1}^{k_i} \left(\sum_{t=1}^s m(z_1^{(1)})^2 \langle (A_1^{(1)})_{xx} \odot M_{[s,t]}^{(i)} \rangle \langle M_{(t,\dots,k_i,1,\dots,s)}^{(i)} \rangle \right. \right. \\
 &\quad \left. \left. + \sum_{t=s}^{k_i} m(z_1^{(1)})^2 \langle A_1^{(1)} \odot M_{(t,\dots,k_i,1,\dots,s)}^{(i)} \rangle \langle M_{[s,t]}^{(i)} \rangle \right) \langle A_1^{(1)} \rangle \right] + \mathcal{O} \left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \prod_{l=1}^{p+1} \eta_*^{k_l - a_l/2}} \right).
 \end{aligned}$$

Adding the contribution for all three terms in (4.32) together allows rewriting (4.31) as

$$\begin{aligned}
 &N^{p+1} \mathbb{E} \left((\langle T_1^{(1)} \rangle - \mathbb{E} \langle T_1^{(1)} \rangle) X_{\alpha_2} \dots X_{\alpha_{p+1}} \right) \\
 &= \sum_{i=2}^{p+1} \mathbf{m}_2 [T_1^{(1)} | T_1^{(i)}, \dots, T_{k_i}^{(i)}] \cdot N^{p-1} \mathbb{E} \left(\prod_{r \neq 1,j} X_{\alpha_r} \right) + \mathcal{O} \left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \prod_{l=1}^{p+1} \eta_*^{k_l - a_l/2}} \right),
 \end{aligned}$$

where we used (3.7) with $k = 1$ and $\ell = k_i$. As the remaining product only consists of $p - 1$ factors, we can apply the induction hypothesis (4.30) and conclude

$$N^{p+1} \mathbb{E} \left(\prod_{j=1}^{p+1} X_{\alpha_j} \right) = \sum_{Q \in \text{Pair}([p+1])} \prod_{(i,j) \in Q} \mathbf{m}_2 [\alpha_i | \alpha_j] + \mathcal{O} \left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \prod_{l=1}^{p+1} \eta_*^{k_l - a_l/2}} \right) \quad (4.34)$$

for the case $\alpha_1 = \{(z_1^{(1)}, A_1^{(1)})\}$.

Assume next that (4.34) holds for all multi-indices α_1 that consist of up to $k_1 - 1$ pairs $(z_j^{(1)}, A_j^{(1)})$, and consider α_1 of length k_1 . As in the base case, we replace the factor X_{α_1} in the product by its leading term using (4.24). This gives

$$N^{p+1} \mathbb{E} \left(X_{\alpha_1} \dots X_{\alpha_{p+1}} \right)$$

$$\begin{aligned}
 &= N^{p+1} m(z_1^{(1)}) \mathbb{E} \left(\left(- \langle \underline{WT_1^{(1)}} \dots T_{k_1}^{(1)} \rangle \right. \right. \\
 &\quad + \langle T_2^{(1)} \dots T_{k_1-1}^{(1)} G_{k_1}^{(1)} A_{k_1}^{(1)} A_1^{(1)} \rangle - \mathbb{E} \langle T_2^{(1)} \dots T_{k_1-1}^{(1)} G_{k_1}^{(1)} A_{k_1}^{(1)} A_1^{(1)} \rangle \\
 &\quad + \frac{\kappa_4}{N} \mathcal{E}[T_2^{(1)}, \dots, T_{k_1-1}^{(1)}, G_{k_1}^{(1)} A_{k_1}^{(1)} A_1^{(1)}] \\
 &\quad + \sum_{j=1}^{k-1} (\langle T_1^{(1)} \dots T_{j-1}^{(1)} G_j^{(1)} \rangle - \mathbb{E} \langle T_1^{(1)} \dots T_{j-1}^{(1)} G_j^{(1)} \rangle) \\
 &\quad + \frac{\kappa_4}{N} \mathcal{E}[T_1^{(1)}, \dots, T_{j-1}^{(1)}, G_j^{(1)}] \mathbf{m}_1[T_j^{(1)}, \dots, T_{k_1}^{(1)}] \\
 &\quad + \sum_{j=2}^k \mathbf{m}_1[T_1^{(1)}, \dots, T_{j-1}^{(1)}, G_j^{(1)}] (\langle T_j^{(1)} \dots T_{k_1}^{(1)} \rangle - \mathbb{E} \langle T_j^{(1)} \dots T_{k_1}^{(1)} \rangle + \frac{\kappa_4}{N} \mathcal{E}[T_j^{(1)}, \dots, T_{k_1}^{(1)}]) \\
 &\quad \left. + \frac{\kappa_4}{N} \mathcal{E}[T_1^{(1)}, \dots, T_{k_1}^{(1)}] \right) X_{\alpha_2} \dots X_{\alpha_{p+1}} \Big) + \mathcal{O} \left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \prod_{l=1}^{p+1} \eta_*^{k_l - a_l / 2}} \right),
 \end{aligned} \tag{4.35}$$

where the underlined term can again be treated by cumulant expansion. Similar to (4.33), we evaluate

$$\begin{aligned}
 &N^{p+1} \mathbb{E} \left(- m(z_1^{(1)}) \langle \underline{WT_1^{(1)}} \dots T_{k_1}^{(1)} \rangle X_{\alpha_2} \dots X_{\alpha_{p+1}} \right) \\
 &= m(z_1^{(1)}) \sum_{i=2}^{p+1} \left(\sum_{j=1}^{k_i} \mathbf{m}_1[T_1^{(1)}, \dots, T_{k_1}^{(1)}, T_j^{(i)}, \dots, T_{j-1}^{(i)}, G_j^{(i)}] \right. \\
 &\quad + \kappa_4 \sum_{r=1}^{k_1} \sum_{s=k+1}^{k+\ell} \left(\sum_{t=1}^s \langle M_{[r]}^{(1)} \odot M_{(s, \dots, k_i, 1, \dots, t)}^{(i)} \rangle \langle (M_{[r, k]}^{(1)} A_k^{(1)}) \odot M_{[t, s]}^{(i)} \rangle \right. \\
 &\quad \left. \left. + \sum_{t=s}^{k_i} \langle M_{[r]}^{(1)} \odot M_{[s, t]}^{(i)} \rangle \langle (M_{[r, k]}^{(1)} A_k^{(1)}) \odot M_{(t, \dots, k_i, 1, \dots, s)}^{(i)} \rangle \right) \right) N^{p-1} \mathbb{E} \left(\prod_{r \neq 1, i} X_{\alpha_r} \right) \\
 &\quad + \mathcal{O} \left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \prod_{l=1}^{p+1} \eta_*^{k_l - a_l / 2}} \right),
 \end{aligned} \tag{4.36}$$

and apply the induction hypothesis (4.30) for the product of $p - 1$ factors. Note that any terms remaining in (4.35) now involve at most $k_1 - 1$ resolvents in the first factor. Hence, (4.34) applies and we obtain, e.g.,

$$\begin{aligned}
 &N^{p+1} \mathbb{E} \left(\left(\langle T_2^{(1)} \dots T_{k_1-1}^{(1)} G_{k_1}^{(1)} A_{k_1}^{(1)} A_1^{(1)} \rangle - \mathbb{E} \langle T_2^{(1)} \dots T_{k_1-1}^{(1)} G_{k_1}^{(1)} A_{k_1}^{(1)} A_1^{(1)} \rangle \right) \prod_{r=2}^{p+1} X_{\alpha_r} \right) \\
 &= \sum_{r=2}^{p+1} \mathbf{m}_2[T_2^{(1)}, \dots, T_{k_1-1}^{(1)}, G_{k_1}^{(1)} A_{k_1}^{(1)} A_1^{(1)} | T_1^{(r)}, \dots, T_{k_i}^{(r)}] \left(\sum_{Q \in \text{Pair}([p+1] \setminus \{1, i\})} \prod_{(i, j) \in Q} \mathbf{m}_2[\alpha_i | \alpha_j] \right) \\
 &\quad + \mathcal{O} \left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \prod_{l=1}^{p+1} \eta_*^{k_l - a_l / 2}} \right).
 \end{aligned}$$

Moreover, note that (4.36) contains the source term of the recursion for $\mathcal{E}[\cdot]$ (cf. Lemma 2.3), which we can combine with the terms involving $\mathcal{E}[\cdot]$ in (4.35). Using (3.1), the terms cancel. We conclude that (4.35) evaluates to

$$\begin{aligned}
 &N^{p+1} \mathbb{E} \left(X_{\alpha_1}^{(k_1)} \dots X_{\alpha_{p+1}}^{(k_{p+1})} \right) \\
 &= \sum_{r=2}^{p+1} \left[m(z_1^{(1)}) \left(\sum_{j=1}^{k_r} \mathbf{m}_1[T_1^{(1)}, \dots, T_{k_1}^{(1)}, T_j^{(r)}, \dots, T_{j-1}^{(r)}, G_j^{(r)}] \right) \right]
 \end{aligned}$$

$$\begin{aligned}
 & + \kappa_4 \sum_{r=1}^{k_1} \sum_{s=k+1}^{k+\ell} \left(\sum_{t=1}^s \langle M_{[r]}^{(1)} \odot M_{(s,\dots,k_i,1,\dots,t)}^{(i)} \rangle \langle (M_{[r,k]}^{(1)} A_k^{(1)}) \odot M_{[t,s]}^{(i)} \rangle \right. \\
 & + \sum_{t=s}^{k_i} \langle M_{[r]}^{(1)} \odot M_{[s,t]}^{(i)} \rangle \langle (M_{[r,k]}^{(1)} A_k^{(1)}) \odot M_{(t,\dots,k_i,1,\dots,s)}^{(i)} \rangle \left. \right) \\
 & + \mathfrak{m}_2[T_2^{(1)}, \dots, T_{k_1-1}^{(1)}, G_{k_1}^{(1)} A_{k_1}^{(1)} A_1^{(1)} | T_1^{(r)}, \dots, T_{k_r}^{(r)}] \\
 & + \sum_{j=2}^k \mathfrak{m}_1[T_1^{(1)}, \dots, T_{j-1}^{(1)}, G_j^{(1)}] \mathfrak{m}_2[T_j^{(1)}, \dots, T_{k_1}^{(1)} | T_1^{(r)}, \dots, T_{k_r}^{(r)}] \\
 & + \sum_{j=2}^k \mathfrak{m}_2[T_1^{(1)}, \dots, T_{j-1}^{(1)}, G_j^{(1)} | T_1^{(r)}, \dots, T_{k_r}^{(r)}] \mathfrak{m}_1[T_j^{(1)}, \dots, T_{k_1}^{(1)}] \left. \right) \\
 & \times \left(\sum_{Q \in \text{Pair}([p+1] \setminus \{1, i\})} \prod_{(i,j) \in Q} \mathfrak{m}_2[\alpha_i | \alpha_j] \right) + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \prod_{l=1}^{p+1} \eta_*^{k_l - a_l/2}}\right) \\
 & = \sum_{Q \in \text{Pair}([p+1])} \prod_{(i,j) \in Q} \mathfrak{m}_2[\alpha_i | \alpha_j] + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N} \eta_* \prod_{l=1}^{p+1} \eta_*^{k_l - a_l/2}}\right)
 \end{aligned}$$

where the last equation follows from (3.7). Hence, (4.34) also holds for α_1 of length k_1 . Moreover, (4.30) stays true if the product on the left-hand side contains $p + 1$ factors, which concludes the proof of (3.15). \square

4.5 Proof of Theorem 2.7 (multi-point functional CLT)

The proof of the multi-point functional CLT in Theorem 2.7 consists of two parts. In the first step, we use Helffer-Sjöstrand representation (see [16]) to express $f_1(W) \dots f_k(W)$ as an integral of products of resolvents at different spectral parameters. This relates the linear statistics Y_α back to the resolvent chains studied in Section 3. The second step is the computation of the leading terms, which establishes the covariance structure in (2.9).

By eigenvalue rigidity (see e.g., [21, Thm. 7.6] or [23]), the spectrum of W is contained in $[-2 - \varepsilon, 2 + \varepsilon]$ for any small $\varepsilon > 0$ with very high probability. In particular, we have $(f \cdot \chi)(W) = f(W)$ with very high probability for any smooth cutoff function χ that is, e.g., equal to one on $[-5/2, 5/2]$ and equal to zero on $[-3, 3]^c$. It is thus sufficient to consider $f_j \in H_0^p([-3, 3]) =: H_0^p$, i.e., Sobolev functions on \mathbb{R} that are non-zero only on $[-3, 3]$. Moreover, recall that every deterministic matrix A_j in the product $F_{[1,k]}$ can be decomposed as $A_j = \langle A_j \rangle \text{Id} + \hat{A}_j$ with $\langle \hat{A}_j \rangle = 0$. We thus assume w.l.o.g. that the deterministic matrices A_j are either traceless or equal to the identity matrix. Moreover, we restrict the following argument to the case $a = k$, i.e., all deterministic matrices are traceless, and fix $p = \lceil k/2 \rceil + 1$ (resp. $q = \lceil \ell/2 \rceil + 1$) throughout. The proof in the general case is analogous and hence omitted.

Let $f \in H_0^p$ and define the *almost analytic extension* of f of order p by

$$f_{\mathbb{C}}(z) = f_{\mathbb{C},p}(x + i\eta) := \left[\sum_{j=0}^{p-1} \frac{(i\eta)^j}{j!} f^{(j)}(x) \right] \tilde{\chi}(N^\gamma \eta) \tag{4.37}$$

where $\tilde{\chi}$ is a smooth cutoff function that is equal to one on $[-5, 5]$ and vanishes on $[-10, 10]^c$. Note that (4.37) together with (2.2) implies the bound

$$\int_{\mathbb{R}} |\partial_{\bar{z}} f_{\mathbb{C},p}(x + i\eta)| dx \lesssim \eta^{p-1} \|f\|_{H^p} \lesssim \eta^{p-1} N^{\gamma p}. \tag{4.38}$$

By Helffer-Sjöstrand representation, we have

$$f(\lambda) = \frac{1}{\pi} \int_{\mathbb{C}} \frac{\partial_{\bar{z}} f_{\mathbb{C}}(z)}{\lambda - z} d^2z, \tag{4.39}$$

where $d^2z = dx d\eta$ denotes the Lebesgue measure on $\mathbb{C} \equiv \mathbb{R}^2$ with $z = x + i\eta$ and $\partial_{\bar{z}} = (\partial_x + i\partial_{\eta})/2$. Applying (4.39) for f_1, \dots, f_k , respectively, we obtain

$$Y_{\alpha}^{(k,a)} = \frac{N}{\pi^k} \int_{\mathbb{C}^k} \left[\prod_{j=1}^k (\partial_{\bar{z}}(f_j)_{\mathbb{C}})(z_j) \right] \left(\langle G(z_1) \dots G(z_k) A_k \rangle - \mathbb{E} \langle G(z_1) \dots G(z_k) A_k \rangle \right) d^2z_{[k+\ell]} \tag{4.40}$$

with $d^2z_{[k+\ell]} = d^2z_1 \dots d^2z_{k+\ell}$. Let $c > 0$ and define

$$\eta_0 := N^{-1+c}.$$

We start by showing that the contribution from the regime $|\eta_j| \leq \eta_0$ for some $j \in [k]$ to the integral (4.40) is negligible. W.l.o.g. assume that $|\eta_j| \leq \eta_0$ only for the single index $j = 1$. The general case is similar and yields an even smaller bound (cf. proof of [13, Thm. 2.6]). Our key tool is the following variant of Stokes' theorem

$$\int_{-10}^{10} \int_{\tilde{\eta}}^{10} \partial_{\bar{z}} \Phi(x + i\eta) h(x + i\eta) dx d\eta = \frac{1}{2i} \int_{-10}^{10} \Phi(x + i\tilde{\eta}) h(x + i\tilde{\eta}) dx, \tag{4.41}$$

which holds for any $\tilde{\eta} \in [0, 10]$, and for any $\Phi, h \in H^1(\mathbb{C}) \equiv H^1(\mathbb{R}^2)$ such that $\partial_{\bar{z}} h = 0$ on the domain of integration and Φ vanishes on the left, right, and top boundary of the domain of integration. Applying (4.41) repeatedly for the variables z_2, \dots, z_k and introducing the interval notation $dx_{[i,j]} = dx_i dx_{i+1} \dots dx_j$ as well as $d\eta_{[i,j]} = d\eta_i d\eta_{i+1} \dots d\eta_j$ for $i < j$, we obtain

$$\begin{aligned} & \left| \int dx_{[k]} \int_{\substack{|\eta_i| \geq \eta_0 \\ i \in [2,k]}} d\eta_{[2,k]} \int_{-\eta_0}^{\eta_0} d\eta_1 \left[\prod_{j=1}^k (\partial_{\bar{z}}(f_j)_{\mathbb{C}})(z_j) \right] \left(\langle G(z_1) A_1 \dots G(z_k) A_k \rangle \right. \right. \\ & \quad \left. \left. - \mathbb{E} \langle G(z_1) A_1 \dots G(z_k) A_k \rangle \right) \right| \\ &= \frac{1}{2^{k-1}} \left| \int dx_{[k]} \int_{-\eta_0}^{\eta_0} d\eta_1 (\partial_{\bar{z}}(f_1)_{\mathbb{C}})(z_1) \left[\prod_{j=2}^k (f_j)_{\mathbb{C}}(x_j + i\eta_0) \right] \right. \\ & \quad \left. \times \left(\langle G(z_1) A_1 G(x_2 + i\eta_0) \dots G(x_k + i\eta_0) A_k \rangle - \mathbb{E} \langle G(z_1) A_1 G(x_2 + i\eta_0) \dots G(x_k + i\eta_0) A_k \rangle \right) \right| \\ &=: I_1 + I_2, \end{aligned}$$

where I_1 and I_2 contain the $|\eta_1| \leq \eta_r := N^{-5k}$ and the $\eta_r \leq |\eta_1| \leq \eta_0$ regime of the η_1 integration, respectively. For the smallest values of η_1 , the trivial estimate

$$|\langle G(z_1) A_1 \dots G(z_k) A_k \rangle| \leq \prod_j \|G(z_j) A_j\| \leq \prod_j |\eta_j|^{-1}$$

together with (4.38) implies that

$$I_1 \lesssim N^{-k}$$

as the x_j -integral of $(f_j)_{\mathbb{C}}(x_j + i\eta_0)$ for $j \in [2, k]$ is of order one due to Assumption 2.1. For I_2 , we use the bound

$$|\langle G(z_1) A_1 \dots G(z_k) A_k \rangle| \prec N^{k/2-1} \prod_{j \in [k]} \frac{1}{\rho(x_j + iN^{-2/3})} \left(1 + \frac{1}{N|\eta_j|} \right)$$

from [12, Lem. 6.1]. Here, $\rho(z) := \pi^{-1}|\Im m(z)|$ for $z \in \mathbb{C} \setminus \mathbb{R}$ denotes the harmonic extension of the semicircle density with $\rho(x + i0) = \rho_{sc}(x)$. This yields

$$I_2 \prec \eta_0(N\eta_0)^{k/2} \|f_1\|_{H^p}.$$

Overall, we conclude that

$$\begin{aligned} Y_\alpha^{(k,a)} &= \frac{N}{\pi^k} \int_{\mathbb{R}^k} dx_{[k]} \int_{[\eta_0,10]^k} d\eta_{[k]} \left[\prod_{j=1}^k (\partial_{\bar{z}}(f_j)_{\mathbb{C}})(z_j) \right] \\ &\quad \times \left(\langle G(z_1)A_1 \dots G(z_k)A_k \rangle - \mathbb{E} \langle G(z_1)A_1 \dots G(z_k)A_k \rangle \right) \\ &\quad + \mathcal{O}_{\prec} \left(\eta_0(N\eta_0)^{k/2} \max_j \|f_j\|_{H^p} \right) \end{aligned} \tag{4.42}$$

Equation (4.42) now reduces the proof of the multi-point functional CLT for general f_1, \dots, f_k to the CLT for resolvents in Theorem 3.8.

It remains to compute the covariance structure (2.9). By (4.42) and Lemma 2.5, we have

$$\begin{aligned} &\mathbb{E} \left(Y_\alpha^{(k,a)} Y_\beta^{(\ell,b)} \right) \\ &= \frac{1}{\pi^{k+\ell}} \int_{\mathbb{R}^k} dx_{[k]} \int_{[\eta_0,10]^k} d\eta_{[k]} \left[\prod_{i=1}^k (\partial_{\bar{z}}(f_i)_{\mathbb{C},p})(z_i) \right] \int_{\mathbb{R}^\ell} dx_{[k+1,k+\ell]} \int_{[\eta_0,10]^\ell} d\eta_{[k+1,k+\ell]} \\ &\quad \times \left[\prod_{j=k+1}^{\ell} (\partial_{\bar{z}}(f_j)_{\mathbb{C},q})(z_j) \right] \mathfrak{m}_2[G(z_1)A_1, \dots, G(z_k)A_k | G(z_{k+1})A_{k+1}, \dots, G(z_{k+\ell})A_{k+\ell}] \\ &\quad + \mathcal{O}_{\prec} \left(\frac{N^\varepsilon \max_{i \in [k]} \|f_i\|_{H^p} \max_{j \in [k+1,k+\ell]} \|f_j\|_{H^q}}{\sqrt{N}} \right) \end{aligned} \tag{4.43}$$

where we estimated the error coming from Lemma 2.5 as

$$\begin{aligned} &\frac{1}{\pi^{k+\ell}} \int_{\mathbb{R}^{k+\ell}} dx_{[k+\ell]} \int_{[\eta_0,10]^{k+\ell}} d\eta_{[k+\ell]} \left[\prod_{j=1}^{k+\ell} (\partial_{\bar{z}}(f_j)_{\mathbb{C}})(z_j) \right] \\ &\quad \times \left(N^2 \mathbb{E} \left[\left(\langle G(z_1) \dots A_k \rangle - \mathbb{E} \langle G(z_1) \dots A_k \rangle \right) \left(\langle G(z_{k+1}) \dots A_{k+\ell} \rangle - \mathbb{E} \langle G(z_{k+1}) \dots A_{k+\ell} \rangle \right) \right] \right. \\ &\quad \left. - \mathfrak{m}_2[G(z_1)A_1, \dots, G(z_k)A_k | G(z_{k+1})A_{k+1}, \dots, G(z_{k+\ell})A_{k+\ell}] \right) \\ &= \mathcal{O} \left(\frac{N^\varepsilon \eta_0^{3/2} \max_{i \in [k]} \|f_i\|_{H^p} \max_{j \in [k+1,k+\ell]} \|f_j\|_{H^q}}{\sqrt{N}} \right). \end{aligned}$$

More precisely, we considered the regime $\eta_1 \leq \dots \leq \eta_k$ and $\eta_{k+1} \leq \dots \leq \eta_{k+\ell}$, as all other regimes give the same contribution by symmetry, and applied (4.41) for the variables $i \in [2, k]$ and $j \in [k + 2, k + \ell]$. Estimating the remaining $\partial_{\bar{z}}(f_1)_{\mathbb{C},p}(z_1)$ and $\partial_{\bar{z}}(f_{k+1})_{\mathbb{C},q}(z_{k+1})$ using (4.38), we obtain a bound of order $|\eta_0|^{p+q}$. Recall that applying Lemma 2.5 yields an $(\sqrt{N}\eta_*\eta_*^{k+\ell-(a+b)/2})^{-1}$ error, where $\eta_* = \min_j \eta_j = \eta_0$ due to the choice of domain of integration. \square

4.6 Proof of Corollary 2.9 (limiting covariance in Theorem 2.7)

Equation (2.14) is immediate from the explicit formula for $\mathfrak{m}_{GUE}[\cdot]$ in [40, Thm. 2.4]. Moreover, the proof of (2.15) identical to [13, Lem. 4.1], as the integral defining Φ_π only involves first-order free cumulants. We hence only focus on the proof of (2.17). Abbreviate $U = U_1 \cup U_2$ and note that

$$\Phi_{\pi_1 \times \pi_2, U_1 \times U_2}(f_1, \dots, f_{k+\ell})$$

$$\begin{aligned}
 &= \frac{1}{\pi^{k+\ell}} \left(\int_{\mathbb{R}^{|U_1|}} \int_{[\eta_0, 10]^{|U_1|}} \left[\prod_{j \in U_1} (\partial_{\bar{z}}(f_j)_{\mathbb{C}})(z_j) \right] m_{\circ\circ}[U_1|U_2] \prod_{j \in U_2} d^2 z_j \right) \\
 &\quad \times \prod_{\substack{B \in \pi_1 \cup \pi_2 \\ \text{not marked}}} \left(\int_{\mathbb{R}^{|B|}} \int_{[\eta_0, 10]^{|B|}} \left[\prod_{j \in B} (\partial_{\bar{z}}(f_j)_{\mathbb{C}})(z_j) \right] m_{\circ}[B] \prod_{j \in B} d^2 z_j \right),
 \end{aligned}$$

where we set again $d^2 z_j = dx_j d\eta_j$ for $z_j = x_j + i\eta_j$. In particular, the product in the last line can again be evaluated using [13, Lem. 4.1]. It remains to compute the integral involving $m_{\circ\circ}[U_1|U_2]$. We claim that

$$\begin{aligned}
 &\frac{1}{\pi^{|U|}} \int_{\mathbb{R}^{|U_1|}} \int_{[\eta_0, 10]^{|U_1|}} \left[\prod_{j \in U_1} (\partial_{\bar{z}}(f_j)_{\mathbb{C}})(z_j) \right] m[U_1|U_2] \prod_{j \in U_2} d^2 z_j \\
 &= \text{sc}[i_1, \dots, i_n | i_{n+1}, \dots, i_{n+m}] + \mathcal{O}\left(\eta_0^2 (N\eta_0)^{(|U|)/2} \max_{i \in [k]} \|f_i\|_{H^p} \max_{j \in [k+1, k+\ell]} \|f_j\|_{H^q}\right) \quad (4.44)
 \end{aligned}$$

where $U_1 = \{i_1, \dots, i_n\}$ and $U_2 = \{i_{n+1}, \dots, i_{n+m}\}$. The corresponding result for $\text{sc}_{\circ\circ}[\cdot|\cdot]$ is then immediate from the second-order moment-cumulant relation (2.13).

To establish (4.44), we use the explicit integral representation for $\tilde{m}[\cdot|\cdot]$ from Corollary 3.11 and rewrite the resulting multi-integral involving the kernel (2.19). Let again $U := U_1 \cup U_2$. Noting that both $|x - z_j|$ and $|y - z_j|$ are bounded from below by η_0 for any j and recalling the bound (4.38) for the almost analytic extension, we have

$$\begin{aligned}
 &\int_{[-2, 2]^2} \int_{\mathbb{R}^{|U_1|}} \int_{[\eta_0, 10]^{|U_1|}} \left| \left[\prod_{j \in U_1} (\partial_{\bar{z}}(f_j)_{\mathbb{C}})(z_j) \right] \left(\sum_{i \in U_1} \frac{1}{(x - z_i)^2} \cdot \prod_{j \neq i} \frac{1}{x - z_j} \right) \right. \\
 &\quad \times \left. \left(\sum_{i \in U_2} \frac{1}{(y - z_i)^2} \cdot \prod_{j \neq i} \frac{1}{y - z_j} \right) \frac{u(x, y)}{2} \right| \left[\prod_{j \in U_2} d^2 z_j \right] dx dy < \infty
 \end{aligned}$$

for any fixed N , as the z_j integrations are bounded by an (N -dependent) constant, which is integrable w.r.t. to the measure $\nu(dx, dy) = u(x, y) dx dy$ (cf. Remark to Corollary 3.11). Hence, Fubini's theorem allows interchanging the order of integration and move the x and y integrations inside. A brief calculation using (4.41) yields

$$\int_{\mathbb{R}} \int_{[\eta_r, 10]} (\partial_{\bar{z}}(f_j)_{\mathbb{C}})(z_j) \frac{1}{x - z_j} d^2 z_j = \pi f(x) + \mathcal{O}(\eta_r)$$

with $\eta_r = N^{-5k}$, and we further have

$$\frac{1}{\pi} \int_{\mathbb{R}} \int_{[\eta_0, 10]} (\partial_{\bar{z}}(f_j)_{\mathbb{C}})(z_j) \left(\frac{1}{x - z_j} \right)^2 d^2 z_j = f'(x) + \mathcal{O}(\eta_r)$$

using integration by parts. Hence,

$$\begin{aligned}
 &\frac{1}{\pi^{|U|}} \int_{\mathbb{R}^{|U_1|}} \int_{[\eta_0, 10]^{|U_1|}} \left[\prod_{j \in U_1} (\partial_{\bar{z}}(f_j)_{\mathbb{C}})(z_j) \right] \\
 &\quad \times \left(\sum_{i \in U_1} \frac{1}{(x - z_i)^2} \cdot \prod_{j \neq i} \frac{1}{x - z_j} \right) \left(\sum_{i \in U_2} \frac{1}{(y - z_i)^2} \cdot \prod_{j \neq i} \frac{1}{y - z_j} \right) \left[\prod_{j \in U_2} d^2 z_j \right] \\
 &= \left(\sum_{i \in U_1} f'_i(x) \cdot \prod_{j \neq i} f_j(x) \right) \left(\sum_{i \in U_2} f'_i(y) \cdot \prod_{j \neq i} f_j(y) \right) \\
 &\quad + \mathcal{O}\left(\eta_0^2 (N\eta_0)^{(|U|)/2} \max_{i \in [k]} \|f_i\|_{H^p} \max_{j \in [k+1, k+\ell]} \|f_j\|_{H^q}\right)
 \end{aligned}$$

such that (4.44) follows from the Leibniz rule. Recall the $[\eta_r, \eta_0]$ regime can be added back in exchange for an $\mathcal{O}(\eta_0^2 (N\eta_0)^{(|U|)/2} \max_{i \in [k]} \|f_i\|_{H^p} \max_{j \in [k+1, k+\ell]} \|f_j\|_{H^q})$ error. This yields (4.44), which concludes the proof of (2.17). \square

4.7 Proof of Corollary 2.13 (application to thermalization)

Throughout the proof, we set $\tilde{f}_j(x) = e^{it_j x} \chi(x)$ with a symmetric smooth cutoff function χ that is equal to one on $[-5/2, 5/2]$ and equal to zero on $[-3, 3]^c$. By eigenvalue rigidity, the functions $\tilde{f}_j(W)$ and $f_j(W) = e^{it_j W}$ coincide with high probability and we can use them interchangeably. The deterministic approximation as well as the Gaussian fluctuations around it are now immediate from [12, Cor. 2.7] and Theorem 2.7, respectively. Note that we use the multi-point functional CLT for the macroscopic regime due to $t \in \mathbb{R}$ being N -independent. The limiting variance can be read off from Corollary 2.9. Observing that

$$\overline{\langle A_1(t)A_2 \rangle} = \langle e^{itW} A_1^* e^{-itW} A_2^* \rangle$$

we set $f_1(x) = e^{itx}$, $f_2(x) = e^{-itx}$, $f_3(x) = e^{itx}$, and $f_4(x) = e^{-itx}$ as well as $A_3 = A_1^*$, and $A_4 = A_2^*$ to apply (2.14). As A_1 and A_2 are assumed to be traceless, only the terms corresponding to the permutations

$$\pi \in \{(14)(23), (13)(24), (1)(24)(3), (14)(2)(3), (1)(24)(3), (13)(2)(4)\} \subset \overline{NCP}(2, 2) \tag{4.45}$$

as well as the terms corresponding to the marked partitions

$$\pi \in \left\{ \{ \underline{1}, \underline{2} \} \times \{ \{ \underline{3}, \underline{4} \} \}, \{ \underline{1}, \underline{2} \} \times \{ \{ \underline{3}, \underline{4} \} \}, \{ \underline{1}, \underline{2} \} \times \{ \{ \underline{3}, \underline{4} \} \}, \{ \underline{1}, \underline{2} \} \times \{ \{ \underline{3}, \underline{4} \} \} \right\} \tag{4.46}$$

contribute to the limiting variance of $\langle A_1(t)A_2 \rangle$. Note that the marked blocks in (4.46) are distinguished by underlining.

It remains to discuss the $t \rightarrow \infty$ limit. We have

$$\int_{-2}^2 e^{itx} \rho_{sc}(x) dx = \frac{J_1(2t)}{t}$$

where J_1 is a Bessel function of the first kind obeying the asymptotics

$$J_1(x) = -\cos\left(x + \frac{\pi}{2}\right) \sqrt{\frac{2}{\pi x}} + \mathcal{O}\left(\frac{1}{x^{3/2}}\right), \quad x \gg 1.$$

In particular,

$$sc_o[1] = sc[1] = \frac{J_1(2t)}{t} = \mathcal{O}\left(\frac{1}{t^{3/2}}\right), \quad t \gg 1.$$

Hence, it readily follows that the term corresponding to $\langle A_1 A_3 \rangle \langle A_2 A_4 \rangle = \langle |A_1|^2 \rangle \langle |A_2|^2 \rangle$ is the largest among the contributions from $\overline{NCP}(2, 2)$ for large t , giving

$$sc_o[1, 4] sc_o[2, 3] = \left(1 - \frac{J_1(2t)J_1(2t)}{t^2}\right) \left(1 - \frac{J_1(2t)J_1(2t)}{t^2}\right) = 1 + \mathcal{O}\left(\frac{1}{t^3}\right), \quad t \gg 1,$$

where we also used the symmetry $J_1(-x) = -J_1(x)$. Moreover, we obtain that, e.g.,

$$sc_o[1] sc_o[2, 3] sc_o[4] = \mathcal{O}\left(\frac{1}{t^3}\right), \quad t \gg 1,$$

with the remaining permutations in (4.45) yielding contributions of comparable or lower order in the $t \rightarrow \infty$ limit.

Lastly, we consider the marked partitions in (4.46), which correspond to the term of $\text{Var}[\xi]$ that contains $\langle A_1 A_2 \rangle \langle A_3 A_4 \rangle = |\langle A_1 A_2 \rangle|^2$. As we work in the macroscopic regime of Theorem 2.7, Equation (2.18) coincides with the limiting covariance structure of the CLT in [37, Sect. 2]. Hence, we obtain, e.g.,

$$sc[1|4] = \frac{1}{2\pi^2} \int_{-2}^2 \int_{-2}^2 \frac{1 - \cos(t(x-y))}{(x-y)^2} \frac{4-xy}{\sqrt{4-x^2}\sqrt{4-y^2}} dx dy. \tag{4.47}$$

In particular, the cutoff χ does not enter the computation. Note that the expressions for $\text{sc}[1|4]$ and $\text{sc}[2|3]$ resp. $\text{sc}[1|3]$ and $\text{sc}[2|4]$ yield the same contribution by symmetry. The integral on the right-hand side of (4.47) is finite, however, it will grow with t as $t \rightarrow \infty$. To identify the asymptotics, we distinguish between the contributions of the bulk regime

$$\frac{4 - xy}{\sqrt{4 - x^2}\sqrt{4 - y^2}} = \mathcal{O}(1),$$

and the edge regime where the denominator $(\sqrt{4 - x^2}\sqrt{4 - y^2})^{-1}$ becomes singular. As

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_{-2}^2 \int_{-2}^2 \frac{1 - \cos(t(x - y))}{(x - y)^2} = 4\pi,$$

the contribution of the bulk is readily identified to be $\mathcal{O}(t)$. In the edge regime, we expand the square root in the denominator and further consider the contributions around the diagonal ($|x - y| \lesssim t^{-1}$) and away from it separately whenever x and y are close to the same value. This also yields a bound of order $\mathcal{O}(t)$, implying that $\text{sc}[1|4] = \mathcal{O}(t)$. Recalling the identity $\text{sc}_{\circ\circ}[1|4] = \text{sc}[1|4] - \text{sc}_{\circ}[1, 4]$ from (2.13), we obtain

$$\text{sc}_{\circ\circ}[1|4]\text{sc}_{\circ}[2]\text{sc}_{\circ}[3] = \mathcal{O}\left(\frac{1}{t^2}\right), \quad t \gg 1.$$

The other marked partitions in 4.46 give rise to terms of comparable order. Summing up all contributions yields (2.21). The proof of (2.22) is analogous and hence omitted. \square

A Proof of Corollary 3.4 (meta argument)

Recall from Lemma 2.3 that

$$\mathbb{E}\langle T_1 \dots T_k \rangle = \mathbf{m}_1[T_1, \dots, T_k] + \frac{\kappa_4}{N} \mathcal{E}[T_1, \dots, T_k] + \mathcal{O}\left(\frac{N^\varepsilon}{N \sqrt{N} \eta_*^{k-a/2}}\right),$$

i.e., $\mathcal{E}[T_1, \dots, T_k]$ constitutes the first subleading term of $\mathbb{E}\langle T_1 \dots T_k \rangle$. In particular, we have

$$N(\mathbb{E}\langle T_1 \dots T_k \rangle - \mathbf{m}_1[T_1, \dots, T_k]) = \kappa_4 \mathcal{E}[T_1, \dots, T_k] + \mathcal{O}\left(\frac{N^\varepsilon}{\sqrt{N} \eta_*^{k-a/2}}\right), \quad (\text{A.1})$$

where the quantity on the left-hand side of (A.1) satisfies the properties stated in Corollary 3.4. For $\mathbb{E}\langle T_1 \dots T_k \rangle$, this is immediate from the cyclicity of the trace and the resolvent identity $G_k G_1 = \frac{G_k - G_1}{z_k - z_1}$, while the corresponding properties for $\mathbf{m}_1[\cdot]$ follow from (1.17) and [13, Lem. 5.4]. Note that (3.4) is a special case of (3.3) and that (3.5) is obtained by iterating (3.4). Once the formula (3.5) is established, the permutation symmetry readily follows from the divided difference structure. Hence, Corollary 3.4(iii) follows from (i) and (ii).

It remains to show that $\mathcal{E}[\cdot]$ satisfies the same cyclicity and divided difference properties as the quantity on the left-hand side of (A.1). Note that simply taking the $N \rightarrow \infty$ limit and applying Lemma 2.3 is not sufficient if $A_j \neq \text{Id}$ for some j , as the deterministic matrices are themselves N -dependent quantities. Instead, let $L \in \mathbb{N}$ and consider the $NL \times NL$ Wigner matrix \mathscr{W} as well as the deterministic matrices $\mathscr{A}_1, \dots, \mathscr{A}_k \in \mathbb{C}^{NL \times NL}$. Here, \mathscr{W} is defined using the same random variables χ_d, χ_{od} as W (i.e., $\sqrt{N}W$ and $\sqrt{NL}\mathscr{W}$ have the same entry distribution) and we define

$$\mathscr{A}_j := A_j \otimes \text{Id}_{L \times L}, \quad j = 1, \dots, k,$$

where \otimes denotes the tensor product, i.e.,

$$\mathcal{A}_j = \begin{pmatrix} (A_j)_{11}\text{Id}_{L \times L} & \cdots & (A_j)_{1N}\text{Id}_{L \times L} \\ \vdots & \ddots & \vdots \\ (A_j)_{N1}\text{Id}_{L \times L} & \cdots & (A_j)_{NN}\text{Id}_{L \times L} \end{pmatrix}.$$

Next, let $\mathcal{G}_j := (\mathcal{W} - z_j)^{-1}$ and $\mathcal{T}_j := \mathcal{G}_j \mathcal{A}_j$ for $j = 1, \dots, k$ and denote by $\mathfrak{m}_1[\mathcal{T}_1, \dots, \mathcal{T}_k]$ the deterministic approximation of $\langle \mathcal{T}_1 \dots \mathcal{T}_k \rangle$. As both \mathcal{W} and W are Wigner matrices and $\langle \mathcal{A}_i \mathcal{A}_j \rangle = \langle (A_i A_j) \otimes \text{Id}_{L \times L} \rangle = \langle A_i A_j \rangle$ by definition of the tensor product, it follows from the closed form of $\mathfrak{m}_1[\cdot]$ in [13, Thm. 2.6] that

$$\mathfrak{m}_1[\mathcal{T}_1, \dots, \mathcal{T}_k] = \mathfrak{m}_1[T_1, \dots, T_k]. \tag{A.2}$$

Similarly, the $1/N$ error term of $\mathbb{E}\langle \mathcal{T}_1 \dots \mathcal{T}_k \rangle$ is given by $\mathcal{E}[T_1, \dots, T_k]$, since (A.2) ensures that we obtain the same recursion from Definition 3.1.

We thus conclude that

$$\begin{aligned} & |\mathcal{E}[T_1, \dots, T_k] - \mathcal{E}[T_2, \dots, T_k, T_1]| \\ & \leq |\mathcal{E}[T_1, \dots, T_k] - N(\mathbb{E}\langle \mathcal{T}_1 \dots \mathcal{T}_k \rangle - \mathfrak{m}_1[\mathcal{T}_1, \dots, \mathcal{T}_k])| \\ & \quad + |\mathcal{E}[T_2, \dots, T_k, T_1] - N(\mathbb{E}\langle \mathcal{T}_2 \dots \mathcal{T}_k, \mathcal{T}_1 \rangle - \mathfrak{m}_1[\mathcal{T}_2, \dots, \mathcal{T}_k, \mathcal{T}_1])| \\ & = \mathcal{O}\left(\frac{(NL)^\varepsilon}{\sqrt{NL\eta_*} \eta_*^{k-a/2}}\right) \end{aligned}$$

by Lemma 2.3 for any $\eta_* \gg (NL)^{-1}$. Letting $L \rightarrow \infty$ while keeping all other parameters $N, z_1, \dots, z_k, A_1, \dots, A_k$ fixed yields

$$\mathcal{E}[T_1, \dots, T_k] = \mathcal{E}[T_2, \dots, T_k, T_1],$$

i.e., $\mathcal{E}[\cdot]$ is cyclic as claimed in part (i) of Corollary 3.4. Similarly, we obtain that

$$\left| \mathcal{E}[T_1, \dots, T_{k-1}, G_k] - \frac{\mathcal{E}[T_2, \dots, T_{k-1}, G_k A_1] - \mathcal{E}[T_1, \dots, T_{k-1}]}{z_k - z_1} \right| = \mathcal{O}\left(\frac{(NL)^\varepsilon}{\sqrt{NL\eta_*} \eta_*^{k-a/2}}\right)$$

whenever $z_1 \neq z_k$ and $A_k = \text{Id}$, which implies (3.3). Recalling that (i) and (ii) of Corollary 3.4 imply (iii), the proof is complete. \square

References

- [1] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, 2009. MR2760897
- [2] Z. D. Bai and J. W. Silverstein. CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.*, 32(1A):553–605, 2004. MR2040792
- [3] Z. D. Bai and J. Yao. On the convergence of the spectral empirical process of Wigner matrices. *Bernoulli*, 11:1059–1092, 2005. MR2189081
- [4] Z. Bao and Y. He. Quantitative CLT for linear eigenvalue statistics of Wigner matrices. *Preprint, arXiv:2103.05402*, 2021. MR4677731
- [5] Z. Bao, K. Schnelli, and Y. Xu. Central limit theorem for mesoscopic eigenvalue statistics of the free sum of matrices. *Int. Math. Res. Not.*, 2022(7):5320–5382, 2022. MR4403963
- [6] F. Bekerman, T. Leblé, and S. Serfaty. CLT for fluctuations of β -ensembles with general potential. *Electron. J. Probab.*, 23:1–31, 2018. MR3885548
- [7] F. Bekerman and A. Lodhia. Mesoscopic central limit theorem for general β -ensembles. *Ann. Inst. H. Poincaré Probab. Stat.*, 54(4):1917–1938, 2018. MR3865662

- [8] G. Borot and A. Guionnet. Asymptotic expansion of β matrix models in the one-cut regime. *Commun. Math. Phys.*, 317:447–483, 2013. MR3010191
- [9] P. Bourgade, K. Mody, and M. Pain. Optimal local law and central limit theorem for β -ensembles. *Commun. Math. Phys.*, 390:1017–1079, 2022. MR4389077
- [10] G. Cipolloni, L. Erdős, and D. Schröder. Central limit theorem for linear eigenvalue statistics of non-Hermitian random matrices. *Comm. Pure Appl. Math.*, 76:946–1034, 2019. MR4569609
- [11] G. Cipolloni, L. Erdős, and D. Schröder. Eigenstate thermalization hypothesis for Wigner matrices. *Commun. Math. Phys.*, 388(2):1005–1048, 2021. MR4334253
- [12] G. Cipolloni, L. Erdős, and D. Schröder. Optimal multi-resolvent local laws for Wigner matrices. *Electron. J. Probab.*, 27:1–38, 2022. MR4479913
- [13] G. Cipolloni, L. Erdős, and D. Schröder. Thermalisation for Wigner matrices. *J. Funct. Anal.*, 282(8), 2022. MR4372147
- [14] G. Cipolloni, L. Erdős, and D. Schröder. Functional central limit theorems for Wigner matrices. *Ann. Appl. Probab.*, 33(1):447–489, 2023. MR4551555
- [15] B. Collins, J. Mingo, P. Śniady, and R. Speicher. Second order freeness and fluctuations of random matrices: III. higher order freeness and free cumulants. *Documenta Math.*, 12:1–70, 2007. MR2302524
- [16] E. B. Davies. The functional calculus. *J. London Math. Soc.*, 52(1):166–176, 1995. MR1345723
- [17] J. Deutsch. Quantum statistical mechanics in a closed system. *Phys. Rev. A*, 43:2046–2049, 1991.
- [18] M. Diaz, A. Jaramillo, and J. C. Pardo. Fluctuations for matrix-valued Gaussian processes. *Ann. Henri Poincaré*, 58(4):2216–2249, 2022. MR4492977
- [19] M. Diaz and J.A. Mingo. On the analytic structure of second-order non-commutative probability spaces and functions of bounded Fréchet variation. *Random Matrices: Theory Appl.*, 12(1):2250044, 2022. MR4552697
- [20] L. Erdős and H. C. Ji. Functional CLT for non-Hermitian random matrices. *Ann. Inst. H. Poincaré Probab. Statist.* 59(4):2083–2105, 2023. MR4663517
- [21] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. The local semicircle law for a general class of random matrices. *Electron. J. Probab.*, 18:1–58, 2013. MR3068390
- [22] L. Erdős, T. Krüger, and D. Schröder. Random matrices with slow correlation decay. *Forum Math. Sigma*, 7, E8, 2019. MR3941370
- [23] L. Erdős, H.-T. Yau, and J. Yin. Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.*, 229:1435–1515, 2012. MR2871147
- [24] M. Ledoux G. Lambert and C. Webb. Quantitative normal approximation of linear statistics of β -ensembles. *Ann. Probab.*, 47(5):2619–2685, 2019. MR4021234
- [25] A. Guionnet. Large deviation upper bounds and central limit theorems for non-commutative functionals of Gaussian large random matrices. *Ann. Inst. H. Poincaré Probab. Stat.*, 38:341–384, 2002. MR1899457
- [26] Y. He. Mesoscopic linear statistics of Wigner matrices of mixed symmetry class. *J. Stat. Phys.*, 175:932–959, 2019. MR3959983
- [27] Y. He and A. Knowles. Mesoscopic eigenvalue statistics of Wigner matrices. *Ann. Appl. Probab.*, 27(3):1510–1550, 2017. MR3678478
- [28] Y. He and A. Knowles. Mesoscopic eigenvalue density correlations of Wigner matrices. *Probab. Theory Relat. Fields*, 177:147–216, 2020. MR4095015
- [29] H. C. Ji and J. O. Lee. Gaussian fluctuations for linear spectral statistics of deformed Wigner matrices. *Random Matrices: Theory Appl.*, 9(3):2050011, 2020. MR4119597
- [30] K. Johansson. On fluctuations of eigenvalues of random Hermitian matrices. *Duke Math. J.*, 91(1):151–204, 1998. MR1487983
- [31] A. M. Khorunzhy, B. A. Khoruzhenko, and L. A. Pastur. On the $1/N$ corrections to the Green functions of random matrices with independent entries. *J. Phys. A Math. Gen.*, 28:L31, 1995. MR1325832

- [32] A. M. Khorunzhy, B. A. Khoruzhenko, and L. A. Pastur. Asymptotic properties of large random matrices with independent entries. *J. Math. Phys.*, 37:5033–5060, 1996. MR1411619
- [33] B. Landon and P. Sosoe. Almost-optimal bulk regularity conditions in the CLT for Wigner matrices. *Preprint, arXiv:2204.03419*, 2022.
- [34] Y. Li, K. Schnelli, and Y. Xu. Central limit theorem for mesoscopic eigenvalue statistics of deformed Wigner matrices and sample covariance matrices. *Ann. Inst. H. Poincaré Probab. Statist.*, 57(1):506–546, 2021. MR4255183
- [35] Y. Li and Y. Xu. On fluctuations of global and mesoscopic linear statistics of generalized Wigner matrices. *Bernoulli*, 27(2):1057–1076, 2021. MR4255226
- [36] A. Lytova. On non-Gaussian limiting laws for certain statistics of Wigner matrices. *Zh. Mat. Fiz. Anal. Geom.*, 9:536–581, 2013. MR3155024
- [37] A. Lytova and L. Pastur. Central limit theorem for linear eigenvalue statistics of the Wigner and the sample covariance random matrices. *Metrika*, 69:153–172, 2009. MR2481919
- [38] C. Male, J. A. Mingo, S. Peché, and R. Speicher. Joint global fluctuations of complex Wigner and deterministic matrices. *Random Matrices: Theory Appl.*, 11(2):2250015, 2022. MR4414193
- [39] J. A. Mingo and R. Speicher. *Free Probability and Random Matrices*. Vol. 35, Fields Institute Research Monographs, Springer, New York, 2017. MR3585560
- [40] J. Reker. Fluctuation moments for regular functions of Wigner matrices. *Math. Phys. Anal. Geom.* 27:10, 2024. MR4761437
- [41] V. Riabov. Mesoscopic eigenvalue statistics for Wigner-type matrices. *Preprint, arXiv:2301.01712*, 2023.
- [42] Z. Rudnick and P. Sarnak. The behavior of eigenstates of arithmetic hyperbolic manifolds. *Comm. Math. Phys.*, 161:195–213, 1994. MR1266075
- [43] M. Shcherbina. Central limit theorem for linear eigenvalue statistics of the Wigner and sample covariance random matrices. *Zh. Mat. Fiz. Anal. Geom.*, 7:176–192, 2011. MR2829615
- [44] M. Shcherbina. Fluctuations of linear eigenvalue statistics of β matrix models in the multi-cut regime. *J. Stat. Phys.*, 151:1004–1034, 2013. MR3063494
- [45] J. W. Silverstein. Weak convergence of random functions defined by the eigenvectors of sample covariance matrices. *Ann. Probab.*, 18(3):1174–1194, 1990. MR1062064
- [46] J. W. Silverstein. Weak convergence of a collection of random functions defined by the eigenvectors of large dimensional random matrices. *Random Matrices: Theory Appl.*, 11(4):2250033, 2022. MR4480828
- [47] Y. G. Sinai and A. B. Soshnikov. Central limit theorem for traces of large random symmetric matrices with independent matrix elements. *Bull. Brazilian Math. Soc.*, 29:1–24, 1998. MR1620151
- [48] Y. G. Sinai and A. B. Soshnikov. A refinement of Wigner’s semicircle law in a neighborhood of the spectrum edge for random symmetric matrices. *Funct. Anal. its Appl.*, 32:114–131, 1998. MR1647832
- [49] A. B. Soshnikov. The central limit theorem for local linear statistics in classical compact groups and related combinatorial identities. *Ann. Probab.*, 28(3):1353–1370, 2000. MR1797877
- [50] P. Sosoe and P. Wong. Regularity conditions in the CLT for linear eigenvalue statistics of Wigner matrices. *Adv. Math.*, 249:37–87, 2013. MR3116567

Acknowledgments. I am very grateful to László Erdős for suggesting the topic and many valuable discussions during my work on the project. I would also like to thank the two anonymous referees for their careful reading of the manuscript and detailed feedback.

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 (LOCKSS¹)
- Easy interface (EJMS²)

Economical model of EJP-ECP

- Non profit, sponsored by IMS³, BS⁴, ProjectEuclid⁵
- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund⁶ (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

¹LOCKSS: Lots of Copies Keep Stuff Safe <http://www.lockss.org/>

²EJMS: Electronic Journal Management System: <https://vtex.lt/services/ejms-peer-review/>

³IMS: Institute of Mathematical Statistics <http://www.imstat.org/>

⁴BS: Bernoulli Society <http://www.bernoulli-society.org/>

⁵Project Euclid: <https://projecteuclid.org/>

⁶IMS Open Access Fund: <https://imstat.org/shop/donation/>