Central Limit Theorems for Random Matrices: From Resolvents to Free Probability

by

Jana Reker

June, 2024

A thesis submitted to the
Graduate School
of the
Institute of Science and Technology Austria
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Committee in charge:

Ylva Götberg, Chair
László Erdős
Robert Seiringer
Antti Knowles

The thesis of Jana Reker, titled *Central Limit Theorems for Random Matrices: From Resolvents to Free Probability*, is approved by:

Supervisor: László Erdős, ISTA

Signature: _____

Committee Member: Robert Seiringer, ISTA

Signature: _____

Committee Member: Antti Knowles, University of Geneva

Signature: _____

Defense Chair: Ylva Götberg, ISTA

Signature: _____

Signed page is on file

ISTA Thesis, ISSN: 2663-337X

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: _____

Jana Reker
June, 2024

Signed page is on file

Abstract

This thesis is structured into two parts. In the first part, we consider the random variable $X := \mathrm{Tr}(f_1(W)A_1 \dots f_k(W)A_k)$ where W is an $N \times N$ Hermitian Wigner matrix, $k \in \mathbb{N}$, and we choose (possibly N-dependent) regular functions $f_1, \dots, f_k$ as well as bounded deterministic matrices $A_1, \dots, A_k$. In this context, we prove a functional central limit theorem on macroscopic and mesoscopic scales, showing that the fluctuations of X around its expectation are Gaussian and that the limiting covariance structure is given by a deterministic recursion. We further 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 Cipolloni, Erdős and Schröder [40] to products of arbitrary length $k \geq 2$. Analyzing the underlying combinatorics leads to a non-recursive formula for the variance of X as well as the covariance of X and $Y := \mathrm{Tr}(f_{k+1}(W)A_{k+1} \dots f_{k+\ell}(W)A_{k+\ell})$ of similar build. When restricted to polynomials, these formulas reproduce recent results of Male, Mingo, Peché, and Speicher [107], showing that the underlying combinatorics of non-crossing partitions and annular non-crossing permutations continue to stay valid beyond the setting of second-order free probability theory. As an application, we consider the fluctuation of $\mathrm{Tr}(\mathrm{e}^{\mathrm{i}tW}A_1\mathrm{e}^{-\mathrm{i}tW}A_2)/N$ around its thermal value $\mathrm{Tr}(A_1)\,\mathrm{Tr}(A_2)/N^2$ when t is large and give an explicit formula for the variance.

The second part of the thesis collects three smaller projects focusing on different random matrix models. In the first project, we show that a class of weakly perturbed Hamiltonians of the form $H_\lambda = H_0 + \lambda W$, where W is a Wigner matrix, exhibits prethermalization. That is, the time evolution generated by H_λ relaxes to its ultimate thermal state via an intermediate prethermal state with a lifetime of order λ^{-2}. As the main result, we obtain a general relaxation formula, expressing the perturbed dynamics via the unperturbed dynamics and the ultimate thermal state. The proof relies on a two-resolvent global law for the deformed Wigner matrix H_λ.

The second project focuses on correlated random matrices, more precisely on a correlated $N \times N$ Hermitian random matrix with a polynomially decaying metric correlation structure. A trivial a priori bound shows that the operator norm of this model is stochastically dominated by $\sqrt{N}$. However, by calculating the trace of the moments of the matrix and using the summable decay of the cumulants, the norm estimate can be improved to a bound of order one.

In the third project, we consider a multiplicative perturbation of the form $UA(t)$ where U is a unitary random matrix and $A = \mathrm{diag}(t, 1, ..., 1)$. This so-called UA model was first introduced by Fyodorov [73] for its applications in scattering theory. We give a general description of the eigenvalue trajectories obtained by varying the parameter t and introduce a flow of deterministic domains that separates the outlier resulting from the rank-one perturbation from the typical eigenvalues for all sub-critical timescales. The results are obtained under generic assumptions on U that hold for various unitary random matrices, including the circular unitary ensemble (CUE) in the original formulation of the model.

Acknowledgements

First of all, I would like to express my sincere gratitude towards my PhD advisor László Erdős, whose dedication to all aspects of academic life has been truly inspiring. I am thankful for his honest feedback, his tireless support, and many markers run out of ink in mathematical discussions.

I would also like to thank my thesis committee members, Antti Knowles and Robert Seiringer, for their support throughout my PhD studies. Many thanks also to Ylva Götberg for agreeing to act as the defense chair and to my faculty mentor, Beatriz Vicoso, for her encouragement and for chairing my qualifying exam.

I would further like to express my gratitude to the members of the Erdős group, present and past, my fellow PhD students, and my colleagues. After almost four years at IST Austria, I cannot imagine a more positive and collaborative research environment, and I am very thankful for the many interesting discussions, seminars, and courses we shared. Special thanks go to Guillaume Dubach, Joscha Henheik, and Vova Riabov for many shared head scratches and the fruitful collaboration on joint projects, and to Torben Krüger for our interesting exchange about correlated random matrices. Many thanks also to our A2P Birgit Oosthuizen-Noczil for her help in navigating any administrative hurdles along the way.

I gratefully acknowledge the financial support of my advisor's ERC Advanced Grant "RMT-Beyond" No. 101020331 and the great hospitality I experienced during academic visits throughout my PhD studies.

Last but not least, I would like to thank my family and friends, particularly my mother, late father, and boyfriend, for their love and support throughout my studies and beyond.

About the Author

Jana Reker started her undergraduate studies in mathematics at TU Braunschweig through an early studies program that allowed her to take university classes parallel to her last three years of school. Following these part-time studies, she joined TU Braunschweig as a regular student and completed a BSc and MSc in mathematics with a minor in physics. Afterwards, she obtained a Dr. rer. nat. while working and researching at Ulm University's Institute of Mathematical Finance. Jana joined ISTA in September 2020 for her PhD studies in the research group of László Erdős. Her research interests lie at the intersection of probability theory, analysis, and mathematical physics, with a focus on random matrices and free probability.

Included Publications

- J. Reker. Multi-point functional central limit theorem for Wigner matrices. *Preprint, arXiv.2307.11028*, 2023.
- J. Reker. Fluctuation moments for regular functions of Wigner matrices. *Math. Phys. Anal. Geom.*, 27:10, 2024.
- L. Erdős, J. Henheik, J. Reker, and V. Riabov. Prethermalization for deformed Wigner matrices. *Preprint, arXiv:2310.06677*, 2023.
- J. Reker. On the operator norm of a Hermitian random matrix with correlated entries. *Random Matrices: Theory Appl.*, 11(4):2250036, 2022.
- G. Dubach and J. Reker. Dynamics of a rank-one multiplicative perturbation of a unitary matrix. *To appear in: Random Matrices: Theory Appl., online ready available at https://doi.org/10.1142/S2010326324500072*, 2024. (Preprint version)

Contents

List of Figures

1 Introduction

Random matrices, i.e., matrices with random variables as entries, first appeared in Wishart's work on sample covariance matrices [141]. Several decades later, the pioneering work of Wigner [140] connected the study of Hermitian random matrices to properties of the eigenvalue and gap statistics of quantum Hamiltonians in physics, opening up general questions of universality. Apart from applications in physics, random matrices are encountered in many different fields, e.g., number theory, finance, data science, and biology (see, e.g., [5]). This chapter aims to introduce the random matrix models considered in this thesis and place the results in the general context.

1.1 Hermitian Random Matrices

Throughout the thesis, we consider $N \times N$ random matrices with entries taking values in the real or complex numbers. As most results concern matrices $W = W^*$ with a Hermitian symmetry, we start the discussion by considering Wigner matrices and some of their generalizations. To set the notation, we give an overview of the definitions. Note that defining an $N \times N$ Hermitian random matrix only requires prescribing the distribution of the $N(N+1)/2$ entries in the upper triangle, as the symmetry condition identifies the remaining entries.

GOE (Gaussian Orthogonal Ensemble) Matrices $W = W^*$ for which the entries w_{ab} in the upper triangle are independent centered real Gaussian random variables satisfying $\mathbb{E}w_{aa}^2 = 2/N$ and $\mathbb{E}w_{ab}^2 = 1/N$ for $a < b$.

GUE (Gaussian Unitary Ensemble) Matrices $W = W^*$ for which the entries in the upper triangle of W independent centered Gaussian variables, the diagonal entries w_{aa} take values in $\mathbb{R}$ and satisfy $\mathbb{E}w_{aa} = 1/N$, and the off-diagonal entries w_{ab} for $a < b$ take values in $\mathbb{C}$ and satisfy $\mathbb{E}w_{ab}^2 = 0$ and $\mathbb{E}|w_{ab}|^2 = 1/N$.

Wigner matrices Matrices $W = W^*$ for which the entries in the upper triangle are independent and there is an N-independent centered real random variable χ_d and an N-independent centered real or 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 $\mathbb{E}|\chi_{od}|^2 = 1$ as well as the existence of $C_p > 0$ for all $p \in \mathbb{N}$ such that $\mathbb{E}|\chi_d|^p + \mathbb{E}|\chi_{od}|^p \leq C_p$.

Deformed Wigner matrices Matrices of the form $H := A + W$, where $A = A^* = \mathbb{E}H$ is deterministic and W is a Wigner matrix.

Wigner-type matrices The entries of the upper triangle of W are independent, have moments of all orders, and satisfy $\mathbb{E}w_{ij} = 0$ as well as $c/N \leq \mathbb{E}|w_{ij}|^2 \leq C/N$. Depending on the source, Wigner-type models may include a diagonal deformation.

Correlated Wigner matrices The entries of W have moments of all orders and the covariance operator $S[R] := \mathbb{E}WRW$ satisfies

$$\frac{c}{N}\operatorname{Tr}[R] \leq S[R] \leq \frac{C}{N}\operatorname{Tr}[R]$$

for any positive semidefinite matrix R in the sense of quadratic forms. Depending on the source, correlated models may allow for variability in the distribution of the entries similar to the Wigner-type model and a diagonal or self-adjoint deformation.

Note that both the GOE and GUE are special cases of Wigner matrices. One obtains the respective generalizations mentioned above by relaxing the centeredness, identical distribution, and independence condition of the original model. Note that the assumptions can be relaxed even further, leading to models such as heavy-tailed or Lévy matrices [10, 3], random band matrices [28], or sparse random matrices [19, 22]. Most of the results in this thesis (Chapters 2 and 3) focus on Wigner matrices, but we also consider a deformed (Chapter 4) and a correlated model (Chapter 5). The perturbed random matrix model considered in Chapter 6 of the thesis is introduced in Section 1.2 below.

For these models, we study the spectrum of the random matrix. In the Hermitian case, the symmetry assumption ensures that the eigenvalues $\lambda_1, \dots, \lambda_N$ of W are real. They are described by the *empirical eigenvalue* measure or *empirical density of states*

$$\mu_N(\mathrm{d}x) := \frac{1}{N} \sum_{j=1}^{N} \delta_{\lambda_j}(\mathrm{d}x),$$

where δ_{x_0} denotes the Dirac measure at $x_0 \in \mathbb{R}$. In this context, we are interested in a deterministic approximation μ of μ_N as N tends to infinity, i.e., the *limiting density of states*, and the fluctuations in the spectrum of the random matrix. Note that these questions can be posed on different scales. In this thesis, we consider results on Hermitian random matrices on the *macroscopic scale*, i.e., results that concern all or an order N number of eigenvalues simultaneously, and the *mesoscopic scale*, i.e., those concerning intervals that contain N^ϵ eigenvalues for some $\epsilon \in (0, 1)$. For simplicity, we mainly focus on Wigner matrices for the following discussion.

1.1.1 Macroscopic Scale: Wigner's Semicircle Law

One of the most celebrated results concerning the spectrum of random matrices is the almost sure weak convergence of the empirical eigenvalue measure of a Wigner matrix to a deterministic probability measure. More precisely, for a wide range of N-independent test functions, it holds with probability one that

$$\lim_{N\to\infty} \int_{\mathbb{R}} f(x)\mu_N(\mathrm{d}x) = \int_{-2}^{2} f(x)\rho_{sc}(x)\mathrm{d}x,$$

where ρ_{sc} denotes the density of the semicircle law given by

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

We interpret the weak convergence of μ_N to some deterministic measure μ as a law of large numbers-type result on the macroscopic scale. Throughout the thesis, we refer to results of this form as *global laws*.

Wigner's original proof of this result (see [140]) uses a simple tree-counting argument to show that

$$\mathbb{E}\Big[\int_{\mathbb{R}} x^k \mu_N(\mathrm{d}x)\Big] = \mathbb{E}\Big[\frac{1}{N}\operatorname{Tr}[W^k]\Big] = \begin{cases} C_{k/2} + \mathcal{O}(\frac{1}{N}), & k \text{ even}, \\ \mathcal{O}(\frac{1}{N}), & k \text{ odd}, \end{cases} \tag{1.1}$$

under a suitable moment condition. Here, C_n denotes the n-th Catalan number given by

$$C_n := \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{(n+1)!n!}.$$

Using that the moments of the semicircle law satisfy

$$\int_{-2}^{2} x^k \rho_{sc}(x)\mathrm{d}x = \begin{cases} C_{k/2}, & k \text{ even}, \\ 0, & k \text{ odd}, \end{cases}$$

and a bound for the variance similar to (1.1), the limit μ is readily identified. Note, however, that this approach requires either solving the moment problem or a suitable ansatz to find the correct distribution. To identify μ from the moments directly, we may use the *Stieltjes transform*

$$m_\mu(z) = \int_{\mathbb{R}} \frac{\mu(\mathrm{d}x)}{x - z}$$

for $\Im z > 0$, i.e., for $z \in \mathbb{C}$ in the upper half plane. Expanding the fraction into a geometric series and inserting the limit of the moments in (1.1) yields

$$m_\mu(z) = -\sum_{k=0}^{\infty} \frac{1}{z^{k+1}} \int_{\mathbb{R}} x^k \mathrm{d}\mu(x) = \frac{-z + \sqrt{z^2 - 4}}{2}, \tag{1.2}$$

where the branch of the complex square root is chosen such that $\Im m_\mu > 0$ whenever $\Im z > 0$ and $m_\mu(z) \to 0$ as $|z| \to \infty$. From (1.2), the underlying measure is recovered by inverting the Stieltjes transform. We revisit Wigner's moment method in Chapter 5, where we derive bounds on the moments to estimate the operator norm of a correlated model. Moreover, computations similar to (1.1) play a key role in free probability (cf. Section 1.3).

1.1.2 Mesoscopic Scales: Local Laws

While global laws yield information on the limiting density of states as a whole, it is further possible to resolve the approximation locally. This added resolution is achieved through a more involved use of the Stieltjes transform, which connects to the resolvent of the random matrix through the identity

$$m_{\mu_N}(z) = \int_{\mathbb{R}} \frac{\mu_N(\mathrm{d}x)}{x - z} = \langle (W - z)^{-1} \rangle, \tag{1.3}$$

where $\langle \cdot \rangle := \mathrm{Tr}[\cdot]/N$ denotes the normalized trace. As the point-wise convergence of $(m_{\mu_N})_N$ is equivalent to the weak convergence if the underlying measures $(\mu_N)_N$, the global law in Section 1.1.1 immediately translates to a concentration result for $\langle (W-z)^{-1} \rangle$ with N-independent spectral parameter z. We further note that

$$\Im m_\mu(x + \mathrm{i}\eta) = \int_{\mathbb{R}} \frac{\eta}{(x-\lambda)^2 + \eta^2} \mu(\mathrm{d}\lambda) = \pi(P_\eta * \mu)(x), \tag{1.4}$$

i.e., the imaginary part of the Stieltjes transform at $z = x + \mathrm{i}\eta$ yields, up to a factor of π, the convolution of μ with the Poisson-kernel of width η. We may hence interpret η as a scaling parameter that allows zooming into the spectrum around $x \in \mathbb{R}$.

Together, the observations (1.3) and (1.4) form the starting point of the *resolvent method*. It is the key tool in proving concentration results for $(m_{\mu_N}(z))_N$ with η (much) smaller

than order one but still above the typical eigenvalue spacing. We refer to results of this type as *local laws*. For example, the local law for Wigner matrices covers the scales $N^{-1} \ll \eta \ll 1$ in the bulk of the spectrum and $N^{-2/3} \ll \eta \ll 1$ around the edges $x = \pm 2$ (see [67]). Similar results hold for the deformed, Wigner-type, and correlated models introduced above (see [69]). Moreover, recent works [36, 39, 37, 38, 31, 32] go beyond the setting of one resolvent suggested by (1.3) and consider the deterministic approximation of $G_1 A_1 \dots G_k A_k$ for $k \in \mathbb{N}$, where $G_j := (W - z_j)^{-1}$ denote resolvents of a Wigner matrix W at possibly different spectral parameters $z_1, \dots, z_k$, and $A_1, \dots, A_k$ are bounded deterministic matrices.

Once a local law is established, one obtains a powerful tool for deducing further properties of the spectrum of the random matrix. For example, a local semicircle law constitutes the first step of the celebrated three-step strategy of Erdős, Schlein, and Yau, which establishes that the asymptotic local eigenvalue statistics of Gaussian Wigner matrices are *universal* for all Wigner matrices in the sense that the universality class is only determined by the symmetry (real symmetric vs. complex Hermitian) and does not depend on the exact entry distribution (see [65, 66, 67]).

For the following discussion, two properties of the spectrum of Wigner matrices are of particular interest. First, the eigenvalues $\lambda_1, \dots, \lambda_N$ exhibit *level repulsion*, i.e., with high probability, cannot be too close to one another. This effect is easiest seen in the Gaussian case, where the joint density is given explicitly as

$$\rho(\lambda_1, \dots, \lambda_N) = C \prod_{j<k} (\lambda_j - \lambda_k)^\beta \exp\Big(- \frac{N\beta}{2} \sum_{n=1}^{N} \lambda_n^2 \Big).$$

Here, the choice of $\beta = 1, 2$ in the formula for $\rho(\lambda_1, \dots, \lambda_N)$ corresponds to the real symmetric (GOE) and complex Hermitian case (GUE), respectively, while the constant C ensures the normalization to a probability density. The same repulsive effect is observed for general Wigner matrices, although its proof is more involved than for the Gaussian case (see [64]). Additionally, the eigenvalues of a Wigner matrix are *rigid*, i.e., with high probability close to their expected location γ_j given by

$$\frac{j}{N} = \int_{-\infty}^{\gamma_j} \rho_{sc}(x) \mathrm{d}x,$$

with an error that is only slightly bigger than the typical local eigenvalue spacing (see, e.g., [67]). For example, we have

$$|\lambda_j - \gamma_j| < \frac{N^\epsilon}{N}$$

for any $\epsilon > 0$ and eigenvalues in the bulk of the spectrum, i.e., $\alpha N \leq j \leq (1 - \alpha)N$ for some $\alpha \in (0, 1)$. Both level repulsion and eigenvalue rigidity indicate that the random variables $\lambda_1, \dots, \lambda_N$ are strongly correlated. This is already evident from comparing a sample of eigenvalues of a random matrix with points that are drawn independently following the semicircle law (see Fig. 1.1). Observe that the histogram on the left appears more uniform while the independent sample on the right exhibits more fluctuations.

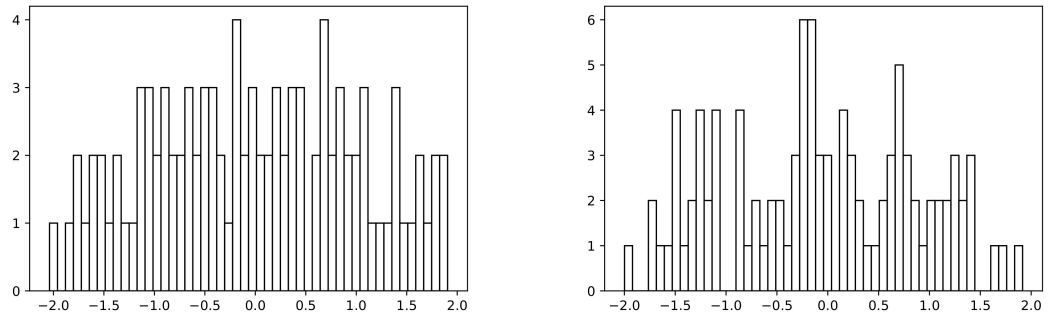

Fig. 1.1. Histograms showing the eigenvalues of a 100×100 GOE matrix (left) and 100 independent samples from the semicircle law (right).

1.1.3 Fluctuations in the Spectrum

With global and local laws representing law of large numbers-type results on different scales, the natural next question concerns the fluctuation around the deterministic approximation. Due to the strong dependence of the eigenvalues, classical central limit theorem-type results aimed at independent or weakly dependent random variables do not apply. However, the linear statistics

$$\operatorname{Tr} f(W) = \sum_{j=1}^{N} f(\lambda_j), \tag{1.5}$$

with an N-independent regular test function $f : \mathbb{R} \to \mathbb{R}$ have a variance of order one and, in fact, satisfy a central limit theorem (CLT) with a Gaussian limit for many random matrix models (see [105] for the Wigner case). Remarkably, the effect of the dependent random variables only manifests in the anomalous scaling, and removing the usual $\sqrt{N}$ factor renders the fluctuations of (1.5) still Gaussian.

Similar to the resolution to mesoscopic scales in the local laws, one may also choose an N-independent, compactly supported function g and consider (1.5) for test functions

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

that live on a scale of $N^{-\gamma}$, $\gamma \in (0, 1)$, around a fixed energy $E \in [-2, 2]$ in the spectrum. Again, a central limit theorem with a Gaussian limit holds for Wigner matrices (see [88]), but also for deformed and Wigner-type models (see [33, 125, 124]).

We remark that the statistics in (1.5) are inherently tracial, i.e., only involve the eigenvalues of the underlying random matrix. The recent results of Cipolloni, Erdős, and Schröder [40] consider centered linear statistics of the form

$$\begin{aligned} L_N(f, A) &:= \operatorname{Tr}[f(W)A] - \mathbb{E}\operatorname{Tr}[f(W)A] \\ &= \sum_{j=1}^{N} f(\lambda_j)\langle \mathbf{u}_j, A\mathbf{u}_j\rangle - \mathbb{E}\Big[\sum_{j=1}^{N} f(\lambda_j)\langle \mathbf{u}_j, A\mathbf{u}_j\rangle\Big], \end{aligned} \tag{1.6}$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_N$ denote the normalized eigenvectors of W and A is a bounded deterministic matrix. Testing against A allows studying $f(W)$ as a matrix rather than merely its trace and, therefore, goes beyond the features described by (1.5). In particular, the

linear statistics $L_N(f, A)$ in (1.6) are sensitive to the eigenvectors of W as well as the eigenvalues. Splitting the matrix A according to the decomposition

$$A = \langle A \rangle I + \mathring{A}_d + \mathring{A}_{od}$$

into a tracial, a traceless diagonal, and a traceless off-diagonal part, the corresponding modes of $L_N(f, A)$ in (1.6) are approximately distributed as

$$\begin{aligned}\Big(L_N(f, I), \sqrt{C_N} L_N(f, \mathring{A}_d), \sqrt{C_N} L_N(f, \mathring{A}_{od})\Big) \\ \stackrel{m}{=} \Big(\xi_{tr}(f), \xi_d(f, \mathring{A}_d), \xi_{od}(f, \mathring{A}_{od})\Big) + \mathcal{O}_m\Big(\sqrt{\frac{C_N}{N}}\Big),\end{aligned}$$

where $\stackrel{m}{=}$ and $\mathcal{O}_m(\cdot)$ indicate that the respective relations hold in the sense of moments. The right-hand side is given by three independent centred N-dependent Gaussian processes $\xi_{tr}(f)$, $\xi_d(f, \mathring{A}_d)$, and $\xi_{od}(f, \mathring{A}_{od})$. Here, the factor C_N is of order one for the macroscopic regime but becomes N-dependent for mesoscopic scales. In the bulk regime, it behaves approximately as N^γ, showing that the modes corresponding to $\mathring{A}_d$ and $\mathring{A}_{od}$ fluctuate on a smaller scale than the tracial mode corresponding to $\langle A \rangle \mathrm{Id}$. We consider a generalization of this theorem to the multi-resolvent setting of [39, 37] in Chapters 2 and 3.

1.2 Perturbations

For applications in physics, random matrix models like the ones introduced in Section 1.1 serve as toy models for more complex real-world systems. Studying the behavior of such a model under a small perturbation gives insight into the stability of its properties and can aid in detecting general underlying mechanisms. In this thesis, we study two models of this type: a deterministic system perturbed by a Wigner random matrix (Chapter 4) and a multiplicative rank-one perturbation of a unitary random matrix model (Chapter 6).

1.2.1 Rank-One Perturbations

One way to describe a small change in a system is by considering rank-one resp. general low-rank perturbations. For this, the nature of the perturbation is often chosen such that it breaks a symmetry of the original model, e.g.,

$$M(t) := W + \mathrm{i} t v v^* \tag{1.7}$$

for some $t \geq 0$, where W is either a $N \times N$ GUE (see, e.g., [78, 79, 75]) or Wigner matrix (see [57]) and $v \in \mathbb{C}^N$ is a (random) unit vector. Another example of a perturbation of a normal operator resulting in a loss of normality is the UA model introduced in [73], which can be seen as the multiplicative analog of (1.7). It is given by

$$M(t) := UA, \quad A = \mathrm{diag}(t, 1, \ldots, 1) \tag{1.8}$$

for $t \in [0, 1]$, where the random matrix U in (1.8) belongs to the Circular Unitary Ensemble (CUE), i.e., U is Haar-distributed on the unitary group. Models of this kind naturally arise in scattering theory and open quantum systems (see, e.g., the review [80]).

For both (1.7) and (1.8), the rank-one perturbation induces an *outlier* in the spectrum, i.e., an eigenvalue that behaves differently from the others. The transition is governed

by the strength of the perturbation. In particular, there is a critical scale at which the outlier becomes distinguishable. For the models in (1.7) and (1.8), these critical scales are $t = 1 + \mathcal{O}(N^{-1/3})$ (see [57]) and $t = \mathcal{O}(N^{-1/2})$, respectively. The emergence of the outlier for the UA model is studied in more detail in Chapter 6 below. We further remark that evaluating models like (1.7) and (1.8) at criticality may be of independent interest, as the scaling allows for explicit computations, e.g., of the eigenvalue distribution (see [72] and the review [71]).

1.2.2 Random Perturbations

Apart from adding a perturbation to a random matrix model, we may also consider deterministic systems subject to a random perturbation, e.g., given by a Wigner matrix. From a physics point of view, these random matrices describe a mean-field setting in which hopping from any state of the system to any other state is statistically equally likely. Perturbing a deterministic system by a Wigner matrix hence couples all modes, possibly breaking a prescribed symmetry or rendering an integrable system chaotic.

A particular example are quantum Hamiltonians of the form

$$H_\lambda = H_0 + \lambda W \tag{1.9}$$

where H_0 is a Hermitian deterministic matrix describing an initial system, W is a Wigner matrix, and $\lambda > 0$ is a (small) parameter regulating the strength of the perturbation. In the language of Section 1.1, H_λ in (1.9) is a deformed Wigner matrix with an added scaling induced by λ. Let $A \in \mathbb{C}^{N\times N}$ denote an observable and $\rho \in \mathbb{C}^{N\times N}$ be a state that we time evolve under H_λ to obtain $\rho_\lambda(t) = \mathrm{e}^{-\mathrm{i}tH_\lambda}\rho\mathrm{e}^{\mathrm{i}tH_\lambda}$. In this setting, the unperturbed time evolution of the quantum expectation value $\mathrm{Tr}[\rho_0(t)A]$ strongly depends on all its constituents and might hence exhibit qualitatively different and generally complex behavior. In contrast, the perturbed evolution obtained by letting $\lambda > 0$ relaxes to equilibrium via a robust mechanism well-described in the physics literature. However, a rigorous mathematical treatment has not been available so far.

A particularly interesting case occurs if both the perturbed and unperturbed systems equilibrate but do not approach the same limiting value. In this case, we may observe *prethermalization* (see, e.g., [114, 47, 48]), i.e., the equilibration of the perturbed dynamics happens in two stages. More precisely, this is the case if the free evolution $\langle A\rangle_{\rho_0(t)}$ has a limiting value $\langle A\rangle_{\rho_{\mathrm{pre}}}$, which does not agree with the microcanonical average $\langle A\rangle_{\mathrm{mc},\lambda}$ of the Hamiltonian H_λ. In this case, we find

$$\langle A\rangle_{\rho_\lambda(t)} \approx \begin{cases} \langle A\rangle_{\rho_{\mathrm{pre}}} & \text{for} \quad t \ll \lambda^{-2}, \\ \langle A\rangle_{\mathrm{mc},\lambda} & \text{for} \quad t \gg \lambda^{-2}. \end{cases}$$

Prethermalization often occurs in systems where H_0 has an additional symmetry (or conserved quantity) that is broken by the perturbation (cf. [115]). We prove prethermalization for the system in (1.9) in Chapter 4.

1.3 Connections to Free Probability

The field of free probability is a relatively young mathematical discipline that emerged in the 1980s due to the seminal work of Voiculescu, who observed that certain abstract

concepts in the study of operator algebras exhibit parallels to classical probability theory and isolated its core concept of *freeness* (see, e.g., [8, 111] for an introduction). While the research questions in this thesis are only incidentally motivated by applications in free probability, several results in Chapters 2 and 3 are formulated using free probability concepts. In this section, we give an overview of the definitions and connect them to the study of random matrices.

1.3.1 General Setup

Similar to Kolmogorov's axioms in classical probability theory, the study of free probability starts by setting up a suitable structure. Instead of a classical probability space, we now consider an unital algebra where the role of the probability measure is played by a suitably normalized functional. In this setting, one can then introduce freeness as the analog of independence in classical probability theory. We summarize the definitions below.

(Tracial) non-commutative probability space A tuple $(\mathcal{A}, \varphi_1)$ of a unital associative algebra $\mathcal{A}$ and a (tracial) linear functional $\varphi_1 : \mathcal{A} \to \mathbb{C}$ with $\varphi_1(1_\mathcal{A}) = 1$ where $1_\mathcal{A}$ denotes the one element of the algebra.

(Non-commutive) distribution The distribution of an element $a \in \mathcal{A}$ is defined through its moments $(\varphi_1(a^k))_k$. Mixed moments are defined analogously.

Freeness Unital subalgebras $(\mathcal{A}_i)_{i\in I}$ are called *free*, if alternating products of centered elements are centered, i.e., if $\varphi_1(a_1 \dots a_k) = 0$ whenever $\varphi_1(a_1) = \dots = \varphi_1(a_k) = 0$ and consecutive factors a_i and a_{i+1} are taken from different subalgebras. We may also consider the freeness of individual elements by considering the subalgebras generated by them.

In the context of random matrices, the most important example of a non-commutative probability space is $(\mathcal{A}, \varphi_1) = (\mathcal{M}_{N\times N}(L^{\infty-}(\Omega, \mathbb{P})), \mathbb{E}\langle\cdot\rangle)$, where $(\Omega, \mathbb{P})$ is a classical probability space, $\mathcal{M}_{N\times N}(S)$ denotes the $N \times N$-matrices with entries in S, and the space

$$L^{\infty-}(\Omega, \mathbb{P}) := \bigcap_{1\le p<\infty} L^p(\Omega, \mathbb{P})$$

contains all random variables with all finite moments. Recall that we denote the normalized trace by $\langle\cdot\rangle$. By definition, $\mathcal{M}_{N\times N}(L^{\infty-}(\Omega, \mathbb{P}))$ contains Wigner matrices and their generalizations introduced in Section 1.1, but also deterministic matrices.

When studying $N \times N$ matrices, the appropriate notion of freeness is that of *asymptotic freeness*, i.e., requiring $\varphi_1(a_1 \dots a_k)$ to vanish in the $N \to \infty$ limit instead of evaluating it for a fixed N. A theorem of Voiculescu (see, e.g. [8, Thm. 5.4.5]) asserts that asymptotic freeness holds for independent Wigner matrices. Moreover, families $(W_j)_j$ and $(A_j)_j$ of independent Wigner and deterministic matrices are asymptotically free (see [111, Sect. 4, Thm. 20]), and the mixed moments $\mathbb{E}\langle W_1 A_1 \dots W_k A_k\rangle$ can be expressed using non-crossing partitions. Recall that a partition of $\{1, \dots, k\}$ is called crossing if there are two blocks S_1, S_2 with $x_1, y_1 \in S_1$ and $x_2, y_2 \in S_2$ that satisfy $x_1 < x_2 < y_1 < y_2$. Otherwise, it is called non-crossing. Similar formulas arise in the local laws for $\langle f_1(W)A_1 \dots f_k(W)A_k\rangle$ in [39, 37], showing that the underlying combinatorics also apply for functions of Wigner matrices.

1.3.2 Higher-Order Structures

Analogous to the notion of non-commutative probability space and freeness centered around the expectation, one can introduce second-order quantities tailored to the study of fluctuations. Note, however, that the definition of second-order freeness is sensitive to the symmetry class of the underlying random matrix ensemble (see [119]). For simplicity, we mainly focus on the complex Hermitian case below.

Second-order non-commutative probability space A triplet $(\mathcal{A}, \varphi_1, \varphi_2)$ that consists of a tracial non-commutative probability space $(\mathcal{A}, \varphi_1)$ and a functional $\varphi_2 : \mathcal{A} \times \mathcal{A} \to \mathbb{C}$ that is bilinear, tracial in both arguments, symmetric under the interchanging of its arguments, and satisfies $\varphi_2(a, 1_\mathcal{A}) = \varphi_2(1_\mathcal{A}, a) = 0$ for all $a \in \mathcal{A}$.

Second-order distribution The second-order distribution of an element $a \in \mathcal{A}$ is characterized in terms of its *fluctuation moments* $(\varphi_2(a^k, a^\ell))_{k,\ell}$. Mixed moments are defined analogously.

Second-order freeness Unital subalgebras $(\mathcal{A}_i)_{i\in I}$ are called *free of second order* if they are free and additionally

$$\varphi_2(a_1 \dots a_m, b_1, \dots, b_n) = \delta_{m,n} \sum_{k=0}^{n-1} \prod_{i=1}^{n} \varphi_1(a_i b_{k-i}) \tag{1.10}$$

whenever $(m, n) \neq (1, 1)$ and the products $a_1 \dots a_m$ resp. $b_1 \dots b_n$ are cyclically alternating, i.e., consecutive factors, as well as a_1 and a_m resp. b_1 and b_n belong to different subalgebras, respectively. The index of b_{k-i} on the right-hand side of (1.10) is interpreted modulo n. We may also consider the second-order freeness of individual elements by considering the subalgebras generated by them.

Note that the space $(\mathcal{A}, \varphi_1) = (\mathcal{M}_{N\times N}(L^{\infty-}(\Omega, \mathbb{P})), \mathbb{E}\langle\cdot\rangle)$ endowed with the covariance functional $\varphi_2(\cdot, \cdot) = \mathrm{Cov}(\mathrm{Tr}(\cdot), \mathrm{Tr}(\cdot))$ yields a second-order non-commutative probability space. When considering $N \times N$ matrices, we are again interested in the $N \to \infty$ limit, i.e., *asymptotic second-order freeness*. However, compared to the first-order setting, we observe a breaking of universality. While deterministic and GUE resp. GOE matrices are asymptotically free of second order in the respective sense (see [113] and [118]), the same does not hold for general Wigner matrices (see [107, Eq. (3)]). Here, the fourth cumulant of the entry distribution enters the computation of the fluctuation moments, resulting in an additional term that prevents (1.10) from holding in the $N \to \infty$ limit.

The above construction can be iterated to construct higher-order non-commutative probability spaces and a suitable notion of freeness (cf. [42]). While the definitions are expressed in terms of moments, one may also introduce an analog of cumulants by adapting their combinatorial definition from classical probability theory. For an element a in a non-commutative probability space $(\mathcal{A}, \varphi_1)$ with moments $(\alpha_k)_k = (\varphi_1(a^k))_k$, the sequence $(\kappa_k)_k$ of free cumulants is given by

$$\alpha_k = \sum_{\pi \in NCP(k)} \prod_{B\in\pi} \kappa_{|B|} \tag{1.11}$$

where $NCP(k)$ denotes the set of non-crossing partitions of the set $\{1, \dots, k\}$. Mixed cumulants are defined analogously and, by definition, vanish when evaluated for free elements. Relations similar to (1.11) also hold for second (see [111, Chap. 5]) and higher-order non-commutative probability spaces (see [26]) using a suitable generalization of non-crossing partitions.

1.4 Overview of Results

The remainder of the thesis consists of largely unmodified versions of the published or submitted papers, presented in thematic rather than chronological order, that were prepared during the author's PhD studies. Among them, [122] and [123] together form the main project on a multi-point functional central limit theorem for Wigner matrices while the remaining three publications [121, 58, 60] represent different smaller projects. To guide the reader, we now briefly summarize the main results of these works and place them in the general context established in Sections 1.1 to 1.3.

1.4.1 Multi-Point Functional Central Limit Theorem for Wigner Matrices

The articles [122, 123] establish a functional CLT that characterizes the fluctuations of the random variable

$$X = \mathrm{Tr}(f_1(W)A_1 \dots f_k(W)A_k) \tag{1.12}$$

around the expectation identified in [39, 37] and identifies the limiting covariance structure. Here, W is an $N \times N$ Hermitian Wigner matrix, $f_1, \dots, f_k$ are suitable (possibly N-dependent) regular functions and $A_1, \dots, A_k \in \mathbb{C}^{N\times N}$ are bounded deterministic matrices. The result generalizes the main result of [40] to general $k \geq 2$.

In [122], we show that the fluctuations of X around its expectation are asymptotically Gaussian and give explicit error bounds in terms of the scaling of $f_1, \dots, f_k$ and the number of traceless matrices among $A_1, \dots, A_k$. The result is first established for the case $f_j(W) = (W - z_j)^{-1}$ using the resolvent method. However, compared to the statistics involving only one function in [88, 40], the limiting covariance structure is now obtained through a recursion. The solution of this recursion is supplied in [123] using a combinatorial ansatz based on graphs closely related to a generalization of non-crossing partitions from free probability theory. Using this explicit description of the limiting covariance, the extension to Sobolev functions $f_1, \dots, f_k$ follows from the Helffer-Sjöstrand formula. The CLT shown in [122, 123] covers the macroscopic and the bulk regime of the mesoscopic scale for general Wigner matrices, with the two papers highlighting different aspects of the theorem and its proof.

Apart from the result for general test functions, some special cases are of independent interest. In [122], we consider the case $f_j(W) = \mathrm{e}^{\mathrm{i}tW}$, which links the statement of the multi-point functional CLT to the fluctuations in the thermalization problems considered in [39, 37]. Moreover, we discuss the case of $f_1, \dots, f_k$ being polynomials in [123], as the resulting random variable X in (1.12) naturally arises in the context of free probability theory. Here, the covariance formula in the multi-point functional CLT simplifies to reproduce known results from [111, 118, 107].

1.4.2 Prethermalization for Deformed Wigner Matrices

In [60], we show that, under suitable assumptions on the initial Hamiltonian H_0, the perturbed system introduced in (1.9) exhibits prethermalization. We further establish that the formulas describing the relaxation of the system in the physics literature (see, e.g., [47, 48, 115] and references therein) are generally valid for short ($t \ll \lambda^{-2}$) and long kinetic times ($t \gg \lambda^{-2}$) and can be extended to the intermediate regimes under a regularity condition on the quadratic forms $\langle \mathbf{u}_j, A\mathbf{u}_j \rangle$ of overlaps with the eigenvectors $\mathbf{u}_j$

of H_0. The primary technical input is a suitable global law for the deformed Wigner matrix H_λ from which the desired formulas are then obtained by extracting the leading term of the deterministic approximation in an appropriate limit.

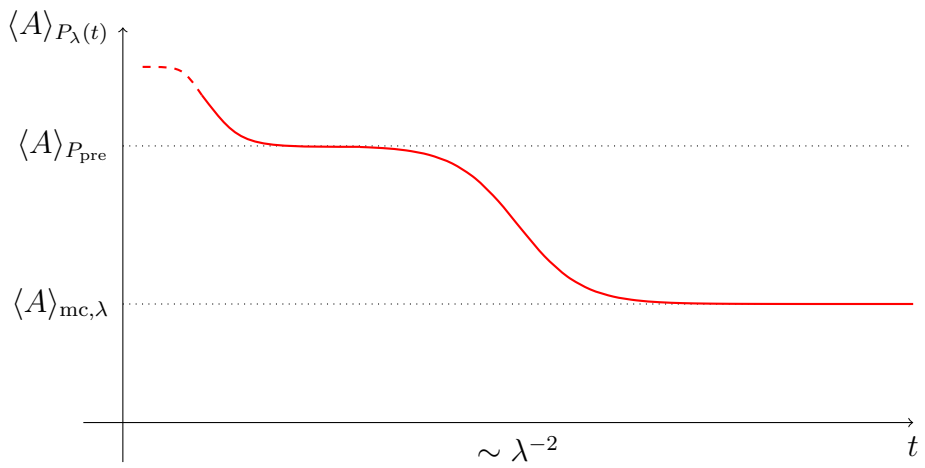

Fig. 1.2. Schematic graph of the prethermalization process. As $\langle A \rangle_{\mathrm{mc},\lambda}$ is by definition different from $\langle A \rangle_{\rho_{\mathrm{pre}}}$, the relaxation of $\langle A \rangle_{\rho_\lambda(t)}$ towards $\langle A \rangle_{\mathrm{mc},\lambda}$ happens via an intermediate prethermal value $\langle A \rangle_{\rho_{\mathrm{pre}}}$ in two steps, whose time scales are separated.

1.4.3 Operator Norm of a Hermitian Random Matrix with Correlated Entries

In [121], we consider a Hermitian random matrix H with a polynomially decaying metric correlation structure similar to the model in [69], but with a slower, merely summable correlation decay. Using a generalization of Wigner's moment method, we show that the operator norm of H is stochastically dominated by a constant.

1.4.4 Dynamics of a Rank-One Multiplicative Perturbation of a Unitary Matrix

In [58], we consider the UA model (1.8) for $t \in [-1, 1]$ and a general class of unitary random matrices U. By interpreting t as a time parameter, we obtain a dynamical description of the model, allowing us to interpret the eigenvalues as random trajectories in the unit disc. The results of the paper are twofold. The first main result identifies a random function $\mathcal{W}$, which distinguishes whether a complex number z lies in the spectrum of the UA model at any given time t. This function constitutes the main tool to describe the trajectories, showing that they do not intersect with probability one and can be described by a system of ordinary differential equations assuming only the (almost sure) simplicity of the spectrum of the underlying unitary matrix U.

In the second main result, we identify the critical time scale for the emergence of the outlier induced by the rank-one perturbation and show that it is well-separated from the typical eigenvalues at all sub-critical timescales. The proof is based on an application of Rouché's theorem, allowing us to identify a flow of deterministic domains that bound the spectrum with high probability. Contrary to the analysis of the weakly non-Hermitian model (1.7) in [57], however, the symmetry of the UA model does not allow to compare $\mathcal{W}$ to a deterministic function in this part of the proof. Instead, we retain some randomness

by working with the affine part of $\mathcal{W}$, replacing the local law used in [57] with a suitable regularity assumption. The results apply to various unitary random various unitary random matrices, including the CUE in the original formulation of the model.

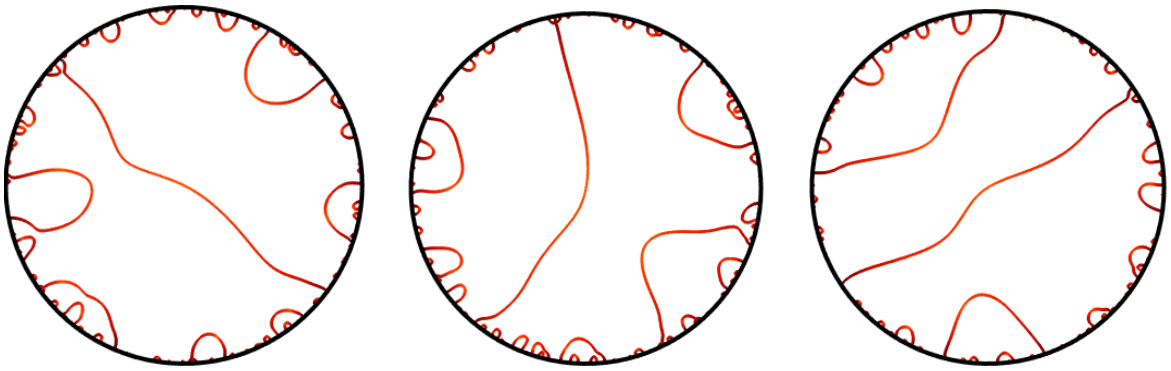

Fig. 1.3. Trajectories of the eigenvalues of the UA model of size 100×100. The evolution of the time parameter $t \in (-1, 1)$ is represented by shades of red.

2 Multi-Point Functional Central Limit Theorem for Wigner Matrices

Consider the random variable $\mathrm{Tr}(f_1(W)A_1 \dots f_k(W)A_k)$ *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 [40] to products of arbitrary length* $k \geq 2$*. As an application, we consider the fluctuation of* $\mathrm{Tr}(\mathrm{e}^{\mathrm{i}tW}A_1\mathrm{e}^{-\mathrm{i}tW}A_2)/N$ *around its thermal value* $\mathrm{Tr}(A_1)\,\mathrm{Tr}(A_2)/N^2$ *when* t *is large and give an explicit formula for the variance.*

Published as J. Reker. Multi-point functional central limit theorem for Wigner matrices. *Preprint, arXiv.2307.11028*, 2023.

2.1 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 $\mathrm{Tr}\, f(W) = \sum_{j=1}^N f(\lambda_j)$ with a regular test function $f : \mathbb{R} \to \mathbb{R}$ have a variance of order one (see [95]) and, in fact, satisfy a central limit theorem with a Gaussian limit, as shown, e.g., in [96] for the Wigner case and in [93] for invariant ensembles, see also [132, 133]. 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 emphasize that this question is well-studied for Wigner matrices, see, e.g., [85, 12, 104, 128, 135, 17, 99] and that recent work by Diaz and Mingo [53] 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

$$f(x) = g(N^\gamma(x - E)) \tag{2.1}$$

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 [88, 87, 89], 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 [91, 101], generalized Wigner [102], and Wigner-type [125] matrices, sample covariance matrices [11, 101], Haar distributed random matrices on the classical compact groups [130, 131, 134], β-ensembles [27, 129, 21, 20, 82, 30], free sums [18], and non-Hermitian random matrices [35, 61]. See also [40] 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

$$\operatorname{Tr}[f(W)A] - \mathbb{E}\operatorname{Tr}[f(W)A] = \sum_{j=1}^N f(\lambda_j)\langle \mathbf{u}_j, A\mathbf{u}_j\rangle - \mathbb{E}[\dots], \tag{2.2}$$

the normalized eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_N$ of W enter into the problem. In the Wigner case, Lytova [103] obtained a CLT for (2.2) 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 [40] extended these results to all mesoscopic scales and further established that decomposing the matrix A in (2.2) according to

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

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 $\mathring{A} = A - \langle A\rangle \mathrm{Id}$, the traceless part of A, respectively. Moreover, the results in [40] 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 (2.2) 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 (2.2) 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

$$Y := \langle f_1(W)A_1 f_2(W)A_2 \dots f_k(W)A_k\rangle - \mathbb{E}\langle f_1(W)A_1 f_2(W)A_2 \dots f_k(W)A_k\rangle, \tag{2.3}$$

show that Y satisfies a CLT with a Gaussian limit and give the limiting covariance structure as well as explicit error estimates. This generalizes [40, Thm. 2.4] to arbitrary $k \geq 1$. We refer to the result as a *multi-point functional CLT*. Similar to the results in [40], 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 (2.1). 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 (2.3) are asymptotically independent whenever the total number of traceless matrices involved is odd.

A key ingredient for studying the fluctuation of (2.3) 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. [37, 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 [39, 37]. 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 [39] 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., [111, 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 [39, 37] 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 [123] to show that the parallels to free probability identified in [39] 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 [111, Ch. 5] and [42] 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 [107]. To avoid introducing additional notation, we work with the recursive definitions in the present paper and only refer to the formulas in [123] for explicit computations and examples.

Lastly, as an application of the functional CLT, we consider the special case $f_j(x) = \mathrm{e}^{\mathrm{i}t_j x}$ 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

$$A(t) := \mathrm{e}^{\mathrm{i}tW} A \mathrm{e}^{-\mathrm{i}tW}$$

describes the Heisenberg time evolution of A. In this context, applying the CLT for the linear statistics (2.3) 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 [40, Thm. 2.3]) due to their connection to the fluctuations in the eigenstate thermalization hypothesis (see [51]) which is referred to as quantum unique ergodicity in mathematics (see [126], further references can be found in [36]). For $k \geq 2$, the statistics in (2.3) 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

$$\langle A_1(t)A_2\rangle = \langle A_1 A_2\rangle \frac{J_1(2t)^2}{t^2} + \frac{\xi(t)}{N} + \mathcal{O}\Big(\frac{N^\varepsilon}{N^{3/2}}\Big) \tag{2.4}$$

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 (2.4) was established in the recent paper [39] 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 $N \to \infty$ 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 2.1.1. We then briefly recall the optimal multi-resolvent local law [37, Thm. 2.5], which constitutes one of the key tools for the analysis. The main results of the paper are then given in Section 2.2. 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 2.2.4). Considering the fluctuations of the statistics in (2.3), we then establish a CLT and give an explicit formula for the limiting covariance (Theorem 2.2.7, Corollary 2.2.9). This is followed by a discussion of the result, including the asymptotics in the mesoscopic regime (Theorem 2.2.10), sufficient conditions for two modes to be asymptotically independent (Corollary 2.2.11) as well as the case of multiple independent Wigner matrices. We conclude Section 2.2 by applying the functional CLT to thermalization problems. In Section 2.3, 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 2.3.1), 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 2.3.6). The role of the limiting covariance in the theorem is played by a recursively defined set function $\mathfrak{m}_2[\cdot|\cdot]$ (Definition 2.3.4). Lastly, the proofs are given in Section 2.4. To keep the presentation concise, some routine calculations are deferred to the appendix.

Acknowledgements: I am very grateful to László Erdős for suggesting the topic and many valuable discussions during my work on the project. Partially supported by ERC Advanced Grant "RMTBeyond" No. 101020331.

2.1.1 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 $\overline{z}$. Moreover, $\|\cdot\|$ denotes the operator norm, $\mathrm{Tr}(\cdot)$ is the usual trace and $\langle\cdot\rangle = N^{-1}\,\mathrm{Tr}(\cdot)$. We further denote the covariance of two complex random variables Y_1, Y_2 by $\mathrm{Cov}(Y_1, Y_2)$ and follow the convention

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

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 $\langle a, b]$ or $[a, b\rangle$ 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 $\mathring{A} := A - \langle A\rangle\mathrm{Id}$ where Id denotes the identity matrix. Further, $\mathbf{a} := \mathrm{diag}A$ 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 $G_{\langle a,b]}$ and $G_{[a,b\rangle}$ 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 [39, 37]. 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

$$T_{[k]} := T_1 \dots T_k = G_1 A_1 \dots G_k A_k, \quad T_{[a,b]} := T_a T_{a+1} \dots T_b. \tag{2.5}$$

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 (2.5) *resolvent chain* of length k.

Throughout the paper, we assume W to be an $N \times N$ complex[1] Wigner matrix satisfying the following assumptions.

Assumption 2.1.1. *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

$$\kappa_4 := \mathbb{E}|\chi_{od}|^4 - 2 \tag{2.6}$$

for the normalized fourth cumulant of the off-diagonal entries. Note that the notation matches [40], 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 [40] is studied for macroscopic scales in the companion paper [123], and the necessary modifications for an extension to mesoscopic scales are sketched.

The eigenvalue density profile of W is described by the semicircle law

$$\rho_{sc}(x) := \frac{\sqrt{4 - x^2}}{2\pi} \mathbb{1}_{[-2,2]}(x) \tag{2.7}$$

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

$$m(z) := \int \frac{\rho_{sc}(x)}{x - z} \mathrm{d}x, \quad z \in \mathbb{C} \setminus \mathbb{R}. \tag{2.8}$$

[1] The 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

$$-\frac{1}{m(z)} = m(z) + z, \quad \Im z \Im(m(z)) > 0 \tag{2.9}$$

and that its derivative satisfies

$$m'(z) = \frac{m(z)^2}{1 - m(z)^2}. \tag{2.10}$$

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

$$q_{i,j} = \frac{m_i m_j}{1 - m_i m_j}, \tag{2.11}$$

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

$$m[z_1, \dots, z_k] := \frac{m[z_2, \dots, z_k] - m[z_1, \dots, z_{k-1}]}{z_k - z_1} \tag{2.12}$$

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

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

where $m^{(k-1)}$ is the $(k-1)$-th derivative of the function m in (2.8). 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 (2.11) coincides with $m[i, j]$.

2.1.2 Preliminaries: Multi-Resolvent Local Laws

Before considering the fluctuations, we briefly recall the optimal multi-resolvent local law [37, 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 2.1.2 (Stochastic domination)**.** *Let*

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

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*

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

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 function[2] $M_{[k]} = M_{[1,k]}$ through the recursion

$$M_{[k]} = m_1\Big(A_1 M_{[2,k]} + q_{1,k}\langle A_1 M_{[2,k]}\rangle + \sum_{j=2}^{k-1}\langle M_{[1,j]}\rangle\Big(M_{[j,k]} + q_{1,k}\langle M_{[j,k]}\rangle\Big)\Big) \tag{2.13}$$

with initial condition $M_\emptyset = 0$. We remark that an explicit (non-recursive) formula for $M_{[k]}$ was derived in [39, Thm. 2.6], however, we will not use it in the present paper. Analogously to (2.13), 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

$$M_S = M_{(s_1,\dots,s_k)}. \tag{2.14}$$

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

Theorem 2.1.3 (Multi-resolvent local law, [37, Thm. 2.5])**.** *Fix $\zeta > 0$ and $k \in \mathbb{N}$. Let $z_1, \dots, z_k \in \mathbb{C} \setminus \mathbb{R}$ with $\max_j |z_j| \le N^{100}$ and $d := \min_j \operatorname{dist}(z_j, [-2,2])$, 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 $\eta_* := \min_j |\Im z_j| \ge N^{-1+\zeta}$. Recalling that $T_j = G_j A_j$, we have the averaged local law*

$$\langle T_{[1,k]}\rangle = \langle M_{[k]} A_k\rangle + \mathcal{O}_\prec\Big(\frac{1}{N\eta_*\ \eta_*^{k-a/2-1}}\Big), \tag{2.15}$$

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

$$\langle \mathbf{x}, T_{[1,k\rangle}G_k\mathbf{y}\rangle = \langle \mathbf{x}, M_{[k]}\mathbf{y}\rangle + \mathcal{O}_\prec\Big(\frac{1}{\sqrt{N\eta_*}\ \eta_*^{k-a/2-1}}\Big). \tag{2.16}$$

As we frequently encounter $\langle M_{[k]}A_k\rangle$ in the following sections, we introduce the notation[3]

$$\mathfrak{m}_1[T_1, \dots, T_k] = \mathfrak{m}_1[z_1, A_1, \dots, z_k, A_k] := \langle M_{[k]}A_k\rangle \tag{2.17}$$

and remark that the function $\mathfrak{m}_1[\cdot]$ satisfies a recursion similar to (2.13). The arguments in the notation $\mathfrak{m}_1[T_1, \dots, T_k]$ indicate the deterministic approximation of $\langle T_1 \dots T_k\rangle$. Whenever $A_1 = \dots = A_k = \mathrm{Id}$, it follows that the deterministic approximation is given by the iterated divided differences, i.e.,

$$\mathfrak{m}_1[G_1, \dots, G_k] = m[1, \dots, k], \tag{2.18}$$

which can be seen from the resolvent identity

$$G_j G_{j-1} = \frac{G_j - G_{j-1}}{z_j - z_{j-1}} \tag{2.19}$$

and the averaged local law (2.15). Note that (2.15) and (2.16) 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 (2.14).

We further note the following a priori bounds for $m[\cdot]$, $\mathfrak{m}_1[\cdot]$, and $M_{[\cdot]}$ (cf. Lemma 2.4 and Appendix A of [37]).

[2]Note 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.

[3]Note that $\mathfrak{m}_1[T_1, \dots, T_n]$ 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 2.1.4. *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 $d := \min_j \operatorname{dist}(z_j, [-2,2]) \leq 1$. Then,*

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

where a' denotes the number of traceless matrices among $A_1, \dots, A_{k-1}$ and $\lceil x \rceil$ denotes the upper integrer part of $x \in \mathbb{R}$. Generically, the above bounds are sharp[4] when not all $\Im z_j$ have the same sign.

Theorem 2.1.3 together with the optimality of the bounds in Lemma 2.1.4 asserts that the deterministic $M_{[k]}$ is indeed the leading order approximation of $T_{[1,k\rangle}G_k$. In particular, the error terms in (2.15) and (2.16) are smaller than the natural upper bound on their leading term by a factor of $(N\eta_*)^{-1}$ and $(N\eta_*)^{-1/2}$, respectively.

2.2 Main Results

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

$$Y_\alpha^{(k,a)} := \langle f_1(W)A_1 \dots f_k(W)A_k\rangle - \mathbb{E}\langle f_1(W)A_1 \dots f_k(W)A_k\rangle, \tag{2.20}$$

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 2.2.1 (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

$$f_j(x) := g_j(N^{\gamma_j}(x - E_j)). \tag{2.21}$$

Note that Assumption 2.2.1 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

[4]The 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 [37].

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 2.1.3 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. Extending the results in this section to complex-valued $g_1, \dots, g_k$ only requires minor modifications to the proofs in Section 2.4. Moreover, as one can decompose any matrix A_j in Y_α as $A_j = \langle A_j \rangle \mathrm{Id} + \mathring{A}_j$, 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 2.2.1. We also 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

$$X_\alpha^{(k,a)} := \langle T_{[1,k]} \rangle - \mathbb{E}\langle T_{[1,k]} \rangle = \langle G_1 A_1 \dots G_k A_k \rangle - \mathbb{E}\langle G_1 A_1 \dots G_k A_k \rangle, \tag{2.22}$$

as a special case of (2.20). By a suitable functional calculus (cf. [49]), information on (2.22) carries over to the general statistics (2.20), thus yielding a key tool for the proof of our main results. We, therefore, consider the analog of the results in Sections 2.2.1 and 2.2.2 for the resolvent case separately in Section 2.3. Throughout the paper, we write

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

for the multi-index in (2.22) 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 (2.21), 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 2.2.1 for the spectral parameters now reads as follows.

Assumption 2.2.2 (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 $z_j = E_j + \mathrm{i}\eta_j$ 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.

2.2.1 The $1/N$ 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 $\langle T_{[1,k]}\rangle - \mathfrak{m}_1[T_1,\dots,T_k]$, i.e., it relates X_α in (2.22) to the bounds in the local law (2.15). The proof of Lemma 2.2.3 is carried out in Section 2.4.2.

Lemma 2.2.3. *Let $k \in \mathbb{N}$, W be a Wigner matrix satisfying Assumption 2.1.1, and fix spectral parameters $z_1,\dots,z_k$ satisfying Assumption 2.2.2 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 2.3.1 below) such that*

$$\mathbb{E}\langle T_1 \dots T_k\rangle = \mathfrak{m}_1[T_1,\dots,T_k] + \frac{\kappa_4}{N}\mathcal{E}[T_1,\dots,T_k] + \mathcal{O}\Big(\frac{N^\varepsilon}{N\sqrt{N\eta_*}\,\eta_*^{k-a/2}}\Big) \tag{2.23}$$

with $\mathfrak{m}_1[\cdot]$ as in (2.17), κ_4 as in (2.6), and $\eta_* := \min_j |\Im z_j|$.

We remark that (cf. Lemma 2.3.2 below)

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

i.e., the error term in (2.23) is indeed smaller than the deterministic leading term. A discussion of the properties of $\mathcal{E}[\cdot]$ is included in Section 2.3.1 below. We now give the expansion of $\mathbb{E}\langle F_{[1,k]}\rangle$.

Theorem 2.2.4. *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 2.1.1 and let $f_1,\dots,f_k$ be test functions satisfying Assumption 2.2.1 with $p = k - \lfloor a/2\rfloor + 1$. Then, for any $\varepsilon > 0$, we have the expansion*

$$\begin{aligned}\mathbb{E}\langle F_{[1,k]}\rangle = &\int_{\mathbb{R}^k}\int_{[0,10]^k}\Big[\prod_{j=1}^k(\partial_{\overline{z}}(f_j)_{\mathbb{C},p})(z_j)\Big]\mathfrak{m}_1[G(z_1)A_1,\dots,G(z_k)A_k]\mathrm{d}\eta_{[k]}\mathrm{d}x_{[k]}\\ &+\frac{\kappa_4}{N\pi^k}\int_{\mathbb{R}^k}\int_{[0,10]^k}\Big[\prod_{j=1}^k(\partial_{\overline{z}}(f_j)_{\mathbb{C},p})(z_j)\Big]\mathcal{E}[G(z_1)A_1,\dots,G(z_k)A_k]\mathrm{d}\eta_{[k]}\mathrm{d}x_{[k]}\\ &+\mathcal{O}\Big(\frac{N^\varepsilon \max_j\|f_j\|_{H^p}}{N^{3/2}}\Big)\end{aligned} \tag{2.24}$$

where we write $z_j = x_j + \mathrm{i}\eta_j$, $\mathrm{d}x_{[k]} = \mathrm{d}x_1\mathrm{d}x_2\dots,\mathrm{d}x_k$ as well as $\mathrm{d}\eta_{[k]} = \mathrm{d}\eta_1\mathrm{d}\eta_2\dots\mathrm{d}\eta_k$, and $(f_j)_{\mathbb{C},p}$ denotes the almost analytic extension of f_j of order p.

Theorem 2.2.4 follows from Lemma 2.2.3 and the Helffer-Sjöstrand formula (see [49]). As a similar argument will be used for the more involved proof of the multi-point functional CLT in Theorem 2.2.7, we omit the details here.

It follows from (93) in [40] that $\mathcal{E}[T_1]$ is given by

$$\mathcal{E}[T_1] = \langle A_1\rangle\frac{m_1^5}{1-m_1^2} = \langle A_1\rangle m_1' m_1^3. \tag{2.25}$$

Hence, computing the second integral in (2.24) shows that the $1/N$ term $\mathcal{E}[f_1(W)A_1]$ of $\mathbb{E}\langle f_1(W)A_1\rangle$ is

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

where ρ_{sc} denotes the density of the semicircle law in (2.7). We remark that this formula was already included in [40, Thm. 2.4]. Theorem 2.2.4 hence generalizes Equation (21) in [40] to arbitrary $k \geq 1$ in the setting of Assumptions 2.1.1 and 2.2.1.

2.2.2 Statement of the Multi-Point Functional CLT

We now state our main result, the multi-point functional CLT for the statistics Y_α in (2.20). To give the limiting covariance structure explicitly, we introduce a set function $\mathfrak{m}_2[\cdot|\cdot]$ to play the role of the deterministic approximation of the (appropriately scaled) covariance[5] 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 $T_{[1,k\rangle}G_k$ (see Theorem 2.1.3 as well as (2.17)). Recall that we use (z_j, A_j) and T_j interchangeably. In particular, we may write

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

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

Lemma 2.2.5. *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 2.1.1. Pick two sets of spectral parameters $z_1,\ldots,z_k$ and $z_{k+1},\ldots,z_{k+\ell}$ that either both satisfy Case 1 or both satisfy Case 2 of Assumption 2.2.2, and denote $\eta_* = \min_j|\Im z_j|$. Moreover, pick deterministic matrices $A_1,\ldots,A_{k+\ell}$ with $\|A_j\| \lesssim 1$ such that a matrices among $A_1,\ldots,A_k$ and b matrices among $A_{k+1},\ldots,A_{k+\ell}$ are traceless. Then there exists a set function $\mathfrak{m}_2[\cdot|\cdot]$ (defined recursively in Definition 2.3.4 below) such that*

$$N^2\mathbb{E}X_\alpha^{(k,a)}X_\beta^{(\ell,b)} = \mathfrak{m}_2[\alpha|\beta] + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\,\eta_*^{k-a/2}\eta_*^{\ell-b/2}}\Big) \tag{2.26}$$

for any $\varepsilon > 0$.

We remark that (cf. (2.56) below)

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

i.e., the error term in (2.26) is smaller than the deterministic leading term at least by a factor $(N\eta_*)^{-1/2}$. 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 2.3.2 below. Moreover, explicit (non-recursive) formulas for $\mathfrak{m}_2[\cdot|\cdot]$ are derived in the companion paper [123]. We further recall the following definition.

[5] Note 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.

Definition 2.2.6. *Consider two functions of the Wigner matrix W in Assumption 2.1.1, 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*

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

for any smal $\varepsilon > 0$, where the implicit constant in $\mathcal{O}(\cdot)$ only depends on the polynomial $\mathcal{P}$ and the constants in Assumption 2.1.1.

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

Theorem 2.2.7 (Multi-point functional CLT)**.** *Under the assumptions of Theorem 2.2.4 it holds that, for any $\varepsilon > 0$, the centered statistics* (2.20) *are approximately distributed (in the sense of moments) as*

$$NY_\alpha^{(k,a)} = \xi(\alpha) + \mathcal{O}\Big(\frac{N^\varepsilon \max_j \|f_j\|_{H^p}}{\sqrt{N}}\Big) \tag{2.27}$$

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

$$\begin{aligned}
&\mathbb{E}[\xi(\alpha)\xi(\beta)] \qquad (2.28)\\
&= \frac{1}{\pi^{k+\ell}} \int_{\mathbb{R}^k} \mathrm{d}x_{[k]} \int_{[0,10]^k} \mathrm{d}\eta_{[k]} \Big[\prod_{i=1}^{k} (\partial_{\bar z}(f_i)_{\mathbb{C},p})(z_i) \Big] \int_{\mathbb{R}^\ell} \mathrm{d}x_{[k+1,k+\ell]} \int_{[0,10]^\ell} \mathrm{d}\eta_{[k+1,k+\ell]} \\
&\quad \times \Big[\prod_{j=k+1}^{\ell} (\partial_{\bar z}(f_j)_{\mathbb{C},q})(z_j) \Big] \mathfrak{m}_2[G(z_1)A_1, \dots, G(z_k)A_k | G(z_{k+1})A_{k+1}, \dots, G(z_{k+\ell})A_{k+\ell}].
\end{aligned}$$

Here, $z_j = x_j + \mathrm{i}\eta_j$, $\mathrm{d}x_{[i,j]} = \mathrm{d}x_i \mathrm{d}x_{i+1} \dots, \mathrm{d}x_j$ as well as $\mathrm{d}\eta_{[i,j]} = \mathrm{d}\eta_i \mathrm{d}\eta_{i+1} \dots, \mathrm{d}\eta_j$ 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 2.2.1 with $q = \ell - \lfloor b/2 \rfloor + 1$. Recall that $\mathfrak{m}_2[\cdot|\cdot]$ was introduced in Lemma 2.2.5.

The key ingredient for the proof of Theorem 2.2.7 is the case of all f_j being resolvents, which we discuss in detail in Section 2.3.2 below. The full multi-point functional CLT is then obtained from the resolvent CLT using the Helffer-Sjöstrand formula (cf. [49]). We carry out the argument in Section 2.4.5.

2.2.3 Discussion of the Multi-Point Functional CLT for Mesoscopic Scales

In this section, we consider the mesoscopic regime of Theorem 2.2.7 (Case 2 of Assumption 2.2.1) in more detail. We start by noting that the set function $\mathfrak{m}_2[\cdot|\cdot]$ in Lemma 2.2.5 can be decomposed as

$$\mathfrak{m}_2[\cdot|\cdot] = \mathfrak{m}_{GUE}[\cdot|\cdot] + \kappa_4 \mathfrak{m}_\kappa[\cdot|\cdot], \tag{2.29}$$

with 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 (2.29) induces a similar decomposition for the limiting covariance in (2.28). 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 (2.28) 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 2.3.5 below. Recall that we may interpret the resolvent $G(z)$ as a function rescaled to scale $|\Im z|^{-1}$ around $\Re z$ to match (2.21).

Example 2.2.8. For $k = \ell = 1$, we have by (92) of [40] (or Definition 2.3.4 below) that

$$\begin{aligned}\mathfrak{m}_2[T_1|T_2] := \langle A_1 A_2\rangle \frac{m_1^2 m_2^2}{(1-m_1m_2)} + \langle \mathbf{a}_1\mathbf{a}_2\rangle \cdot \kappa_4 m_1^3 m_2^3 \\ + \langle A_1\rangle\langle A_2\rangle\Big(\frac{m_1' m_2'}{(1-m_1m_2)^2} - \frac{m_1^2 m_2^2}{(1-m_1m_2)} + 2\kappa_4 m_1 m_1' m_2 m_2' - \kappa_4 m_1^3 m_2^3\Big),\end{aligned} \tag{2.30}$$

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 explicit form

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

of the solution to (2.9) shows that $|(1-m_1m_2)^{-1}| \lesssim \eta_*^{-1}$. Hence, the function $\mathfrak{m}_2[T_1|T_2]$ in (2.30) satisfies the bounds

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

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

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

i.e., the two parts of $\mathfrak{m}_2[\cdot|\cdot]$ (cf. decomposition in (2.29)) 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 2.2.7 may be tedious to compute using the recursive definition of $\mathfrak{m}_2[\cdot|\cdot]$ alone. In the companion paper [123], we consider the recursion defining $\mathfrak{m}_2[\cdot|\cdot]$ in detail and derive explicit formulas. Combining [123, Thm. 2.4] with (2.28) 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 2.2.9 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 $A_1, \ldots, A_k$ and an expression in the test functions $f_1, \ldots, f_{k+\ell}$, respectively. Moreover, the combinatorics underlying the summation mirror those encountered in second-order free probability theory. The proof of Corollary 2.2.9 is given in Section 2.4.6.

To state the result, we denote by $NCP(k)$ the set of non-crossing partitions of $[k]$, by $\overrightarrow{NCP}(k,\ell)$ 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 $\pi \in \overrightarrow{NCP}(k,\ell)$, 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

$$h[S] = \sum_{\pi \in NCP(S)} \prod_{B\in\pi} h_\circ[B], \tag{2.31}$$

$$h[S_1|S_2] = \sum_{\pi \in \overrightarrow{NCP}(|S_1|,|S_2|)} \prod_{B\in\pi} h_\circ[B] + \sum_{\substack{\pi_1\times\pi_2 \in NCP(|S_1|)\times NCP(|S_2|), \\ U_1\in\pi_1, U_2\in\pi_2 \text{ marked}}} h_{\circ\circ}[U_1|U_2] \prod_{\substack{B\in\pi_1\setminus U_1 \\ \cup \pi_2\setminus U_2}} h_\circ[B]. \tag{2.32}$$

The full definitions and some illustrative examples are, e.g., given in [123, Sect. 1] or [111].

Corollary 2.2.9. *Consider the setup of Theorem 2.2.7 for $\kappa_4 = 0$. Then,*

$$\mathbb{E}\xi(\alpha)\xi(\beta) = \sum_{\pi \in \overrightarrow{NCP}(k,\ell)} \Big(\prod_{B\in K(\pi)} \Big\langle \prod_{j\in B} A_j \Big\rangle \Big) \prod_{B\in\pi} \Phi_{\pi,B}(f_j | j \in B) \tag{2.33}$$
$$+ \sum_{\substack{\pi_1\times\pi_2 \in NCP(k)\times NCP(\ell),\\ U_1\in\pi_1, U_2\in\pi_2 \text{ marked}}} \Big(\prod_{\substack{B_1\in K(\pi_1),\\ B_2\in K(\pi_2)}} \Big\langle \prod_{j\in B_1} A_j \Big\rangle \Big\langle \prod_{j\in B_2} A_j \Big\rangle \Big) \Phi_{\pi_1\times\pi_2, U_1\times U_2}(f_1,\dots,f_{k+\ell}).$$

The functions $\Phi_{\pi,B}$ and $\Phi_{\pi_1\times\pi_2,U_1\times U_2}$ in (2.33) are given by

$$\Phi_{\pi,B}(f_j | j\in B) := \mathrm{sc}_\circ[B], \tag{2.34}$$

where $\mathrm{sc}_\circ[\cdot]$ denotes the first-order free cumulant function associated with

$$\mathrm{sc}[i_1,\dots,i_n] := \int_{-2}^{2} \Big[\prod_{j=1}^{n} f_{i_j}(x) \Big] \rho_{sc}(x)\mathrm{d}x, \tag{2.35}$$

with ρ_{sc} as in (2.7), and

$$\Phi_{\pi_1\times\pi_2,U_1\times U_2}(f_1,\dots,f_{k+\ell}) := \mathrm{sc}_{\circ\circ}[U_1|U_2] \prod_{\substack{B_1\in\pi_1\setminus U_1,\\ B_2\in\pi_2\setminus U_2}} \mathrm{sc}_\circ[B_1]\mathrm{sc}_\circ[B_2]. \tag{2.36}$$

Here, $\mathrm{sc}_{\circ\circ}[\cdot|\cdot]$ denotes the second-order free cumulants associated with $\mathrm{sc}[\cdot]$ in (2.35) and

$$\mathrm{sc}[i_1,\dots,i_n|i_{n+1},\dots,i_{n+m}] := \frac{1}{2}\int_{-2}^{2}\int_{-2}^{2} \Big(\prod_{j=1}^{n} f_{i_j}(x) \Big)' \Big(\prod_{j=1}^{m} f_{i_{n+j}}(y) \Big)' u(x,y)\mathrm{d}x\mathrm{d}y, \tag{2.37}$$

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

$$u(x,y) := \frac{1}{4\pi^2} \ln \Big[\frac{(\sqrt{4-x^2}+\sqrt{4-y^2})^2(xy+4-\sqrt{4-x^2}\sqrt{4-y^2})}{(\sqrt{4-x^2}-\sqrt{4-y^2})^2(xy+4+\sqrt{4-x^2}\sqrt{4-y^2})} \Big]. \tag{2.38}$$

We remark that the structure of (2.33) resembles the formula in [107, 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 [123]. In particular, applying Theorem 2.2.7 and (2.33) for $f_1(x) = \cdots = f_{k+\ell}(x) = x$ reproduces the corresponding formulas in [107] (cf. [123, Cor. 2.11]).

Theorem 2.2.7 and Corollary 2.2.9 identify the limiting process $\xi(\alpha)$ in terms of the test functions $f_1,\dots,f_{k+\ell}$. Similar to [40, Sect. 2.3] in the case $k=\ell=1$, we can make use of the mesoscopic scaling (2.21) 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 $\mathrm{sc}[\cdot]$ and $\mathrm{sc}[\cdot|\cdot]$ 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

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

Theorem 2.2.10 (Bulk scaling asymptotics)**.** *Under the assumptions of Theorem 2.2.7, pick test functions $f_1,\dots,f_{k+\ell}$ that satisfy Assumption 2.2.1 with some $\delta,\gamma_j > 0$.*

(i) Whenever $\gamma_1 = \cdots = \gamma_{n+m} = \gamma$ 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

$$N^{\gamma}\mathrm{sc}[i_1, \dots, i_{n+m}] = \rho_{sc}(E_0)\Big\langle \prod_{j=1}^{n} g_{i_j}, \prod_{j=n+1}^{n+m} g_{i_j} \Big\rangle_{L^2} + \mathcal{O}\Big(N^{-\gamma}\Big),$$
$$\mathrm{sc}[i_1, \dots, i_n | i_{n+1}, \dots, i_{n+m}] = \frac{1}{4\pi^2}\Big\langle \prod_{j=1}^{n} g_{i_j}, \prod_{j=n+1}^{n+m} g_{i_j} \Big\rangle_{\dot{H}^{1/2}} + \mathcal{O}\Big(N^{-\gamma}\Big).$$

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

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

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

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

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

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 $|\mathrm{supp} g_j|$.

The proof of Theorem 2.2.10 follows from the definition of $\mathrm{sc}[\cdot]$ in (2.35) and (careful) integration by parts of (2.37). 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 2.3.6 below. Next, we discuss conditions under which two modes Y_α and Y_β in Theorem 2.2.7 are asymptotically independent. The case $k = \ell = 1$ of Corollary 2.2.11 was already discussed in [40, Thm. 2.13].

Corollary 2.2.11 (Independent modes)**.** *Under the assumptions of Theorem 2.2.7, 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 2.2.7.*

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

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

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

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

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

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

where $p = k - \lfloor a/2\rfloor + 1$ and $q = \ell - \lfloor b/2\rfloor + 1$, i.e., the leading deterministic term in (2.28) has the same size as the error term in Theorem 2.2.7.

The proof of Corollary 2.2.11 is immediate from Theorem 2.2.10 as well as the remark on independent modes in the resolvent CLT below Theorem 2.3.6.

Remark (Multiple independent Wigner matrices). The high-probability sense of Theorem 2.2.7 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 [39, 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 2.2.1 with $p=3$. Then,

$$\begin{aligned}
&N^2\mathbb{E}\Big[\Big(\langle f_1(W)A_1f_2(W')A_2\rangle - \mathbb{E}\langle f_1(W)A_2f_2(W')A_2\rangle\Big)\\
&\quad\times\Big(\langle f_3(W)A_3f_4(W')A_4\rangle - \mathbb{E}\langle f_3(W)A_3f_4(W')A_4\rangle\Big)\Big]\\
&= \mathrm{sc}_{\circ\circ}[1|3]\langle A_1f_2(W')A_2\rangle\langle A_3f_4(W')A_4\rangle + \mathrm{sc}_{\circ}[1,3]\langle A_1f_2(W')A_2A_3f_4(W')A_4\rangle\\
&\quad+ N^2\mathbb{E}\Big[\Big(\langle A_1f_2(W')A_2\rangle - \mathbb{E}\langle A_2f_2(W')A_2\rangle\Big)\Big(\langle A_3f_4(W')A_4\rangle - \mathbb{E}\langle A_3f_4(W')A_4\rangle\Big)\Big]\\
&\quad\times\mathrm{sc}_{\circ}[1]\mathrm{sc}_{\circ}[3] + \mathcal{O}\Big(\frac{N^{\varepsilon}\max\{\|f_1\|_{H^3},\|f_3\|_{H^3}\}}{\sqrt{N}}\Big)\\
&= \langle A_1A_2\rangle\langle A_3A_4\rangle(\mathrm{sc}_{\circ\circ}[1|3]\mathrm{sc}_{\circ}[2]\mathrm{sc}_{\circ}[4] + \mathrm{sc}_{\circ\circ}[2|4]\mathrm{sc}_{\circ}[1]\mathrm{sc}_{\circ}[3])\\
&\quad+\langle A_1A_4\rangle\langle A_2A_3\rangle\mathrm{sc}_{\circ}[1,3]\mathrm{sc}_{\circ}[2,4] + \langle A_1A_2A_3A_4\rangle\mathrm{sc}_{\circ}[1,3]\mathrm{sc}_{\circ}[2]\mathrm{sc}_{\circ}[4]\\
&\quad+\langle A_2A_1A_4A_3\rangle\mathrm{sc}_{\circ}[1]\mathrm{sc}_{\circ}[2,4]\mathrm{sc}_{\circ}[3] + \mathcal{O}\Big(\frac{N^{\varepsilon}\max_j\|f_j\|_{H^3}}{\sqrt{N}}\Big).
\end{aligned}$$

In the first step, we conditioned on W' and applied Corollary 2.2.9, treating W', and hence $f_j(W')$, as deterministic. After computing the leading term, Corollary 2.2.9 is applied again for W'. Lastly, the remaining terms are identified using the local law [37, Cor. 2.7], which yields a total of five summands. In contrast, if $W = W'$, all terms on the right-hand side of (2.28) may contribute. For $k=\ell=2$, this yields 27 terms in total (cf. [123, 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 resp. *non-mixing* marked partitions for n colors, which was also established in [107] for the special case $f_1(x) = \cdots = f_k(x) = x$ and general Wigner matrices.

2.2.4 Application to Thermalization Problems

We now specialize Theorem 2.2.7 to the functions $f_j(x) = \mathrm{e}^{\mathrm{i}t_jx}$ with (N-independent) numbers $t_j\in\mathbb{R}$. Recall that

$$A(t) := \mathrm{e}^{\mathrm{i}tW}A\mathrm{e}^{-\mathrm{i}tW} \tag{2.39}$$

describes the Heisenberg time evolution of an observable A and that it follows from [39, 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 2.2.7, 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 $t \to \infty$ limit after letting $N \to \infty$. As the randomness in $\langle A_1(t)A_2\rangle$ cancels out if either $A_1 = \mathrm{Id}$ or $A_2 = \mathrm{Id}$, 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 2.2.12. *Let $\kappa_4 = 0$, $\langle A_1\rangle = \langle A_2\rangle = 0$, and $\|A_1\|, \|A_2\| \lesssim 1$. Then,*

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

where $A_1(t)$ is as defined in (2.39), J_1 is a Bessel function of the first kind, and $\xi(t)$ is a centered Gaussian random variable. In the $t \to \infty$ limit, the variance of $\xi(t)$ satsifies the asymptotics

$$\mathrm{Var}[\xi(t)] = \langle|A_1|^2\rangle\langle|A_2|^2\rangle + \mathcal{O}\Big(\frac{1}{t^2}\Big) \tag{2.40}$$

and we further obtain

$$\mathbb{E}\xi(t_1)\overline{\xi(t_2)} = \langle|A_1|^2\rangle\langle|A_2|^2\rangle\Big(\frac{J_1(2(t_1-t_2))}{t_1-t_2}\Big)^2 + \mathcal{O}\Big(\frac{1}{(\min\{t_1,t_2\})^2}\Big). \tag{2.41}$$

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

Corollary 2.2.12 follows directly from Corollary 2.2.9 by specifying the test functions. We carry out the details in Section 2.4.7.

Remark. The $\mathcal{O}(t^{-2})$ term in (2.40) can be given more explicitly, as the limiting variance can be read off from Corollary 2.2.9 or [104, Sect. 2] (since t is fixed). We obtain

$$\begin{aligned}
\mathrm{Var}[\xi(t)] = {}& \langle|A_1|^2\rangle\langle|A_2|^2\rangle\Big(1-\frac{J_1(2t)^2}{t^2}\Big)^2 \\
&+ |\langle A_1A_2^*\rangle|^2\Big[\Big(\frac{J_1(4t)}{2t} - \frac{J_1(2t)^2}{t^2}\Big)^2 + 2\mathrm{sc}_{\circ\circ}[+,+]\frac{J_1(2t)^2}{t^2} - 2\mathrm{sc}_{\circ\circ}[+,-]\frac{J_1(2t)^2}{t^2}\Big] \\
&+ \Big(\langle A_1A_2A_1^*A_2^*\rangle + \langle A_2A_1A_2^*A_1^*\rangle\Big)\Big(\frac{J_1(4t)}{2t} - \frac{J_1(2t)^2}{t^2}\Big)\frac{J_1(2t)}{t^2} \\
&+ \Big(\langle|A_1A_2|^2\rangle + \langle|A_2A_1|^2\rangle\Big)\Big(1-\frac{J_1(2t)^2}{t^2}\Big)\frac{J_1(2t)^2}{t^2}
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{sc}_{\circ\circ}[+,+] &= \frac{1}{2\pi^2}\int_{-2}^2\int_{-2}^2\Big(\frac{\mathrm{e}^{\mathrm{i}tx}-\mathrm{e}^{\mathrm{i}ty}}{x-y}\Big)^2\frac{4-xy}{\sqrt{4-x^2}\sqrt{4-y^2}}\mathrm{d}x\mathrm{d}y - \frac{J_1(4t)}{2t} + \frac{J_1(2t)^2}{t^2}, \\
\mathrm{sc}_{\circ\circ}[+,-] &= \frac{1}{2\pi^2}\int_{-2}^2\int_{-2}^2\frac{1-\cos(t(x-y))}{(x-y)^2}\frac{4-xy}{\sqrt{4-x^2}\sqrt{4-y^2}}\mathrm{d}x\mathrm{d}y - 1 + \frac{J_1(2t)^2}{t^2}.
\end{aligned}$$

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 $|\langle A_1A_2^*\rangle|^2$ term is $\mathcal{O}(t^{-3})$ while the double integrals behave like t for $t \to \infty$. Hence the entire $|\langle A_1A_2^*\rangle|^2$ 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 $\mathcal{O}(t^{-6})$.

2.3 Central Limit Theorem for Resolvents

In this section, we supply the recursive definitions of the set functions $\mathcal{E}[\cdot]$ and $\mathfrak{m}_2[\cdot|\cdot]$ introduced in Lemmas 2.2.3 and 2.2.5, respectively, and study their properties. We further state the analog of Theorem 2.2.7 for the resolvent case, which constitutes the main ingredient for the proof of the multi-point functional CLT.

2.3.1 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 2.3.1. *Let $(T_1,\dots,T_k)$ be an ordered set of $\mathbb{C}^{N\times N}$ matrices of the form $T_j = G_jA_j$. 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)*

$$\begin{aligned}
&\mathcal{E}[T_1,\dots,T_k]\\
&= m_1\Big(\mathcal{E}[T_2,\dots,T_{k-1},T_kA_1] + q_{1,k}\mathcal{E}[T_2,\dots,T_{k-1},G_kA_1]\langle A_k\rangle\\
&\quad+\sum_{j=1}^{k-1}\mathcal{E}[T_1,\dots,T_{j-1},G_j]\Big(\mathfrak{m}_1[T_j,\dots,T_k] + q_{1,k}\mathfrak{m}_1[T_j,\dots,T_{k-1},G_k]\langle A_k\rangle\Big)\\
&\quad+\sum_{j=2}^{k}\mathfrak{m}_1[T_1,\dots,T_{j-1},G_j]\Big(\mathcal{E}[T_j,\dots,T_k] + q_{1,k}\mathcal{E}[T_j,\dots,T_{k-1},G_k]\langle A_k\rangle\Big)\\
&\quad+\sum_{1\le r\le s\le t\le k}\langle M_{[r]}\odot M_{[s,t]}\rangle\langle M_{[r,s]}\odot(M_{[t,k]}A_k)\rangle\\
&\quad+q_{1,k}\sum_{1\le r\le s\le t\le k}\langle M_{[r]}\odot M_{[s,t]}\rangle\langle M_{[r,s]}\odot M_{[t,k]}\rangle\langle A_k\rangle\Big)
\end{aligned}\tag{2.42}$$

and the initial condition $\mathcal{E}[\emptyset]=0$. Recall that $\odot$ denotes the Hadamard product, $M_{[\cdot]}$ was defined through the recursion in (2.13)*, $\mathfrak{m}_1[\cdot]$ was defined in* (2.17)*, and $q_{1,k} = \frac{m_1m_k}{1-m_1m_k}$.*

We remark that $M_{[k]}$ is diagonal whenever $A_1=\cdots=A_k=\mathrm{Id}$. In this case, the last two lines of (2.42) are readily evaluated and the recursion simplifies to

$$\begin{aligned}
\mathcal{E}[G_1,\dots,G_k] = \frac{m_1}{1-m_1m_k}\Big(&\mathcal{E}[G_2,\dots,G_k] + \sum_{j=1}^{k-1}\mathcal{E}[G_1,\dots,G_j]m[j,\dots,k]\\
&+\sum_{j=2}^{k}m[1,\dots,j]\mathcal{E}[G_j,\dots,G_k]\\
&+\sum_{1\le r\le s\le t\le k}m[1,\dots,r]m[r,\dots,s]m[s,\dots,t]m[t,\dots,k]\Big),
\end{aligned}$$

where $m[\cdot]$ denotes the iterated divided differences in (2.12). 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 2.4.1.

Lemma 2.3.2. *Under the assumptions of Lemma 2.2.3, let a among the matrices $A_1,\dots,A_k$ be traceless. Then,*

$$|\mathcal{E}[T_1,\dots,T_k]| \lesssim \frac{1}{\eta_*^{k-1-\lceil a/2\rceil}}.\tag{2.43}$$

Generically, the bound is sharp if not all $\Im z_j$ have the same sign.

Remark. Lemma 2.3.2 shows that the sub-leading term in (2.23) 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 (2.23) in the form

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

for the proof of the CLT in the resolvent case in Section 2.4.4, a more careful resolution of the error is not needed.

Note that applying (2.42) once yields the formula

$$\mathcal{E}[T_1] = \langle A_1\rangle \frac{m_1^5}{1-m_1^2} = \langle A_1\rangle m_1' m_1^3, \tag{2.44}$$

and we readily reobtain (93) of [40] by Lemma 2.2.3.

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 (2.23), to derive further identities for $\mathcal{E}[\cdot]$ (cf. the "meta argument" below [37, Lem. 4.1]). They are listed in Corollary 2.3.3 below and the proof is given in Appendix 2.A. In fact, these identities can also be proven from Definition 2.3.1 directly, however, using (2.23) allows for a shorter proof. Note that $\mathfrak{m}[\cdot]$ satisfies the same properties by [39, Lem. 5.4].

Corollary 2.3.3. *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 $A_k = \mathrm{Id}$, we have

$$\mathcal{E}[T_1,\dots,T_{k-1},G_k] = \frac{\mathcal{E}[T_2,\dots,T_{k-1},G_kA_1] - \mathcal{E}[T_1,\dots,T_{k-1}]}{z_k - z_1}. \tag{2.45}$$

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

$$\mathcal{E}[G_1,\dots,G_k] = \frac{\mathcal{E}[G_2,\dots,G_k] - \mathcal{E}[G_1,\dots,G_{k-1}]}{z_k - z_1} \tag{2.46}$$

and we have the closed formula

$$\mathcal{E}[G_1,\dots,G_k] = \sum_{j=1}^k \prod_{i\neq j} \frac{1}{z_i - z_j}\mathcal{E}[G_j] = \sum_{j=1}^k \prod_{i\neq j} \frac{m[i,j]}{m_i - m_j}\mathcal{E}[G_j] \tag{2.47}$$

with $m[\cdot]$ as in (2.18) 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.

2.3.2 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 (2.22), 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 [123].

Definition 2.3.4. *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

$$\mathfrak{m}_2[(T_i, i \in B_1)|\emptyset] = \mathfrak{m}_2[\emptyset|(T_j, j \in B_2)] = 0. \tag{2.48}$$

(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

$$\begin{aligned}
&\mathfrak{m}_2[T_1, \dots, T_k|T_{k+1}, \dots, T_{k+\ell}] \\
&= m_1\Bigg(\mathfrak{m}_2[T_2, \dots, T_{k-1}, G_k A_k A_1|T_{k+1}, \dots, T_{k+\ell}] \\
&\quad + q_{1,k}\mathfrak{m}_2[T_2, \dots, T_{k-1}, G_k A_1|T_{k+1}, \dots, T_{k+\ell}]\langle A_k\rangle \\
&\quad + \sum_{j=1}^{k-1} \mathfrak{m}_2[T_1, \dots, T_{j-1}, G_j|T_{k+1}, \dots, T_{k+\ell}]\Big(\mathfrak{m}_1[T_j, \dots, T_k] \\
&\quad + q_{1,k}\mathfrak{m}_1[T_j, \dots, T_{k-1}, G_k]\langle A_k\rangle\Big) \\
&\quad + \sum_{j=2}^{k} \mathfrak{m}_1[T_1, \dots, T_{j-1}, G_j]\Big(\mathfrak{m}_2[T_j, \dots, T_k|T_{k+1}, \dots, T_{k+\ell}] \\
&\quad + q_{1,k}\mathfrak{m}_2[T_j, \dots, T_{k-1}, G_k|T_{k+1}, \dots, T_{k+\ell}]\langle A_k\rangle\Big) + \mathfrak{s}_{GUE} + \mathfrak{s}_\kappa\Bigg)
\end{aligned} \tag{2.49}$$

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

$$\begin{aligned}
\mathfrak{s}_{GUE} := \sum_{j=1}^{\ell}\Big(&\mathfrak{m}_1[T_1, \dots, T_k, T_{k+j}, \dots, T_{k+j-1}, G_{k+j}] \\
&+ q_{1,k}\mathfrak{m}_1[T_1, \dots, T_{k-1}, G_k, T_{k+j}, \dots, T_{k+j-1}, G_{k+j}]\langle A_k\rangle\Big)
\end{aligned} \tag{2.50}$$

$$\begin{aligned}\mathfrak{s}_\kappa := \kappa_4 \sum_{r=1}^{k} \sum_{s=k+1}^{k+\ell} \Big(\sum_{t=k+1}^{s} \langle M_{[r]} \odot M_{(s,\dots,k+\ell,k+1,\dots,t)} \rangle \langle (M_{[r,k]} A_k) \odot M_{[t,s]} \rangle \\ + \sum_{t=s}^{k+\ell} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle (M_{[r,k]} A_k) \odot M_{(t,\dots,k+\ell,k+1,\dots,s)} \rangle \Big) \\ + \kappa_4 q_{1,k} \sum_{r=1}^{k} \sum_{s=k+1}^{k+\ell} \Big(\sum_{t=k+1}^{s} \langle M_{[r]} \odot M_{(s,\dots,k+\ell,k+1,\dots,t)} \rangle \langle M_{[r,k]} \odot M_{[t,s]} \rangle \\ + \sum_{t=s}^{k+\ell} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle M_{[r,k]} \odot M_{(t,\dots,k+\ell,k+1,\dots,s)} \rangle \Big) \langle A_k \rangle. \end{aligned} \tag{2.51}$$

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 (2.8)*, $M_{(\dots)}$ was introduced in* (2.14)*, and $\mathfrak{m}_1[\cdot]$ was defined in* (2.17)*.*

Remark. The special role of m_1 in (2.49) is a result of the identity (2.74) used for the proof of Lemma 2.2.5 in Section 2.4.3 below. Similar to the recursion for $\mathfrak{m}_1[\cdot]$ in [37, Lem. 4.1], it is possible to derive a version of (2.49) for every $j = 2, \dots, k$ that singles out the factor m_j instead of m_1 on the right-hand side, i.e., (2.49) 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 (2.29), where $\mathfrak{m}_{GUE}[\cdot|\cdot]$ satisfies (2.49) for $\kappa_4 = 0$, and $\kappa_4 \mathfrak{m}_\kappa[\cdot|\cdot]$ satisfies (2.49) without $\mathfrak{s}_{GUE}$. We remark that by [39, 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 (2.49), eventually making $\mathfrak{m}_2[\cdot|\cdot]$ a function of the same quantities. Similarly, we have the decomposition

$$\widetilde{m}[\cdot|\cdot] = \widetilde{m}_{GUE}[\cdot|\cdot] + \kappa_4 \widetilde{m}_\kappa[\cdot|\cdot], \tag{2.52}$$

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

$$\widetilde{m}[1, \dots, k | k+1, \dots, k+\ell] := \mathfrak{m}_2[G_1, \dots, G_k | G_{k+1}, \dots, G_{k+\ell}] \tag{2.53}$$

in the special case $A_1 = \cdots = A_k = \mathrm{Id}$.

Next, we consider the size of $\mathfrak{m}_2[\cdot|\cdot]$. We have the following bounds, which we prove in Section 2.4.1.

Lemma 2.3.5. *Under the assumptions of Lemma 2.2.5, we have the estimates*

$$\big| \mathfrak{m}_{GUE}[T_1, \dots, T_k | T_{k+1}, \dots, T_{k+\ell}] \big| \lesssim \frac{1}{\eta_*^{k+\ell-\lceil (a+b)/2 \rceil}}, \tag{2.54}$$

$$\big| \mathfrak{m}_{\kappa}[T_1, \dots, T_k | T_{k+1}, \dots, T_{k+\ell}] \big| \lesssim \frac{1}{\eta_*^{k+\ell-1-\lceil (a+b)/2 \rceil}}. \tag{2.55}$$

Generically, both bounds are sharp if not all $\Im z_j$ have the same sign. In particular, $\mathfrak{m}_\kappa[\cdot|\cdot]$ is dominated by $\mathfrak{m}_{GUE}[\cdot|\cdot]$ on all mesoscopic scales and it holds that

$$\big| \mathfrak{m}_{2}[T_1, \dots, T_k | T_{k+1}, \dots, T_{k+\ell}] \big| \lesssim \frac{1}{\eta_*^{k+\ell-\lceil (a+b)/2 \rceil}}. \tag{2.56}$$

After this preparation, we state a CLT for resolvents which generalizes [40, 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 2.2.5 and its proof as input. We give it in Section 2.4.4.

Theorem 2.3.6 (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 2.1.1. 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 2.2.2, 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,*

$$N^p \mathbb{E}\Big(\prod_{j=1}^{p} X_{\alpha_j}^{(k_j,a_j)} \Big) = \sum_{Q \in Pair([p])} \prod_{\{i,j\} \in Q} \mathfrak{m}_2[\alpha_i | \alpha_j] + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*} \ \prod_{l=1}^{p} \eta_*^{k_l - a_l/2}} \Big) \tag{2.57}$$

for any $\varepsilon > 0$. Here, $\mathfrak{m}_2[\cdot|\cdot]$ is as in Definition 2.3.4 and $Pair(S)$ denotes the pairings of a set S. Equation (2.57) 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 $N\eta_* \to \infty$.

Remark (Independent modes). Note that (2.56) implies that two modes $X_{\alpha_1}^{(k_1,a_1)}$ and $X_{\alpha_2}^{(k_2,a_2)}$ in Theorem 2.3.6 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 for the macroscopic regime, i.e., if $\eta_* \gtrsim 1$. In the case $k = 1$, we may also write the deterministic matrix as

$$A = \langle A \rangle \mathrm{Id} + \mathring{A}_d + \mathring{A}_{od} \tag{2.58}$$

with A_d and A_{od} denoting the diagonal and off-diagonal part of $\mathring{A} = A - \langle A \rangle \mathrm{Id}$, respectively. The three resulting modes $\langle A \rangle \operatorname{Tr} f(W)$, $\operatorname{Tr} f(W)\mathring{A}_d$, and $\operatorname{Tr} f(W)\mathring{A}_{od}$ are asymptotically uncorrelated as $N\eta_* \to \infty$ (cf. [40, 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 $B_1, B_2 \in \{\mathrm{Id}, \mathring{A}_d, \mathring{A}_{od}\}$, this phenomenon is exclusive to the $k = 1$ case and decomposing $A_1, \dots, A_k$ for $k \geq 2$ according to (2.58) 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 (2.58) gives rise to 9 distinct modes. Here, computing the covariance, e.g., of $\operatorname{Tr}[G_1 \mathring{A}_d G_2 \mathring{B}_d]$ and $\operatorname{Tr}[G_1 \mathring{A}_{od} G_2 \mathring{B}_{od}]$, 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 2.3.3, we may use identities that are valid on the random matrix side, i.e., the left-hand side of (2.26), to derive further identities among the recursively defined quantities $m_2[\cdot|\cdot]$ and $\mathfrak{m}_2[\cdot|\cdot]$. The proof is analogous to the proof of Corollary 2.3.3 and hence omitted. We refer to Appendix 2.A for the setup of the necessary "meta argument".

Corollary 2.3.7. *Let $S_1, S_2 \neq \emptyset$ be two ordered multi-sets. Then*

(i) $m[S_1|S_2]$ is invariant under any permutation of the elements of S_1 as well as S_2.

(ii) $\mathfrak{m}_2[\cdot|\cdot]$ is cyclic in the sense that

$$\mathfrak{m}_2[(T_j, j \in S_1)|T_1, \dots, T_k] = \mathfrak{m}_2[(T_j, j \in S_1)|T_2, \dots, T_k, T_1].$$

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

$$\widetilde{m}[S_1|1, \dots, k] = \frac{\widetilde{m}[S_1|2, \dots, k] - \widetilde{m}[S_1|1, \dots, k-1]}{z_k - z_1}, \tag{2.59}$$

and we have the closed formula

$$\begin{aligned}\widetilde{m}[S_1|S_2] &= \sum_{(s,t)\in S_1\times S_2}\Big(\prod_{\substack{i\in S_1,\\ i\neq s}}\frac{1}{z_i-z_s}\prod_{\substack{j\in S_2,\\ j\neq t}}\frac{1}{z_j-z_t}\Big)\widetilde{m}[s|t]\\ &= \sum_{(s,t)\in S_1\times S_2}\Big(\prod_{\substack{i\in S_1,\\ i\neq s}}\frac{m[i,s]}{m_i-m_s}\prod_{\substack{j\in S_2,\\ j\neq t}}\frac{m[j,t]}{m_j-m_t}\Big)\widetilde{m}[s|t] \qquad (2.60)\end{aligned}$$

with $\widetilde{m}[s|t]=\frac{m_s'm_t'}{(1-m_sm_t)^2}$. *Recall that* $\widetilde{m}[\cdot|\cdot]$ *was defined in* (2.53).

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

$$\begin{aligned}&\mathfrak{m}_2[(T_j, j\in S_1)|T_1,\dots,T_{k-1},G_k]\\ &=\frac{\mathfrak{m}_2[(T_j, j\in S_1)|T_2,\dots,T_{k-1},G_kA_1]-\mathfrak{m}_2[(T_j, j\in S_1)|T_1,\dots,T_{k-1}]}{z_k-z_1}. \qquad (2.61)\end{aligned}$$

Moreover, we have the following alternative integral representation for $\widetilde{m}_{GUE}[\cdot|\cdot]$ (cf. decomposition in (2.52)). The proof of Corollary 2.3.8 is carried out in Section 2.4.3 below.

Corollary 2.3.8. *Let* $k,\ell\in\mathbb{N}$. *Then,*

$$\begin{aligned}&\widetilde{m}_{GUE}[1,\dots,k|k+1,\dots,k+\ell] \qquad (2.62)\\ &=\frac{1}{2}\int\int\Big(\sum_{i=1}^{k}\frac{1}{(x-z_i)^2}\cdot\prod_{j\neq i}\frac{1}{x-z_j}\Big)\Big(\sum_{i=k+1}^{k+\ell}\frac{1}{(y-z_i)^2}\cdot\prod_{j\neq i}\frac{1}{y-z_j}\Big)u(x,y)\mathrm{d}x\mathrm{d}y\end{aligned}$$

with the kernel $u:[-2,2]\times[-2,2]\to\mathbb{R}$ *in* (2.38).

Remark. 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 (2.9) is given by $\mathcal{B}(z_1,z_2)=1-m(z_1)m(z_2)$, we can also express (2.38) in terms of 1×1 determinants as

$$u(x,y)=-\frac{1}{2\pi^2}\Re\Big(\ln(\det[\mathcal{B}(x+\mathrm{i}0,y+\mathrm{i}0)])-\ln(\det[\mathcal{B}(x+\mathrm{i}0,y+\mathrm{i}0)])\Big)$$

to match the formulas in [125, Sect. 7] for $k=\ell=1$ and $A_1=A_2=\mathrm{Id}$. Note that W being a GUE matrix corresponds to the choice $\beta=2$ and $\mathcal{C}^{(4)}=0$ in the notation of [125].

2.4 Proofs

2.4.1 Proof of Lemmas 2.3.2 and 2.3.5 (Size of $\mathcal{E}[\cdot]$ and $\mathfrak{m}_2[\cdot|\cdot]$)

In this section, we prove the estimates identifying the size of the deterministic approximations $\mathcal{E}[\cdot]$ and $\mathfrak{m}_2[\cdot|\cdot]$. We start by noting two bounds for $q_{1,2}$ that are used for both proofs.

Lemma 2.4.1. *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,*

$$q_{1,2}=m[1,2]=\frac{m_1m_2}{1-m_1m_2}\lesssim\begin{cases}\zeta^{-1}, & \text{if } \Im z_1,\Im z_2 \text{ have the same sign,}\\ \eta_*^{-1}, & \text{if } \Im z_1,\Im z_2 \text{ have opposite signs.}\end{cases}$$

The estimates in Lemma 2.4.1 are immediate from the explicit form $m(z) = (-z + \sqrt{z^2-4})/2$ of the solution to (2.9). Next, we establish the estimate for $\mathcal{E}[\cdot]$.

Proof of Lemma 2.3.2. We show (2.43) by induction. As the base case $k = 1$ readily follows from (2.25), assume that the bound in (2.43) 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 $A_k = \mathrm{Id}$ and $\Im z_1$ and $\Im z_k$ have opposite signs. Here, (2.43) follows immediately by using Corollary 2.3.3(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 (2.45).

In the remaining cases, (2.43) follows from the recursion (2.42), which allows rewriting $\mathcal{E}[T_1, \dots, T_k]$ in terms of $\mathfrak{m}_1[\cdot]$ 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 2.4.1 yields

$$q_{1,k} \lesssim \frac{\pi}{\min\{|\Im m_1|, |\Im m_k|\}}.$$

Note that $\Im m(z)$ can be bounded from below independently of η_* for both the macroscopic scale and the the bulk regime of the mesoscopic scales (cf. Assumption 2.2.2). Thus, estimating the right-hand side of (2.42) using Lemma 2.1.4 and the induction hypothesis yields the claim. In particular, we obtain η_* with the exponent $-(k-1-\lceil \frac{a}{2} \rceil)$ 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 (2.42) do not contribute. Hence, (2.43) again follows from Lemma 2.1.4 and the induction hypothesis. □

Next, we show the estimates for $\mathfrak{m}_{GUE}[\cdot|\cdot]$, $\mathfrak{m}_\kappa[\cdot|\cdot]$, and $\mathfrak{m}_2[\cdot|\cdot]$. To illustrate the tools at hand, the bound for $\mathfrak{m}_{GUE}[\cdot|\cdot]$ is obtained from the explicit formula in [123, Thm. 2.4] while the bound for $\mathfrak{m}_\kappa[\cdot|\cdot]$ is proved using the recursion (2.49). We start with a lemma.

Lemma 2.4.2. *Under the assumptions of Lemma 2.2.5, let $A_1 = \cdots = A_{k+\ell} = \mathrm{Id}$ and $\kappa_4 = 0$. Then,*

$$|\widetilde{m}_{GUE}[1, \dots, k | k+1, \dots, k+\ell]| \lesssim \frac{1}{\eta_*^{k+\ell}} \tag{2.63}$$

$$|m_{\circ\circ}[1, \dots, k | k+1, \dots, k+\ell]| \lesssim \frac{1}{\eta_*^{k+\ell}} \tag{2.64}$$

where $m_{\circ\circ}[\cdot|\cdot]$ denotes the second-order free cumulant function associated with the iterated divided differences $m[\cdot]$ and $\widetilde{m}_{GUE}[\cdot|\cdot]$. Generically, both bounds are sharp if not all $\Im z_j$ have the same sign.

Proof of Lemma 2.4.2. The bound (2.63) follows by induction on k and ℓ. As the base case $k = \ell = 1$ is covered by Example 2.2.8, assume that the bound for $\widetilde{m}_{GUE}[\cdot|\cdot]$ holds for up to $k-1$ indices in the first argument and a fixed number ℓ indices in the second argument. Recall that $\widetilde{m}_{GUE}[\cdot|\cdot]$ is symmetric under the interchanging of its arguments by Definition 2.3.4(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 2.4.1, where the right-hand side can be bounded from above independently of η_* under Assumption 2.2.2. Rewriting $\widetilde{m}_{GUE}[1,\dots,k|k+1,\dots,k+\ell]$ using the recursion (2.49), the bound (2.63) follows directly from the induction hypothesis and the estimate for $m[\cdot]$ from Lemma 2.1.4.

Case 2: $\Im z_1$ and $\Im z_k$ have opposite signs: Recalling that $\widetilde{m}_{GUE}[\cdot|\cdot]$ has a divided difference structure (cf. Corollary 2.3.7), it follows from the induction hypothesis that

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

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 (2.63). This concludes the induction step.

The bound (2.64) for $m_{\circ\circ}[\cdot|\cdot]$ is an immediate consequence of the second-order moment-cumulant relation (2.32) as well as the estimates in (2.63) and Lemma 2.1.4. □

Proof of Lemma 2.3.5. Given Lemma 2.4.2, the bound (2.54) readily follows from [123, 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 $k = \ell = 2$. We have

$$\begin{aligned}
\mathfrak{m}_{GUE}[T_1,T_2|T_3,T_4] = {} & \langle A_1\rangle\langle A_2\rangle\langle A_3\rangle\langle A_4\rangle m_{\circ\circ}[1,2|3,4] + \langle A_1A_2\rangle\langle A_3\rangle\langle A_4\rangle m_{\circ\circ}[1|3,4]m_{\circ}[2] \\
& + \langle A_1A_2\rangle\langle A_3A_4\rangle m_{\circ\circ}[1|3]m_{\circ}[2]m_{\circ}[4] + \dots \\
& + \langle A_1\rangle\langle A_2A_4\rangle\langle A_3\rangle m_{\circ}[1,2,3,4] \\
& + \langle A_1\rangle\langle A_2A_3A_4\rangle m_{\circ}[1,2,3]m_{\circ}[4] + \dots,
\end{aligned}$$

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 $\widetilde{m}_{GUE}[\cdot|\cdot]$. By Lemmas 2.1.4 and 2.4.2, 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

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

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

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

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 $|K(\pi_2)| \leq l - \lceil b/2 \rceil$. In particular, the two parts of $\mathfrak{m}_{GUE}[\cdot|\cdot]$ only contribute equally to (2.54) if $\lceil \frac{a}{2} \rceil + \lceil \frac{b}{2} \rceil = \lceil \frac{a+b}{2} \rceil$.

The bound (2.55) follows by induction on k and ℓ. As the base case $k = \ell = 1$ is again covered by Example 2.2.8, 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 $\beta = \{(z_{k+1}, A_{k+1}), \ldots, (z_{k+\ell}, A_{k+\ell})\}$. 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 2.3.2, we distinguish two cases depending on the deterministic matrices $A_1, \ldots, A_k$.

Case 1 ($\exists j$ such that $A_j = \mathrm{Id}$): We start by noting that $\mathfrak{m}_{GUE}[\cdot|\cdot]$ satisfies (i)-(iv) of Corollary 2.3.7. This implies that the same holds for $\mathfrak{m}_\kappa[\cdot|\cdot] = \mathfrak{m}_2[\cdot|\cdot] - \mathfrak{m}_{GUE}[\cdot|\cdot]$. Using the divided difference structure, we rewrite $\mathfrak{m}_\kappa[T_1, \ldots, T_k | T_{k+1}, \ldots, T_{k+\ell}]$ either as a contour integral

$$\mathfrak{m}_\kappa[T_1, \ldots, T_{j-1}, G_j, T_{j+1}, \ldots, T_k|\beta] \\ = \frac{1}{2\pi \mathrm{i}} \int_{\mathbb{R}} \frac{\mathfrak{m}_\kappa[\ldots, T_{j-1}, G(x+\mathrm{i}\eta)A_{j+1}, \ldots|\beta] - \mathfrak{m}_\kappa[\ldots, T_{j-1}, G(x-\mathrm{i}\eta)A_{j+1}, \ldots|\beta]}{(x + \mathrm{i}s\eta - z_j)(x + \mathrm{i}s\eta - z_{j+1})} \mathrm{d}x \tag{2.65}$$

if $\Im z_j$ and $\Im z_{j+1}$ have the same sign, or as

$$\mathfrak{m}_\kappa[T_1, \ldots, T_{j-1}, G_j, T_{j+1}, \ldots, T_k|\beta] \\ = \frac{\mathfrak{m}_\kappa[T_1, \ldots, T_{j-1}, T_{j+1}, \ldots, T_k|\beta] - \mathfrak{m}_\kappa[T_1, \ldots, T_{j-1}, T_j A_{j+1}, T_{j+2}, \ldots, T_k|\beta]}{z_j - z_{j+1}} \tag{2.66}$$

if $\Im z_j$ and $\Im z_{j+1}$ have opposite signs. Estimating (2.65) and (2.66) using the induction hypothesis yields (2.55).

Case 2 (all $A_1, \ldots, A_k$ traceless): As $\mathfrak{m}_\kappa[\cdot|\cdot]$ solves the recursion (2.49) without the source term $\mathfrak{s}_{GUE}$ we can rewrite $\mathfrak{m}_\kappa[T_1, \ldots, T_k|T_{k+1}, \ldots, T_{k+\ell}]$ 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 (2.49) vanish. The desired estimate thus readily follows from the induction hypothesis and Lemma 2.1.4. For the source term $\mathfrak{s}_\kappa$ in (2.51), we obtain, e.g.,

$$\Big| \langle M_{[r]} \odot M_{(s,\ldots,k+\ell,k+1,\ldots,t)} \rangle \langle M_{[r,k]} \odot M_{[t,s]} \rangle \Big| \\ \leq \frac{1}{N^2} \sum_{x,y \in [N]} \Big| (M_{[r]})_{xx} (M_{(s,\ldots,k+\ell,k+1,\ldots,t)})_{xx} (M_{[r,k]})_{yy} (M_{[t,s]})_{yy} \Big| \\ \leq \Big(\frac{1}{\eta_*} \Big)^{k+\ell-\lceil (a+b)/2 \rceil}$$

for any $1 \le r \le k$ and $k+1 \le s \le t \le k+\ell$. Recall that $\lceil \frac{x}{2} \rceil + \lceil \frac{y}{2} \rceil \ge \lceil \frac{x+y}{2} \rceil$ 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 (2.55).

The bound (2.56) for $\mathfrak{m}_2[\cdot|\cdot]$ is immediate form the decomposition (2.29) using the estimates (2.54) and (2.55). □

2.4.2 Proof of Lemma 2.2.3 (Expansion for $\mathbb{E}\langle T_{[1,k]} \rangle$)

We use proof by induction to establish (2.23). As the base case $\mathbb{E}\langle T_\emptyset \rangle$ is trivial, assume that the expansion in Lemma 2.2.3 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 $A_j = \mathrm{Id}$, i.e., that are of the form $T_{[1,k]} = T_{[1,j\rangle} G_j G_{j+1} T_{\langle j+1,k]}$ 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 $G_j G_{j+1}$ 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 $s := \operatorname{sign}(\Im z_j) = \operatorname{sign}(\Im z_{j+1})$. By the residue theorem, we can write the product $G_j G_{j+1}$ as a contour integral (cf. [37, Lem. 3.2])

$$G_j G_{j+1} = \frac{1}{\pi} \int_{\mathbb{R}} \frac{\Im G(x+\mathrm{i}\eta)}{(x+\mathrm{i}s\eta - z_j)(x+\mathrm{i}s\eta - z_{j+1})} \mathrm{d}x, \quad G_j = G(z_j), \tag{2.67}$$

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 $\mathfrak{m}_1[\cdot]$ and $\mathcal{E}[\cdot]$ have a similar representation, as

$$\begin{aligned} &\mathfrak{m}_1[T_1, \dots, T_{j-1}, G_j, T_{j+1}, \dots, T_k] \\ &= \frac{1}{2\pi\mathrm{i}} \int_{\mathbb{R}} \frac{\mathfrak{m}_1[\dots, T_{j-1}, G(x+\mathrm{i}\eta)A_{j+1}, \dots] - \mathfrak{m}_1[\dots, T_{j-1}, G(x-\mathrm{i}\eta)A_{j+1}, \dots]}{(x+\mathrm{i}s\eta - z_j)(x+\mathrm{i}s\eta - z_{j+1})} \mathrm{d}x \end{aligned} \tag{2.68}$$

by the residue theorem and [39, Lem. 4.4] as well as

$$\begin{aligned} &\mathcal{E}[T_1, \dots, T_{j-1}, G_j, T_{j+1}, \dots, T_k] \\ &= \frac{1}{2\pi\mathrm{i}} \int_{\mathbb{R}} \frac{\mathcal{E}[\dots, T_{j-1}, G(x+\mathrm{i}\eta)A_{j+1}, \dots] - \mathcal{E}[\dots, T_{j-1}, G(x-\mathrm{i}\eta)A_{j+1}, \dots]}{(x+\mathrm{i}s\eta - z_j)(x+\mathrm{i}s\eta - z_{j+1})} \mathrm{d}x \end{aligned} \tag{2.69}$$

using Corollary 2.3.3(ii). Recall that the properties listed in Corollary 2.3.3 can be derived directly from the recursion (2.42), i.e., their proof is independent of Lemma 2.2.3 and the "meta argument" in Appendix 2.A. Rewriting the left-hand side of (2.23) using (2.67) and (2.68), we obtain an integral involving a resolvent chain of length $k-1$. Hence, by the induction hypothesis,

$$\begin{aligned} &\mathbb{E}(\langle T_{[1,k]} \rangle - \mathfrak{m}_1[T_1, \dots, T_k]) \\ &= \frac{\kappa_4}{2\pi\mathrm{i}N} \int_{\mathbb{R}} \frac{\mathcal{E}[\dots, T_{j-1}, G(x+\mathrm{i}\eta)A_{j+1}, \dots] - \mathcal{E}[\dots, T_{j-1}, G(x-\mathrm{i}\eta)A_{j+1}, \dots]}{(x+\mathrm{i}s\eta - z_j)(x+\mathrm{i}s\eta - z_{j+1})} \mathrm{d}x \\ &\quad + \mathcal{O}\Big(\frac{N^\varepsilon}{N\sqrt{N\eta_*}\,\eta_*^{k-1-a/2}} \Big). \end{aligned}$$

Evaluating the integral using (2.69) gives (2.23) as desired.

Case 2 ($\Im z_j$ and $\Im z_{j+1}$ have opposite signs): Applying the resolvent identity (2.19) and the divided difference structure of $\mathfrak{m}_1[\cdot]$ (cf. [39, Lem. 5.4]) yields

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

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 (2.45).

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

$$\underline{Wf(W)} := Wf(W) - \widetilde{\mathbb{E}}\widetilde{W}(\partial_{\widetilde{W}} f)(W) \tag{2.70}$$

with $\partial_{\widetilde{W}}$ denoting the directional derivative in direction $\widetilde{W}$ and $\widetilde{W}$ denoting an independent GUE matrix with expectation $\widetilde{\mathbb{E}}$. By construction, the renormalization in (2.70) 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 (2.70) for the resolvent $f(W) = (W-z)^{-1}$ yields the formulas

$$\begin{aligned}
\underline{WG_1} &= WG_1 + \langle G_1\rangle G_1, \\
\underline{WT_1\dots T_k} &= \underline{WG_1}A_1T_{[2,k]} + \sum_{j=2}^{k}\langle T_{[1,j\rangle}G_j\rangle T_{[j,k]},
\end{aligned} \tag{2.71}$$

and we further recall the identities

$$\langle G_1 - m_1\rangle = \frac{1}{1-m_1^2}(-m_1\langle\underline{WG_1}\rangle + m_1\langle G_1 - m_1\rangle^2) \tag{2.72}$$

$$\langle T_1\rangle - \mathfrak{m}_1[T_1] = -m_1\langle\underline{WT_1}\rangle + m_1^2\langle G_1 - m_1\rangle\langle A_1\rangle + m_1\langle G_1 - m_1\rangle\langle T_1 - \mathfrak{m}_1[T_1]\rangle \tag{2.73}$$

from (96) in [40]. To complete the induction step, we need the analog of (2.73) 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 (2.42) up to an $\mathcal{O}(N^\varepsilon/(\sqrt{N\eta_*}\eta_*^{k-a/2}))$ error.

A brief calculation (see the proof of the local law [39, Thm. 3.4] or [37, Lem. 4.1]) yields

$$\begin{aligned}
\langle T_{[1,k]}\rangle = m_1\Big(&- \langle\underline{WT_{[1,k]}}\rangle + \langle T_{[2,k]}A_1\rangle + \sum_{j=2}^{k-1}\langle T_{[1,j\rangle}G_j\rangle\langle T_{[j,k]}\rangle + \langle G_1 - m_1\rangle\langle T_{[1,k]}\rangle \\
&+ \langle T_{[1,k\rangle}G_k\rangle\langle (G_k - m_k)A_k\rangle\Big).
\end{aligned} \tag{2.74}$$

Next, we rewrite the equation to the form

$$\begin{aligned}
&\Big(1+\mathcal{O}_\prec\Big(\frac{1}{N\eta_*}\Big)\Big)(\langle T_{[1,k]}\rangle - \mathfrak{m}_1[T_1,\dots,T_k]) \\
&= m_1\Big(-\langle \underline{WT_{[1,k]}}\rangle + (\langle T_{[2,k\rangle}G_kA_kA_1\rangle - \mathfrak{m}_1[T_2,\dots,T_{k-1},G_kA_kA_1]) \\
&\quad + \sum_{j=1}^{k-1}(\langle T_{[1,j\rangle}G_j\rangle - \mathfrak{m}_1[T_1,\dots,T_{j-1},G_j])\mathfrak{m}_1[T_j,\dots,T_k] \\
&\quad + \sum_{j=2}^{k}\mathfrak{m}_1[T_1,\dots,T_{j-1},G_j](\langle T_{[j,k]}\rangle - \mathfrak{m}_1[T_j,\dots,T_k]) \\
&\quad + \sum_{j=2}^{k}(\langle T_{[1,j\rangle}G_j\rangle - \mathfrak{m}_1[T_1,\dots,T_{j-1},G_j])(\langle T_{[j,k]}\rangle - \mathfrak{m}_1[T_j,\dots,T_k])\Big),
\end{aligned} \tag{2.75}$$

where we applied (2.15) for $\langle G_1 - m_1\rangle$ on the right-hand side. Recall that $\langle A_k\rangle = 0$ and $a = k$ in the case considered. Moving the factor $(1+\mathcal{O}_\prec((N\eta_*)^{-1})$ to the right-hand side, multiplying (2.75) with N and taking the expectation yields

$$\begin{aligned}
N\mathbb{E}(\langle T_{[1,k]}\rangle - \mathfrak{m}_1[T_1,\dots,T_k]) = \Big(m_1+\mathcal{O}_\prec\Big(\frac{1}{N\eta_*}\Big)\Big)\Big(&- N\mathbb{E}\langle \underline{WT_{[1,k]}}\rangle + \kappa_4\mathcal{E}[T_1,\dots,T_k] \\
&- \kappa_4 \sum_{1\le r\le s\le t\le k}\langle M_{[r]}\odot M_{[s,t]}\rangle\langle M_{[r,s]}\odot(M_{[t,k]}A_k)\rangle\Big) \\
&+ \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\ \eta_*^{k/2}}\Big)
\end{aligned}$$

by the induction hypothesis, (2.42) and the expansion $\frac{1}{1+x} = 1+\mathcal{O}(x)$. We further applied (2.15) for the last line of (2.75) to obtain

$$(\langle T_{[1,j\rangle}G_j\rangle - \mathfrak{m}_1[T_1,\dots,T_{j-1},G_j])(\langle T_{[j,k]}\rangle - \mathfrak{m}_1[T_j,\dots,T_k]) = \mathcal{O}_\prec\Big(\frac{1}{N^2\eta_*^{k/2+1}}\Big).$$

It follows that

$$N\mathbb{E}(\langle T_{[1,j\rangle}G_j\rangle - \mathfrak{m}_1[T_1,\dots,T_{j-1},G_j])(\langle T_{[j,k]}\rangle - \mathfrak{m}_1[T_j,\dots,T_k]) = \mathcal{O}\Big(\frac{N^\varepsilon}{(N\eta_*)\ \eta_*^{k/2}}\Big),$$

i.e., the term is indeed part of the error. Hence, (2.23) is established if

$$N\mathbb{E}\langle \underline{WT_{[1,k]}}\rangle = -\kappa_4 \sum_{1\le r\le s\le t\le k}\langle M_{[r]}\odot M_{[s,t]}\rangle\langle M_{[r,s]}\odot(M_{[t,k]}A_k)\rangle + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\ \eta_*^{k/2}}\Big). \tag{2.76}$$

By cumulant expansion, the underlined term on the left-hand side of (2.76) is given by

$$N\mathbb{E}\langle \underline{WT_{[1,k]}}\rangle = \sum_{n\ge 2}\sum_{x,y\in[N]}\sum_{\nu\in\{xy,yx\}^n}\frac{\kappa(xy,\nu)}{n!}\mathbb{E}\partial_\nu(T_{[1,k]})_{yx}, \tag{2.77}$$

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 (2.77) is canceled out by the renormalization (2.70).

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 2.1.4, as the factor N^2 obtained from the double summation is canceled by the bound for the cumulant. Note that since

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

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 2.1.4) 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

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

and identify the term as part of the error in (2.23). It remains to consider the $n = 2$ and $n = 3$ contribution to (2.77). Recall that we write $T_{[i,j]} = T_i \dots T_j$ for $i \leq j$ and $T_{[i,i\rangle} = T_\emptyset = 0$.

Estimate for the $n = 2$ term of (2.77): Evaluating the derivative yields, e.g.,

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

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

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

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

$$N^{-3/2} \sum_{x,y\in[N]} (T_{[1,r\rangle} G_r)_{yx} (T_{[r,s\rangle} G_s)_{yy} (T_{[s,k]})_{xx} \tag{2.78}$$

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 (2.78) into a deterministic term satisfying the bounds in Lemma 2.1.4 and a fluctuation that is controlled by the local law (2.16). Recalling that

$$\|M_{[i,j]}A_j\| \lesssim \|M_{[i,j]}\| \lesssim \frac{1}{\eta_*^{(j-i)/2}} \tag{2.79}$$

for $i \leq j$ by Lemma 2.1.4, as $M_{[i,j]}$ involves $j-i$ traceless matrices and $\|A_j\| \lesssim 1$ by assumption, and that

$$\max_{x,y\in[N]} |(T_{[i,j\rangle}G_j - M_{[i,j]})A_j)_{xy}| = \mathcal{O}_\prec\Big(\frac{1}{\sqrt{N\eta_*}\ \eta_*^{(j-i)/2}}\Big) \tag{2.80}$$

by the isotropic local law (2.16), we obtain the bound

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

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

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

and we can include this term in the error in (2.23). Recall that we consider $N(\mathbb{E}\langle T_{[1,k]} - \mathfrak{m}_1[T_1,\dots,T_k]\rangle$, i.e., the error terms obtained differ from (2.23) by a factor of N.

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

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

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

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

by (2.79). Recalling that $T_{[1,r\rangle}G_r$ contains $r-1$ traceless matrices by assumptions, we obtain

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

from (2.79) and (2.16). Hence, the term can be included in the error in (2.23).

It remains to estimate the two terms in (2.78) that involve one deterministic and one fluctuation term each. By the Cauchy-Schwarz inequality, we obtain

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

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

$$\|\mathrm{dvec}(T_{[r,s\rangle}G_s - M_{[r,s]})\| \le \sqrt{N}\max_{x\in[N]}|(T_{[r,s\rangle}G_s - M_{[r,s]})_{xx}| = \mathcal{O}_\prec\Big(\frac{1}{\eta_*^{(s-r+1)/2}}\Big),$$

$$\|T_{[1,r\rangle}G_r\| \le \|M_{[r]}\| + N\max_{x,y\in[N]}|(T_{[1,r\rangle}G_r - M_{[r]})_{xy}| = \mathcal{O}_\prec\Big(\frac{1}{\eta_*^{(r-1)/2}} + \frac{\sqrt{N}}{\eta_*^{r/2}}\Big)$$

by (2.79), (2.80), and the fact that $\|B\| \le N\max_{x,y}|B_{xy}|$ for any matrix $B \in \mathbb{C}^{N\times N}$. Overall, we obtain

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

i.e., the term is part of the error in (2.23). 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 (2.77) can be included in the error term in (2.23).

Computation of the $n = 3$ term of (2.77): 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 (2.23) by decomposing each resolvent chain entry into a deterministic and a fluctuation part. Note that $M_{(\cdot)}$ is not necessarily diagonal, i.e., applying (2.16) 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,

$$\begin{aligned}\partial_\nu(T_{[1,k]})_{yx} = &-\sum_{1\le t\le s\le r\le k}(T_{[1,t\rangle}G_t)_{yy}(T_{[t,s\rangle}G_s)_{xx}(T_{[s,r\rangle}G_r)_{yy}(T_{[r,k]})_{xx}\\ &-\sum_{1\le r\le s\le t\le k}(T_{[1,r\rangle}G_r)_{yy}(T_{[r,s\rangle}G_s)_{xx}(T_{[s,t\rangle}G_t)_{yy}(T_{[t,k]})_{xx} - \dots \qquad (2.81)\end{aligned}$$

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

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

where the last equality follows from the definition of the Hadamard product.

Adding the contributions to (2.77) together, the cumulant expansion evaluates to

$$-N\mathbb{E}\langle \underline{W T_{[1,k]}}\rangle = \kappa_4 \sum_{1\le r\le s\le t\le k} \Big(\frac{1}{N^2}\sum_{x,y}(M_{[r]})_{yy}(M_{[r,s]})_{xx}(M_{[s,t]})_{yy}(M_{[t,k]}A_k)_{xx}\Big) + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\,\eta_*^{k-a/2}}\Big).$$

We conclude that $N(\mathbb{E}\langle T_{[1,k]}\rangle - \mathfrak{m}_1[T_1,\dots,T_k])$ coincides with the right-hand side of (2.42) up to an error term and an additional factor κ_4. Applying the recursion thus yields

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

as claimed. □

2.4.3 Proof of Lemma 2.2.5 (Deterministic Approximation 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 2.2.5.

Lemma 2.4.3 (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*

$$|\partial_\nu X_\alpha^{(k,a)}| = \mathcal{O}_\prec\Big(\frac{1}{N\eta_*^{k-a/2}}\Big). \tag{2.82}$$

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

$$\partial_{xy}X_\alpha^{(k,a)} = -\frac{1}{N}\Big[\sum_{r=1}^k (M_{(r,\dots,k,1,\dots,r)})_{yx} + \mathcal{O}_\prec\Big(\frac{1}{\sqrt{N\eta_*}\,\eta_*^{k-a/2}}\Big)\Big], \tag{2.83}$$

$$\partial_{vw}\partial_{xy}X_\alpha^{(k,a)} = \frac{1}{N}\Big[\sum_{r=1}^k\Big(\sum_{s=1}^r (M_{(r,\dots,k,1,\dots,s)})_{wx}(M_{[s,r]})_{yv} + \sum_{s=r}^k (M_{[r,s]})_{wx}(M_{(s,\dots,k,1,\dots,r)})_{yv}\Big) + \mathcal{O}_\prec\Big(\frac{1}{\sqrt{N\eta_*}\,\eta_*^{k-a/2}}\Big)\Big]. \tag{2.84}$$

Proof. The bounds in (2.83) and (2.84) follow directly from (2.16), Lemma 2.1.4, and

$$\partial_{xy}\langle T_{[1,k]}\rangle = -\frac{1}{N}\sum_{r=1}^k (T_{[r,k]}T_{[1,r\rangle}G_r)_{yx},$$

$$\partial_{vw}\partial_{xy}\langle T_{[1,k]}\rangle = \frac{1}{N}\sum_{r=1}^k\Big(\sum_{s=1}^r (T_{[r,k]}T_{[1,s\rangle}G_s)_{wx}(T_{[s,r\rangle}G_r)_{yv} + \sum_{s=r}^k (T_{[r,s\rangle}G_s)_{wx}(T_{[s,k]}T_{[1,r\rangle}G_r)_{yv}\Big).$$

The claim (2.82) follows inductively by (2.16) and Lemma 2.1.4. Note that (2.82) is also true for $\nu = \emptyset$ since

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

as a consequence of (2.15) and Lemma 2.2.3. □

Proof of Lemma 2.2.5. We use proof by induction on the length of the multi-indices α and β. First, note that the left-hand side of (2.26) 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 (2.48), 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 (2.26) 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 deterministic approximation of $N^2\mathbb{E}X_\alpha^{(k)}X_\beta^{(\ell)}$ is given by $\mathfrak{m}_1[T_1,\dots,T_k|T_{k+1}\dots T_{k+\ell}]$ 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 $A_j = \mathrm{Id}$, i.e., that are of the form $T_{[1,k]} = T_{[1,j\rangle}G_jG_{j+1}T_{\langle j+1,k]}$ 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_jG_{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 $s := \operatorname{sign}(\Im z_j) = \operatorname{sign}(\Im z_{j+1})$ and recall from (2.67) that the product G_jG_{j+1} can be rewritten as a contour integral using the residue theorem. We obtain a similar contour integral formula for $\mathfrak{m}_2[\cdot|\cdot]$ from Corollary 2.3.7(iii), namely

$$\begin{aligned}&\mathfrak{m}_2[T_1,\dots,T_{j-1},G_j,T_{j+1},\dots,T_k|\beta] \qquad (2.85)\\ &= \frac{1}{2\pi\mathrm{i}}\int_{\mathbb{R}} \frac{\mathfrak{m}_2[\dots,T_{j-1},G(x+\mathrm{i}\eta)A_{j+1},\dots|\beta] - \mathfrak{m}_2[\dots,T_{j-1},G(x-\mathrm{i}\eta)A_{j+1},\dots|\beta]}{(x+\mathrm{i}s\eta-z_j)(x+\mathrm{i}s\eta-z_{j+1})}\mathrm{d}x.\end{aligned}$$

Recall that the properties of $\mathfrak{m}_2[\cdot|\cdot]$ stated in Corollary 2.3.7 are a consequence of the recursion (2.49) and thus independent of Lemma 2.2.5. Rewriting the left-hand side of (2.26) using (2.68) 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 $A_j = \mathrm{Id}$, i.e., the number of traceless matrices a is unchanged. By the induction hypothesis, we thus obtain

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

where the integral evaluates to $\mathfrak{m}_2[\alpha|\beta]$ by (2.85). Hence, (2.26) holds in the case considered.

Case 2: $\Im z_j$ and $\Im z_{j+1}$ have opposite signs: Applying (2.19) and the divided difference structure of $\mathfrak{m}_2[\cdot|\cdot]$, it follows that

$$\begin{aligned}&N^2\mathbb{E}X_\alpha^{(k,a)}X_\beta^{(\ell,b)}\\ &= N^2\Big(\frac{(\langle T_{[1,j\rangle}G_jA_{j+1}T_{\langle j+1,k]}\rangle - \mathbb{E}\langle T_{[1,j\rangle}G_jA_{j+1}T_{\langle j+1,k]}\rangle)}{z_j-z_{j+1}}\\ &\qquad - \frac{(\langle T_{[1,j\rangle}T_{[j+1,k]}\rangle - \mathbb{E}\langle T_{[1,j\rangle}T_{[j+1,k]}\rangle)}{z_j-z_{j+1}}\Big)X_\beta^{(\ell,b)}\end{aligned}$$

$$= \frac{\mathfrak{m}_2[T_1,\dots,T_{j-1},T_{j+1},\dots,T_k|\beta] - \mathfrak{m}_2[T_1,\dots,T_{j-1},T_jA_{j+1},T_{j+2},\dots,T_k|\beta]}{z_j - z_{j+1}}$$
$$+ \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\eta_*^{k+\ell-(a+b)/2-1}|z_j-z_{j+1}|}\Big)$$
$$= \mathfrak{m}_2[T_1,\dots,T_k|\beta] + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\ \eta_*^{k+\ell-(a+b)/2}}\Big),$$

since $|z_j - z_{j+1}| \geq 2\eta_*$ in the case considered. This concludes the proof of (2.26) for resolvent chains that contain at least one deterministic matrix $A_j = \mathrm{Id}$.

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 2.2.3, 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 (2.49) up to an $N^\varepsilon/(\sqrt{N\eta_*}\eta_*^{k-a/2}\eta_*^{\ell-b/2})$ error.

Combining (2.75) and Lemma 2.2.3 yields the starting point

$$\begin{aligned} X_\alpha^{(k,a)} = m_1\Big(&- (\langle \underline{WT_{[1,k]}}\rangle - \mathbb{E}\langle \underline{WT_{[1,k]}}\rangle) + (\langle T_{[2,k\rangle}G_kA_kA_1\rangle - \mathbb{E}\langle T_{[2,k\rangle}G_kA_kA_1\rangle) \\ &+ \sum_{j=1}^{k-1}(\langle T_{[1,j\rangle}G_j\rangle - \mathbb{E}\langle T_{[1,j\rangle}G_j\rangle)\mathfrak{m}_1[T_j,\dots,T_k] \\ &+ \sum_{j=2}^{k}\mathfrak{m}_1[T_1,\dots,T_{j-1},G_j](\langle T_{[j,k]}\rangle - \mathbb{E}\langle T_{[j,k]}\rangle)\Big) + \mathcal{O}_\prec\Big(\frac{1}{N\sqrt{N\eta_*}\ \eta_*^{k/2}}\Big). \end{aligned} \tag{2.86}$$

Next, multiply (2.86) by $N^2X_\beta^{(\ell)}$ 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

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

where we used the recursion (2.49) to introduce $\mathfrak{m}_2[\alpha|\beta]$. It remains to compute

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

By cumulant expansion, we obtain

$$\begin{aligned} N^2\mathbb{E}\Big(\langle \underline{WT_{[1,k]}}\rangle X_\beta^{(\ell)}\Big) = &-\sum_{j=1}^{\ell}\mathbb{E}\langle T_{[1,k]}T_{[k+j,k+\ell]}T_{[1,k+j\rangle}G_{k+j}\rangle \\ &+ N\sum_{n=2}^{3}\sum_{x,y\in[N]}\sum_{\nu\in\{xy,yx\}^n}\frac{\kappa(xy,\nu)}{n!}\mathbb{E}\partial_\nu\Big((T_{[1,k]})_{yx}X_\beta^{(\ell)}\Big) \\ &+ \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\ \eta_*^{(k+\ell)/2}}\Big), \end{aligned} \tag{2.87}$$

where the $n = 1$ term was evaluated using (2.70) 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 2.1.4 and 2.4.3. By the isotropic local law (2.16) we have

$$\mathbb{E}\langle T_{[1,k]} T_{[k+j,k+\ell]} T_{[1,k+j\rangle} G_{k+j}\rangle = \mathfrak{m}_1[T_1, \ldots, T_k, T_{k+j}, \ldots, T_{k+j-1}, G_{k+j}] + \mathcal{O}\Big(\frac{N^\varepsilon}{(N\eta_*)\ \eta_*^{(k+\ell)/2}}\Big),$$

which yields an error term that can be included in the error in (2.26). Hence, it only remains to consider the contributions for $n = 2$ and $n = 3$.

Estimate for the $n = 2$ term of (2.87): Evaluating the derivative $\partial_\nu((T_{[1,k]})_{yx} X_\beta^{(\ell)})$ always yields at least one xy or yx entry of a resolvent chain, which we can use to apply isotropic resummation (cf. estimate for (2.78) above). Note that $X_\beta^{(\ell)}$ always contributes a factor N^{-1} due to Lemma 2.4.3, either from evaluating derivatives using (2.83) and (2.84), or from the trivial estimate (2.82). The additional factor N in front of the x, y summation in (2.87) is thus balanced out. Overall, we obtain

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

In particular, the $n = 2$ term can be included in the error in the last line of (2.87).

Computation of the $n = 3$ term of (2.87): By applying the Leibniz rule, we distribute the derivatives as

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

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 2.2.3. Using (2.81) and (2.16), it follows that

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

since $X_\beta^{(\ell)}$ is centered. We hence include the term in the error in (2.87).

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 (2.87). It remains to

consider the case $\nu_1 = (yx)$ and $\nu_2 \in \{(xy, yx), (yx, xy)\}$. Here, we compute

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

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 (2.87) by isotropic resummation. Applying (2.16) and recalling that $\kappa(xy,\nu) = \kappa_4/N^2$ in the case considered, it follows that

$$\begin{aligned}&N\sum_{x,y}\frac{\kappa_4}{N^2}\mathbb{E}\Big(\big(\partial_{\nu_1}(T_{[1,k]})_{yx}\big)\big(\partial_{\nu_2}X_\beta^{(\ell)}\big)\Big)\\ &= -\frac{\kappa_4}{N^2}\sum_{x,y}\Big(\sum_{r=1}^{k}\sum_{s=k+1}^{k+\ell}(M_{[r]})_{yy}(M_{[r,k]}A_k)_{xx}\Big(\sum_{t=k+1}^{s}(M_{(s,\dots,k+\ell,k+1,\dots,t)})_{yy}(M_{[t,s]})_{xx}\\ &\quad + \sum_{t=s}^{k+\ell}(M_{[s,t]})_{yy}(M_{(t,\dots,k+\ell,k+1,\dots,s)})_{xx}\Big)\Big) + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\,\eta_*^{k-a/2}\eta_*^{\ell-b/2}}\Big)\end{aligned}\tag{2.88}$$

which can again be rewritten using the definition of the Hadamard product. Note that we obtain six terms of the form (2.88) 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 (2.87).

Adding the contributions to (2.87) back together, the cumulant expansion evaluates to

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

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 (2.49). Applying the recursion yields

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

which completes the induction step. □

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

Proof of Corollary 2.3.8. We use proof by induction, starting with the base case $k = \ell = 1$. While we could establish (2.62) directly using a similar computation as in [40] or [125], it is much quicker to read off the kernel from the existing literature. To do so, note that the function $\widetilde{m}_{GUE}[1|2]$ and its counterpart for W being a GOE matrix only differ by a constant factor. More precisely, evaluating (92) of [40] for GOE ($\sigma = 1$, $\kappa_4 = 0$, $\widetilde{\omega}_2 = 0$) yields

$$\lim_{N\to\infty} N^2 \mathrm{Cov}(\langle G_1 - \mathbb{E}G_1\rangle, \overline{\langle G_2 - \mathbb{E}G_2\rangle}) = 2 \cdot \frac{m_1' m_2'}{(1 - m_1 m_2)^2} = 2 \cdot \widetilde{m}_{GUE}[1|2]. \tag{2.89}$$

This allows us to obtain the desired integral representation for $\widetilde{m}_{GUE}[1|2]$ from [52, Thm. 2.3], which only applies to real symmetric Gaussian matrices. Applying the theorem yields

$$\lim_{N\to\infty} N^2 \mathrm{Cov}(\langle G_1 - \mathbb{E}G_1\rangle, \overline{\langle G_2 - \mathbb{E}G_2\rangle}) = \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{1}{(x - z_1)^2}\frac{1}{(y - z_2)^2} u(x, y)\mathrm{d}x\mathrm{d}y, \tag{2.90}$$

where u is given by

$$u(x, y) = \frac{\sqrt{4 - x^2}\sqrt{4 - y^2}}{\pi^2} \cdot \int_0^1 \frac{1 - w^2}{w^2(x - y)^2 - wxy(1 - w)^2 + (1 - w^2)^2}\mathrm{d}w. \tag{2.91}$$

In the notation of [52], Equation (2.90) 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 [52, Thm. 2.3]. As (2.90) 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 (2.91) can be carried out explicitly, yielding the form in (2.38).

Comparing (2.89) and (2.90), we obtain (2.62) in the case $k = \ell = 1$.

The induction step uses the divided difference structure from Corollary 2.3.7(iii) and is similar to the proof of [39, Lem. 4.1]. Assume that the integral representation (2.62) holds for $\widetilde{m}_{GUE}[S_1|S_2]$ with $|S_1| = k$ and $|S_2| \le \ell$ for some fixed $k \ge 1$. We start by rewriting

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

where the induction hypothesis applies to both summands in the nominator, 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 (2.62) as claimed. Since $\widetilde{m}_{GUE}[S_1|S_2] = \widetilde{m}_{GUE}[S_2|S_1]$ by definition, the same argument applies for the other entry of $\widetilde{m}_{GUE}[\cdot|\cdot]$. □

2.4.4 Proof of Theorem 2.3.6 (CLT for Resolvents)

We use proof by induction over the number of factors on the left-hand side of (2.57). 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 2.2.5, which give $\mathbb{E}X_{\alpha_1} = 0$ and

$$N^2\mathbb{E}(X_{\alpha_1}X_{\alpha_2}) = \mathfrak{m}_2[\alpha_1|\alpha_2] + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\ \eta_*^{k_1-a_1/2}\eta_*^{k_2-a_2/2}}\Big),$$

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

$$N^p\mathbb{E}\Big(\prod_{j=1}^{p} X_{\alpha_j}\Big) = \sum_{Q\in Pair(\{1,\dots n\})}\ \prod_{(i,j)\in Q} \mathfrak{m}_2[\alpha_i|\alpha_j] + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\ \prod_{l=1}^{p}\eta_*^{k_l-a_l/2}}\Big) \tag{2.92}$$

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

$$N^{p+1}\mathbb{E}\Big((\langle T_1^{(1)}\rangle - \mathbb{E}\langle T_1^{(1)}\rangle)X_{\alpha_2}\dots X_{\alpha_{p+1}}\Big). \tag{2.93}$$

Combining (2.73) and Lemma 2.2.3 yields

$$\begin{aligned}\langle T_1^{(1)}\rangle - \mathbb{E}\langle T_1^{(1)}\rangle = -m(z_1^{(1)})\Big(\langle \underline{WT_1^{(1)}}\rangle + q_{1,1}^{(1)}\langle \underline{WG_1^{(1)}}\rangle\langle A_1^{(1)}\rangle\Big)\\ - \frac{\kappa_4}{N}\mathcal{E}[T_1] + \mathcal{O}\Big(\frac{N^\varepsilon}{N\ \sqrt{N\eta_*}\ \eta_*^{k_1-a_1/2}}\Big),\end{aligned} \tag{2.94}$$

where the superscript in $q^{(1)}$ indicates that the arguments are taken from α_1. We use (2.94) to replace the first factor in (2.93). The underlined terms are again treated by cumulant expansion. This yields

$$\begin{aligned}&- m(z_1^{(1)})N^{p+1}\mathbb{E}\Big(\langle \underline{WT_1}\rangle X_{\alpha_2}\dots X_{\alpha_{p+1}}\Big)\\ &= m(z_1^{(1)})\sum_{i=2}^{p+1}\sum_{j=1}^{k_i}\mathbb{E}\Big(\langle T_1^{(1)}T_j^{(i)}\dots T_{k_i}^{(i)}T_1^{(i)}\dots T_{j-1}^{(i)}G_j^{(i)}\rangle\cdot N^{p-1}\prod_{r\neq 1,i}X_{\alpha_r}\Big)\\ &\quad - m(z_1^{(1)})N^p\sum_{n\geq 2}\sum_{x,y\in[N]}\sum_{\nu\in\{xy,yx\}^n}\frac{\kappa(xy,\nu)}{n!}\mathbb{E}\partial_\nu\Big((T_1^{(1)})_{yx}X_{\alpha_2}\dots X_{\alpha_{p+1}}\Big),\end{aligned} \tag{2.95}$$

where the $n = 1$ contribution was again evaluated using (2.70) and the computations in the proof of Lemma 2.4.3. 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 (2.15), we have

$$\begin{aligned}\mathbb{E}\langle T_1^{(1)}T_j^{(i)}\dots T_{k_i}^{(i)}T_1^{(i)}\dots T_{j-1}^{(i)}G_j^{(i)}\rangle = \mathfrak{m}_1[T_1^{(1)}, T_j^{(i)}, \dots, T_{j-1}^{(i)}, G_j^{(i)}]\\ + \mathcal{O}\Big(\frac{N^\varepsilon}{N\eta_*\ \eta_*^{k_i-(a_1+a_i)/2}}\Big).\end{aligned}$$

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 2.2.5 to conclude that the contributions for $n \geq 4$ in (2.95) are sub-leading. Similarly, most of the terms arising for $n = 2$ and $n = 3$ in (2.95) were already computed in the proof of Lemma 2.2.5 and their treatment here is analogous. We, therefore, focus on the differences between (2.87) and (2.95) 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 (2.95): Compared to (2.87), 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

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

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 (2.95): We distinguish three cases for $|\nu_1|$.

Case 1 ($|\nu_1| = 3$): After evaluating the derivative, we apply (2.16) to get

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

In particular, adding the above terms for the cumulant expansions resulting from the two underlined terms in (2.94) yields

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

Case 2 ($|\nu_1| = 1$): Whenever the two remaining derivatives act on different factors of the product $X_{\alpha_2}\dots X_{\alpha_{p+1}}$, (2.83) 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 (2.95). Otherwise, we evaluate the term similar to (2.88). Let again $\nu_2 = \nu \setminus \nu_1$. For $i = 2,\dots,p+1$, it follows that

$$\begin{aligned}
&N^p \sum_{x,y} \frac{\kappa_4}{N^2} \mathbb{E}\Big(\big(\partial_{\nu_1}(T_1)_{yx}\big)X_{\alpha_2}\dots X_{\alpha_{i-1}}\big(\partial_{\nu_2}X_{\alpha_i}\big)X_{\alpha_{i+1}}\dots X_{\alpha_{p+1}}\Big) \\
&= -\kappa_4 \sum_{s=1}^{k_i}\Big(\sum_{t=1}^{s} m(z_1^{(1)})^2\langle A_1^{(1)}\odot M_{[s,t]}^{(i)})\rangle\langle M_{(t,\dots,k_i,1,\dots,s)}^{(i)}\rangle \\
&\qquad + \sum_{t=s}^{k_i} m(z_1^{(1)})^2\langle A_1^{(1)}\odot M_{(s,\dots,k_i,1,\dots,t)}^{(i)})\rangle\langle M_{[t,s]}^{(i)})\rangle\Big)\mathbb{E}\Big(\prod_{r\neq 1,i} X_{\alpha_r}\Big) \\
&\quad + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\ \prod_{l=1}^{p+1}\eta_*^{k_l-a_l/2}}\Big),
\end{aligned}$$

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 (2.87). We omit the details.

Overall, the cumulant expansion (2.95) evaluates to

$$\begin{aligned}
&N^{p+1}\mathbb{E}\Big(- m(z_1^{(1)})\langle \underline{WT_1^{(1)}}\rangle X_{\alpha_2}\dots X_{\alpha_{p+1}}\Big)\\
&= N^{p-1}\mathbb{E}\Big(\prod_{r\neq 1,j} X_{\alpha_r}\Big)\cdot m(z_1^{(1)})\sum_{i=2}^{p+1}\Big[\sum_{j=1}^{k_i}\mathfrak{m}_1[T_1^{(1)},T_j^{(i)}\dots T_{j-1}^{(i)},G_j^{(i)}]\\
&\quad + q_{1,1}^{(1)}\sum_{j=1}^{k_i}\mathfrak{m}_1[G_1^{(1)},T_j^{(i)}\dots T_{j-1}^{(i)},G_j^{(i)}]\langle A_1^{(1)}\rangle\\
&\quad + \kappa_4\sum_{s=1}^{k_i}\Big(\sum_{t=1}^{s} m(z_1^{(1)})^2\langle A_1^{(1)}\odot M_{[s,t]}^{(i)}\rangle\langle M_{(t,\dots,k_i,1,\dots,s)}^{(i)}\rangle\\
&\quad + \sum_{t=s}^{k_i} m(z_1^{(1)})^2\langle A_1^{(1)}\odot M_{(t,\dots,k_i,1,\dots,s)}^{(i)}\rangle\langle M_{[s,t]}^{(i)}\rangle\Big)\\
&\quad + \kappa_4 q_{1,1}^{(1)}\sum_{s=1}^{k_i}\Big(\sum_{t=1}^{s} m(z_1^{(1)})^2\langle A_1^{(1)})_{xx}\odot M_{[s,t]}^{(i)})\rangle\langle M_{(t,\dots,k_i,1,\dots,s)}^{(i)}\rangle\\
&\quad + \sum_{t=s}^{k_i} m(z_1^{(1)})^2\langle A_1^{(1)}\odot M_{(t,\dots,k_i,1,\dots,s)}^{(i)}\rangle\langle M_{[s,t]}^{(i)}\rangle\Big)\langle A_1^{(1)}\rangle\Big] + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\ \prod_{l=1}^{p+1}\eta_*^{k_l-a_l/2}}\Big).
\end{aligned}$$

Adding the contribution for all three terms in (2.94) together allows rewriting (2.93) as

$$\begin{aligned}
&N^{p+1}\mathbb{E}\Big((\langle T_1^{(1)}\rangle - \mathbb{E}\langle T_1^{(1)}\rangle)X_{\alpha_2}\dots X_{\alpha_{p+1}}\Big)\\
&= \sum_{i=2}^{p+1}\mathfrak{m}_1[T_1^{(1)}|T_1^{(i)},\dots,T_{k_i}^{(i)}]\cdot N^{p-1}\mathbb{E}\Big(\prod_{r\neq 1,j} X_{\alpha_r}\Big) + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\ \prod_{l=1}^{p+1}\eta_*^{k_l-a_l/2}}\Big),
\end{aligned}$$

where we used (2.49) with $k = 1$ and $\ell = k_i$. As the remaining product only consists of $p-1$ factors, we can apply the induction hypothesis (2.92) and conclude

$$N^{p+1}\mathbb{E}\Big(\prod_{j=1}^{p+1} X_{\alpha_j}\Big) = \sum_{Q\in Pair([p+1])}\ \prod_{(i,j)\in Q}\mathfrak{m}_1[\alpha_i|\alpha_j] + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\ \prod_{l=1}^{p+1}\eta_*^{k_l-a_l/2}}\Big) \tag{2.96}$$

for the case $\alpha_1 = \{(z_1^{(1)}, A_1^{(1)})\}$. Assume next that (2.96) 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 (2.86). This gives

$$\begin{aligned}
&N^{p+1}\mathbb{E}\Big(X_{\alpha_1}\dots X_{\alpha_{p+1}}\Big)\\
&= N^{p+1}m(z_1^{(1)})\mathbb{E}\Big(\Big(- \langle\underline{WT_1^{(1)}\dots T_{k_1}^{(1)}}\rangle\\
&\quad + (\langle T_2^{(1)}\dots T_{k_1-1}^{(1)}G_{k_1}^{(1)}A_{k_1}^{(1)}A_1^{(1)}\rangle - \mathbb{E}\langle T_2^{(1)}\dots T_{k_1-1}^{(1)}G_{k_1}^{(1)}A_{k_1}^{(1)}A_1^{(1)}\rangle\\
&\quad + \frac{\kappa_4}{N}\mathcal{E}[T_2^{(1)},\dots,T_{k_1-1}^{(1)},G_{k_1}^{(1)}A_{k_1}^{(1)}A_1^{(1)}])
\end{aligned} \tag{2.97}$$

$$
\begin{aligned}
&+\sum_{j=1}^{k-1}(\langle T_1^{(1)}\dots T_{j-1}^{(1)}G_j^{(1)}\rangle-\mathbb{E}\langle T_1^{(1)}\dots T_{j-1}^{(1)}G_j^{(1)}\rangle\\
&+\frac{\kappa_4}{N}\mathcal{E}[T_1^{(1)},\dots,T_{j-1}^{(1)},G_j^{(1)}])\mathfrak{m}_1[T_j^{(1)},\dots,T_{k_1}^{(1)}]\\
&+\sum_{j=2}^{k}\mathfrak{m}_1[T_1^{(1)},\dots,T_{j-1}^{(1)},G_j^{(1)}](\langle T_j^{(1)}\dots T_{k_1}^{(1)}\rangle-\mathbb{E}\langle T_j^{(1)}\dots T_{k_1}^{(1)}\rangle+\frac{\kappa_4}{N}\mathcal{E}[T_j^{(1)},\dots,T_{k_1}^{(1)}])\Big)\\
&+\frac{\kappa_4}{N}\mathcal{E}[T_1^{(1)},\dots,T_{k_1}^{(1)}]\Big)X_{\alpha_2}\dots X_{\alpha_{p+1}}\Big)+\mathcal{O}\Big(\frac{N^{\varepsilon}}{\sqrt{N\eta_*}\ \prod_{l=1}^{p+1}\eta_*^{k_l-a_l/2}}\Big),
\end{aligned}
$$

where the underlined term can again be treated by cumulant expansion. Similar to (2.95), we evaluate

$$
\begin{aligned}
&N^{p+1}\mathbb{E}\Big(-m(z_1^{(1)})\langle\underline{WT_1^{(1)}\dots T_{k_1}^{(1)}}\rangle X_{\alpha_2}\dots X_{\alpha_{p+1}}\Big)\\
&=m(z_1^{(1)})\sum_{i=2}^{p+1}\Bigg(\sum_{j=1}^{k_i}\mathfrak{m}_1[T_1^{(1)},\dots,T_{k_1}^{(1)},T_j^{(i)},\dots,T_{j-1}^{(i)},G_j^{(i)}]\\
&\quad+\kappa_4\sum_{r=1}^{k_1}\sum_{s=k+1}^{k+\ell}\Big(\sum_{t=1}^{s}\langle M_{[r]}^{(1)}\odot M_{(s,\dots,k_i,1,\dots,t)}^{(i)}\rangle\langle(M_{[r,k]}^{(1)}A_k^{(1)})\odot M_{[t,s]}^{(i)}\rangle\\
&\quad+\sum_{t=s}^{k_i}\langle M_{[r]}^{(1)}\odot M_{[s,t]}^{(i)}\rangle\langle(M_{[r,k]}^{(1)}A_k^{(1)})\odot M_{(t,\dots,k_i,1,\dots,s)}^{(i)}\rangle\Big)\Bigg)N^{p-1}\mathbb{E}\Big(\prod_{r\neq1,i}X_{\alpha_r}\Big)\\
&\quad+\mathcal{O}\Big(\frac{N^{\varepsilon}}{\sqrt{N\eta_*}\ \prod_{l=1}^{p+1}\eta_*^{k_l-a_l/2}}\Big),
\end{aligned}
\tag{2.98}
$$

and apply the induction hypothesis (2.92) for the product of $p-1$ factors. Note that any terms remaining in (2.97) now involve at most k_1-1 resolvents in the first factor. Hence, (2.96) applies and we obtain, e.g.,

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

Moreover, note that (2.98) contains the source term of the recursion for $\mathcal{E}[\cdot]$ (cf. Lemma 2.2.3), which we can combine with the terms involving $\mathcal{E}[\cdot]$ in (2.97). Using (2.42), the terms cancel. We conclude that (2.97) evaluates to

$$
\begin{aligned}
&N^{p+1}\mathbb{E}\Big(X_{\alpha_1}^{(k_1)}\dots X_{\alpha_{p+1}}^{(k_{p+1})}\Big)\\
&=\sum_{r=2}^{p+1}\Big[m(z_1^{(1)})\Big(\sum_{j=1}^{k_r}\mathfrak{m}_1[T_1^{(1)},\dots,T_{k_1}^{(1)},T_j^{(r)},\dots,T_{j-1}^{(r)},G_j^{(r)}]\\
&\quad+\kappa_4\sum_{r=1}^{k_1}\sum_{s=k+1}^{k+\ell}\Big(\sum_{t=1}^{s}\langle M_{[r]}^{(1)}\odot M_{(s,\dots,k_i,1,\dots,t)}^{(i)}\rangle\langle(M_{[r,k]}^{(1)}A_k^{(1)})\odot M_{[t,s]}^{(i)}\rangle\\
&\quad+\sum_{t=s}^{k_i}\langle M_{[r]}^{(1)}\odot M_{[s,t]}^{(i)}\rangle\langle(M_{[r,k]}^{(1)}A_k^{(1)})\odot M_{(t,\dots,k_i,1,\dots,s)}^{(i)}\rangle\Big)\\
&\quad+\mathfrak{m}_2[T_2^{(1)},\dots,T_{k_1-1}^{(1)},G_{k_1}^{(1)}A_{k_1}^{(1)}A_1^{(1)}|T_1^{(r)},\dots,T_{k_r}^{(r)}]
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{j=2}^{k}\mathfrak{m}_1[T_1^{(1)},\dots,T_{j-1}^{(1)},G_j^{(1)}]\mathfrak{m}_2[T_j^{(1)},\dots,T_{k_1}^{(1)}|T_1^{(r)},\dots,T_{k_r}^{(r)}]\\
&+\sum_{j=2}^{k}\mathfrak{m}_2[T_1^{(1)},\dots,T_{j-1}^{(1)},G_j^{(1)}|T_1^{(r)},\dots,T_{k_r}^{(r)}]\mathfrak{m}_1[T_j^{(1)},\dots,T_{k_1}^{(1)}]\Big)\Big]\\
&\times\Big(\sum_{Q\in Pair([p+1]\setminus\{1,i\})}\prod_{(i,j)\in Q}\mathfrak{m}_2[\alpha_i|\alpha_j]\Big)+\mathcal{O}\Big(\frac{N^{\varepsilon}}{\sqrt{N\eta_*}\ \prod_{l=1}^{p+1}\eta_*^{k_l-a_l/2}}\Big)\\
=&\sum_{Q\in Pair([p+1])}\prod_{(i,j)\in Q}\mathfrak{m}_2[\alpha_i|\alpha_j]+\mathcal{O}\Big(\frac{N^{\varepsilon}}{\sqrt{N\eta_*}\ \prod_{l=1}^{p+1}\eta_*^{k_l-a_l/2}}\Big)
\end{aligned}
$$

where the last equation follows from (2.49). Hence, (2.96) also holds for α_1 of length k_1. Moreover, (2.92) stays true if the product on the left-hand side contains $p+1$ factors, which concludes the proof of (2.57). □

2.4.5 Proof of Theorem 2.2.7 (Multi-Point Functional CLT)

The proof of the multi-point functional CLT in Theorem 2.2.7 consists of two parts. In the first step, we use Helffer-Sjöstrand representation (see [49]) 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 2.3. The second step is the computation of the leading terms, which establishes the covariance structure in (2.28).

By eigenvalue rigidity (see e.g., [62, Thm. 7.6] or [68]), 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., Soboloev 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 $A_j=\langle A_j\rangle\mathrm{Id}+\mathring{A}_j$ with $\langle\mathring{A}_j\rangle=0$. 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

$$
f_{\mathbb{C}}(z)=f_{\mathbb{C},p}(x+\mathrm{i}\eta):=\Big[\sum_{j=0}^{p-1}\frac{(\mathrm{i}\eta)^j}{j!}f^{(j)}(x)\Big]\widetilde{\chi}(N^{\gamma}\eta) \tag{2.99}
$$

where $\widetilde{\chi}$ is a smooth cutoff function that is equal to one on $[-5,5]$ and vanishes on $[-10,10]^c$. Note that (2.99) together with (2.21) implies the bound

$$
\int_{\mathbb{R}}|\partial_{\overline{z}}f_{\mathbb{C},p}(x+\mathrm{i}\eta)|\mathrm{d}x\lesssim\eta^{p-1}\|f\|_{H^p}\lesssim\eta^{p-1}N^{\gamma p}. \tag{2.100}
$$

By Helffer-Sjöstrand representation, we have

$$
f(\lambda)=\frac{1}{\pi}\int_{\mathbb{C}}\frac{\partial_{\overline{z}}f_{\mathbb{C}}(z)}{\lambda-z}\mathrm{d}^2z, \tag{2.101}
$$

where $\mathrm{d}^2 z = \mathrm{d}x\mathrm{d}\eta$ denotes the Lebesgue measure on $\mathbb{C} \equiv \mathbb{R}^2$ with $z = x + \mathrm{i}\eta$ and $\partial_{\bar{z}} = (\partial_x + \mathrm{i}\partial_\eta)/2$. Applying (2.101) for $f_1, \dots, f_k$, respectively, we obtain

$$Y_\alpha^{(k,a)} = \frac{N}{\pi^k} \int_{\mathbb{C}^k} \Big[\prod_{j=1}^k (\partial_{\bar{z}} (f_j)_{\mathbb{C}})(z_j) \Big] \Big(\langle G(z_1) \dots G(z_k) A_k \rangle - \mathbb{E} \langle G(z_1) \dots G(z_k) A_k \rangle \Big) \mathrm{d}^2 z_{[k+\ell]} \tag{2.102}$$

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

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

We start by showing that the contribution from the regime $|\eta_j| \le \eta_0$ for some $j \in [k]$ to the integral (2.102) is negligible. W.l.o.g. assume that $|\eta_j| \le \eta_0$ only for the single index $j = 1$. The general case is similar and yields an even smaller bound (cf. proof of [39, Thm. 2.6]). Our key tool is the following variant of Stokes' theorem

$$\int_{-10}^{10} \int_{\widetilde{\eta}}^{10} \partial_{\bar{z}} \Phi(x + \mathrm{i}\eta) h(x + \mathrm{i}\eta) \mathrm{d}x \mathrm{d}\eta = \frac{1}{2\mathrm{i}} \int_{-10}^{10} \Phi(x + \mathrm{i}\widetilde{\eta}) h(x + \mathrm{i}\widetilde{\eta}) \mathrm{d}x, \tag{2.103}$$

which holds for any $\widetilde{\eta} \in [0, 10]$, 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 (2.103) repeatedly for the variables $z_2, \dots, z_k$ and introducing the interval notation $\mathrm{d}x_{[i,j]} = \mathrm{d}x_i \mathrm{d}x_{i+1} \dots \mathrm{d}x_j$ as well as $\mathrm{d}\eta_{[i,j]} = \mathrm{d}\eta_i \mathrm{d}\eta_{i+1} \dots \mathrm{d}\eta_j$ for $i < j$, we obtain

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

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

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

together with (2.100) implies that

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

as the x_j-integral of $(f_j)_{\mathbb{C}}(x_j + \mathrm{i}\eta_0)$ for $j \in [2, k]$ is of order one due to Assumption 2.2.1. For I_2, we use the bound

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

from [37, 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 $\rho(x + \mathrm{i}0) = \rho_{sc}(x)$. This yields

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

Overall, we conclude that

$$\begin{aligned} Y_\alpha^{(k,a)} = \frac{N}{\pi^k} \int_{\mathbb{R}^k} \mathrm{d}x_{[k]} \int_{[\eta_0,10]^k} \mathrm{d}\eta_{[k]} \Big[\prod_{j=1}^k (\partial_{\overline{z}} (f_j)_{\mathbb{C}})(z_j) \Big] \\ \times \Big(\langle G(z_1) A_1 \dots G(z_k) A_k \rangle - \mathbb{E} \langle G(z_1) A_1 \dots G(z_k) A_k \rangle \Big) \\ + \mathcal{O}_\prec \Big(\eta_0 (N\eta_0)^{k/2} \max_j \|f_j\|_{H^p} \Big) \end{aligned} \tag{2.104}$$

Equation (2.104) now reduces the proof of the multi-point functional CLT for general $f_1, \dots, f_k$ to the CLT for resolvents in Theorem 2.3.6.

It remains to compute the covariance structure (2.28). By (2.104) and Lemma 2.2.5, we have

$$\begin{aligned} &\mathbb{E}\Big(Y_\alpha^{(k,a)} Y_\beta^{(\ell,b)} \Big) \\ &= \frac{1}{\pi^{k+\ell}} \int_{\mathbb{R}^k} \mathrm{d}x_{[k]} \int_{[\eta_0,10]^k} \mathrm{d}\eta_{[k]} \Big[\prod_{i=1}^k (\partial_{\overline{z}} (f_i)_{\mathbb{C},p})(z_i) \Big] \int_{\mathbb{R}^\ell} \mathrm{d}x_{[k+1,k+\ell]} \int_{[\eta_0,10]^\ell} \mathrm{d}\eta_{[k+1,k+\ell]} \\ &\quad \times \Big[\prod_{j=k+1}^{\ell} (\partial_{\overline{z}} (f_j)_{\mathbb{C},q})(z_j) \Big] \mathfrak{m}_2[G(z_1)A_1, \dots, G(z_k)A_k | G(z_{k+1})A_{k+1}, \dots, G(z_{k+\ell})A_{k+\ell}] \\ &\quad + \mathcal{O}_\prec \Big(\frac{N^\varepsilon \max_{i\in[k]} \|f_i\|_{H^p} \max_{j\in[k+1,k+\ell]} \|f_j\|_{H^q}}{\sqrt{N}} \Big) \end{aligned} \tag{2.105}$$

where we estimated the error coming from Lemma 2.2.5 as

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

More precisely, we considered the regime $\eta_1 \le \dots \le \eta_k$ and $\eta_{k+1} \le \dots \le \eta_{k+\ell}$, as all other regimes give the same contribution by symmetry, and applied (2.103) for the variables $i \in [2, k]$ and $j \in [k+2, k+\ell]$. Estimating the remaining $\partial_{\overline{z}} (f_1)_{\mathbb{C},p}(z_1)$ and $\partial_{\overline{z}} (f_{k+1})_{\mathbb{C},q}(z_{k+1})$ using (2.100), we obtain a bound of order $|\eta_0|^{p+q}$. Recall that applying Lemma 2.2.5 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. □

2.4.6 Proof of Corollary 2.2.9 (Limiting Covariance in Theorem 2.2.7)

Equation (2.33) is immediate from the explicit formula for $\mathfrak{m}_{GUE}[\cdot|\cdot]$ in [123, Thm. 2.4]. Moreover, the proof of (2.34) identical to [39, Lem. 4.1], as the integral defining Φ_π

only involves first-order free cumulants. We hence only focus on the proof of (2.36). Abbreviate $U = U_1 \cup U_2$ and note that

$$\begin{aligned}
&\Phi_{\pi_1\times\pi_2,U_1\times U_2}(f_1,\dots,f_{k+\ell}) \\
&= \frac{1}{\pi^{k+\ell}}\Big(\int_{\mathbb{R}^{|U|}}\int_{[\eta_0,10]^{|U|}}\Big[\prod_{j\in U}(\partial_{\overline{z}}(f_j)_{\mathbb{C}})(z_j)\Big]m_{\circ\circ}[U_1|U_2]\prod_{j\in U}\mathrm{d}^2z_j\Big) \\
&\quad\times\prod_{\substack{B\in\pi_1\cup\pi_2\\ \text{not marked}}}\Big(\int_{\mathbb{R}^{|B|}}\int_{[\eta_0,10]^{|B|}}\Big[\prod_{j\in B}(\partial_{\overline{z}}(f_j)_{\mathbb{C}})(z_j)\Big]m_{\circ}[B]\prod_{j\in B}\mathrm{d}^2z_j\Big),
\end{aligned}$$

where we set again $\mathrm{d}^2z_j = \mathrm{d}x_j\mathrm{d}\eta_j$ for $z_j = x_j + \mathrm{i}\eta_j$. In particular, the product in the last line can again be evaluated using [39, Lem. 4.1]. It remains to compute the integral involving $m_{\circ\circ}[U_1|U_2]$. We claim that

$$\begin{aligned}
&\frac{1}{\pi^{|U|}}\int_{\mathbb{R}^{|U|}}\int_{[\eta_0,10]^{|U|}}\Big[\prod_{j\in U}(\partial_{\overline{z}}(f_j)_{\mathbb{C}})(z_j)\Big]m[U_1|U_2]\prod_{j\in U}\mathrm{d}^2z_j \\
&= \mathrm{sc}[i_1,\dots,i_n|i_{n+1},\dots,i_{n+m}] + \mathcal{O}\Big(\eta_0^2(N\eta_0)^{(|U|)/2}\max_{i\in[k]}\|f_i\|_{H^p}\max_{j\in[k+1,k+\ell]}\|f_j\|_{H^q}\Big) \quad (2.106)
\end{aligned}$$

where $U_1 = \{i_1,\dots,i_n\}$ and $U_2 = \{i_{n+1},\dots,i_{n+m}\}$. The corresponding result for $\mathrm{sc}_{\circ\circ}[\cdot|\cdot]$ is then immediate from the second-order moment-cumulant relation (2.32).

To establish (2.106), we use the explicit integral representation for $\widetilde{m}[\cdot|\cdot]$ from Corollary 2.3.8 and rewrite the resulting multi-integral involving the kernel (2.38). 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 (2.100) for the almost analytic extension, we have

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

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 $\nu(\mathrm{d}x,\mathrm{d}y) = u(x,y)\mathrm{d}x\mathrm{d}y$ (cf. Remark to Corollary 2.3.8). Hence, Fubini's theorem allows interchanging the order of integration and move the x and y integrations inside. A brief calculation using (2.103) yields

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

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

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

using integration by parts. Hence,

$$\begin{aligned}
&\frac{1}{\pi^{|U|}}\int_{\mathbb{R}^{|U|}}\int_{[\eta_0,10]^{|U|}}\Big[\prod_{j\in U}(\partial_{\overline{z}}(f_j)_{\mathbb{C}})(z_j)\Big] \\
&\quad\times\Big(\sum_{i\in U_1}\frac{1}{(x-z_i)^2}\cdot\prod_{j\neq i}\frac{1}{x-z_j}\Big)\Big(\sum_{i\in U_2}\frac{1}{(y-z_i)^2}\cdot\prod_{j\neq i}\frac{1}{y-z_j}\Big)\Big[\prod_{j\in U}\mathrm{d}^2z_j\Big]
\end{aligned}$$

$$= \Big(\sum_{i\in U_1} f_i'(x) \cdot \prod_{j\neq i} f_j(x) \Big) \Big(\sum_{i\in U_2} f_i'(y) \cdot \prod_{j\neq i} f_j(y) \Big)$$
$$+ \mathcal{O}\Big(\eta_0^2 (N\eta_0)^{(|U|)/2} \max_{i\in[k]} \|f_i\|_{H^p} \max_{j\in[k+1,k+\ell]} \|f_j\|_{H^q} \Big)$$

such that (2.106) 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 (2.106), which concludes the proof of (2.36). □

2.4.7 Proof of Corollary 2.2.12 (Application to Thermalization)

Throughout the proof, we set $\widetilde{f}_j(x) = \mathrm{e}^{\mathrm{i}t_j x}\chi(x)$ 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 $\widetilde{f}_j(W)$ and $f_j(W) = \mathrm{e}^{\mathrm{i}t_j W}$ 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 [37, Cor. 2.7] and Theorem 2.2.7, 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 2.2.9. Observing that

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

we set $f_1(x) = \mathrm{e}^{\mathrm{i}tx}$, $f_2(x) = \mathrm{e}^{-\mathrm{i}tx}$, $f_3(x) = \mathrm{e}^{\mathrm{i}tx}$, and $f_4(x) = \mathrm{e}^{-\mathrm{i}tx}$ as well as $A_3 = A_1^*$, and $A_4 = A_2^*$ to apply (2.33). As A_1 and A_2 are assumed to be traceless, only the terms corresponding to the permutations

$$\pi \in \{(14)(23), (13)(24), (1)(24)(3), (14)(2)(3), (1)(24)(3), (13)(2)(4)\} \subset \overrightarrow{NCP}(2,2) \tag{2.107}$$

as well as the terms corresponding to the marked partitions

$$\pi \in \Big\{ \big\{\{\underline{1}\},\{2\}\big\} \times \big\{\{\underline{3}\},\{4\}\big\}, \big\{\{\underline{1}\},\{2\}\big\} \times \big\{\{3\},\{\underline{4}\}\big\}, \big\{\{1\},\{\underline{2}\}\big\} \times \big\{\{\underline{3}\},\{4\}\big\},$$
$$\big\{\{1\},\{\underline{2}\}\big\} \times \big\{\{3\},\{\underline{4}\}\big\} \Big\} \tag{2.108}$$

contribute to the limiting variance of $\langle A_1(t) A_2 \rangle$. Note that the marked blocks in (2.108) are distinguished by underlining.

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

$$\int_{-2}^{2} \mathrm{e}^{\mathrm{i}tx} \rho_{sc}(x) \mathrm{d}x = \frac{J_1(2t)}{t}$$

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

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

In particular,

$$\mathrm{sc}_\circ[1] = \mathrm{sc}[1] = \frac{J_1(2t)}{t} = \mathcal{O}\Big(\frac{1}{t^{3/2}}\Big), \quad t \gg 1.$$

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 $\overrightarrow{NCP}(2,2)$ for large t, giving

$$\mathrm{sc}_\circ[1,4]\mathrm{sc}_\circ[2,3] = \Big(1 - \frac{J_1(2t)J_1(2t)}{t^2}\Big)\Big(1 - \frac{J_1(2t)J_1(2t)}{t^2}\Big) = 1 + \mathcal{O}\Big(\frac{1}{t^3}\Big), \quad t \gg 1,$$

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

$$\mathrm{sc}_\circ[1]\mathrm{sc}_\circ[2,3]\mathrm{sc}_\circ[4] = \mathcal{O}\Big(\frac{1}{t^3}\Big), \quad t \gg 1,$$

with the remaining permutations in (2.107) yielding contributions of comparable or lower order in the $t \to \infty$ limit.

Lastly, we consider the marked partitions in (2.108), which correspond to the term of $\mathrm{Var}[\xi]$ that contains $\langle A_1A_2\rangle\langle A_3A_4\rangle = |\langle A_1A_2\rangle|^2$. As we work in the macroscopic regime of Theorem 2.2.7, Equation (2.37) coincides with the limiting covariance structure of the CLT in [104, Sect. 2]. Hence, we obtain, e.g.,

$$\mathrm{sc}[1|4] = \frac{1}{2\pi^2}\int_{-2}^{2}\int_{-2}^{2}\frac{1-\cos(t(x-y))}{(x-y)^2}\frac{4-xy}{\sqrt{4-x^2}\sqrt{4-y^2}}\mathrm{d}x\mathrm{d}y. \tag{2.109}$$

In particular, the cutoff χ does not enter the computation. Note that the expressions for $\mathrm{sc}[1|4]$ and $\mathrm{sc}[2|3]$ resp. $\mathrm{sc}[1|3]$ and $\mathrm{sc}[2|4]$ yield the same contribution by symmetry. The integral on the right-hand side of (2.109) is finite, however, it will grow with t as $t \to \infty$. 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

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

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 $\mathrm{sc}[1|4] = \mathcal{O}(t)$. Recalling the identity $\mathrm{sc}_{\circ\circ}[1|4] = \mathrm{sc}[1|4] - \mathrm{sc}_\circ[1,4]$ from (2.32), we obtain

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

The other marked partitions in 2.108 give rise to terms of comparable order. Summing up all contributions yields (2.40). The proof of (2.41) is analogous and hence omitted. □

2.A Proof of Corollary 2.3.3 (Meta Argument)

Recall from Lemma 2.2.3 that

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

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

$$N(\mathbb{E}\langle T_1\dots T_k\rangle - \mathfrak{m}_1[T_1,\dots,T_k]) = \kappa_4\mathcal{E}[T_1,\dots,T_k] + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N\eta_*}\,\eta_*^{k-a/2}}\Big), \tag{2.110}$$

where the quantity on the left-hand side of (2.110) satisfies the properties stated in Corollary 2.3.3. 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 $\mathfrak{m}_1[\cdot]$ follow from (2.17) and [39, Lem. 5.4]. Note that (2.46) is a special case of (2.45) and that (2.47) is obtained by iterating (2.46). Once the formula (2.47) is established, the permutation symmetry readily follows from the divided difference structure. Hence, Corollary 2.3.3(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 (2.110). Note that simply taking the $N \to \infty$ limit and applying Lemma 2.2.3 is not sufficient if $A_j \neq \mathrm{Id}$ 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

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

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

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

Next, let $\mathscr{G}_j := (\mathscr{W} - z_j)^{-1}$ and $\mathscr{T}_j := \mathscr{G}_j \mathscr{A}_j$ for $j = 1, \dots, k$ and denote by $\mathscr{m}_1[\mathscr{T}_1, \dots, \mathscr{T}_k]$ the deterministic approximation of $\langle \mathscr{T}_1 \dots \mathscr{T}_k\rangle$. As both $\mathscr{W}$ and W are Wigner matrices and $\langle \mathscr{A}_i \mathscr{A}_j\rangle = \langle (A_i A_j) \otimes \mathrm{Id}_{L\times L}\rangle = \langle A_i A_j\rangle$ by definition of the tensor product, it follows from the closed form of $\mathfrak{m}_1[\cdot]$ in [39, Thm. 2.6] that

$$\mathscr{m}_1[\mathscr{T}_1, \dots, \mathscr{T}_k] = \mathfrak{m}_1[T_1, \dots, T_k]. \tag{2.111}$$

Similarly, the $1/N$ error term of $\mathbb{E}\langle \mathscr{T}_1 \dots \mathscr{T}_k\rangle$ is given by $\mathcal{E}[T_1, \dots, T_k]$, since (2.111) ensures that we obtain the same recursion from Definition 2.3.1.

We thus conclude that

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

by Lemma 2.2.3 for any $\eta_* \gg (NL)^{-1}$. Letting $L \to \infty$ 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 (a) of Corollary 2.3.3. Similarly, we obtain that

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

whenever $z_1 \neq z_k$ and $A_k = \mathrm{Id}$, which implies (2.45). Recalling that (i) and (ii) of Corollary 2.3.3 imply (iii), the proof is complete. □

3 Fluctuation Moments for Regular Functions of Wigner Matrices

We compute the deterministic approximation for mixed fluctuation moments of products of deterministic matrices and general Sobolev functions of Wigner matrices. Restricting to polynomials, our formulas reproduce recent results of Male, Mingo, Peché, and Speicher [107], showing that the underlying combinatorics of non-crossing partitions and annular non-crossing permutations continue to stay valid beyond the setting of second-order free probability theory. The formulas obtained further characterize the variance in the functional central limit theorem given in the recent companion paper [122] and thus allow identifying the fluctuation around the thermal value in certain thermalization problems.

Published as J. Reker. Fluctuation moments for regular functions of Wigner matrices. *Math. Phys. Anal. Geom.*, 27:10, 2024.

3.1 Introduction

In his seminal work [140], Wigner established that the empirical spectral measure of certain random matrix ensembles converges, as the dimension goes to infinity, to the semicircle distribution. Since then, many variations and extensions of this result have been considered, yielding a variety of asymptotic phenomena for a wide range of random matrix models. One particular example is the fact that the resolvent $G(z) = (W - z)^{-1}$ of a large Hermitian random matrix W tends to concentrate around a deterministic matrix $M = M(z)$ for spectral parameters $z \in \mathbb{C}$ even just slightly away from the real axis (see, e.g., [37] and references therein for a collection of recent results). It was recently shown (see [39, 37]) that a similar concentration holds for alternating products of the form

$$F_{[1,k]} := f_1(W)A_1 \dots f_k(W)A_k. \tag{3.1}$$

Here, $A_1, \dots, A_k$ are bounded deterministic matrices and $f_1, \dots, f_k$ are regular test functions, allowing in particular for $f_j(W) = G(z_j)$. Apart from resolvents, two choices of f_j are of special interest in this setting. First, consider the case $f_j(x) = \mathrm{e}^{\mathrm{i}t_j x}$ with $t_j \in \mathbb{R}$. Interpreting the Wigner matrix W as the Hamiltonian of a mean-field quantum system and the deterministic bounded matrix A as an observable, the quantity

$$A(t) := \mathrm{e}^{\mathrm{i}tW} A \mathrm{e}^{-\mathrm{i}tW}$$

describes the Heisenberg time evolution of A. For $k \geq 2$, (3.1) hence describes the time evolution of multiple observables, possibly to different times, in the same quantum system. Considering the regime where the differences between the individual times are large links the matrix product to thermalization problems in mathematical physics (see [39, Sect. 2.1] and [37, Rem. 2.8]). Second, products of the form (3.1) with $f_j(W)$ replaced by (polynomials of) the random matrix itself play a key role in free probability theory,

as they characterize the joint non-commutative probability distribution of Wigner and deterministic matrices.

We remind the reader that a (tracial *first-order*) non-commutative probability space is a pair $(\mathcal{A}, \varphi_1)$ consisting of a complex unital algebra $\mathcal{A}$ and a tracial linear functional $\varphi_1 : \mathcal{A} \to \mathbb{C}$ with $\varphi_1(1_\mathcal{A}) = 1$, where $1_\mathcal{A}$ is the unit element of the algebra. One particular example is the space $(\mathcal{A}, \varphi_1) = (\mathcal{M}_{N\times N}(L^{\infty-}(\Omega, \mathbb{P})), \mathbb{E}\langle\cdot\rangle)$ of $N \times N$ random matrices, where $(\Omega, \mathbb{P})$ is a classical probability space, $\mathcal{M}_{N\times N}(S)$ denotes the $N \times N$-matrices with entries in S, the space

$$L^{\infty-}(\Omega, \mathbb{P}) := \bigcap_{1\leq p<\infty} L^p(\Omega, \mathbb{P})$$

contains all random variables with all finite moments, and $\langle\cdot\rangle$ denotes the normalized trace. Note that this definition includes deterministic and Wigner matrices. In this context, the non-commutative probability distribution of $a \in \mathcal{A}$ is characterized in terms of its moments $(\varphi_1(a^k))_k$ with the joint distribution of multiple elements of $\mathcal{A}$ being defined analogously. Recent work by Cipolloni, Erdős and Schröder [39] established that the structure of the limit of $\mathbb{E}\langle F_{[1,k]}\rangle$ as in (3.1) matches the formulas obtained in free probability, and reproduces known results 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., [111, Sect. 4.4]) in the case $f_j(x) = x$. More precisely, in the large N limit, the leading-order term $\mathfrak{m}_1[F_{[1,k]}]$ of $\mathbb{E}\langle F_{[1,k]}\rangle$ is of the form

$$\mathfrak{m}_1[F_{[1,k]}] := \sum_{\pi\in NCP([k])} \Big(\prod_{B\in\pi} \Big\langle \prod_{j\in B} A_j \Big\rangle \Big) \Phi_\pi^{(1)}(f_1, \dots, f_k), \tag{3.2}$$

where $NCP([k])$ denotes the non-crossing partitions of the cyclically ordered set $\{1, \dots, k\}$ and the functions $\Phi_\pi^{(1)}$ only depend on $f_1, \dots, f_k$ and $\pi \in NCP([k])$. Hence, the right-hand side of (3.2) is a sum of terms that factorize into a contribution of the deterministic matrices resp. the test functions appearing in the product (3.1) with the underlying combinatorics matching the results obtained for the case $f_j(x) = x$ in free probability theory. Note, however, that resolvents and functions with an N-dependent mesoscopic scaling are typically not accessible in free probability as many of the standard techniques rely on explicit moment computations for polynomials. The results in [39] thus show that the underlying combinatorics continue to apply in a more general context. In particular, evaluating (3.2) for $f_j(x) = \mathrm{e}^{\mathrm{i}t_j x}$ with $t_j \in \mathbb{R}$ implies a closed formula for the thermal value of quantities such as $\langle \mathrm{e}^{\mathrm{i}tW} A_1 \mathrm{e}^{-\mathrm{i}tW} A_2\rangle$, which allows for a direct analysis of the $t \gg 1$ regime relevant for thermalization (see, e.g., [39, Cor. 2.9]).

After considering the concentration of (3.1), the next natural step is to study the fluctuations around the deterministic value. It is well-known that the linear statistics $\operatorname{Tr} f(W) = \sum_{j=1}^N f(\lambda_j)$ with a regular test function $f : \mathbb{R} \to \mathbb{R}$ have a variance of order one (first observed in [95]) and, in fact, satisfy a central limit theorem (CLT) with a Gaussian limit, as shown, e.g., in [96] for the Wigner case and in [93] for invariant ensembles. By now, the statistics $\operatorname{Tr} f(W)$ are well-studied on both macroscopic and mesoscopic scales (see, e.g., [85, 12, 105, 104, 128, 135, 27, 129, 89, 17, 99] for the Wigner case and [40, 122] for further references on previous results for Wigner matrices and other models). However, while the fluctuations of $\operatorname{Tr}[f(W)A]$ are known for general regular functions f (see [103] and [40]), traces of products of the form (3.1) for $k \geq 2$ have so far only been studied for f_j being polynomials in the context of second-order freeness (see, e.g., [111, Ch. 5] or [42, 106, 107]).

We remind the reader that a *second-order* non-commutative probability space is a triplet $(\mathcal{A}, \varphi_1, \varphi_2)$, where the functional $\varphi_2 : \mathcal{A} \times \mathcal{A} \to \mathbb{C}$ is bilinear, tracial in both arguments, symmetric under the interchanging of its arguments, and satisfies $\varphi_2(a, 1_\mathcal{A}) = \varphi_2(1_\mathcal{A}, a) = 0$ for all $a \in \mathcal{A}$. The second-order probability distribution of $a \in \mathcal{A}$ is characterized in terms of $(\varphi_2(a^k, a^\ell))_{k,\ell}$, called the *fluctuation moments*, with the joint moments of multiple elements again being defined analogously. As a canonical example, we remark that $\mathcal{M}_{N\times N}(L^{\infty-}(\Omega, \mathbb{P}))$ may be endowed with the functional $\varphi_2(\cdot,\cdot) = \mathrm{Cov}(\mathrm{Tr}(\cdot), \mathrm{Tr}(\cdot))$, to make it a second-order probability space. In contrast to the first-order structure, the fluctuation moments are sensitive to the symmetry class of the underlying Wigner matrix and explicitly involve the fourth cumulant of the entry distribution (see [107, Thm. 6], as well as [118, 119, 40]). In particular, we observe a breaking of universality compared to the first-order problem of computing $\mathbb{E}\langle\cdot\rangle$. The joint fluctuation moments of Wigner and deterministic matrices are explicitly known (cf. [111, Thm. 13 of Ch. 5] for the GUE case and [107, Thm. 6] for general Wigner matrices).

A functional CLT for traces of products of the form (3.1) has recently been established in the companion paper [122] and the limiting covariance is derived using a recursion. In the present paper, we supply the combinatorial argument necessary to obtain the solution to the recursion and compute the limiting covariance explicitly. In particular, we show that if W is a GUE matrix, the leading order term $\mathfrak{m}_2[F_{[1,k]}|F_{[k+1,k+\ell]}]$ of the covariance of $\mathrm{Tr}(F_{[1,k]})$ and $\mathrm{Tr}(F_{[k+1,k+\ell]})$ (with $F_{[k+1,k+\ell]} = f_{k+1}A_{k+1}\dots f_{k+\ell}A_{k+\ell}$ of the same build as (3.1)) is given by

$$\begin{aligned}\mathfrak{m}_2[F_{[1,k]}|F_{[k+1,k+\ell]}] &= \sum_{\pi\in\overrightarrow{NCP}(k,\ell)} \Big(\prod_{B\in\pi}\Big\langle\prod_{j\in B}A_j\Big\rangle\Big)\Phi^{(2)}_\pi(f_1,\dots,f_{k+\ell}) \\ &+ \sum_{\pi_1\times\pi_2\in NCP(k)\times NCP(\ell)} \Big(\prod_{B_1\in\pi_1,B_2\in\pi_2}\Big\langle\prod_{j\in B_1}A_j\Big\rangle\Big\langle\prod_{j\in B_2}A_j\Big\rangle\Big)\Phi^{(2)}_{\pi_1\times\pi_2}(f_1,\dots,f_{k+\ell}).\end{aligned} \tag{3.3}$$

Here, $\overrightarrow{NCP}(k,\ell)$ denotes the non-crossing permutations of the (k,ℓ)-annulus and the functions $\Phi^{(2)}_\pi$ resp. $\Phi^{(2)}_{\pi_1\times\pi_2}$ only depend on $f_1,\dots,f_{k+\ell}$ and the underlying permutation resp. partition. Similar to (3.2), we thus obtain a sum of terms that factorize into a contribution of the deterministic matrices resp. the test functions appearing in the product (3.1) with the underlying combinatorics again matching the results obtained for the case $f_j(x) = x$ in free probability theory (see [113]). Moreover, we show that the overall structure of (3.3) continues to hold if W is chosen to be a Wigner matrix with $W_{ij} \overset{d}{=} N^{-1/2}\chi_{od}$ for $i < j$ and $W_{jj} \overset{d}{=} N^{-1/2}\chi_d$ for general entry distributions χ_{od} and χ_d. In the general case, however, the sum in the first line of the right-hand side of (3.3) splits into four summands $\Phi^{(GUE)}_\pi$, $\kappa_4\Phi^{(\kappa)}_\pi$, $\sigma\Phi^{(\sigma)}_\pi$, and $\widetilde{\omega}_2\Phi^{(\omega)}_\pi$ which have different prefactors in terms of the deterministic matrices $A_1,\dots,A_{k+\ell}$. Here, $\Phi^{(GUE)}_\pi$ corresponds to the GUE case in (3.3) and the remaining contributions are associated with the parameters

$$\kappa_4 = \mathbb{E}|\chi_{od}|^4 - 2, \quad \sigma = \mathbb{E}\chi_{od}^2, \quad \widetilde{\omega}_2 = \mathbb{E}\chi_d^2 - 1 - \sigma \tag{3.4}$$

of the Wigner matrix W. A similar decomposition is also observed for $\Phi^{(2)}_{\pi_1\times\pi_2}$ in (3.3). In particular, we find that the closed expression obtained from solving the recursion in [122] has the same overall structure as the formulas in [107, Thm. 6]. This shows that the analogies [39] established in the first-order setting have a counterpart for the second-order structures. Our combinatorial approach further allows us to give the functions in (3.3) in a closed form, thus yielding a fully explicit formula for the limiting covariance in the GUE case.

We remark that the main results of the present paper, i.e., combinatorial formulas for $\mathfrak{m}_2[F_{[1,k]}|F_{[k+1,k+\ell]}]$ such as (3.3), are applied to obtain an explicit limiting covariance structure for the multi-point functional CLT [122, Thm. 2.7]. Here, replacing the recursive definition of the limiting variance by a closed formula allows for an easier application of the theorem, e.g., to thermalization problems in physics. Although the form of the limiting variance in the $k = 2$ case considered in [122, Cor. 2.12] is sufficiently simple and could be derived without additional combinatorial tools, characterizing the fluctuation for general $k \geq 2$ does require them to obtain an explicit formula for the variance. We further emphasize that the main results in the companion paper [122] are of analytic nature and that their main technical difficulty lies in including functions with a mesoscopic scaling of the form

$$f_j(x) = g_j(N^\gamma(x - E)) \tag{3.5}$$

where g_j is a regular N-independent function, $E \in \mathbb{R}$ lies in the bulk of the limiting spectrum of W, and $N^{-\gamma}$ is larger than the typical eigenvalue spacing around E. In contrast, we assume all test functions to be N-independent in the present paper and focus on the combinatorial structures arising in the multi-point functional CLT. While an extension to the functions in (3.5) is possible using the techniques from [122], restricting to the macroscopic regime allows for a cleaner presentation of the results. It further facilitates working with more general assumptions on the Wigner matrix W. Note that [122, As. 1.1] corresponds to setting $\sigma = \widetilde{\omega}_2 = 0$ in (3.4), while Assumption 3.1.1 below matches the setting of [107, 40] with general $\sigma \in [-1, 1]$ and $\widetilde{\omega}_2 \geq -2$, thus generalizing the formulas from [122].

We conclude the section with a brief overview of the paper. After introducing some commonly used notations, the assumptions on the Wigner matrix W are given in Assumption 3.1.1. We then give a brief overview of the combinatorics needed to identify the deterministic approximation of $\langle T_1 \dots T_k \rangle$ where $T_j := G(z_j)A_j$, and the multi-resolvent local laws needed for the analysis of the fluctuations (Section 3.1.2) as well as the definitions from free probability that are used to characterize the limiting covariance of $\langle T_1 \dots T_k \rangle - \mathbb{E}\langle T_1 \dots T_k \rangle$ and $\langle T_{k+1} \dots T_{k+\ell} \rangle - \mathbb{E}\langle T_{k+1} \dots T_{k+\ell} \rangle$ (Section 3.1.3). To prepare for the statements of our main results, we give a CLT for the case that all functions f_j are resolvents (Theorem 3.2.3). The role of the limiting covariance in the theorem is played by a recursively defined set function $\mathfrak{m}_2[\cdot|\cdot]$ (Definition 3.2.1), which is the main object of interest in the present paper. We study the recursion in detail in Section 3.2.2 and obtain the main results, explicit combinatorial formulas for its solution (Theorems 3.2.4, 3.2.6 to 3.2.8). In Section 3.2.3, we extend the CLT to more general test functions (Theorem 3.2.9 and Corollary 3.2.10) to discuss the connection to free probability theory in detail. In particular, we apply the results to the case $f_j(x) = x$ and show that the limiting covariance in the functional CLT reduces to the formula for the joint fluctuation moments of GUE and deterministic matrices (Corollary 3.2.11) as given in [107]. Lastly, the proofs are given in Sections 3.3 and 3.4. To keep the presentation concise, some routine calculations are deferred to the appendix. This includes the proof of a multi-resolvent global law with transposes (Appendix 3.A.1) and the proof of the resolvent CLT in Theorem 3.2.3 (Appendix 3.A.2) which are similar to previous results in [39] and [122], respectively.

Acknowledgements: I am very grateful to László Erdős for suggesting the topic and many valuable discussions during my work on the project. Partially supported by ERC Advanced Grant "RMTBeyond" No. 101020331.

3.1.1 General Notation

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 $\overline{z}$. Moreover, $\|\cdot\|$ denotes the operator norm, $\mathrm{Tr}(\cdot)$ is the usual trace and $\langle\cdot\rangle = N^{-1}\,\mathrm{Tr}(\cdot)$. We further denote the covariance of two complex random variables X_1, X_2 by $\mathrm{Cov}(X_1, X_2)$ and follow the convention

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

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 $[k] = \{1, \ldots, k\}$ and adopt the interval notation $[a, b] = \{a, a+1, \ldots, b\}$. We further write $\langle a, b]$ or $[a, b\rangle$ to indicate that a or b are excluded from the interval, respectively. Ordered sets are denoted by $(\ldots)$ instead of $\{\ldots\}$. Sets of the form $[k]$ and intervals are treated as ordered sets.

Given a matrix $A \in \mathbb{C}^{N\times N}$, the traceless part of A is denoted by $\mathring{A} := A - \langle A\rangle \mathrm{Id}$ where Id denotes the identity matrix. Further, $\mathbf{a} := \mathrm{diag}A$ 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 spectral parameters $z_1, \ldots, z_k \in \mathbb{C} \setminus \mathbb{R}$, we write the corresponding resolvents as $G_j = G(z_j) := (W - z_j)^{-1}$ and index products of resolvents using the interval notation

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

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 $G_{\langle a,b]}$ and $G_{[a,b\rangle}$ 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 [39, 37]. As we often consider alternating products of resolvents with deterministic matrices $A_1, \ldots, A_k$, define $T_j := G_j A_j$ and apply the same interval notation as above to write

$$T_{[k]} := T_1 \ldots T_k = G_1 A_1 \ldots G_k A_k, \quad T_{[a,b]} := T_a T_{a+1} \ldots T_b. \tag{3.6}$$

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 (3.6) *resolvent chain* of length k.

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

Assumption 3.1.1. *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 real or 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 $\mathbb{E}|\chi_{od}|^2 = 1$ 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*

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

We remark that Assumption 3.1.1 matches the model considered in [40] and [107]. Compared to the conditions $\mathbb{E}\chi_{od}^2 = 0$ and $\mathbb{E}\chi_d^2 = 1$ in [122], we allow for arbitrary values of

the parameters $\sigma = \mathbb{E}\chi_{od}^2 \in [-1,1]$ and $\omega_2 = \mathbb{E}\chi_d^2 \geq 0$. This description includes real symmetric Wigner ensembles such as GOE ($\sigma = 1$) as well as matrices of the form $W = D + \mathrm{i}S$ where D is a diagonal matrix and S is skew-symmetric ($\sigma = -1$). We further introduce the notation

$$\kappa_4 := \mathbb{E}|\chi_{od}|^4 - 2 \tag{3.7}$$

for the normalized fourth cumulant of the off-diagonal entries as well as

$$\widetilde{\omega_2} := \omega_2 - 1 - \sigma. \tag{3.8}$$

The eigenvalue density profile of W is described by the semicircle law

$$\rho_{sc}(x) := \frac{\sqrt{4-x^2}}{2\pi} \mathbb{1}_{[-2,2]}(x) \tag{3.9}$$

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

$$m(z) := \int \frac{\rho_{sc}(x)}{x-z} \mathrm{d}x, \quad z \in \mathbb{C} \setminus \mathbb{R}. \tag{3.10}$$

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

$$-\frac{1}{m(z)} = m(z) + z, \quad \Im z \Im(m(z)) > 0 \tag{3.11}$$

and that its derivative satisfies

$$m'(z) = \frac{m(z)^2}{1 - m(z)^2}. \tag{3.12}$$

Given fixed $z_1, \ldots, z_k \in \mathbb{C} \setminus \mathbb{R}$, set $m_j = m(z_j)$ and $m_j' = m'(z_j)$, respectively. We further introduce

$$q_{i,j} = \frac{m_i m_j}{1 - m_i m_j}, \tag{3.13}$$

and remark that $q_{j,j} = m_j'$ whenever $i = j$.

3.1.2 Preliminaries Part 1: First-Order Quantities

In this section, we briefly summarize the definitions and results from [37, 39] which are needed to characterize the deterministic approximation of $\langle T_{[1,k]} \rangle$.

Definition 3.1.2 (Non-crossing partitions)**.** *Let S be a finite (cyclically) ordered set of integers. We call a partition π of the set S* crossing *if there exist blocks $B \neq B'$ in π with $a, b \in B$, $c, d \in B'$, and $a < c < b < d$, otherwise we call it* non-crossing*. The set of non-crossing partitions is denoted by $NCP(S)$ and we abbreviate $NCP(k) := NCP([k])$. For each non-crossing partition $\pi = \{B_1, \ldots, B_n\}$, set $|\pi| := n$ for the number of blocks in the partition.*

Recall that non-crossing partitions have an alternative geometrical definition: Arrange the elements of S equidistantly in clockwise order on the circle and for each $\pi \in B$ consider the convex hull P_B of the points $s \in B$. Then π is non-crossing if and only if the polygons $\{P_B | B \in \pi\}$ are pair-wise disjoint. Because of this, we also call the elements of $NCP(k)$ *disk non-crossing* to distinguish them from their annulus analog defined below. We further recall the definition of the Kreweras complement (see Fig. 3.1 for an example).

Definition 3.1.3 (Kreweras complement, disk case)**.** *Let $S \subset \mathbb{N}$ be a finite set of integers equidistantly arranged in clockwise order on the circle and label the midpoints of the arcs between the points $s \in S$ also by the elements of S. We arrange the new labels such that the arc s follows the point s in clockwise order. Let $\pi \in NCP(S)$. Then the* (disk) Kreweras complement *of π, denoted by $K(\pi)$, is the element of $NCP(S)$ such that r, s belong to the same block of $K(\pi)$ if and only if the arcs labeled r, s are in the same connected component in the complement $D \setminus \cup_{B\in\pi} P_B$ of the polygons $\{P_B | B \in \pi\}$ in the labeled disk D.*

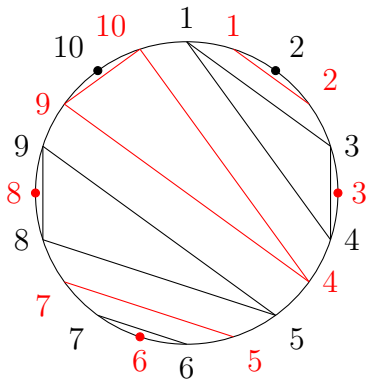

Fig. 3.1. The non-crossing partition $\pi = \{\{1, 3, 4\}, \{2\}, \{5, 8, 9\}, \{6, 7\}, \{10\}\}$ (black) and its Kreweras complement $K(\pi) = \{\{1, 2\}, \{3\}, \{4, 9, 10\}, \{5, 7\}, \{6\}, \{8\}\}$ (red). Singleton sets are indicated by a black or red dot, respectively.

Observe that $D\setminus\cup_{B\in\pi} P_B$ has $|S|-|\pi|+1$ connected components, hence $|\pi|+|K(\pi)| = |S|+1$. Further, $K^2 = K \circ K$ recovers π up to a rotation of D, i.e., $K^2(\pi)$ is the partition where for $S = \{s_1, \ldots, s_k\}$ the elements in each block of π are shifted by $s_1 \mapsto s_2 \mapsto \cdots \mapsto s_k \mapsto s_1$. In particular, taking the Kreweras complement is invertible as a map on $NCP(S)$.

Definition 3.1.4 (Free cumulant function)**.** *Fix $k \in \mathbb{N}$, denote the power set*[1] *of $[k]$ by $\mathcal{P}([k])$ and let $(f_n)_n$ be a family of functions such that f_n maps n-element sets in $\mathcal{P}([k])$ to $\mathbb{C}$. We define the* (first-order) free cumulants *$(f_{\circ,n})_n$ associated with the sequence $(f_n)_n$ through the relation*

$$f_{|S|}[S] = \sum_{\pi\in NCP(S)} \prod_{B\in\pi} f_{\circ,|B|}[B] \tag{3.14}$$

for any $S \subseteq [k]$. For simplicity. we associate the sequences $(f_n)_n$ (resp. $(f_{\circ,n})_n$) with a single function f (resp. $f_\circ$) by setting $f[s_1, \ldots, s_n] := f_n[\{s_1, \ldots, s_n\}]$ (resp. $f_\circ[s_1, \ldots, s_n] := f_{\circ,n}[\{s_1, \ldots, s_n\}]$) and call $f_\circ$ the (first-order) free cumulant function *associated to f.*

We emphasize that Definition 3.1.4 does not require the functions f_n to have any particular symmetries. However, in the free probability literature, f usually arises from a family of tracial functionals and is hence symmetric under the cyclic permutation of its entries (cf., e.g., [111, Ch. 2]). The implicit relation in (3.14) can be recursively turned into an explicit definition of $f_\circ$. Alternatively, we may also invert (3.14) explicitly using the Möbius function associated with the lattice of non-crossing partitions. Recall that $NCP(S)$ is

[1]Recall that $[k]$ is treated as an ordered set. Hence, the subsets in $\mathcal{P}([k])$ carry a natural ordering inherited from $[k]$.

a lattice with respect to the refinement order, i.e., the partial order in which $\pi \leq \nu$ if and only if for each $B \in \pi$ there exists $B' \in \nu$ with $B \subset B'$. Moreover, there are unique maximal and minimal elements given by $0_S := \{\{s\}|s \in S\}$ and $1_S := \{S\}$, respectively. The free cumulant function can then be written as

$$f_\circ[S] = \sum_{\pi \in NPC[S]} \mu(\pi, \mathbf{1}_S) \prod_{B \in \pi} f[B], \quad \mu(\pi,\nu) := \begin{cases} 1, & \pi = \nu, \\ -\sum_{\pi < \tau \leq \nu} \mu(\tau,\nu), & \pi < \nu, \end{cases} \tag{3.15}$$

using the Möbius function $\mu : \{(\pi,\nu)|\pi \leq \nu \in NCP(S)\} \to \mathbb{Z}$ that is recursively defined by (3.15). We remark that $\mu(\pi, 1_S)$ can be given in a closed form using the Catalan numbers (see, e.g., [39, Lem. 2.16]).

The following choice for the function f is of particular interest. Recall that m denotes the Stieltjes transform of the semicircle law (3.9).

Definition 3.1.5 (Divided differences)**.** *For finite sets $\{z_1, \ldots, z_k\} \subset \mathbb{C} \setminus \mathbb{R}$ we recursively define*

$$m[z_1, \ldots, z_k] := \frac{m[z_2, \ldots, z_k] - m[z_1, \ldots, z_{k-1}]}{z_k - z_1}.$$

The definition readily extends to multi-sets $\{z_1, \ldots, z_k\} \subset \mathbb{C} \setminus \mathbb{R}$ by setting

$$m[\underbrace{z_j, \ldots, z_j}_{k \text{ times}}] := \frac{m^{(k-1)}(z_j)}{(k-1)!}$$

whenever an element z_j occurs with a multiplicity greater than one. Here, $m^{(k-1)}$ denotes the $(k-1)$-th derivative of the function m in (3.10)*. Note that $m[\cdot]$ is well-defined in the sense that $m[z_1, \ldots, z_k]$ is independent of the ordering of the multi-set $\{z_1, \ldots, z_k\}$. We abbreviate $m[1, \ldots, k] := m[z_1, \ldots, z_k]$.*

We emphasize that $m[\cdot]$, and hence $m_\circ[\cdot]$, have full permutation symmetry, which is much more than what was assumed for f in Definition 3.1.4. The following example illustrates the combinatorial formulas (3.14) and (3.15) for $f = m[\cdot]$.

Example 3.1.6 (First-order free cumulants). In the case $k = 1$ we have $m[1] = m(z_1)$. For $k = 2$, the only non-crossing partitions are (12) and (1)(2) such that

$$m_\circ[1,2] = m[1,2] - m_1 m_2, \quad m_j := m[j] = m[z_j]$$

while for $k = 3$ we have

$$m_\circ[1,2,3] = m[1,2,3] - m_1 m[2,3] - m_2 m[1,3] - m_3 m[1,2] + 2m_1 m_2 m_3.$$

The quantities m and $m_\circ$ were studied in detail in [39], yielding a close connection to *non-crossing graphs*. We recall the definition and give an example in Fig. 3.2 below. These graphs are planar. For later convenience, we use a slightly more general notion of planar graphs throughout the paper than the standard literature by allowing for self-connections (loops) and multi-edges.

Definition 3.1.7 (Disk non-crossing graphs)**.** *Let $S \subset \mathbb{N}$ be a finite (cyclically) ordered set of integers equidistantly arranged in clockwise order on the circle. We call an undirected planar graph (S, E) on the vertex set S without loops or multi-edges* (disk) crossing *if there exist two edges $(a, b), (c, d) \in E$ with $a < c < b < d$, otherwise we call it* (disk) non-crossing[2]*. The set of all* (disk) non-crossing graphs *with vertex set S is denoted by $NCG(S)$ and we denote the subset of connected graphs as $NCG_c(S)$. Whenever $S = [k]$, abbreviate $NCG(k) := NCG([k])$.*

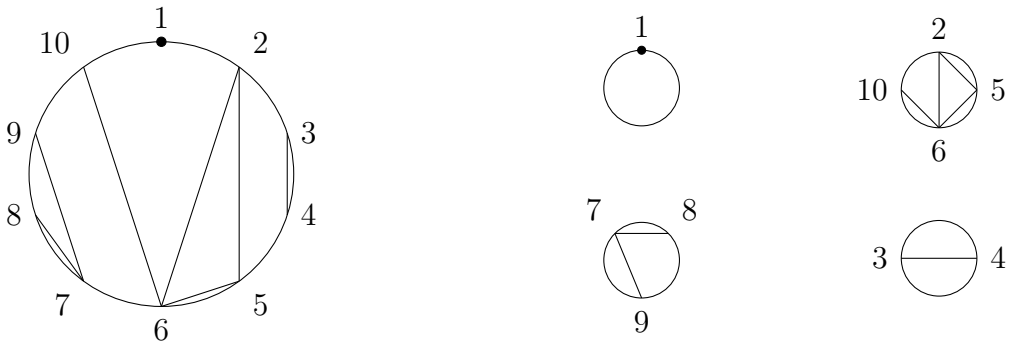

Fig. 3.2. An element of $NCG(10)$ and its connected subgraphs.

Emphasizing that Definition 3.1.7 lives on a disk is important, as we later introduce a non-crossing property on the annulus. Whenever both definitions are used together, we use the specifications *disk* non-crossing and *annular* non-crossing to distinguish the underlying geometry. By construction, every $\Gamma \in NCG(S)$ induces a non-crossing partition with blocks representing the vertices in the connected components of Γ. Further, any connected component of Γ is itself a (disk) non-crossing graph.

Lemma 5.2 of [39] proves the representations

$$m[S] = \Big(\prod_{s\in S} m_s\Big) \sum_{\Gamma\in NCG(S)} \prod_{(i,j)\in E(\Gamma)} q_{i,j}, \tag{3.16}$$

$$m_\circ[S] = \Big(\prod_{s\in S} m_s\Big) \sum_{\Gamma\in NCG_c(S)} \prod_{(i,j)\in E(\Gamma)} q_{i,j} \tag{3.17}$$

in terms of the *weights* $q_{i,j}$ in (3.13). Here, $E(\Gamma)$ is the edge set of the graph Γ. Note that (3.16) and (3.17) are still well-defined if S is an ordered multi-set, i.e., if some elements are repeated. In this case, we consider $NCG(|S|)$ instead of $NCG(S)$ and use the one-to-one correspondence between the (possibly repeated) labels $\{s|s \in S\}$ and $\{1, \ldots, |S|\}$ to obtain a uniquely defined right-hand side (see Fig. 3.3 below).

[2]The edges of a disk non-crossing graph Γ can be drawn in the interior of the disk without intersecting, i.e., Γ is a planar graph drawn inside a labeled disk.

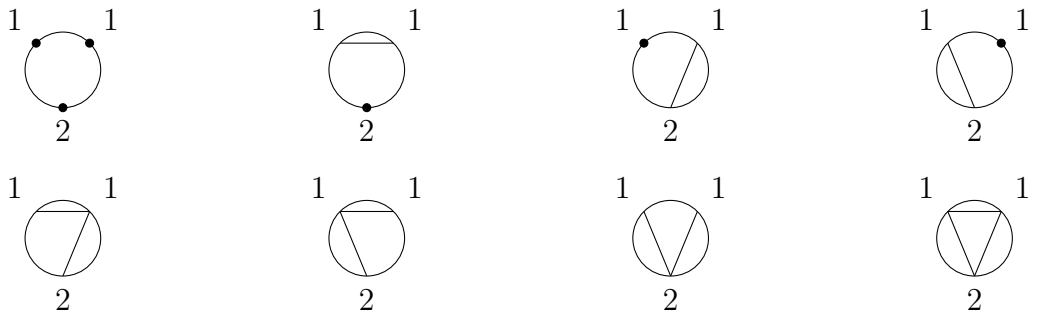

Fig. 3.3. The graphs contributing to the sum on the right-hand side of (3.16) (top and bottom row) and (3.17) (bottom row only) for the multi-set $S = (1, 1, 2)$. Note that the prefactor is given by $\prod_{s\in S} m_s = m_1^2 m_2$.

A key technical tool in the proof of our main results is the optimal multi-resolvent local law [37, Thm. 2.5]. As we only work on macroscopic scales, i.e., with N-independent spectral parameters, in the present paper, we state the result in the form of a global law and omit the dependence on $\eta_* = \min|\Im z_j|$. Recall the commonly used definition of stochastic domination.

Definition 3.1.8 (Stochastic domination)**.** *Let*

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

be two families of non-negative random variables that are indexed by N and possibly some other parameter u in some (possibly N-dependent) domain $U^{(N)}$. 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*

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

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

Theorem 3.1.9 (Macroscopic version of [37, Thm. 2.5])**.** *Fix $k \in \mathbb{N}$ and pick spectral parameters $z_1, \dots, z_k \in \mathbb{C}\setminus\mathbb{R}$ with $|\Im z_j| \gtrsim 1$ and $\max_j |z_j| \le N^{100}$ as well as deterministic matrices $A_1, \dots, A_k \in \mathbb{C}^{N\times N}$ with $\|A_i\| \lesssim 1$. Define*[3]

$$M_{[k]} := \sum_{\pi\in NCP(k)} \Big(\prod_{\substack{B\in K(\pi),\\ k\notin B}} \Big\langle \prod_{j\in B} A_j\Big\rangle \prod_{i\in B(k)\setminus\{k\}} A_i\Big)\Big(\prod_{B\in\pi} m_\circ[B]\Big), \tag{3.18}$$

where $B(k)$ is the block in $K(\pi)$ that contains k. Recalling that $T_j = G_j A_j$, we have the averaged[4] *local law*

$$\langle T_{[1,k]}\rangle = \langle M_{[k]} A_k\rangle + \mathcal{O}_\prec\Big(\frac{1}{N}\Big), \tag{3.19}$$

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

$$\langle \mathbf{x}, T_{[1,k\rangle} G_k \mathbf{y}\rangle = \langle \mathbf{x}, M_{[k]}\mathbf{y}\rangle + \mathcal{O}_\prec\Big(\frac{1}{\sqrt{N}}\Big). \tag{3.20}$$

[3]The noncommutative product $\prod_{j\in B} A_j$ for $B = (j_1, \dots, j_r)$ is defined as $A_{j_1} \dots A_{j_r}$ in the order inherited from $B \subseteq S$.

[4]Recall that $\langle\cdot\rangle$ carries a N^{-1} normalization factor, but $\langle\cdot,\cdot\rangle$ does not. For $A := |\mathbf{y}\rangle\langle\mathbf{x}|$ with two unit vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$, we hence have $\langle \mathbf{x}, G(z)\mathbf{y}\rangle = \mathrm{Tr}[G(z)A]$ and $\langle G(z)A\rangle = N^{-1}\langle \mathbf{x}, G(z)\mathbf{y}\rangle$.

As we frequently encounter $\langle M_{[k]} A_k \rangle$ in the following sections, we introduce the notation

$$\mathfrak{m}_1[T_1, \dots, T_k] = \mathfrak{m}_1[z_1, A_1, \dots, z_k, A_k] := \langle M_{[k]} A_k \rangle. \tag{3.21}$$

In particular,

$$\mathfrak{m}_1[T_1, \dots, T_k] = \sum_{\pi \in NCP(k)} \Big(\prod_{B \in K(\pi)} \Big\langle \prod_{j \in B} A_j \Big\rangle \Big) \Big(\prod_{B \in \pi} m_\circ[B] \Big) \tag{3.22}$$

and we have $\mathfrak{m}_1[G_1, \dots, G_k] = m[1, \dots, k]$ as a consequence of (3.14). Comparing (3.22) with the formulas in [111, Sect. 4.4], the set functions m and $m_\circ$ may be interpreted as the moments resp. free cumulants characterizing the limitig non-commutative probability distribution of resolvents $(G_j)_j$ while $\mathfrak{m}_1[\cdot]$ describes the limiting joint distribution of resolvents $(G_j)_j$ and deterministic matrices $(A_j)_j$.

We further apply (3.19) or (3.20) for a 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 denoted as

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

with the same definition as in (3.18).

Remark. The quantities $m[1, \dots, k]$, $m_\circ[1, \dots, k]$, $\mathfrak{m}_1[T_1, \dots, T_k]$, $(M_{[k]})_{ij}$, and $\|M_{[k]}\|$ are of order one for any $k \in \mathbb{N}$ and $i, j \in [N]$ in the macroscopic regime (cf. Lemma 2.4 and Appendix A of [37]). Theorem 3.1.9 asserts that the deterministic $M_{[k]}$ is the leading order approximation of $T_{[1,k\rangle} G_k$. In particular, the error terms in (3.19) and (3.20) are smaller than the natural upper bound on their leading term by a factor of $1/N$ and $1/\sqrt{N}$, respectively.

We further need a generalization of the averaged local law (3.19) that includes transposes.

Theorem 3.1.10 (Global law for resolvent chains with transposes)**.** *Let $k \in \mathbb{N}$ and pick spectral parameters $z_1, \dots, z_k$ with $|\Im(z_j)| \gtrsim 1$ and $\max_j |z_j| \le N^{100}$ as well as deterministic matrices $A_1, \dots, A_k$ with $\|A_j\| \lesssim 1$. Moreover, let $G_j^\sharp$ denote either the resolvent $G_j = G(z_j)$ or its transpose G_j^t and denote by $\#$ the binary vector that has a one in j-th position if $G_j^\sharp = G_j^t$ and a zero otherwise. Then,*

$$\langle G_1^\sharp A_1 \dots G_k^\sharp A_k \rangle = \sum_{\pi \in NCP(k)} \Big(\prod_{B \in K(\pi)} \Big\langle \prod_{j \in B} A_j \Big\rangle \Big) \Big(\prod_{B \in \pi} m_\circ^{\#,\sigma}[B] \Big) + \mathcal{O}_\prec\Big(\frac{1}{N}\Big) \tag{3.23}$$

where $m_\circ^{\#,\sigma}[\cdot]$ denotes the free cumulants associated with the set function $m^{\#,\sigma}[\cdot]$ by Definition 3.1.4. Here, $m^{\#,\sigma}[\cdot]$ is defined to satisfy $m^{\#,\sigma}[\emptyset] = 0$ as well as the recursion

$$m^{\#,\sigma}[1, \dots, k] = m_1 (1 + q_{1,k}^\sharp) \Big(m^{\#,\sigma}[2, \dots, k] + \sum_{j=2}^{k} c_{1,j} m^{\#,\sigma}[1, \dots, j] m^{\#,\sigma}[j, \dots, k] \Big) \tag{3.24}$$

with $q_{1,k}^\sharp = q_{1,k} = \frac{m_1 m_k}{1 - m_1 m_k}$ whenever $\#_1 = \#_k$, i.e., either both G_1 and G_k occur as transposes in the product $G_1^\sharp \dots G_k^\sharp$ or neither of them, and $q_{1,k}^\sharp = \frac{\sigma m_1 m_k}{1 - \sigma m_1 m_k}$ otherwise. Similarly, $c_{1,j} = 1$ whenever $\#_1 = \#_j$ and $c_{1,j} = \sigma$ otherwise. Recall that $\sigma = \mathbb{E}\chi_{od}^2$ where χ_{od} is the real or complex random variable that specifies the distribution of the off-diagonal entries of the Wigner matrix W.

The proof of Theorem 3.1.10 is, modulo careful bookkeeping of the transposes, similar to the proof of the averaged local law in [39, Thm. 3.4]. For the convenience of the reader, a brief sketch of the argument is included in Appendix 3.A.1. We remark that the same result may be obtained on mesoscopic scales with optimal error bounds following the strategy of [37] (cf. [37, Rem. 2.2]) and that several examples in the cases $k \in \{2, 3\}$ are considered in Proposition 3.4 and Remark 3.5 of [36] as well as in Propositions 3.3 and 3.4 of [40].

Note that $\sigma = 1$ implies that the matrix W is real and its resolvent satisfies $G_j^t = G_j$. Hence, the statement of Theorem 3.1.10 reduces to that of an averaged global law for real symmetric Wigner matrices in this case. Due to the structural similarity between (3.22) and (3.23), we will slightly abuse notation and write the right-hand side of (3.23) as

$$\mathfrak{m}_1[G_1^\sharp A_1, \dots, G_k^\sharp A_k] := \sum_{\pi \in NCP(k)} \Big(\prod_{B \in K(\pi)} \Big\langle \prod_{j \in B} A_j \Big\rangle \Big) \Big(\prod_{B \in \pi} m_\circ^{\#,\sigma}[B] \Big). \tag{3.25}$$

Moreover, by Definition 3.1.4, we have $\mathfrak{m}_1[G_1^\sharp, \dots, G_k^\sharp] = m^{\#,\sigma}[1, \dots, k]$ and (3.24) reduces to the divided differences in Definition 3.1.5 whenever $\#$ is the zero vector.

3.1.3 Preliminaries Part 2: Second-Order Quantities

In this section, we give an overview of the definitions from free probability that are used in later sections (roughly following the notation of [111, Ch. 5]) as well as some related quantities appearing in the CLTs.

Recall that the key picture for describing the expectation of $\langle T_{[1,k]} \rangle$ is a disk with the labels $1, \dots, k$ organized in clockwise order along its boundary. In a very similar spirit, the key picture for describing the corresponding second-order object, i.e., the covariance of $\langle T_{[1,k]} \rangle$ and $\langle T_{[k+1,k+\ell]} \rangle$, consists of two concentric labeled circles. Let $k, \ell \in \mathbb{N}$ and arrange the numbers $1, \dots, k$ equidistantly in clockwise order on the outer circle and the numbers $k+1, \dots, k+\ell$ equidistantly in counter-clockwise order on the inner circle. We refer to the planar domain between these two circles together with the labeled points on its boundary as *the* (k, ℓ)*-annulus* (see Fig. 3.4 below). The labeled points will often serve as vertices of a graph. In this case, any edges connecting two points are drawn inside the annulus.

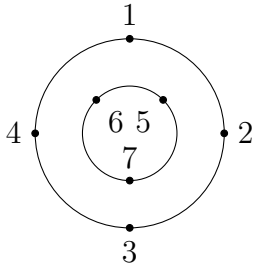

Fig. 3.4. The labels of the $(4, 3)$-annulus.

Definition 3.1.11 (Annular non-crossing permutations)**.** *Let $k, \ell \in \mathbb{N}$. We call a permutation of $[k+\ell]$ an* annular non-crossing permutation *if we can draw its cycles*[5] *on the (k,ℓ)-annulus such that the following conditions (see [111, Def. 5 in Ch. 5]) are satisfied:*

(i) Non-crossing property: The cycles do not cross.

(ii) Standardness: Each cycle encloses a region in the annulus that is homeomorphic to the disk with boundary oriented clockwise (in particular, the cycles follow the orientation of the numbering of the circles).

(iii) Connectedness: At least one cycle connects both circles.

The set of annular non-crossing permutations is denoted by $\overrightarrow{NCP}(k,\ell)$. Any cycle that connects both circles is referred to as connecting cycle*.*

We remark that $\overrightarrow{NCP}(k,\ell)$ can be fully characterized by the avoidance of certain crossing patterns (cf. analogous geometric characterization of $NCP(k)$ below Definition 3.1.2) and an algebraic analog of the standardness condition. This equivalent definition is discussed, e.g., in [112, Sect. 3], but we will not use it here.

Definition 3.1.12 (Annular non-crossing partitions)**.** *Let $k, \ell \in \mathbb{N}$. We call the partitions induced by the cycles of $\overrightarrow{NCP}(k,\ell)$* annular non-crossing partitions*. The set of annular non-crossing partitions is denoted by $NCP(k,\ell)$. A block that arises from a connecting cycle is referred to as* connecting block*.*

While there is a one-to-one correspondence between the non-crossing partitions of the disk in Definition 3.1.2 and the permutations of $[k]$ avoiding the same crossing pattern, there is a crucial difference between non-crossing partitions and permutations on the (k,ℓ)-annulus. In particular, there is no bijective mapping between a permutation in $\overrightarrow{NCP}(k,\ell)$ and the partition of $[k+\ell]$ induced by its cycles, as, e.g., both permutations (123) and (132) correspond to the partition $\{\{1,2,3\}\}$, but give rise to different pictures due to the orientation induced by Definition 3.1.11(ii) (see Fig. 3.5 below). In general, a permutation always uniquely determines the underlying partition, but a partition can be obtained from more than one permutation. This happens if and only if there is exactly one connecting block (cf. [112, Sect. 4]).

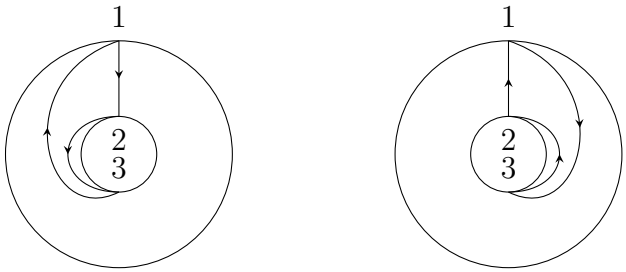

Fig. 3.5. The non-crossing permutations (123) and (132) on the $(1,2)$-annulus. They are different as permutations, but their cycles induce the same non-crossing partition.

[5]Recall that every permutation has a unique cycle decomposition. We represent a cycle $(abc\ldots x)$ by an oriented graph with edge set $\{(a,b),(b,c),\ldots,(x,a)\}$.

Lastly, we consider partitions arising from permutations that respect the non-crossing property and standardness condition but do not have a connecting cycle. In this case, we may consider the permutation restricted to each circle separately, i.e., as an element of $NCP(k) \times NCP(\ell)$, and introduce an artificial connection by *marking* one block on each circle.

Definition 3.1.13 (Marked non-crossing partition)**.** *Consider $\pi \in NCP(k) \times NCP(\ell)$ that naturally splits into $\pi = \pi_1 \times \pi_2$ with $\pi_1 \in NCP(k)$, $\pi_2 \in NCP(\ell)$. We pick one block of π_1 and one of π_2, respectively, and mark them by underlining. The resulting object is referred to as a* marked non-crossing partition*.*

Marking a block on each circle allows us to artificially introduce a connecting block by considering the union of the two marked blocks. As a consequence, any marked non-crossing partition can be associated with a unique element of $NCP(k, \ell)$. We further note that there are $|\pi_1| \cdot |\pi_2|$ possibilities to mark the blocks of $\pi = \pi_1 \times \pi_2 \in NCP(k) \times NCP(\ell)$. For example, $\{\{\underline{1}\}, \{2\}\} \times \{\{\underline{3}\}\}$ and $\{\{1\}, \{\underline{2}\}\} \times \{\{\underline{3}\}\}$ are considered different marked partitions although both arise from $\{\{1\}, \{2\}\} \times \{\{3\}\} \in NCP(2) \times NCP(1)$ (see Fig. 3.6 below).

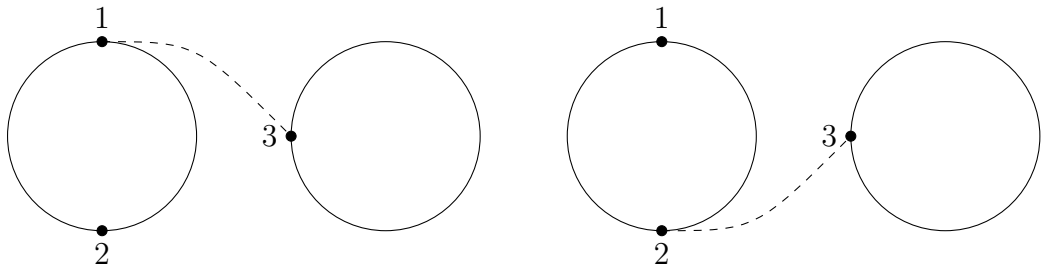

Fig. 3.6. A visualization of $\{\{\underline{1}\}, \{2\}\} \times \{\{\underline{3}\}\}$ and $\{\{1\}, \{\underline{2}\}\} \times \{\{\underline{3}\}\}$. Marking is indicated by a dashed line.

We further recall from [111, Ch. 5] that there is a second-order analog of Definition 3.1.4.

Definition 3.1.14 (Second-order free cumulant function)**.** *Let $(f_{m,n})_{m,n}$ be a sequence of functions such that $f_{m,n}$ maps tuples (S_1, S_2) of two finite (cyclically) ordered sets of integers with $|S_1| = m$ and $|S_2| = n$ to $\mathbb{C}$. Assume further that the functions are symmetric under interchanging of the two arguments in the sense that $f_{m,n}[S_1|S_2] = f_{n,m}[S_2|S_1]$ and cyclic, i.e., $f_{m,n}[S_1|\{s_1, \ldots, s_n\}] = f_{m,n}[S_1|\{s_2, \ldots, s_n, s_1\}]$. Moreover, we assume that $f_{m,0}[S_1|\emptyset] = f_{0,n}[\emptyset|S_2] = 0$. Similar to Definition 3.1.4, we associate the sequence $(f_{m,n})_{m,n}$ with a single function f by setting*

$$f[s_1, \ldots, s_m|s_{m+1}, \ldots, s_{m+n}] := f_{m,n}[\{s_1, \ldots, s_m\}|\{s_{m+1}, \ldots, s_{m+n}\}].$$

We implicitly define the second-order free cumulant function *of f as the unique map $f_{\circ\circ}$ defined on pairs of finite (cyclically) ordered sets (U_1, U_2) that satisfies*

$$f[S_1|S_2] = \sum_{\pi \in \overrightarrow{NCP}(|S_1|,|S_2|)} \prod_{B \in \pi} f_\circ[B] + \sum_{\substack{\pi_1 \times \pi_2 \in NCP(|S_1|) \times NCP(|S_2|), \\ U_1 \in \pi_1, U_2 \in \pi_2 \text{ marked}}} f_{\circ\circ}[U_1|U_2] \prod_{\substack{B \in \pi_1 \setminus U_1 \\ \cup \pi_2 \setminus U_2}} f_\circ[B] \tag{3.26}$$

for any finite S_1, S_2. Here, $f_\circ$ is the first-order free cumulant function introduced in Definition 3.1.4.

Note that we use a set function f that is symmetric under the interchanging of its arguments instead of its skew-symmetric version $f[S_1|S_2] = \overline{f[S_2|S_1]}$ typically used in the free probability literature to mimic the covariance functional (cf. [111, Ch. 5]). This choice will simplify the computations by reducing the number of complex conjugates arising in the intermediate steps.

Similar to (3.14), the implicit relation (3.26) may be turned into an explicit definition of $f_{\circ\circ}$ by recursion. Note that the term $f_{\circ\circ}[[k]\,|\,[k+1,k+\ell]]$ in formula (3.26) with $f[[k]\,|\,[k+1,k+\ell]]$ on the left-hand side only occurs for the marked partition $\{\{\underline{1,\ldots,k}\}\} \times \{\{\underline{k+1,\ldots,k+\ell}\}\}$ and hence always has coefficient one, so we can express it in terms of f, $f_\circ$, and the previously identified values of $f_{\circ\circ}$. This shows that $f_{\circ\circ}$ is well-defined. Although we will not rely on Möbius inversion to express $f_{\circ\circ}$, we remark that it is possible to include both $\overrightarrow{NCP}(k,\ell)$ and the marked elements of $NCP(k) \times NCP(\ell)$ into one common definition, the *non-crossing partitioned permutations*, which can be endowed with a partial ordering and hence render it suitable for Möbius inversion. This would allow to rewrite the right-hand side of (3.26) to a structure similar to (3.14) and obtain a closed formula similar to (3.15). We refer to Sections 4 and 5 of [42] for the full construction.

Similar to (3.14) above, the relation (3.26) is applied for one particular choice of f built up from the Stieltjes transform $m(z)$. We will later see that the set function $\widetilde{m}[\cdot|\cdot]$ defined below arises as the deterministic approximation of the (appropriately scaled) covariance of $\langle G_{[1,k]}\rangle$ and $\langle G_{[k+1,k+\ell]}\rangle$ in a similar way that the divided differences $m[\cdot]$ arise for the expectation of $\langle G_{[1,k]}\rangle$. In particular, $\widetilde{m}[\cdot|\cdot]$ satisfies the symmetry and cyclicity assumption in Definition 3.1.14. We give a recursive definition of $\widetilde{m}[\cdot|\cdot]$ for now, however, closed formulas are later obtained in Section 3.2.2.

Definition 3.1.15. *Let $S_1 = (z_1,\ldots,z_{k'}) \subset \mathbb{C}\setminus\mathbb{R}$ and $S_2 = (z_{k'+1},\ldots,z_{k'+\ell'}) \subset \mathbb{C}\setminus\mathbb{R}$ be two finite ordered multi-sets. We define $\widetilde{m}[\cdot|\cdot]$ to be the set function taking values in $\mathbb{C}$ with the properties (i)-(iii) listed below. Similar to $m[\cdot]$ in Definition 3.1.5, we interpret $\widetilde{m}[\cdot|\cdot]$ as a function of the indices of the spectral parameters.*

(i) Symmetry: $\widetilde{m}[\cdot|\cdot]$ 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

$$\widetilde{m}[(i,z_i\in B_1)|(j,z_j\in B_2)] = \widetilde{m}[(j,z_j\in B_2)|(i,z_i\in B_1)].$$

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

$$\widetilde{m}[(i,z_i\in B_1)|\emptyset] = \widetilde{m}[\emptyset|(j,j\in B_2)] = 0. \tag{3.27}$$

(iii) Recursion: Let $B_1 \subseteq S_1$ and $B_2 \subseteq S_2$ be ordered subsets with $|B_1| = k \le k'$ and $|B_2| = \ell \le \ell'$ elements, respectively. For simplicity, we index them by $[k]$ and $[k+1,k+\ell]$. The function $\widetilde{m}[\cdot|\cdot]$ satisfies the following linear recursion

$$\begin{aligned}&\widetilde{m}[1,\ldots,k|k+1,\ldots,k+\ell]\\ &= \frac{m_1}{1-m_1m_k}\Bigg(\widetilde{m}[2,\ldots,k|k+1,\ldots,k+\ell]\end{aligned}$$

$$+\sum_{j=1}^{k-1}\widetilde{m}[1,\dots,j|k+1,\dots,k+\ell]m[j,\dots,k]$$
$$+\sum_{j=2}^{k}m[1,\dots,j]\widetilde{m}[j,\dots,k|k+1,\dots,k+\ell]+s_{GUE}+s_\kappa+s_\sigma+s_\omega\Bigg) \tag{3.28}$$

where the source terms in the last line are given by

$$s_{GUE}:=\sum_{j=1}^{\ell}m[1,\dots,k,k+j,\dots,k+\ell,k+1,\dots,k+j]$$
$$s_\kappa:=\kappa_4\sum_{r=1}^{k}\sum_{k+1\le s\le t\le k+\ell}\Big(m[1,\dots,r]m[r,\dots,k]m[s,\dots,t]$$
$$\times m[t,\dots,k+\ell,k+1,\dots,s]\Big)$$
$$s_\sigma:=\sigma\sum_{j=1}^{\ell}m^{\#,\sigma}[1,\dots,k,k+j,\dots,k+\ell,k+1,\dots,k+j]$$
$$s_\omega:=\widetilde{\omega_2}\sum_{j=1}^{\ell}m[1,\dots,k]m[k+j,\dots,k+\ell,k+1,\dots k+j].$$

Here, we wrote out the underlying multi-set in the definition of s_κ to indicate that it evaluates to $m[s,s]$ instead of m_s if $t=s$ and the vector $\#\in\{0,1\}^{k+\ell+1}$ is given by $\#_1=\dots=\#_k=0$ and $\#_{k+1}=\dots=\#_{k+\ell+1}=1$. Recall that $m[\cdot]$ denotes the divided differences as introduced in Definition 3.1.5 and that $m^{\#,\sigma}[\cdot]$ was introduced in Theorem 3.1.10.

Note that the recursion for $\widetilde{m}[\cdot|\cdot]$ is linear with different types of source terms in the last line of (3.28). Therefore, we may introduce the decomposition

$$\widetilde{m}[\cdot|\cdot]=\widetilde{m}_{GUE}[\cdot|\cdot]+\kappa_4\widetilde{m}_\kappa[\cdot|\cdot]+\sigma\widetilde{m}_\sigma[\cdot|\cdot]+\widetilde{\omega}_2\widetilde{m}_\omega[\cdot|\cdot] \tag{3.29}$$

where $\widetilde{m}_{GUE}[\cdot|\cdot]$ satisifes (3.28) for $\kappa_4=\sigma=\widetilde{\omega}_2=0$ and $\kappa_4\widetilde{m}_\kappa[\cdot|\cdot]$, $\sigma\widetilde{m}_\sigma[\cdot|\cdot]$, and $\widetilde{\omega}_2\widetilde{m}_\omega[\cdot|\cdot]$ satisfy (3.28) with s_κ, s_σ, and s_ω as the only source term, respectively.

Note that the right-hand side of (3.28) only contains divided differences and $\widetilde{m}[B_1|B_2]$ for $|B_1|+|B_2|<k+\ell$, so (3.28) indeed defines $\widetilde{m}[\cdot|\cdot]$ recursively. The symmetry assumption in (i) then extends (3.28) to the second entry of $\widetilde{m}[\cdot|\cdot]$. Moreover, all source terms in the last line of (3.28) are fully expressable as a function of $m_1,\dots,m_{k+\ell}$ by (3.16), making $\widetilde{m}[\cdot|\cdot]$ eventually a function of $m_1,\dots,m_{k+\ell}$ as well.

As an example, setting $\sigma=\widetilde{\omega}_2=0$ and applying the recursion once gives

$$\begin{aligned}\widetilde{m}[1|2]&=\frac{m_1^2m_2^2}{(1-m_1^2)(1-m_2^2)(1-m_1m_2)^2}+\kappa_4\frac{m_1^3m_2^3}{(1-m_1^2)(1-m_2^2)}\\&=\frac{m_1'm_2'}{(1-m_1m_2)^2}+\kappa_4m_1m_1'm_2m_2'\end{aligned} \tag{3.30}$$

with $m_i'=m'(z_i)$. We remark that $\widetilde{m}_{GUE}[1|2]$, seen as a function of (z_1,z_2), is sometimes referred to as the second-order Cauchy transform of the GUE ensemble in the free probability literature (cf. [53]). The corresponding first-order object is $-m(z)$, which is obtained by applying the usual Cauchy transform to the semicircle law.[6]

[6]The Cauchy transform of a probability measure μ is given by $c(z)=\int_{\mathbb{R}}\frac{\mu(\mathrm{d}x)}{z-x}$ and hence only differs from the Stieltjes transform by a sign.

We consider another special case in the following example.

Example 3.1.16. Whenever $\kappa_4 = \sigma = \widetilde{\omega}_2 = 0$ and one argument of $\widetilde{m}_{GUE}[\cdot|\cdot]$ is a singleton set, only the fourth line of (3.28) gives a non-zero contribution. Rewriting this term using (3.16) (cf. Lemma 3.1.17 below) yields the closed formula

$$\begin{aligned}\widetilde{m}_{GUE}[1|2,\dots,\ell+1] &= \frac{m_1}{1-m_1^2}\sum_{j=2}^{\ell+1} m[1,\dots,j,\dots,\ell+1,j] \\ &= \Big(\prod_{s=1}^{\ell+1} m_s\Big)\Big(\sum_{\Gamma\in NCG([\ell+1])}\prod_{(a,b)\in E(\Gamma)} q_{a,b}\Big)\sum_{j=2}^{\ell+1}\frac{m_1'}{m_1}\frac{m_j'}{m_j}\Big(1+\sum_{i\neq j} q_{i,j}\Big).\end{aligned} \tag{3.31}$$

For $\ell = 2$, we hence obtain

$$\widetilde{m}_{GUE}[1|2,3] = \frac{m_1'(m_2'm_3(1-m_1m_3)+m_2m_3'(1-m_1m_2))}{(1-m_1m_2)^2(1-m_1m_3)^2(1-m_2m_3)}.$$

Note that the right-hand side of (3.31) is fully expressed terms of non-crossing graphs on a labeled disk. This is because $\overrightarrow{NCP}(1,\ell)$ and marked elements of $NCP(1)\times NCP(\ell)$ can be reduced to disk non-crossing partitions in this special case. In particular, the orientation of the circles is not relevant for this example.

The proof of (3.31) is immediate from the following combinatorial lemma which may be of independent interest. We give its proof in Appendix 3.B.1.

Lemma 3.1.17. *For $j\in\{1,\dots,k\}$, $k\geq 1$, we have*

$$m[1,\dots,j,\dots,k,j] = m[1,\dots,k]\Big(1+\sum_{l\in[k]\setminus\{j\}} q_{j,l}\Big)\frac{m_j'}{m_j} \tag{3.32}$$

with $q_{i,j}$ as in (3.13).

We also give some examples to illustrate the combinatorial formula (3.26) for the choice $f[\cdot|\cdot] = \widetilde{m}_{GUE}[\cdot|\cdot]$.

Example 3.1.18 (Second-order free cumulants). Let $\kappa_4 = \sigma = \widetilde{\omega}_2 = 0$. In the case $k=\ell=1$, the only non-crossing annular permutation is (12) and there is also only one option for marking $\{\{1\}\}\times\{\{2\}\}$, namely $\{\{\underline{1}\}\}\times\{\{\underline{2}\}\}$. Rearranging (3.26), we thus get

$$m_{\circ\circ}[1|2] = \widetilde{m}[1|2] - m[1,2] + m_1m_2.$$

Similarly, considering $k=1$ and $\ell=2$ yields

$$\begin{aligned}m_{\circ\circ}[1|2,3] = \widetilde{m}[1|2,3] - \widetilde{m}[1|2]m_3 - \widetilde{m}[1|3]m_2 - 2m[1,2,3] \\ + 2m_1m[2,3] + 2m_2m[1,3] + 2m_3m[1,2] - 4m_1m_2m_3.\end{aligned}$$

In the case $k=\ell=2$, there are 18 distinct non-crossing annular permutations (see Fig. 3.7 below) and 4 elements in $NCP(2)\times NCP(2)$. However, the second sum in (3.26) consists of 9 terms in total due to the marking of the blocks as, e.g., $\{\{\underline{1}\},\{2\}\}\times\{\{\underline{3,4}\}\}$ and $\{\{1\},\{\underline{2}\}\}\times\{\{\underline{3,4}\}\}$ correspond to $m_{\circ\circ}[1|3,4]m_2$ and $m_{\circ\circ}[2|3,4]m_1$, respectively, which do not need to coincide. In total, the formula defining $m_{\circ\circ}[1,2|3,4]$ has 27 terms on the right-hand side of (3.26).

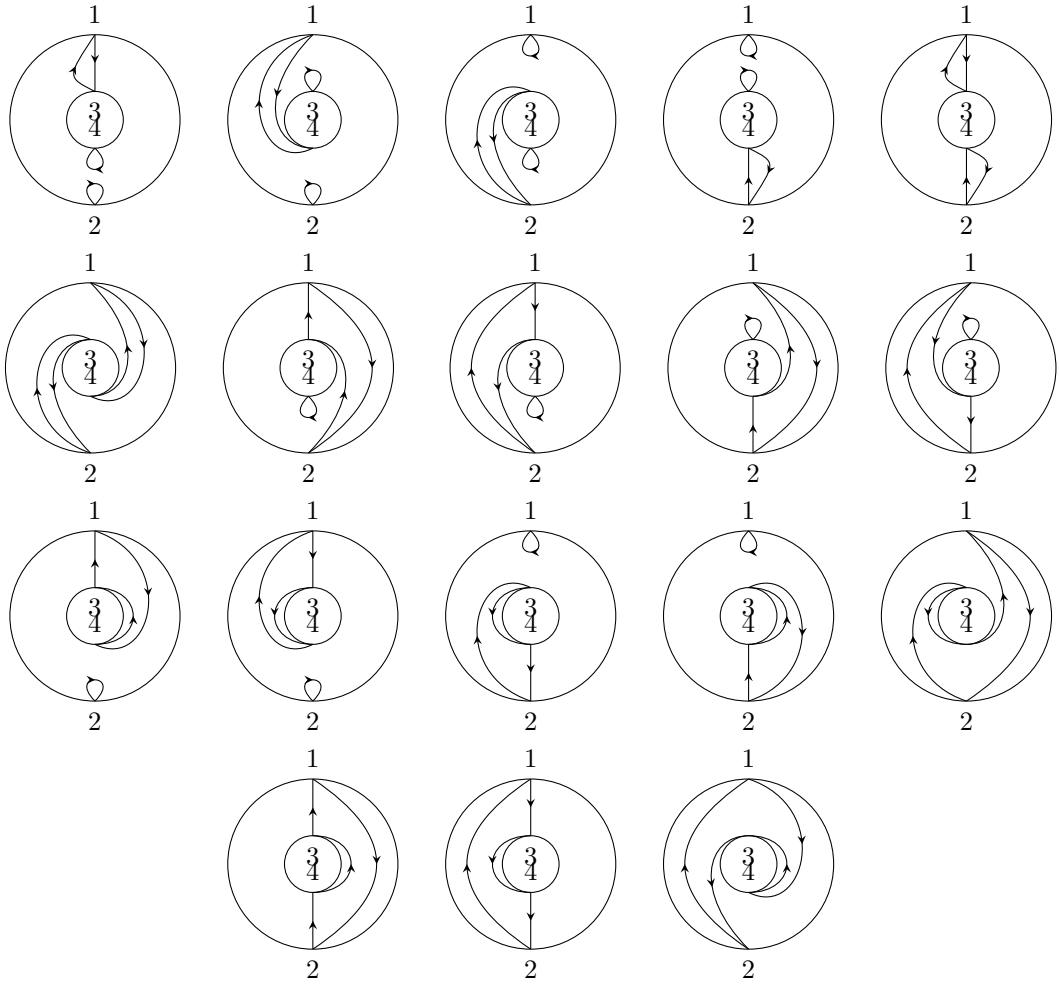

Fig. 3.7. The 18 elements of $\overrightarrow{NCP}(2,2)$.

We conclude this section by recalling the Kreweras complement for annular non-crossing permutations from [112] (see Fig. 3.8 below for an example). Similarly to the disk case, taking the Kreweras complement is an invertible map on $\overrightarrow{NCP}(k,\ell)$.

Definition 3.1.19 (Kreweras complement, annulus case)**.** *Consider the (k,ℓ)-annulus and label the midpoints of the arcs between the points $1,\dots,k+\ell$ (black in Fig. 3.8) also by $1,\dots,k+\ell$ (red in Fig. 3.8). Respecting the orientation of the two circles, we arrange the new labels such that the arc s follows the point s. Let $\pi \in \overrightarrow{NCP}(k,\ell)$ be visualized on the (k,ℓ)-annulus as in Definition 3.1.11. The* (annular) Kreweras complement *$K(\pi) \in \overrightarrow{NCP}(k,\ell)$ is defined as the maximal annular non-crossing permutation on $[k+\ell]$ that can be drawn using only the labels at the midpoints of the arcs and without intersecting the cycles of π. In particular, each cycle of $K(\pi)$ again encloses a region in the annulus that is homeomorphic to the disk with boundary oriented clockwise. In this context, we consider an annular non-crossing permutation maximal if none of its cycles can be extended (by merging cycles) without inducing a crossing.*

Note that $|\pi| + |K(\pi)| = k+\ell$ for any $\pi \in \overrightarrow{NCP}(k,\ell)$ (see, e.g., [112, Sect. 6]). We remark that while defining the annular Kreweras complement on the level of partitions would also be possible, the resulting map does not have the same properties as in the disk case (see, e.g., [112, Sect. 1] for a discussion). Therefore, we will only consider the

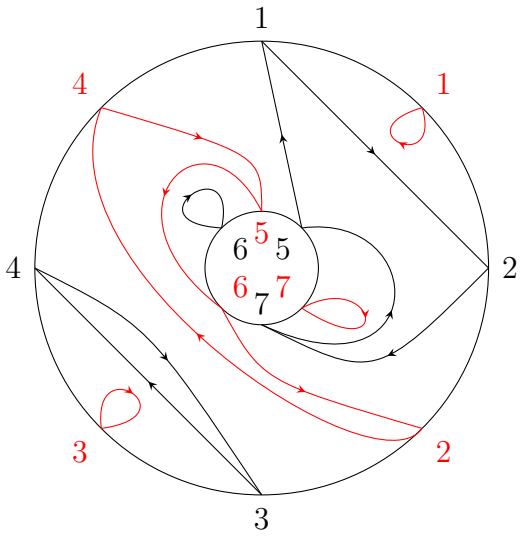

Fig. 3.8. The annular non-crossing permutation $\pi = (1275)(34)(6)$ in black and its Kreweras complement $K(\pi) = (1)(2456)(3)(7)$ in red.

annular Kreweras complement for permutations. Note that one can further assign a unique Kreweras complement to any marked non-crossing partition π arising from some element $\pi_1 \times \pi_2 \in NCP(k) \times NCP(\ell)$ by applying Definition 3.1.3 circle-wise. In this case, we write $K(\pi) = K(\pi_1) \times K(\pi_2)$.

3.2 Main Results

The main focus of the present paper lies in determining the limiting covariance structure arising in the CLT for the centered statistics

$$X_\alpha := \langle T_{[1,k]} \rangle - \mathbb{E}\langle T_{[1,k]} \rangle = \langle G_1 A_1 \dots G_k A_k \rangle - \mathbb{E}\langle G_1 A_1 \dots G_k A_k \rangle, \tag{3.33}$$

$$Y_\alpha := \langle f_1(W) A_1 \dots f_k(W) A_k \rangle - \mathbb{E}\langle f_1(W) A_1 \dots f_k(W) A_k \rangle. \tag{3.34}$$

Here, $\alpha = ((z_1, A_1), \dots, (z_k, A_k))$ resp. $\alpha := ((f_1, A_1), \dots, (f_k, A_k))$ is a multi-index containing bounded deterministic matrices $A_1, \dots, A_k$ and either the spectral parameters $z_1, \dots, z_k \in \mathbb{C} \setminus \mathbb{R}$ with $|\Im z_j| \gtrsim 1$ appearing in the resolvents or the test functions $f_1, \dots, f_k \in H^{k+1}(\mathbb{R})$ with $\|f_j\| \lesssim 1$. Whenever we need to refer to the number k of resolvents (resp. test functions) in the product X_α (resp. Y_α), we carry the parameter k as a superscript and write $X_\alpha^{(k)}$ (resp. $Y_\alpha^{(k)}$). Recall that we set $T_j = G_j A_j = G(z_j) A_j$ as well as $T_{[i,j]} = T_i T_{i+1} \dots T_j$. Similarly, 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 $F_\emptyset = 0$.

3.2.1 Resolvent Central Limit Theorem and Recursion

We start by identifying the joint distribution of multiple $X_{\alpha_i}^{(k_i)}$ with different k_i and α_i. To state the limiting covariance structure, we introduce a recursively defined set function $\mathfrak{m}_2[\cdot|\cdot]$, which we later identify as the deterministic approximation of the (appropriately scaled) covariance of $\langle T_{[1,k]} \rangle$ and $\langle T_{[k+1,k+\ell]} \rangle$ similar to $M_{[k]}$ and $\mathfrak{m}_1[\cdot]$ arising

for the expectation of $T_{[1,k\rangle}G_k$ (see Theorem 3.1.9 as well as (3.21) and (3.22))[7]. Note that $\alpha = ((z_1, A_1), \ldots, (z_k, A_k))$ contains the same information on the spectral parameters and deterministic matrices involved as the set of matrices $(T_j, j \in [k])$. We will, therefore, occasionally abuse notation and use (z_j, A_j) and $T_j = G_j A_j$ interchangeably. In particular, we write

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

where the two multi-indices α and β index the spectral parameters and deterministic matrices in $T_1, \ldots, T_k$ and $T_{k+1}, \ldots, T_{k+\ell}$, respectively. At this point, we only give a recursive definition for $\mathfrak{m}_2[\cdot|\cdot]$, however, explicit formulas are later obtained in Section 3.2.2. Note that the case $\sigma = \widetilde{\omega}_2 = 0$ of Definition 3.2.1 was already given in [122].

Definition 3.2.1. *Let $S_1 = (T_1, \ldots, T_{k'})$ and $S_2 = (T_{k'+1}, \ldots, T_{k'+\ell'})$ 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

$$\mathfrak{m}_2[(T_i, i \in B_1)|\emptyset] = \mathfrak{m}_2[\emptyset|(T_j, j \in B_2)] = 0. \tag{3.35}$$

(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

$$\begin{aligned}
&\mathfrak{m}_2[T_1, \ldots, T_k|T_{k+1}, \ldots, T_{k+\ell}] \\
&= m_1\Bigg(\mathfrak{m}_2[T_2, \ldots, T_{k-1}, G_k A_k A_1|T_{k+1}, \ldots, T_{k+\ell}] \\
&\quad + q_{1,k}\mathfrak{m}_2[T_2, \ldots, T_{k-1}, G_k A_1|T_{k+1}, \ldots, T_{k+\ell}]\langle A_k\rangle \\
&\quad + \sum_{j=1}^{k-1} \mathfrak{m}_2[T_1, \ldots, T_{j-1}, G_j|T_{k+1}, \ldots, T_{k+\ell}]\Big(\mathfrak{m}_1[T_j, \ldots, T_k] \\
&\quad + q_{1,k}\mathfrak{m}_1[T_j, \ldots, T_{k-1}, G_k]\langle A_k\rangle\Big) \\
&\quad + \sum_{j=2}^{k} \mathfrak{m}_1[T_1, \ldots, T_{j-1}, G_j]\Big(\mathfrak{m}_2[T_j, \ldots, T_k|T_{k+1}, \ldots, T_{k+\ell}] \\
&\quad + q_{1,k}\mathfrak{m}_2[T_j, \ldots, T_{k-1}, G_k|T_{k+1}, \ldots, T_{k+\ell}]\langle A_k\rangle\Big) + \mathfrak{s}_{GUE} + \mathfrak{s}_\kappa + \mathfrak{s}_\sigma + \mathfrak{s}_\omega\Bigg)
\end{aligned} \tag{3.36}$$

[7]Note 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.

where the source terms $\mathfrak{s}_{GUE}$, $\mathfrak{s}_\kappa$, $\mathfrak{s}_\sigma$, and $\mathfrak{s}_\omega$ are given by

$$\begin{aligned}
\mathfrak{s}_{GUE} &:= \sum_{j=1}^{\ell} \Big(\mathfrak{m}_1[T_1, \dots, T_k, T_{k+j}, \dots, T_{k+j-1}, G_{k+j}] \\
&\qquad + q_{1,k} \mathfrak{m}_1[T_1, \dots, T_{k-1}, G_k, T_{k+j}, \dots, T_{k+j-1}, G_{k+j}] \langle A_k \rangle \Big) \qquad (3.37) \\
\mathfrak{s}_\kappa &:= \kappa_4 \sum_{r=1}^{k} \sum_{s=k+1}^{k+\ell} \Big(\sum_{t=k+1}^{s} \langle M_{[r]} \odot M_{(s,\dots,k+\ell,k+1,\dots,t)} \rangle \langle (M_{[r,k]} A_k) \odot M_{[t,s]} \rangle \\
&\qquad + \sum_{t=s}^{k+\ell} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle (M_{[r,k]} A_k) \odot M_{(t,\dots,k+\ell,k+1,\dots,s)} \rangle \Big) \\
&\quad + \kappa_4 q_{1,k} \sum_{r=1}^{k} \sum_{s=k+1}^{k+\ell} \Big(\sum_{t=k+1}^{s} \langle M_{[r]} \odot M_{(s,\dots,k+\ell,k+1,\dots,t)} \rangle \langle M_{[r,k]} \odot M_{[t,s]} \rangle \\
&\qquad + \sum_{t=s}^{k+\ell} \langle M_{[r]} \odot M_{[s,t]} \rangle \langle M_{[r,k]} \odot M_{(t,\dots,k+\ell,k+1,\dots,s)} \rangle \Big) \langle A_k \rangle. \qquad (3.38) \\
\mathfrak{s}_\sigma &:= \sigma \sum_{j=1}^{\ell} \mathfrak{m}_1[T_1 \dots, T_k, G_{k+j}^t A_{k+j-1}^t, \dots, G_{k+\ell}^t A_{k+1}^t, \dots, G_{k+j-1}^t A_{k+j}^t, G_{k+j}] \\
&\quad + q_{1,k} \sigma \sum_{j=1}^{\ell} \mathfrak{m}_1[T_1 \dots, T_{k-1}, G_k, G_{k+j}^t A_{k+j-1}^t, \dots \\
&\qquad \dots, G_{k+\ell}^t A_{k+1}^t, \dots, G_{k+j-1}^t A_{k+j}^t, G_{k+j}] \langle A_k \rangle \qquad (3.39) \\
\mathfrak{s}_\omega &:= \widetilde{\omega_2} \sum_{j=1}^{\ell} \langle (M_{[k]} A_k) \odot M_{(k+j,\dots,k+\ell,k+1,\dots k+j)} \rangle \\
&\quad + q_{1,k} \widetilde{\omega_2} \sum_{j=1}^{\ell} \langle M_{[k]} \odot M_{(k+j,\dots,k+\ell,k+1,\dots k+j)} \rangle \langle A_k \rangle \qquad (3.40)
\end{aligned}$$

Recall that $\odot$ denotes the Hadamard product, $q_{1,k}$ was defined in (3.13)*, and $M_{(\dots)}$ was defined in Theorem 3.1.9. Moreover, recall that $\mathfrak{m}_1[\cdot]$ was defined in* (3.21) *and the notation with transposes was introduced in* (3.25)*.*

Note that setting $A_1 = \dots = A_{k+\ell} = \mathrm{Id}$ reduces (3.36) to (3.28), showing that

$$\mathfrak{m}_2[G_1, \dots, G_k | G_{k+1}, \dots, G_{k+\ell}] = m[1, \dots, k | k+1, \dots, k+\ell].$$

We use the linearity of the recursion and the different types of source terms to introduce the decomposition

$$\mathfrak{m}_2[\cdot|\cdot] = \mathfrak{m}_{GUE}[\cdot|\cdot] + \kappa_4 \mathfrak{m}_\kappa[\cdot|\cdot] + \sigma \mathfrak{m}_\sigma[\cdot|\cdot] + \widetilde{\omega}_2 \mathfrak{m}_\omega[\cdot|\cdot], \qquad (3.41)$$

where $\mathfrak{m}_{GUE}[\cdot|\cdot]$ satisfies (3.36) for $\kappa_4 = \sigma = \widetilde{\omega}_2 = 0$, and $\kappa_4 \mathfrak{m}_\kappa[\cdot|\cdot]$, $\sigma \mathfrak{m}_\sigma[\cdot|\cdot]$ resp. $\widetilde{\omega}_2 \mathfrak{m}_\omega[\cdot|\cdot]$ satisfy (3.36) with $\mathfrak{s}_\kappa$, $\mathfrak{s}_\sigma$ resp. $\mathfrak{s}_\omega$ as only source term. Note that $\mathfrak{s}_{GUE} + \mathfrak{s}_\kappa + \mathfrak{s}_\sigma + \mathfrak{s}_\omega$ in (3.36) is fully expressible as a function of $A_1, \dots, A_{k+\ell}$ and $m_1, \dots, m_{k+\ell}$ by (3.22), (3.17) and Lemma 3.A.2, eventually making $\mathfrak{m}_2[\cdot|\cdot]$ a function of the same quantities.

Recall that we set $X_\alpha = \langle T_1 \dots T_k \rangle - \mathbb{E}\langle T_1 \dots T_k \rangle$ with $\alpha = ((z_1, A_1), \dots, (z_k, A_k))$. Before stating the CLT for X_α, we note the following definition.

Definition 3.2.2. *Consider two functions of the Wigner matrix W in Assumption 3.1.1, 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*

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

for any smal $\varepsilon > 0$, where the implicit constant in $\mathcal{O}(\cdot)$ only depends on the polynomial $\mathcal{P}$ and the constants in Assumption 3.1.1.

We now give a CLT for X_α in (3.33). As the main interest of the present paper is the deterministic approximation $\mathfrak{m}_2[\cdot|\cdot]$, we restrict the discussion of the CLT to the macroscopic regime ($|\Im z_i| \gtrsim 1$) for technical simplicity. Note that this implies in particular that $\mathfrak{m}_2[\cdot|\cdot]$ as well as its components $\mathfrak{m}_{GUE}[\cdot|\cdot]$, $\mathfrak{m}_\kappa[\cdot|\cdot]$, $\mathfrak{m}_\sigma[\cdot|\cdot]$, and $\mathfrak{m}_\omega[\cdot|\cdot]$ are of order one. The proof of Theorem 3.2.3 is analogous to that of [122, Thm. 3.6]. For the convenience of the reader, we include the necessary modifications for adapting the proof in [122] to the generalized model in Assumption 3.1.1 in Appendix 3.A.2.

Theorem 3.2.3 (Macroscopic 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 3.1.1. For each $j = 1, \dots, p$ pick a set of spectral parameters $z_1^{(j)}, \dots, z_{k_j}^{(j)}$ with $|\Im z_i^{(j)}| \gtrsim 1$ and $\max_j |z_j| \leq N^{100}$ as well as deterministic matrices $A_1^{(j)}, \dots, A_{k_j}^{(j)}$ with $\|A_i^{(j)}\| \lesssim 1$. Then,*

$$N^p \mathbb{E}\Big(\prod_{j=1}^{p} X_{\alpha_j} \Big) = \sum_{Q \in Pair([p])} \prod_{\{i,j\} \in Q} \mathfrak{m}_2[\alpha_i|\alpha_j] + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N}} \Big) \tag{3.42}$$

for any $\varepsilon > 0$. Here, $Pair(S)$ denotes the pairings of a set S and $\mathfrak{m}_2[\cdot|\cdot]$ is a set function that satisfies Definition 3.2.1. Equation (3.42) *establishes an asymptotic version of Wick's rule and hence identifies the joint limiting distribution of the random variables $(X_{\alpha_j})_j$ as asymptotically complex Gaussian in the sense of moments in the limit $N \to \infty$.*

By (3.41), Theorem 3.2.3 implies in particular that

$$\lim_{N\to\infty} N^2 \mathbb{E}\Big(X_\alpha X_\beta \Big) = \mathfrak{m}_{GUE}[\alpha|\beta] + \kappa_4 \mathfrak{m}_\kappa[\alpha|\beta] + \sigma \mathfrak{m}_\sigma[\alpha|\beta] + \widetilde{\omega}_2 \mathfrak{m}_\omega[\alpha|\beta].$$

We remark that $\mathfrak{m}_2[\cdot|\cdot]$ is cyclic in the sense that

$$\mathfrak{m}_2[(T_j, j \in S_1)|T_1, \dots, T_k] = \mathfrak{m}_2[(T_j, j \in S_1)|T_2, \dots, T_k, T_1]$$

and that further

$$\begin{aligned} &\mathfrak{m}_2[(T_j, j \in S_1)|T_1, \dots, T_{k-1}, G_k] \\ &= \frac{\mathfrak{m}_2[(T_j, j \in S_1)|T_2, \dots, T_{k-1}, G_k A_1] - \mathfrak{m}_2[(T_j, j \in S_1)|T_1, \dots, T_{k-1}]}{z_k - z_1}. \end{aligned}$$

whenever $z_1 \neq z_k$, $A_k = \mathrm{Id}$, and $\sigma = 0$. These identities can be obtained from the "meta argument" below [37, Lem. 4.1] (see also [122, Cor. 3.7]) using that the analogous formulas for the original resolvent chains are trivially true by resolvent identities. However, any additional information on $\mathfrak{m}_2[\cdot|\cdot]$ has to be obtained from the recursion (3.36) directly.

3.2.2 Solution of the Recursion

After identifying $\mathfrak{m}_2[\alpha|\beta]$ as the deterministic approximation of $\mathbb{E}[X_\alpha X_\beta]$, we consider Definition 3.2.1 in detail. In this section, we derive the main result, which is a solution to the (deterministic) recursion (3.36). This characterizes the overall structure of the function $\mathfrak{m}_2[\cdot|\cdot]$ and yields explicit combinatorial formulas to replace the recursive definition in applications. Making use of the linearity of the recursion and the decomposition (3.41), it is sufficient to consider the components $\mathfrak{m}_{GUE}[\cdot|\cdot]$, $\mathfrak{m}_\kappa[\cdot|\cdot]$,$\mathfrak{m}_\sigma[\cdot|\cdot]$, and $\mathfrak{m}_\omega[\cdot|\cdot]$ separately. We start by studying $\mathfrak{m}_{GUE}[\cdot|\cdot]$. The proof consists of two steps that are carried out in Section 3.3.

Theorem 3.2.4. *Let $\alpha = ((z_1, A_1), \dots, (z_k, A_k))$ and $\beta = ((z_{k+1}, A_{k+1}), \dots, (z_{k+\ell}, A_{k+\ell}))$ for some $k, \ell \in \mathbb{N}$. Then,*

$$\begin{aligned}\mathfrak{m}_{GUE}[\alpha|\beta] = &\sum_{\pi \in \overline{NCP}(k,\ell)} \Big(\prod_{B\in K(\pi)} \Big\langle \prod_{j\in B} A_j \Big\rangle \Big) \prod_{B\in\pi} m_\circ[B] \\ &+ \sum_{\substack{\pi_1\times\pi_2 \in NCP(k)\times NCP(\ell),\\ U_1\in\pi_1, U_2\in\pi_2 \text{ marked}}} \Big(\prod_{\substack{B_1\in K(\pi_1),\\ B_2\in K(\pi_2)}} \Big\langle \prod_{j\in B_1} A_j \Big\rangle \Big\langle \prod_{j\in B_2} A_j \Big\rangle \Big) m_{\circ\circ}[U_1|U_2] \\ &\times \prod_{\substack{B_1\in\pi_1\setminus U_1,\\ B_2\in\pi_2\setminus U_2}} m_\circ[B_1] m_\circ[B_2] \end{aligned} \tag{3.43}$$

with $m_\circ$ and $m_{\circ\circ}$ being the first and second-order free cumulant functions as defined in (3.14) *and* (3.26)*, respectively.*

Observe that the right-hand side of (3.43) reduces to the combinatorial expression in (3.26) if $A_1 = \cdots = A_{k+\ell} = \mathrm{Id}$.

Remark. Note that the right-hand side of (3.43) is symmetric with respect to interchanging of $((z_1, A_1), \dots, (z_k, A_k))$ and $((z_{k+1}, A_{k+1}), \dots, (z_{k+\ell}, A_{k+\ell}))$, which is consistent with the symmetry of $\mathfrak{m}_2[\cdot|\cdot]$ in Definition 3.2.1(i). We can check this directly from (3.43) by observing that there is a one-to-one correspondence between non-crossing permutations of the (k, ℓ)-annulus and those of the (ℓ, k)-annulus. This follows from drawing the cycles of the permutation as curves on the respective annuli and observing that interchanging the inner and outer circle with a conformal map (e.g., by inversion to a concentric circle between the outer and inner circle) preserves the standardness and non-crossing property of the picture (cf. Definition 3.1.15). Moreover, this symmetry of $\widetilde{m}_{GUE}[\cdot|\cdot]$ implies that $m_{\circ\circ}[\cdot|\cdot]$ is also invariant under interchanging $((z_1, A_1), \dots, (z_k, A_k))$ and $((z_{k+1}, A_{k+1}), \dots, (z_{k+\ell}, A_{k+\ell}))$ since $\widetilde{m}_{GUE}[\cdot|\cdot]$ determines $m_{\circ\circ}[\cdot|\cdot]$ uniquely by Definition 3.1.14.

Example 3.2.5 (Asymptotics of covariances for GUE). We consider a special case of Theorem 3.2.3. Let $p = 2$, $k_1 = k_2 = 1$, and assume that W is a GUE matrix[8]. By decomposing A_1 and A_2 into a tracial and a traceless part, the deterministic approximation for the covariance follows directly from [40, Thm. 4.1], giving

$$\begin{aligned} &N^2 \mathbb{E}(\langle T_1\rangle - \mathbb{E}\langle T_1\rangle)(\langle T_2\rangle - \mathbb{E}\langle T_2\rangle) \\ &\quad = \langle A_1A_2\rangle \frac{m_1^2 m_2^2}{(1-m_1m_2)} + \langle A_1\rangle\langle A_2\rangle \Big(\frac{m_1' m_2'}{(1-m_1m_2)^2} - \frac{m_1^2 m_2^2}{(1-m_1m_2)} \Big) + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N}}\Big) \end{aligned}$$

[8]Note that we are only using that W satisfies Assumption 3.1.1 and $\kappa_4 = \sigma = \widetilde{\omega}_2 = 0$.

$$= \langle A_1 A_2 \rangle m_\circ[1,2] + \langle A_1 \rangle \langle A_2 \rangle m_{\circ\circ}[1|2] + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N}}\Big),$$

where the last equation follows from the formulas in Examples 3.1.6 and 3.1.18. Note that the error bound $\Psi/\sqrt{L}$ in [40, Eq. (91)] evaluates to $\mathcal{O}(1/\sqrt{N})$ on macroscopic scales. We remark that the deterministic leading term matches the formula for $\mathfrak{m}_{GUE}[T_1|T_2]$ obtained from applying (3.36) to the initial condition $\mathfrak{m}_{GUE}[T_1|\emptyset] = 0$.

Next, we consider the recursion for $\mathfrak{m}_\kappa[\cdot|\cdot]$. We obtain a closed solution similar to Theorem 3.2.4, i.e., a sum of terms that factorizes into two parts depending only on the deterministic matrices $A_1, \dots, A_{k+\ell}$ and the spectral parameters $z_1, \dots, z_{k+\ell}$, respectively. The proof of Theorem 3.2.6 is given in Section 3.4.1 below.

Theorem 3.2.6. *Let $\alpha = ((z_1, A_1), \dots, (z_k, A_k))$ and $\beta = ((z_{k+1}, A_{k+1}), \dots, (z_{k+\ell}, A_{k+\ell}))$ for some $k, \ell \in \mathbb{N}$. Then there exist*

(i) a family $(\psi_{\pi,B})_{B\in\pi}$ of functions $\psi_{B,\pi} : \mathbb{C}^{|B|} \to \mathbb{C}$ for every $\pi \in \overrightarrow{NCP}(k,\ell)$ and

(ii) a family $(\psi_{\pi,U_1,U_2})_{U_1\subset[k],U_2\subset[k+1,k+\ell]}$ of functions $\psi_{\pi,U_1,U_2} : \mathbb{C}^{|U_1|} \times \mathbb{C}^{|U_2|} \to \mathbb{C}$ that are invariant under interchanging of the two arguments as well as functions $(\psi_{\pi_1,B_1})_{B_1\in\pi_1\setminus U_1}$ and $(\psi_{\pi_2,B_2})_{B_2\in\pi_2\setminus U_2}$ with $\psi_{\pi_i,B_i} : \mathbb{C}^{|B_i|} \to \mathbb{C}$ for every $\pi = \pi_1 \times \pi_2 \in NCP(k) \times NCP(\ell)$ with marked blocks $U_1 \in \pi_1$ and $U_2 \in \pi_2$

such that

$$\begin{aligned}\mathfrak{m}_\kappa[\alpha|\beta] = {} & \sum_{\pi\in\overrightarrow{NCP}(k,\ell)} \prod_{B\in K(\pi)} \Big\langle \Big(\prod_{j\in B\cap[k]} A_j \Big) \odot \Big(\prod_{j\in B\cap[k+1,k+\ell]} A_j \Big) \Big\rangle \prod_{B\in\pi} \psi_{\pi,B}(z_j | j \in B) \\ & + \sum_{\substack{\pi=\pi_1\times\pi_2\in NCP(k)\times NCP(\ell),\\ U_1\in\pi_1, U_2\in\pi_2 \text{ marked}}} \Big(\prod_{\substack{B_1\in K(\pi_1),\\ B_2\in K(\pi_2)}} \Big\langle \prod_{j\in B_1} A_j \Big\rangle \Big\langle \prod_{j\in B_2} A_j \Big\rangle \Big) \\ & \times \psi_{\pi,U_1,U_2}(z_j | j \in U_1 \cup U_2) \prod_{\substack{B_1\in\pi_1\setminus U_1,\\ B_2\in\pi_2\setminus U_2}} \psi_{\pi_1,B_1}(z_j | j \in B_1) \psi_{\pi_2,B_2}(z_j | j \in B_2),\end{aligned} \tag{3.44}$$

where $\odot$ denotes the Hadamard product.

Theorem 3.2.6 is a purely structural result which shows that $\mathfrak{m}_\kappa[\cdot|\cdot]$ can be written in terms of non-crossing permutations and partitions similar to $\mathfrak{m}_{GUE}[\cdot|\cdot]$. Despite the obvious similarities between (3.43) and (3.44), considering the minimal example

$$\begin{aligned}\mathfrak{m}_2[T_1|T_2] &= \mathfrak{m}_{GUE}[T_1|T_2] + \kappa_4 \mathfrak{m}_\kappa[T_1|T_2] \\ &= \langle A_1 A_2 \rangle \frac{m_1^2 m_2^2}{(1-m_1m_2)} + \langle A_1 \rangle \langle A_2 \rangle \Big(\frac{m_1' m_2'}{(1-m_1m_2)^2} - \frac{m_1^2 m_2^2}{(1-m_1m_2)} \Big) \\ &\quad + \kappa_4 \Big(\langle \mathbf{a}_1 \mathbf{a}_2 \rangle m_1^3 m_2^3 + \langle A_1 \rangle \langle A_2 \rangle (2 m_1 m_1' m_2 m_2' - m_1^3 m_2^3) \Big)\end{aligned}$$

with $\sigma = \widetilde{\omega}_2 = 0$ already shows that the functions ψ_i describing the dependence on the spectral parameters do not coincide with the free cumulant functions $m_\circ[\cdot]$ and $m_{\circ\circ}[\cdot|\cdot]$ in general. However, (3.36) implies that the functions ψ_i themselves satisfy a recursion. Theorem 3.2.6 hence reduces the computation of $\mathfrak{m}_\kappa[\cdot|\cdot]$ from iterating (3.36) to an inductive computation of the part that only depends on the spectral parameters.

We continue by deriving an explicit formula for $\mathfrak{m}_\sigma[\cdot|\cdot]$. As the source term $\mathfrak{s}_\sigma$ in the corresponding recursion is, up to transposes, identical to $\mathfrak{s}_{GUE}$, the solution of the recursion is analogous to Theorem 3.2.4 but uses $m^{\#,\sigma}[\cdot]$ instead of the iterated divided differences $m[\cdot]$. We give the proof in Section 3.4.2.

Theorem 3.2.7. *Let $\alpha = ((z_1, A_1), \dots, (z_k, A_k))$, $\beta = ((z_{k+1}, A_{k+1}), \dots, (z_{k+\ell}, A_{k+\ell}))$ for some $k, \ell \in \mathbb{N}$ and abbreviate*

$$\widetilde{m}_\sigma[1, \dots, k | k+1, \dots, k+\ell] := \mathfrak{m}_\sigma[G_1, \dots G_k | G_{k+1}, \dots, G_{k+\ell}]$$

in the special case $A_1 = \cdots = A_{k+\ell}$. If $B \in \pi$ is a connecting cycle of $\pi \in \overrightarrow{NCP}(k, \ell)$ decomposed as $B = (i_1, \dots, i_r) \circ (j_1, \dots, j_s)$ with $i_1, \dots, i_r \subset [k]$ and $j_1, \dots, j_s \subset [k+1, k+\ell]$, we introduce the notation

$$B_\sigma := (i_1, \dots, i_r) \circ (j_s, \dots, j_1).$$

Then,

$$\begin{aligned}\mathfrak{m}_\sigma[\alpha|\beta] = &\sum_{\pi \in \overrightarrow{NCP}(k,\ell)} \Big(\prod_{B \in K(\pi)} \Big\langle \prod_{j \in B \cap [k]} A_j \Big(\prod_{j \in B \cap [k+1, k+j]} A_j \Big)^t \Big\rangle \Big) \prod_{B \in \pi} m_\circ^{\#,\sigma}[B_\sigma] \\ &+ \sum_{\substack{\pi_1 \times \pi_2 \in NCP(k) \times NCP(\ell), \\ U_1 \in \pi_1, U_2 \in \pi_2 \text{ marked}}} \Big(\prod_{\substack{B_1 \in K(\pi_1), \\ B_2 \in K(\pi_2)}} \Big\langle \prod_{j \in B_1} A_j \Big\rangle \Big\langle \prod_{j \in B_2} A_j \Big\rangle \Big) \\ &\quad \times (m_\sigma)_{\circ\circ}[U_1 | U_2] \prod_{\substack{B_1 \in \pi_1 \setminus U_1, \\ B_2 \in \pi_2 \setminus U_2}} m_\circ^{\#,\sigma}[B_1] m_\circ^{\#,\sigma}[B_2]\end{aligned} \tag{3.45}$$

where $\# = (0, \dots, 0, 1, \dots, 1)$ with the number of zeros and ones matching the number of labels on the inner and outer circle involved in B, respectively. Moreover, $m_\circ^{\#,\sigma}[\cdot]$ denotes the free cumulant function associated with $m^{\#,\sigma}[\cdot]$ via (3.14) and $(m_\sigma)_{\circ\circ}[\cdot|\cdot]$ denotes the second-order free cumulant function asssociated to $\widetilde{m}_\sigma[\cdot|\cdot]$ and $m^{\#,\sigma}[\cdot]$ via (3.26), respectively.

Note that the set function $\mathfrak{m}_\sigma[\cdot|\cdot]$ satisfies the same factorization property as $\mathfrak{m}_{GUE}[\cdot|\cdot]$ and $\mathfrak{m}_\kappa[\cdot|\cdot]$. For $k = \ell = 1$, it readily follows that

$$\mathfrak{m}_\sigma[T_1|T_2] = \langle A_1 A_2^t \rangle \frac{m_1^2 m_2^2}{1 - \sigma m_1 m_2} + \langle A_1 \rangle \langle A_2 \rangle \Big(\frac{m_1' m_2'}{(1 - \sigma m_1 m_2)^2} - \frac{m_1^2 m_2^2}{1 - \sigma m_1 m_2} \Big).$$

We further remark that (3.41) evaluates to $\mathfrak{m}_2[\cdot|\cdot] = \mathfrak{m}_{GUE}[\cdot|\cdot] + \mathfrak{m}_\sigma[\cdot|\cdot]$ in the case of GOE matrices ($\sigma = 1$ and $\kappa_4 = \widetilde{\omega}_2 = 0$).

It remains to consider $\mathfrak{m}_\omega[\cdot|\cdot]$. The proof of Theorem 3.2.8 is given in Section 3.4.2.

Theorem 3.2.8. *Let $\alpha = ((z_1, A_1), \dots, (z_k, A_k))$ and $\beta = ((z_{k+1}, A_{k+1}), \dots, (z_{k+\ell}, A_{k+\ell}))$ for some $k, \ell \in \mathbb{N}$. Then there exist*

(i) a family $(\Psi_{\pi,B})_{B \in \pi}$ of functions $\Psi_{B,\pi} : \mathbb{C}^{|B|} \to \mathbb{C}$ for every $\pi \in \overrightarrow{NCP}(k, \ell)$ and

(ii) a family $(\Psi_{\pi,U_1,U_2})_{U_1 \subset [k], U_2 \subset [k+1,k+\ell]}$ of functions $\Psi_{\pi,U_1,U_2} : \mathbb{C}^{|U_1|} \times \mathbb{C}^{|U_2|} \to \mathbb{C}$ that are invariant under interchanging of the two arguments, and functions $(\Psi_{\pi_1,B_1})_{B_1 \in \pi_1 \setminus U_1}$ resp. $(\Psi_{\pi_2,B_2})_{B_2 \in \pi_2 \setminus U_2}$ with $\Psi_{\pi_i,B_i} : \mathbb{C}^{|B_i|} \to \mathbb{C}$ for every $\pi = \pi_1 \times \pi_2 \in NCP(k) \times NCP(\ell)$ with marked blocks $U_1 \in \pi_1$ and $U_2 \in \pi_2$

such that

$$\begin{aligned}\mathfrak{m}_\omega[\alpha|\beta] &= \sum_{\pi\in\overrightarrow{NCP}(k,\ell)} \prod_{B\in K(\pi)} \Big\langle \Big(\prod_{j\in B\cap[k]} A_j\Big)\odot\Big(\prod_{j\in B\cap[k+1,k+\ell]} A_j\Big)\Big\rangle \prod_{B\in\pi}\Psi_{B,\pi}(z_j|j\in B)\\ &+ \sum_{\substack{\pi=\pi_1\times\pi_2\in NCP(k)\times NCP(\ell),\\ U_1\in\pi_1,U_2\in\pi_2 \text{ marked}}} \Big(\prod_{\substack{B_1\in K(\pi_1),\\ B_2\in K(\pi_2)}} \Big\langle\prod_{j\in B_1} A_j\Big\rangle\Big\langle\prod_{j\in B_2} A_j\Big\rangle\Big) \qquad (3.46)\\ &\times \Psi_{\pi,U_1,U_2}(z_j|j\in U_1\cup U_2)\prod_{\substack{B_1\in\pi_1\setminus U_1,\\ B_2\in\pi_2\setminus U_2}}\Psi_{\pi_1,B_1}(z_j|j\in B_1)\Psi_{\pi_2,B_2}(z_j|j\in B_2).\end{aligned}$$

Similar to Theorem 3.2.6, Theorem 3.2.8 is a pure structural result. It shows that the last contribution $\mathfrak{m}_\omega[\cdot|\cdot]$ satisfies the same factorization property as $\mathfrak{m}_{GUE}[\cdot|\cdot]$, $\mathfrak{m}_\kappa[\cdot|\cdot]$, and $\mathfrak{m}_\sigma[\cdot|\cdot]$. In the case $k=\ell=1$, we have the formula

$$\mathfrak{m}_\omega[T_1|T_2] = \langle \mathbf{a}_1\mathbf{a}_2\rangle m_1^2 m_2^2 + \langle A_1\rangle\langle A_2\rangle(m_1' m_2' - m_1^2 m_2^2).$$

It further follows from (3.36) that the functions Ψ_i themselves satisfy a recursion. Theorem 3.2.8 thus reduces the computation of $\mathfrak{m}_\omega[\cdot|\cdot]$ to the inductive computation of the part that only depends on the spectral parameters.

3.2.3 General Test Functions and Applications to Free Probability

We conclude the discussion by comparing the explicit formulas from Section 3.2.2 to the free probability results in [107]. To do so, we generalize the CLT for resolvents in Theorem 3.2.3 to a full multi-point functional CLT for (N-independent) test functions $f_1,\dots,f_k$, i.e., a CLT for the statistics Y_α in (3.34). Given Theorem 3.2.3 and the formulas fron Section 3.2.2, the proof follows from a classical application of the Helffer-Sjöstrand formula (analogous to the proof of [122, Thm. 2.7]) and is hence omitted. Note that we restrict Theorem 3.2.9 to real-valued test functions only for simplicity. Extending the results in this section to complex-valued test functions only requires minor modifications to the argument.

Theorem 3.2.9 (Macroscopic multi-point functional CLT)**.** *Let $k\in\mathbb{N}$ and pick deterministic matrices $A_1,\dots,A_k\in\mathbb{C}^{N\times N}$ with $\|A_j\|\lesssim 1$. Let further W be a Wigner matrix satisfying Assumption 3.1.1 and let $f_1,\dots,f_k\in H^{k+1}(\mathbb{R})$ be real-valued compactly supported test functions with $\|f_j\|\lesssim 1$. Then, for any $\varepsilon>0$, the centered statistics* (3.34) *are approximately distributed (in the sense of moments) as*

$$NY_\alpha^{(k,a)} = \xi(\alpha) + \mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N}}\Big) \qquad (3.47)$$

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

$$\begin{aligned}\mathbb{E}\xi(\alpha)\xi(\beta) &= \sum_{\pi\in\overrightarrow{NCP}(k,\ell)}\Big(\prod_{B\in K(\pi)}\Big\langle\prod_{j\in B} A_j\Big\rangle\Big)\Phi_\pi^{(GUE)}(f_1,\dots,f_{k+\ell})\\ &+ \kappa_4\sum_{\pi\in\overrightarrow{NCP}(k,\ell)}\prod_{B\in K(\pi)}\Big\langle\Big(\prod_{j\in B\cap[k]} A_j\Big)\odot\Big(\prod_{j\in B\cap[k+1,k+\ell]} A_j\Big)\Big\rangle\Phi_\pi^{(\kappa)}(f_1,\dots,f_{k+\ell})\end{aligned}$$

$$+\sigma \sum_{\pi\in\overrightarrow{NCP}(k,\ell)} \Big(\prod_{B\in K(\pi)} \Big\langle \prod_{j\in B\cap[k]} A_j \Big(\prod_{j\in B\cap[k+1,k+j]} A_j \Big)^t \Big\rangle \Big) \Phi_\pi^{(\sigma)}(f_1,\dots,f_{k+\ell}) \tag{3.48}$$

$$+\widetilde{\omega}_2 \sum_{\pi\in\overrightarrow{NCP}(k,\ell)} \prod_{B\in K(\pi)} \Big\langle \Big(\prod_{j\in B\cap[k]} A_j \Big) \odot \Big(\prod_{j\in B\cap[k+1,k+\ell]} A_j \Big) \Big\rangle \Phi_\pi^{(\omega)}(f_1,\dots,f_{k+\ell})$$

$$+ \sum_{\substack{\pi_1\times\pi_2\in NCP(k)\times NCP(\ell),\\ U_1\in\pi_1, U_2\in\pi_2 \text{ marked}}} \Big(\prod_{\substack{B_1\in K(\pi_1),\\ B_2\in K(\pi_2)}} \Big\langle \prod_{j\in B_1} A_j \Big\rangle \Big\langle \prod_{j\in B_2} A_j \Big\rangle \Big) \Phi_{\pi_1\times\pi_2, U_1\times U_2}(f_1,\dots,f_{k+\ell}).$$

Here, β denotes another multi-index of length ℓ containing the deterministic matrices $A_{k+1},\dots,A_{k+\ell}$ satisfying $\|A_j\| \lesssim 1$ and the test functions $f_{k+1},\dots,f_{k+\ell} \in H^{\ell+1}(\mathbb{R})$. The functions $\Phi_\pi^{(\cdot)}$ and $\Phi_{\pi_1\times\pi_2,U_1\times U_2}$ in (3.48) *can be computed recursively and only depend on the underlying permutation resp. marked partition, the functions $f_1,\dots,f_{k+\ell}$ and the model parameters κ_4, σ, and $\widetilde{\omega}_2$.*

For the later applications, we note the following formulas for the case $\kappa_4 = \sigma = \widetilde{\omega}_2 = 0$. Corollary 3.2.10 below is proven in [122] by explicitly evaluating the integrals that are obtained from the Helffer-Sjöstrand formula in the proof of the functional CLT.

Corollary 3.2.10 (Cor. 2.9 in [122])**.** *Consider Theorem 3.2.9 for a GUE matrix*[9] *W. In this case, we have*

$$\Phi_\pi^{(GUE)}(f_1,\dots,f_{k+\ell}) := \prod_{B\in\pi} \mathrm{sc}_\circ[B], \tag{3.49}$$

where $\mathrm{sc}_\circ[\cdot]$ *denotes the free cumulant function associated with*

$$\mathrm{sc}[i_1,\dots,i_n] := \int_{-2}^{2} \Big[\prod_{j=1}^{n} f_{i_j}(x) \Big] \rho_{sc}(x)\mathrm{d}x, \tag{3.50}$$

with ρ_{sc} as in (3.9)*, and*

$$\Phi_{\pi_1\times\pi_2,U_1\times U_2}(f_1,\dots,f_{k+\ell}) := \mathrm{sc}_{\circ\circ}[U_1|U_2] \prod_{\substack{B_1\in\pi_1\setminus U_1,\\ B_2\in\pi_2\setminus U_2}} \mathrm{sc}_\circ[B_1]\mathrm{sc}_\circ[B_2], \tag{3.51}$$

where $\mathrm{sc}_{\circ\circ}[\cdot|\cdot]$ *denotes the second-order free cumulants associated with* $\mathrm{sc}[\cdot]$ *in* (3.50) *and*

$$\mathrm{sc}[i_1,\dots,i_n|i_{n+1},\dots,i_{n+m}] := \frac{1}{2}\int_{-2}^{2}\int_{-2}^{2} \Big(\prod_{j=1}^{n} f_{i_j}(x) \Big)' \Big(\prod_{j=1}^{m} f_{i_{n+j}}(y) \Big)' u(x,y)\mathrm{d}x\mathrm{d}y \tag{3.52}$$

by Definition 3.1.14. The kernel $u : [-2,2]\times[-2,2] \to \mathbb{R}$ is given by

$$u(x,y) := \frac{1}{4\pi^2} \ln\Big[\frac{(\sqrt{4-x^2}+\sqrt{4-y^2})^2(xy+4-\sqrt{4-x^2}\sqrt{4-y^2})}{(\sqrt{4-x^2}-\sqrt{4-y^2})^2(xy+4+\sqrt{4-x^2}\sqrt{4-y^2})} \Big]. \tag{3.53}$$

We remark that the formula (3.53) also appears in [52] and [125] (see also [122, Cor. 3.8]).

Whenever $f_j(x) = x$ for all $j = 1,\dots,k+\ell$ or, more generally, f_j is an (N-independent) polynomial[10], the $N\to\infty$ limit of (3.48) describes the second-order limiting distribution of GUE and deterministic matrices in free probability. It is readily checked that Theorem 3.2.9 indeed coincides with the free probability literature in this case. The computations to obtain Corollary 3.2.11 are included in Appendix 3.B.2.

[9] We only use that W satisfies Assumption 3.1.1 and $\kappa_4 = \sigma = \widetilde{\omega}_2 = 0$.

[10] We implicitly assume f_j to be compactly supported by setting $\widetilde{f}_j(x) = f_j(x)\chi(x)$, where f_j is a polynomial supported on all of $\mathbb{R}$ and χ is a smooth cutoff function that is equal to one on $[-5/2,5/2]$ and equal to zero on $[-3,3]^c$. Since $f_j(W) = \widetilde{f}_j(W)$ with high probability by eigenvalue rigidity, we may use f_j and $\widetilde{f}_j$ interchangeably here.

Corollary 3.2.11. *Under the assumptions of Corollary 3.2.10 let* $f_1(x) = \cdots = f_{k+\ell}(x) = x$, *i.e.,* $Y_\alpha^{(k,a)} = \langle WA_1 \dots WA_k \rangle - \mathbb{E}\langle \dots \rangle$ *and* $Y_\beta^{(\ell,b)} = \langle WA_{k+1} \dots WA_{k+\ell} \rangle - \mathbb{E}\langle \dots \rangle$. *Then,*

$$\lim_{N\to\infty} N^2 \mathbb{E}\Big(Y_\alpha^{(k,a)} Y_\beta^{(\ell,b)}\Big) = \sum_{\pi \in \overrightarrow{NCP}_2(k,\ell)} \Big(\prod_{B\in K(\pi)} \Big\langle \prod_{j\in B} A_j \Big\rangle \Big) \tag{3.54}$$

where $\overrightarrow{NCP}_2(k,\ell)$ *denotes the pairings in* $\overrightarrow{NCP}(k,\ell)$.

We remark that the limit in (3.54) reproduces the well-known result of second-order freeness of GUE and deterministic matrices from [113]. Moreover, computations similar to Corollaries 3.2.10 and 3.2.11 for the GOE ($\sigma = 1$ and $\kappa_4 = \widetilde{\omega}_2 = 0$) yield

$$\begin{aligned}\lim_{N\to\infty} N^2 \mathbb{E}\Big(Y_\alpha^{(k,a)} Y_\beta^{(\ell,b)}\Big) = &\sum_{\pi \in \overrightarrow{NCP}_2(k,\ell)} \Big(\prod_{B\in K(\pi)} \Big\langle \prod_{j\in B} A_j \Big\rangle \Big) \\ &+ \sum_{\pi \in \overrightarrow{NCP}_2(k,\ell)} \Big(\prod_{B\in K(\pi)} \Big\langle \prod_{j\in B\cap[k]} A_j \Big(\prod_{j\in B\cap[k+1,k+j]} A_j \Big)^t \Big\rangle \Big)\end{aligned}$$

which is the (real) second-order freeness of GOE and deterministic matrices established in [118]. Apart from these two special cases, the deterministic approximation in Theorem 3.2.9 mirrors the overall structure of the joint second-order distribution of Wigner and deterministic matrices described in [107, Eq. (3)]. We remark that resolvents and functions with an N-dependent mesoscopic scaling (3.5) as considered in [122] are usually not accessible in free probability theory as many of the standard techniques rely on explicit moment computations. Theorem 3.2.9 and its mesoscopic analog [122, Thm. 2.7] thus show that the underlying combinatorics of non-crossing annular permutations and marked partitions are, in fact, more general.

We further remark that the parallels between Theorem 3.2.9 and [107] continue to hold if we consider multiple independent Wigner matrices instead of one matrix W. More precisely, for n independent GUE (or GOE) matrices, the underlying combinatorial structure is given by the so-called *non-mixing* annular non-crossing permutations resp. *non-mixing* marked partitions for n colors (cf. Remark below [122, Cor. 2.11]).

3.3 Proof of Theorem 3.2.4 (Formula for $\mathfrak{m}_{GUE}[\cdot|\cdot]$)

The proof of Theorem 3.2.4 consists of two main steps:

1.) Solve the recursion for $\mathfrak{m}_{GUE}[\cdot|\cdot]$ in the case $A_1 = \cdots = A_{k+\ell} = \mathrm{Id}$. As $m[\cdot]$ can be expressed in terms of non-crossing graphs on the disk, we can rewrite the coefficients and the source term of the recursion in terms of graphs. By defining a suitable family of graphs on the (k,ℓ)-annulus, we extend this representation to obtain an ansatz for $\widetilde{m}_{GUE}[\cdot|\cdot]$ and, by Definition 3.1.14, also an expression for the second-order free cumulant function.

2.) Using the formula for $\widetilde{m}_{GUE}[\cdot|\cdot]$ obtained in Step 1, write out the right-hand side of (3.43) and show that it satisfies the recursion for $\mathfrak{m}_{GUE}$. Checking that this ansatz also satisfies the remaining properties in Definition 3.2.1 yields equality between the two sides of (3.43).

We address the two steps in Sections 3.3.1 and 3.3.2, respectively.

3.3.1 Step 1: Graphs

As we only consider $\widetilde{m}_{GUE}$ throughout this section, let $\kappa_4 = \sigma = \widetilde{\omega} = 0$ and thus $\widetilde{m}_{GUE}[\cdot|\cdot] = \widetilde{m}[\cdot|\cdot]$. Recall from (3.16) and (3.17) that both $m[\cdot]$ and $m_\circ[\cdot]$ are expressable in terms of disk non-crossing graphs. In this section, we give analogous combinatorial formulas for $\widetilde{m}[\cdot|\cdot]$ and $m_{\circ\circ}[\cdot|\cdot]$. For this task, we define a new, albeit closely related, multi-set of graphs on the (k,ℓ)-annulus. We start by introducing a transformation that translates between the disk and the annulus picture (cf. [111, Thm. 8 in Ch. 5]).

Definition 3.3.1. *Fix $k,\ell \in \mathbb{N}$, $1 \le j \le \ell$, and consider a disk with the $k+\ell+1$ labels $1,\dots,k,k+j,\dots,k+\ell,k+1,\dots,k+j$ equidistantly placed around its boundary in clockwise order. We define a map τ, refered to as* mediating map*, that takes this picture to the (k,ℓ)-annulus as follows:*

1.) Use a homeomorphic continuous deformation, e.g., a conformal map, to map the disk and its labels to the $(k,\ell+1)$-annulus with a slit located between 1 and k on the outer circle and the two copies of $k+j$ on the inner circle.

2.) Remove the slit to obtain an annulus.

3.) Merge the two copies of the label $k+j$.

We visualize τ for an example in Fig. 3.9 below. The two labels $k+j$ are denoted as 6 and 6' to distinguish between them more easily.

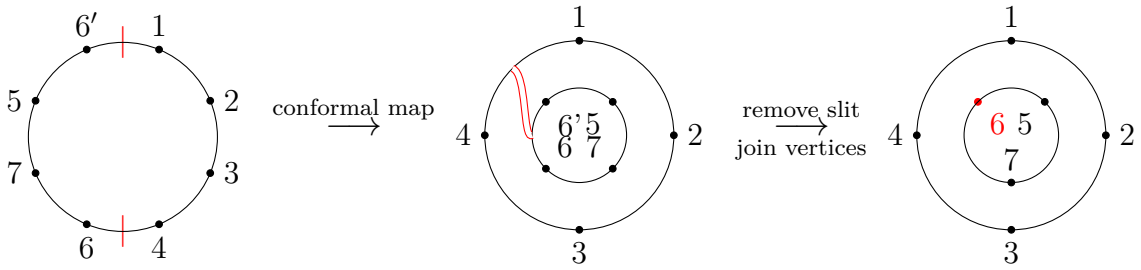

Fig. 3.9. The geometry of the transformation τ for $k=4$, $\ell=3$, and $j=2$.

The map τ induces a transformation of any graph Γ defined on the labeled disk to a graph defined on the (k,ℓ)-annulus. We denote the resulting annulus graph as $\tau(\Gamma)$. By construction, $\tau(\Gamma)$ is planar whenever Γ is a disk non-crossing graph. Recall that we use a slightly more general notion of planar graphs than the standard literature by allowing for loops and multi-edges. We give an example in Fig. 3.10 below. For better visibility, the loop arising in the last step is moved from between the two $(1,3)$ edges to the right.

We can now introduce the family of graphs $\mathcal{G}(k,\ell)$ that constitute the key tool in the proof of Theorem 3.2.4. In analogy to the disk non-crossing graphs in Definition 3.1.7, we require the elements of $\mathcal{G}(k,\ell)$ to be drawn on the (k,ℓ)-annulus with the vertices placed around the boundary and the edges drawn in the interior of the annulus (see Fig. 3.10 and Example 3.3.4 below).

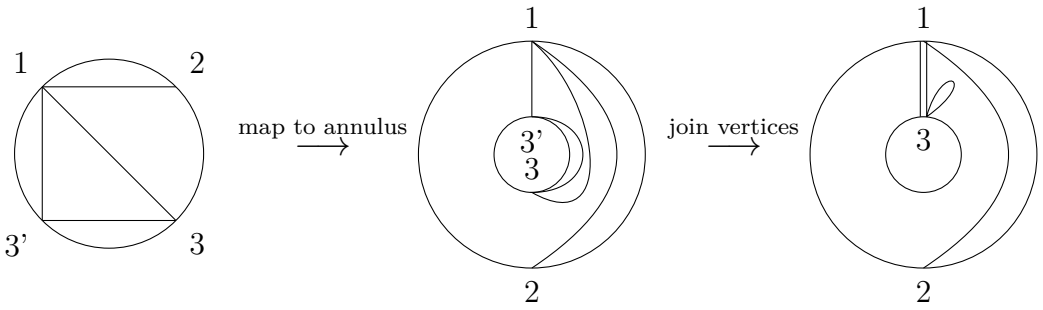

Fig. 3.10. Construction of $\tau(\Gamma)$ for a graph $\Gamma \in NCG(\{1,2,3,3'\})$.

Definition 3.3.2. *For $k, \ell \in \mathbb{N}$ we define $\mathcal{G}([k],[k+1,k+\ell])$ to be the multi-set[11] of undirected, planar graphs on the (k,ℓ)-annulus with vertex set $\{1,\dots,k+\ell\}$ and possible loops or double edges that is obtained from the following recursive construction:*

(i) For any $S_1 \subset [k]$ and $S_2 \subset [k+1,k+\ell]$ we have

$$\mathcal{G}(S_1,\emptyset) = \mathcal{G}(\emptyset,S_2) = \emptyset. \tag{3.55}$$

(ii) The multi-set $\mathcal{G}([k],[k+1,k+\ell])$ can be constructed from the multi-sets $\mathcal{G}(S_1,[k+1,k+\ell])$ with $S_1 \subsetneq [k]$ as follows: We define

$$\mathcal{G}_{\neg(1,k)} := \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3 \cup \mathcal{G}_4, \tag{3.56}$$

as the disjoint union of the sets

$$\mathcal{G}_1 := \Big\{\Gamma \cup \{1\} \Big| \Gamma \in \mathcal{G}([2,k],[k+1,k+\ell])\Big\}$$

$$\mathcal{G}_2 := \bigcup_{j=2}^{k} \Big\{\Gamma = \Gamma_1 \cup \Gamma_2 \Big| \Gamma_1 \in NCG([1,j]) \text{ with edge } (1,j),\ \Gamma_2 \in \mathcal{G}([j,k],[k+1,k+\ell])\Big\},$$

$$\mathcal{G}_3 := \bigcup_{j=1}^{k-1} \Big\{\Gamma = \Gamma_1 \cup \Gamma_2 \Big| \Gamma_1 \in \mathcal{G}([1,j],[k+1,k+\ell]) \text{ with edge } (1,j),\ \Gamma_2 \in NCG([j,k])\Big\},$$

$$\mathcal{G}_4 := \bigcup_{j=1}^{\ell} \tau\Big(\Big\{\Gamma \in NCG(\{1,\dots,k,k+j,\dots,k+\ell,k+1,\dots,k+j-1,k+j\}) \Big| \Gamma \text{ has edge } (1,k+j)\Big\}\Big).$$

Here, the union $\Gamma \cup \{1\}$ in $\mathcal{G}_1$ is to be understood as adding a separated vertex 1 to Γ while the union $\Gamma_1 \cup \Gamma_2$ in $\mathcal{G}_2$ and $\mathcal{G}_3$ refers to the graph with the vertex set $\{1,\dots,k+\ell\}$ and the edge set given by the union of the edge sets of Γ_1 and Γ_2, respectively. Further, recall τ from Definition 3.3.1. We remark that all elements of $\mathcal{G}_1,\dots,\mathcal{G}_4$ are planar graphs. Next, let

$$\mathcal{G}_{(1,k)} := \Big\{\Gamma \cup \{(1,k)\} \Big| \Gamma \in \mathcal{G}_{\neg(1,k)}\Big\},$$

[11] The graph Γ may be obtained in several different ways from the recursive construction. This is reflected by the multiplicity of Γ in the multi-set $\mathcal{G}([k],[k+1,k+\ell])$.

where the union is to be understood as adding an edge $(1,k)$ to each graph in $\mathcal{G}_{\neg(1,k)}$. Observe that the resulting graphs are again planar. Finally, we define

$$\mathcal{G}([k],[k+1,k+\ell]) := \mathcal{G}_{(1,k)} \cup \mathcal{G}_{\neg(1,k)}, \tag{3.57}$$

where any graphs that occur more than once are counted with multiplicity. In particular, whenever the same graph occurs in both $\mathcal{G}_{(1,k)}$ and $\mathcal{G}_{\neg(1,k)}$, the multiplicity of $\Gamma \in \mathcal{G}([k],[k+1,k+\ell])$ is the total number of occurrences of Γ in both subsets.

(iii) $\mathcal{G}(S_1,S_2) = \mathcal{G}(S_2,S_1)$ *for any $S_1 \subset [k]$ and $S_2 \subset [k+1,k+\ell]$ (in the sense that there is a well-defined bijective mapping that takes each element of $\mathcal{G}(S_1,S_2)$ to its counterpart).*

We abbreviate $\mathcal{G}(k,\ell) := \mathcal{G}([k],[k+1,k+\ell])$ and refer to its elements as good graphs. *The subset of connected good graphs is denoted by $\mathcal{G}_c(k,\ell)$. Any edge (i,j) with $i \in [k]$ and $j \in [k+1,k+\ell]$ is referred to as* connecting edge.

We give a schematic overview of the elements in $\mathcal{G}(k,\ell)$ in Fig. 3.11 below. To avoid overcrowding, labels and edges that are not prescribed by the definition are left out.

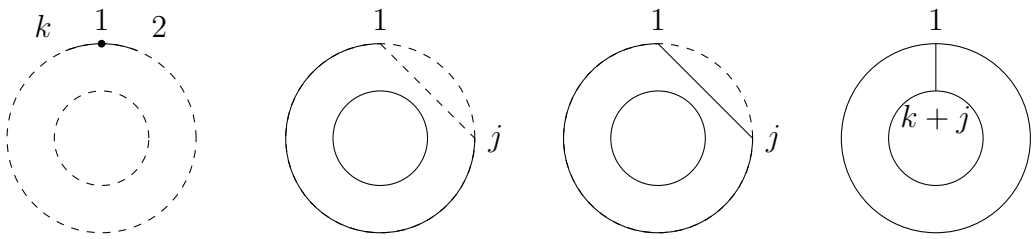

Fig. 3.11. Schematic visualization of the elements of $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_3$, and $\mathcal{G}_4$ (left to right). Different subgraphs are indicated by different line styles.

Remark. The occurrence of multi-edges or loops is inherent to the construction of $\mathcal{G}(k,\ell)$, which can be seen from a simple counting argument. First, note that an element of $\mathcal{G}(k,\ell)$ can have at most $2(k+\ell)$ edges by construction. To see this, observe that the elements of $\mathcal{G}_{\neg(1,k)}$ with the highest number of edges lie in $\mathcal{G}_4$ and that $\Gamma \in \mathcal{G}_4$ has the same number of edges as the underlying disk non-crossing graph. As any disk non-crossing graph on n vertices has at most $2n-3$ edges (realized by a triangulation), the maximal number of edges for $\Gamma \in \mathcal{G}_4$ is $2(k+\ell+1)-3 = 2(k+\ell)-1$. As (3.57) may add another edge to Γ, the maximum for $\mathcal{G}(k,\ell)$ is $2(k+\ell)$ edges. On the other hand, the maximal number of edges in a planar graph on the (k,ℓ)-annulus without multi-edges or loops is only $2(k+\ell)$ if $k,\ell \geq 3$ (again realized by a triangulation). It is readily seen that such a graph has strictly less than $2(k+\ell)$ edges if either k or ℓ is one or two. As the construction in Definition 3.3.2 does not introduce crossings and begins with the cases $k,\ell \leq 2$, the difference between the two maxima must be reflected as multi-edges or loops.

We further remark that $\mathcal{G}(k,\ell)$ is a genuine multi-set, i.e., some graphs appear with multiplicity larger than one, unless $k=\ell=1$ (cf. Lemma 3.3.5(d) and Fig. 3.13). The key property shared by all elements of $\mathcal{G}(k,\ell)$ is that each graph arises from some disk non-crossing graph along the recursive construction. Therefore, we may interpret $\mathcal{G}_4$ as a kind of source term. In view of Lemma 3.3.3 below, the multi-set $\mathcal{G}(k,\ell)$ gives an annulus

analog to the disk non-crossing graphs, as it plays the same role in the combinatorial description of $\widetilde{m}[\cdot|\cdot]$ as $NCG(S)$ does for $m[\cdot]$.

Lemma 3.3.3. *For fixed $k, \ell \in \mathbb{N}$, we have*

$$\widetilde{m}[1,\dots,k|k+1,\dots,k+\ell] = \Big(\prod_{s=1}^{k+\ell} m_s\Big) \sum_{\Gamma\in\mathcal{G}(k,\ell)} \prod_{(i,j)\in E(\Gamma)} q_{i,j}. \tag{3.58}$$

Proof. As $q_{i,j} = q_{j,i}$ by (3.13) and $\mathcal{G}(k,\ell) = \mathcal{G}(\ell,k)$ by Definition 3.3.2(iii), it readily follows that the right-hand side of (3.58) is symmetric under the interchanging $[k]$ and $[k+1,k+\ell]$. Similar to the proof of [39, Lem. 5.2], we use the combinatorial formula (3.58) as an ansatz to solve the recursion given in (3.27) and (3.28). First, observe that

$$\Big(\prod_{s=1}^{k} m_s\Big) \sum_{\Gamma\in\mathcal{G}(S_1,\emptyset)} \prod_{(i,j)\in E(\Gamma)} q_{i,j} = \Big(\prod_{s=k+1}^{k+\ell} m_s\Big) \sum_{\Gamma\in\mathcal{G}(\emptyset,S_2)} \prod_{(i,j)\in E(\Gamma)} q_{i,j} = 0$$

for any $S_1 \subset [k]$ and $S_2 \subset [k+1,k+\ell]$ due to the sums being empty. Hence, the initial condition (3.27) is satisfied.

It remains to check (3.28). We introduce the notation

$$q_\Gamma := \prod_{(i,j)\in E(\Gamma)} q_{i,j}$$

and conclude from the decompositions (3.57) and (3.56) that

$$\begin{aligned}\sum_{\Gamma\in\mathcal{G}(k,\ell)} q_\Gamma &= (1+q_{1,k}) \sum_{\Gamma\in\mathcal{G}_{\neg(1,k)}} q_\Gamma \\ &= (1+q_{1,k})\Big(\sum_{\Gamma\in\mathcal{G}_1} q_\Gamma + \sum_{\Gamma\in\mathcal{G}_2} q_\Gamma + \sum_{\Gamma\in\mathcal{G}_3} q_\Gamma + \sum_{\Gamma\in\mathcal{G}_4} q_\Gamma\Big).\end{aligned} \tag{3.59}$$

Noting that the vertex 1 in $\Gamma \in \mathcal{G}_1$ has no adjacent edges, we may write

$$\sum_{\Gamma\in\mathcal{G}_1} q_\Gamma = \sum_{\Gamma\in\mathcal{G}([2,k],[k+1,k+\ell])} q_\Gamma.$$

Further, the transformation τ only changes the geometry underlying a graph Γ, but does not influence its edge set. By the definition of $\mathcal{G}_4$, any $\Gamma \in NCG([1,\dots,k+\ell,k+j])$ used in the construction must have at least one edge $(1,k+j)$. As a consequence, the product q_Γ always includes the factor $q_{1,k+j} = m_1 m_{k+j}(1+q_{1,k+j})$. This yields

$$\sum_{\Gamma\in\mathcal{G}_4} q_\Gamma = \sum_{j=1}^{\ell} \Big(m_1 m_{k+j} \sum_{\Gamma\in NCG([1,\dots,k+\ell,k+j])} q_\Gamma\Big).$$

Note that the identity $q_{1,k+j} = m_1 m_j (1+q_{1,k+j})$ allows writing summations restricted to graphs with an edge $(1,k+j)$ on the left-hand side into an unrestricted sum over all graphs on the right-hand side. We use the same trick for $\mathcal{G}_2$ and $\mathcal{G}_3$ as $q_{\Gamma_1\cup\Gamma_2} = q_{\Gamma_1} q_{\Gamma_2}$ for the union of graphs introduced in Definition 3.3.2. With these replacements, (3.59) can

be written as

$$\begin{aligned}
&\frac{1}{1+q_{1,k}} \sum_{\Gamma\in\mathcal{G}(k,\ell)} q_\Gamma \\
&= \sum_{\Gamma\in\mathcal{G}([2,k],[k+1,k+\ell])} q_\Gamma + \sum_{j=2}^{k} m_1 m_j \Big(\sum_{\Gamma\in NCG([1,j])} q_\Gamma \Big) \Big(\sum_{\Gamma\in\mathcal{G}([j,k],[k+1,k+\ell])} q_\Gamma \Big) \\
&\quad + \sum_{j=1}^{k-1} m_1 m_j \Big(\sum_{\Gamma\in\mathcal{G}([1,j],[k+1,k+\ell])} q_\Gamma \Big) \Big(\sum_{\Gamma\in NCG([j,k])} q_\Gamma \Big) \\
&\quad + \sum_{j=1}^{\ell} m_1 m_{k+j} \Big(\sum_{\Gamma\in NCG([1,\ldots,k+\ell,k+j])} q_\Gamma \Big).
\end{aligned}$$

Multiplying both sides with $\prod_{s=1}^{k+\ell} m_s$ and noting that $1+q_{1,j} = (1-m_1m_j)^{-1}$, we see that the right-hand side of (3.58) satisfies (3.28) as claimed. □

Example 3.3.4. We have $|\mathcal{G}(1,1)| = 8$. The graphs are visualized in Fig. 3.12 below.

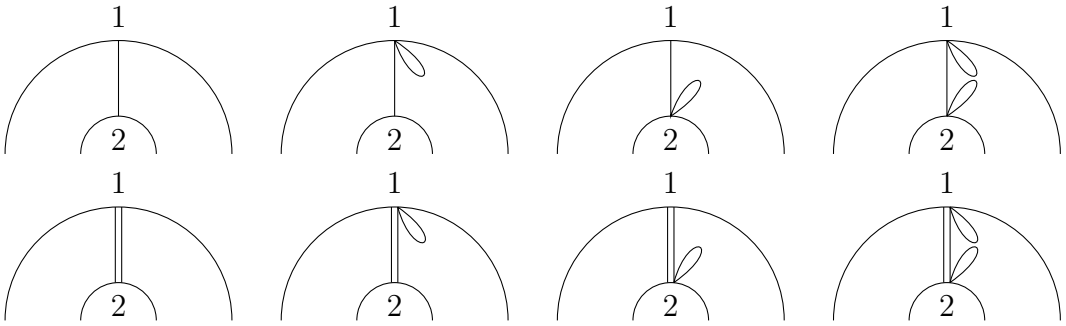

Fig. 3.12. The elements of $\mathcal{G}(1,1)$.

In particular, we readily reobtain (3.30) by evaluating q_Γ for every graph in the above list. Note that decomposing (3.30) into the form $m_1m_2 \sum_{\Gamma\in\mathcal{G}(1,1)} q_\Gamma$ is possible in multiple ways. However, picking graphs that contain a connecting edge yields the set $\mathcal{G}(1,1)$ in Fig. 3.12 as the smallest possible set.

We state a few properties of the elements of $\mathcal{G}(k,\ell)$. In Lemma 3.3.5, we focus on general characteristics of $\Gamma\in\mathcal{G}(k,\ell)$ as a planar graph drawn on the (k,ℓ)-annulus. The properties of $\mathcal{G}(k,\ell)$ that are needed for the proof of the combinatorial formula for $m_{\circ\circ}$ are given in the separate Lemma 3.3.6.

Lemma 3.3.5. *Let $k,\ell\in\mathbb{N}$.*

(a) The connected components of any $\Gamma\in\mathcal{G}(k,\ell)$ give rise to a non-crossing partition of the (k,ℓ)-annulus. In particular, Γ contains a connecting edge if $k,\ell\geq 1$.

(b) $\Gamma\in\mathcal{G}(k,\ell)$ may have at most two loops. Loops only occur at vertices that are adjacent to a connecting edge. Two loops on vertices on the same circle do not occur.

(c) $\Gamma\in\mathcal{G}(k,\ell)$ may have up to $k+\ell-1$ double edges. Double edges are either connecting edges or adjacent to a connecting edge. Edges with a multiplicity higher than two do not occur.

(d) $\mathcal{G}(k,\ell)$ is a genuine multi-set unless $k=\ell=1$.

Proof. (a) We use proof by induction. First, note that the elements of $\mathcal{G}(S_1,\emptyset)=\emptyset$ and $\mathcal{G}(\emptyset,S_2)=\emptyset$ with $S_1\subseteq[k]$ and $S_2\subseteq[k+1,k+\ell]$ clearly give rise to an annular non-crossing partition. Moreover, Example 3.3.4 establishes the claim in the case $k=\ell=1$ and shows that any $\Gamma\in\mathcal{G}(1,1)$ contains at least one connecting edge.

Assume next that the elements of $\mathcal{G}(S_1,[k+1,k+\ell])$ give rise to an annular non-crossing partition for any $S_1\subset[k]$ with $|S_1|\leq k-1$ and a fixed $\ell\geq 1$. We aim to show that the connected components of any $\Gamma\in\mathcal{G}(k,l)$ also correspond to the blocks of some $\pi\in NCP(k,\ell)$. Due to the symmetry induced by Definition 3.3.2(iii), this is enough to establish the induction step.

By definition, the vertices 1 and k lie next to each other on the outer circle. Hence, adding an edge $(1,k)$ may connect two connected components, but cannot introduce a crossing in the partition obtained from them. It is, therefore, sufficient to check the claim for elements of $\mathcal{G}_{\neg(1,k)}$ or, equivalently, for the sets $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_3$, and $\mathcal{G}_4$ in (3.56). First, note that the transformation τ indeed takes the disk partition induced by the disk non-crossing graph to an annular non-crossing partition. This is due to the continuous homeomorphism used in the definition of τ. Moreover, the edge $(1,k+j)$ prescribed for $\Gamma\in NCG(\{1,\dots,k,k+j,\dots,k+j-1,k+j\})$ by the definition is mapped to a connecting edge, ensuring that the resulting partition has a connecting block.

By construction, the graphs in $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$ contain an element of $\mathcal{G}(S_1,[k+1,k+\ell])$ with some $S_1\subset[k]$ as a subgraph. Applying the induction hypothesis for $\mathcal{G}([2,k],k+1,k+\ell])$ and noting that a separate vertex 1 only adds a singleton set to the underlying partition, we can conclude that $\mathcal{G}_1$, too, behaves as claimed. The argument for $\mathcal{G}_2$ and $\mathcal{G}_3$ is similar. Here, the key observation is that the connected components of the added disk non-crossing graph induce a non-crossing partition of an interval placed along the outer circle. Recalling that all elements of $NCP(k,\ell)$ have at least one connecting block, the corresponding connected component of $\Gamma\in\mathcal{G}(k,\ell)$ must contain a connecting edge.

(b) It is readily seen that the elements of $\mathcal{G}(S_1,\emptyset)=\emptyset$ and $\mathcal{G}(\emptyset,S_2)=\emptyset$ with $S_1\subseteq[k]$ and $S_2\subseteq[k+1,k+\ell]$, as well as all $\Gamma\in\mathcal{G}(1,1)$ have the claimed structure. Moreover, adding an edge $(1,k)$ to a graph cannot introduce a loop unless $k=1$, so it is again sufficient to establish the induction step for $\mathcal{G}_{\neg(1,k)}$.

By Definition 3.1.7, a disk non-crossing graph does not contain any loops. Hence, any loops of $\Gamma\in\mathcal{G}_1\cup\mathcal{G}_2\cup\mathcal{G}_3$ must occur in the subgraph in $\mathcal{G}(S_1,[k+1,k+\ell])$ with $S_1\subset[k]$ used to construct Γ. As the induction hypothesis applies to this subgraph, we conclude that Γ has at most two loops that satisfy the claimed placement rules, respectively. For $\Gamma\in\mathcal{G}_4$, note that applying τ to an element of $NCG(\{1,\dots,k,k+j,\dots,k+j-1,k+j\})$ yields an annulus graph with a loop if and only if the disk non-crossing graph contains an edge $(k+j,k+j)$. As the latter must also contain an edge $(1,k+j)$ by definition, a loop in Γ is always adjacent to a connecting edge.

We remark that the loops in $\Gamma\in\mathcal{G}(k,\ell)$ are a consequence of the definition of τ rather than an artifact of the recursive construction of $\mathcal{G}(k,\ell)$, i.e., they can be created only once. In particular, Γ can only contain a loop if it arises from $\mathcal{G}_4$ or its analog in a previous iteration. Hence, the construction cannot yield a graph with more than two loops or more than one loop per circle, respectively. In particular, any graph in $\mathcal{G}(k,\ell)$ containing more than one loop is necessarily obtained from an element in $\mathcal{G}(1,1)$ with two loops.

(c) Again, the elements of $\mathcal{G}(S_1, \emptyset) = \emptyset$ and $\mathcal{G}(\emptyset, S_2) = \emptyset$ with $S_1 \subseteq [k]$ and $S_2 \subseteq [k+1, k+\ell]$, as well as all $\Gamma \in \mathcal{G}(1,1)$ have the claimed structure. We further note that any double edges of $\Gamma \in \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$ must occur in the subgraph in $\mathcal{G}(S_1, [k+1, k+\ell])$ with $S_1 \subset [k]$ used to construct Γ and that the induction hypothesis applies to the latter. For $\Gamma \in \mathcal{G}_4$, a double edge occurs if and only if the corresponding element of $NCG(\{1, \ldots, k, k+j, \ldots, k+j-1, k+j\})$ has a vertex i that shares an edge with both copies of $k+j$. However, there is at most one such vertex in $[k]$ and $[k+1, k+\ell]$, respectively, as having two vertices i, i' connected to both copies of $k+j$ in either set induces a crossing. In particular, any double edge shares a vertex with the edge $(1, k+j) \in \Gamma$ prescribed by the definition. Hence, all graphs in $\mathcal{G}_{\neg(1,k)}$ have the claimed structure.

It remains to consider $\mathcal{G}_{(1,k)}$, i.e., to add an edge $(1, k)$ to the graphs considered previously. In the case $j = k$ of $\mathcal{G}_2$, this doubles an existing $(1, k)$ edge. By part (a), any such graph must also have an edge connecting k with a vertex in $[k+1, k+\ell]$. This shows that the placement rule for double edges is satisfied. Further, at most one doubled edge can be added with each application of the recursion, i.e., there are at most $k+\ell-1$ double edges in total. It is readily checked that two is indeed the highest edge multiplicity possible.

(d) As the case $k = \ell = 1$ has already been discussed in Example 3.3.4, let $k = 2$, $\ell = 1$, and consider $\tau(NCG(\{1, 2, 3, 3\}))$. We relabel the vertices as $1, 2, 3, 3'$ to distinguish between the two copies of the doubled vertex 3 more easily. Observe that the graphs Γ_1 and Γ_2 with edge sets $\{(1, 3'), (2, 3')\}$ and $\{(1, 3'), (2, 3)\}$, respectively, give rise to the same element of $\mathcal{G}(2, 1)$, namely the annulus graph with edge set $\{(1, 3), (2, 3)\}$ (see Fig. 3.13 below).

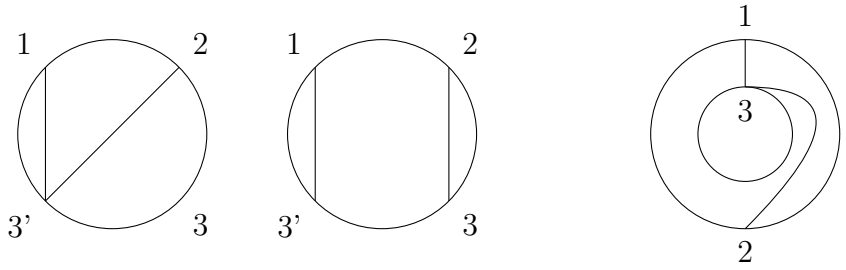

Fig. 3.13. Two disk non-crossing graphs that give rise to the same annulus graph.

Hence, $\mathcal{G}(1, 2)$ is indeed a multi-set. By Definition 3.3.2, either $\mathcal{G}(1, 2)$ or $\mathcal{G}(2, 1)$ is used in the construction of $\mathcal{G}(k, \ell)$ for $k, \ell > 2$, i.e., the construction yields again a multi-set. □

For the following discussion, we introduce the disjoint decomposition

$$\mathcal{G}(k, \ell) = \mathcal{G}_{dec}(k, \ell) \cup \mathcal{G}_{\neg dec}(k, \ell), \tag{3.60}$$

where $\mathcal{G}_{dec}(k, \ell)$ denotes the graphs in $\mathcal{G}(k, \ell)$ that have double edges or loops and $\mathcal{G}_{\neg dec}(k, \ell)$ denotes the graphs that do not have either. We refer to the elements of $\mathcal{G}_{dec}(k, \ell)$ as *decorated* graphs.

Lemma 3.3.6. *Let $k, \ell \in \mathbb{N}$.*

(a) Whenever $\Gamma \in \mathcal{G}(k, \ell)$ has more than one connected component that contains a connecting edge, every connected component of Γ can be uniquely identified with a disk non-crossing graph.

(b) Whenever $\Gamma \in \mathcal{G}(k,\ell)$ has exactly one connected component with a connecting edge and Γ is not decorated, the same identification as in (a) holds, but it is no longer unique. If Γ_1 denotes the connected component of Γ that contains a connecting edge and $U_1 \cup U_2$ with $U_1 \subset [k]$ and $U_2 \subset [k+1, k+\ell]$ is the vertex set of Γ_1, there are $|U_1| \cdot |U_2|$ different ways to identify the connected components of Γ with a disk non-crossing graph.

Lemma 3.3.6 translates between a graph $\Gamma \in \mathcal{G}(k,\ell)$, the partition induced by its connected components and the cycle structure arising in (3.43). We give a schematic of the construction of the disk graphs in Fig. 3.14 below. Recall from Definition 3.1.11 that any cycle of an annular non-crossing permutation encloses a region homeomorphic to the unit disk with the boundary oriented clockwise.

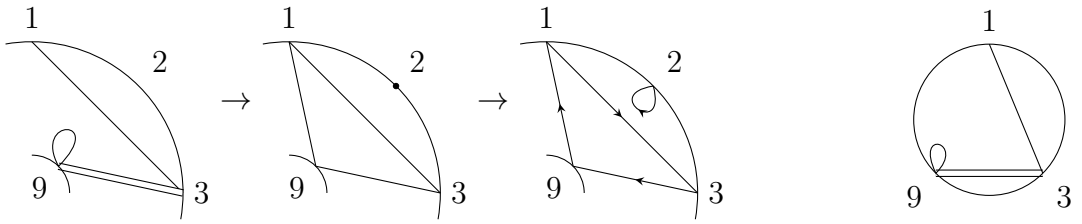

Fig. 3.14. A subgraph of $\Gamma \in \mathcal{G}(k,\ell)$ with the induced partition π_Γ and a possible permutation π'_Γ (left) as well as the disk non-crossing graph obtained for a connected component of Γ (right).

Note that the assignment of the orientation in the second step on the left of Fig. 3.14 is not unique if π_Γ has only one connecting block (cf. [112, Prop. 4.6], see Fig. 3.5 for an example). We remark that (a) and (b) are almost complementary cases and that only decorated graphs with exactly one connected component containing a connecting edge are not covered by Lemma 3.3.6. The proof of (a) further shows that any $\Gamma \in \mathcal{G}(k,\ell)$ with at least two connected components containing a connecting edge cannot be decorated.

We give the full construction behind the schematics in Fig. 3.14 below.

Proof of Lemma 3.3.6. (b) It follows from Lemma 3.3.5(a) that the connected components of Γ give rise to a partition $\pi_\Gamma \in NCP(k,\ell)$. By Propositions 4.4. and 4.6 of [112], any element of $NCP(k,\ell)$ may be identified with an annular non-crossing permutation, i.e., its blocks can be given an orientation (see the left of Fig. 3.14). This orientation of the individual cycles is naturally induced by the orientation of the inner and outer circle. However, the identification between $NCP(k,\ell)$ and $\overrightarrow{NCP}(k,\ell)$ is only unique whenever the underlying partition has more than one connecting block. If there is only one connecting block $U_1 \cup U_2$ with $U_1 \subset [k]$, $U_2 \subset [k+1, k+\ell]$, there are $|U_1| \cdot |U_2|$ possibilities to identify π_Γ with an annular non-crossing permutation (cf. Fig. 3.5). We fix one $\pi'_\Gamma \in \overrightarrow{NCP}(k,\ell)$ that is associated with π_Γ in this way.

By definition, the cycles of π'_Γ can be drawn on the (k,ℓ)-annulus such that each cycle encloses a region between the circles homeomorphic to the disk with boundary oriented clockwise. Adding the elements of the cycle around the boundary yields a labeled disk as in Definition 3.1.7. We may use the same transformation to map a connected component of Γ to a disk non-crossing graph (see Fig. 3.15 below). Note that this transformation cannot induce any crossings of the edges of Γ, however, any loops or double edges of the

original graph are kept. Hence, we only obtain a disk non-crossing graph in the sense of Definition 3.1.7 if Γ does not contain any loops or double edges to begin with.

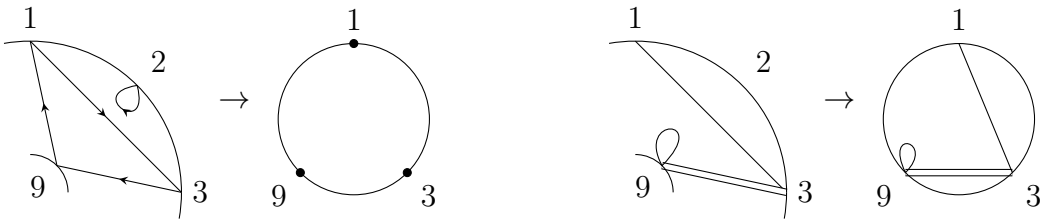

Fig. 3.15. Transformation of a cycle of $\pi'_\Gamma \in \overrightarrow{NCP}(k,\ell)$ to a circle (left) and the induced transformation of a connected component of $\Gamma \in \mathcal{G}(k,\ell)$ into a disk graph (right).

(a) Assume next that $\Gamma \in \mathcal{G}(k,\ell)$ has at least two connected components that contain a connecting edge. It follows from Definition 3.3.2 that this structure can only arise from $\mathcal{G}_4$ in (3.56) or its analog in a previous iteration of the recursive definition. Since adding subgraphs that only live on one circle of the (k,ℓ)-annulus does not interfere with the following argument, assume w.l.o.g. that Γ arises from $\mathcal{G}_4$ directly. To have two connecting blocks, Γ must have at least two connecting edges (i_1,i_2) and (i'_1,i'_2) where $i_1,i'_1 \in [k]$, $i_2,i'_2 \in [k+1,k+\ell]$ and $i_1 \neq i'_1$, $i_2 \neq i'_2$. Note that either (i_1,i_2) or (i'_1,i'_2) may coincide with the edge $(1,k+j)$ prescribed by the definition.

Let $\widetilde{\Gamma} \in NCG(\{1,\dots,k,k+j,\dots,k+j-1,k+j\})$ denote a disk non-crossing graph such that $\tau(\widetilde{\Gamma}) = \Gamma$. Here, τ denotes the map introduced in Definition 3.3.1. By construction, $\widetilde{\Gamma}$ also has two edges edges (i_1,i_2) and (i'_1,i'_2) with $i_1,i'_1 \in [k]$, $i_2,i'_2 \in [k+1,k+\ell]$ and $i_1 \neq i'_1$, $i_2 \neq i'_2$. This structure imposes several restrictions on $\widetilde{\Gamma}$, as can be seen from Fig. 3.16 below.

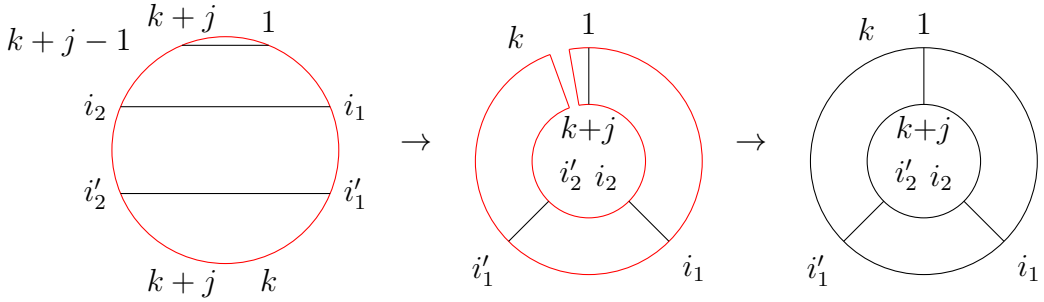

Fig. 3.16. A schematic visualization of $\widetilde{\Gamma}$ (left) and its mapping to the original graph $\Gamma = \tau(\widetilde{\Gamma})$ (right).

First, there cannot be an edge $(1,k) \in \widetilde{\Gamma}$ without violating the non-crossing condition. Together with (3.57), this implies that $\Gamma = \tau(\widetilde{\Gamma})$ has at most a single edge $(1,k)$. Further, $\widetilde{\Gamma}$ cannot contain a vertex that connects to both copies of $k+j$. This implies that $\Gamma = \tau(\widetilde{\Gamma})$ cannot have any double edges. Lastly, Γ cannot have any loops, as $\widetilde{\Gamma}$ containing an edge $(k+j,k+j)$ would induce a crossing, too. Hence, any $\Gamma \in \mathcal{G}(k,\ell)$ with more than one connected component containing a connecting edge has only single edges and no loops. □

So far, we have only considered the non-crossing annular partition induced by an element of $\mathcal{G}(k,\ell)$. The following lemma allows us to partially reverse this relation and explicitly construct a graph that is associated with a given $\pi \in NCP(k,\ell)$.

Lemma 3.3.7. *For every $\pi \in NCP(k,\ell)$ there is at least one $\Gamma \in \mathcal{G}(k,\ell)$ for which the vertex sets of the connected components coincide with the blocks of π. If π has exactly one connecting block, then there is at least one such graph in $\mathcal{G}_{dec}(k,\ell)$ and one in $\mathcal{G}_{\neg dec}(k,\ell)$, respectively.*

We briefly sketch the construction of an element on $\mathcal{G}(k,\ell)$ from a given annular non-crossing partition. For the example, we assume that the sketched connecting block is the only one in the partition. After completing the steps sketched in Fig. 3.17, the remaining connected components of the graph are readily added using steps corresponding to $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$ in Definition 3.3.2 and the symmetry under interchanging of the inner and outer circle.

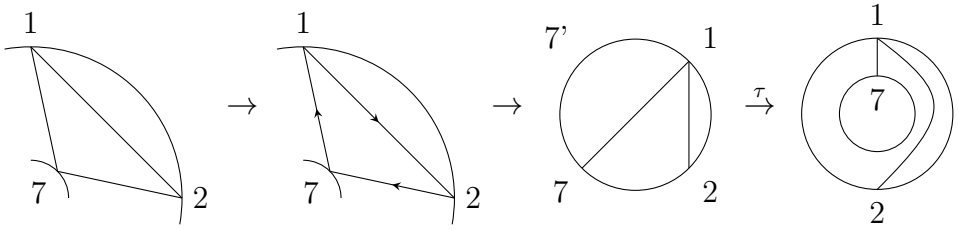

Fig. 3.17. The construction of an element of $\mathcal{G}(2,1)$ (right) from the connecting block of an annular non-crossing partition (left).

Proof of Lemma 3.3.7. Fix $\pi \in NCP(k,\ell)$ and assume first that π has at least two connecting blocks. W.l.o.g. let 1 and k be assigned to different blocks of π. Indeed, if 1 and k occurred in the same block, we may split it into two disjoint parts that contain 1 and k, respectively, and later add an edge $(1,k)$ to the graph obtained from this modified partition. Further, assume w.l.o.g. that 1 is contained in a connecting block. If $i > 1$ were the smallest element that occurs in a connecting block, we may further remove any blocks containing $\{1,\dots,i-1\}$ from the partition and later add a suitable subgraph. Note that the latter is possible by only using steps corresponding to $\mathcal{G}_1$ and $\mathcal{G}_2$ in Definition 3.3.2.

Under these assumptions, there exists $\widetilde{\pi} \in NCP(\{1,\dots,k,k+j,\dots,k+j-1,k+j\})$ such that the transformation τ in Definition 3.3.1 maps the blocks of $\widetilde{\pi}$ to the blocks of π. This is readily seen from Fig. 3.9, as 1 and k being in different blocks implies that none of the blocks of π intersects the slit between 1 and k. We pick j such that 1 and $k+j$ are in the same block of π and may further choose $\widetilde{\pi}$ such that one copy of the doubled label $k+1$ occurs as a singleton set. Finally, define $\widetilde{\Gamma} \in NCG(\{1,\dots,k,k+1,\dots,k+\ell,k+1\})$ by considering each block $B = \{i_1,\dots,i_n\} \in \widetilde{\pi}$ separately and adding the edges $(i_1,i_2),\dots,(i_1,i_n)$ to the graph. By construction, the vertex sets of the connected components of $\widetilde{\Gamma}$ coincide with the blocks of $\widetilde{\pi}$. Considering $\Gamma_\pi = \tau(\widetilde{\Gamma})$ now yields the element of $\mathcal{G}(k,\ell)$ with the claimed properties (cf. Fig. 3.17, where the procedure is sketched for a single block).

Next, consider $\pi \in NCP(k,\ell)$ that has only one connecting block $U = U_1 \cup U_2$ with $U_1 \subseteq [k]$, $U_2 \subseteq [k+1,k+\ell]$. Similar to the first case, we can construct a disk non-crossing graph Γ_U for which $\tau(\Gamma_U) \in \mathcal{G}_c(|U_1|,|U_2|)$. As $\tau(\Gamma_U)$ is only required to have one connecting edge, there is also a choice for Γ_U containing a $(k+j,k+j)$ edge (cf. Fig. 3.15). In particular, we obtain at least one graph in $\mathcal{G}_{c,\neg dec}(|U_1|,|U_2|)$ and one graph in $\mathcal{G}_{c,dec}(|U_1|,|U_2|)$, respectively. The remaining connected subgraphs of Γ_π are then added recursively by alternating between adding an isolated vertex (cf. $\mathcal{G}_1$) and a disk non-crossing graph (cf. $\mathcal{G}_2$ and $\mathcal{G}_3$) to $\tau(\Gamma_U)$. Note that starting from $\tau(\Gamma_U) \in \mathcal{G}_{c,\neg dec}(|U_1|,|U_2|)$

yields a graph that is not decorated while starting from $\tau(\Gamma_U) \in \mathcal{G}_{c,dec}(|U_1|, |U_2|)$ yields a decorated graph. □

Using the properties of $\Gamma \in \mathcal{G}(k,\ell)$ from Lemmas 3.3.6 and 3.3.7, we obtain an explicit non-recursive combinatorial formula for $m_{\circ\circ}$.

Lemma 3.3.8. *Let $k, \ell \in \mathbb{N}$. Then,*

$$m_{\circ\circ}[1,\dots,k|k+1,\dots,k+\ell] = \Big(\prod_{s=1}^{k+\ell} m_s\Big) \sum_{\Gamma\in\mathcal{G}_c(k,\ell)} c_\Gamma q_\Gamma, \tag{3.61}$$

with suitable constants $c_\Gamma \in \mathbb{Z}$. In particular, $c_\Gamma = 1$ for $\Gamma \in \mathcal{G}_{dec}(k,\ell)$.

Note that the constants c_Γ for $\Gamma \in \mathcal{G}_{\neg dec}(k,\ell)$ are readily obtained from the multiplicity of the corresponding graph in the multi-set, however, their exact values are not needed for the proof of Theorem 3.2.4.

Proof. We start by observing that any connected component of $\Gamma \in \mathcal{G}(k,\ell)$ with a connecting edge is itself an annular non-crossing graph. Further, a connected component of Γ that only involves vertices from either $[k]$ or $[k+\ell]$ cannot contain loops or double edges (cf. (b) and (c) of Lemma 3.3.5) and we may identify it with a disk non-crossing graph.

To simplify notation, decompose any $B \subseteq S_1 \cup S_2$ as a union $B_1 \cup B_2$ with $B_1 \subseteq [k]$, $B_2 \subseteq [k+1, k+\ell]$ and associate it with a tuple (B_1, B_2) if neither of the subsets is empty. This allows us to use the common notation $NCG(B)$ for both disk and annular graphs by setting $NCG(B) = \mathcal{G}(B_1, B_2)$ whenever B contains elements from both $[k]$ and $[k+1, k+\ell]$, and $NCG(B) = NCG(B_1)$ resp. $NCG(B) = NCG(B_2)$ if it does not. Splitting the sum in (3.58) according to the underlying partition, we can write

$$\begin{aligned}\widetilde{m}[1,\dots,k|k+1,\dots,k+\ell] &= \Big(\prod_{s=1}^{k+\ell} m_s\Big) \sum_{\Gamma\in\mathcal{G}(k,\ell)} q_\Gamma \\ &= \sum_{\pi\in NCP(k,\ell)} \prod_{B\in\pi} \Big[\Big(\prod_{s\in B} m_s\Big) \sum_{\Gamma\in NCG_c(B)} q_\Gamma\Big] \end{aligned} \tag{3.62}$$

where $NCG_c(B)$ denotes the connected graphs in $NCG(B)$. By Lemma 3.3.7, none of the sums on the right-hand side are empty. However, there may be multiple permutations in $\overrightarrow{NCP}(k,\ell)$ as well as a marked element of $NCP(k) \times NCP(\ell)$ associated with a given annular non-crossing partition $\pi \in NCP(k,\ell)$. To obtain the same structure as in (3.26), we thus need to decompose the sum over $\pi \in NCP(k,\ell)$ on the right-hand side of (3.62) further.

Distinguishing by the number of connecting blocks of π yields

$$\begin{aligned}&\sum_{\pi\in NCP(k,\ell)} \prod_{B\in\pi} \Big[\Big(\prod_{s\in B} m_s\Big) \sum_{\Gamma\in NCG_c(B)} q_\Gamma\Big] \\ &= \sum_{\substack{\pi\in NCP(k,\ell),\\ 1 \text{ conn. block}}} \prod_{B\in\pi} \Big[\Big(\prod_{s\in B} m_s\Big) \sum_{\Gamma\in NCG_c(B)} q_\Gamma\Big] + \sum_{\substack{\pi\in NCP(k,\ell),\\ \geq 2 \text{ conn. blocks}}} \prod_{B\in\pi} \Big[\Big(\prod_{s\in B} m_s\Big) \sum_{\Gamma\in NCG_c(B)} q_\Gamma\Big].\end{aligned} \tag{3.63}$$

As an annular non-crossing partition with at least two connecting blocks uniquely corresponds to an annular non-crossing permutation (cf. [112, Prop. 4.4]), we may replace

$NCP(k,\ell)$ by $\overrightarrow{NCP}(k,\ell)$ in the summation if we interpret a cycle as an ordered set. Together with Lemma 3.3.6 and (3.17), it follows that

$$\begin{aligned}
&\sum_{\substack{\pi\in NCP(k,\ell),\\ \geq 2 \text{ conn. blocks}}} \prod_{B\in\pi}\Big[\Big(\prod_{s\in B} m_s\Big)\sum_{\Gamma\in NCG_c(B)} q_\Gamma\Big] \\
&= \sum_{\substack{\pi\in \overrightarrow{NCP}(k,\ell),\\ \geq 2 \text{ conn. cycles}}} \prod_{B\in\pi}\Big[\Big(\prod_{s\in B_1\cup B_2} m_s\Big)\sum_{\Gamma\in NCG_c(B_1\cup B_2)} q_\Gamma\Big] \qquad (3.64)\\
&= \sum_{\substack{\pi\in \overrightarrow{NCP}(k,\ell),\\ \geq 2 \text{ conn. cycles}}} \prod_{B\in\pi} m_\circ[B_1\cup B_2].
\end{aligned}$$

where we used that $m_\circ$ is invariant under permutation of the spectral parameters $(z_j)_j$.

Applying a similar argument for the partitions with one connecting block U, interpreted as (U_1,U_2) with $U_1 = U\cap[k], U_2 = U\cap[k+1,k+\ell]$, yields

$$\begin{aligned}
&\sum_{\substack{\pi\in NCP(k,\ell),\\ 1 \text{ conn. block}}} \prod_{B\in\pi}\Big[\Big(\prod_{s\in B} m_s\Big)\sum_{\Gamma\in NCG_c(B)} q_\Gamma\Big] \\
&= \sum_{\substack{\pi\in NCP(k,\ell),\\ 1 \text{ conn. block}}} \Big[\Big(\prod_{s\in U} m_s\Big)\sum_{\Gamma\in\mathcal{G}_c(U)} q_\Gamma\Big]\prod_{B\in\pi\setminus\{U\}}\Big[\Big(\prod_{s\in B} m_s\Big)\sum_{\Gamma\in NCG_c(B)} q_\Gamma\Big] \\
&= \sum_{\substack{\pi\in NCP(k,\ell),\\ 1 \text{ conn. block}}} \Big[\Big(\prod_{s\in U} m_s\Big)\sum_{\Gamma\in\mathcal{G}_c(U)} q_\Gamma\Big]\prod_{B\in\pi\setminus\{U\}} m_\circ[B] \qquad (3.65)
\end{aligned}$$

by (3.17). Note that $B\neq U$ cannot involve both sets $[k]$ and $[k+1,k+\ell]$, as U is the only connecting block. Using (3.60), we decompose

$$\sum_{\Gamma\in\mathcal{G}_c(U)} q_\Gamma = \sum_{\Gamma\in\mathcal{G}_{c,dec}(U)} q_\Gamma + \sum_{\Gamma\in\mathcal{G}_{c,\neg dec}(U)} q_\Gamma$$

and recall from Lemma 3.3.7 that neither sum on the right-hand side is empty. Note that the split induced by (3.60) also decomposes the right-hand side of (3.65) into two terms.

Next, consider the term corresponding to $\mathcal{G}_{c,\neg dec}(U)$ and recall that any $\Gamma\in\mathcal{G}_{c,\neg dec}(U)$ can be identified with a disk non-crossing graph by Lemma 3.3.6(b). However, this identification is not unique. Decompose U into $U_1 = U\cap[k], U_2 = U\cap[k+1,k+\ell]$. Then there are $|U_1|\cdot|U_2|$ different disk graphs that can be obtained from a given $\Gamma\in\mathcal{G}_{c,\neg dec}(U)$. As the resulting graphs only differ in the labeling of the vertices and $m_\circ$ is invariant under permutation of its arguments, the contribution of each graph to the sum is the same. Recall that the number of annular non-crossing permutations arising from $\pi\in NCP(k,\ell)$ is equal to $|U_1|\cdot|U_2|$ by [112, Prop. 4.6]. We thus write

$$\sum_{\Gamma\in\mathcal{G}_{c,\neg dec}(U)} q_\Gamma = |U_1|\cdot|U_2| \sum_{\Gamma\in NCG_c(U_1\cup U_2)} q_\Gamma + \sum_{\Gamma\in\mathcal{G}_{c,\neg dec}(U)} c_\Gamma q_\Gamma.$$

with suitable constants $c_\Gamma\in\mathbb{Z}$. In particular, we do not necessarily have $c_\Gamma = 1$. This can be seen from considering the element of $\mathcal{G}(2,1)$ that has the edge set $\{(1,2),(1,3),(2,3)\}$. Hence,

$$\sum_{\Gamma\in\mathcal{G}_c(U)} q_\Gamma = |U_1|\cdot|U_2| \sum_{\Gamma\in NCG_c(U_1\cup U_2)} q_\Gamma + \sum_{\Gamma\in\mathcal{G}_c(U)} c_\Gamma q_\Gamma. \qquad (3.66)$$

where we set $c_\Gamma = 1$ for $\Gamma \in \mathcal{G}_{c,dec}(U)$. Note that the contribution of the first term of (3.66) to (3.65) is

$$\begin{aligned}&\sum_{\substack{\pi\in NCP(k,\ell),\\ 1 \text{ conn. block } U}} \Big[\Big(\prod_{s\in U} m_s\Big)\cdot|U_1|\cdot|U_2| \sum_{\Gamma\in NCG_c(U_1\cup U_2)} q_\Gamma\Big] \prod_{B\in\pi\setminus\{U\}} m_\circ[B]\\ &= \sum_{\substack{\pi\in\overrightarrow{NCP}(k,\ell),\\ 1 \text{ conn. block}}} \prod_{B\in\pi} m_\circ[B_1\cup B_2]. \end{aligned} \tag{3.67}$$

Putting everything together, the right-hand side of (3.63) reads

$$\begin{aligned}\widetilde{m}[1,\dots,k|k+1,\dots,k+\ell] &= \sum_{\pi\in\overrightarrow{NCP}(k,\ell)} \prod_{B\in\pi} m_\circ[B] \\ &+ \sum_{\substack{\pi\in NCP(k)\times NCP(\ell),\\ U_1,U_2 \text{ marked}}} \prod_{B\in\pi\setminus\{U_1,U_2\}} m_\circ[B_1\cup B_2]\Big[\Big(\prod_{s\in U} m_s\Big)\sum_{\Gamma\in\mathcal{G}_c(U)} c_\Gamma q_\Gamma\Big].\end{aligned} \tag{3.68}$$

with $c_\Gamma \in \mathbb{Z}$ as in (3.66) and $U = U_1 \cup U_2$. The first term on the right-hand side of (3.68) is the sum of (3.64) and (3.67). The second term is obtained from the second term in (3.66) by noting that U is the only connecting block of the partition, i.e., any block of $\pi \setminus \{U\}$ only lives on one of the circles. Observing that this matches the structure in (3.26), we obtain (3.61) by comparing term-by-term. □

We finally have all the necessary tools to prove Theorem 3.2.4.

3.3.2 Step 2: Conclusion

Proof of Theorem 3.2.4. Let $\mathfrak{f}$ denote the right-hand side of (3.43). We recall that the initial condition (3.35) is immediate from Definition 3.3.2(ii) and that the symmetry in Definition 3.2.1(i) follows from the remark on Theorem 3.2.4 above. Hence, it only remains to check that $\mathfrak{f}$ satisfies the recursion (3.36). To simplify notation, we interpret $\mathfrak{f}$ as a function of the multi-indices α and β.

Similar to the proof of [39, Lem. 4.4], we use (3.17) and (3.61) to write out the first and second-order free cumulant functions in terms of suitable graphs. This yields

$$\begin{aligned}\frac{\mathfrak{f}[\alpha|\beta]}{m_1\dots m_{k+\ell}} &= \sum_{\pi\in\overrightarrow{NCP}(k,\ell)} \Big(\prod_{B\in K(\pi)} \Big\langle \prod_{j\in B} A_j\Big\rangle\Big) \prod_{B\in\pi}\Big(\sum_{\Gamma\in NCG_c(B_1\cup B_2)} q_\Gamma\Big) \\ &+ \sum_{\substack{\pi\in NCP(k)\times NCP(\ell),\\ U_1,U_2 \text{ marked}}} \Big(\prod_{B\in K(\pi)} \Big\langle \prod_{j\in B} A_j\Big\rangle\Big)\Big(\sum_{\Gamma\in NCG_c(U_1,U_2)} c_\Gamma q_\Gamma\Big) \\ &\times \prod_{B\in\pi\setminus\{U_1,U_2\}}\Big(\sum_{\Gamma\in NCG_c(B_1\cup B_2)} q_\Gamma\Big).\end{aligned} \tag{3.69}$$

To collect the terms involving the deterministic matrices $A_1,\dots,A_k$, we define

$$\mathcal{F}(\pi) := \prod_{B\in K(\pi')} \Big\langle \prod_{j\in B} A_j\Big\rangle \tag{3.70}$$

for any $\pi \in NCP(k,\ell)$ that has more than one connecting block. Here, $\pi' \in \overrightarrow{NCP}(k,\ell)$ denotes the unique permutation for which the blocks of π' coincide with π. Further, we

define

$$\mathcal{F}(\pi) := \sum_{\substack{\pi' \in \overrightarrow{NCP}(k,\ell) \\ blocks(\pi')=\pi}} \Big(\prod_{B \in K(\pi')} \Big\langle \prod_{j \in B} A_j \Big\rangle \Big) + \prod_{B \in K(\pi'')} \Big\langle \prod_{j \in B} A_j \Big\rangle \tag{3.71}$$

for any $\pi \in NCP(k,\ell)$ that has exactly one connecting block U. Here, π'' is the marked partition obtained from π by splitting the connecting block into $U_1 = U \cap [k]$ and $U_2 = U \cap [k+1, k+\ell]$, and marking U_1, U_2 on the respective circles.

Next, we decompose the sum over $\overrightarrow{NCP}(k,\ell)$ in (3.69) according to the number of connecting cycles in the permutation and rewrite the right-hand side as

$$\frac{\mathfrak{f}[\alpha|\beta]}{m_1 \dots m_{k+\ell}} = \sum_{\pi \in NCP(k,\ell)} \mathcal{F}(\pi) \prod_{B \in \pi} \Big(\sum_{\Gamma \in NCG_c(B)} q_\Gamma \Big)$$

using (3.70) and (3.71). Recall that any connected component of $\Gamma \in \mathcal{G}(k,\ell)$ can either be identified with a disk non-crossing graph or itself satisfies Definition 3.3.2. We can thus interpret $NCG_c(B)$ as a connected component of a bigger graph and rewrite

$$\sum_{\pi \in NCP(k,\ell)} \mathcal{F}(\pi) \prod_{B \in \pi} \Big(\sum_{\Gamma \in NCG_c(B)} q_\Gamma \Big) = \sum_{\Gamma \in \mathcal{G}(k,\ell)} \mathcal{F}(\pi_\Gamma) q_\Gamma. \tag{3.72}$$

Here, π_Γ denotes the partition arising from the vertex sets of the connected components of Γ.

Using (3.57) from Definition 3.3.2, decompose the right-hand side of (3.72) as

$$\sum_{\Gamma \in \mathcal{G}(k,\ell)} \mathcal{F}(\pi_\Gamma) q_\Gamma = \sum_{\Gamma \in \mathcal{G}_{\neg(1,k)}} q_\Gamma \Big(\mathcal{F}(\pi_\Gamma) + q_{1,k} \mathcal{F}(\pi_{\Gamma \cup \{(1,k)\}}) \Big). \tag{3.73}$$

Recall that $\Gamma \cup \{(1,k)\}$ denotes the graph obtained from adding an edge $(1,k)$ to Γ and that, therefore, $q_{\Gamma \cup \{(1,k)\}} = q_{1,k} q_\Gamma$. Next, apply (3.56) to split the right-hand side of (3.73) into contributions corresponding to $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_3$, and $\mathcal{G}_4$. This yields

$$\begin{aligned}
\sum_{\Gamma \in \mathcal{G}(k,\ell)} \mathcal{F}(\pi_\Gamma) q_\Gamma = & \sum_{\Gamma \in \mathcal{G}([2,k],[k+1,k+\ell])} q_\Gamma \Big(\mathcal{F}(\pi_\Gamma) + q_{1,k} \mathcal{F}(\pi_{\Gamma \cup \{(1,k)\}}) \Big) \\
& + \sum_{j=2}^{k} \sum_{\Gamma \in NCG_{(1,j)}([1,j]) \times \mathcal{G}([j,k],[k+1,k+\ell])} q_\Gamma \Big(\mathcal{F}(\pi_\Gamma) + q_{1,k} \mathcal{F}(\pi_{\Gamma \cup \{(1,k)\}}) \Big) \\
& + \sum_{j=1}^{k-1} \sum_{\Gamma \in \mathcal{G}_{(1,j)}([1,j],[k+1,k+\ell]) \times NCG([j,k])} q_\Gamma \Big(\mathcal{F}(\pi_\Gamma) + q_{1,k} \mathcal{F}(\pi_{\Gamma \cup \{(1,k)\}}) \Big) \\
& + \sum_{j=1}^{\ell} \sum_{\Gamma \in NCG_{(1,k+j)}(\{1,k,k+j,\dots,k+\ell,\dots,k+j\})} q_\Gamma \Big(\mathcal{F}(\pi_\Gamma) + q_{1,k} \mathcal{F}(\pi_{\Gamma \cup \{(1,k)\}}) \Big),
\end{aligned} \tag{3.74}$$

where the edges prescribed in the definitions of $\mathcal{G}_2$, $\mathcal{G}_3$, and $\mathcal{G}_4$ are added as a subscript to NCG and $\mathcal{G}$, respectively. It remains to compare (3.74) with (3.36).

Computation of the first line of (3.74): Recall that $\sum_{\Gamma \in \mathcal{G}_1} q_\Gamma = \sum_{\Gamma \in \mathcal{G}([2,k],[k+1,k+\ell])} q_\Gamma$. As any $\Gamma \in \mathcal{G}_1$ has an isolated vertex 1, the set $\{1\}$ appears as a singleton block of the underlying partition π_Γ. Let π'_Γ denote an annular non-crossing permutation with blocks

given by π_Γ. Then $(\dots k1\dots) \in K(\pi'_\Gamma)$, i.e., 1 and k appear in the same cycle of the Kreweras complement. This gives

$$\prod_{B\in K(\pi'_\Gamma)} \Big\langle \prod_{j\in B} A_j \Big\rangle = \prod_{B\in K(\pi'_\Gamma)|_{(1,k]\cup[k+1,k+\ell]}} \Big\langle \prod_{j\in B} A'_j \Big\rangle \tag{3.75}$$

where $A'_j = A_j$ for $j = 2,\dots k-1$ and $j \in [k+1, k+\ell]$, but $A'_k = A_k A_1$. Considering $\mathcal{F}(\pi_{\Gamma\cup(1,k)})$, note that adding an edge $(1,k)$ to the graph Γ implies that $(k) \in K(\pi_{\Gamma\cup\{(1,k)\}})$, i.e., $\langle A_k\rangle$ always occurs as a separate factor in $\mathcal{F}$. Hence, whenever π_Γ has more than one connecting block we can use (3.70) to evaluate

$$\begin{aligned}&\mathcal{F}(\pi_\Gamma) + q_{1,k}\mathcal{F}(\pi_{\Gamma\cup\{(1,k)\}})\\ &= \prod_{B\in K(\pi'_\Gamma)|_{\langle 1,k]\cup[k+1,k+\ell]}} \Big\langle \prod_{j\in B} A'_j \Big\rangle + q_{1,k}\langle A_k\rangle \prod_{B\in K(\pi'_\Gamma)|_{(1,k)\cup[k+1,k+\ell]}} \Big\langle \prod_{j\in B} A_j \Big\rangle\end{aligned} \tag{3.76}$$

with A'_j as in (3.75). In the remaining cases, π_Γ has only one connecting block U, and $\mathcal{F}$ is a sum of two terms. As the first term of (3.71) can be evaluated similarly to (3.76), we only consider the second term. Let π''_Γ denote the element of $NCP(k)\times NCP(\ell)$ in which the blocks $U_1 = U\cap[k]$ and $U_2 = U\cap[k+1,k+\ell]$ are marked. Recall that the marking does not influence the Kreweras complement, which is taken for both circles separately here. As 1 and k lie on the same circle, we can argue as in the permutation case. It follows that

$$\begin{aligned}&\Big(\prod_{s=1}^{k+\ell} m_s\Big) \sum_{\Gamma\in\mathcal{G}([2,k],[k+1,k+\ell])} q_\Gamma\Big(\mathcal{F}(\pi_\Gamma) + q_{1,k}\mathcal{F}(\pi_{\Gamma\cup\{(1,k)\}})\Big)\\ &= m_1\Big(F[T_2,\dots,T_{k-1},G_kA_kA_1|T_{k+1},\dots,T_{k+\ell}]\\ &\quad + q_{1,k}F[T_2,\dots,T_{k-1}G_kA_1|T_{k+1},\dots,T_{k+\ell}]\langle A_k\rangle\Big).\end{aligned} \tag{3.77}$$

Computation of the second line of (3.74): For the contribution arising from $\mathcal{G}_2$, recall that $\Gamma = \Gamma_1\cup\Gamma_2$ with $\Gamma_1 \in NCG_{(1,j)}([1,j])$ and $\Gamma_2 \in \mathcal{G}([j,k],[k+1,k+\ell])$ implies $q_\Gamma = q_{\Gamma_1}q_{\Gamma_2}$, i.e., the weights factorize. Further, the term involving Γ_1 evaluates to

$$\sum_{\Gamma_1\in NCG_{(1,j)}([1,j])} q_{\Gamma_1} = \frac{q_{1,j}}{1+q_{1,j}} \sum_{\Gamma_1\in NCG[1,j])} q_{\Gamma_1} = m_1m_j \sum_{\Gamma_1\in NCG[1,j])} q_{\Gamma_1} \tag{3.78}$$

as q_{Γ_1} always includes a factor $q_{1,j}$ if $\Gamma_1 \in NCG_{(1,j)}([1,j])$.

Let $\pi_\Gamma \in NCP(j)\times NCP(k-j+1,\ell)$ denote the partition associated with a graph $\Gamma \in NCG_{(1,j)}([1,j])\times\mathcal{G}([j,k],[k+1,k+\ell])$ in the second term of (3.74) and let π'_Γ is be an annular non-crossing permutation with blocks given by π_Γ. Since the edge $(1,j)$ must occur in Γ, the vertices 1 and j must be associated with same cycle of π'_Γ. Hence, the elements in $[1,j\rangle$ and $[j,k]$ are in different cycles of $K(\pi'_\Gamma)$. Moreover, none of $1,\dots,j-1$ can be part of a connecting cycle in $K(\pi'_\Gamma)$. This implies the decomposition

$$\prod_{B\in K(\pi'_\Gamma)} \Big\langle \prod_{j\in B} A_j \Big\rangle = \Big(\prod_{B\in K(\pi'_\Gamma|_{[1,j]})} \Big\langle \prod_{i\in B\setminus\{j\}} A_i \Big\rangle\Big)\Big(\prod_{B\in K(\pi'_\Gamma|_{[j,k]\cup[k+1,k+\ell]})} \Big\langle \prod_{i\in B} A_i \Big\rangle\Big).$$

Again, the only difference between $\mathcal{F}(\pi_\Gamma)$ and $\mathcal{F}(\pi_{\Gamma\cup\{(1,k)\}})$ is the fact that $\langle A_k\rangle$ must occur as a separate factor in the second case. Thus, we can argue as in (3.76) and evaluate

$$\begin{aligned}\mathcal{F}(\pi_\Gamma)+q_{1,k}\mathcal{F}(\pi_{\Gamma\cup\{(1,k)\}}) &= \Big(\prod_{B\in K(\pi_\Gamma|_{[1,j]})}\Big\langle\prod_{i\in B\setminus\{j\}}A_i\Big\rangle\Big)\Big(\prod_{B\in K(\pi'_\Gamma|_{[j,k]\cup[k+1,k+\ell]})}\Big\langle\prod_{i\in B}A_i\Big\rangle\Big)\\ &\quad+q_{1,k}\langle A_k\rangle\Big(\prod_{B\in K(\pi_\Gamma|_{[1,j]})}\Big\langle\prod_{i\in B\setminus\{j\}}A_j\Big\rangle\Big)\\ &\quad\times\Big(\prod_{B\in K(\pi'_\Gamma|_{[j,k]\cup[k+1,k+\ell]})}\Big\langle\prod_{i\in B\setminus\{k\}}A_i\Big\rangle\Big)\end{aligned}\tag{3.79}$$

whenever π_Γ has more than one connecting block and (3.70) applies.

In the remaining cases, $\mathcal{F}(\pi_\Gamma)$ is evaluated using (3.71). Here, the first term can be treated similarly to (3.79), leaving only the contribution of the marked partition π''_Γ. Recalling that $[1,j\rangle$ and $[j,k]$ lie on the same circle and that $K(\pi''_\Gamma)$ is evaluated circle-wise, we can argue as in the permutation case. In particular,

$$\begin{aligned}\mathcal{F}(\pi_\Gamma) &= \sum_{\substack{\pi'_\Gamma\in\overrightarrow{NCP}(k,\ell)\\ blocks(\pi'_\Gamma)=\pi_\Gamma}}\Big(\prod_{B\in K(\pi'_\Gamma|_{[1,j]})}\Big\langle\prod_{i\in B\setminus\{j\}}A_i\Big\rangle\Big)\Big(\prod_{B\in K(\pi'_\Gamma|_{[j,k]\cup[k+1,k+\ell]})}\Big\langle\prod_{i\in B}A_i\Big\rangle\Big)\\ &\quad+\Big(\prod_{B\in K(\pi''_\Gamma|_{[1,j]})}\Big\langle\prod_{i\in B\setminus\{j\}}A_i\Big\rangle\Big)\Big(\prod_{B\in K(\pi''_\Gamma|_{[j,k]\cup[k+1,k+\ell]})}\Big\langle\prod_{i\in B}A_i\Big\rangle\Big)\\ &= \Big(\prod_{B\in K(\pi_\Gamma|_{[1,j]})}\Big\langle\prod_{i\in B\setminus\{j\}}A_i\Big\rangle\Big)\mathcal{F}(\pi_{\Gamma_2})\end{aligned}$$

i.e., $\mathcal{F}$ factorizes similar to (3.79). Here, Γ_2 denotes the subgraph of $\Gamma\in\mathcal{G}_2$ that lies in $\mathcal{G}([j,k],[k+1,k+\ell])$. Using (3.22) and (3.17), we evaluate

$$\begin{aligned}&\Big(\prod_{s=1}^{k+\ell}m_s\Big)\sum_{\Gamma\in NCG_{(1,j)}([1,j])\times\mathcal{G}([j,k],[k+1,k+\ell])}q_\Gamma\Big(\mathcal{F}(\pi_\Gamma)+q_{1,k}\mathcal{F}(\pi_{\Gamma\cup\{(1,k)\}})\Big)\\ &= m_1\mathfrak{m}_1[T_1,\dots,T_{j-1},G_j]\Big(F[T_j,\dots,T_k|T_{k+1},\dots,T_{k+\ell}]\\ &\qquad+q_{1,k}F[T_j,\dots,T_{k-1},G_k|T_{k+1},\dots,T_{k+\ell}]\langle A_k\rangle\Big).\end{aligned}\tag{3.80}$$

Computation of the third line of (3.74): The contribution from $\mathcal{G}_3$ can be treated similarly to the second line of (3.74).

Computation of the fourth line of (3.74): For the term that arises from $\mathcal{G}_4$, recall that τ only influences the geometry of the graph, but not its edge set. Hence, the summation reduces to the underlying disk non-crossing graphs and we can evaluate it using (3.22) and (3.17). As the only difference between $\mathcal{F}(\pi_\Gamma)$ and $\mathcal{F}(\pi_{\Gamma\cup\{(1,k)\}})$ is again the fact that $\langle A_k\rangle$ must occur as a separate factor in the second case, we obtain

$$\begin{aligned}&\Big(\prod_{s=1}^{k+\ell}m_s\Big)\sum_{\Gamma\in NCG_{(1,k+j)}(\{1,k,k+j,\dots,k+\ell,\dots,k+j\})}q_\Gamma\Big(\mathcal{F}(\pi_\Gamma)+q_{1,k}\mathcal{F}(\pi_{\Gamma\cup\{(1,k)\}})\Big)\\ &= m_1\Big(\mathfrak{m}_1[T_1,\dots,T_k,T_{k+j},\dots,T_{k+j-1},G_{k+j}]\\ &\qquad+q_{1,k}\mathfrak{m}_1[T_1,\dots,T_{k-1},G_k,T_{k+j},\dots,T_{k+j-1},G_{k+j}]\langle A_k\rangle\Big).\end{aligned}\tag{3.81}$$

Similar to (3.78), Γ containing an edge $(1, k+j)$ ensures that the contribution has the prefactor m_1.

Putting (3.77), (3.80), and (3.81) together, we see that (3.74) is equivalent to (3.36) with $\mathfrak{f}[\cdot|\cdot]$ in place of $\mathfrak{m}_2[\cdot|\cdot]$. We conclude that $\mathfrak{f}[\cdot|\cdot]$ satisfies the same symmetry and initial condition as $\mathfrak{m}_2[\cdot|\cdot]$, as well as the same recursion. Recall that these three properties uniquely identify $\mathfrak{f}[\alpha|\beta]$ and $\mathfrak{m}_2[\alpha|\beta]$ for any multi-indices α, β. It now readily follows by induction that $\mathfrak{f}[\alpha|\beta]$ and $\mathfrak{m}_2[\alpha|\beta]$ indeed coincide for all α, β, i.e., that $\mathfrak{f}[\cdot|\cdot] = \mathfrak{m}_2[\cdot|\cdot]$, as claimed. □

3.4 Proof of the Formulas for $\mathfrak{m}_\kappa[\cdot|\cdot]$, $\mathfrak{m}_\sigma[\cdot|\cdot]$, and $\mathfrak{m}_\omega[\cdot|\cdot]$

3.4.1 Proof of Theorem 3.2.6

We use proof by induction over the length of the multi-indices and the recursion for $\mathfrak{m}_\kappa[\cdot|\cdot]$ from (3.36). First, recall that by definition of $NCP(k)$, $NCP(\ell)$, and $\overrightarrow{NCP}(k,\ell)$ both sums on the right-hand side of (3.44) are empty whenever either α or β is empty. Observing that Definition 3.2.1(i) together with Theorem 3.2.4 implies that $\mathfrak{m}_\kappa[S_1|\emptyset] = \mathfrak{m}_\kappa[\emptyset|S_2] = 0$, the base case is established. As the formula on the right-hand side of (3.44) is further symmetric under the interchanging α and β (cf. remark below Theorem 3.2.4), it is sufficient to only carry out the induction step for one of the arguments in $\mathfrak{m}_\kappa[\cdot|\cdot]$.

Fix $k, \ell \in \mathbb{N}$ and assume that (3.44) holds for multi-indices α of length $1, \dots, k-1$ and β of length ℓ. Recalling that $\kappa_4\mathfrak{m}_\kappa[\cdot|\cdot]$ satisfies (3.36) with $\mathfrak{s}_\kappa$ in (3.38) as the only source term, we can use the recursion to express $\mathfrak{m}_\kappa[\alpha|\beta]$ in terms of $\mathfrak{m}_1[\cdot]$ and $\mathfrak{m}_\kappa[\alpha'|\beta]$ with multi-indices α' of length $1, \dots, k-1$. The claim thus follows by applying the induction hypothesis and showing that the expression obtained from the recursion can be rewritten to match the structure on the right-hand side of (3.44). We start by considering the summands on the right-hand side of (3.36) separately and then check that their sum, i.e., $\mathfrak{m}_\kappa[\cdot|\cdot]$, is of the same form. To facilitate keeping track of the individual contributions, we start with the terms on the right-hand side of (3.36) that do not contain the prefactor $q_{1,k}$ and abbreviate

$$\begin{aligned}
K_1 &:= m_1\mathfrak{m}_\kappa[T_2, \dots, T_{k-1}, G_kA_kA_1|T_{k+1}, \dots, T_{k+\ell}] \\
K_2^{(j)} &:= m_1\mathfrak{m}_\kappa[T_1, \dots, T_{j-1}, G_j|T_{k+1}, \dots, T_{k+\ell}]\mathfrak{m}_1[T_j, \dots, T_k], \quad j \in [k-1] \\
K_3^{(j)} &:= m_1\mathfrak{m}_1[T_1, \dots, T_{j-1}, G_j]\mathfrak{m}_\kappa[T_j, \dots, T_k|T_{k+1}, \dots, T_{k+\ell}], \quad j \in [2,k] \\
K_4^{(r,s,t)} &:= m_1\langle M_{[r]} \odot M_{[s,t]}\rangle\langle (M_{[r,k]}A_k) \odot M_{(t,\dots,k+\ell,k+1,\dots,s)}\rangle, \\
&\qquad\qquad r \in [k],\ k+1 \le s \le t \le k+\ell.
\end{aligned}$$

As the argument is similar, we fix j resp. r, s, t and omit the superscripts for the following discussion.

Structure of K_1: By the induction hypothesis, $\mathfrak{m}_\kappa[T_2, \dots, T_{k-1}, G_kA_kA_1|T_{k+1}, \dots, T_{k+\ell}]$ factorizes into expressions involving only deterministic matrices or spectral parameters, respectively, and further has the structure specified on the right-hand side of (3.44) in terms of $z_2, \dots, z_{k+\ell}$ and the matrices $A_2, \dots, A_{k-1}, A_kA_1, A_{k+1}, \dots, A_{k+\ell}$. As a consequence, the matrices A_k and A_1 always occur together in the matrix products. On the level of the indices of the deterministic matrices, we may reinterpret this as an element $\pi \in \overrightarrow{NCP}(k,\ell)$ with a cycle $(\dots k1\dots)$ or an element of $\pi \in NCP(k) \times NCP(\ell)$ with

a block $\{\dots, k, 1, \dots\}$ depending on the rest of the underlying structure. The treament of the two cases is identical and we consider them in parallel. As the indices 1 and k always occur together in $K(\pi)$, the index 1 must occur separated in $K^{-1}(\pi)$, either as a fixed point (1) or a singleton set $\{1\}$, to match the structure in (3.44). Note that the spectral parameter z_1 only appears in the prefactor m_1. Hence, setting the functions $\psi_{K^{-1}(\pi),(1)}$ resp. $\psi_{K^{-1}(\pi),\{1\}}$ equal to $m(z_1)$ yields the missing contribution. It follows that K_1 matches the structure on the right-hand side of (3.44). Note that all ψ_i associated with permutations without a fixed point (1) and partitions without a singleton block $\{1\}$, respectively, are equal to zero for K_1.

Structure of K_2: We apply the induction hypothesis for $\mathfrak{m}_\kappa[T_1, \dots, T_{j-1}, G_j | T_{k+1}, \dots, T_{k+\ell}]$ as well as (3.22) for $\mathfrak{m}_1[T_j, \dots, T_k]$ to rewrite K_2 as a sum of terms that naturally factorize into expressions involving only $z_1, \dots, z_{k+\ell}$ or $A_1, \dots, A_{k+\ell}$, respectively. It remains to check for the structure on the right-hand side of (3.44), i.e., that each summand can be associated with an annular non-crossing permutation π or a marked element $\pi \in NCP(k) \times NCP(\ell)$ such that the terms involving spectral parameters factorize according to cycles resp. blocks of π and the contribution of the deterministic matrices factorizes according to the cycles resp. blocks in the Kreweras complement $K(\pi)$. As treatment of the cases $\pi \in \overrightarrow{NCP}(k, \ell)$ and $\pi \in NCP(k) \times NCP(\ell)$ is identical, we consider them in parallel.

Observe that the induction hypothesis and (3.22) already prescribe the desired complement structure for the indices of the spectral parameters and deterministic matrices occurring in $m_1 \mathfrak{m}_\kappa[T_1, \dots, T_{j-1}, G_j | T_{k+1}, \dots, T_{k+\ell}]$ and $\mathfrak{m}_1[T_j, \dots, T_k]$, respectively. As we may visualize the elements of $NCP(j)$ with their Kreweras complement on an interval by cutting the boundary of the labeled disk (cf. Fig. 3.1), we can draw both factors of K_2 onto the same annulus. The result is visualized on the left of Fig. 3.18 below. For simplicity, we omit most of the intermediate labels and only add some of the matrices associated with the vertices on the midpoints of the arcs between the labels in red.

Note that the interval is placed such that the orientation inherited from the disk aligns with the orientation of the underlying annulus. Moreover, Fig. 3.18 matches the picture of a (k, ℓ)-annulus up to the label j occurring twice and the label on the midpoint of the arch connecting the two copies of j that is associated with the identity matrix. Since this identity matrix does not influence the value of K_2, we may remove the label corresponding to Id from the picture in Fig. 3.18. This leaves $k + \ell$ labels at the midpoints of arches along the annulus, one associated with each matrix $A_1, \dots, A_{k+\ell}$. On the level of the indices of the deterministic matrices, each term in in K_2 can thus be identified with an element $\pi \in \overrightarrow{NCP}(k, \ell)$ resp. an element of $\pi \in NCP(k) \times NCP(\ell)$. As the two labels j now occur next to each other, we merge them to obtain the (k, ℓ)-annulus as the structure underlying the indices of the spectral parameters. The result is visualized on the right of Fig. 3.18 below. Note that the labels now match Definition 3.1.19 exactly.

It is readily checked that the cycle resp. block structure obtained from merging the two labels j matches $K^{-1}(\pi)$ and that any cycles resp. blocks that were connected in this step can be interpreted as a cycle of an element in $\overrightarrow{NCP}(k, \ell)$ resp. a block of an element of $NCP(k) \times NCP(\ell)$. In particular, the contribution of the spectral parameters factorizes as claimed on the right-hand side of (3.44). Comparing the permutations resp. partitions contributing to each $K_2^{(j)}$ and noting that the right-hand side of (3.44) is linear in each ψ_i, the sum $\sum_j K_2^{(j)}$ also has the desired structure.

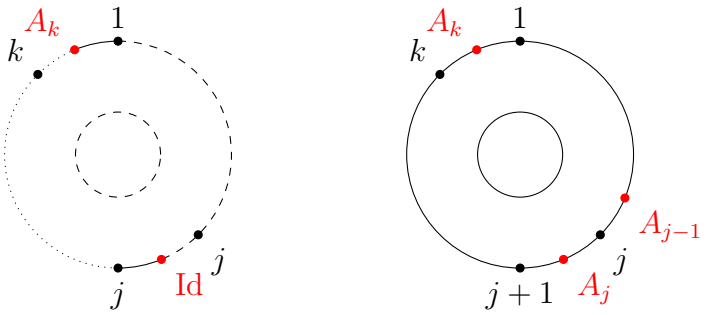

Fig. 3.18. Visualization of the indices in $m_1\mathfrak{m}_\kappa[T_1,\dots,T_{j-1},G_j|T_{k+1},\dots,T_{k+\ell}]$ (dashed) and in $\mathfrak{m}_1[T_j,\dots,T_k]$ (dotted) before the rewriting (left) and the indices after merging the two labels j (right).

Structure of K_3: The treatment of K_3 is analogous to that of K_2.

Structure of K_4: From the formula for $M_{(\dots)}$ in (3.18), it follows that K_4 can be written as a sum of terms that naturally factorize into expressions involving only $z_1,\dots,z_{k+\ell}$ or $A_1,\dots,A_{k+\ell}$, respectively. In particular, the part that involves deterministic matrices always consists of two factors of the form $\langle(\prod_{j\in I_1}A_j)\odot(\prod_{j\in I_2}A_j)\rangle$ with index sets $I_1\subseteq[k]$ and $I_2\subseteq[k+1,\dots,k+\ell]$ due to the Hadamard product in K_4. It remains to check for the structure on the right-hand side of (3.44).

Using the same trick as for K_2 and K_3, we can visualize the terms involved in K_4 on the same annulus. The result is sketched in Fig. 3.19 below. Again, we omit most of the labels and add the matrices associated with the vertices on the midpoints of the arcs in red. To avoid overcrowding the labels in the interior of the inner circle, we further visualize the two circles separately.

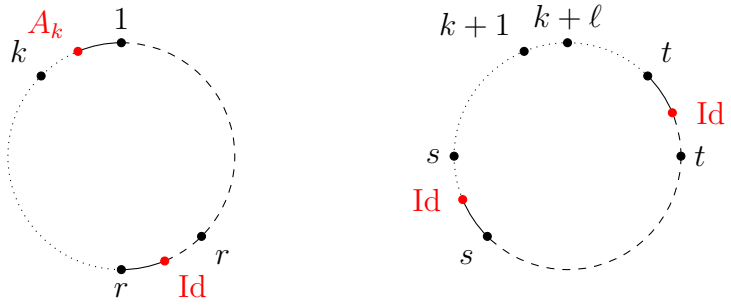

Fig. 3.19. Visualization of the indices in $m_1\langle M_{[r]}\odot M_{[s,t]}\rangle$ (dashed) and $\langle(M_{[r,k]}A_k)\odot M_{(t,\dots,k+\ell,k+1,\dots,s)}\rangle$ (dotted) on the outer (left) and inner circle (right) of the (k,ℓ)-annulus (pictured separately).

Observe that whenever a label occurs twice, both copies are placed next to one another and the label at the midpoint of the arch connecting them is always associated with the identity matrix. As the identity matrices do not contribute to K_4, we may argue as before and remove the corresponding labels from the picture in Fig. 3.19. Note that this leaves $k+\ell$ labels at the midpoints of arches along the annulus, one associated with each matrix $A_1,\dots,A_{k+\ell}$, such that we can reinterpret the structure on the level of the indices of the deterministic matrices as an element $\pi\in\overrightarrow{NCP}(k,\ell)$ resp. $\pi\in NCP(k)\times NCP(\ell)$. Since the two copies of r, s, and t are now placed right next to their counterpart, we merge

them to obtain a (k,ℓ)-annulus as the structure underlying the indices of the spectral parameters. It is readily checked that the cycle resp. block structure obtained from this step matches $K^{-1}(\pi)$ and that any cycles resp. blocks that were connected by merging two identical labels can be interpreted as a cycle of an element in $\overrightarrow{NCP}(k,\ell)$ resp. a block of an element of $NCP(k)\times NCP(\ell)$. In particular, whenever the merging of the labels creates two new cycles, the cycles can be drawn onto the (k,ℓ)-annulus without crossing. It follows that the contribution of the spectral parameters factorizes as claimed on the right-hand side of (3.44). Comparing the permutations resp. partitions contributing to each $K_4^{(r,s,t)}$ and noting that the right-hand side of (3.44) is linear in each ψ_i, summing over r,s,t preserves the structure of the term.

The remaining terms with prefactor $q_{1,k}\langle A_k\rangle$ can be rewritten similarly. As the factor $\langle A_k\rangle$ always occurs separately, the structure underlying the indices of the deterministic matrices is an element $\pi\in\overrightarrow{NCP}(k,\ell)$ resp. $\pi\in NCP(k)\times NCP(\ell)$ with a fixed point (k) resp. a singleton set $\{k\}$. Hence, K^{-1} must contain a cycle $(\dots k1\dots)$ resp. a block $\{\dots,k,1,\dots\}$ in which 1 and k occur together. Recalling that $q_{1,k}=\frac{m_1m_k}{1-m_1m_k}$, it is ensured that the part of the term depending on $z_1,\dots,z_k$ contains a factor $\psi_i(\dots,z_k,z_1,\dots)$. With this modification, we can use the same argument as for $K_1,\dots,K_4$ to conclude that all terms contributing to the recursion match the structure on the right-hand side of (3.44). Comparing the permutations resp. partitions contributing to each term and recalling that the right-hand side of (3.44) is linear in each ψ_i, it follows that the same holds for their sum, i.e., for $\mathfrak{m}_\kappa[\alpha|\beta]$. This concludes the induction step. □

3.4.2 Proof of Theorems 3.2.7 and 3.2.8

The proofs of Theorems 3.2.7 and 3.2.8 are similar to those of Theorems 3.2.4 and 3.2.6, respectively. We, therefore, mainly focus on the necessary modifications below.

Proof of Theorem 3.2.7. The overall argument is analogous to the proof of Theorem 3.2.4 in Sections 3.3.1 and 3.3.2 up to the definition of the graphs appearing in the combinatorial formula for $\widetilde{m}_\sigma[\cdot|\cdot]$. Recalling that the structure of the recursion for $\widetilde{m}_\sigma[\cdot|\cdot]$ is similar to the one for $\widetilde{m}_{GUE}[\cdot|\cdot]$, the multi-set of graphs can be constructed as described in Definition 3.3.2. However, since the source term involves $m^{\#,\sigma}[\cdot]$, the resulting multi-set of graphs will carry the same kind of vertex coloring. By replacing $\mathcal{G}_4$ in Definition 3.3.2 by

$$\mathcal{G}_4^\sigma := \bigcup_{j=1}^{\ell}\tau\Big(\Big\{\Gamma\in NCG^{\#}(\{1,\dots,k,k+j,k+j-1,\dots,k+1,k+\ell,k+j\})\Big|\\ \Gamma\text{ has edge }(1,k+j)\Big\}\Big).$$

where $\#=(0,\dots,0,1,\dots,1)$ with k zeros and $k+\ell+1$ ones, we obtain a family $\mathcal{G}^\sigma(k,\ell)$ of graphs with the desired properties. Recall that map τ was introduced in Definition 3.3.1. An example in the current setting is given in Fig. 3.20 below. Note the different arrangement of the indices on the left compared to Fig. 3.9, which results from the structure of the source term in the recursion for $\widetilde{m}_\sigma[\cdot|\cdot]$. Further, τ does not influence the coloring of the vertices.

Similar to the proof of Lemma 3.3.3, we obtain

$$\widetilde{m}_\sigma[1,\dots,k|k+1,\dots,k+\ell]=\Big(\prod_{s=1}^{k+\ell}m_s\Big)\sum_{\Gamma\in\mathcal{G}^\sigma(k,\ell)}\prod_{(i,j)\in E(\Gamma)}q_{i,j}^{\sharp}$$

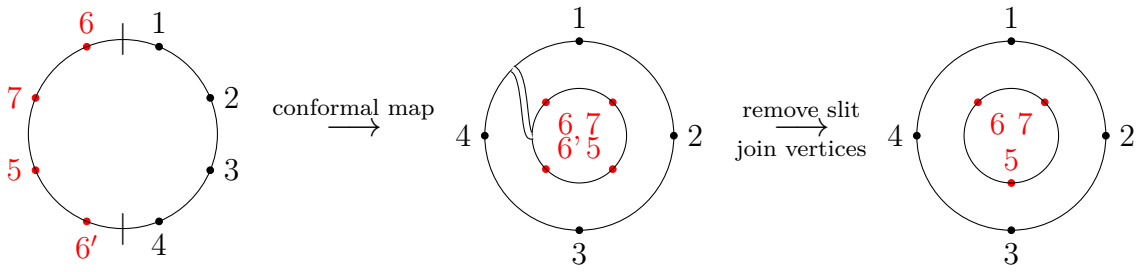

Fig. 3.20. The geometry of the transformation τ for $k = 4$, $\ell = 3$, and $j = 2$.

where $q^{\sharp}_{i,j}$ is as in Lemma 3.A.2, i.e., $q^{\sharp}_{i,j} = q_{i,j} = \frac{m_i m_j}{1-m_i m_j}$ whenever the edge (i, j) connects two vertices of the same color and $q^{\sharp}_{i,j} = \frac{\sigma m_i m_j}{1-\sigma m_i m_j}$ otherwise. From here, the remaining steps are carried out as in the proof of Theorem 3.2.4. □

Proof of Theorem 3.2.8. We use again proof by induction. As the recursions for $\mathfrak{m}_\omega[\cdot|\cdot]$ and $\mathfrak{m}_\kappa[\cdot|\cdot]$ are the same up to the source term, it only remains to show that

$$K_5^{(j)} := \langle (M_{[k]} A_k) \odot M_{(k+j,\dots,k+\ell,k+1,\dots k+j)} \rangle, \ j \in [\ell]$$

is of the form (3.46) and that summing up the contributions on the right-hand side of the recursion (3.36) does not break the structure. We start by noting that each $K_5^{(j)}$ contains the matrices $A_1, \dots, A_{k+\ell}$ exactly once and that the indices involved in the first and second factor of the Hadamard product match the indices on the outer and inner circle of the (k, ℓ)-annulus, respectively. The desired structure now follows from (3.22) and the fact that any annular non-crossing permutation of the (k, ℓ)-annulus can be decomposed uniquely into a composition of two permutations that only act on the inner and outer circle, respectively. Note that the index $k + j$ occurring twice in $M_{(k+j,\dots,k+\ell,k+1,\dots k+j)}$ does not influence the structure of the Kreweras complement. When visualizing the term on a labeled disk, the two vertices labeled $k + j$ lie next to each other and the vertex on the midpoint of the arch connecting the two copies is associated with the identity matrix, i.e., its effect is not visible when the matrix product in $M_{(k+j,\dots,k+\ell,k+1,\dots k+j)}$ is evaluated. □

3.A Proofs for the CLT in the Resolvent Case

3.A.1 Proof of Theorem 3.1.10 (Global Law with Transposes)

The proof of Theorem 3.1.10 is, modulo careful bookkeeping of the transposes, similar to the proof of the averaged local law in [39, Thm. 3.4]. In particular, an explicit formula for $m^{\#,\sigma}[\cdot]$ and the associated free cumulants is obtained along the way. For the convenience of the reader, we give a brief overview of the necessary changes. As there is nothing to prove if $\sigma = 1$ and the case $\sigma = -1$ needs an additional argument, consider $\sigma \in (-1, 1)$ first. We start by introducing a renormalization that captures the general second moment structure of W more precisely (cf. [36] and [40]). Let

$$\underline{Wf(W)} := Wf(W) - \widetilde{\mathbb{E}}\widetilde{W}(\partial_{\widetilde{W}} f)(W) \tag{3.82}$$

with $\widetilde{W}$ an independent copy of W. This yields, e.g.,

$$\underline{WG_1} = WG_1 + \langle G_1 \rangle G_1 + \frac{\sigma}{N} G_1^t G_1 + \frac{\widetilde{\omega_2}}{N} \mathrm{diag} G_1 G_1, \tag{3.83}$$

$$\underline{WT_1\dots T_k} = \underline{WG_1}A_1T_{[2,k]} + \sum_{j=2}^{k}\Big(\langle T_{[1,j\rangle}G_j\rangle T_{[j,k]} + \frac{\sigma}{N}(T_{[1,j\rangle}G_j)^tT_{[j,k]} + \frac{\widetilde{\omega_2}}{N}\mathrm{diag}T_{[1,j\rangle}G_jT_{[j,k]}\Big). \tag{3.84}$$

Note that a simpler renormalization can be obtained by choosing $\widetilde{W}$ to be an independent GUE matrix instead of an independent copy of W. For the formulas (3.83) and (3.84) above, the difference between the two renormalizations is negligible. However, it becomes significant if the matrix product involves at least one transpose of a resolvent (cf. [36, Rem. 4.3]). As an example, consider the normalized trace of

$$\underline{WG_1^t} = WG_1^t + \frac{1}{N}G_1G_1^t + \sigma\langle G_1\rangle G_1^t + \frac{\widetilde{\omega_2}}{N}\mathrm{diag}G_1G_1^t$$

where $\sigma\langle G_1\rangle^2 \sim 1$.

To obtain (3.23), we rewrite $G_1^\sharp A_1\dots G_k^\sharp A_k$ in terms of $\underline{WG_1^\sharp A_1\dots G_k^\sharp A_k}$ and estimate the resulting terms. We start by considering the case $A_1 = \dots = A_k = \mathrm{Id}$. For $k = 2$, we obtain, e.g.,

$$\underline{WG_1G_2^t} = \underline{WG_1}G_2^t + \frac{1}{N}(G_1G_2^t)^tG_2^t + \sigma\langle G_1G_2^t\rangle G_2^t + \frac{\widetilde{\omega_2}}{N}\mathrm{diag}G_1G_2^tG_2^t$$

which yields

$$\Big(1-\sigma m_1m_2+\mathcal{O}_\prec\Big(\frac{1}{N}\Big)\Big)\langle G_1G_2^t\rangle = m_1m_2 - m_1\underline{WG_1G_2^t} + \frac{m_1}{N}\Big(\sigma\langle G_1^tG_1G_2^t\rangle + \widetilde{\omega_2}\langle\mathrm{diag}G_1G_1G_2^t + \mathrm{diag}G_1G_2^tG_2^t\rangle\Big) + \mathcal{O}_\prec\Big(\frac{1}{N}\Big)$$

by (3.83), the resolvent identity $WG - zG = \mathrm{Id}$, (3.11), and the local law for $\langle G_j\rangle$. Noting that $|\langle\underline{WG_1^tG_2}\rangle| = \mathcal{O}_\prec(N^{-1})$ following the bounds in [36, Thm. 4.1] and that the term with prefactor N^{-1} is also of lower order (cf. [40, Eq. (69)–(70)]), we obtain (3.23) with

$$\mathfrak{m}_1[G_1,G_2^t] = m^{(0,1),\sigma}[1,2] = \frac{m_1m_2}{1-\sigma m_1m_2}.$$

Note that the only effect of the inclusion of transposes lies in the emergence of the stability factor $1-\sigma m_1m_2$. However, as $|1-\sigma m_1m_2| > 1-|\sigma|$ for $|\sigma|<1$, the term is bounded from below and does not change the outcome of the estimates. The analog for general $k \geq 2$ follows by induction over the number of resolvents. Similar to [39, Thm. 3.4], the above argument allows us to extract a recursion that is satisfied by the deterministic approximation of $\langle G_1^\sharp\dots G_k^\sharp\rangle$ up to an $\mathcal{O}_\prec(N^{-1})$ error. By defining $m^{\#,\sigma}[\cdot]$ to satisfy this recursion exactly, i.e., defining it by (3.24), yields the desired statement. The recursion for $\mathfrak{m}_1[G_1^\sharp A_1,\dots,G_k^\sharp A_k]$ is obtained analogously.

The explicit formula (3.23) now follows by solving this recursion. As the treatment of the deterministic matrices is completely analogous to [39, Thm. 3.4], we restrict the discussion to the case $A_1 = \dots = A_k = \mathrm{Id}$, i.e., to (3.24). Here, we obtain that $m^{\#,\sigma}[\cdot]$ and the associated free cumulants have a representation in terms of non-crossing graphs that mirror (3.16) and (3.17).

Definition 3.A.1 (Bicolored NCG)**.** *Let $k \in \mathbb{N}$ and denote by $\#$ a binary vector of length k. For every $\Gamma \in NCG(k)$, color the vertex $j \in \{1,\dots,k\}$ red if the j-th entry of $\#$ is 1, otherwise color it black. We call the resulting set of graphs* bicolored (disk) non-crossing graphs *on $\{1,\dots,k\}$ and denote it by $NCG^\#(k)$.*

Lemma 3.A.2. *Let $k \in \mathbb{N}$ and fix a binary vector $\#$ of length k. Then*

$$m^{\#,\sigma}[1,\dots,k] = \Big(\prod_{s=1}^{k} m_s\Big) \sum_{\Gamma \in NCG^{\#}(k)} q_\Gamma^{\#} \tag{3.85}$$

$$m_\circ^{\#,\sigma}[1,\dots,k] = \Big(\prod_{s=1}^{k} m_s\Big) \sum_{\Gamma \in NCG_c^{\#}(k)} q_\Gamma^{\#} \tag{3.86}$$

where $NCG_c^{\#}(k)$ denotes the connected graphs in $NCG^{\#}(k)$ and $q_\Gamma := \prod_{(i,j)\in\mathbb{E}(\Gamma)} q_{i,j}^{\sharp}$ with $E(\Gamma)$ denoting the edge set of Γ. The edge weights $q_{i,j}^{\sharp}$ are such that $q_{i,j}^{\sharp} := q_{i,j} = \frac{m_i m_j}{1-m_i m_j}$ whenever the edge (i,j) connects two vertices of the same color and $q_{i,j}^{\sharp} := \frac{\sigma m_i m_j}{1-\sigma m_i m_j}$ whenever (i,j) connects a red vertex to a black one.

The proof is analogous to [39, Lem. 5.2] using the recursive structure of $NCG(k)$. Note that the modified edge weights as well as the factors $q_{1,k}^{\sharp}$ and $c_{1,j}$ in the recursion account for the stability factor $1-\sigma m_i m_j$ arising whenever the product $G_1^{\sharp}\dots G_k^{\sharp}$ involves both resolvents and their transposes. We illustrate Lemma 3.A.2 with an example.

Example 3.A.3. Let $k=3$ and pick spectral parameters $z_1, z_2, z_3 \in \mathbb{C}$ with $|\Im z_j| \gtrsim 1$ as well as $A_1 = A_2 = A_3 = \mathrm{Id}$. By Theorem 3.1.10, the deterministic approximation of $\langle G_1 G_2^t G_3^t\rangle$ is given by $m^{\#,\sigma}[1,2,3]$ with $\# = (0,1,1)$. We visualize the elements of the set $NCG^{\#}(3)$ in Fig. 3.21 below. For a better overview, the edges are drawn as solid or dashed according to their contribution to $q_\Gamma^{\#}$. We use this sketch to compute

$$m^{\#,\sigma}[1,2,3] = m_1 m_2 m_3 \sum_{\Gamma\in NCG^{\#}} = \frac{m_1 m_2 m_3}{(1-\sigma m_1 m_2)(1-\sigma m_1 m_3)(1-m_2 m_3)}.$$

Note that we thus reobtain [40, Eq. (49)] on macroscopic scales if $z_2 = z_3$.

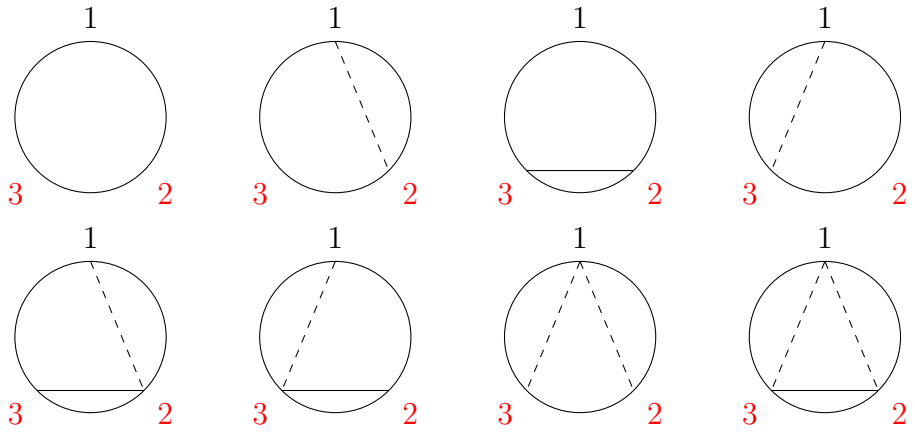

Fig. 3.21. The elements of $NCG^{\#}(3)$ for $\# = (0,1,1)$. The solid edges contribute a factor of $\frac{m_i m_j}{1-m_i m_j}$ to $q_\Gamma^{\#}$ while dashed edges contribute a factor $\frac{\sigma m_i m_j}{1-\sigma m_i m_j}$.

It remains to consider the case $\sigma = -1$. Here, the Wigner matrix is of the form $W = D + \mathrm{i}S$ with a diagonal matrix D and a skew-symmetric matrix S. Note that whenever the diagonal part is equal to zero, the resolvent $R(z_j) = (\mathrm{i}S - z_j)^{-1}$ satisfies $R(z_j)^t = -R(-z_j)$. This allows treating the proof of (3.23) for $\langle R(z_1)^{\sharp}\dots R(z_k)^{\sharp}\rangle$ with $R(z_j)^{\sharp} \in \{R(z_j), R(z_j)^t\}$

analogous to the case $\langle G_1 \dots G_k \rangle$. Recall that this requires in particular a local law for $\langle R(z_j) \rangle$, which holds even if the Wigner matrix has zero diagonal. The general case follows from the bound

$$\langle G_1^\sharp \dots G_k^\sharp \rangle = \langle R(z_1)^\sharp \dots R(z_k)^\sharp \rangle + \mathcal{O}_\prec\Big(\frac{1}{N}\Big) \tag{3.87}$$

which reduces the proof of (3.23) to the previously considered case $D = 0$. The proof of (3.87) is given for $\langle G_1 G_2^t \rangle$ in [40, App. B]. By expanding $G_3^\sharp, \dots, G_k^\sharp$ analogously, the argument readily extends to $k \geq 3$ resolvents. □

3.A.2 Proof of Therorem 3.2.3 (CLT for Resolvents)

We follow the general outline of the proof of [122, Thm. 3.6] to identify the necessary modifications for the general model described by Assumption 3.1.1. As the formulas in the general case are derived analogously, we assume w.l.o.g. $\langle A_k \rangle = 0$ throughout to keep the following equations short. The first step is the analog of [122, Lem. 2.3] that characterizes the difference between $\langle T_{[1,k]} \rangle - \mathfrak{m}_1[T_1, \dots, T_k]$ and $\langle T_{[1,k]} \rangle - \mathbb{E}\langle T_{[1,k]} \rangle$, thus connecting the statistics X_α back to the local law (3.19).

Lemma 3.A.4. *Let $k \in \mathbb{N}$ and fix spectral parameters $z_1, \dots, z_n$ with $|\Im z_j| \gtrsim 1$ and $\max_j |z_j| \leq N^{100}$ as well as bounded deterministic matrices $A_1, \dots, A_k$ such that $\langle A_k \rangle = 0$. Then,*

$$\mathbb{E}\langle T_1 \dots T_k \rangle = \mathfrak{m}_1[T_1, \dots, T_k] + \frac{1}{N}\mathcal{E}[T_1, \dots, T_k] + \mathcal{O}\Big(\frac{N^\varepsilon}{N^{3/2}}\Big) \tag{3.88}$$

with $\mathfrak{m}_1[\cdot]$ as in (3.21) and a set function $\mathcal{E}[\cdot]$ that satisfies $\mathcal{E}[\emptyset] = 0$ as well as the recursion

$$\begin{aligned}
\mathcal{E}[T_1, \dots, T_k] = m_1\Big(&\mathcal{E}[T_2, \dots, T_{j-1}, G_k A_k A_1] + \sum_{j=1}^{k-1} \mathcal{E}[T_1, \dots, T_{j-1}, G_j]\mathfrak{m}_1[T_j, \dots, T_k] \\
&+ \sum_{j=2}^{k} \mathfrak{m}_1[T_1, \dots, T_{j-1}, G_j]\mathcal{E}[T_j, \dots, T_k] + \frac{\widetilde{\omega_2}}{N}\sum_{j=1}^{k}\langle M_{[j]} \odot (M_{[j,k]}A_k)\rangle \\
&+ \frac{\sigma}{N}\sum_{j=1}^{k} \mathfrak{m}_1[G_j^t A_{j-1}^t, \dots, G_2^t A_1^t, G_1^t, T_j, \dots, T_k] \\
&+ \kappa_4 \sum_{1 \leq r \leq s \leq t \leq k} \langle M_{[r]} \odot M_{[s,t]}\rangle\langle M_{[r,s]} \odot (M_{[t,k]}A_k)\rangle\Big),
\end{aligned}$$

where $\odot$ denotes the Hadamard product and $\mathfrak{m}_2[\cdot|\cdot]$ is interpreted as in (3.25).

Proof. We note the following modifications to the proof of [122, Lem. 2.3]. Using the renormalization in (3.82) with $\widetilde{W}$ being an independent copy of W (compared to the independent GUE matrix $\widetilde{W}$ used in [122]), we obtain the relation

$$\begin{aligned}
&\langle T_{[1,k]} \rangle - \mathfrak{m}_1[T_1, \dots, T_k] \\
&= m_1\Big(- \underline{\langle W T_{[1,k]} \rangle} + (\langle T_{[2,k)} G_k A_k A_1 \rangle - \mathfrak{m}_1[T_2, \dots, T_{k-1}, G_k A_k A_1]) \\
&\quad + \sum_{j=1}^{k-1} (\langle T_{[1,j\rangle} G_j \rangle - \mathfrak{m}_1[T_1, \dots, T_{j-1}, G_j])\mathfrak{m}_1[T_j, \dots, T_k]
\end{aligned}$$

$$+\sum_{j=2}^{k}\mathfrak{m}_1[T_1,\dots,T_{j-1},G_j](\langle T_{[j,k]}\rangle-\mathfrak{m}_1[T_j,\dots,T_k]) \tag{3.89}$$
$$+\frac{\sigma}{N}\sum_{j=1}^{k}\langle(T_{[1,j\rangle}G_j)^tT_{[j,k]}\rangle+\frac{\widetilde{\omega_2}}{N}\sum_{j=1}^{k}\langle\mathrm{diag}T_{[1,j\rangle}G_jT_{[j,k]}\rangle\Big)+\mathcal{O}_\prec\Big(\frac{1}{N^2}\Big).$$

Recall that the deterministic approximation $\mathfrak{m}_1[\cdot]$ is independent of the value of σ and $\widetilde{\omega}_2$ (cf. [39, Thm. 3.4]). The terms in the last line of (3.89) involve an additional factor of N^{-1}. Applying Theorem 3.1.10 as well as the isotropic local law (3.20) thus yields

$$\langle(T_{[1,j\rangle}G_j)^tT_{[j,k]}\rangle=\mathfrak{m}_1[G_j^tA_{j-1}^t,\dots,G_2^tA_1^t,G_1^t,T_j,\dots,T_k]+\mathcal{O}_\prec\Big(\frac{1}{N}\Big)$$
$$\langle\mathrm{diag}T_{[1,j\rangle}G_jT_{[j,k]}\rangle=\frac{1}{N}\sum_{r=1}^{N}(T_{[1,j\rangle}G_j)_{rr}(T_{[j,k]})_{rr}=\frac{1}{N}\sum_{r=1}^{N}(M_{[j]})_{rr}(M_{[j,k]}A_k)_{rr}+\mathcal{O}_\prec\Big(\frac{1}{\sqrt{N}}\Big)$$

and we can replace the respective terms in (3.89) by their deterministic approximation. Note that this also changes the error term to $\mathcal{O}_\prec(N^{-3/2})$. Computing the expectation of (3.89) is now analogous to [122, Lem. 2.3]. In particular, the cumulant expansion

$$N\mathbb{E}\langle\underline{WT_{[1,k]}}\rangle=\sum_{n\geq2}\sum_{x,y\in[N]}\sum_{\nu\in\{xy,yx\}^n}\frac{\kappa(xy,\nu)}{n!}\mathbb{E}\partial_\nu(T_{[1,k]})_{yx}$$

yields the same result, since the term does not involve any second moments of the entries of W. □

Next, we identify the limiting covariance of two modes X_α and X_β. This yields an analog of [122, Lem. 2.5] for the model in Assumption 3.1.1 on macroscopic scales and thus constitutes the base case for the induction argument in the proof of Theorem 3.2.3.

Lemma 3.A.5. *Fix $k,\ell\in\mathbb{N}$ and let α,β be two multi-indices of length k and ℓ, respectively. Assume that the Wigner matrix W satisfies Assumption 3.1.1 and pick spectral parameters $z_1,\dots,z_{k+\ell}$ with $|\Im z_j|\gtrsim 1$ and $\max_j|z_j|\leq N^{100}$ as well as deterministic matrices $A_1,\dots,A_{k+\ell}$ with $\|A_j\|\lesssim 1$. Then, for any $\varepsilon>0$,*

$$N^2\mathbb{E}X_\alpha^{(k,a)}X_\beta^{(\ell,b)}=\mathfrak{m}_2[\alpha|\beta]+\mathcal{O}\Big(\frac{N^\varepsilon}{\sqrt{N}}\Big),$$

with $\mathfrak{m}_2[\cdot|\cdot]$ as introduced in Definition 3.2.1.

Proof of Lemma 3.A.5. The proof is analogous to that of [122, Lem. 2.5] with one key difference. When evaluating the cumulant expansion

$$N^2\mathbb{E}\Big(\langle\underline{WT_{[1,k]}}\rangle X_\beta^{(\ell)}\Big)=N\sum_{n\geq1}\sum_{x,y\in[N]}\sum_{\nu\in\{xy,yx\}^n}\frac{\kappa(xy,\nu)}{n!}\mathbb{E}\partial_\nu\Big((T_{[1,k]})_{yx}X_\beta^{(\ell)}\Big),$$

the second moment structure of the model in Assumption 3.1.1 has to be taken into account for evaluating the $n=1$ term. We obtain

$$N\sum_{x,y\in[N]}\sum_{\nu\in\{xy,yx\}}\kappa(xy,\nu)\mathbb{E}\partial_\nu\Big((T_{[1,k]})_{yx}X_\beta^{(\ell)}\Big)$$
$$=-\sum_{j=1}^{\ell}\mathbb{E}\Big(\langle T_{[1,k]}T_{[k+j,k+\ell]}T_{[1,k+j\rangle}G_{k+j}\rangle+\sigma\langle T_{[1,k]}(T_{[k+j,k+\ell]}T_{[1,k+j\rangle}G_{k+j})^t\rangle$$
$$+\widetilde{\omega_2}\langle\mathrm{diag}T_{[1,k]}\mathrm{diag}T_{[k+j,k+\ell]}T_{[1,k+j\rangle}G_{k+j}\rangle\Big),$$

which, compared to the computation in the proof of [122, Lem. 2.5], additionally involves σ and $\widetilde{\omega_2}$. The deterministic approximation of the terms follow from (3.19), Theorem 3.1.10 and (3.20), respectively, which yields the source terms $\mathfrak{s}_{GUE}$, $\mathfrak{s}_\sigma$, and $\mathfrak{s}_\omega$ in (3.36). The rest of the expansion evaluates exactly as its counterpart in [122] since the term does not involve any second moments of the entries of W. □

The rest of the proof of Theorem 3.2.3 is analogous to that of [122, Thm. 3.6], i.e., we apply induction on the number of factors in (3.42) using $\mathbb{E}X_\alpha = 0$ and Lemma 3.A.5 as base cases. □

3.B Additional Proofs and Computations

3.B.1 Proof of Lemma 3.1.17

Using (3.16) for the ordered multi-set $S = \{z_1, \dots, z_j, \dots, z_k, z_j\}$, rewrite the left-hand side of (3.32) as

$$m[1, \dots, j, \dots, k, j] = m_1 \dots m_{j-1} m_j^2 m_{j+1} \dots m_k \sum_{\Gamma \in NCG[\{1,\dots,k,j\}]} \prod_{(a,b)\in E(\Gamma)} q_{a,b}. \tag{3.90}$$

Recalling that $m[S]$ is invariant under any permutation of the elements of S, we pick an ordering in which the two j's occur in two consecutive positions and visualize the corresponding non-crossing graphs by equidistantly arranging the vertices on a circle. In this picture, the edge e connecting both j's cannot be involved in any crossing, even in an arbitrary planar graph on the given vertices (see left of Fig. 3.22). Hence, for any non-crossing graph with edges $\{e_1, \dots, e_n\} \not\ni e$, the graph with edge set $\{e_1, \dots, e_n, e\}$ is also non-crossing. In particular, every non-crossing graph that involves e has a counterpart that does not. Next, note that there is at most one vertex l among $1, \dots, j-1, j+1, \dots, k$ that is connected to both copies of j, as having two distinct vertices l, l' with this property results in a crossing (see right of Fig. 3.22).

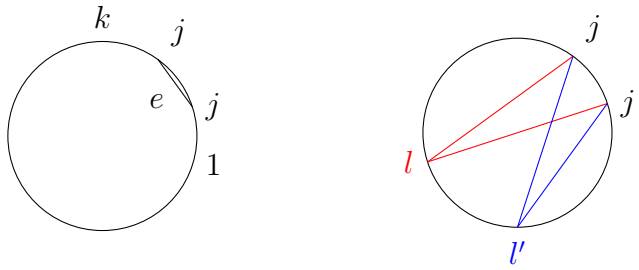

Fig. 3.22. An edge between two consecutive vertices never results in a crossing (left), but connecting two distinct vertices to both copies of j always does (right).

Consider the disjoint decomposition

$$NCG(\{1, \dots, j, \dots, k, j\}) = \mathcal{S}_1 \cup \Big(\bigcup_{l \in [k]\setminus\{j\}} \mathcal{S}_2^{(l)} \Big)$$

where $\mathcal{S}_2^{(l)}$ contains all $\Gamma \in NCG(\{1, \dots, j, \dots, k, j\})$ for which the vertex $l \in [k] \setminus \{j\}$ is connected to both copies of j and $\mathcal{S}_1$ contains any remaining Γ. This implies

$$\sum_{\Gamma \in NCG[\{1,\dots,k,j\}]} \prod_{(a,b)\in E(\Gamma)} q_{a,b} = \sum_{\Gamma \in \mathcal{S}_1} \prod_{(a,b)\in E(\Gamma)} q_{a,b} + \sum_{l \in [k]\setminus\{j\}} \sum_{\Gamma \in \mathcal{S}_2^{(l)}} \prod_{(a,b)\in E(\Gamma)} q_{a,b}. \tag{3.91}$$

Next, merge both vertices with the label j. The resulting graphs are either an element of $NCG(k)$ or arise from an element of $NCG(k)$ by adding a loop (j,j). Recalling that

$$q_{j,j} = \frac{m_j^2}{1-m_j^2} = m_j' \tag{3.92}$$

where the second equality follows from (3.12), we can factor out $(1+m_j')$ on the right-hand side of (3.91) and reduce to summation over $NCG(k)$ without further restriction. Note that this also results in the edge (l,j) doubling whenever $\Gamma \in \mathcal{S}_2^{(l)}$, i.e., we obtain an extra factor of $q_{j,l}$ in this case. Hence,

$$\sum_{\Gamma\in NCG[\{1,\dots,k,j\}]}\prod_{(a,b)\in E(\Gamma)} q_{a,b} = (1+m_j')\Big(\sum_{\Gamma\in NCG[\{1,\dots,k\}]}\prod_{(a,b)\in E(\Gamma)} q_{a,b}\Big)\Big(1+\sum_{l\in[k]\setminus\{j\}} q_{j,l}\Big).$$

Noting that (3.92) is equivalent to $m_j(1+m_j') = \frac{m_j'}{m_j}$ and applying (3.16) to recover $m[1,\dots,k]$ yields (3.32). □

3.B.2 Proof of Corollary 3.2.11

We start by evaluating $\mathrm{sc}[i_1,\dots,i_n]$ and $\mathrm{sc}[i_1,\dots,i_n|i_{n+1},\dots,i_{n+m}]$ before considering the associated free cumulant functions. First, note that

$$\mathrm{sc}[i_1,\dots,i_n] = \int_{-2}^{2} x^n \rho_{sc}(x)\mathrm{d}x = \begin{cases} 0, & \text{if } n \text{ odd},\\ C_{n/2}, & \text{if } n \text{ even},\end{cases}$$

where $C_0, C_1, C_2, \dots$ denote the Catalan numbers. In particular, $\mathrm{sc}[i_1,\dots,i_n]$ coincides with the number of non-crossing pairings of the set $[n]$. It readily follows that $\mathrm{sc}_\circ[i_1,i_2] = 1$ and that $\mathrm{sc}_\circ[i_1,\dots,i_n] = 0$ whenever n is odd. Moreover, $\mathrm{sc}_\circ[i_1,\dots,i_n] = 0$ for any even $n \geq 4$. The latter follows inductively from (3.14) by writing

$$\mathrm{sc}_\circ[i_1,\dots,i_n] = \mathrm{sc}[i_1,\dots,i_n] - \sum_{\pi\in NCP(n)\setminus\{[n]\}}\prod_{B\in\pi}\mathrm{sc}_\circ[B], \tag{3.93}$$

and observing that only pairings contribute to the sum on the right-hand side of (3.93), i.e., the two terms cancel.

Next, we compute $\mathrm{sc}[i_1,\dots,i_n|i_{n+1},\dots,i_{n+m}]$. Observe that by (3.26) and Theorem 3.2.9,

$$\mathrm{sc}[i_1,\dots,i_n|i_{n+1},\dots,i_{n+m}] = \lim_{N\to\infty}\mathbb{E}\Big[(\operatorname{Tr}W^n - \mathbb{E}W^n)(\operatorname{Tr}W^m - \mathbb{E}W^m)\Big].$$

The limit on the right-hand side is well-known in the free probability literature (see, e.g., [107]) and hence readily identified as the number of non-crossing pairings of the (n,m)-annulus. Solving (3.26) for $\mathrm{sc}_{\circ\circ}[i_1,\dots,i_n|i_{n+1},\dots,i_{n+m}]$, it follows inductively that $\mathrm{sc}_{\circ\circ}[i_1,\dots,i_n|i_{n+1},\dots,i_{n+m}] = 0$ for any n,m. Hence, $\Phi_{\pi_1\times\pi_2,U_1\times U_2}(f_1,\dots,f_{k+\ell}) = 0$ and

$$\Phi_\pi(f_1,\dots,f_{k+\ell}) = \begin{cases} 1, & \text{if } \pi \in \overrightarrow{NCP}_2(k,\ell),\\ 0, & \text{otherwise},\end{cases}$$

which is the claim. Note that the error in (3.48) evaluates to $\mathcal{O}(N^{\varepsilon-1/2})$ as $\|f_i\|_{H^p}$ for $i=1,\dots,k$ resp. $\|f_j\|_{H^q}$ for $j=k+1,\dots,k+\ell$ are N-independent constants in the macroscopic regime. □

4 Prethermalization for Deformed Wigner Matrices

We prove that a class of weakly perturbed Hamiltonians of the form $H_\lambda = H_0 + \lambda W$, with W being a Wigner matrix, exhibits prethermalization*. That is, the time evolution generated by H_λ relaxes to its ultimate thermal state via an intermediate prethermal state with a lifetime of order λ^{-2}. Moreover, we obtain a general relaxation formula, expressing the perturbed dynamics via the unperturbed dynamics and the ultimate thermal state. The proof relies on a two-resolvent law for the deformed Wigner matrix H_λ.*

Published as L. Erdős, J. Henheik, J. Reker, and V. Riabov. Prethermalization for deformed Wigner matrices. *Preprint, arXiv:2310.06677*, 2023.

4.1 Introduction

It is well-known (see, e.g., [84]) that certain macroscopic observables in an isolated quantum system with many interacting degrees of freedom tend to equilibrate, i.e., their expectation values become essentially constant at large times. However, if the system is coupled to the environment (reservoir), then the process of relaxation to equilibrium may take different forms depending on the properties of the initial system and the structure of the perturbation.

In this work, we consider a weakly coupled system of the form

$$H_\lambda := H_0 + \lambda W, \tag{4.1}$$

where H_0 is a single-body or a many-body *Hamiltonian*, W is an energy preserving (Hermitian) *perturbation*, and λ is a small coupling constant. For our phenomenological study, we consider a mean-field random perturbation that couples all modes. Following the extensive physics literature [140, 51, 116, 45, 47, 48], we choose the perturbation W to be a *Wigner random matrix*, i.e., a random matrix with centered, independent identically distributed (i.i.d.) entries (modulo the Hermitian symmetry).
The central object of our study is the perturbed time evolution of the quantum expectation value

$$\langle A \rangle_{P_\lambda(t)} := \mathrm{Tr}[P_\lambda(t)A] \tag{4.2}$$

of an *observable* A, compared to the unperturbed evolution $\langle A \rangle_{P_0(t)} = \mathrm{Tr}[P_0(t)A]$, which is considered known. Here

$$P_\lambda(t) := \mathrm{e}^{-\mathrm{i}tH_\lambda} P \mathrm{e}^{\mathrm{i}tH_\lambda} \qquad \text{resp.} \qquad P_0(t) := \mathrm{e}^{-\mathrm{i}tH_0} P \mathrm{e}^{\mathrm{i}tH_0} \tag{4.3}$$

denote the Heisenberg time evolution of an initial *state* P governed by the (un)perturbed Hamiltonian. We point out that the *unperturbed* evolution strongly depends on all its constituents and hence, it might exhibit qualitatively different and generally complex behavior.

In contrast, the *perturbed* system relaxes to equilibrium via a robust mechanism, and it can be described by a fairly simple general *relaxation formula*

$$\langle A\rangle_{P_\lambda(t)} \approx \langle A\rangle_{\widetilde{P}_\lambda} + |g_\lambda(t)|^2\big[\langle A\rangle_{P_0(t)} - \langle A\rangle_{\widetilde{P}_\lambda}\big], \qquad g_\lambda(t) := \mathrm{e}^{-\alpha\lambda^2 t}, \quad \alpha > 0, \tag{4.4}$$

where $\widetilde{P}_\lambda$ is the thermal state of the composite system (4.1). In this form, Eq. (4.4) is first mentioned in [116, Eq. (40)], where it describes the time dependence of the expectation of an observable in a nonintegrable system after perturbation by a random matrix.

The relaxation formula (4.4) shows convergence to the thermal state at an exponential rate on time scales of order λ^{-2} (in agreement with *Fermi's golden rule*), but it also carries more refined information about the role of the unperturbed dynamics in the process. A particularly interesting case occurs if both the perturbed and unperturbed systems equilibrate but do not approach the same limiting value. This often happens if H_0 has an additional symmetry (conserved quantity) that is broken by the perturbation. If the time scale λ^{-2} of the perturbed equilibration is smaller than that of the unperturbed one, then the former robust process eclipses the latter. In particular, the precise form of $\langle A\rangle_{P_0(t)}$ in (4.4) is irrelevant whenever the prefactor $|g_\lambda(t)|^2$ is already exponentially small. In the opposite case, however, the equilibration of the perturbed dynamics happens in two stages. This phenomenon, known as *prethermalization* in the physics literature, was first described in a paper by Moeckel and Kehrein [114]. We remark, however, that this terminology was already used to describe a different phenomenon a few years earlier [23].

Nowadays, prethermalization has been extensively studied both experimentally (see, e.g., the review [100]) and theoretically (e.g., in [25, 54, 83, 97, 120, 138, 108], see also the review [115]). Reimann and Dabelow [45] studied the first relaxation stage of a prethermalization process, which is governed by H_0. More precisely, assuming that $P_0(t)$ relaxes to a non-thermal steady state, they find that the perturbed time evolution $\langle A\rangle_{P_\lambda(t)}$ with a sufficiently weak perturbation ($\lambda \ll 1$) closely follows the unperturbed time evolution $\langle A\rangle_{P_0(t)}$ for times $t \ll \lambda^{-2}$. In particular, the perturbed time evolution $\langle A\rangle_{P_\lambda(t)}$ is close to the non-thermal steady state of H_0 for times $1 \ll t \ll \lambda^{-2}$.[1] The authors of [45] further extended their principal approach to a general study of relaxation theory for perturbed quantum dynamics in [47, 48]. These works now include all times and also the *strong* coupling regime (in case of banded matrices), which yields a characteristic power-law time decay (given more precisely by a Bessel function) instead of the exponential decay in (4.4). The theoretical model is then applied to several examples and compare the prediction to numerical and experimental works (see also Dabelow's PhD thesis [44] for further details). Finally, we also mention that prethermalization in the form of the existence of an effectively conserved quantity for very long times has been rigorously established in [1] for periodically driven quantum systems if the frequency is large compared with the size of the driving potential.

In this paper, we approach prethermalization from the viewpoint of random matrix theory, interpreting the unperturbed Hamiltonian $H_0 \equiv H_0(N) \in \mathbb{C}^{N\times N}$ as a fixed sequence of bounded self-adjoint deterministic matrices and the perturbation $W \equiv W(N)$ as an $N \times N$ Wigner random matrix. Our Hamiltonian H_λ in the setting of (4.1) is also called *deformed Wigner matrix* in random matrix theory, or it can be viewed as a Wigner random matrix

[1] Unperturbed systems H_0, for which the time evolution $\langle A\rangle_{P_0(t)}$ does not approach the microcanonical prediction of equilibrium statistical mechanics, but nevertheless have a large-t limit (a non-thermal steady state), are studied in [15]. In a sense, these systems exhibit prethermalization, although they do not approach thermal equilibrium after the first steady state is reached.

with nonzero expectation. Wigner matrices are encountered in many related physics models, e.g., the recent rigorous study of thermalization problems [36, 39, 37]. Here, the key technical result is a strong concentration property of the resolvent $G(z) = (H_\lambda - z)^{-1}$ or products of several resolvents around their deterministic approximation. Such results are commonly called *multi-resolvent global or local laws*, depending on the distance of the spectral parameter from the spectrum. For example, a typical *two-resolvent* law computes

$$\langle\!\langle G(z_1)A_1G(z_2)A_2\rangle\!\rangle = \langle\!\langle (H_\lambda - z_1)^{-1}A_1(H_\lambda - z_2)^{-1}A_2\rangle\!\rangle \tag{4.5}$$

to leading order in N, where $\langle\!\langle\cdot\rangle\!\rangle$ denotes the normalized trace, $z_1, z_2 \in \mathbb{C} \setminus \mathbb{R}$, and $A_1, A_2 \in \mathbb{C}^{N\times N}$ are deterministic matrices. Using functional calculus, the resolvents can be replaced by more general and even N-dependent functions, thus linking (4.5) to the Heisenberg time evolution. Recent work [33] establishes a multi-resolvent local law for deformed Wigner matrices in the bulk regime of the spectrum, which motivated our study of perturbed quantum systems.

4.1.1 Description of the Main Results

The principal goal of this work is a rigorous proof of the relaxation formula (Corollary 4.2.6) and prethermalization (Corollary 4.2.8) for perturbed quantum Hamiltonians of the form (4.1). We thereby assume that the unperturbed Hamiltonian H_0 has a (locally) regular limiting density of states ρ_0 around a reference energy E_0 and only energies in a microscopically large but macroscopically small interval $I_\Delta := [E_0 - \Delta, E_0 + \Delta]$ are populated by the initial state P (similar assumptions are made in [45, 47, 48]). We then show the following corollaries of our main Theorem 4.2.4:

Cor. 4.2.6: The relaxation formula (4.4) holds generally for short and long kinetic times, i.e. $t \ll \lambda^{-2}$ and $t \gg \lambda^{-2}$, corresponding to $|g_\lambda(t)|^2 \approx 1$ and $|g_\lambda(t)|^2 \approx 0$, respectively. At intermediate times, $t \sim \lambda^{-2}$ it is generally *not* valid, unless the quadratic forms $\langle \boldsymbol{u}_j, A\boldsymbol{u}_j\rangle$ of overlaps with the eigenvectors $\boldsymbol{u}_j$ of H_0 behave regularly in j (cf. Definition 4.2.5). This happens, e.g., if H_0 satisfies the Eigenstate Thermalization Hypothesis (ETH).

Cor. 4.2.8: Assuming that the unperturbed time evolution has a long time limit $\langle A\rangle_{P_0(t)} \overset{t\to\infty}{\longrightarrow} \langle A\rangle_{P_{\mathrm{pre}}}$, such that the *prethermal state* P_{pre} is distinguishable from the thermal state $\widetilde{P}_\lambda$ of the perturbed system, i.e. $\langle A\rangle_{P_{\mathrm{pre}}} \neq \langle A\rangle_{\widetilde{P}_\lambda}$ (cf. Definition 4.2.7), we show the characteristic two-step relaxation of a prethermalization process; see Fig. 4.1.

4.1.2 Outline of the Paper

The general task in this paper is to approximately evaluate the random time evolution $\langle A\rangle_{P_\lambda(t)}$ from (4.2). This is carried out in several steps summarized schematically in Fig. 4.2. First, in Theorem 4.2.4 (a) in Section 4.2.2, the true time evolution $P_\lambda(t)$ is expressed as a linear combination of the unperturbed time evolution $P_0(t)$ and another deterministic time-dependent object $\widetilde{P}_{\lambda,t}$ that is conceptually simpler than $P_\lambda(t)$. Then, in Theorem 4.2.4 (b), we identify a time-independent state $\widetilde{P}_\lambda$ as the large time limit of $\widetilde{P}_{\lambda,t}$. Combining both parts of Theorem 4.2.4, we arrive at Corollary 4.2.6, which establishes the relaxation formula (4.4) at small and large kinetic times. As mentioned above, at intermediate times, it holds only for observables having the *local overlap regularity (LOR) property* (see Definition 4.2.5). In the subsequent Section 4.2.3, dropping

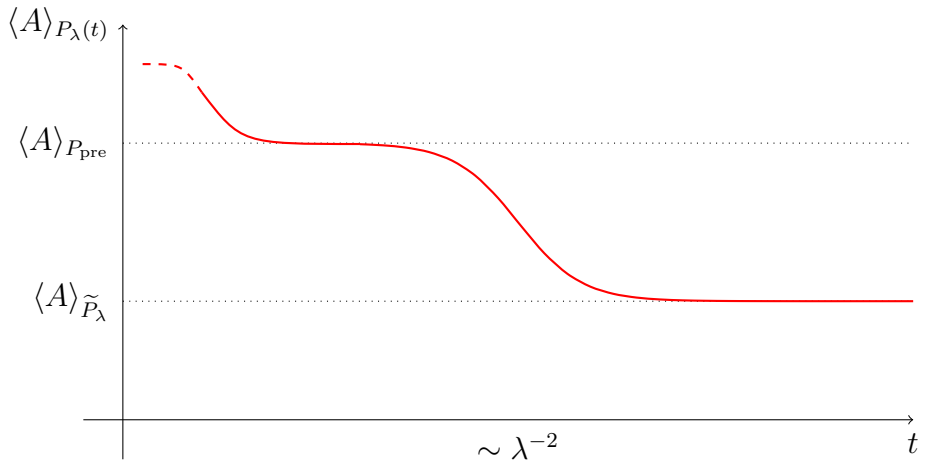

Fig. 4.1. Depicted is a schematic graph of the prethermalization process: For times $1 \ll t \ll \lambda^{-2}$, the perturbed time evolution of the quantum expectation $\langle A\rangle_{P_\lambda(t)}$ (see (4.2)) is close to the quantum expectation of A in the prethermal state P_{pre}, i.e. $\langle A\rangle_{P_\lambda(t)} \approx \langle A\rangle_{P_{\mathrm{pre}}}$. For times $t \gg \lambda^{-2}$, we have that $\langle A\rangle_{P_\lambda(t)}$ is close to the limiting thermal quantum expectation, i.e. $\langle A\rangle_{P_\lambda(t)} \approx \langle A\rangle_{\widetilde{P}_\lambda}$. This value is typically different from the prethermal quantum expectation $\langle A\rangle_{P_{\mathrm{pre}}}$. This means, the ultimate relaxation of $\langle A\rangle_{P_\lambda(t)}$ towards $\langle A\rangle_{\widetilde{P}_\lambda}$ happens via an intermediate *prethermal* value $\langle A\rangle_{P_{\mathrm{pre}}}$ in two steps, whose time scales are separated by λ^{-2}.

the LOR property, but assuming additionally that the unperturbed Hamiltonian H_0 and the initial state P have the *prethermalization property* (see Definition 4.2.7), we obtain the characteristic two-scale relaxation of $P_\lambda(t)$ towards $\widetilde{P}_\lambda$ via an intermediate *prethermal state* P_{pre} (see Corollary 4.2.8 and Fig. 4.1). As an additional result, in Theorem 4.2.9 in Section 4.2.4 we relate $\widetilde{P}_\lambda$ to the microcanonical ensemble of H_λ, called $P_\lambda^{(\mathrm{mc})}$, which is independent of the initial state P. Finally, our results are illustrated by two simple examples in Section 4.2.5.

While most proofs are given in Section 4.3, some auxiliary results and additional proofs are deferred to Appendix 4.A.

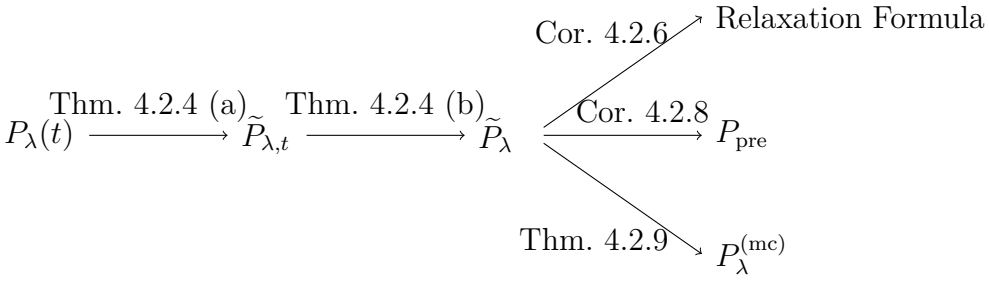

Fig. 4.2. The structure of our main results.

4.1.3 Notation

For positive quantities f, g we write $f \lesssim g$ (or $f = \mathcal{O}(g)$) and $f \sim g$ if $f \leq Cg$ or $cg \leq f \leq Cg$, respectively, for some constants $c, C > 0$ which only depend on the constants appearing in the moment condition (see (4.7)) and the definition of the set

of admissible energies (see (4.10)). In informal explanations, we frequently use the notation $f \ll g$, which indicates that f is "much smaller" than g. Moreover, we shall also write $w \approx z$ to indicate the closeness of $w, z \in \mathbb{C}$ with a not precisely specified error.

For any natural number n we set $[n] := \{1, 2, \ldots, n\}$. Matrix entries are indexed by lowercase Roman letters $a, b, c, \ldots$ and $i, j, k, \ldots$ from the beginning or the middle of the alphabet and unrestricted sums over $a, b, c, \ldots$ and $i, j, k, \ldots$ are always understood to be over $[N] = \{1, \ldots, N\}$.

We denote vectors by bold-faced lowercase Roman letters $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^N$, for some $N \in \mathbb{N}$. Vector and matrix norms, $\|\boldsymbol{x}\|$ and $\|A\|$, indicate the usual Euclidean norm and the corresponding induced matrix norm. For any $N \times N$ matrices A, B we use the notations $\langle\langle A \rangle\rangle := N^{-1}\mathrm{Tr}A$ to denote the normalized trace of A and $\langle A \rangle_B := \mathrm{Tr}[AB]$ is the trace of the product AB. We denote the spectrum of a matrix or operator A by $\sigma(A)$. Moreover, for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^N$ and matrices $A \in \mathbb{C}^{N \times N}$ we define

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \sum_i \overline{x}_i y_i \,, \qquad A_{\boldsymbol{xy}} := \langle \boldsymbol{x}, A\boldsymbol{y} \rangle.$$

For a unit vector $\boldsymbol{v} \in \mathbb{C}^N$ we shall also use the notation $|\boldsymbol{v}\rangle \langle \boldsymbol{v}|$ for the projection onto the one-dimensional subspace spanned by $\boldsymbol{v}$.

Finally, we use the concept of "with very high probability" *(w.v.h.p.)* meaning that for any fixed $C > 0$, the probability of an N-dependent event is bigger than $1 - N^{-C}$ for $N \geq N_0(C)$. We introduce the notion of *stochastic domination* (see e.g. [62]): given two families of non-negative random variables

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

indexed by N (and possibly some parameter u in some parameter space $U^{(N)}$), we say that X is stochastically dominated by Y, if for all $\xi, C > 0$ we have

$$\sup_{u \in U^{(N)}} \mathbf{P}\left[X^{(N)}(u) > N^{\xi} Y^{(N)}(u)\right] \leq N^{-C} \tag{4.6}$$

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

Acknowledgments

We thank Peter Reimann and Lennart Dabelow for helpful comments.

4.2 Main Results

In Section 4.2.1, we give the precise definition of Wigner random matrices and the assumptions on the Hamiltonian, observables and states under consideration in (4.1) to (4.3). Afterward, we formulate our main results in Sections 4.2.2–4.2.4. Finally, in Section 4.2.5 we discuss our findings in the context of two simple examples.

4.2.1 Assumptions

We begin with formulating the assumption on the Wigner matrix W.

Assumption 4.2.1 (Wigner matrix). *Let $W \equiv W(N) = (w_{ij})_{i,j\in[N]}$ from (4.1) be a real symmetric or complex Hermitian random matrix $W = W^*$ with independent entries distributed according to the laws $w_{ij} \overset{\mathrm{d}}{=} N^{-1/2}\chi_{\mathrm{od}}$ for $i < j$ and $w_{jj} \overset{\mathrm{d}}{=} N^{-1/2}\chi_{\mathrm{d}}$. The random variables χ_{od} and χ_{d} satisfy the following assumptions:[2] We assume that χ_{d} is a centered real random variable, and χ_{od} is a real or complex random variable with $\mathbb{E}\chi_{\mathrm{od}} = 0$ and $\mathbb{E}|\chi_{\mathrm{od}}|^2 = 1$. Furthermore, we assume the existence of higher moments, namely*

$$\mathbb{E}|\chi_{\mathrm{d}}|^p + \mathbb{E}|\chi_{\mathrm{od}}|^p \leq C_p, \tag{4.7}$$

for all $p \in \mathbb{N}$, where C_p are positive constants.

For concreteness, we focus on the complex case with the additional assumptions $\mathbb{E}\chi_{\mathrm{od}}^2 = 0$ and $\mathbb{E}|\chi_{\mathrm{d}}|^2 = 1$; all other cases can also be handled as in [36]. The precise conditions on the Wigner matrix only play a role in the underlying *two resolvent global law* (Proposition 4.3.1).

For the Hamiltonian $H_0 \equiv H_0(N)$ in (4.1) we assume the following.

Assumption 4.2.2 (H_0 and its density of states). *The Hamiltonian $H_0 \equiv H_0(N)$ is deterministic, self-adjoint $H_0 = H_0^*$, and uniformly bounded $\|H_0\| \lesssim 1$. We denote the resolvent of H_0 at any spectral parameter $z \in \mathbb{C} \setminus \mathbb{R}$ by*

$$M_0(z) := \frac{1}{H_0 - z}.$$

Moreover, we assume the following:

(i) There exists a compactly supported measurable function $\rho_0 : \mathbb{R} \to [0, +\infty)$ with

$$\int_{\mathbb{R}} \rho_0(x)\mathrm{d}x = 1$$

and two positive sequences $\epsilon_0(N)$ and $\eta_0(N)$, both converging to zero as $N \to \infty$, such that, uniformly in $z \in \mathbb{C}\setminus\mathbb{R}$ with $\eta := |\Im z| \geq \eta_0 \equiv \eta_0(N)$, we have

$$\left\langle\!\left\langle M_0(z) \right\rangle\!\right\rangle = m_0(z) + \mathcal{O}(\epsilon_0) \quad \text{with} \quad \epsilon_0 \equiv \epsilon_0(N)\,. \tag{4.8}$$

Here,

$$m_0(z) := \int_{\mathbb{R}} \frac{\rho_0(x)}{x - z}\mathrm{d}x \tag{4.9}$$

is the Stieltjes transform of ρ_0. We refer to ρ_0 as the limiting density of states*, and to* $\mathrm{supp}(\rho_0)$ *as the* limiting spectrum *of H_0.*

(ii) For small positive constants $\kappa, c > 0$, we define the set of admissible energies $\sigma_{\mathrm{adm}}^{(\kappa,c)}$ *in the limiting spectrum of H_0 by*[3]

$$\sigma_{\mathrm{adm}}^{(\kappa,c)} := \left\{ x \in \mathrm{supp}(\rho_0) : \inf_{|y-x|\leq\kappa} \rho_0(y) > c,\ \|\rho_0\|_{C^{1,1}([x-\kappa,x+\kappa])} \leq 1/c \right\}. \tag{4.10}$$

We assume that for some positive $\kappa, c > 0$, $\sigma_{\mathrm{adm}}^{(\kappa,c)}$ is not empty.

[2] A careful examination of our proof reveals that the entries of W need not be distributed identically. Indeed, only the matching of the second moments is necessary, but higher moments can differ.

[3] Here, $C^{1,1}(J)$ denotes the set of continuously differentiable functions with a Lipschitz-continuous derivative on an interval J, equipped with the norm $\|f\|_{C^{1,1}(J)} := \|f\|_{C^1(J)} + \sup_{x,y\in J: x\neq y} \frac{|f'(x)-f'(y)|}{|x-y|}$.

Assuming that the set of admissible energies in (4.10) is non-empty ensures that there is a part of the limiting spectrum $\mathrm{supp}(\rho_0)$, where the limiting density of states ρ_0 behaves regularly, i.e. it is strictly positive and sufficiently smooth. Finally, we formulate our assumptions on the observables and states considered in (4.2) and (4.3).

Assumption 4.2.3 (States and observables)**.** *Given Assumption 4.2.2, we first pick a* reference energy

$$E_0 \in \sigma_{\mathrm{adm}}^{(\kappa_0,c_0)} \quad \text{for some} \quad \kappa_0, c_0 > 0, \tag{4.11}$$

and further introduce $I_\delta := [E_0 - \delta, E_0 + \delta]$ *for any* $0 < \delta < \kappa_0$*. Moreover, take an* energy width $\Delta \in (0, \frac{1}{6}\kappa_0)$ *and let* $\Pi_\Delta := \mathbb{1}_{I_\Delta}(H_0)$ *be the spectral projection of* H_0 *onto the interval* I_Δ*. Then, we assume the following:*

(i) The (deterministic) initial state $P \equiv P(N) \in \mathbb{C}^{N\times N}$ *in* (4.3) *is a state in the usual sense (*$P = P^*$*,* $0 \leq P \leq 1$*, and* $\mathrm{Tr}[P] = 1$*), and is localized in* I_Δ*, i.e.*

$$P = \Pi_\Delta P \Pi_\Delta \,. \tag{4.12}$$

(ii) The observable $A \equiv A(N) \in \mathbb{C}^{N\times N}$ *is deterministic, self-adjoint and satisfies the bound* $\|A\| \lesssim 1$*.*

Note that we assume only the state P to be localized in I_Δ and not the observable. However, by inspecting the proof, we see that A and P play essentially symmetric roles, and thus, our results hold verbatim if we assume localization of A instead of P. Moreover, for ease of notation, we drop the N-dependence of all the involved matrices.

4.2.2 Relaxation of Perturbed Quantum Dynamics

In this section, we present our main result on the time evolution of the random quantum expectation $\langle A\rangle_{P_\lambda(t)}$ from (4.2). Its relaxation is described in two steps, hence Parts (a) and (b) in the following theorem. In the first step, we eliminate the randomness and identify the leading deterministic part of $\langle A\rangle_{P_\lambda(t)}$ in terms of M_0, the unperturbed resolvent. In the second step, we consider the short and long-time limits of the leading term. Further explanatory comments come after the theorem and in Remark 4.2.2 below.

Theorem 4.2.4 (Relaxation of perturbed dynamics)**.** *Let* $H_\lambda = H_0 + \lambda W$ *be a perturbed Hamiltonian like in* (4.1) *with* $\lambda > 0$*, whose constituents satisfy Assumptions 4.2.1 and 4.2.2, respectively. Pick a reference energy* E_0 *like in* (4.11)*. Let* P *be a state satisfying Assumption 4.2.3 (i) for some energy width* $\Delta > 0$ *and* A *an observable satisfying Assumption 4.2.3 (ii). Then, using the notations* (4.2) *and* (4.3)*, we have the following two approximation statements:*

(a) [Relaxation in the kinetic limit] *The perturbed dynamics* $\langle A\rangle_{P_\lambda(t)}$ *satisfies*

$$\langle A\rangle_{P_\lambda(t)} = |g_\lambda(t)|^2 \langle A\rangle_{P_0(t)} + \langle A\rangle_{\widetilde{P}_{\lambda,t}} + \mathcal{E}, \tag{4.13}$$

where we denoted

$$g_\lambda(t) := \mathrm{e}^{-\alpha\lambda^2 t} \quad \text{with} \quad \alpha := \pi\rho_0(E_0). \tag{4.14}$$

Moreover, we introduced[4]

$$\widetilde{P}_{\lambda,t} := \frac{\int_{\mathbb{R}}\int_{\mathbb{R}} \Im M_0(x + \mathrm{i}\alpha\lambda^2)\, K_{\lambda,t}(x-y)\, \langle \Im M_0(y + \mathrm{i}\alpha\lambda^2)\rangle_P \,\mathrm{d}x\mathrm{d}y}{\int_{\mathbb{R}} \mathrm{Tr}[\Im M_0(x + \mathrm{i}\alpha\lambda^2)]\, \langle \Im M_0(x + \mathrm{i}\alpha\lambda^2)\rangle_P \,\mathrm{d}x}, \tag{4.15}$$

[4]Recall that the imaginary part of a matrix $B \in \mathbb{C}^{N\times N}$ is given by $\Im B = \frac{1}{2i}(B - B^*)$.

with an explicit kernel given by

$$K_{\lambda,t}(u) := \frac{1}{\pi}\frac{2\alpha\lambda^2}{u^2+(2\alpha\lambda^2)^2}\left(2\alpha\lambda^2 t\frac{\sin(tu)}{tu} - \cos(tu) + \mathrm{e}^{-2\alpha\lambda^2 t}\right). \tag{4.16}$$

Finally, we have $\mathcal{E} = \mathcal{O}(\mathcal{E}_0) + \mathcal{O}_\prec(C(\lambda,t)/\sqrt{N})$ *for some constant* $C(\lambda,t) > 0$ *and for every fixed* $T \in (0,\infty)$*, the deterministic error* $\mathcal{E}_0 = \mathcal{E}_0(\lambda,t,\Delta,N)$ *satisfies*

$$\lim_{\Delta\to 0}\lim_{\substack{t\to\infty\\ \lambda\to 0\\ \lambda^2 t=T}}\lim_{N\to\infty}\mathcal{E}_0 = 0. \tag{4.17}$$

(b) [Long and short kinetic time limit] *Defining*

$$\widetilde{P}_\lambda := \frac{\int_{\mathbb{R}} \Im M_0(x+\mathrm{i}\alpha\lambda^2)\,\langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle_P\,\mathrm{d}x}{\int_{\mathbb{R}} \mathrm{Tr}[\Im M_0(x+\mathrm{i}\alpha\lambda^2)]\,\langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle_P\,\mathrm{d}x}, \tag{4.18}$$

it holds that

$$\langle A\rangle_{\widetilde{P}_{\lambda,t}} = \left(1-|g_\lambda(t)|^2\right)\langle A\rangle_{\widetilde{P}_\lambda} + \mathcal{R}, \tag{4.19}$$

where, for every fixed $T \in (0,\infty)$*, the error term* $\mathcal{R} = \mathcal{R}(\lambda,t,\Delta,N)$ *satisfies*

$$\limsup_{\Delta\to 0}\limsup_{\substack{t\to\infty\\ \lambda\to 0\\ \lambda^2 t=T}}\limsup_{N\to\infty}|\mathcal{R}| \lesssim T\mathrm{e}^{-2\alpha T}. \tag{4.20}$$

We point out that the error $\mathcal{E}$ in (4.13) naturally consists of two parts, a deterministic and a stochastic one. The stochastic part of order $\mathcal{O}_\prec(C(\lambda,t)/\sqrt{N})$ is obtained from a *global law* for two resolvents of the random matrix H_λ (see (4.46) below); the deterministic part $\mathcal{O}(\mathcal{E}_0)$ is obtained from estimating the deterministic leading term in (4.47). Note that the error $\mathcal{R}$ is small compared to the first term in the rhs. of (4.19) only in the regime where T is large, in particular $\langle A\rangle_{\widetilde{P}_{\lambda,t}}$ converges to $\langle A\rangle_{\widetilde{P}_\lambda}$ exponentially fast. In the small T regime, both terms on the right-hand side of (4.19) vanish linearly in T. We chose the above formulation (4.19) because, in this way, it relates directly to the relaxation formula (4.4) (see Corollary 4.2.6 below).

Remark. We have two further comments on Theorem 4.2.4.

(i) The triple limits in (4.17) and (4.20) consist of a thermodynamic limit ($N\to\infty$), a kinetic limit or *van Hove limit* ($t\to\infty$ and $\lambda\to 0$ while $\lambda^2 t$ is fixed), and an infinitesimal spectral localization ($\Delta\to 0$). Note that the *kinetic time parameter* $T = \lambda^2 t$ is natural, as the time scale prescribed for relaxation in the physics literature, e.g., by explicit analysis of the Pauli master equation in [115, Sect. 5.2.6], is $\mathcal{O}(\lambda^{-2})$. The $\Delta\to 0$ limit is needed only to ensure that the mean level spacing is approximately constant near E_0 on the scale Δ. If the density of states is flat, the $\Delta\to 0$ limit can be omitted. We emphasize that the error terms in Theorem 4.2.4 are explicit in the sense that their dependence on the scaling parameters N, t, λ, and Δ is tracked throughout the proof. The limit in (4.17) is then the natural order of limits in which these errors vanish. Finally, we remark that the explicitly tracked errors allow for certain combined limits, although for simplicity, we do not pursue these extensions.

(ii) The idea behind (4.18)–(4.20) is that the kernel (4.16) is an approximate delta function with $T = \lambda^2 t$-dependent magnitude. More precisely, its Fourier transform[5]

$$\widehat{K}_{\lambda,t}(p) = \frac{1}{\sqrt{2\pi}} \left(1 - \mathrm{e}^{-2\alpha\lambda^2(t-|p|)}\right) \mathbb{1}(|p| \leq t)$$

converges uniformly in compact intervals to a constant. That is, for every fixed $T \in (0,\infty)$ and for every compact set $\Omega \subset \mathbb{R}$, we have

$$\lim_{\substack{t\to\infty \\ \lambda\to 0 \\ \lambda^2 t = T}} \sup_{p\in\Omega} \left| \widehat{K}_{\lambda,t}(p) - \frac{1}{\sqrt{2\pi}} \left(1 - \mathrm{e}^{-2\alpha T}\right) \right| = 0.$$

However, since $x \mapsto \Im M_0(x + \mathrm{i}\alpha\lambda^2)$ is only regular on scale λ^2, the approximation

$$K_{\lambda,t}(x-y) \approx \left(1 - \mathrm{e}^{-2\alpha\lambda^2 t}\right)\delta(x-y),$$

used in heuristically obtaining (4.19) from (4.15) and (4.18), is *not* generally valid unless T is very small or very large, as (4.20) indicates.

In the remaining part of the current Section 4.2.2, we connect Theorem 4.2.4 to the relaxation formula (4.4). As a preparation, we formulate the following *local overlap regularity property*, required in Corollary 4.2.6 (c) below. For this purpose, let μ_j and $\boldsymbol{u}_j$ denote the eigenvalues and corresponding normalized eigenvectors of

$$H_0 = \sum_j \mu_j \left|\boldsymbol{u}_j\right\rangle \left\langle \boldsymbol{u}_j\right| . \tag{4.21}$$

Definition 4.2.5 (Local overlap regularity (LOR))**.** *Let the Hamiltonian H_0 be as in Assumption 4.2.2. We say that an observable A satisfying Assumption 4.2.3 (ii) has the* local overlap regularity (LOR) property *if and only if the eigenvector overlaps $\langle \boldsymbol{u}_j, A\boldsymbol{u}_j\rangle$ are approximately constant in the following sense: There exists a constant $\mathfrak{A} \in \mathbb{R}$ such that*[6]

$$\langle \boldsymbol{u}_j, A\boldsymbol{u}_j\rangle = \mathfrak{A} + \mathcal{O}\big(\mathcal{E}_{\mathrm{LOR}}\big) \quad \textit{for all} \quad j \in \mathbb{N} \quad \textit{with} \quad \mu_j \in I_{2\Delta}\,, \tag{4.22}$$

where the error $\mathcal{E}_{\mathrm{LOR}} = \mathcal{E}_{\mathrm{LOR}}(\Delta, N)$ satisfies

$$\lim_{\Delta\to 0} \lim_{N\to\infty} \mathcal{E}_{\mathrm{LOR}} = 0.$$

[5] We use the convention that the Fourier transform of $f \in L^1(\mathbb{R})$ is defined as $\widehat{f}(p) := (2\pi)^{-1/2} \int_{\mathbb{R}} f(x)\mathrm{e}^{-\mathrm{i}px}\mathrm{d}x$.

[6] In fact, it is sufficient to assume that the overlaps are regular in j in the sense that there exists a uniformly equicontinuous sequence $(\mathfrak{A}_N)_{N\in\mathbb{N}}$ of functions $\mathfrak{A}_N : I_{2\Delta} \to \mathbb{R}$ such that

$$\langle \boldsymbol{u}_j, A\boldsymbol{u}_j\rangle = \mathfrak{A}_N(\mu_j) + \mathcal{O}\big(\mathcal{E}_{\mathrm{LOR}}\big) \quad \text{for all} \quad j \in \mathbb{N} \quad \text{with} \quad \mu_j \in I_{2\Delta}.$$

Alternatively, we could also assume that, for every fixed $\Delta > 0$ small enough, the overlaps $\langle \boldsymbol{u}_j, A\boldsymbol{u}_j\rangle$ are well approximated by $\mathfrak{A}_N(\mu_j)$ in ℓ^p-sense for some $p \geq 1$, i.e.,

$$\frac{1}{N} \sum_{j:\mu_j\in I_{2\Delta}} |\langle \boldsymbol{u}_j, A\boldsymbol{u}_j\rangle - \mathfrak{A}_N(\mu_j)|^p \overset{N\to\infty}{\longrightarrow} 0\,.$$

The case $p = 2$ is reminiscent of the so called *weak ETH* studied in [24, 46].

The LOR property (4.22) is satisfied, e.g., if H_0 satisfies the Eigenstate Thermalization Hypothesis [51, 136] (see also the discussion in [47]). For general systems, the ETH remains an unproven hypothesis. We remark, however, that it has been rigorously proven for a large class of mean-field random matrices H_0 (see [36, 33]), including Wigner matrices and their deformations.

The following corollary collects our rigorous results on the relaxation formula (4.4). The first two parts (items (a) and (b) below) immediately follow from Theorem 4.2.4 (a) and (b). The third part, item (c), involves the LOR property of the observable A and requires a separate argument, provided in Section 4.3.4.

Corollary 4.2.6 (Relaxation formula)**.** *Under the assumptions and using the notations of Theorem 4.2.4, it holds that*

$$\langle A\rangle_{P_\lambda(t)} = \langle A\rangle_{\widetilde{P}_\lambda} + |g_\lambda(t)|^2 \left[\langle A\rangle_{P_0(t)} - \langle A\rangle_{\widetilde{P}_\lambda}\right] + \mathcal{R} + \mathcal{E}. \tag{4.23}$$

In particular, we have the following:

(a) [Short kinetic time behavior] *Let* $0 < T \lesssim 1$*. Then, it holds that*

$$\limsup_{\Delta\to 0} \limsup_{\substack{t\to\infty \\ \lambda\to 0 \\ \lambda^2 t = T}} \limsup_{N\to\infty} \left|\langle A\rangle_{P_\lambda(t)} - \langle A\rangle_{P_0(t)}\right| \lesssim T \quad \text{almost surely (a.s.)} \tag{4.24}$$

(b) [Long kinetic time behavior] *Let* $T \gtrsim 1$*. Then it holds that (recall* $\alpha = \pi\rho_0(E_0)$*)*

$$\limsup_{\Delta\to 0} \limsup_{\substack{t\to\infty \\ \lambda\to 0 \\ \lambda^2 t = T}} \limsup_{N\to\infty} \left|\langle A\rangle_{P_\lambda(t)} - \langle A\rangle_{\widetilde{P}_\lambda}\right| \lesssim T\mathrm{e}^{-2\alpha T} \quad \text{a.s.} \tag{4.25}$$

Moreover, additionally assuming that A *has the LOR property from Definition 4.2.5:*

(c) [Intermediate kinetic times under LOR] *For every fixed* $T \in (0,\infty)$ *it holds that*

$$\lim_{\Delta\to 0} \lim_{\substack{t\to\infty \\ \lambda\to 0 \\ \lambda^2 t = T}} \lim_{N\to\infty} \left[|\mathcal{R}| + |\mathcal{E}|\right] = 0 \quad \text{a.s.}, \tag{4.26}$$

i.e., the relaxation formula (4.4) *is valid at* all *kinetic times* $T \in (0,\infty)$*.*

Summarizing Corollary 4.2.6, we have that the relaxation formula (4.4) generally holds in the two limiting regimes (a) $|g_\lambda(t)|^2 \approx 0$ and (b) $|g_\lambda(t)|^2 \approx 1$, i.e. $T \ll 1$ or $T \gg 1$, respectively. In between, (4.4) is valid under the additional assumption that A has the LOR property from Definition 4.2.5, as this allows for the improved bound (4.26) on $\mathcal{R}$ compared to (4.20). However, *without* this regularity assumption, only the bound (4.20) (i.e., (4.24) and (4.25)) can hold, which indicates that the relaxation formula (4.4) is *not* generally valid for intermediate kinetic times $T \sim 1$. Indeed, it is easy to construct a counterexample. Finally, we remark that Corollary 4.2.6 (c) holds verbatim if the state P satisfies the LOR condition instead of the observable A. This simply follows by inspecting the proof in Section 4.3.4.

Remark. We have two further comments on Corollary 4.2.6.

(i) The relaxation formula (4.23) is in perfect agreement with the main result of Dabelow and Reimann, see [47, Eq. (16)]. In fact, the state $\widetilde{P}_\lambda$ defined in (4.18) agrees with $\widetilde{\rho}$ from [47, Eq. (16)], named the "washed out descendant of the so-called diagonal ensemble" [47].

(ii) In fact, recalling (4.21), the proof of Theorem 4.2.4 (b) in Section 4.3.3 reveals that (see (4.111)) the error $\mathcal{R}$ in (4.19) and (4.23) is given by

$$\mathcal{R} = \frac{1}{r}\sum_{j,k}\langle \boldsymbol{u}_j, A\boldsymbol{u}_j\rangle\langle \boldsymbol{u}_k, P\boldsymbol{u}_k\rangle \mathfrak{R}_{\lambda,t}(\mu_j-\mu_k), \tag{4.27}$$

where we denoted

$$r := \int_{\mathbb{R}} \mathrm{Tr}[\Im M_0(x+\mathrm{i}\alpha\lambda^2)]\,\langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle_P \,\mathrm{d}x \tag{4.28}$$

and

$$\mathfrak{R}_{\lambda,t}(u) := \pi \mathrm{e}^{-2\alpha\lambda^2 t}\frac{2\alpha\lambda^2}{u^2+(2\alpha\lambda^2)^2}\left(1-\cos(tu)-2\alpha\lambda^2 t\frac{\sin(tu)}{tu}\right). \tag{4.29}$$

The explicit error term (4.27)–(4.29) is in precise agreement with the error term in [47], see Eqs. (17) and (18). In particular, assuming the LOR property (4.22) for A (or P), the smallness of $\mathcal{R}$ in (4.26) for all kinetic times $T \in (0,\infty)$ is a consequence of the fact that $\int_{\mathbb{R}} \mathfrak{R}_{\lambda,t}(u)\mathrm{d}u = 0$.

4.2.3 Prethermalization

In this section, we specialize the general relaxation theory of perturbed quantum dynamics from Theorem 4.2.4 to a class of unperturbed Hamiltonians H_0 and states P which *has the prethermalization property* in the following sense.

Definition 4.2.7 (Prethermalization property)**.** *Let the Hamiltonian H_0 and the state P be defined as in Assumptions 4.2.2 and 4.2.3. We say that (H_0, P) has the* prethermalization property *if and only if there exists a state P_{pre} (called the* prethermal state*) such that we have the following:*

(a) The unperturbed time evolution $P_0(t)$ converges to P_{pre}, i.e., for all[7] observables A satisfying Assumption 4.2.3 (ii), it holds that

$$\lim_{t\to\infty}\lim_{N\to\infty}\left[\langle A\rangle_{P_0(t)} - \langle A\rangle_{P_{\mathrm{pre}}}\right] = 0. \tag{4.30a}$$

(b) There exists an observable A_0 (satisfying Assumption 4.2.3 (ii)) which distinguishes P_{pre} from $\widetilde{P}_\lambda$ (cf. (4.18)), i.e. there exists a constant $\mathfrak{c}_{\mathrm{pre}} > 0$ such that

$$\liminf_{\lambda\to 0}\liminf_{N\to\infty}\left|\langle A_0\rangle_{P_{\mathrm{pre}}} - \langle A_0\rangle_{\widetilde{P}_\lambda}\right| \geq \mathfrak{c}_{\mathrm{pre}}. \tag{4.30b}$$

[7] A common variant of this requirement in the physics literature is to assume the validity of (4.30a) only for *local* observables, i.e. supported in a finite region of an underlying space (see, e.g., [115, Section 5.2]).

We emphasize that (H_0, P) having the *prethermalization property* is a purely deterministic condition, i.e., in particular, it does not depend on the Wigner matrix W. In the physics literature (see, e.g., [108] but also [115, 100, 97, 15]), the prethermalization property is generally expected to be satisfied if H_0 is an *integrable* Hamiltonian having at least one additional conserved quantity Q for which $[H_0, Q] = 0$.[8] This symmetry is then broken by a generic perturbation W, i.e. $[W, Q] \neq 0$. In the presence of M conserved quantities $(Q_k)_{k=1}^M$, a good candidate for the prethermal state P_{pre} is given by the so called *generalized Gibbs ensemble (GGE)*

$$P_{\mathrm{GGE}} = \frac{\mathrm{e}^{-\sum_{k=1}^M \lambda_k Q_k}}{\operatorname{Tr} \mathrm{e}^{-\sum_{k=1}^M \lambda_k Q_k}},$$

where the parameters λ_k are chosen in such a way that $\operatorname{Tr} Q_k P_{\mathrm{GGE}} = \operatorname{Tr} Q_k P$ for all $k \in [M]$ (see, e.g., [97] and [115, Section 5.1]). Exemplary pairs (H_0, P) and observables A_0 satisfying the conditions in Definition 4.2.7 are given in Section 4.2.5.

Assuming that (H_0, P) has the prethermalization property, Theorem 4.2.4 reads as follows.

Corollary 4.2.8 (Prethermalization)**.** *Under the assumptions of Theorem 4.2.4, let further (H_0, P) have the prethermalization property from Definition 4.2.7. Then, recalling the notations from Theorem 4.2.4, it holds that*

$$\langle A \rangle_{P_\lambda(t)} = \langle A \rangle_{\widetilde{P}_\lambda} + |g_\lambda(t)|^2 \left[\langle A \rangle_{P_{\mathrm{pre}}} - \langle A \rangle_{\widetilde{P}_\lambda} \right] + \mathcal{R} + \mathcal{E}'. \tag{4.31}$$

We have $\mathcal{E}' = \mathcal{O}(\mathcal{E}_0') + \mathcal{O}_\prec(C(\lambda, t)/\sqrt{N})$ for some constant $C(\lambda, t) > 0$ and for every fixed $T \in (0, \infty)$, the deterministic errors $\mathcal{E}_0' = \mathcal{E}_0'(\lambda, t, \Delta, N)$ and $\mathcal{R} = \mathcal{R}(\lambda, t, \Delta, N)$ satisfy

$$\lim_{\Delta \to 0} \lim_{\substack{t \to \infty \\ \lambda \to 0 \\ \lambda^2 t = T}} \lim_{N \to \infty} \mathcal{E}_0' = 0 \qquad \text{and} \qquad \limsup_{\Delta \to 0} \limsup_{\substack{t \to \infty \\ \lambda \to 0 \\ \lambda^2 t = T}} \limsup_{N \to \infty} |\mathcal{R}| \lesssim T \mathrm{e}^{-2\alpha T}.$$

We remark that the error term $\mathcal{E}_0'$ contributing in (4.31) consists of two parts, $\mathcal{E}_0' = \mathcal{E}_0 + \mathcal{E}_{\mathrm{pre}}$, with $\mathcal{E}_0 = \mathcal{E}_0(\lambda, t, \Delta, N)$ from Theorem 4.2.4 and $\mathcal{E}_{\mathrm{pre}} = \mathcal{E}_{\mathrm{pre}}(t, N)$ being the (absolute value of the) error in (4.30a). Note that (4.31) in particular implies the following small and large T behaviors:

$$\begin{aligned} &\limsup_{\Delta \to 0} \limsup_{\substack{t \to \infty \\ \lambda \to 0 \\ \lambda^2 t = T}} \limsup_{N \to \infty} \left| \langle A \rangle_{P_\lambda(t)} - \langle A \rangle_{P_{\mathrm{pre}}} \right| \lesssim T \quad \text{for} \quad T \lesssim 1 \quad \text{a.s.}, \\ \text{and} \quad &\limsup_{\Delta \to 0} \limsup_{\substack{t \to \infty \\ \lambda \to 0 \\ \lambda^2 t = T}} \limsup_{N \to \infty} \left| \langle A \rangle_{P_\lambda(t)} - \langle A \rangle_{\widetilde{P}_\lambda} \right| \lesssim T \mathrm{e}^{-2\alpha T} \quad \text{for} \quad T \gtrsim 1 \quad \text{a.s.} \end{aligned} \tag{4.32}$$

Moreover, (4.30b) ensures that $\langle A \rangle_{P_{\mathrm{pre}}} \neq \langle A \rangle_{\widetilde{P}_\lambda}$ for at least one observable $A = A_0$, which, together with (4.32) establishes Fig. 4.1 as a schematic graph of a prethermalization process.

4.2.4 Connection to the Microcanonical Ensemble

Under an additional regularity assumption on $x \mapsto \langle\!\langle \Im M_0(x + \mathrm{i}\alpha\lambda^2) A \rangle\!\rangle$ we can relate the state $\widetilde{P}_\lambda$ from (4.18) to the microcanonical ensemble.

[8] We use the usual notation for the commutator, i.e. $[A, B] := AB - BA$ for all matrices A, B.

Theorem 4.2.9 (Microcanonical average)**.** *Under the assumptions of Theorem 4.2.4, let us further assume that*

$$h \equiv h(\lambda, N) : x \mapsto \langle\!\langle \Im M_0(x + \mathrm{i}\alpha\lambda^2) A \rangle\!\rangle$$

is a Lipschitz continuous map on I_Δ with Lipschitz constant $\mathrm{Lip}_{I_\Delta}(h)$ bounded in the sense that

$$\limsup_{\Delta\to 0} \limsup_{\lambda\to 0} \limsup_{N\to\infty} \mathrm{Lip}_{I_\Delta}(h) \lesssim 1. \tag{4.33}$$

Then,

$$\langle A \rangle_{\widetilde{P}_\lambda} = \langle A \rangle_{P_\lambda^{(\mathrm{mc})}} + \mathcal{E}_{\mathrm{mc}} \qquad \textit{with} \qquad P_\lambda^{(\mathrm{mc})} := \frac{\Im M_0(E_0 + \mathrm{i}\alpha\lambda^2)}{\mathrm{Tr}[\Im M_0(E_0 + \mathrm{i}\alpha\lambda^2)]}, \tag{4.34}$$

where the error $\mathcal{E}_{\mathrm{mc}} = \mathcal{E}_{\mathrm{mc}}(\lambda, \Delta, N)$ satisfies

$$\lim_{\Delta\to 0} \lim_{\lambda\to 0} \lim_{N\to\infty} \mathcal{E}_{\mathrm{mc}} = 0.$$

We emphasize that $P_\lambda^{(\mathrm{mc})}$ is completely independent of the initial state P. Moreover, as mentioned above and already indicated by the notation, we can interpret $\langle A \rangle_{P_\lambda^{(\mathrm{mc})}}$ from (4.34) as the microcanonical average of H_λ at energy E_0. The reason underlying this interpretation is that for any normalized eigenvector $\boldsymbol{v}_\lambda$ of H_λ with eigenvalue E_λ very close to E_0, it holds that[9]

$$\langle \boldsymbol{v}_\lambda, A\boldsymbol{v}_\lambda \rangle \approx \frac{\mathrm{Tr}[\Im M_0(E_0 + \mathrm{i}\alpha\lambda^2) A]}{\mathrm{Tr}[\Im M_0(E_0 + \mathrm{i}\alpha\lambda^2)]} = \langle A \rangle_{P_\lambda^{(\mathrm{mc})}}.$$

This means, $P_\lambda^{(\mathrm{mc})}$ is a close effective approximation to the actual projection $|\boldsymbol{v}_\lambda\rangle \langle \boldsymbol{v}_\lambda|$ onto the eigenspace spanned by $\boldsymbol{v}_\lambda$.

4.2.5 Examples

In this section, we give two examples of physical settings where prethermalization occurs and connect them to our assumptions. Note that both examples are one-dimensional, however, the extension to higher dimensions is straightforward.

Next-Nearest Neighbor Hopping

For $N \in \mathbb{N}$ even, we consider the Laplacian-like Hamiltonian H_0 acting on functions $\psi \in \ell^2(\mathbb{Z}/(N\mathbb{Z}))$ as

$$(H_0\psi)(x) := 2\psi(x) - \psi(x-2) - \psi(x+2) \tag{4.35}$$

where $x - 2$ and $x + 2$ are interpreted mod N. Note that H_0 is similar to the discrete Laplacian with periodic boundary condition but induces next-nearest neighbor hopping instead of nearest neighbor hopping. In particular, H_0 conserves parity in the sense that functions that are only supported on the even or odd points of $\mathbb{Z}/(N\mathbb{Z})$, respectively, remain invariant, and thus its spectrum has an additional two-fold degeneracy.

[9] Indeed, taking, say, $E_\lambda \in I_{\Delta/2}$ it can be rigorously shown that the difference $\langle \boldsymbol{v}_\lambda, A\boldsymbol{v}_\lambda \rangle - \langle A \rangle_{P_\lambda^{(\mathrm{mc})}}$ vanishes in the triple limit (4.17). More precisely, this follows from the *Eigenstate Thermalization Hypothesis (ETH)* for the random matrix $H_\lambda = H_0 + \lambda W$ (see [33, Theorem 2.6]) and using assumption (4.33) together with (4.42) and (4.133).

Similar to the routine computations done for the discrete Laplacian, one can readily check the following:

- The Hamiltonian H_0 is bounded, $\|H_0\| \lesssim 1$.
- Its spectrum is given by $\sigma(H_0) = \{2(1-\cos(2p_j)) : p_j = 2\pi j/N\}_{j\in[N]} \subset [0,4]$.
- The limiting density of states as $N \to \infty$ evaluates to

$$\rho_0(x) = \frac{1}{\pi\sqrt{x(4-x)}} \mathbb{1}_{[0,4]}(x) \tag{4.36}$$

  which is compactly supported and satisfies the regularity assumptions in Assumption 4.2.2 for x bounded away from 0 and 4.

In this setting, we fix k such that the eigenvalue $2(1-\cos(2p_k))$ satisfies $p_k \in (0, \pi/2)$. Now take $P := |\boldsymbol{u}_k\rangle \langle \boldsymbol{u}_k|$ with $\boldsymbol{u}_k$ being the normalized eigenvector of H_0 supported on the *even* sub-lattice corresponding to the eigenvalue $2(1-\cos(2p_k))$. By construction, for every observable A satisfying Assumption 4.2.3, we have

$$\langle A \rangle_{P_0(t)} = \langle A \rangle_P = \langle A \rangle_{P_{\mathrm{pre}}}, \quad \text{for all} \quad t \geq 0,$$

since $[P, H_0] = 0$. Hence, the symmetry implies that $P_{\mathrm{pre}} = P$. In particular, for $A := \mathbf{1}_{\mathrm{odd}}$ being the identity operator on the *odd* sub-lattice, its prethermal value is given by $\langle A \rangle_{P_{\mathrm{pre}}} = 0$. Moreover, by spectral decomposition of $H_0 = \sum_j \mu_j |\boldsymbol{u}_j\rangle \langle \boldsymbol{u}_j|$, we obtain

$$\langle\!\langle \Im M_0(x + \mathrm{i}\lambda^2\alpha) A \rangle\!\rangle = \frac{1}{N} \sum_j \langle \boldsymbol{u}_j, A\boldsymbol{u}_j \rangle \frac{\alpha\lambda^2}{|x-\mu_j|^2 + (\alpha\lambda^2)^2} > 0, \tag{4.37}$$

which implies that $\langle A \rangle_{\widetilde{P}_\lambda} \neq \langle A \rangle_{P_{\mathrm{pre}}}$ (recall the definition of $\widetilde{P}_\lambda$ in (4.18)) for $A = \mathbf{1}_{\mathrm{odd}}$. Hence, we deduce that (H_0, P) has the prethermalization property from Definition 4.2.7.

Free Spinless Fermions on a Lattice

As our second example, we consider a model of spinless fermions in a periodic one-dimensional lattice of even length N (cf. [47, App. B]), which can be seen as a many-body analog of the first example (although with nearest neighbor hopping instead of next-nearest neighbor hopping). Let

$$H_0 = \frac{1}{\sqrt{N}} \sum_j c_j^\dagger c_{j+1} + c_{j+1}^\dagger c_j, \tag{4.38}$$

where $c_j^\dagger$ and c_j denote the fermionic creation and annihilation operators at site j, and the summation indices are considered modulo N. Note that the Hamiltonian in (4.38) conserves the particle number. It is readily checked that H_0 admits a limiting density of states which is not compactly supported but has fast decaying (Gaussian) tails. As the regularity assumptions in Assumption 4.2.2 (ii) are satisfied, this example is still sufficiently close to our theory to be described by it reasonably well.

In this setting, pick ψ_j as the orthonormal eigenfunctions of the discrete Laplacian describing nearest neighbor hopping with periodic boundary conditions (i.e. the analog of (4.35) with ± 1 instead of ± 2) corresponding to the eigenvalues $2(1-\cos(p_j))$ with

$$p_j := \frac{2\pi j}{N}, \quad \frac{j}{N} \in \left[\frac{1}{8}, \frac{3}{8}\right] \cup \left[\frac{5}{8}, \frac{7}{8}\right].$$

We then construct $P := |\psi\rangle\langle\psi|$ as a rank-1 projection onto an eigenstate of H_0 by taking

$$\psi := \bigwedge_j \psi_j$$

as a Slater determinant of the $N/2$ one-particle wave functions ψ_j. This ensures that P satisfies Assumption 4.2.3 (i), as the density of states (which is the same as (4.36)) is regular in such intervals. Noting that $[P, H_0] = 0$, we obtain

$$\langle A\rangle_{P_0(t)} = \langle A\rangle_P = \langle A\rangle_{P_{\mathrm{pre}}}, \quad \text{for all} \quad t \geq 0,$$

for every observable A satisfying Assumption 4.2.3. Hence, $P_{\mathrm{pre}} = P$, similar to the first example. In particular, for $A = \mathbf{1}_{\mathcal{H}_{N/2}^{\perp}}$ being the identity on the orthogonal complement of the $N/2$-particle sector of the Fock space, the prethermal value is given by $\langle A\rangle_{P_{\mathrm{pre}}} = 0$. Moreover, by spectral decomposition of H_0, similarly to (4.37), we find that $\langle A\rangle_{\widetilde{P}_\lambda} \neq \langle A\rangle_{P_{\mathrm{pre}}}$. Hence, we deduce that (H_0, P) has the prethermalization property from Definition 4.2.7.

4.3 Proofs

In this section, we provide the proofs of our main results formulated in Section 4.2. We begin by giving the proof of Theorem 4.2.4, which we organize in three steps:

(i) In Section 4.3.1, as the first step, we approximate the random time evolution $\langle A\rangle_{P_\lambda(t)}$ by a deterministic object, up to an error vanishing as $N \to \infty$ with very high probability. This is done using a suitable *global law* for two resolvents of the random matrix H_λ (see Proposition 4.3.1 below).

(ii) The deterministic object resulting from Step (i) consists of two terms, a regular and a singular one. In Section 4.3.2, we evaluate these terms up to errors captured by $\mathcal{E}$ (see Proposition 4.3.2). This proves Theorem 4.2.4 (a).

(iii) As the third and final step in Section 4.3.3, we examine the behavior of the singular term in the limits $T \to 0$ and $T \to \infty$ for $T := \lambda^2 t$ (see Proposition 4.3.6). This proves Theorem 4.2.4 (b).

Afterwards, we give the proofs of Corollary 4.2.6 and Theorem 4.2.9 in Sections 4.3.4 and 4.3.5, respectively. The proof of Corollary 4.2.8 is immediate from Definition 4.2.7 and Theorem 4.2.4 and hence omitted.

4.3.1 Step (i): Global Law

Ignoring λ for a minute, let $H := D + W$ such that $D \in \mathbb{C}^{N\times N}$ is a self-adjoint deterministic matrix with $\|D\| \lesssim 1$ and W is a Wigner matrix satisfying Assumption 4.2.1. We refer to H as a *deformed Wigner matrix*. It is well known [4, 69, 6, 7], that the resolvent $G(z) := (H - z)^{-1}$ of H at spectral parameter $z \in \mathbb{C} \setminus \mathbb{R}$ is very well approximated by a deterministic matrix M, which is the unique solution to the *Matrix Dyson Equation (MDE)*[10]

$$-\frac{1}{M(z)} = z - D + \langle\!\langle M(z)\rangle\!\rangle, \quad \text{with} \quad \Im M(z)\Im z > 0. \tag{4.39}$$

[10]The MDE in the context of mean-field random matrices was introduced in [4] and extensively analyzed in [6].

In particular (see [69, Theorem 2.1]), for $|\Im z| \gtrsim 1$ and arbitrary deterministic matrix $B \in \mathbb{C}^{N\times N}$ with $\|B\| \lesssim 1$, it holds that

$$\Big|\langle\!\langle (G(z) - M(z))B \rangle\!\rangle\Big| \prec \frac{1}{N}. \tag{4.40}$$

For our purposes, it is not sufficient to approximate only a single resolvent in the sense of (4.40). Instead, we need to establish the deterministic approximation to $\langle \boldsymbol{x}, G(z_1)BG(z_2)\boldsymbol{y}\rangle$ with two deterministic vectors $\boldsymbol{x}, \boldsymbol{y}$. This is the content of the following proposition, the proof of which is given in Appendix 4.A.2.

Proposition 4.3.1 (Isotropic two-resolvent global law for deformed Wigner matrices)**.** *Let $H := D + W$ be an $N \times N$ deformed Wigner matrix (as in Assumption 4.2.1) with a bounded self-adjoint deformation $D \in \mathbb{C}^{N\times N}$. Pick $B \in \mathbb{C}^{N\times N}$ a deterministic matrix with $\|B\| \lesssim 1$, deterministic vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^N$ with $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1$ as well as $z_1, z_2 \in \mathbb{C}$ with $\min\{|\Im z_1|, |\Im z_2|\} \geq 1$.[11] Denote further $G_i = G(z_i) = (H - z_i)^{-1}$. Then,*

$$\Big|\langle \boldsymbol{x}, G_1BG_2\boldsymbol{y}\rangle - \Big\langle \boldsymbol{x}, \Big(M_1BM_2 + \frac{M_1M_2\langle\!\langle M_1BM_2\rangle\!\rangle}{1 - \langle\!\langle M_1M_2\rangle\!\rangle}\Big)\boldsymbol{y}\Big\rangle\Big| \prec \frac{1}{\sqrt{N}}, \tag{4.41}$$

where we denoted $M_i = M(z_i)$ with $M(z) \in \mathbb{C}^{N\times N}$ being the solution of (4.39)*.*

We now apply Proposition 4.3.1 to resolvents $G_\lambda(z) := (H_\lambda - z)^{-1}$ of our concrete deformed Wigner random matrix $H_\lambda = H_0 + \lambda W$. First, the scaling by λ leads to the following version of the MDE (4.39):

$$-\frac{1}{M_\lambda(z)} = z - H_0 + \lambda^2\langle\!\langle M_\lambda(z)\rangle\!\rangle, \quad \Im M_\lambda(z)\Im z > 0. \tag{4.42}$$

Next, inspecting the proof of Proposition 4.3.1 in Appendix 4.A.2, we find that (4.41) becomes

$$\begin{aligned}&\Big|\Big(G_\lambda(z_1)BG_\lambda(z_2)\Big)_{\boldsymbol{xy}} \\ &\quad - \Big(M_\lambda(z_1)BM_\lambda(z_2) + \lambda^2\frac{M_\lambda(z_1)M_\lambda(z_2)\langle\!\langle M_\lambda(z_1)BM_\lambda(z_2)\rangle\!\rangle}{1 - \lambda^2\langle\!\langle M_\lambda(z_1)M_\lambda(z_2)\rangle\!\rangle}\Big)_{\boldsymbol{xy}}\Big| \prec \frac{C(\lambda, c)}{\sqrt{N}},\end{aligned} \tag{4.43}$$

for $z_1, z_2 \in \mathbb{C}\setminus\mathbb{R}$ with $\min\{|\Im z_1|, |\Im z_2|\} \geq c$ for some $c > 0$. The positive constant $C(\lambda, c)$ in (4.43) depends only on its arguments λ and c.[12]

For the proof of Theorem 4.2.4, we now employ (4.43) as follows: Applying residue calculus allows to rewrite $\langle A\rangle_{P_\lambda(t)}$ as the contour integral

$$\langle A\rangle_{P_\lambda(t)} = \langle \mathrm{e}^{-\mathrm{i}tH_\lambda}P\mathrm{e}^{\mathrm{i}tH_\lambda}A\rangle = \frac{1}{(2\pi\mathrm{i})^2}\oint_{\gamma_1}\oint_{\gamma_2}\mathrm{e}^{\mathrm{i}t(z_1-z_2)}\operatorname{Tr}\Big[G_\lambda(z_1)AG_\lambda(z_2)P\Big]\mathrm{d}z_1\mathrm{d}z_2 \tag{4.44}$$

where γ_1 and γ_2 are two semicircles (each being the complex conjugate of the other) with some large radius $R \gtrsim 1$ (see Fig. 4.3 below). We further define the contours such that

[11] As used also in (4.43) below, the constant one can be replaced by any other N-independent $c > 0$.

[12] In fact, by examining the proof of Proposition 4.3.1, it can easily be seen that the dependence on both small parameters λ and c is at most (inverse) polynomial.

the distance between the flat part of the semicircles and the real line is t^{-1}. Note that we have

$$\sigma(H_\lambda) \subseteq \sigma(H_0) + [-(2+\epsilon)\lambda, (2+\epsilon)\lambda] \quad \text{w.v.h.p.} \tag{4.45}$$

for any fixed $\epsilon > 0$ by standard perturbation theory, using that $\|W\| \leq 2 + \epsilon$ with very high probability (see, e.g., [62, Theorem 7.6]). In particular, the contours encircle the spectrum of H_λ completely if R is chosen large enough.

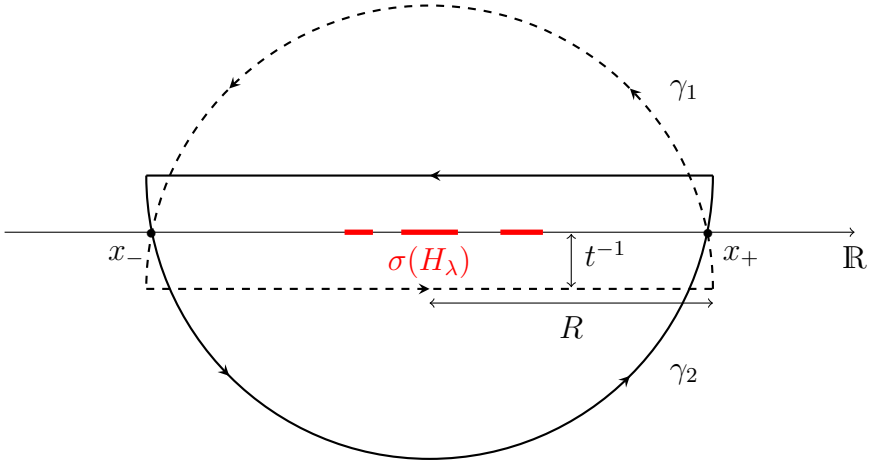

Fig. 4.3. Sketch of the contours γ_1 and γ_2 from (4.44). The intersections of γ_1 (and γ_2) with the real line are denoted by $x_\pm$.

Writing $P = \sum_j p_j \, |\boldsymbol{p}_j\rangle \langle \boldsymbol{p}_j|$ in spectral decomposition and using that the $p_j \in [0,1]$ sum to one by Assumption 4.2.3 (i), the global law (4.43) applied to $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{p}_j$ implies[13]

$$\begin{aligned} &\Big| \mathrm{Tr}[G_\lambda(z_1) A G_\lambda(z_2) P] \\ &\quad - \mathrm{Tr}\left[M_\lambda(z_1) A M_\lambda(z_2) P + \lambda^2 \frac{M_\lambda(z_1) M_\lambda(z_2) P \langle\!\langle M_\lambda(z_1) A M_\lambda(z_2) \rangle\!\rangle}{1 - \lambda^2 \langle\!\langle M_\lambda(z_1) M_\lambda(z_2) \rangle\!\rangle} \right] \Big| \prec \frac{C(\lambda,t)}{\sqrt{N}} \end{aligned} \tag{4.46}$$

uniformly for z_1, z_2 along the contours γ_1, γ_2 for any fixed $\lambda > 0$. Just as in (4.43), $C(\lambda, t)$ denotes a positive constant depending only on λ and t. Therefore, combining (4.44) with (4.46), we find that

$$\begin{aligned} \langle A \rangle_{P_\lambda(t)} = {} & \frac{1}{(2\pi \mathrm{i})^2} \oint_{\gamma_1} \oint_{\gamma_2} \mathrm{e}^{\mathrm{i}t(z_1 - z_2)} \, \mathrm{Tr}\left[M_\lambda(z_1) A M_\lambda(z_2) P \right] \mathrm{d}z_1 \mathrm{d}z_2 \\ & + \frac{1}{(2\pi \mathrm{i})^2} \oint_{\gamma_1} \oint_{\gamma_2} \mathrm{e}^{\mathrm{i}t(z_1 - z_2)} \lambda^2 \frac{\langle\!\langle M_\lambda(z_1) A M_\lambda(z_2) \rangle\!\rangle \, \mathrm{Tr}\left[M_\lambda(z_1) P M_\lambda(z_2) \right]}{1 - \lambda^2 \langle\!\langle M_\lambda(z_1) M_\lambda(z_2) \rangle\!\rangle} \mathrm{d}z_1 \mathrm{d}z_2 \\ & + \mathcal{O}_\prec\left(\frac{C(\lambda,t)}{\sqrt{N}} \right). \end{aligned} \tag{4.47}$$

To establish (4.13), our main task thus lies in evaluating the right-hand side of (4.47). For simplicity, we refer to the integrals in the first and second line of (4.47) as the *regular* and the *singular* term, respectively.

[13] Note that the true bound in the averaged law is of order N^{-1}, i.e., by a factor $N^{-1/2}$ better than (4.46). However, for the following applications, our weaker bound is sufficient.

4.3.2 Step (ii): Evaluation of the Regular and Singular Term and Proof of Theorem 4.2.4 (a)

We organize the result of our computation of (4.47) in the following proposition.

Proposition 4.3.2 (Evaluation of the regular and singular term)**.** *Under the assumptions of Theorem 4.2.4 and letting γ_1, γ_2 be the contours in Fig. 4.3, we have (recalling $\alpha = \pi\rho_0(E_0)$)*

$$\begin{aligned}\frac{1}{(2\pi\mathrm{i})^2}\oint_{\gamma_1}\oint_{\gamma_2}\mathrm{e}^{\mathrm{i}t(z_1-z_2)}\operatorname{Tr}\left[M_\lambda(z_1)AM_\lambda(z_2)P\right]\mathrm{d}z_1\mathrm{d}z_2\\ = \mathrm{e}^{-2\alpha\lambda^2 t}\langle A\rangle_{P_0(t)}+\mathcal{O}(\mathcal{E}_{\mathrm{reg}}),\end{aligned}\tag{4.48}$$

for the regular term and

$$\begin{aligned}\widetilde{\mathcal{F}}_{\mathrm{sing}} := \frac{\lambda^2}{(2\pi\mathrm{i})^2}\oint_{\gamma_1}\oint_{\gamma_2}\mathrm{e}^{\mathrm{i}t(z_1-z_2)}\frac{\left\langle\!\left\langle M_\lambda(z_1)AM_\lambda(z_2)\right\rangle\!\right\rangle\operatorname{Tr}\left[M_\lambda(z_1)PM_\lambda(z_2)\right]}{1-\lambda^2\left\langle\!\left\langle M_\lambda(z_1)M_\lambda(z_2)\right\rangle\!\right\rangle}\mathrm{d}z_1\mathrm{d}z_2\\ = \langle A\rangle_{\widetilde{P}_{\lambda,t}}+\mathcal{O}(\mathcal{E}_{\mathrm{sing}})\end{aligned}\tag{4.49}$$

for the singular term, with some error terms $\mathcal{E}_{\mathrm{reg/sing}} = \mathcal{E}_{\mathrm{reg/sing}}(\lambda, t, \Delta, N)$ in (4.48) *and* (4.49) *satisfying* (4.17)*. The explicit form of $\mathcal{E}_{\mathrm{reg}}$ is given in* (4.57) *and* (4.59)*, while the explicit form of $\mathcal{E}_{\mathrm{sing}}$ is given in* (4.67)*.*

Plugging (4.48) and (4.49) into (4.47), we immediately conclude Theorem 4.2.4 (a) after setting $\mathcal{E}_0 := \mathcal{E}_{\mathrm{reg}} + \mathcal{E}_{\mathrm{sing}}$ and including the error term from (4.47) into $\mathcal{E}$. □

It thus remains to give the proof of Proposition 4.3.2, i.e. its two parts (4.48) and (4.49). This is done in Sections 4.3.2 and 4.3.2, respectively.

Proof of (4.48)

The main contribution to the integral in (4.48) comes from the regime[14] where z_1 and z_2 are close to E_0. Hence, as a first approximation we use the replacements $\langle\!\langle M_\lambda(z_1)\rangle\!\rangle \approx \overline{m_0(E_0)}$ and $\langle\!\langle M_\lambda(z_2)\rangle\!\rangle \approx m_0(E_0)$ in (4.42), which leads to

$$M_\lambda(z_1) \approx \frac{1}{H_0 - z_1 - \lambda^2\overline{m_0(E_0)}} \qquad \text{and} \qquad M_\lambda(z_2) \approx \frac{1}{H_0 - z_2 - \lambda^2 m_0(E_0)}.\tag{4.50}$$

Applying the replacements in (4.50) for the term in (4.48) yields

$$\begin{aligned}&\frac{1}{(2\pi\mathrm{i})^2}\oint_{\gamma_1}\oint_{\gamma_2}\mathrm{e}^{\mathrm{i}t(z_1-z_2)}\operatorname{Tr}\left[\frac{1}{H_0 - z_1 - \lambda^2\overline{m_0(E_0)}}A\frac{1}{H_0 - z_2 - \lambda^2 m_0(E_0)}P\right]\mathrm{d}z_1\mathrm{d}z_2\\ &= \operatorname{Tr}\left[\mathrm{e}^{\mathrm{i}t(H_0-\lambda^2\overline{m_0(E_0)})}A\mathrm{e}^{-\mathrm{i}t(H_0-\lambda^2 m_0(E_0))}P\right]\\ &= \mathrm{e}^{-2\Im m_0(E_0)\lambda^2 t}\operatorname{Tr}\left[\mathrm{e}^{\mathrm{i}tH_0}A\mathrm{e}^{-\mathrm{i}tH_0}P\right] = \mathrm{e}^{-2\alpha\lambda^2 t}\,\langle A\rangle_{P_0(t)},\end{aligned}\tag{4.51}$$

since $\Im m_0(E_0) = \pi\rho_0(E_0)$, from simple residue calculus for $\lambda > 0$ small enough, using that $|m_0(E_0)| \lesssim 1$ and γ_1, γ_2 encircle the spectrum of H_0. We have thus extracted the

[14] Recall that the flat pieces of the contours γ_1 and γ_2 from Fig. 4.3 lie on the lower and upper half plane, respectively.

main term in (4.48), and it remains to estimate the errors resulting from the replacements in (4.50).

Recall (see (4.21)) that μ_j and $\boldsymbol{u}_j$ are the eigenvalues and the respective orthonormalized eigenvectors of H_0, i.e.

$$H_0 = \sum_j \mu_j \left|\boldsymbol{u}_j\right\rangle \left\langle \boldsymbol{u}_j\right| . \tag{4.52}$$

Then, by means of (4.42), spectral decomposition (4.52) of H_0 and using Assumption 4.2.3 (i) together with $[H_0, \Pi_\Delta] = 0$ and $\Pi_\Delta^2 = \Pi_\Delta$, we have that

$$\text{lhs. of (4.48)} = \operatorname{Tr}\left[\widetilde{\Theta}_1 A \widetilde{\Theta}_2 P\right] = \sum_{\mu_i,\mu_j \in I_\Delta} \langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle\langle \boldsymbol{u}_j, P\boldsymbol{u}_i\rangle \widetilde{\vartheta}_1(i)\widetilde{\vartheta}_2(j), \tag{4.53}$$

where we denoted

$$\widetilde{\Theta}_1 := \sum_{\mu_j \in I_\Delta} \left|\boldsymbol{u}_j\right\rangle \left\langle \boldsymbol{u}_j\right| \widetilde{\vartheta}_1(j) \quad \text{and} \quad \widetilde{\Theta}_2 := \sum_{\mu_j \in I_\Delta} \left|\boldsymbol{u}_j\right\rangle \left\langle \boldsymbol{u}_j\right| \widetilde{\vartheta}_2(j) \tag{4.54}$$

with

$$\begin{aligned} \widetilde{\vartheta}_1(j) &:= \frac{1}{2\pi \mathrm{i}} \oint_{\gamma_1} \frac{\mathrm{e}^{\mathrm{i}tz_1}}{\mu_j - z_1 - \lambda^2 \langle\!\langle M_\lambda(z_1)\rangle\!\rangle} \mathrm{d}z_1, \\ \widetilde{\vartheta}_2(j) &:= \frac{1}{2\pi \mathrm{i}} \oint_{\gamma_2} \frac{\mathrm{e}^{-\mathrm{i}tz_2}}{\mu_j - z_2 - \lambda^2 \langle\!\langle M_\lambda(z_2)\rangle\!\rangle} \mathrm{d}z_2. \end{aligned} \tag{4.55}$$

Note that, by symmetry of the contours γ_1 and γ_2, we have $\overline{\widetilde{\vartheta}_1(j)} = \widetilde{\vartheta}_2(j)$ and $\widetilde{\Theta}_1^* = \widetilde{\Theta}_2$.

The key to approximating (4.53) is the following lemma, whose proof is given at the end of the current Section 4.3.2.

Lemma 4.3.3 (First replacement lemma)**.** *Using the above notations and assumption, denote*

$$\Theta_1 := \sum_{\mu_j \in I_\Delta} \left|\boldsymbol{u}_j\right\rangle \left\langle \boldsymbol{u}_j\right| \vartheta_1(j) \quad \textit{with} \quad \vartheta_1(j) := \frac{1}{2\pi \mathrm{i}} \oint_{\gamma_1} \frac{\mathrm{e}^{\mathrm{i}tz_1}}{\mu_j - z_1 - \lambda^2 \overline{m_0(E_0)}} \mathrm{d}z_1$$

and $\Theta_2 := \Theta_1^*$ *via* $\vartheta_2(j) := \overline{\vartheta_1(j)}$*, analogously to* (4.54) *and* (4.55)*. Then it holds that*

$$\sup_{\mu_i \in I_\Delta} \left|\widetilde{\vartheta}_1(i) - \vartheta_1(i)\right| + \sup_{\mu_j \in I_\Delta} \left|\widetilde{\vartheta}_2(j) - \vartheta_2(j)\right| \lesssim \widetilde{\mathcal{E}}_{\mathrm{reg}} \tag{4.56}$$

for sufficiently small $\lambda > 0$ *and* N *large enough (dependent on* λ*, cf. Lemma 4.A.1). Here, recalling* (4.8) *for the definition of* ϵ_0*, we denoted*

$$\widetilde{\mathcal{E}}_{\mathrm{reg}} := \lambda^2 t\, \Delta + \lambda\left(1 + \lambda^2 t\right) + \frac{\lambda}{\Delta}\left(1 + \frac{\lambda}{\Delta}\right) + \lambda^2 t\, \epsilon_0. \tag{4.57}$$

Therefore, by writing $\widetilde{\Theta} = \Theta + (\widetilde{\Theta} - \Theta)$ in (4.53), we find the lhs. of (4.48) to be given by

$$\operatorname{Tr}\left[\Theta_1 A \Theta_2 P\right] + \operatorname{Tr}\left[(\widetilde{\Theta}_1 - \Theta_1) A \Theta_2 P\right] + \operatorname{Tr}\left[\Theta_1 A (\widetilde{\Theta}_2 - \Theta_2) P\right] + \operatorname{Tr}\left[(\widetilde{\Theta}_1 - \Theta_1) A (\widetilde{\Theta}_2 - \Theta_2) P\right]. \tag{4.58}$$

The first term in (4.58) precisely yields the result of (4.51) using Assumption 4.2.3 (i). Using $\|A\| \lesssim 1$ and $\operatorname{Tr}[P] = 1$, the second and third term in (4.58) can be estimated by (a constant times)

$$\|\Theta_1\|\, \|\widetilde{\Theta}_2 - \Theta_2\| + \|\widetilde{\Theta}_1 - \Theta_1\|\, \|\Theta_2\| \lesssim \widetilde{\mathcal{E}}_{\mathrm{reg}}.$$

Here we used (4.56) and that $\|\Theta_1\| \leq 1$ and $\|\Theta_2\| \leq 1$ as follows by the explicit expressions

$$\Theta_1 = \mathrm{e}^{\mathrm{i}t(\Pi_\Delta H_0 \Pi_\Delta - \lambda^2 \overline{m_0(E_0)})} \quad \text{and} \quad \Theta_2 = \mathrm{e}^{-\mathrm{i}t(\Pi_\Delta H_0 \Pi_\Delta - \lambda^2 m_0(E_0))}$$

and $\Im m_0(E_0) \geq 0$. Similarly, applying (4.56) again, the fourth term in (4.58) is bounded by $\mathcal{O}(\widetilde{\mathcal{E}}_{\mathrm{reg}}^2)$. Collecting all four terms of (4.57), this concludes the proof of (4.48) with

$$\mathcal{E}_{\mathrm{reg}} := \widetilde{\mathcal{E}}_{\mathrm{reg}} + \widetilde{\mathcal{E}}_{\mathrm{reg}}^2. \tag{4.59}$$

It remains to give the proof of Lemma 4.3.3.

Proof of Lemma 4.3.3. As $\overline{\widetilde{\vartheta}_1(j)} = \widetilde{\vartheta}_2(j)$ and $\vartheta_2(j) := \overline{\vartheta_1(j)}$, we only estimate $\widetilde{\vartheta}_2(j) - \vartheta_2(j)$ for arbitrary but fixed index j such that $\mu_j \in I_\Delta$. Moreover, for ease of notation, we completely drop the subscript 2.

As a first step, we split the contour into three parts:

$$\gamma = \Gamma_1 \dot{+} \Gamma_2 \dot{+} \Gamma_3, \tag{4.60}$$

where Γ_1 is the horizontal part of γ with $\Re z \in I_{2\Delta}$, Γ_2 is the horizontal part of γ with $\Re z \notin I_{2\Delta}$ and Γ_3 consists of the great arc of radius R (cf. Fig. 4.3). We now estimate these three parts separately.

For the first part, Γ_1, we have (using the notation $m_\lambda(z) = \langle\!\langle M_\lambda(z) \rangle\!\rangle$ from Lemma 4.A.1)[15]

$$\begin{aligned} &\left| \int_{\Gamma_1} \mathrm{e}^{-\mathrm{i}tz} \left[\frac{1}{\mu_j - z - \lambda^2 m_\lambda(z)} - \frac{1}{\mu_j - z - \lambda^2 m_0(E_0)} \right] \mathrm{d}z \right| \\ &\quad \lesssim \int_{\Gamma_1} \frac{\lambda^2 \left(1/t + \lambda + \Delta + \epsilon_0\right)}{|\mu_j - z - \lambda^2 m_\lambda(z)|\,|\mu_j - z - \lambda^2 m_0(E_0)|} |\mathrm{d}z| \lesssim \lambda^2 t \left(1/t + \lambda + \Delta + \epsilon_0\right), \end{aligned} \tag{4.61}$$

uniformly in $\mu_j \in I_\Delta$. To go to the second line, we used that

$$|m_\lambda(z) - m_0(E_0)| \lesssim 1/t + \lambda + \Delta + \epsilon_0.$$

This follows by adding and subtracting $m_\lambda(E_0)$ and using $|m_\lambda(z) - m_\lambda(E_0)| \lesssim \Delta + 1/t$ (using $|m_\lambda'(z)| \lesssim 1$ for $\Re z \in I_{2\Delta}$; cf. the last estimate in (4.129) from Lemma 4.A.1) and $|m_\lambda(E_0) - m_0(E_0)| \lesssim \lambda + \epsilon_0$ (using that (4.133) holds down to the real line by combining it with (4.134)). For the final bound, we employed a Schwarz inequality for the integral and estimated the resulting integrals

$$\int_{\Gamma_1} \frac{|\mathrm{d}z|}{|\mu_j - z - \lambda^2 m_\lambda(z)|^2} \lesssim (1 + \lambda^2)\, t \lesssim t \quad \text{and} \quad \int_{\Gamma_1} \frac{|\mathrm{d}z|}{|\mu_j - z - \lambda^2 m_0(E_0)|^2} \lesssim t,$$

by a change of variables $z \to z + \lambda^2 m_\lambda(z)$ using that $|m_\lambda'(z)| \lesssim 1$ for $z \in \Gamma_1$ by means of (4.129) together with $|\Im[z + \lambda^2 m_\lambda(z)]| \geq t^{-1}$, and $|m_0(E_0)| \lesssim 1$ together with $|\Im[z + \lambda^2 m_0(E_0)]| \geq t^{-1}$, respectively.

We now turn to the second part, i.e. the integral similar to the left-hand side of (4.61) but on the contour Γ_2. By means of $|m_0(E_0)| \lesssim 1$ and $|m_\lambda(z)| \leq \lambda^{-1}$ (see the first estimate in (4.128)) we bound $|m_\lambda(z) - m_0(E_0)| \lesssim \lambda^{-1}$. Using $|m_0(E_0)| \lesssim 1$ and $|m_\lambda(z)| \leq \lambda^{-1}$

[15] Here and in the following, $|\mathrm{d}z|$ denotes the total variation of the complex measure $\mathrm{d}z$.

again, together with $\operatorname{dist}(\mu_j, \Gamma_2) \gtrsim \Delta$, we find this second part to be bounded by (a constant times)

$$\int_{\Gamma_2} \frac{\lambda}{|\mu_j - z|^2} |\mathrm{d}z| \left(1 + \frac{\lambda}{\Delta}\right) \lesssim \frac{\lambda}{\Delta} \left(1 + \frac{\lambda}{\Delta}\right), \tag{4.62}$$

again uniformly in $\mu_j \in I_\Delta$.

Finally, we estimate the third part. By the exact same reasoning as for Γ_2, we arrive at the bound (4.62) with Δ replaced by R and Γ_2 replaced by Γ_3. Hence, using that the radius R of the semicircle is larger than one (see Fig. 4.3), we find the third part to be bounded by $\mathcal{O}(\lambda/R)$, uniformly in $\mu_j \in I_\Delta$. Combining this with the error terms in (4.61) and (4.62), this concludes the proof. □

Proof of (4.49)

Recall that $\widetilde{\mathcal{F}}_{\mathrm{sing}}$ denotes the singular term defined in (4.49). To carry out the analog of the approximation (4.50), we observe that the resolvent identity for H_0 implies

$$\frac{z_{1,\lambda} - z_{2,\lambda}}{z_1 - z_2} M_0(z_{1,\lambda}) M_0(z_{2,\lambda}) = \frac{M_0(z_{1,\lambda}) - M_0(z_{2,\lambda})}{z_1 - z_2} = \frac{1}{\pi} \int_{\mathbb{R}} \frac{\Im M_0(x_\lambda)}{(x - z_1)(x - z_2)} \mathrm{d}x, \tag{4.63}$$

where we introduce the notation $z_{1,\lambda} := z_1 + \lambda^2 \overline{m_0(E_0)}$, $z_{2,\lambda} := z_2 + \lambda^2 m_0(E_0)$, and $x_\lambda := x + \lambda^2 m_0(E_0)$. Here, the second equality follows from the contour representation of the resolvent M_0 of H_0, namely

$$M_0(z) = \frac{1}{\pi} \int_{\mathbb{R}} \frac{\Im M_0(x + \mathrm{i}\eta)}{x + \mathrm{i}\eta - z} \mathrm{d}x, \quad \Im z > \eta > 0. \tag{4.64}$$

On the other hand, subtracting two instances of the MDE (4.42) yields

$$\frac{M_\lambda(z_1) M_\lambda(z_2)}{1 - \lambda^2 \langle\!\langle M_\lambda(z_1) M_\lambda(z_2) \rangle\!\rangle} = \frac{M_\lambda(z_1) - M_\lambda(z_2)}{z_1 - z_2} = \frac{1}{\pi} \int_{\mathbb{R}} \frac{\Im M_\lambda(x)}{(x - z_1)(x - z_2)} \mathrm{d}x, \tag{4.65}$$

where we used the matrix-valued analog of the Stieltjes representation for $M_\lambda(z)$ (cf. [6, Prop. 2.1]),

$$M_\lambda(z) = \frac{1}{\pi} \int_{\mathbb{R}} \frac{\Im M_\lambda(x)}{x - z} \mathrm{d}x, \quad z \in \mathbb{C} \setminus \mathbb{R}. \tag{4.66}$$

In particular, identities (4.63) and (4.65) suggest that the appropriate approximation for the factor $1 - \lambda^2 \langle\!\langle M_\lambda(z_1) M_\lambda(z_2) \rangle\!\rangle$ in the denominator of (4.49) is $(z_1 - z_2)(z_{1,\lambda} - z_{2,\lambda})^{-1}$. Indeed, we prove that the following estimate holds.

Lemma 4.3.4. *Under the Assumption 4.2.2 and 4.2.3, the singular term $\widetilde{\mathcal{F}}_{\mathrm{sing}}$ defined in* (4.49) *satisfies*

$$\left| \widetilde{\mathcal{F}}_{\mathrm{sing}} - \mathcal{F}_{\mathrm{sing}} \right| \lesssim \mathcal{E}_{\mathrm{sing}} := (\Delta + \epsilon_0)(1 + \lambda^2 t) + \lambda \Big(1 + \lambda^2 t + \Delta^{-1} \log t\Big)^2, \tag{4.67}$$

where the quantity $\mathcal{F}_{\mathrm{sing}}$ is given by

$$\mathcal{F}_{\mathrm{sing}} := \frac{\lambda^2}{(2\pi\mathrm{i})^2} \oint_{\gamma_1} \oint_{\gamma_2} \mathrm{e}^{\mathrm{i}t(z_1 - z_2)} \frac{z_{1,\lambda} - z_{2,\lambda}}{z_1 - z_2} \times \Big\langle\!\Big\langle M_0(z_{1,\lambda}) A M_{0(2,\lambda)} \Big\rangle\!\Big\rangle \operatorname{Tr}\Big[M_0(z_{1,\lambda}) P M_0(z_{2,\lambda})\Big] \mathrm{d}z_1 \mathrm{d}z_2. \tag{4.68}$$

We defer the proof of Lemma 4.3.4 to the end of the current Section 4.3.2, and proceed to analyze the right-hand side of (4.68).

Applying the identity (4.63) to both traces in the integrand of (4.68), we obtain the expression

$$\mathcal{F}_{\text{sing}} = \int_{\mathbb{R}} \int_{\mathbb{R}} \Big\langle\!\Big\langle \Im M_0(x_\lambda) A \Big\rangle\!\Big\rangle \operatorname{Tr}[\Im M_0(y_\lambda) P] F_{\lambda,t}(x,y) \mathrm{d}x \mathrm{d}y, \tag{4.69}$$

where the function $F_{\lambda,t}(x,y)$ is defined as

$$F_{\lambda,t}(x,y) := \frac{\lambda^2}{(2\pi \mathrm{i})^2 \pi^2} \oint_{\gamma_1} \oint_{\gamma_2} \frac{\mathrm{e}^{\mathrm{i}t(z_1-z_2)}\,(z_1-z_2)}{(x-z_1)(x-z_2)(y-z_1)(y-z_2)(z_{1,\lambda}-z_{2,\lambda})} \mathrm{d}z_2 \mathrm{d}z_1. \tag{4.70}$$

Recall the contours γ_1 and γ_2 from Fig. 4.3, and that $z_{1,\lambda} - z_{2,\lambda} = z_1 - z_2 - 2\mathrm{i}\alpha\lambda^2$. Evaluating the contour integration over γ_2 in (4.70) yields

$$\begin{aligned} F_{\lambda,t}(x,y) = \frac{\lambda^2}{\pi^2} \frac{1}{2\pi \mathrm{i}} \oint_{\gamma_1} \Bigg(& \frac{-\mathrm{e}^{\mathrm{i}t(z_1-x)} \chi(x)}{(x-y)(y-z_1)(x_\lambda - z_{1,\lambda})} + \frac{\mathrm{e}^{\mathrm{i}t(z_1-y)} \chi(y)}{(x-y)(x-z_1)(y_\lambda - z_{1,\lambda})} \\ & + \frac{2\mathrm{i}\alpha\lambda^2 \mathrm{e}^{-2\alpha\lambda^2 t}}{(x-z_1)(y-z_1)(x_\lambda - z_{1,\lambda})(y_\lambda - z_{1,\lambda})} \Bigg) \mathrm{d}z_1 - F^{\text{out}}_{\lambda,t}(x,y), \end{aligned} \tag{4.71}$$

where we define $\chi(z) := \mathbb{1}_{\Omega_2}(z)$, and Ω_2 is the compact connected component of $\mathbb{C} \backslash \gamma_2$, and the function $F^{\text{out}}_{\lambda,t}(x,y)$ is defined as

$$F^{\text{out}}_{\lambda,t}(x,y) := \frac{\lambda^2}{\pi^2} \frac{1}{2\pi \mathrm{i}} \oint_{\gamma_1} \frac{2\mathrm{i}\alpha\lambda^2 \mathrm{e}^{-2\alpha\lambda^2 t} \big(1 - \chi(z_1 - 2\mathrm{i}\alpha\lambda^2)\big)}{(x-z_1)(y-z_1)(x_\lambda - z_{1,\lambda})(y_\lambda - z_{1,\lambda})} \mathrm{d}z_1. \tag{4.72}$$

We proceed to show that $F^{\text{out}}_{\lambda,t}(x,y)$ contributes at most an $\mathcal{O}(\lambda^4(1 + |\log\lambda|^2 + (\lambda^2 t)^2))$ error to the right-hand side of (4.69). Let $\gamma_{1,c}$ denote the set $\{z_1 \in \gamma_1 : z_1 - 2\mathrm{i}\alpha\lambda^2 \notin \Omega_2\}$, then the contribution of $F^{\text{out}}_{\lambda,t}(x,y)$ to the integral in (4.69) is given by

$$\mathcal{E}_{\gamma_{1,c}} := \frac{\alpha\lambda^4 \mathrm{e}^{-2\alpha\lambda^2 t}}{\pi^3} \int_{\gamma_{1,c}} \int_{\mathbb{R}} \frac{\Big\langle\!\Big\langle \Im M_0(x_\lambda) A \Big\rangle\!\Big\rangle \mathrm{d}x}{(x-z_1)(x_\lambda - z_{1,\lambda})} \int_{\mathbb{R}} \frac{\operatorname{Tr}[\Im M_0(y_\lambda) P] \mathrm{d}y}{(y-z_1)(y_\lambda - z_{1,\lambda})} \mathrm{d}z_1. \tag{4.73}$$

Note that by choosing the radii of the contours $R \gtrsim 1$ large enough, we can assume that

$$\operatorname{dist}(\sigma(H_0), \gamma_{1,c}) \gtrsim R. \tag{4.74}$$

Using the spectral decomposition of H_0 from (4.52), the fact that $\int_{\mathbb{R}} \Im[(\mu_j - x_\lambda)] \mathrm{d}x = \pi$, and Assumption 4.2.3, we conclude that

$$\left| \int_{\mathbb{R}} \Big\langle\!\Big\langle \Im M_0(x_\lambda) A \Big\rangle\!\Big\rangle \mathrm{d}x \right| + \left| \int_{\mathbb{R}} \operatorname{Tr}\Big[\Im M_0(y_\lambda) P\Big] \mathrm{d}y \right| \lesssim 1 + \|A\| \lesssim 1. \tag{4.75}$$

Further, the spectral decomposition for H_0 implies that for all x with $\operatorname{dist}(x, \sigma(H_0)) \gtrsim 1$,

$$\Big|\operatorname{Tr}\Big[\Im M_0(x_\lambda) P\Big]\Big| + \Big|\Big\langle\!\Big\langle \Im M_0(x_\lambda) A \Big\rangle\!\Big\rangle\Big| \lesssim \lambda^2 |x - E_0|^{-2}. \tag{4.76}$$

Let $x_\pm$ denote the intersections of the contour γ_1 with $\mathbb{R}$ (see Fig. 4.3), and define $\mathbb{D}$ to be a union of two small disks around $x_\pm$, $\mathbb{D} := \{z \in \mathbb{C} : \min\{|z - x_-|, |z - x_+|\} \lesssim 1\}$. Applying the Cauchy-Schwarz inequality to the z_1 integration in (4.73), and using the

estimates (4.74)-(4.76) to bound the contribution coming from the outside of $\mathbb{D}$, and applying the estimates (4.74) and (4.76) for all $x, y \in \mathbb{R} \cap \mathbb{D}$, we obtain

$$\left|\mathcal{E}_{\gamma_{1,c}}\right| \lesssim \lambda^4 \int_{\gamma_{1,c}\cap\mathbb{D}} \left| \frac{\lambda^2}{R^2} \int_{\mathbb{R}\cap\mathbb{D}} \frac{\mathrm{d}x}{|x - z_1||x_\lambda - z_{1,\lambda}|} \right|^2 |\mathrm{d}z_1| + \mathcal{O}\Big(R^{-3}\lambda^4\Big). \tag{4.77}$$

For z_1 on the horizontal linear segment of $\gamma_{1,c} \cap \mathbb{D}$, we use that $\Im z_1 = -1/t$ to obtain

$$\frac{\lambda^2}{R^2} \int_{\mathbb{R}\cap\mathbb{D}} \frac{\mathrm{d}x}{|x - z_1||x_\lambda - z_{1,\lambda}|} \lesssim \frac{1 + \lambda^2 t}{R^2}, \tag{4.78}$$

On the other hand, for z_1 lying on the circular arc parts of $\gamma_{1,c} \cap \mathbb{D}$, we compute

$$\frac{\lambda^2}{R^2} \int_{\mathbb{R}\cap\mathbb{D}} \frac{\mathrm{d}x}{|x - z_1||x_\lambda - z_{1,\lambda}|} \lesssim \frac{\lambda^2}{R^2} \frac{|\log \lambda^2| + \Big|\log|\eta_1|\Big| + \Big|\log|\eta_1 - 2\alpha\lambda^2|\Big|}{|\eta_1| + \lambda^2}, \tag{4.79}$$

where $\eta_1 := \Im z_1$. Squaring the estimates (4.78) and (4.79) and integrating them over the respective parts of $\gamma_{1,c} \cap \mathbb{D}$, we conclude from (4.77) that

$$\left|\mathcal{E}_{\gamma_{1,c}}\right| \lesssim R^{-3}\lambda^4\Big(1 + (\lambda^2 t)^2 + |\log \lambda|^2\Big). \tag{4.80}$$

Next, using residue calculus, we compute the first term on the right-hand side of (4.71), i.e., the contour integral over γ_1, to obtain

$$\begin{aligned} F_{\lambda,t}(x,y) = \frac{K_{\lambda,t}(x-y)}{\alpha\pi} - F^{\mathrm{out}}_{\lambda,t}(x,y) + \frac{K_{\lambda,t}(x-y)}{\alpha\pi}\Big(\chi(x)\chi(y) - 1\Big) \\ + \frac{1}{2\pi^2} \frac{2\lambda^2 \mathrm{e}^{-2\alpha\lambda^2 t}\Big(\chi(x) - \chi(y)\Big)^2}{|x-y|^2 + (2\alpha\lambda^2)^2} + \frac{1}{\pi^2} \frac{2\mathrm{i}\alpha\lambda^4 \mathrm{e}^{-2\alpha\lambda^2 t}}{|x-y|^2 + (2\alpha\lambda^2)^2} \frac{\chi(x) - \chi(y)}{x - y}, \end{aligned} \tag{4.81}$$

where $K_{\lambda,t}$ is the kernel defined in (4.16). As we have proved above, the contribution of $F^{\mathrm{out}}_{\lambda,t}(x,y)$ to the integral in (4.69) admits the bound (4.80). Similarly, using the estimates (4.75) and (4.76), it is straightforward to check that the third and fourth terms on the right-hand side of (4.81) contribute at most $\mathcal{O}(\lambda^4)$ to the right-hand side of (4.69), while the last term contributes at most $\mathcal{O}(\lambda^2)$. Therefore,

$$\mathcal{F}_{\mathrm{sing}} = \langle A \rangle_{\widetilde{P}_{\lambda,t}} \Big(1 + \mathcal{O}(\epsilon_0 + \Delta + \lambda^2/\Delta)\Big) + \mathcal{O}(\lambda^2), \tag{4.82}$$

where we used the estimate $\pi\alpha = \pi^2\rho_0(E_0) = N^{-1}r\Big(1 + \mathcal{O}(\epsilon_0 + \Delta + \lambda^2/\Delta)\Big)$ that follows from Lemma 4.A.2 for r defined in (4.28), and performed a change of variables $x \to x - \lambda^2 \Re m_0(E_0)$ and $y \to y - \lambda^2 \Re m_0(E_0)$. We note that the N^{-1} prefactor results from the different normalization of the trace in (4.28) and (4.140).

Furthermore, Proposition 4.3.6 below implies that under the Assumptions 4.2.2 and 4.2.3, we have $|\langle A \rangle_{\widetilde{P}_{\lambda,t}}| \lesssim 1$. Since the proof of Proposition 4.3.6 is independent of the statement of (4.49), this concludes the proof of (4.49).

We proceed to prove Lemma 4.3.4.

Proof of Lemma 4.3.4. Define the sequences of overlaps $\mathfrak{a}_j := \langle \boldsymbol{u}_j, A\boldsymbol{u}_j \rangle$, and $\mathfrak{p}_k := \langle \boldsymbol{u}_k, P\boldsymbol{u}_k \rangle$ where we recall from (4.52) that $\boldsymbol{u}_j$'s are the eigenvectors of H_0. Observe that the Assumption 4.2.3 implies that

$$\|\mathfrak{a}\|_\infty \lesssim 1, \quad \mathfrak{p}_k \geq 0 \quad \text{and} \quad \|\mathfrak{p}\|_1 = 1. \tag{4.83}$$

Using the spectral decomposition (4.52) of H_0 and the identity (4.65), we rewrite $\widetilde{\mathcal{F}}_{\text{sing}}$ in the following form,

$$\widetilde{\mathcal{F}}_{\text{sing}} = \sum_{j,k} \frac{\mathfrak{a}_j \mathfrak{p}_k}{N} \frac{1}{\pi} \int_{\mathbb{R}} \Im \widetilde{\nu}_j(x) \lambda^2 \big|\widetilde{\omega}_k(x)\big|^2 \mathrm{d}x, \tag{4.84}$$

where $\widetilde{\nu}_j(z) := (\mu_j - z - \lambda^2 m_\lambda(z))^{-1}$ for $z \in \mathbb{C}$, and the functions $\widetilde{\omega}_k(x)$ are defined by the improper integrals[16]

$$\widetilde{\omega}_k(x) := \frac{1}{2\pi \mathrm{i}} \oint_{\gamma_2} \frac{\mathrm{e}^{\mathrm{i}tz} \widetilde{\nu}_k(z)}{x - z} \mathrm{d}z, \quad x \in \mathbb{R}. \tag{4.85}$$

Here we adhere to the convention $m_\lambda(x) := \lim_{\eta \to +0} m_\lambda(x + \mathrm{i}\eta)$.

The key estimates for proving (4.67) are collected in the following Lemma that we prove at the end of the subsection.

Lemma 4.3.5 (Second replacement lemma)**.** *Define the functions*

$$\omega_k(x) := \frac{1}{2\pi \mathrm{i}} \oint_{\gamma_2} \frac{\mathrm{e}^{-\mathrm{i}tz}}{x - z} \frac{1}{\mu_k - z_\lambda} \mathrm{d}z, \quad k \in [N], \tag{4.86}$$

where we denote $z_\lambda := z + \lambda^2 m_0(E_0)$. Then under the Assumptions 4.2.2 and 4.2.3, the estimates

$$\big|\omega_k(x) - \widetilde{\omega}_k(x)\big| \lesssim \frac{(\Delta + \epsilon_0)(1 + \lambda^2 t)}{|\mu_k - x_\lambda|} + 1 + \lambda^2 t + \Delta^{-1} \log t, \quad x \in I_{3\Delta}, \tag{4.87}$$

$$\big|\omega_k(x)\big| + \big|\widetilde{\omega}_k(x)\big| \lesssim \frac{1}{|\mu_k - x_\lambda|} + 1 + \lambda^2 t + \Delta^{-1} \log t, \quad x \in I_{3\Delta}, \tag{4.88}$$

$$\big|\omega_k(x)\big| + \big|\widetilde{\omega}_k(x)\big| \lesssim \Delta^{-1} \log t + R^{-1}, \quad x \in [-\tfrac{1}{2}R, \tfrac{1}{2}R] \backslash I_{3\Delta}, \tag{4.89}$$

hold for all k with $\mu_k \in I_\Delta$.

Observe that applying the identities (4.63), (4.65), and the spectral decomposition of H_0 to (4.69) yields

$$\widetilde{\mathcal{F}}_{\text{sing}} - \mathcal{F}_{\text{sing}} = \mathcal{E}_{\text{sing},1} + \mathcal{E}_{\text{sing},2}, \tag{4.90}$$

where, recalling $x_\lambda := x - \lambda^2 m_0(E_0)$, the quantities $\mathcal{E}_{\text{sing},1}$ and $\mathcal{E}_{\text{sing},2}$ are defined as

$$\mathcal{E}_{\text{sing},1} := \sum_{j,k} \frac{\mathfrak{a}_j \mathfrak{p}_k}{N} \frac{1}{\pi} \int_{\mathbb{R}} \Im\big[\widetilde{\nu}_j(x) - (\mu_j - x_\lambda)^{-1}\big] \lambda^2 \big|\widetilde{\omega}_k(x)\big|^2 \mathrm{d}x, \tag{4.91}$$

$$\mathcal{E}_{\text{sing},2} := \sum_{j,k} \frac{\mathfrak{a}_j \mathfrak{p}_k}{N} \frac{1}{\pi} \int_{\mathbb{R}} \Im\big[(\mu_j - x_\lambda)^{-1}\big] \lambda^2 \Big(\big|\widetilde{\omega}_k(x)\big|^2 - \big|\omega_k(x)\big|^2\Big) \mathrm{d}x. \tag{4.92}$$

First, we estimate the quantity $\mathcal{E}_{\text{sing},1}$ defined in (4.91). In the regime $x \in I_{3\Delta}$, the bounds in (4.129) imply that

$$\big|\widetilde{\nu}_j(x) - (\mu_j - x_\lambda)^{-1}\big| \lesssim (\Delta + \epsilon_0) \frac{\lambda^2}{|\mu_j - x_\lambda|^2}, \quad x \in I_{3\Delta}. \tag{4.93}$$

[16]The integral in (4.85) diverges logarithmically as x approaches the intersection of the contour γ_2 with the real line. However, the contribution of such singularities to the $\mathrm{d}x$ integral on the right-hand side of (4.84) is negligible.

On the other hand, it is straightforward to check that the regime $|x| \geq \frac{1}{2}R$ contributes at most $\mathcal{O}(\lambda^2 R^{-2})$ to the integral on the right-hand side of (4.91). We note that the logarithmic singularity resulting from the contour γ_2 intersecting the real line is removed by the x integration.

Therefore, estimates (4.83), (4.88), (4.89), and (4.93) imply that

$$\begin{aligned}\left|\mathcal{E}_{\mathrm{sing},1}\right| &\lesssim \sum_{j,k} \frac{|\mathfrak{a}_j|\mathfrak{p}_k}{N} \int_{I_{3\Delta}} \frac{\lambda^2(\Delta+\epsilon_0)}{|\mu_j - x_\lambda|^2} \frac{\lambda^2}{|\mu_k - x_\lambda|^2} \mathrm{d}x \\ &\quad + \lambda^2(1+\lambda^2 t + \Delta^{-1}\log t)^2 \sum_j \frac{1}{N} \int_{\mathbb{R}} \Im\left[\widetilde{\nu}_j(x) + (\mu_j - x_\lambda)^{-1}\right] \mathrm{d}x \qquad (4.94) \\ &\lesssim \sum_{j,k} \frac{|\mathfrak{a}_j|\mathfrak{p}_k}{N} \int_{\mathbb{R}} \frac{\lambda^2(\Delta+\epsilon_0)}{|\mu_j - x_\lambda|^2} \frac{\lambda^2}{|\mu_k - x_\lambda|^2} \mathrm{d}x + \lambda^2\left(1+\lambda^2 t + \Delta^{-1}\log t\right)^2,\end{aligned}$$

where in the fist step we used the bound $|\Im[\widetilde{\nu}_j(x) - (\mu_j - x_\lambda)^{-1}]| \leq \Im[\widetilde{\nu}_j(x) + (\mu_j - x_\lambda)^{-1}]$ to estimate the contribution coming form the second term on the right-hand side of (4.88) and the regime $x \in [-\frac{1}{2}R, \frac{1}{2}R] \setminus I_{3\Delta}$; and in the second step we used that $\int_{\mathbb{R}} \Im[(\mu_j - x_\lambda)^{-1}]\mathrm{d}x = \pi$ and $\int_{\mathbb{R}} \frac{1}{N}\sum_j \Im\widetilde{\nu}_j(x)\mathrm{d}x = \int_R \Im m_\lambda(x)\mathrm{d}x = \pi$ (see, e.g., Proposition 2.1 and Eq. (2.9) in [6]). Computing the integral in the second term on the right-hand side of (4.94) explicitly, and using the spectral decomposition of H_0, we deduce from the admissibility of E_0 that

$$\sum_{j,k} \frac{|\mathfrak{a}_j|\mathfrak{p}_k}{N} \int_{\mathbb{R}} \frac{\lambda^2(\Delta+\epsilon_0)}{|\mu_j - x_\lambda|^2} \frac{\lambda^2 \mathrm{d}x}{|\mu_k - x_\lambda|^2} \lesssim (\Delta+\epsilon_0) \sup_{\mu_k \in I_\Delta} \left\langle\!\left\langle \Im M_0(\mu_k + 2\mathrm{i}\lambda^2\alpha) \right\rangle\!\right\rangle \lesssim \Delta + \epsilon_0. \quad (4.95)$$

Here we employed (4.83) and the estimate $\langle \mathfrak{a}, X\mathfrak{p}\rangle \lesssim \|\mathfrak{a}\|_\infty \|\mathfrak{p}\|_1 \sup_{k \in \mathrm{supp}(\mathfrak{p})} \sum_j |X_{jk}|$. Hence, we conclude that

$$\left|\mathcal{E}_{\mathrm{sing},1}\right| \lesssim \Delta + \epsilon_0 + \lambda^2\left(1+\lambda^2 t + \Delta^{-1}\log t\right)^2. \quad (4.96)$$

We proceed to estimate the quantity $\mathcal{E}_{\mathrm{sing},2}$ defined in (4.92). We note again that the contribution of the regime $|x| \geq \frac{1}{2}R$ to the integral on the right-hand side of (4.92) is bounded by $\mathcal{O}(\lambda^2 R^{-2})$. Therefore, combining the estimates (4.87), (4.88), and (4.89) yields the bound

$$\left|\mathcal{E}_{\mathrm{sing},2}\right| \lesssim (\Delta+\epsilon_0)(1+\lambda^2 t) + \lambda\left(1+\lambda^2 t + \Delta^{-1}\log t\right)^2, \quad (4.97)$$

obtained similarly to (4.94) and (4.96). Together with (4.90), the bound (4.96) and (4.97) conclude the proof of (4.67). □

It remains to prove Lemma 4.3.5.

Proof of Lemma 4.3.5. Throughout the proof we assume that $k \in [N]$ satisfies $\mu_k \in I_\Delta$, and $x \in \mathbb{R}$ satisfies $|x| \leq \frac{1}{2}R$. We introduce the auxiliary quantities

$$\check{\nu}_k(z) := \frac{1}{\mu_k - z - \lambda^2 m_\lambda(\mu_k)}, \quad \check{\omega}_k(x) := \frac{1}{2\pi\mathrm{i}} \oint_{\gamma_2} \frac{\mathrm{e}^{-\mathrm{i}tz}}{x - z} \check{\nu}_k(z)\mathrm{d}z. \quad (4.98)$$

An explicit computation using the residue calculus reveals that

$$\check{\omega}_k(x) = \frac{\mathrm{e}^{-\mathrm{i}tx} - \mathrm{e}^{-\mathrm{i}t(\mu_k - \lambda^2 m_\lambda(\mu_k))}}{\mu_k - x - \lambda^2 m_\lambda(\mu_k)}, \quad \omega_k(x) = \frac{\mathrm{e}^{-\mathrm{i}tx} - \mathrm{e}^{-\mathrm{i}t(\mu_k - \lambda^2 m_0(E_0))}}{\mu_k - x_\lambda}. \quad (4.99)$$

Furthermore, using the bound in (4.129), we obtain

$$\Big|\omega_k(x) - \check{\omega}_k(x)\Big| \lesssim \left(\frac{1}{|\mu_k - x_\lambda|^2} + \frac{t}{|\mu_k - x_\lambda|}\right)\lambda^2|\mu_k - E_0| \lesssim \frac{\Delta(1+\lambda^2 t)}{|\mu_k - x_\lambda|}, \tag{4.100}$$

where we additionally applied the estimate

$$\Big|y + \mathcal{O}(\eta)\Big| + \eta \sim |y| + \eta, \quad y \in \mathbb{R},\ \eta > 0. \tag{4.101}$$

We decompose the contour $\gamma_2 = \Gamma_1 \dot{+} \Gamma_2 \dot{+} \Gamma_3$ according to (4.60). It is straightforward to check that for $\nu^{\#}(z)$ denoting one of $\widetilde{\nu}_k(z)$, $\check{\nu}_k(z)$ or $(\mu_k - z_\lambda)^{-1}$,

$$\left|\int_{\Gamma_2 \dot{+} \Gamma_3} \frac{\mathrm{e}^{-\mathrm{i}tz}}{x - z}\nu_k^{\#}(z)\mathrm{d}z\right| \lesssim \frac{\log t}{\Delta} + \frac{1}{R}, \tag{4.102}$$

where we used that $\Im z = t^{-1}$ for all $z \in \Gamma_1$. Therefore, rewriting the left-hand sides of (4.87)-(4.89) using the integral definitions (4.85) and(4.86), it suffices to estimate the contributions coming from the segment $\Gamma_1 \subset \gamma_2$.

Using (4.101) and (4.129), we deduce that for all $z \in \Gamma_1$, defined in (4.60),

$$\Big|\check{\nu}_k(z) - \widetilde{\nu}_k(z)\Big| \lesssim \frac{\lambda^2}{|\mu_k - z_\lambda|}. \tag{4.103}$$

Integrating the bound (4.103) then yields

$$\left|\int_{\Gamma_1} \frac{\mathrm{e}^{-\mathrm{i}tz}}{x - z}\Big[\widetilde{\nu}_k(z) - \check{\nu}_k(z)\Big]\mathrm{d}z\right| \lesssim \Big(1 + \lambda^2 t\Big)\mathbb{1}_{x \in I_{3\Delta}} + \Big(\Delta^{-1}\log t\Big)\mathbb{1}_{x \notin I_{3\Delta}}, \tag{4.104}$$

which, together with (4.100), (4.102) immediately implies (4.87) after writing $\widetilde{\omega}_k - \omega_k = (\widetilde{\omega}_k - \check{\omega}_k) + (\check{\omega}_k - \omega_k)$.

On the other hand, noting that $|\omega_k(x)| + |\check{\omega}_k(x)| \lesssim |\mu_k - x_\lambda|^{-1}$ by (4.99), and combining the estimates (4.100) and (4.102) yields (4.88) and (4.89). This concludes the proof of Lemma 4.3.5. □

4.3.3 Step (iii): Limiting Behavior of the Singular Term and Proof of Theorem 4.2.4 (b)

We organize the result of approximating $\langle A\rangle_{\widetilde{P}_{\lambda,t}}$ in the following proposition.

Proposition 4.3.6. *Under the assumptions of Theorem 4.2.4, and with $\widetilde{P}_\lambda$ defined as in* (4.18)*, we have that, for any fixed $T \in (0,\infty)$ and recalling $\alpha = \pi\rho_0(E_0)$*

$$\limsup_{\Delta\to 0}\limsup_{\substack{t\to\infty\\ \lambda\to 0\\ \lambda^2 t = T}}\limsup_{N\to\infty}\Big|\langle A\rangle_{\widetilde{P}_{\lambda,t}} - (1 - \mathrm{e}^{-2\alpha\lambda^2 t})\langle A\rangle_{\widetilde{P}_\lambda}\Big| \lesssim T\mathrm{e}^{-2\alpha T}.$$

Given Proposition 4.3.6, Theorem 4.2.4 (b) immediately follows. □

Proof of Proposition 4.3.6. First, we observe that representing $\Im M_\lambda$ in spectral decomposition of H_0, the quantity $\langle A\rangle_{\widetilde{P}_{\lambda,t}}$ with $P_{\lambda,t}$ defined in (4.15), can be rewritten in the from

$$\langle A\rangle_{\widetilde{P}_{\lambda,t}} = \frac{1}{r}\sum_{j,k}\mathfrak{a}_j\mathfrak{p}_k\int_{\mathbb{R}}\varphi_{\alpha\lambda^2}(x - \mu_j)\Big(K_{\lambda,t} * \varphi_{\alpha\lambda^2}\Big)(x - \mu_k)\mathrm{d}x, \tag{4.105}$$

where $r = \int_{\mathbb{R}} \mathrm{Tr}[\Im M_0(x + \mathrm{i}\alpha\lambda^2)]\langle \Im M_0(x + \mathrm{i}\alpha\lambda^2)\rangle_P \mathrm{d}x > 0$ has already been introduced in Remark 4.2.2 (ii), and we denoted $\varphi_\eta := \Im[(x - \mathrm{i}\eta)^{-1}]$. Recall that $\mu_j, \boldsymbol{u}_j$ are the eigenvalues and the respective eigenvectors of H_0, and $\mathfrak{a}_j := \langle \boldsymbol{u}_j, A\boldsymbol{u}_j\rangle$, $\mathfrak{p}_j := \langle \boldsymbol{u}_j, P\boldsymbol{u}_j\rangle$. Applying the Parseval-Plancherel identity to the right-hand side of (4.105) yields

$$\langle A\rangle_{\widetilde{P}_{\lambda,t}} = \frac{1}{r}\sum_{j,k} \mathfrak{a}_j\mathfrak{p}_k\Phi_{\lambda,t}(\mu_j - \mu_k), \quad \Phi_{\lambda,t}(u) := \frac{\pi^2}{(2\pi)^{1/2}}\int_{\mathbb{R}} \mathrm{e}^{-2\alpha\lambda^2|q| - \mathrm{i}uq}\,\widehat{K_{\lambda,t}}(q)\mathrm{d}q, \tag{4.106}$$

where we used the fact that $\widehat{\varphi_\eta}(q) = (\frac{\pi}{2})^{1/2}\mathrm{e}^{-\eta|q|}$, $\eta > 0$ (recall Footnote 5).

A direct computation starting with (4.16) reveals that

$$\widehat{K_{\lambda,t}}(q) = \begin{cases} (2\pi)^{-1/2}\Big(1 - \mathrm{e}^{-2\alpha\lambda^2(t-|q|)}\Big) & \text{for} \quad |p| \le t\,, \\ 0 & \text{for} \quad |p| > t\,, \end{cases} \tag{4.107}$$

and implies that $\Phi_{\lambda,t}(u)$ admits the explicit expression

$$\Phi_{\lambda,t}(u) = \Big(1 - \mathrm{e}^{-2\alpha\lambda^2 t}\Big)\pi\varphi_{2\alpha\lambda^2}(u) + \mathfrak{R}_{\lambda,t}(u), \tag{4.108}$$

where the function $\mathfrak{R}_{\lambda,t}(u)$ is defined by

$$\mathfrak{R}_{\lambda,t}(u) := \pi\mathrm{e}^{-2\alpha\lambda^2 t}\varphi_{2\alpha\lambda^2}(u)\bigg(1 - \cos(tu) - 2\alpha\lambda^2 t\frac{\sin(tu)}{tu}\bigg). \tag{4.109}$$

Observe that the contribution of the first term on the right-hand side of (4.108) to $\langle A\rangle_{\widetilde{P}_{\lambda,t}}$ is given by $(1 - \mathrm{e}^{-2\lambda^2\alpha t})\langle A\rangle_{\widetilde{P}_\lambda}$, since

$$\langle A\rangle_{\widetilde{P}_\lambda} = \frac{1}{r}\sum_{j,k} \mathfrak{a}_j\mathfrak{p}_k\int_{\mathbb{R}} \varphi_{\alpha\lambda^2}(x - \mu_j)\varphi_{\alpha\lambda^2}(x - \mu_k)\mathrm{d}x. \tag{4.110}$$

Here we used the definition of the state $\widetilde{P}_\lambda$ in (4.18), and the Parseval-Plancherel identity. The key observation is that the contribution of the remaining $\mathfrak{R}_{\lambda,t}(\mu_j - \mu_k)$ term

$$\mathcal{R} := \sum_{j,k} \frac{\mathfrak{a}_j\mathfrak{p}_k}{r}\mathfrak{R}_{\lambda,t}(\mu_j - \mu_k) \tag{4.111}$$

in (4.108) to $\langle A\rangle_{\widetilde{P}_{\lambda,t}}$ admits the bound[17]

$$\Big|\mathcal{R}\Big| \le \sup_{k\in\mathrm{supp}(\mathfrak{p})} \frac{1}{r}\sum_j \Big|\mathfrak{R}_{\lambda,t}(\mu_j - \mu_k)\Big| \cdot \|\mathfrak{a}\|_\infty\|\mathfrak{p}\|_1\,. \tag{4.112}$$

Observe that there exists a constant $C > 0$ such that for any $\xi > 0$ and $t > 0$, we have that $\varphi_\xi(u)(1 - \cos(tu)) \le C\xi t\varphi_{1/t}(u)$ for all $u \in \mathbb{R}$. This follows immediately from the fact that the function $s \mapsto (s^2 + 1)(1 - \cos s)/s^2$ is uniformly bounded on $\mathbb{R}$. Therefore, the function $\mathfrak{R}$ admits the bound

$$|\mathfrak{R}_{\lambda,t}(u)| \le 2\pi\alpha\lambda^2 t\,\mathrm{e}^{-2\alpha\lambda^2 t}\bigg(C\varphi_{1/t}(u) + \varphi_{2\alpha\lambda^2}(u)\bigg), \quad u \in \mathbb{R}. \tag{4.113}$$

[17]Inequality (4.112) can be interpreted as a discrete analog of Young's convolution inequality, which can not be evoked directly since the eigenvalues μ_j do not form a group under addition.

Summing the bound (4.113) over $u = \mu_j$ yields

$$\frac{1}{r}\sum_j \Big|\mathfrak{R}_{\lambda,t}(\mu_j - \mu_k)\Big| \lesssim \lambda^2 t\, \mathrm{e}^{-2\alpha\lambda^2 t}\frac{N}{r}\Im\langle\!\langle M_0(\mu_k + 2\mathrm{i}\alpha\lambda^2) + M_0(\mu_k + \mathrm{i}/t)\rangle\!\rangle. \tag{4.114}$$

Using the localization of the state P as in (4.12), the admissibility of E_0 in (4.11), and the first line of (4.140) to deduce that $r \sim N(1 + \mathcal{O}(\epsilon_0 + \Delta + \lambda^2/\Delta))$, we obtain

$$\sup_{k\in\mathrm{supp}(\mathfrak{p})}\frac{1}{r}\sum_j \Big|\mathfrak{R}_{\lambda,t}(\mu_j - \mu_k)\Big| \lesssim \lambda^2 t\, \mathrm{e}^{-2\alpha\lambda^2 t}\Big(1 + \epsilon_0\Big)\Big(1 + \epsilon_0 + \Delta + \lambda^2/\Delta\Big). \tag{4.115}$$

This concludes the proof of Proposition 4.3.6. □

4.3.4 Relaxation formula: Proof of Corollary 4.2.6

Estimates (4.24) and (4.25) in items (a) and (b), respectively, follow immediately from Theorem 4.2.4.

To prove (4.26) in item (c), observe that plugging the estimate (4.22) from the Definition 4.2.5 of local overlap regularity into (4.111) yields

$$\begin{aligned}\Big|\mathcal{R}\Big| \lesssim |\mathfrak{A}| \sup_{k\in\mathrm{supp}(\mathfrak{p})}\bigg|\frac{1}{N}\sum_{\mu_j\in I_{2\Delta}}\mathfrak{R}_{\lambda,t}(\mu_j - \mu_k)\bigg| + \sup_{k\in\mathrm{supp}(\mathfrak{p})}\bigg|\sum_{\mu_j\in I_{2\Delta}}\frac{\mathfrak{a}_j - \mathfrak{A}}{N}\mathfrak{R}_{\lambda,t}(\mu_j - \mu_k)\bigg| \\ + \sup_{k\in\mathrm{supp}(\mathfrak{p})}\bigg|\frac{1}{N}\sum_{\mu_j\notin I_{2\Delta}}\mathfrak{a}_j\mathfrak{R}_{\lambda,t}(\mu_j - \mu_k)\bigg|,\end{aligned} \tag{4.116}$$

where we used that $|r| \sim N\,(1 + \mathcal{O}(\epsilon_0 + \Delta + \lambda^2/\Delta))$ by the first line of (4.140) from Lemma 4.A.2. Applying the estimates analogous to (4.112) and (4.115) to the second sum on the right-hand side of (4.116), we deduce the bound

$$\sup_{k\in\mathrm{supp}(\mathfrak{p})}\bigg|\sum_{\mu_j\in I_{2\Delta}}\frac{\mathfrak{a}_j - \mathfrak{A}}{N}\mathfrak{R}_{\lambda,t}(\mu_j - \mu_k)\bigg| \lesssim \Big|\mathcal{E}_{\mathrm{LOR}}\Big|\cdot\Big(1 + \lambda^2/\Delta\Big) \tag{4.117}$$

Note that by (4.83) and the uniform bound

$$|\mathfrak{R}_{\lambda,t}(u)| \lesssim \frac{\lambda^2}{\Delta^2}, \quad \text{for } |u| \gtrsim \Delta$$

following from (4.113), the tail sum, i.e., the second line of (4.116), admits the estimate

$$\sup_{k\in\mathrm{supp}(\mathfrak{p})}\bigg|\frac{1}{N}\sum_{\mu_j\notin I_{2\Delta}}\mathfrak{a}_j\mathfrak{R}_{\lambda,t}(\mu_j - \mu_k)\bigg| \lesssim \frac{\lambda^2}{\Delta^2}. \tag{4.118}$$

Therefore, it remains to estimate the first term on the right-hand side of (4.116). Since the function $\mathfrak{R}_{\lambda,t}(u)$ is holomorphic in u for $|\Im u| \le \alpha\lambda^2$, we obtain the following series of estimates,

$$\begin{aligned}\frac{1}{N}\sum_{\mu_j\in I_{2\Delta}}\mathfrak{R}_{\lambda,t}(\mu_j - \mu_k) &= \frac{1}{2\pi\mathrm{i}}\oint_\gamma \mathfrak{R}_{\lambda,t}(z - \mu_k)\langle\!\langle M_0(z)\rangle\!\rangle\mathrm{d}z + \mathcal{O}\bigg(\frac{\lambda^2}{\Delta^2}\bigg) \\ &= \frac{1}{2\pi\mathrm{i}}\oint_\gamma \mathfrak{R}_{\lambda,t}(z - \mu_k)m_0(z)\mathrm{d}z + \mathcal{O}\bigg(\eta_0 + \frac{\epsilon_0}{\lambda^2} + \frac{\lambda^2}{\Delta^2}\bigg) \\ &= \int_{I_{2\Delta}}\mathfrak{R}_{\lambda,t}(u - \mu_k)\rho_0(u)\mathrm{d}u + \mathcal{O}\bigg(\eta_0 + \frac{\epsilon_0}{\lambda^2} + \frac{\lambda^2}{\Delta^2}\bigg),\end{aligned} \tag{4.119}$$

where the contour γ is defined to be a rectangle of height $2\eta_0$ and width $4C$ centered at E_0, and the constant $C \sim 1$ is chosen in such a way that $\sigma(H_0) \subset [E_0 - C, E_0 + C]$. Here, in the first step, we used residue calculus and an estimate analogous to (4.118) to extend the sum to all μ_j's. The second step follows by integrating the estimate (4.8) on the horizontal segments of γ and bounding the contribution of the vertical segments of the contour γ by $\mathcal{O}(\eta_0)$. Finally, the third step is a consequence of the Stieltjes representation (4.9) and $|\mathfrak{R}_{\lambda,t}(u)| \lesssim \Delta^{-2}\lambda^2$ for $|u| \gtrsim \Delta$. Using the estimate $\rho_0(u) = \rho_0(E_0) + \mathcal{O}(\Delta)$ for all $u \in I_{2\Delta}$ by admissibility of E_0 as in (4.10), we conclude that

$$\int_{I_{2\Delta}} \mathfrak{R}_{\lambda,t}(u-\mu_k)\rho_0(u)\mathrm{d}u = \rho_0(E_0)\int_{I_{2\Delta}} \mathfrak{R}_{\lambda,t}(u-\mu_k)\mathrm{d}u + \mathcal{O}(\Delta), \tag{4.120}$$

where we used $\int_{\mathbb{R}} |\mathfrak{R}_{\lambda,t}(u)|\mathrm{d}u \lesssim \lambda^2 t\,\mathrm{e}^{-2\alpha\lambda^2 t} \lesssim 1$ as a consequence of (4.113). Moreover, a direct computation starting with (4.109) reveals that

$$\int_{\mathbb{R}} \mathfrak{R}_{\lambda,t}(u-\mu_k)\mathrm{d}u = 0 \qquad \text{and} \qquad \int_{\mathbb{R}\setminus I_{2\Delta}} \Big|\mathfrak{R}_{\lambda,t}(u-\mu_k)\Big|\mathrm{d}u \lesssim \frac{\lambda^2}{\Delta}. \tag{4.121}$$

Hence, combining estimates (4.116)–(4.121) yields

$$\Big|\mathcal{R}\Big| \lesssim \Delta + \eta_0 + \lambda^{-2}\epsilon_0 + \Delta^{-2}\lambda^2, \tag{4.122}$$

which implies the $\mathcal{R}$-part of (4.26); the $\mathcal{E}$-part is an immediate consequence of Theorem 4.2.4 (a).

To complete the proof under the weaker assumption on $\langle \boldsymbol{u}_j, A\boldsymbol{u}_j\rangle$, stated in Footnote 6, we first *uniformly* approximate $\mathfrak{A}_N$ by a real analytic function $\mathfrak{A}_{N,\ell} : I_{2\Delta} \to \mathbb{R}$ with $\ell = \ell(\lambda, t) > 0$ (to be chosen below), which can be analytically extended to $\{z \in \mathbb{C} : \mathrm{dist}(z, I_{2\Delta}) < \ell\}$ and satisfy $\sup_{N\in\mathbb{N}} \|\mathfrak{A}_{N,\ell} - \mathfrak{A}_N\|_\infty \to 0$ as $\ell \to 0$. Such $\mathfrak{A}_{N,\ell}$ can be explicitly constructed, e.g., by convolution of $\mathfrak{A}_N$ with a Gaussian having variance of order ℓ. For ease of notation, we shall now drop the subscript N. Then, the error term $\mathfrak{A} - \mathfrak{A}_\ell$ is easily seen to give a vanishing contribution (as $\ell \to 0$) by means of (4.115). Next, observe that using analyticity of $\mathfrak{A}_\ell$ and reasoning as in the proof for the case of $\mathfrak{A}$ being constant above, we obtain

$$\Big|\mathcal{R}\Big| \lesssim \sup_{\mu_k\in I_\Delta} \left|\int_{I_{2\Delta}} \Big(\mathfrak{A}_\ell(u) - \mathfrak{A}_\ell(\mu_k)\Big)\mathfrak{R}_{\lambda,t}(u-\mu_k)\mathrm{d}u\right| + \Delta + \eta_0 + \lambda^{-2}\epsilon_0 + \Delta^{-2}\lambda^2. \tag{4.123}$$

Since $\mathfrak{A}_\ell(z)$ is analytic in the strip of width ℓ, (4.113) implies that the integral on the right-hand side of (4.123) is bounded by

$$\frac{1}{\ell}\int_{I_{2\Delta}} \left(\frac{\lambda^2|u-\mu_k|}{|u-\mu_k|^2 + (2\alpha\lambda^2)^2} + \frac{t^{-1}|u-\mu_k|}{|u-\mu_k|^2 + t^{-2}}\right)\mathrm{d}u \lesssim \frac{\lambda^2|\log\lambda| + t^{-1}\log t}{\ell}, \tag{4.124}$$

uniformly in k such that $\mu_k \in I_\Delta$. Hence, choosing, say, $\ell := \lambda + t^{-1/2}$, this concludes the proof of Corollary 4.2.6. □

4.3.5 Microcanonical Average: Proof of Theorem 4.2.9

Using (4.18), we start by writing out

$$\langle A\rangle_{\widetilde{P}_\lambda} = \frac{\int_{\mathbb{R}} \langle\!\langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)A\rangle\!\rangle \,\langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle_P\,\mathrm{d}x}{\int_{\mathbb{R}} \langle\!\langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle\!\rangle \,\langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle_P\,\mathrm{d}x}. \tag{4.125}$$

For the denominator, we have

$$\begin{aligned}&\int_{\mathbb{R}} \langle\!\langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle\!\rangle \, \langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle_P \, \mathrm{d}x \\ &= \pi \langle\!\langle \Im M_0(E_0+\mathrm{i}\alpha\lambda^2)\rangle\!\rangle + \mathcal{O}\Big(\epsilon_0 + \Delta + \lambda^2/\Delta\Big)\end{aligned} \tag{4.126}$$

from Lemma 4.A.2. For the numerator, we use that $h(x) = \langle\!\langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)A\rangle\!\rangle$ has uniformly bounded Lipschitz constant for $x \in I_\Delta$ (recall (4.33)). Hence we find

$$\int_{\mathbb{R}} h(x) \, \langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle_P \, \mathrm{d}x = \pi h(E_0) + \mathcal{O}\Big(\Delta + \lambda^2/\Delta\Big) \tag{4.127}$$

completely analogously to (4.141) and (4.142), using Assumption 4.2.3 (i).

Therefore, plugging (4.126) and (4.127) into (4.125), and using Assumption 4.2.2 (ii) together with Lemma 4.A.2, we obtain

$$\langle A\rangle_{\widetilde{P}_\lambda} = \langle A\rangle_{P_\lambda^{(\mathrm{mc})}} + \mathcal{O}(\mathcal{E}_{\mathrm{mc}}) \quad \text{with} \quad \mathcal{E}_{\mathrm{mc}} := \epsilon_0 + \Delta + \lambda^2/\Delta.$$

This concludes the proof of Theorem 4.2.9. □

4.A Auxiliary Results and Additional Proofs

4.A.1 Auxiliary Results

In this subsection, we derive two technical lemmas, which are frequently used throughout the main text.

The first one (Lemma 4.A.1) is concerned with properties of the self-consistent resolvent $M_\lambda(z)$ from the λ-dependent quadratic matrix equation (4.42) using Assumption 4.2.2 on the unperturbed matrix H_0. Recall that $\frac{1}{6}\kappa_0$ denotes the upper bound for the energy width Δ (cf. Assumption 4.2.3).

Lemma 4.A.1. *Let $z := E + \mathrm{i}\eta$ be a spectral parameter in $\mathbb{C}$ with $|z| \le C$, then the solution $M_\lambda(z)$ to (4.42) satisfies the bounds*

$$\langle\!\langle |M_\lambda(z)|^2\rangle\!\rangle \le \lambda^{-2}, \quad \langle\!\langle |M_\lambda(z)|\rangle\!\rangle \le \lambda^{-1}, \quad \Im z > 0. \tag{4.128}$$

Moreover, assuming that $\Im z \ge 0$, $|\Re z - E_0| \le \frac{1}{2}\kappa_0$, and E_0 lies in the admissible spectrum $\sigma_{\mathrm{adm}}^{(\kappa_0,c_0)}$ of H_0 the following estimates

$$\Im m_\lambda(z) \gtrsim 1, \quad |m_\lambda(z)| \lesssim 1, \quad |m_\lambda'(z)| \lesssim 1, \tag{4.129}$$

with $m_\lambda(z) := \langle\!\langle M_\lambda(z)\rangle\!\rangle$ hold for any fixed $0 < \lambda \le \lambda_*$, and all $N \ge N_\lambda \in \mathbb{N}$, uniformly in z.

Proof of Lemma 4.A.1. First, we prove the (4.128) for $\Im z > 0$. Taking the imaginary part of the MDE (4.42), multiplying by $|M_\lambda|^2$ and taking the averaged trace yields

$$\Im m_\lambda(z) = \Big(\Im z + \lambda^2 \Im m_\lambda(z)\Big) \langle\!\langle |M_\lambda(z)|^2\rangle\!\rangle, \tag{4.130}$$

which immediately implies the first bound in (4.128). The second estimate in (4.128) follows from the first one by the Cauchy–Schwarz inequality. To extend (4.128) down to $\Im z = 0$, we first address the regularity[18] of $m_\lambda(z) = \langle\!\langle M_\lambda(z) \rangle\!\rangle$.

Differentiating the MDE (4.42), taking the trace, and evoking the first bound in (4.128), we obtain

$$\big|m'_\lambda(z)\big| = \frac{\Big|\langle\!\langle M_\lambda(z)^2 \rangle\!\rangle\Big|}{\Big|1 - \lambda^2 \langle\!\langle M_\lambda(z)^2 \rangle\!\rangle\Big|} \leq \frac{1}{2\lambda^4 \Big\langle\!\Big\langle (\Im M_\lambda(z))^2 \Big\rangle\!\Big\rangle} \lesssim \frac{1}{\lambda^4 (\Im m_\lambda(z))^2}, \tag{4.131}$$

where we used (4.128) to deduce that

$$|1 - \lambda^2 \langle\!\langle M_\lambda(z)^2 \rangle\!\rangle| \geq 1 - \lambda^2 \Re \langle\!\langle M_\lambda(z)^2 \rangle\!\rangle \geq 2\lambda^2 \langle\!\langle (\Im M_\lambda(z))^2 \rangle\!\rangle,$$

and the positive-definiteness of $\Im M_\lambda(z)$ to obtain the last inequality in (4.131).

Together with the first estimate in (4.128), (4.131) implies that $\lambda^{4/3} \Im m_\lambda(z)$ is uniformly 1/3-Hölder continuous in $\{z \in \mathbb{C} : |\Im z| \geq 0, |z| \leq C\}$. This concludes the proof of (4.128) for $\Im z \geq 0$.

Next, we turn to proving the first estimate in (4.129). It follows from the MDE (4.42) that

$$m_\lambda(z) = \Big\langle\!\Big\langle \big(H_0 - z - \lambda^2 m_\lambda(z)\big)^{-1} \Big\rangle\!\Big\rangle, \tag{4.132}$$

and $\Im z + \lambda^2 \Im m_\lambda(z) \geq \Im z$.

First, assume that $\Im z \geq \eta_0$ (recall (4.8)). It follows from (4.128) that $\lambda^2 |m_\lambda(z)| \leq \lambda$, hence by suitably shrinking the threshold λ_*, we can assume that $|\Re[z + \lambda^2 m_\lambda(z)] - E_0| \leq \frac{3}{4}\kappa_0$ for all z satisfying $\Im z \geq \eta_0$, $|z| \leq C$ and $|\Re z - E_0| \leq \frac{1}{2}\kappa_0$. Therefore

$$m_\lambda(z) = m_0\big(z + \lambda^2 m_\lambda(z)\big) + \mathcal{O}(\epsilon_0) = m_0(z) + \mathcal{O}(\lambda + \epsilon_0), \tag{4.133}$$

where the first step follows by (4.8) and (4.128), and in the second estimate we used that $E_0 \in \sigma_{\mathrm{adm}}^{(\kappa_0, c_0)}$, defined in (4.10). In particular, taking the imaginary part of (4.133), and using the positivity of ρ_0 in the admissible spectrum yields $\Im m_\lambda(z) \gtrsim 1 + \mathcal{O}(\lambda + \epsilon_0)$. Hence, from the 1/3-Hölder continuity of $\lambda^{4/3} \Im m_\lambda(z)$, and (4.133) we deduce that

$$\lambda^2 \Im m_\lambda(z) \gtrsim \lambda^2 + \mathcal{O}(\lambda^3 + \lambda^2 \epsilon_0 + \lambda^{2/3} \eta_0^{1/3}). \tag{4.134}$$

for all z with $|\Re z - E_0| \leq \frac{1}{2}\kappa_0$, $\Im z \geq 0$, and $|z| \leq C$. Therefore, for a suitably small threshold λ_* and any fixed $0 < \lambda \leq \lambda_*$, the first estimate in (4.129) is established for all N satisfying $\eta_0(N)^{1/3} \lesssim \lambda^2$ with the implicit constant depending only on the constant in (4.131). Since $\eta_0(N)$ converges to zero, there exists $N_\lambda \in \mathbb{N}$ such that all $N \geq N_\lambda$ satisfy $\eta_0(N)^{1/3} \lesssim \lambda^2$. Furthermore, we obtain $\Im z + \lambda^2 \Im m_\lambda(z) \gtrsim \lambda^2 \geq \eta_0$, hence the second estimate in (4.129) follows from (4.8) and (4.132).

Finally, we prove the third estimate in (4.129). Differentiating (4.132) with respect to z yields

$$m'_\lambda(z) = \Big\langle\!\Big\langle \big(H_0 - z - \lambda^2 m_\lambda(z)\big)^{-2} \Big\rangle\!\Big\rangle \big(1 + \lambda^2 m'_\lambda(z)\big), \tag{4.135}$$

[18] The comprehensive analysis of the MDE in [4, 6] shows that for any fixed λ, under the additional boundedness assumption $\|M_\lambda\| \lesssim 1$, the operator $M_\lambda(z)$ is 1/3-Hölder continuous with a λ-dependent constant. However, for the purposes of proving the regularity of m_λ, the operator norm bound is not necessary.

Im particular, the first factor on the right-hand side of (4.135) is a normalized trace of $(H_0-\zeta)^{-2}$ with $\zeta := z+\lambda^2 m_\lambda(z)$, satisfying $|\Re\zeta - E_0| \le \frac{3}{4}\kappa_0$ and $\Im\zeta \ge 2c\lambda^2 \ge \eta_0$ for some positive constant $c \sim 1$ by the first and second estimates in (4.129). Hence, integrating (4.8), we deduce that

$$\langle\!\langle (H_0-\zeta)^{-2}\rangle\!\rangle = \int_{\mathbb{R}+\mathrm{i}c\lambda^2} \frac{\Im\langle\!\langle (H_0-w)^{-1}\rangle\!\rangle}{\pi\,(w-\zeta)^2}\mathrm{d}w = \int_{\mathbb{R}+\mathrm{i}c\lambda^2} \frac{\Im m_0(w)}{\pi\,(w-\zeta)^2}\mathrm{d}w + \mathcal{O}\big(\lambda^{-2}\epsilon_0\big). \tag{4.136}$$

Moreover, using the fact that $m_0(w)$ is the Stieltjes transform of the limiting density ρ_0, we conclude that

$$\langle\!\langle (H_0-\zeta)^{-2}\rangle\!\rangle = \frac{1}{\pi}\int_{\mathbb{R}} \frac{\rho_0(x)}{(x-\zeta)^2}\mathrm{d}x + \mathcal{O}\big(\lambda^{-2}\epsilon_0\big), \tag{4.137}$$

In particular, since $\Re\zeta$ lies in the admissible spectrum of H_0, the integral on the right-hand side of (4.137) admits the estimate

$$\left|\frac{1}{\pi}\int_{\mathbb{R}} \frac{\rho_0(x)}{(x-\zeta)^2}\mathrm{d}x\right| = \left|\int_J \frac{\rho_0'(x)-\rho_0'(\Re\zeta)}{x-\zeta}\mathrm{d}x\right| + \left|\int_J \frac{\rho_0'(\Re\zeta)\,\Im\zeta}{|x-\zeta|^2}\mathrm{d}x\right| + \mathcal{O}(\kappa_0^{-2}), \tag{4.138}$$

where $J := [\Re\zeta - c\kappa_0, \Re\zeta + c\kappa_0] \subset I_{\kappa_0}$. Here we used that the kernel $(x-\Re\zeta)/|x-\zeta|^2$ is odd around $\Re\zeta$ and the integrability of ρ_0 to estimate the integral over $\mathbb{R}\setminus J$, while the boundary term resulting from integration by parts over J is bounded by $\kappa_0^{-1}\mathcal{O}(\|\rho_0\|_{C^1(J)}) \lesssim \kappa_0^{-1}$. Observe that the second term on the right-hand side of (4.138) is bounded by $\mathcal{O}(|\rho_0'(\Re\zeta)|)$, while the first term is bounded by $\mathcal{O}(L)$, where L is the Lipschitz constant of ρ_0' on the interval $[E_0-\kappa_0, E_0+\kappa_0]$. Therefore, $|\langle\!\langle (H_0-\zeta)^{-2}\rangle\!\rangle| = \mathcal{O}(1)$. Finally, rearranging the identity (4.135) now yields

$$m_\lambda'(z) = \frac{\Big\langle\!\Big\langle \big(H_0 - z - \lambda^2 m_\lambda(z)\big)^{-2}\Big\rangle\!\Big\rangle}{1-\lambda^2\Big\langle\!\Big\langle \big(H_0 - z - \lambda^2 m_\lambda(z)\big)^{-2}\Big\rangle\!\Big\rangle} = \mathcal{O}(1). \tag{4.139}$$

This concludes the proof of Lemma 4.A.1. □

We conclude this section by evaluating the denominator in (4.15).

Lemma 4.A.2. *Under Assumptions 4.2.2 and 4.2.3 (i) (recalling that $\alpha = \pi\rho_0(E_0)$) it holds that*

$$\begin{aligned}\left|\int_{\mathbb{R}} \langle\!\langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle\!\rangle \langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle_P\,\mathrm{d}x - \pi^2\rho_0(E_0)\right| &= \mathcal{O}\big(\epsilon_0+\Delta+\lambda^2/\Delta\big),\\ \left|\pi^2\rho_0(E_0) - \pi\langle\!\langle \Im M_0(E_0+\mathrm{i}\alpha\lambda^2)\rangle\!\rangle\right| &= \mathcal{O}\big(\epsilon_0+\Delta+\lambda^2/\Delta\big).\end{aligned} \tag{4.140}$$

Proof. We only prove the first relation in (4.140). The argument for the second estimate is analogous and hence omitted.

We have that

$$\begin{aligned}&\int_{\mathbb{R}} \langle\!\langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle\!\rangle \langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle_P\,\mathrm{d}x\\ =&\int_{\mathbb{R}} \Im m_0(x+\mathrm{i}\alpha\lambda^2)\langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle_P\,\mathrm{d}x + \mathcal{O}(\epsilon_0)\\ =&\sum_{\mu_j\in I_\Delta}\langle \boldsymbol{u}_j, P\boldsymbol{u}_j\rangle \int_{\mathbb{R}}\mathrm{d}y\rho_0(y)\int_{\mathbb{R}}\mathrm{d}x \frac{\alpha\lambda^2}{(x-y)^2+(\alpha\lambda^2)^2}\frac{\alpha\lambda^2}{(x-\mu_j)^2+(\alpha\lambda^2)^2} + \mathcal{O}(\epsilon_0)\\ =&\pi^2\sum_{\mu_j\in I_\Delta}\langle \boldsymbol{u}_j, P\boldsymbol{u}_j\rangle \int_{\mathbb{R}}\mathrm{d}y\rho_0(y)\frac{1}{\pi}\frac{2\alpha\lambda^2}{(y-\mu_j)^2+(2\alpha\lambda^2)^2} + \mathcal{O}(\epsilon_0),\end{aligned} \tag{4.141}$$

where we used $\int_{\mathbb{R}} \langle \Im M_0(x+\mathrm{i}\alpha\lambda^2)\rangle_P \,\mathrm{d}x = \pi$ and (4.8) to go to the second line. To go to the third line, we employed spectral decomposition (4.52) of H_0 and used Assumption 4.2.3 (i). Next, using that $\mu_j \in I_\Delta$ and regularity of ρ_0 within $I_{2\Delta}$, we can evaluate the integral in the last line of (4.141) as

$$\int_{\mathbb{R}} \mathrm{d}y \rho_0(y) \frac{1}{\pi} \frac{2\alpha\lambda^2}{(y-\mu_j)^2 + (2\alpha\lambda^2)^2} = \rho_0(E_0) + \mathcal{O}\Big(\Delta + \lambda^2/\Delta\Big), \tag{4.142}$$

by adding and subtracting $\rho_0(E_0)$ and estimating $I_{2\Delta}$ and $I_{2\Delta}^c$ separately.

Combining (4.141) with (4.142) and using that $\mathrm{Tr}[P] = 1$ concludes the argument. □

4.A.2 Proof of Proposition 4.3.1

Our proof closely follows [37, Appendix B], [34, Section 5.2], and [33, Section 6.2],[19] hence, we only give the main steps. Note that we are only interested in a *global law*, i.e. the spectral parameters z_1, z_2 have imaginary parts uniformly bounded away from zero. In particular, we can simply afford the norm bounds $\|G_i\| \lesssim 1$ and $\|M_i\| \lesssim 1$.

As a preparation for our argument, we recall the definition of the *second order renormalization*, denoted by *underline*, from [36]. For functions $f(W), g(W)$ of the random matrix W, we define

$$\underline{f(W)Wg(W)} := f(W)Wg(W) - \widetilde{\mathbb{E}}\Big[(\partial_{\widetilde{W}} f)(W)\widetilde{W}g(W) + f(W)\widetilde{W}(\partial_{\widetilde{W}} g)(W)\Big], \tag{4.143}$$

where $\partial_{\widetilde{W}}$ denotes the directional derivative in the direction of the GUE matrix $\widetilde{W}$ that is independent of W. The expectation is taken w.r.t. the matrix $\widetilde{W}$. Note that if W itself is a GUE matrix, then $\mathbb{E}\underline{f(W)Wg(W)} = 0$, while for W with a general distribution, this expectation is independent of the first two moments of W. In other words, the underline renormalizes the product $f(W)Wg(W)$ to the second order. We remark that the underline (4.143) is a well-defined notation if the 'middle' W to which the renormalization refers is unambiguous. This is the case in our proof, since the functions f, g are resolvents, i.e. not involving explicitly monomials in W.

Moreover, we note that $\widetilde{\mathbb{E}}\widetilde{W}R\widetilde{W} = \langle\!\langle R \rangle\!\rangle$ and furthermore, that the directional derivative of the resolvent is given by $\partial_{\widetilde{W}} G = -G\widetilde{W}G$. For example, in the special case $f(W) = 1$ and $g(W) = (W + D - z)^{-1} = G$, we thus have

$$\underline{WG} = WG + \langle\!\langle G \rangle\!\rangle G$$

by definition of the underline in (4.143). Using this underline notation in combination with the identity $G(W + D - z) = I$ and the defining equation (4.39) for M, we have

$$G = M - M\underline{WG} + M\langle\!\langle G - M \rangle\!\rangle G = M - \underline{GW}M + G\langle\!\langle G - M \rangle\!\rangle M\,. \tag{4.144}$$

Moreover, we have the following lemma, the proof of which is given at the end of this section.

[19] In fact, for z_1, z_2 in the *bulk* of the self consistent density of states $\rho(x) := \pi^{-1} \lim_{\eta\to 0^+} \langle\!\langle \Im M(x+\mathrm{i}\eta) \rangle\!\rangle$, Proposition 4.3.1 has already been proven for so called *regular* matrices A in [33, Proposition 4.4]. Here, we provide the proof uniformly in the spectrum and for arbitrary matrices.

Lemma 4.A.3 (Representation as full underlined, cf. Lemma 5.6 in [34])**.** *Under the notations and assumptions of Proposition 4.3.1, we have that*

$$\left(G_1BG_2-\left(M_1BM_2+\frac{M_1M_2\langle\!\langle M_1BM_2\rangle\!\rangle}{1-\langle\!\langle M_1M_2\rangle\!\rangle}\right)\right)_{\boldsymbol{xy}}=-\big(\underline{G_1B'M_2WG_2}\big)_{\boldsymbol{xy}}+\mathcal{O}_{\prec}\left(\mathcal{E}_{\mathrm{iso}}\right) \tag{4.145}$$

with $\mathcal{E}_{\mathrm{iso}}:=1/\sqrt{N}$ *for some bounded deterministic matrix* $B'\in\mathbb{C}^{N\times N}$.

Having this approximate representation of the lhs. of (4.41) as a full underlined term at hand, we turn to the proof of (4.41) via a (minimalistic) cumulant expansion; see [37, Eq. (4.32)] and [34, Eq. (5.24)].

Let $p\in\mathbb{N}$ be arbitrary. Then, abbreviating the lhs. of (4.145) by $\mathcal{Q}_{\boldsymbol{xy}}$, we obtain

$$\begin{aligned}\mathbb{E}\big|\mathcal{Q}_{\boldsymbol{xy}}\big|^{2p}\lesssim\ &\mathbb{E}\,\widetilde{\Xi}\,\big|\mathcal{Q}_{\boldsymbol{xy}}\big|^{2p-2}+\sum_{|\boldsymbol{l}|+\sum(J\cup J_*)\geq 2}\mathbb{E}\,\Xi(\boldsymbol{l},J,J_*)\big|\mathcal{Q}_{\boldsymbol{xy}}\big|^{2p-1-|J\cup J_*|}\\&+\mathcal{O}_{\prec}\big(\mathcal{E}_{\mathrm{iso}}^{-2p}\big),\end{aligned} \tag{4.146}$$

where the summation in (4.146) is taken over tuples $\boldsymbol{l}\in\mathbb{Z}^2_{\geq 0}$ and multisets of tuples $J,J_*\subset\mathbb{Z}^2_{\geq 0}\setminus\{(0,0)\}$, for which we set $\partial^{(l_1,l_2)}:=\partial_{ab}^{l_1}\partial_{ba}^{l_2}$, $|(l_1,l_2)|=l_1+l_2$, $\sum J=\sum_{\boldsymbol{j}\in J}|\boldsymbol{j}|$. Moreover, we denoted

$$\begin{aligned}\widetilde{\Xi}:=&\frac{\Big|\big(G_1B'G_1BG_2\big)_{\boldsymbol{xy}}\big(G_1G_2\big)_{\boldsymbol{xy}}\Big|+\Big|\big(G_1B'G_2\big)_{\boldsymbol{xy}}\big(G_1BG_2G_2\big)_{\boldsymbol{xy}}\Big|}{N}\\&+\frac{\Big|\big(G_1B'G_2^*B^*G_1^*\big)_{\boldsymbol{xx}}\big(G_2^*G_2\big)_{\boldsymbol{yy}}\Big|+\Big|\big(G_1B'G_1^*\big)_{\boldsymbol{xx}}\big(G_2^*A^*G_1^*G_2\big)_{\boldsymbol{yy}}\Big|}{N},\end{aligned} \tag{4.147}$$

and defined $\Xi(\boldsymbol{l},J,J_*)$ via

$$\begin{aligned}\Xi:=N^{-(|\boldsymbol{l}|+\sum(J\cup J_*)+1)/2}\sum_{ab}&\Big|\partial^{\boldsymbol{l}}\Big[(G_1B')_{\boldsymbol{x}a}\big(G_2\big)_{b\boldsymbol{y}}\Big]\Big|\\&\times\prod_{\boldsymbol{j}\in J}\Big|\partial^{\boldsymbol{j}}\big(G_1BG_2\big)_{\boldsymbol{xy}}\Big|\prod_{\boldsymbol{j}\in J_*}\Big|\partial^{\boldsymbol{j}}\big(G_2^*B^*G_2^*\big)_{\boldsymbol{yx}}\Big|.\end{aligned} \tag{4.148}$$

Now, by a simple norm bound, $\|G\|\lesssim 1$, we find that

$$\widetilde{\Xi}\prec\mathcal{E}_{\mathrm{iso}}^2\,. \tag{4.149}$$

For Ξ, our goal is to show that

$$\Xi(\boldsymbol{l},J,J_*)\prec\mathcal{E}_{\mathrm{iso}}^{1+|J\cup J_*|}\,. \tag{4.150}$$

First, we have the naive bounds

$$\Big|\partial^{\boldsymbol{l}}\Big[(G_1B')_{\boldsymbol{x}a}\big(G_2\big)_{b\boldsymbol{y}}\Big]\Big|+\Big|\partial^{\boldsymbol{j}}\big(G_1BG_2\big)_{\boldsymbol{xy}}\Big|+\Big|\partial^{\boldsymbol{j}}\big(G_2^*B^*G_2^*\big)_{\boldsymbol{yx}}\Big|\prec 1 \tag{4.151}$$

and hence

$$\Xi\prec N^{-(|\boldsymbol{l}|+\sum(J\cup J_*)+1)/2}N^2=N^{(4-|\boldsymbol{l}|)/2}N^{-(1+\sum(J\cup J_*))/2}=N^{(4-|\boldsymbol{l}|)/2}\mathcal{E}_{\mathrm{iso}}^{1+\sum(J\cup J_*)}\,.$$

Thus, for $|\boldsymbol{l}| \geq 4$, we find the naive bounds (4.151) to be sufficient for (4.150), since trivially $|J \cup J_*| \leq \sum(J \cup J_*)$. For $|\boldsymbol{l}| \leq 3$, we perform the summation $\sum_{ab}$ a bit more carefully, e.g., recalling the norm bound $\|G\| \lesssim 1$, as

$$\sum_a |G_{\boldsymbol{x}a}| \leq N^{1/2} \sqrt{\sum_a |G_{\boldsymbol{x}a}|^2} = N^{1/2} (|G|^2)_{\boldsymbol{x}\boldsymbol{x}} \lesssim N^{1/2}$$

instead of the naive $\sum_a |G_{\boldsymbol{x}a}| \lesssim N$. Indeed, using the condition $|\boldsymbol{l}| + \sum(J \cup J_*) \geq 2$, we can check all the cases $|\boldsymbol{l}| \leq 3$ explicitly and find

$$\sum_{ab} \left| \partial^{\boldsymbol{l}} \left[(G_1 B')_{\boldsymbol{x}a} \left(G_2\right)_{b\boldsymbol{y}} \right] \right| \prod_{\boldsymbol{j} \in J} \left| \partial^{\boldsymbol{j}} \left(G_1 B G_2 \right)_{\boldsymbol{x}\boldsymbol{y}} \right| \prod_{\boldsymbol{j} \in J_*} \left| \partial^{\boldsymbol{j}} \left(G_2^* B^* G_2^* \right)_{\boldsymbol{y}\boldsymbol{x}} \right| \prec N^2 \, N^{-(4-|\boldsymbol{l}|)/2} \,,$$

from which we conclude (4.150).

Plugging (4.150) together with (4.149) into (4.146), using a Young inequality and recalling that p was arbitrary, we deduce that

$$|\mathcal{Q}_{\boldsymbol{x}\boldsymbol{y}}| \prec \mathcal{E}_{\mathrm{iso}} = \frac{1}{\sqrt{N}} \,,$$

i.e. we have proven Proposition 4.3.1. □

It remains to give the proof of Lemma 4.A.3.

Proof of Lemma 4.A.3. Applying (4.144) to G_2, we thus find that

$$G_1 \widetilde{B} G_2 = G_1 \widetilde{B} M_2 - G_1 \widetilde{B} M_2 \underline{W G_2} + G_1 \widetilde{B} M_2 \langle\!\langle G_2 - M_2 \rangle\!\rangle G_2$$

for $\widetilde{B} = \mathcal{X}_{12}[B]$, where we introduced the linear operator

$$\mathcal{X}_{12}[C] := \left(1 - \langle\!\langle M_1 \, \cdot \, M_2 \rangle\!\rangle\right)^{-1}[C] \quad \text{for} \quad C \in \mathbb{C}^{2N \times 2N}. \tag{4.152}$$

Extending the underline to the whole product, we obtain

$$\begin{aligned} G_1 \widetilde{B} G_2 =& M_1 \widetilde{B} M_2 + (G_1 - M_1) \widetilde{B} M_2 - \underline{G_1 \widetilde{B} M_2 W G_2} \\ &+ G_1 \widetilde{B} M_2 \langle\!\langle G_2 - M_2 \rangle\!\rangle G_2 + G_1 \langle\!\langle G_1 \widetilde{B} M_2 \rangle\!\rangle G_2, \end{aligned}$$

from which we conclude

$$\begin{aligned} G_1 B G_2 =& M_1 \mathcal{X}_{12}[B] M_2 + (G_1 - M_1) \mathcal{X}_{12}[B] M_2 - \underline{G_1 \mathcal{X}_{12}[B] M_2 W G_2} \\ &+ G_1 \mathcal{X}_{12}[B] M_2 \langle\!\langle G_2 - M_2 \rangle\!\rangle G_2 + G_1 \langle\!\langle (G_1 - M_1) \mathcal{X}_{12}[B] M_2 \rangle\!\rangle G_2. \end{aligned} \tag{4.153}$$

To proceed, we note that

$$B' := \mathcal{X}_{12}[B] = B + \frac{\langle\!\langle M_1 B M_2 \rangle\!\rangle}{1 - \langle\!\langle M_1 M_2 \rangle\!\rangle}$$

has bounded norm, $\|B'\| \lesssim 1$, since for $|\Im z_1|, |\Im z_2| \gtrsim 1$, $\mathcal{X}_{12}$ is a bounded operator (see [34, Lemma B.5] and [33, Appendix A.2]). Then, using the norm bounds $\|G_i\| \lesssim 1$ and $\|M_i\| \lesssim 1$ together with the *single resolvent global law* (4.40) (see also [69, Theorem 2.1]) for the second, fourth and fifth term in (4.153), we conclude the desired. □

5 Operator Norm of a Hermitian Random Matrix with Correlated Entries

We consider a correlated $N \times N$ Hermitian random matrix with a polynomially decaying metric correlation structure. By calculating the trace of the moments of the matrix and using the summable decay of the cumulants, we show that its operator norm is stochastically dominated by one.

Published as J. Reker. On the operator norm of a Hermitian random matrix with correlated entries. *Random Matrices: Theory Appl.*, 11(4):2250036, 2022.

5.1 Introduction

Let H be a Hermitian $N \times N$ random matrix such that $H = \frac{1}{\sqrt{N}}W$, where $W \in \mathbb{C}^{N\times N}$ has matrix elements of order one. For Wigner matrices, i.e., when the entries of W are identically distributed and independent (up to the Hermitian symmetry) with some mild moment condition, it is well-known that $\|H\|$ is bounded uniformly in N with very high probability. In fact, it even converges to 2 under the normalization $\mathbb{E}|W_{ij}|^2 = 1$ (see [13] and [8, Thm 2.1.22], as well as [81], [94], [139] for more quantitative bounds under stronger moment conditions). Similar statements hold true for Wigner-type ensembles which allow different distributions of the matrix entries and a more general variance profile, but retain independence up to the Hermitian symmetry. Here, the operator norm converges to the maximum of the support of the asymptotic density of states (see [63], [117]), which, although possibly different from 2, is deterministic as well. In contrast, if the entries of W are very strongly correlated, the norm of H may be as large as $\sqrt{N}$. Considering dependent random variables furher adds the technical difficulty of tracking the correlations between the matrix entries explicitly throughout the analysis. The spectral properties of correlated random matrices have been studied through global laws (see, e.g., [9], [16], [50], [86], [70], [127] for the discussion of different models and correlation structures), local laws and universality results (see [2], [7],[69]).

In this paper, we assume the entries of W to be correlated following a *polynomially decaying metric correlation structure* as, e.g., considered in [69], but with a weaker, merely summable correlation decay. This dependence structure is characterized by the 2-cumulants of the matrix elements W_{ij} decaying at least as an inverse $2+\epsilon$ power of the distance with respect to a natural metric on the index pairs (i, j). The higher cumulants follow a similar pattern (see Assumption (A3) below). Under these mild decay conditions, we show that $\|H\|$ is essentially bounded with very high probability.

This result was already stated in [69], indicating that an extension of Wigner's moment method applies. In the current paper, we carry out this task which turns out to be rather involved. In [69], this bound was used as an apriori control on $\|H\|$ for the resolvent method, leading to optimal local laws for H. We remark that it is possible the modify

the proof in [69] to obtain the bound on $\|H\|$ directly, i.e., without relying on the current paper. However, an independent proof via the moment method has several advantages. First, it is conceptually much simpler and less technical than the resolvent approach in [69]. Further, it only requires the summability of the 2-cumulants (see exponent $s > 2$ in (5.2) below), while [69] assumed a faster decay ($s > 12$, see [69, Eq. (3a)]). Lastly, the current method can be generalized to even weaker correlation decays, resulting in correlated random matrix ensembles whose norm grows with N but slower than the trivial $\sqrt{N}$ bound.

We start by giving some general notation and the precise assumptions on the matrix W in Section 5.1.1 below. The bound on the operator norm of H is then formulated in Theorem 5.1.5 and its proof is given in Sections 5.2 and 5.3. For simplicity, the argument is carried out only for symmetric $H \in \mathbb{R}^{N\times N}$. The Hermitian case follows analogously and is hence omitted.

Acknowledgment: I am very grateful to László Erdős for suggesting the topic and supervising my work on this project. Partially supported by ERC Advanced Grant "RMTBeyond" No. 101020331.

5.1.1 Notation and Assumptions on the Model

Throughout the paper, boldface indicates vectors $\mathbf{x} \in \mathbb{C}^N$ and their Euclidean norm is denoted by $\|\mathbf{x}\|_2$. Further, the operator norm of a matrix $A \in \mathbb{C}^{N\times N}$ is denoted by $\|A\|$. In the estimates, C (without subscript) denotes a generic constant the value of which may change from line to line. We note the following assumptions on the matrix W.

Assumption 5.1.1 (A1)**.** $\mathbb{E}W_{i,j} = 0$ *for all* $i, j = 1, \dots, N$.

Assumption 5.1.2 (A2)**.** *For all* $q \in \mathbb{N}$ *there exists a constant* μ_q *such that* $\mathbb{E}|W_{i,j}|^q \le \mu_q$ *for all* $i, j = 1, \dots, N$.

The assumption on the correlation decay is given in terms of the multivariate cumulants $\kappa^{(k)}$ of the matrix elements.

Definition 5.1.3 (Cumulants)**.** *Let* $\mathbf{w} = (w_1, \dots, w_n)$ *be a random vector taking values in* $\mathbb{R}^n$. *The* ***cumulants*** κ_m *of* $\mathbf{w}$ *are defined as the Taylor coefficients of the log-characteristic function of* $\mathbf{w}$*, i.e.,*

$$\ln \mathbb{E}[\mathrm{e}^{i\mathbf{t}\cdot\mathbf{w}}] = \sum_m \kappa_m \frac{(i\mathbf{t})^m}{m!},$$

where the sum is taken over all multi-indices $m = (m_1, \dots, m_n) \in \mathbb{N}^n$ *and* $m! = \prod_{j=1}^n (m_j!)$. *For a multiset* $B \subset \{1, \dots, n\}$ *with* $|B| = k$*, we also write* $\kappa^{(k)}(w_j | j \in B)$ *instead of* κ_m*, where* m_i *is the multiplicity of* $i \in B$.

The cumulants of $\mathbf{w}$ satisfy the moment-cumulant relation

$$\mathbb{E}(w_1 \dots w_n) = \sum_{\pi\in\Pi_n} \prod_{B\in\pi} \kappa^{(|B|)}(w_j | j \in B), \tag{5.1}$$

where Π_n denotes the set of partitions of $\{1, \dots, n\}$ (see, e.g., [111]). The complex cumulants arising whenever $\mathbf{w}$ takes values in $\mathbb{C}^n$ can be reduced to the above definition by considering the real and imaginary part of the random variables separately. In particular,

the complex m-cumulant of $w_{j_1}, \dots, w_{j_m}$ is uniquely determined as a linear combination of the m-cumulants of the $2m$ random variables $\Re(w_{j_1}), \Im(w_{j_1}), \dots, \Re(w_{j_m}), \Im(w_{j_m})$. Writing $w_j = x_j + iy_j$ with $j = 1, 2$, we have, e.g.,

$$\kappa^{(2)}(w_1, w_2) = \kappa^{(2)}(x_1, x_2) + i\kappa^{(2)}(x_1, y_2) - i\kappa^{(2)}(y_1, x_2) - \kappa^{(2)}(y_1, y_2).$$

To keep the notation short, we usually view the cumulants as a function of the indices of the matrix elements by identifying $\kappa^{(k)}(W_{a_1,a_2}, \dots)$ with $\kappa^{(k)}(a_1a_2, \dots)$ or $\kappa^{(k)}(\alpha_1, \dots, \alpha_k)$ using $\alpha_1, \dots, \alpha_k \in \{1, \dots, N\}^2$. Further, d denotes the Euclidean distance on $\{1, \dots, N\}^2$ modulo the (Hermitian) symmetry, i.e.,

$$d(a_1a_2, a_3a_4) := \min\{|a_1 - a_3| + |a_2 - a_4|, |a_1 - a_4| + |a_2 - a_3|\}.$$

In this notation, the conditions from the polynomially decaying metric correlation structure can be formulated as follows.

Assumption 5.1.4 (A3)**.** *Whenever W is a real random matrix, the k-cumulants $\kappa^{(k)}$ of the elements of W satisfy*

$$|\kappa^{(2)}(a_1a_2, a_3a_4)| \le \frac{C_{\kappa,2}}{1 + d(a_1a_2, a_3a_4)^s}, \tag{5.2}$$

$$|\kappa^{(k)}(\alpha_1, \dots, \alpha_k)| \le C_{\kappa,k} \prod_{e \in T_{min}} |\kappa^{(2)}(e)|, \ k \ge 3, \tag{5.3}$$

for $s > 2$ and some constants $C_{\kappa,k} > 0$ for $k \ge 2$. Here, T_{min} is the (any) minimal spanning tree[1] *on the complete graph on k vertices labelled by $\alpha_1, \dots, \alpha_k$ with edge weights $d(\alpha_i, \alpha_j)$. If W is a complex random matrix, we assume the above conditions for the real and imaginary part of W separately in the sense that $\kappa^{(k)}(M_{\alpha_1}, \dots, M_{\alpha_k})$ satisfies* (5.2) *or* (5.3)*, respectively, for any combination of $M_\alpha \in \{\Re(W_\alpha), \Im(W_\alpha)\}$.*

Note that a correlation decay of the form (5.3) arises in different statistical physics models (see [59]).

5.1.2 Statement of the Main Result

With the notation established, we give the statement on the operator norm of H as follows.

Theorem 5.1.5. *Under the assumptions* (A1)-(A3)*, we have that for all $\epsilon > 0$, $D > 0$ there exists a suitable constant $C(\epsilon, D)$ such that, for all $N \in \mathbb{N}$,*

$$\mathbb{P}\Big(\|H\| > N^\epsilon\Big) \le C(\epsilon, D)N^{-D}.$$

Remark. In [69], the bound of Theorem 5.1.5 (for $s > 12$) is used as a priori control for the operator norm in proving optimal local laws for random matrices with slow correlation decay. As a consequence of these results, the eigenvalues of H are asymptotically confined to the support of the self-consistent density of states with high probability. This improves upon the above stochastic domination bound by replacing the N^ϵ factor with a large constant (see [69, Cor. 2.3]).

[1] A spanning tree is a graph that connects all the vertices without cycles. Any spanning tree with the lowest possible sum of edge weights is called minimal.

5.1.3 Setup for the Moment Method

We show that the assumptions (A1)-(A3) imply that

$$\mathbb{E}[\tfrac{1}{N}\operatorname{tr}(H^k)] \leq C(k), \quad \forall k \in \mathbb{N}, \tag{5.4}$$

from which the statement of Theorem 5.1.5 follows by an application of Chebyshev's inequality. As H is Hermitian, we have $\|H\|^k \leq \operatorname{tr}(H^k)$ for all even $k \in \mathbb{N}$. This implies

$$\mathbb{P}(\|H\| > N^\epsilon) \leq \frac{\mathbb{E}[\|H\|^k]}{N^{k\epsilon}} \leq \frac{N\mathbb{E}[\frac{1}{N}\operatorname{tr}(H^k)]}{N^{k\epsilon}} \leq \frac{NC(k)}{N^{k\epsilon}}$$

and thus gives the desired bound if k is chosen large enough.

Expanding the term on the left-hand side of (5.4) using the relation (5.1) yields

$$\begin{aligned}\mathbb{E}[\tfrac{1}{N}\operatorname{tr}(H^k)] &= \frac{1}{N}\sum_{a_1,\dots,a_k}\mathbb{E}[H_{a_1,a_2}H_{a_2,a_3}\dots H_{a_k,a_1}] \\ &= N^{-k/2-1}\sum_{\pi\in\Pi_k}\sum_{a_1,\dots,a_k}\prod_{B\in\pi}\kappa^{(|B|)}(a_j a_{j+1}|j\in B),\end{aligned} \tag{5.5}$$

where Π_k denotes the set of partitions of $\{1,\dots,k\}$ and the index $j+1$ is to be interpreted mod k, i.e., if $j=k$, then $a_j a_{j+1} = a_k a_1$. Observe that all terms involving 1-cumulants vanish due to (A1). Hence, one can restrict the sum to partitions π without singleton sets. The cumulant expansion (5.5) is the main difference between the real symmetric and complex Hermitian case, as considering $H \in \mathbb{C}^{N\times N}$ requires replacing the cumulants by their complex counterparts. However, one can always reduce the argument to the real case by considering the real and imaginary parts of the random variables separately. Thus, from now on we assume H to be real. We develop the estimates by first deriving suitable bounds for the products that only involve 2-cumulants, which include the leading terms, and then successively incorporating higher-order cumulants. The fast correlation decay from Assumption (A3) implies that roughly half of the summations over the indices $a_1,\dots,a_k$ yield a factor N, while the other half can be summed up with an N-independent bound. To quantify this behavior, we introduce the following counting rule, which constitutes a useful property of a term involving $n \in \mathbb{N}_0$ summations.

Counting Rule (CR). A term involving several index summations and a product of cumulants satisfies the counting rule if every independent summation over an index $a_{i_1},\dots,a_{i_n}$ yields a contribution of order $\sqrt{N}$.

Whenever one can show that (CR) holds for a given term, the counting rule allows to replace explicit bounds by cruder power counting arguments, which simplifies calculating the contribution to (5.5). Throughout the proof of (5.4), we, therefore, aim to show this property for as many terms as possible. Note, however, that bounding the leading terms requires an extra power of N compared to the counting rule (see the proof of Lemma 5.2.1 below). For these terms, the factor $N^{-k/2-1}$ in (5.5) has to be canceled out completely.

5.1.4 Visualization Using Graphs

Following the discussion of the previous subsection, we aim to bound

$$N^{-k/2-1}\Big|\sum_{a_1,\dots,a_k}\prod_{B\in\pi}\kappa^{(|B|)}(a_j a_{j+1}|j\in B)\Big| \tag{5.6}$$

for any given $\pi \in \Pi_k$. Terms of this form can be visualized by considering a k-gon, where the vertices are labeled by the indices $a_1, \dots, a_k$ and the edges by the successive double indices $(a_1, a_2), \dots, (a_k, a_1)$. In this picture, every j-cumulant combines j distinct edges such that every edge appears in exactly one of the cumulants in the product. We say that a vertex *belongs to a j-cumulant*, if it is adjacent to one of the edges associated with it. A vertex (resp. the corresponding index) which only occurs in a single cumulant is referred to as *internal vertex* (resp. *internal index*). We define the corresponding graphs as follows.

Definition 5.1.6. *The graph given by the regular k-gon with vertices $a_1, \dots, a_k$ is denoted as Γ_k. For any $\pi \in \Pi_k$, we define $\Gamma_k(\pi)$ as the colored graph in which the j edges that belong to the same j-cumulant are assigned the same color.*

Examples of this visualization are given in Fig. 5.1 and Fig. 5.2 below, where the colors are indicated by different linestyles. For later applications of algorithms on the graph, we further introduce an ordering. Assume first that π is a pairing, i.e., $\Gamma_k(\pi)$ has exactly two edges of each color. Let $E_k = \{(a_1, a_2), \dots, (a_k, a_1)\}$ denote the (ordered) set of edges and $\varphi : E_k \to E_k$ be the function that maps every edge to the other one of the same color. Starting with $e_1 = (a_1, a_2)$, we go through the elements of E_k and note the pairing $(e_n, \varphi(e_n))$ whenever $e_n \neq \varphi(e_i)$ for all $i < n$ to obtain $k/2$ pairs of edges. We denote the pairings as an ordered set $C(\pi) := \{(e_1, \varphi(e_1)), \dots, (e_{k/2}, \varphi(e_{k/2}))\}$ and refer to the edges in $C(\pi)$ as *paired edges*. Similarly, we can extract a set $C(\pi)$ of j-tuples, $j \geq 2$, for any partition $\pi \in \Pi_k$.

Definition 5.1.7. *For a partition $\pi \in \Pi_k$, we denote the graph with the (ordered) pairing structure induced by π by the pair $(\Gamma_k, C(\pi))$.*

Note that the above construction induces a one-to-one correspondence between cumulant terms of the form (5.6) and colored k-gons. In particular, both $\Gamma_k(\pi)$ and $(\Gamma_k, C(\pi))$ contain all information on the corresponding term in (5.6).

5.2 Proof of (5.4) for Terms Involving Only 2-Cumulants

In this section, we focus on the following special case.

Lemma 5.2.1. *Let $k \in \mathbb{N}$ be even and $\pi \in \Pi_k$ with $|B| = 2$ for all $B \in \pi$. Then,*

$$N^{-k/2-1} \Big| \sum_{a_1, \dots, a_k} \prod_{B \in \pi} \kappa^{(2)}(a_j a_{j+1} | j \in B) \Big| \leq C(k). \tag{5.7}$$

Assume first that π is chosen such that the 2-cumulants occurring in the term do not involve internal indices. We note the following general estimates whose proofs are elementary from (5.2) and the fact that $s > 2$.

Lemma 5.2.2. *Assume that* (A3) *holds. Then*

$$\sum_{a_1} |\kappa^{(2)}(a_1 a_2, a_3 a_4)| \leq C, \quad \sum_{a_1, a_2} |\kappa^{(2)}(a_1 a_2, a_3 a_4)| \leq C, \quad \sum_{a_1, a_3} |\kappa^{(2)}(a_1 a_2, a_3 a_4)| \leq CN,$$
$$\sum_{a_1, a_2, a_3} |\kappa^{(2)}(a_1 a_2, a_3 a_4)| \leq CN, \quad \sum_{a_1, \dots, a_4} |\kappa^{(2)}(a_1 a_2, a_3 a_4)| \leq CN^2$$

uniformly for any choice of the unsummed indices.

Note that all estimates in Lemma 5.2.2 follow (CR), some of them are even stronger. However, the summation over an internal index would, in general, not obey this counting rule, since $\sum_{a_2=1}^{N}|\kappa^{(2)}(a_1a_2, a_2a_3)|$ may be of order N if $a_1 = a_3$. This is the reason why internal indices are treated separately. The key to estimating (5.7) in the given case is a recursive summation procedure. We demonstrate the approach for the term visualized in Fig. 5.1 below, where different linestyles indicate the edges associated with the same 2-cumulant.

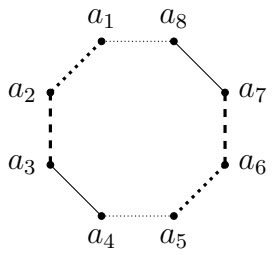

Fig. 5.1. Visualization of $\kappa^{(2)}(a_1a_2, a_5a_6)\kappa^{(2)}(a_2a_3, a_6a_7)\kappa^{(2)}(a_3a_4, a_7a_8)\kappa^{(2)}(a_4a_5, a_8a_1)$

Example 5.2.3. To start the summation, estimate

$$\begin{aligned} S &:= N^{-5}\sum_{a_1,\dots,a_8}|\kappa^{(2)}(a_1a_2, a_5a_6)\kappa^{(2)}(a_2a_3, a_6a_7)\kappa^{(2)}(a_3a_4, a_7a_8)\kappa^{(2)}(a_4a_5, a_8a_1)| \\ &\leq N^{-5}\sum_{a_1,\dots,a_8,a_1'}|\kappa^{(2)}(a_1a_2, a_5a_6)\kappa^{(2)}(a_2a_3, a_6a_7)\kappa^{(2)}(a_3a_4, a_7a_8)\kappa^{(2)}(a_4a_5, a_8a_1')|. \quad (5.8) \end{aligned}$$

Adding the extra summation label may a_1' appear as an unnecessary overestimate, but it gives a dedicated start and end point for the summation algorithm by breaking the cyclic structure of the graph. Next, isolate the 2-cumulant involving a_1 by taking the maximum over the remaining indices a_2, a_5, a_6 for the other factors. Together with a_1, we sum over all labels appearing in $\kappa^{(2)}(a_1a_2, a_5a_6)$. Note that, by Lemma 5.2.2, the counting rule (CR) applies to this sum, giving a factor of $(\sqrt{N})^4 = N^2$ and the estimate

$$S \leq N^{-5}CN^2\max_{a_2,a_5,a_6}\Big(\sum_{a_3,a_4,a_7,a_8,a_1'}|\kappa^{(2)}(a_2a_3, a_6a_7)\kappa^{(2)}(a_3a_4, a_7a_8)\kappa^{(2)}(a_4a_5, a_8a_1')|\Big).$$

From a_1, continue counter-clockwise along the octagon to find the next index to sum over, i.e., a_3. Identifying and isolating $\kappa^{(2)}(a_2a_3, a_6a_7)$, sum over all remaining indices in the factor, i.e., a_3 and a_7. Again, Lemma 5.2.2 justifies the use of (CR), giving a contribution of $(\sqrt{N})^2 = N$ and the estimate

$$S \leq CN^{-2}\max_{a_3,a_5,a_7}\Big(\sum_{a_4,a_8,a_1'}|\kappa^{(2)}(a_3a_4, a_7a_8)\kappa^{(2)}(a_4a_5, a_8a_1')|\Big).$$

Repeating the previous step, continuing along the octagon yields a_4 as the next index. Isolating $\kappa^{(2)}(a_3a_4, a_7a_8)$, we apply Lemma 5.2.2 to perform the summations over a_4 and a_8 using (CR). This yields a factor of $(\sqrt{N})^2 = N$ and

$$S \leq CN^{-1}\max_{a_4,a_5,a_8}\Big(\sum_{a_1'}|\kappa^{(2)}(a_4a_5, a_8a_1')|\Big).$$

Lastly, the summation over $a_{1'}$ in $|\kappa^{(2)}(a_4a_5, a_8a_1')|$ remains. By applying the counting rule again, we obtain a final bound of order $N^{-1/2}$, which can be estimated by a constant as

claimed in Lemma 5.2.1. In particular, the pairing in Fig. 1 gives a subleading contribution to (5.5). Note that introducing an additional index in the first step is affordable, as are any overestimates from using (CR) over the explicit estimates in Lemma 5.2.2. The cruder bounds allow to carry out every step but the initial one following the same pattern, which reduces the final bound to a counting argument.

As this recursive summation procedure relies on Lemma 5.2.2, it cannot be applied directly if summation over internal indices are present. To prepare for the general case, we extend Lemma 5.2.2 to more general objects.

Definition 5.2.4. *For* $\mathbf{x} \in \mathbb{R}^N$*, set*

$$\kappa^{(2)}(\mathbf{x}a_2, a_3a_4) := \sum_{a_1=1}^{N} \kappa^{(2)}(a_1a_2, a_3a_4)x_{a_1}$$

and define any 2-cumulants with one or more indices replaced by a vector analogously.

We collect some estimates for 2-cumulants of this form below, where we follow the convention that vectors only occur in place of the first index of an index pair and only replace unsummed indices. However, the same bounds hold if the second index of the respective pair is replaced instead. As in Lemma 5.2.2, no internal index is summed up.

Lemma 5.2.5. *Assume that* (A3) *holds and let* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$*. Then we have*

$$\begin{aligned}
&\sum_{a_2} |\kappa^{(2)}(\mathbf{x}a_2, a_3a_4)| \leq C\|\mathbf{x}\|_2, \quad \sum_{a_3} |\kappa^{(2)}(\mathbf{x}a_2, a_3a_4)| \leq CN^{1/2}\|\mathbf{x}\|_2,\\
&\sum_{a_2,a_3} |\kappa^{(2)}(\mathbf{x}a_2, a_3a_4)| \leq CN\|\mathbf{x}\|_2, \quad \sum_{a_3,a_4} |\kappa^{(2)}(\mathbf{x}a_2, a_3a_4)| \leq CN\|\mathbf{x}\|_2,\\
&\sum_{a_2,a_3,a_4} |\kappa^{(2)}(\mathbf{x}a_2, a_3a_4)| \leq CN^{3/2}\|\mathbf{x}\|_2
\end{aligned}$$

uniformly for any choice of unsummed indices. Similar bounds hold with two vectors, i.e.,

$$\begin{aligned}
&|\kappa^{(2)}(\mathbf{x}a_2, \mathbf{y}a_4)| \leq C\|\mathbf{x}\|_2\|\mathbf{y}\|_2, \quad \sum_{a_2} |\kappa^{(2)}(\mathbf{x}a_2, \mathbf{y}a_4)| \leq CN^{1/2}\|\mathbf{x}\|_2\|\mathbf{y}\|_2,\\
&\sum_{a_2,a_4} |\kappa^{(2)}(\mathbf{x}a_2, \mathbf{y}a_4)| \leq CN\|\mathbf{x}\|_2\|\mathbf{y}\|_2.
\end{aligned}$$

In particular, the estimates follow (CR).

Proof. Noting that $\max_j |x_j| \leq \|\mathbf{x}\|_2$, the bounds in Lemma 5.2.2 imply that

$$\sum_{a_2} |\kappa^{(2)}(\mathbf{x}a_2, a_3a_4)| \leq \|\mathbf{x}\|_2 \sum_{a_1,a_2} |\kappa^{(2)}(a_1a_2, a_3a_4)| \leq C\|\mathbf{x}\|_2.$$

If the summation over a_2 is replaced by a summation over a_3 or a_4, we obtain a bound of order N instead. Further, an application of the Cauchy-Schwarz inequality leads to

$$\sum_{a_3} |\kappa^{(2)}(\mathbf{x}a_2, a_3a_4)| \leq \|\mathbf{x}\|_2 \sqrt{\sum_{a_1} \Big(\sum_{a_3} |\kappa^{(2)}(a_1a_2, a_3a_4)| \Big)^2} \leq CN^{1/2}\|\mathbf{x}\|_2. \tag{5.9}$$

Recalling that the summation over three indices gives a factor N, the bound for summation over all three indices in $|\kappa^{(2)}(\mathbf{x}a_2, a_3a_4)|$ follows analogously.

For 2-cumulants that involve two vectors, applying (5.2) yields

$$|\kappa^{(2)}(\mathbf{x}a_2, \mathbf{y}a_4)| \leq \sum_{a_1,a_3} \Big(\frac{C_\kappa |x_{a_1} y_{a_3}|}{1 + |a_1 - a_3|^s + |a_2 - a_4|^s} + \frac{C_\kappa |x_{a_1} y_{a_3}|}{1 + |a_1 - a_4|^s + |a_2 - a_3|^s} \Big). \quad (5.10)$$

Set $\varepsilon = \|\mathbf{y}\|_2 / \|\mathbf{x}\|_2$ and estimate the first term as

$$\sum_{a_1,a_3} \frac{|x_{a_1} y_{a_3}|}{1 + |a_1 - a_3|^s + |a_2 - a_4|^s} \leq \sum_{a_1,a_3} \frac{\epsilon |x_{a_1}|^2 + \epsilon^{-1} |y_{a_3}|^2}{1 + |a_1 - a_3|^s} \leq C(\epsilon \|\mathbf{x}\|_2^2 + \epsilon^{-1} \|\mathbf{y}\|_2^2)$$

to obtain an N-independent bound. The estimate of the second term is similar.

Adding a summation over one index, e.g., a_2, two applications of the Cauchy-Schwarz inequality and the third estimate of Lemma 5.2.2 yield

$$\begin{aligned} \sum_{a_2} |\kappa^{(2)}(\mathbf{x}a_2, \mathbf{y}a_4)| &\leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \sqrt{\sum_{a_3} \Big(\sum_{a_1} \sum_{a_2} |\kappa^{(2)}(a_1 a_2, a_3 a_4)| \Big)^2} \\ &\leq C N^{1/2} \|\mathbf{x}\|_2 \|\mathbf{y}\|_2. \end{aligned} \quad (5.11)$$

Finally, it follows

$$\sum_{a_2,a_4} |\kappa^{(2)}(\mathbf{x}a_2, \mathbf{y}a_4)| \leq \sum_{a_1,\dots,a_4} \Big(\frac{C_\kappa |x_{a_1} y_{a_3}|}{1 + |a_1 - a_3|^s + |a_2 - a_4|^s} + \frac{C_\kappa |x_{a_1} y_{a_3}|}{1 + |a_1 - a_4|^s + |a_2 - a_3|^s} \Big).$$

Here, we obtain

$$\sum_{a_1,\dots,a_4=1}^{N} \frac{|x_{a_1} y_{a_3}|}{1 + |a_1 - a_3|^s + |a_2 - a_4|^s} \leq CN \sum_{a_1,a_3=1}^{N} \frac{|x_{a_1} y_{a_3}|}{1 + |a_1 - a_3|^{s-1}} \leq CN \|\mathbf{x}\|_2 \|\mathbf{y}\|_2,$$

and the estimate for the second term is similar. □

Lemma 5.2.5 is the key tool for estimating terms that include summation over internal indices. Consider the matrix $T \in \mathbb{R}^{N \times N}$ defined by its matrix elements

$$T_{a_1,a_3} := \sum_{a_2} T^{(a_2)}_{a_1,a_3} := \sum_{a_2} \kappa^{(2)}(a_1 a_2, a_2 a_3) \quad (5.12)$$

and observe that $|T_{a_1,a_3}| \leq C_\kappa N$ by (5.2), but also $\|T\| \leq CN$, since

$$\|T^{(a_2)} \mathbf{x}\|_2^2 \leq \sum_{a_1} \Big(\sum_{a_3} |\kappa^{(2)}(a_1 a_2, a_2 a_3) x_{a_3}| \Big)^2 \leq \Big(\sum_{a_1} \sum_{a_3} |\kappa^{(2)}(a_1 a_2, a_2 a_3)|^2 \Big) \|\mathbf{x}\|_2^2 \leq C \|\mathbf{x}\|_2^2$$

for $\mathbf{x} \in \mathbb{R}^N$, and $\|T\| \leq \sum_{a_2} \|T^{(a_2)}\|$. Next, we derive similar estimates for the matrix $T^{[j]} \in \mathbb{R}^{N \times N}$ defined for $2 \leq j \leq k-1$ by

$$T^{[j]}_{a_1,a_{2j+1}} := \sum_{a_2,a_{2j}} \kappa^{(2)}(a_1 a_2, a_{2j} a_{2j+1}) T^{j-1}_{a_2,a_{2j}}. \quad (5.13)$$

Note that the superscript corresponds to the total number of 2-cumulants that are rewritten to obtain $T^{[j]}_{a_1,a_{2j+1}}$. Again, we have $|T^{[j]}_{a_1,a_{2j+1}}| \leq CN^j$ by a direct estimate, but also

$$\|T^{[j]}\| \leq CN^j \quad \forall j \in \{2, \dots, k-1\}. \quad (5.14)$$

As the argument is the same in the general case, we consider only $j = 2$. Define the vector $\mathbf{y}^{a_4}$ through $(\mathbf{y}^{a_4})_{a_2} := N^{-1}T_{a_2,a_4}$, $a_2 = 1, \dots, N$ and let

$$T^{[2],a_4} := \sum_{a_2} \kappa^{(2)}(a_1a_2, a_4a_5)T_{a_2,a_4}.$$

Observing that $\|\mathbf{y}^{a_4}\|_2 \leq C$ uniformly in a_4, it follows that

$$\|T^{[2],a_4}\mathbf{x}\|_2^2 \leq \left(N|\kappa^{(2)}(a_1\mathbf{y}^{a_4}, a_4\mathbf{x})|\right)^2 \leq CN^2\|\mathbf{y}^{a_4}\|_2^2\|\mathbf{x}\|_2^2 \leq CN^2\|\mathbf{x}\|_2^2 \tag{5.15}$$

for $\mathbf{x} \in \mathbb{R}^N$, which implies $\|T^{[2]}\| \leq N \max_{a_4} \|T^{[2],a_4}\| \leq CN^2$. Note that the structure of the matrix-vector multiplication in (5.15) does not allow for the usual convention of the vector occurring only as the first index of an index pair.

We demonstrate the approach for treating general products of 2-cumulants for the terms visualized in Fig. 5.2 below.

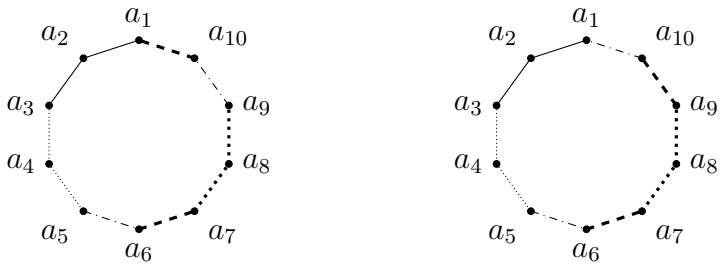

Fig. 5.2. A crossing (left) and a non-crossing pairing (right) for $k = 10$.

Example 5.2.6. First, consider the term on the left of Fig. 2. Recalling the definition of the matrix T from (5.12), rewrite the summation over the internal indices a_2, a_4 and a_7 as

$$\begin{aligned}
&N^{-6} \sum_{a_1,\dots,a_{10}} \kappa^{(2)}(a_1a_2, a_2a_3)\kappa^{(2)}(a_3a_4, a_4a_5)\kappa^{(2)}(a_5a_6, a_9a_{10})\kappa^{(2)}(a_6a_7, a_{10}a_1)\kappa^{(2)}(a_7a_8, a_8a_9) \\
&= N^{-6} \sum_{a_1,a_5,a_6,a_7,a_9,a_{10}} T^2_{a_1,a_5}\kappa^{(2)}(a_5a_6, a_9a_{10})\kappa^{(2)}(a_6a_7, a_{10}a_1)T_{a_7,a_9} \\
&= N^{-3} \sum_{a_1,a_6,a_7,a_{10}} \kappa^{(2)}(\mathbf{x}^{a_1}a_6, \mathbf{y}^{a_7}a_{10})\kappa^{(2)}(a_6a_7, a_{10}a_1),
\end{aligned} \tag{5.16}$$

where the vectors $\mathbf{x}^{a_1}, \mathbf{y}^{a_7} \in \mathbb{R}^N$ in the last step are given by

$$x^{a_1}_{a_5} = \frac{1}{N^2}T^2_{a_1,a_5},\ a_5 = 1, \dots, N, \quad y^{a_7}_{a_9} = \frac{1}{N}T_{a_7,a_9},\ a_9 = 1, \dots, N.$$

To keep the notation consistent with the proof in the general case and Lemma 5.2.5, we introduce the convention that vectors obtained from matrix elements are always defined via the rows of the respective matrix. Recalling that $\|T\| \leq CN$, we have $\|\mathbf{x}^{a_1}\|_2, \|\mathbf{y}^{a_7}\|_2 \leq C$ uniformly for any choice of a_1 and a_7, respectively. Note that one factor of N per power of T is written in front of the sum and that the convention chosen for the vectors ensures that we always replace the first index of an index pair. Further, the sum obtained from (5.16) does not involve summation over internal indices. Modifying the recursive summation procedure from Example 5.2.3 by also taking the maximum

over a_1 in the 2-cumulant involving $\mathbf{x}^{a_1}$ instead of introducing the additional summation label a_1', and using the bounds from Lemmas 5.2.2 and 5.2.5 yields

$$\begin{aligned} &N^{-3} \sum_{a_1,a_6,a_7,a_{10}} |\kappa^{(2)}(\mathbf{x}^{a_1}a_6, \mathbf{y}^{a_7}a_{10})\kappa^{(2)}(a_6a_7, a_{10}a_1)| \\ &\leq N^{-3} \max_{a_1,a_6,a_7,a_{10}} |\kappa^{(2)}(\mathbf{x}^{a_1}a_6, \mathbf{y}^{a_7}a_{10})| \sum_{a_1,a_6,a_7,a_{10}} |\kappa^{(2)}(a_6a_7, a_{10}a_1)| \\ &\leq CN^{-3}N^2 \max_{a_1,a_6,a_7,a_{10}} |\kappa^{(2)}(\mathbf{x}^{a_1}a_6, \mathbf{y}^{a_7}a_{10})| \leq CN^{-1} \leq C, \end{aligned} \tag{5.17}$$

showing that the pairing on the left of Fig. 5.2 gives a sub-leading contribution to (5.5).

In contrast, observe that the non-crossing pairing on the right of Fig. 5.2 needs to be handled differently, since treating T as before yields $\kappa^{(2)}(\mathbf{x}^{a_1}a_6, a_{10}a_1)$ and $\kappa^{(2)}(a_6a_7, \mathbf{y}^{a_7}a_{10})$, which cannot be estimated using Lemma 5.2.5 due to the indices a_1 and a_7 appearing twice in the respective 2-cumulants. However, the terms can be rewritten using the matrices $T^{[j]}$ defined in (5.13). Recalling that $\|T^{[j]}\| \leq CN^j$ from (5.14), we obtain

$$\begin{aligned} &N^{-6}\Big| \sum_{a_1,a_5,a_6,a_7,a_9,a_{10}} T^2_{a_1,a_5}\kappa^{(2)}(a_5a_6, a_{10}a_1)\kappa^{(2)}(a_6a_7, a_9a_{10})T_{a_7,a_9}\Big| \\ &= N^{-6}\Big| \sum_{a_6,a_{10}} T^{[3]}_{a_{10},a_6}T^{[2]}_{a_6,a_{10}}\Big| = N^{-6}\Big|\operatorname{tr}\left(T^{[3]}T^{[2]}\right)\Big| \leq CN^{-6}NN^3N^2 = C. \end{aligned}$$

Hence, the term on the left of Fig. 5.2 yields a leading contribution to (5.5).

After all these preparations, we can give a complete proof of Lemma 5.2.1.

5.2.1 Proof of Lemma 5.2.1

Let $k \geq 2$ be even, $\pi \in \Pi_k$ such that $|B| = 2$ for all $B \in \pi$. The proof of Lemma 5.2.1 is structured into two main steps. In the first step (Lemmas 5.2.7 and 5.2.8), we generalize the strategies from Example 5.2.6 to rewrite the term on the left-hand side to a form that is tractable by a recursive summation procedure. We then carry out the required estimates in the second step (Lemmas 5.2.9 and 5.2.10), showing that the final bound is indeed of order one.

Step 1: Rewriting
The aim of the rewriting step is to obtain a term of the form

$$\sum_{a_1,\dots,a_k} \prod_{B\in\pi} \kappa^{(2)}(a_ja_{j+1}|j \in B) = \sum_{b_1,\dots,b_{k'}} \Big(\prod_{B\in\pi'} \kappa^{(2)}(b_jb_{j+1}|j \in B) \prod_{i\in J} M^{(i)}_{b_{i_1},b_{i_2}} \Big) \tag{5.18}$$

where $k' \leq k$, $M^{(i)} \in \mathbb{R}^{N\times N}$, and π' is chosen such that neither of the 2-cumulants on the right-hand side of (5.18) involves internal indices and such that b_{i_1}, b_{i_2} always appear in some 2-cumulant, but never in the same one. Recall that this structure was also obtained in (5.16) of Example 5.2.6. Whenever the left-hand side of (5.7) does not involve summation over internal indices, choosing $k = k'$, $\pi = \pi'$, and $J = \emptyset$ in (5.18) allows to skip the rewriting step (see Example 5.2.3).

Following the discussion in Section 5.1.4, the term on the left-hand side of (5.7) corresponds to the tuple $(\Gamma_k, C(\pi))$. We aim to visualize the right-hand side of (5.18) as well by associating it with a polygon on the vertices $b_1, \dots, b_{k'}$. In this picture, any matrix element $M^{(i)}_{b_{i_1},b_{i_2}}$ corresponds to an edge connecting the vertices b_{i_1}, b_{i_2} and the two edges

associated with the same 2-cumulant are assigned the same color. We further introduce edge weights w to encode bounds of order N^w on the corresponding matrices in the graph for the following estimates. The rewriting procedure to obtain (5.18) thus corresponds to an algorithm on $(\Gamma_k, C(\pi))$, which we introduce in Lemma 5.2.7 below. An example is given in Fig. 5.3.

Lemma 5.2.7. *Let $k \geq 2$ be even, $\pi \in \Pi_k$ be a pairing and $(\Gamma_k, C(\pi))$ the corresponding graph introduced in Section 5.1.4. Then there exists an edge-weighted graph $\widetilde{\Gamma}_{k'}(\pi)$ on $k' \leq k$ vertices with integer weights $w_i \geq 0$ and an ordered set $\widetilde{C}(\pi)$ of pairings such that the following properties hold.*

(a) Edges that appear in $\widetilde{C}(\pi)$ all have zero weight.

(b) Two edges with nonzero weight are never adjacent to each other.

(c) Two adjacent edges in $\widetilde{\Gamma}_{k'}(\pi)$ that both appear in $\widetilde{C}(\pi)$ belong to different pairs.

(d) Whenever two edges are adjacent to an edge assigned a nonzero weight, they appear in $\widetilde{C}(\pi)$, but belong to different pairs.

Proof. We obtain $(\widetilde{\Gamma}_{k'}(\pi), \widetilde{C}(\pi))$ from $(\Gamma_k, C(\pi))$ by assigning all edges of Γ_k the initial weight zero and carrying out the following reduction algorithm.

Step I Check the set $C(\pi)$ for pairs that involve adjacent edges. If there are any, go through them in the order they appear in $C(\pi)$, replace the corresponding edges of Γ_k and their common vertex by a single edge and remove the pair from $C(\pi)$. Any new edges created in this step are assigned the weight one. This ensures (a) and (c).

Step II Check the new graph for any edges of nonzero weight that are adjacent. If there are any, identify the edges that involve the vertex a_l for the smallest value $l \in \{1, \ldots, k\}$, replace the two edges and their common vertex by a single edge and assign it the sum of the weights of the edges that were replaced. This step is repeated until (b) holds. The validity of (a) and (c) is not changed in the process.

Step III Check the new graph for any edges assigned a nonzero weight, say $w_j > 0$, that are adjacent to two edges belonging to the same pair. If there are any, go through the corresponding pairs in the order they appear in $C(\pi)$ and replace the pair and the weighted edge, as well as the two vertices between them by a single edge. After going through $C(\pi)$ once, remove the pairs that were replaced from $C(\pi)$ and assign the new edges the respective weights $w_j + 1$. As this may generate new subgraphs of the same structure, repeat the step until (d) is satisfied. This does not interfere with (a) and (c), but (b) may not hold any more.

Step IV Repeat Steps II and III until neither can be carried out any further. The graph resulting from this procedure satisfies (b) and (d) simultaneously.

Denoting the (weighted) graph resulting from this procedure by $\widetilde{\Gamma}_{k'}(\pi)$ and the collection of pairs that remain in $C(\pi)$ after Step IV by $\widetilde{C}(\pi)$, we obtain a tuple that satisfies (a) to (d). □

Note that the above algorithm removes all subgraphs that correspond to a non-crossing partition of a subset of $\{1, \ldots, k\}$ and replaces them by an edge with nonzero weight each. In particular, $\widetilde{C}(\pi) = \emptyset$ whenever the pairing π is non-crossing. In this case only one vertex connected to itself by an edge (loop) of weight $k/2$ remains, as the weight assigned to an edge reflects the number of pairings removed from $C(\pi)$ in obtaining it. We demonstrate the algorithm for the example given on the right of Fig. 5.2 below. For simplicity, the vertex labels and edge weights that are equal to zero are left out.

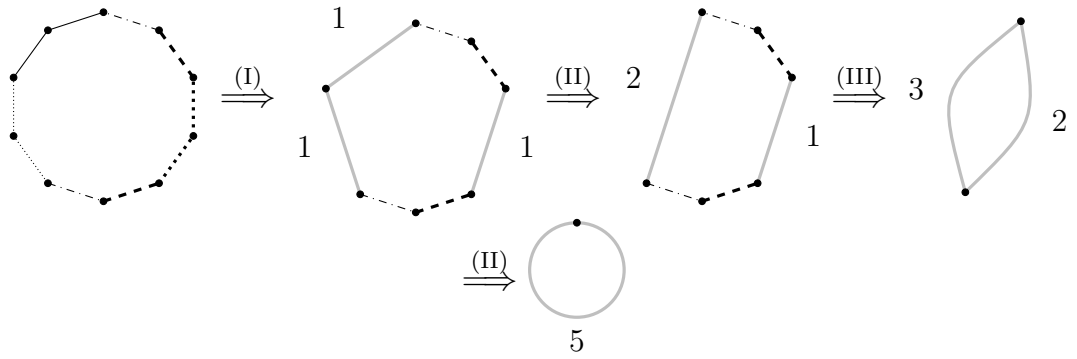

Fig. 5.3. The steps of the reduction algorithm with arrows indicating the steps.

Observe that there is a one-to-one correspondence between the graph $(\widetilde{\Gamma}_{k'}(\pi), \widetilde{C}(\pi))$ obtained in Lemma 5.2.7 and a suitable right-hand side of (5.18). By considering the algebraic counterpart of the graph algorithm, we obtain a bound for the norm of the matrices involved.

Lemma 5.2.8. *Let $k \geq 2$ be even. For every pairing $\pi \in \Pi_k$ exists an expression of the form* (5.18) *with matrices $M^{(i)}$, $i \in J$, that satisfy $\|M^{(i)}\| \leq CN^{w_i}$. The numbers $w_i \geq 1$ correspond to the edge weights of the graph $(\widetilde{\Gamma}_{k'}(\pi), \widetilde{C}(\pi))$ that satisfies (a)-(d) of Lemma 5.2.7.*

Proof. Starting with the left-hand side of (5.18), we carry out the algorithm from Lemma 5.2.7 on the corresponding graph and note the analogous steps for the 2-cumulants. The matrices $M^{(i)}$, $i \in J$, are defined inductively along the rewriting procedure. Recall that every pairing in $C(\pi)$ corresponds to a 2-cumulant and every vertex in Γ_k represents a summation over an index $a_1, \ldots, a_k$. An iteration of Step I thus corresponds to replacing

$$\sum_{a_{l+1}} \kappa^{(2)}(a_l a_{l+1}, a_{l+1} a_{l+2}) = T_{a_l, a_{l+2}},$$

i.e., applying (5.12) to carry out a summation over an internal index. Hence, every new edge of weight one in Step I corresponds to introducing one matrix element of T and defining

$$M^{(i)} := T$$

for all $i \in J$. The vertex that is removed from the graph matches the index over which the summation was carried out. Recalling that $\|T\| \leq CN$, the order of N in the norm bound matches the edge weight as claimed.

Step II corresponds to matrix multiplication. Let $M^{(i_1)}$ and $M^{(i_2)}$ denote two matrices corresponding to adjacent edges of nonzero weight in the graph and let $a_l, a_m, a_n \in$

$\{a_1, \ldots, a_k\}$ with $a_l \neq a_m$ and $a_m \neq a_n$ denote the vertices adjacent to them. An iteration of Step II corresponds to replacing

$$\sum_{a_m} M^{(i_1)}_{a_l,a_m} M^{(i_2)}_{a_m,a_n} = (M^{(i_1)} M^{(i_2)})_{a_l,a_n} =: M^{(i')}_{a_l,a_n},$$

where $M^{(i')}$ denotes the matrix corresponding to the new edge and the vertex that is removed matches the summation index. Since $\|M^{(i')}\| \leq \|M^{(i_1)}\| \, \|M^{(i_2)}\|$, we obtain a norm bound from the initial estimates. In particular, the exponents of N from the previous bounds are added, matching the edge weight prescribed by the graph algorithm.

For Step III, let $M^{(i)}$ denote the matrix corresponding to an edge of weight $w_i > 0$ and assume that the two adjacent edges belong to the same pair, which is represented by the 2-cumulant $\kappa^{(2)}(a_l a_{l+1}, a_m a_{m+1})$. An iteration of Step III corresponds to replacing

$$\sum_{a_{l+1},a_m} M^{(i)}_{a_{l+1},a_m} \kappa^{(2)}(a_l a_{l+1}, a_m a_{m+1}) =: M^{(i')}_{a_l,a_{m+1}}, \tag{5.19}$$

where the matrix $M^{(i')}$ corresponds to the edge of weight $w_i + 1$ introduced in the graph algorithm and the summation indices a_{l+1}, a_m match the vertices that are removed. Note that the norm of the new matrix can be estimated by $\|M^{(i')}\| \leq N\|M^{(i)}\| \leq CN^{w_i+1}$ following an argument similar to (5.15) such that the power of N in the bound again matches the prescribed edge weight. We have thus shown that the reduction algorithm on $(\Gamma_k, C(\pi))$ from Lemma 5.2.7 translates to a rewriting procedure for the corresponding cumulant term. In particular, an edge with weight w_i appearing in any step of the graph algorithm can be associated with a matrix $M^{(i)}$ with $\|M^{(i)}\| \leq CN^{w_i}$. □

Remark. Lemma 5.2.8 extends to arbitrary $k \in \mathbb{N}$ and partitions $\pi \in \Pi_k$ by restricting the algorithm in Lemma 5.2.7 to the pairs of $C(\pi)$. The resulting graph $(\widetilde{\Gamma}_{k'}(\pi), \widetilde{C}(\pi))$ satisfies (a)-(d) with the properties (c) and (d) restricted to the pairs in $\widetilde{C}(\pi)$. Further, (5.18) generalizes to the form

$$\sum_{a_1,\ldots,a_k} \prod_{B\in\pi} \kappa^{(|B|)}(a_j a_{j+1} | j \in B) = \sum_{b_1,\ldots,b_{k'}} \Big(\prod_{B\in\pi'} \kappa^{(|B|)}(b_j b_{j+1} | j \in B) \prod_{i\in J} M^{(i)}_{b_{i_1},b_{i_2}} \Big) \tag{5.20}$$

where none of the 2-cumulants on the right-hand side of (5.20) involve internal indices and the indices b_{i_1}, b_{i_2} never appear in the same 2-cumulant.

After the rewriting procedure, summations over $k' \leq k$ indices remain. Matching (5.18), we rename them as $b_1, \ldots, b_{k'}$ to indicate that the rewriting step is completed.

Step 2: Estimates via summing-in steps

Recall that the reduction algorithm in Lemma 5.2.7 terminates in one of two possible states: either $\widetilde{C}(\pi) = \emptyset$ (if π is non-crossing) or $\widetilde{C}(\pi)$ contains at least two pairings (if π is crossing). We estimate the corresponding contributions to (5.7) separately.

Lemma 5.2.9 (Leading Terms)**.** *Let $k \geq 2$ be even and $\pi \in \Pi_k$ be a non-crossing pairing. Then*

$$N^{-k/2-1} \Big| \sum_{a_1,\ldots,a_k} \prod_{B\in\pi} \kappa^{(2)}(a_j a_{j+1} | j \in B) \Big| \leq C.$$

Proof. Whenever π is non-crossing, the algorithm in Lemma 5.2.7 reduces $(\Gamma_k, C(\pi))$ to a single vertex b_1 connected to itself by an edge of weight $k/2$. By the rewriting part of Lemma 5.2.8, we thus have

$$\sum_{a_1,\ldots,a_k} \prod_{B\in\pi} \kappa^{(2)}(a_j a_{j+1} | j \in B) = \sum_{b_1} M_{b_1,b_1} = \mathrm{tr}(M),$$

where M denotes the matrix corresponding to the loop in $\widetilde{\Gamma}_{k'}(\pi)$. Further, $\|M\| \leq CN^{k/2}$, giving $|\operatorname{tr}(M)| \leq CN^{k/2+1}$ and the claim. □

Lemma 5.2.10 (Subleading Pairings)**.** *Let $k \geq 2$ be even and $\pi \in \Pi_k$ be a crossing pairing. Then*

$$N^{-k/2-1}\Big|\sum_{a_1,\ldots,a_k}\prod_{B\in\pi}\kappa^{(2)}(a_j a_{j+1}|j\in B)\Big| \leq CN^{-1/2}.$$

Proof. Whenever π is crossing, at least two pairings remain in $\widetilde{C}(\pi)$ after the algorithm in Lemma 5.2.7 terminates. As the resulting term is claimed to be subleading, we aim to show (CR) for the entire right-hand side of (5.18), possibly allowing an additional summation, i.e., another factor of $\sqrt{N}$. Let $M^{(i)}$ denote a matrix that corresponds to an edge of weight w_i in $\widetilde{\Gamma}_{k'}(\pi)$, i.e., $\|M^{(i)}\| \leq CN^{w_i}$ following Lemma 5.2.8. By defining

$$\left(\mathbf{x}^{[M^{(i)}]b_l}\right)_{b_{l+1}} := \frac{1}{N^j}M^{(i)}_{b_l,b_{l+1}},\ b_{l+1}=1,\ldots,N \tag{5.21}$$

with $l \in \{1,\ldots,k'\}$ and $k'+1=1$, we obtain a vector that satisfies

$$\|\mathbf{x}^{[M^{(i)}]b_l}\|_2 \leq C \tag{5.22}$$

uniformly for any choice of b_l. Going through the index set J, we rewrite every matrix element on the right-hand side of (5.18) as a normalized vector and collect any separate powers of N in front of the sum.

Next, we sum in the vectors using the notation of Lemma 5.2.5. Going through the newly introduced vectors (5.21) in increasing order of the corresponding l, we replace

$$\sum_{b_{l+1}}\left(\mathbf{x}^{[M^{(i)}]b_l}\right)_{b_{l+1}}\kappa^{(2)}(b_{l+1}b_{l+2},b_m b_{m+1}) = \kappa^{(2)}(\mathbf{x}^{[M^{(i)}]b_l}b_{l+2},b_m b_{m+1}),$$

where $b_m \neq b_{l-1},b_l,b_{l+1},b_{l+2}$ denotes some other summation index. Since all vectors are defined from the rows of the respective matrices in (5.21), we always involve the first index of an index pair and thus replace at most two indices in each 2-cumulant by a vector. As a consequence of Lemmas 5.2.7 and 5.2.8, the resulting term neither involves summation over internal indices (cf. Step I) nor an index that occurs in the same 2-cumulant both as a superscript of a vector and as an argument (cf. Step III). In Fig. 5.4 we visualize the rewriting and summing in procedure for the term on the left of Fig. 5.2. The small arrows in Fig. 5.4 point to the index that will be summed in. Again, edge weights that are equal to zero are left out.

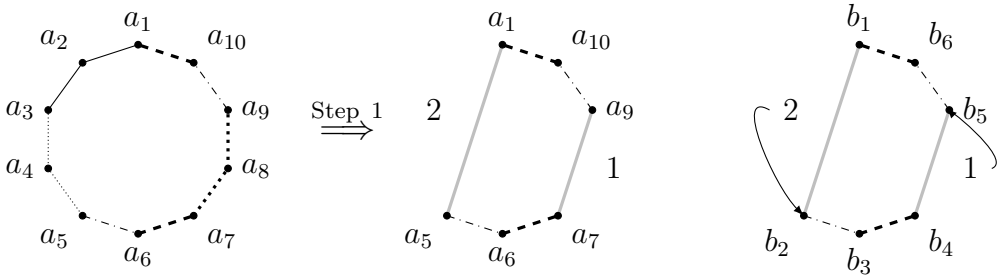

Fig. 5.4. Visualization of the rewriting and summing in for a crossing partition.

The term obtained after the sum-in procedure is now tractable by recursive summation as illustrated in Examples 5.2.3 and 5.2.6. We give the algorithm below.

Step i Identify the index b_l in the sum with the smallest subscript l. After the rewriting and sum-in steps, b_l occurs in exactly two 2-cumulants. We distinguish two cases:

- Whenever both $b_l b_{l+1}$ and $b_{l-1} b_l$ occur as index pairs in different 2-cumulants, apply an estimate similar to (5.8) with b_l in place of a_1, and start the procedure with the 2-cumulant that involves the pair $b_l b_{l+1}$.
- Whenever b_l occurs as a superscript of a vector in one 2-cumulant and as an argument in another, start the procedure with the 2-cumulant that involves b_l as an argument.

Step ii We distinguish two cases depending on the choice in Step i:

- Whenever the 2-cumulant does not involve any vectors, identify its four indices. Separate the summations over these indices from the remainder of the sum by estimating any occurrences of them in other 2-cumulants by a suitable maximum.
- Whenever the 2-cumulant involves vectors, identify the indices that either occur as the superscript of a vector or as arguments. Treat any summation over indices that appear as arguments similar to the first case, then consider the maximum for the remaining indices to isolate the term from the remainder of the sum (see (5.17)).

Next, perform the summations that have been isolated using the uniform bound (5.22) for any vectors that appear. Note that all estimates obtained from Lemmas 5.2.2 and 5.2.5 follow (CR), i.e., the final bound in each iteration reduces to a counting argument.

Step iii Identify the index $b_{l'}$ in the sum which has the smallest subscript l' and carry out Step ii for the 2-cumulant that involves $b_{l'}$ as an argument. Repeat the procedure until all summations have been carried out.

The above argument shows that all summations remaining after the rewriting and sum-in procedure follow (CR), possibly up to a factor of $\sqrt{N}$ if a summation is added in Step i. Hence, the claim follows by showing that (CR) stays valid when the remaining summations as well as the factors N^{w_i} obtained from (5.21) are taken into account. First, observe that $\|T^j\| \leq CN^j$, while a matrix element of T^j is obtained by evaluating

$$T^j_{a_l, a_{l+2j}} = \sum_{a_{l+1}, \ldots, a_{l+2j-1}} \kappa^{(2)}(a_l a_{l+1}, a_{l+1} a_{l+2}) \ldots \kappa^{(2)}(a_{l+2j-2} a_{l+2j-1}, a_{l+2j-1} a_{l+2j})$$

i.e., performing summations over $2j-1$ consecutive indices. Hence, these matrix elements yield a factor $\sqrt{N}$ more than prescribed by (CR) if estimated trivially. Moreover, carrying out a rewriting step corresponding to Step II or III on the graph does not change this fact. Indeed, if the matrix elements of $M^{(1)}$ and $M^{(2)}$ are obtained from carrying out $2j_1 - 1$ and $2j_2 - 1$ summations and the matrices are bounded by CN^{j_1} and CN^{j_2} in norm, respectively, then

$$(M^{(1)} M^{(2)})_{a,c} = \sum_b M^{(1)}_{a,b} M^{(2)}_{b,c},$$

accounts for $2(j_1 + j_2) - 1$ summations while $\|M^{(1)} M^{(2)}\| \leq CN^{j_1 + j_2}$. Similarly, if the matrix elements of $M^{(i)}$ are obtained from carrying out $2j-1$ summations and $\|M^{(i)}\| \leq CN^j$, then the matrix elements of $M^{(i')}$ in (5.19) account for two more summations, i.e., $2(j+1)-1$

in total, while the power of N in the norm bound is increased by one. Since replacing an index by a vector of the form (5.21) after the rewriting of the term also requires carrying out one summation, the additional factor of $\sqrt{N}$ per matrix is balanced out.

Hence, at most $k+1$ powers of $\sqrt{N}$ are collected from bounding the k summations with the extra factor $\sqrt{N}$ being obtained whenever an estimate similar to (5.8) is used in Step i. Thus, the bound is of order $N^{-1/2}$ here, giving the claim and showing that all crossing pairings are subleading. □

5.3 Proof of (5.4) in the General Case

Let now $k \geq 1$ be arbitrary. We aim to show that

$$N^{-k/2-1}\Big|\sum_{a_1,\dots,a_k}\prod_{B\in\pi}\kappa^{(|B|)}(a_j a_{j+1}|j\in B)\Big| \leq C(k) \tag{5.23}$$

for all $\pi \in \Pi_k$. This would complete the proof of (5.4) via (5.5). We start by proving the necessary bounds for j-cumulants with $j \geq 3$ and give the proof of (5.23) in Section 5.3.1. Note that no internal index is summed up below.

Lemma 5.3.1. *Assume that* (A3) *holds. Then*

$$\sum_{a_1,a_2,a_3,a_4}|\kappa^{(3)}(a_1a_2,a_3a_4,a_5a_6)| \leq C, \tag{5.24}$$

$$\sum_{a_1,\dots,a_5}|\kappa^{(3)}(a_1a_2,a_3a_4,a_5a_6)| \leq CN, \tag{5.25}$$

$$\sum_{a_1,\dots,a_6}|\kappa^{(3)}(a_1a_2,a_3a_4,a_5a_6)| \leq CN^2 \tag{5.26}$$

uniformly for any choice of the unsummed indices. In particular, the estimates follow (CR).

Proof. Applying (5.3) to the 3-cumulant in (5.24) and taking the maximum in a suitable way in the resulting terms, we obtain, e.g.,

$$\begin{aligned}&\sum_{a_1,\dots,a_4}|\kappa^{(2)}(a_1a_2,a_3a_4)\kappa^{(2)}(a_3a_4,a_5a_6)|\\ &\leq\Big(\max_{a_3,a_4}\sum_{a_1,a_2}|\kappa^{(2)}(a_1a_2,a_3a_4)|\Big)\sum_{a_3,a_4}|\kappa^{(2)}(a_3a_4,a_5a_6)|,\end{aligned}$$

as well as similar bounds for the terms corresponding to the other possible spanning trees. Hence, the summation over up to two index pairs is bounded by a constant, giving (5.24). The remaining estimates (5.25) and (5.26) readily follow, as the additional summations yield a factor of N or N^2, respectively. □

Observe that the bounds given in (5.24) and (5.25) imply that, e.g.,

$$\sum_{a_2}|\kappa^{(3)}(a_1a_2,a_2a_3,a_4a_5)| \leq \sum_{a_2,a_2'}|\kappa^{(3)}(a_1a_2,a_2'a_3,a_4a_5)| \leq C \leq C\sqrt{N} \tag{5.27}$$

uniformly for any choice of the unsummed indices. In particular, the estimates comply with (CR) and there is no need to perform summations over internal indices separately as in Section 5.2. The only exception occurs for $k=3$, where one obtains

$$\sum_{a_1,a_2,a_3}|\kappa^{(3)}(a_1a_2,a_2a_3,a_3a_1)| \leq \sum_{a_1,a_2,a_3,a_1',a_2',a_3'}|\kappa^{(3)}(a_1a_2,a_2'a_3,a_3'a_1')| \leq CN^2 \tag{5.28}$$

instead of the bound of order $N^{3/2}$ prescribed by the counting rule. As (5.28) is the only term for $k = 3$ due to (A1), and $k/2 + 1 = 5/2$, it follows that (5.4) holds for $k = 3$. Hence, we can exclude the $k = 3$ case from the following analysis and assume that all estimates for 3-cumulants with or without internal indices comply with (CR). We further note the following definition.

Definition 5.3.2. *For $\mathbf{x} \in \mathbb{R}^N$, set*

$$\kappa^{(3)}(\mathbf{x}a_2, a_3a_4, a_5a_6) := \sum_{a_1} \kappa^{(3)}(a_1a_2, a_3a_4, a_5a_6)x_{a_1} \tag{5.29}$$

and define any 3-cumulants with one or more indices replaced by a vector analogously.

Again, we follow the convention that vectors only occur as the first index of every index pair. Similarly to Lemma 5.2.5, we obtain the following estimates.

Lemma 5.3.3. *Assume that* (A3) *holds and let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^N$. Then summation over any number of indices of $|\kappa^{(3)}(\mathbf{x}a_2, a_3a_4, a_5a_6)|$, $|\kappa^{(3)}(\mathbf{x}a_2, \mathbf{y}a_4, a_5a_6)|$, or $|\kappa^{(3)}(\mathbf{x}a_2, \mathbf{y}a_4, \mathbf{z}a_6)|$ satisfies* (CR)*. Thus, for one vector we, e.g., have*

$$\sum_{a_2} |\kappa^{(3)}(\mathbf{x}a_2, a_3a_4, a_5a_6)| \leq CN^{1/2}\|\mathbf{x}\|_2, \tag{5.30}$$

$$\sum_{a_2,a_4} |\kappa^{(3)}(\mathbf{x}a_2, a_3a_4, a_5a_6)| \leq CN\|\mathbf{x}\|_2 \tag{5.31}$$

uniformly for any choice of the unsummed indices. Similar bounds hold if two or three vectors are involved, respectively, e.g.,

$$\sum_{a_5} |\kappa^{(3)}(\mathbf{x}a_2, \mathbf{y}a_4, a_5a_6)| \leq CN^{1/2}\|\mathbf{x}\|_2\|\mathbf{y}\|_2, \tag{5.32}$$

$$\sum_{a_2} |\kappa^{(3)}(\mathbf{x}a_2, \mathbf{y}a_4, \mathbf{z}a_6)| \leq CN^{1/2}\|\mathbf{x}\|_2\|\mathbf{y}\|_2\|\mathbf{z}\|_2.$$

Moreover, we have bounds that are stronger than (CR) *by a factor of $\sqrt{N}$ for every summation that is carried out over two consecutive indices belonging to different index pairs, e.g.,*

$$\sum_{a_4,a_5} |\kappa^{(3)}(\mathbf{x}a_2, a_3a_4, a_5a_6)| \leq CN^{1/2}\|\mathbf{x}\|_2. \tag{5.33}$$

$$\sum_{a_2,\ldots,a_6} |\kappa^{(3)}(\mathbf{x}a_2, a_3a_4, a_5a_6)| \leq CN^{3/2}\|\mathbf{x}\|_2, \tag{5.34}$$

$$\sum_{a_4,a_5} |\kappa^{(3)}(\mathbf{x}a_2, \mathbf{y}a_4, a_5a_6)| \leq CN^{1/2}\|\mathbf{x}\|_2\|\mathbf{y}\|_2. \tag{5.35}$$

In particular, for 3-cumulants, a summation over an internal index also contributes at most a factor of $\sqrt{N}$.

Proof. The proof is divided into three general arguments. First, as the bounds obtained in Lemma 5.3.1 are stronger than (CR), estimating the vector elements by $\max_j |v_j| \leq \|\mathbf{v}\|_2$ for $\mathbf{v} = \mathbf{x}, \mathbf{y}, \mathbf{z}$, respectively, and performing the summation directly yields the desired estimate in most cases. Whenever this argument does not yield a sufficiently strong bound, we apply the Cauchy-Schwarz inequality similar to the proof of Lemma 5.2.5. Lastly, it remains to check that the estimates obtained are strong enough to satisfy the second part of the lemma and also comply with (CR) if summation over internal indices is included.

As the latter can be treated by introducing additional summation labels similar to (5.27), it is enough to derive suitable bounds for distinct $a_1, \dots, a_6$.

Assume first that the 3-cumulant involves only one vector. Here, estimating $|x_j| \leq \|\mathbf{x}\|_2$ and applying (5.24) yields

$$\sum_{a_2,a_3,a_4} |\kappa^{(3)}(\mathbf{x}a_2, a_3a_4, a_5a_6)| \leq \|\mathbf{x}\|_2 \sum_{a_1,\dots,a_4} |\kappa^{(3)}(a_1a_2, a_3a_4, a_5a_6)| \leq C\|\mathbf{x}\|_2, \tag{5.36}$$

which immediately implies (5.30) and (5.31). The same N-independent bound as (5.36) holds if one sums over a_2, a_5, a_6 instead. Whenever all three index pairs are involved in the summation, Lemma 5.3.1 implies a bound of order N, i.e.,

$$\sum_{a_2,\dots,a_5} |\kappa^{(3)}(\mathbf{x}a_2, a_3a_4, a_5a_6)| \leq \|\mathbf{x}\|_2 \sum_{a_1,\dots,a_5} |\kappa^{(3)}(a_1a_2, a_3a_4, a_5a_6)| \leq CN\|\mathbf{x}\|_2, \tag{5.37}$$

and the same estimate holds if the summation over some other index $a_2, \dots, a_5$ is left out instead of a_6. In particular, all summations over two, three and four distinct indices yield a bound of at most order N, which complies with (CR). The validity of (CR) for summation over five distinct indices follows from (5.26).

Now we prove (5.33) and (5.34) by applying (5.3) and estimating one 2-cumulant in the bound similar to (5.9) in the proof of Lemma 5.2.5. This leads to

$$\begin{aligned}&\sum_{a_1,a_3,a_5} |x_{a_1}\kappa^{(2)}(a_1a_2, a_3a_4)\kappa^{(2)}(a_3a_4, a_5a_6)| \\ &\leq \Big(\max_{a_3} \sum_{a_5} |\kappa^{(2)}(a_3a_4, a_5a_6)|\Big) \sum_{a_1,a_3} |x_{a_1}\kappa^{(2)}(a_1a_2, a_3a_4)| \leq CN^{1/2}\|\mathbf{x}\|_2,\end{aligned}$$

which, together with similar estimates obtained for the other possible spanning trees on the complete graph on three vertices, proves (5.33). The estimate in (5.34) follows similarly. Applying (5.27) and (5.37) implies the remaining stronger bounds for the second part of the lemma, which completes the proof for $\kappa^{(3)}(\mathbf{x}a_2, a_3a_4, a_5a_6)$.

Next, assume that two vectors are involved in $\kappa^{(3)}$. Here, estimating $|x_j| \leq \|\mathbf{x}\|_2$ and $|y_j| \leq \|\mathbf{y}\|_2$, applying (5.24) yields

$$\sum_{a_2,a_4} |\kappa^{(3)}(\mathbf{x}a_2, \mathbf{y}a_4, a_5a_6)| \leq \|\mathbf{x}\|_2\|\mathbf{y}\|_2 \sum_{a_1,\dots,a_4} |\kappa^{(3)}(a_1a_2, a_3a_4, a_5a_6)| \leq C\|\mathbf{x}\|_2\|\mathbf{y}\|_2. \tag{5.38}$$

Further, arguing similarly to (5.9) gives

$$\sum_{a_1,a_3,a_5} |x_{a_1}y_{a_3}\kappa^{(2)}(a_3a_4, a_5a_6)\kappa^{(2)}(a_5a_6, a_1a_2)| \leq CN^{1/2}\|\mathbf{x}\|_2\|\mathbf{y}\|_2,$$

which, together with bounds of at most order $N^{1/2}$ for the terms corresponding to the other possible spanning trees, implies (5.32). An analogous estimate holds for summation over a_6, showing the validity of (CR) for a single summation. The remaining cases for summation over distinct indices follow similar to (5.38) by applying (5.25) and (5.26).

Note, however, that the above bounds are again not sufficient to imply (CR) for summation over internal indices. Arguing as in (5.11) yields the stronger estimate

$$\sum_{a_1,a_3,a_4,a_5} |x_{a_1}y_{a_3}\kappa^{(2)}(a_1a_2, a_3a_4)\kappa^{(2)}(a_3a_4, a_5a_6)| \leq CN^{1/2}\|\mathbf{x}\|_2\|\mathbf{y}\|_2.$$

Again, one can estimate similarly for the terms corresponding to the other possible spanning trees, which implies (5.35). Lastly, one needs to check that also

$$\sum_{a_2,a_4,a_5} |\kappa^{(3)}(\mathbf{x}a_2, \mathbf{y}a_4, a_4a_5)| \leq \sum_{a_2,a_4,a_5,a_4'} |\kappa^{(3)}(\mathbf{x}a_2, \mathbf{y}a_4, a_4'a_5)| \leq CN^{3/2}\|\mathbf{x}\|_2\|\mathbf{y}\|_2,$$

which follows from a similar argument. This completes the proof for $\kappa^{(3)}(\mathbf{x}a_2, \mathbf{y}a_4, a_5a_6)$.

For 3-cumulants that involve three vectors, note that writing out the term similar to (5.29) already involves all three index pairs. Whenever two summations are carried out, the estimates obtained from Lemma 5.3.1 comply with (CR), e.g.,

$$\begin{aligned}\sum_{a_2,a_4} |\kappa^{(3)}(\mathbf{x}a_2, \mathbf{y}a_4, \mathbf{z}a_6)| &\leq \|\mathbf{x}\|_2\|\mathbf{y}\|_2\|\mathbf{z}\|_2 \sum_{a_1,\dots,a_5} |\kappa^{(3)}(a_1a_2, a_3a_4, a_5a_6)| \\ &\leq CN\|\mathbf{x}\|_2\|\mathbf{y}\|_2\|\mathbf{z}\|_2,\end{aligned}$$

but a different argument is needed for estimating zero, one, or three summations. Here, arguing as in (5.11) gives

$$\begin{aligned}&\sum_{a_2} \sum_{a_1,a_3,a_5} |x_{a_1}y_{a_3}z_{a_5}\kappa^{(2)}(a_1a_2, a_3a_4)\kappa^{(2)}(a_3a_4, a_5a_6)| \\ &\leq \|\mathbf{z}\|_2 \Big(\max_{a_3} \sum_{a_5} |\kappa^{(2)}(a_3a_4, a_5a_6)| \Big) \sum_{a_1,a_2,a_3} |x_{a_1}y_{a_3}\kappa^{(2)}(a_1a_2, a_3a_4)| \leq CN^{1/2}\|\mathbf{x}\|_2\|\mathbf{y}\|_2\|\mathbf{z}\|_2\end{aligned}$$

together with similar bounds for the terms corresponding to the other possible spanning trees. By a similar argument, we obtain an N-independent bound and a bound of order N for zero and three summations, respectively. As internal indices cannot occur here, the proof of the lemma is complete. □

Lastly, we derive the necessary estimates to incorporate cumulants of order four and higher. As no exceptions similar to (5.28) occur, we directly include summations over internal indices into the bound.

Lemma 5.3.4. *Assume that* (A3) *holds and let* $j \geq 4$. *Then*

$$\sum_{a_1,\dots,a_{2j}} |\kappa^{(j)}(a_1a_2, \dots, a_{2j-1}a_{2j})| \leq CN^2.$$

Proof. After applying (5.3), bound the term by considering all possible spanning trees on the complete graph on j vertices. Starting the summation at the leaves of the respective tree, sum up one index pair of each 2-cumulant. Continuing the procedure until the root of the tree is reached, an N-independent bound is obtained in every step. Finally, a summation over all four indices of the last 2-cumulant remains, which yields the claimed bound of order N^2 by applying the last estimate of Lemma 5.2.2. □

Note that Lemma 5.3.4 in particular implies that

$$\sum_{a_1,\dots,a_j} |\kappa^{(j)}(a_1a_2, a_2a_3, a_3a_4, \dots, a_ja_1)| \leq \sum_{a_1,\dots,a_j,a_1',\dots,a_j'} |\kappa^{(j)}(a_1a_2, a_2'a_3, \dots, a_j'a_1')| \leq CN^2$$

for any $j \geq 4$ such that no special treatment is needed for (CR). Moreover, whenever $j \geq 5$, the bound is stronger than (CR). However, Lemma 5.3.4 only gives a bound of N^2 independent of the number of summations carried out.

5.3.1 Including 3-cumulants

Throughout this section, assume that $\pi \in \Pi_k$ with $|B| \in \{2,3\}$ for all $B \in \pi$. Excluding the cases in Lemma 5.2.1, we obtain the following bound.

Lemma 5.3.5 (Subleading Term with 3-cumulants)**.** *Under assumptions* (A1)-(A3)*, let $k \geq 1$ and choose a partition $\pi \in \Pi_k$ with $|B| \in \{2,3\}$ for all $B \in \pi$ that is not a pairing. Then*

$$N^{-k/2-1} \sum_{a_1,\dots,a_k} \prod_{B \in \pi} |\kappa^{(|B|)}(a_j a_{j+1} | j \in B)| \leq C N^{-1/2}. \tag{5.39}$$

Proof of Lemma 5.3.5. We extend the proof Lemma 5.2.10 to include 3-cumulants and show that (CR) holds up to a factor of $\sqrt{N}$. The modifications to each step are given below.

Step 1: Rewriting
Since the summation over internal indices in 3-cumulants does not interfere with (CR), we restrict to rewriting the 2-cumulants that involve internal indices and obtain the desired form via (5.20). Recall that certain properties only hold for the pairs in $\widetilde{C}(\pi)$ such that, e.g., two edges adjacent to an edge with nonzero weight may still belong to the same 3-tuple. As before, we rename the remaining k' summation indices as $b_1, \dots, b_{k'}$ and denote the matrices obtained along the rewriting procedure as $M^{(i)}$ with i in some index set J. Recall that $\|M^{(i)}\| \leq C N^{w_i}$ where $w_i \neq 0$ is the weight of the corresponding edge in the graph.

Step 2: Sum-In
In this step, we focus on the terms specific to 3-cumulants that are not covered by the argument in Lemma 5.2.10. Consider the possible subgraphs in which an edge with nonzero weight is adjacent to two edges that belong to the same 3-tuple. Here, rewriting and estimating the term using the vectors in (5.21) fails similarly to the second part of Example 5.2.6. We visualize the respective subgraphs in Fig. 5.5 below, where edges belonging to the same 3-tuple are indicated by the same linestyle and w_j (resp. w_{j_1}, w_{j_2}) denotes some nonzero edge weight. For simplicity, edge weights that are equal to zero as well as the labels of most vertices are left out and the remainder of the graph is indicated by horizontal dots.

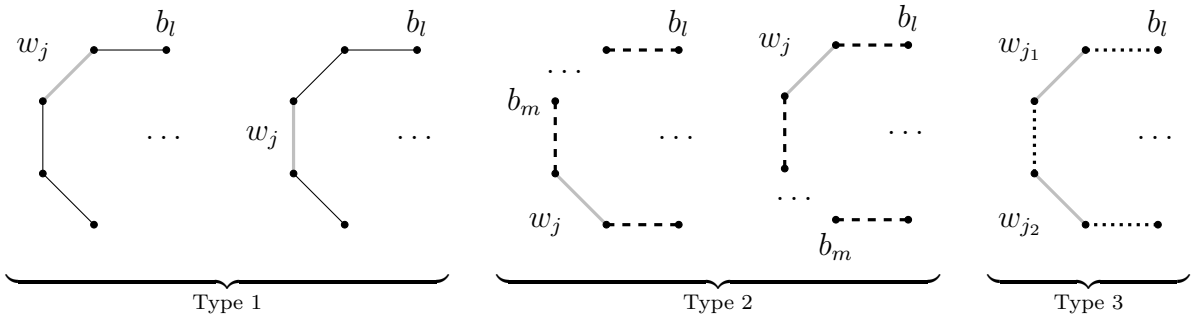

Fig. 5.5. The only subgraphs not covered by previous estimates.

We gather the 3-cumulant and the matrix element(s) corresponding to the respective subgraphs in Fig. 5.5 into one term that we refer to as type 1, 2 or 3, respectively. For example, the left type 1 graph translates to the product

$$M^{(j)}_{b_{l+1},b_{l+2}} \kappa^{(3)}(b_l b_{l+1}, b_{l+2} b_{l+3}, b_{l+3} b_{l+4})$$

in the cumulant calculation while the left type 2 graph translates to

$$M^{(j)}_{b_{m+1},b_{m+2}}\kappa^{(3)}(b_l b_{l+1}, b_m b_{m+1}, b_{m+2}b_{m+3}).$$

Here, $M^{(j)}$ denotes the matrix that corresponds to the edge with weight w_j, respectively. Note that the two solid and two dashed graphs in Fig. 5.5 look very similar, but are still not identical since the cyclic ordering breaks the symmetry. As a result of Lemma 5.2.8, any matrix element that does not occur in a term of type 1, 2 or 3 can be treated using the same sum-in procedure as in the proof of Lemma 5.2.10, i.e., by applying (5.21) and performing the summation over the corresponding b_{l+1} explicitly. Again, defining the vectors by the rows of the respective matrices ensures that only the first index of every index pair in the remaining 2- and 3-cumulants may be replaced by a vector. In particular, estimates for the special cases in Fig. 5.5 and the counting rule are sufficient to cover any term that is encountered. As an example, consider two interlaced type 1 subgraphs: If the second type 1 term is added between the two solid edges of the first, the result can be estimated using only bounds for type 1 and 2, respectively. Otherwise, the sum-in procedure applies, as at least one edge of nonzero weight is no longer adjacent to two edges from the same grouping.

At this point, any separate matrix element occurs in a term of type 1, 2 or 3. Note that also up to one (types 1 and 3) or two (type 2) indices of the 3-cumulant in the respective terms may have been replaced by a vector along the sum-in procedure.

Step 3: Estimates

To estimate the remaining summations, we apply a recursive summation procedure similar to the proof of Lemma 5.2.10. Note that the summation over any number of indices of a 2- or 3-cumulant follows (CR) by Lemmas 5.2.2, 5.2.5, 5.3.1, and 5.3.3, including summations over internal indices in the 3-cumulants and any summation over internal indices of 2-cumulants resulting in the sum-in of the vectors in (5.21). The only terms not covered by the previous arguments are the terms of type 1, 2, and 3. We claim that they, too, follow (CR) and show that applying the bounds from Lemma 5.3.1 or 5.3.3 for the 3-cumulant compensates the additional factor(s) of $\sqrt{N}$ obtained from estimating the matrix element(s) trivially by the norm of the respective matrix.

Assume first that the summation on the left-hand side of (5.23) reduces to one single term of type 1, with the internal index occurring, e.g., at b_4, or of type 3. We obtain from Lemma 5.3.1 and (5.27) that

$$\sum_{b_1,\dots,b_5} |\kappa^{(3)}(b_1b_2, b_3b_4, b_4b_5)| \leq CN^2, \qquad \sum_{b_1,\dots,b_6} |\kappa^{(3)}(b_1b_2, b_3b_4, b_5b_6)| \leq CN^2.$$

Observe that these estimates are better than the bound prescribed by (CR) by one or two factors of $\sqrt{N}$, respectively. Hence, the final bound for the term follows (CR) again.

Whenever a term of type 1, 2, or 3 is encountered along the recursive summation procedure, we argue similarly. Recall that Step ii used in the proof of Lemma 5.2.1 requires summing over all remaining indices for every 2-cumulant that is encountered. For terms of type 1, 2, or 3, we modify this step and sum up all but one (type 1 and 3) or two (type 2) indices instead, always leaving the index b_l for the largest value of l in each connected subgraph unsummed. The index or indices that are left out always belong to another 2- or 3-cumulant or term of type 1, 2, or 3 and are thus considered in later steps of the procedure. The summation rule is visualized in Fig. 5.6 below, where the large dots denote the indices that will be summed up after estimating the matrix element(s) trivially.

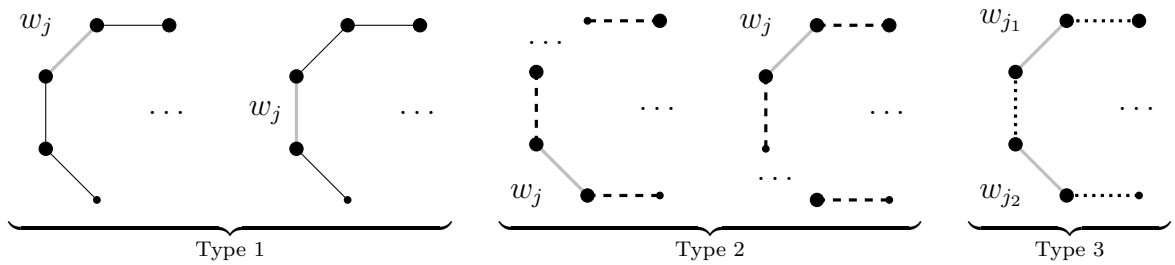

Fig. 5.6. The summation rule.

For the terms of type 1 or 3, leaving one index of the respective 3-cumulant unsummed always allows for a bound of order N by Lemma 5.3.1 or (5.37), i.e., at least $(\sqrt{N})^2$ better than (CR) would give, which compensates for the additional factor of $\sqrt{N}$ or N, respectively, from estimating the matrix elements. The same holds for terms of type 2. Note, however, that due to the dashed subgraphs in Fig. 5.6 having two connected components, we may also encounter the case that only two summations remain. Here, we apply (5.24), (5.33) or (5.35) whenever the 3-cumulant involves zero, one or two vectors, respectively. This allows to estimate the remaining two summations by a bound of at most order $N^{1/2}$, showing that (CR) holds again. Hence, we can apply a similar recursive summation procedure as used in the proof of Lemma 5.2.1 given by the following modified algorithm.

Step i Identify the terms of type 1, 2 or 3 in the summation. We distinguish two cases.

- Whenever terms of type 1, 2 or 3 are present, identify the summation index b_l in them that corresponds to the smallest subscript l .
- Whenever no terms of type 1, 2 or 3 are present, carry out Step i as in the proof of Lemma 5.2.10, possibly choosing an index occuring in a 3-cumulant.

Step ii We distinguish two cases depending on the outcome of Step i.

- Whenever the index b_l occurs in a term of type 1, 2 or 3, estimate the corresponding matrix element(s) using the bound for the matrix norm, then isolate the corresponding 3-cumulant and perform the summation according to Fig. 5.6.
- Whenever the index b_l occurs in a 2- or 3-cumulant, carry out Step ii as in the proof of Lemma 5.2.10, possibly applying Lemma 5.3.1 or 5.3.3 if b_l occurs in a 3-cumulant.

Step iii Identify the index $b_{l'}$ in the remaining sum that corresponds to the smallest subscript l' and carry out Step ii until all summations have been evaluated.

Note that starting the summation with estimating a matrix element breaks the cyclic structure of the graph similarly to (5.8). Further, we have shown that all estimates obtained in Step ii follow (CR). The final bound for the sum obtained at the end of the procedure is thus at most of order $N^{k/2+1/2}$, which is the claim. □

5.3.2 Including Cumulants of Order Four and Higher

With the necessary tools established, we proceed to estimating the summation (5.23) in the general case, i.e., when we have cumulants (tuples) of arbitrary order.

Proof of (5.23). Let $k \in \mathbb{N}$ and $\pi \in \Pi_k$ be arbitrary. Excluding the terms already considered, we can assume that $k \geq 4$ and that the partition includes at least one set of four or more elements. As before, we start by identifying any 2-cumulants with internal indices on the right-hand side of (5.23) and rewriting the term using (5.20). Next, we carry out the sum-in procedure from Lemma 5.3.5 almost unchanged, i.e., (5.21) is applied for every matrix element that can be incorporated into a 2- or 3-cumulant and the matrix elements occurring in a term of type 1, 2 or 3 (see Fig. 5.5) are gathered together with the corresponding 3-cumulant. Note that this step might leave some matrix elements unassigned. We assume first that π is chosen such that this is not the case.

Recall that the counting rule established for 2- and 3-cumulants prescribes that every individual summation yields a factor of $\sqrt{N}$, while the estimate for higher-order cumulants yields a contribution of N^2 when summing over all indices in the respective cumulant. To apply both rules simultaneously, we divide the terms obtained from the rewriting procedure into two factors F and G, where G collects the 2- and 3-cumulants, as well as the terms of type 1, 2 and 3, and the additional factors of N obtained from the N^{-j} normalization in (5.21), while the higher-order cumulants constitute F. Whenever no rewriting is required, we can split the left-hand side of (5.23) directly into

$$F(b_1, \dots, b_k) := \prod_{\substack{B \in \pi \\ |B| \geq 4}} |\kappa^{(|B|)}(b_j b_{j+1} | j \in B)|, \; G(b_1, \dots, b_k) := \prod_{\substack{B \in \pi \\ 2 \leq |B| \leq 3}} |\kappa^{(|B|)}(b_j b_{j+1} | j \in B)|.$$

Next, consider the set $X := \{j : b_j \text{ appears in a cumulant that belongs to } F\}$. Abbreviating $b_X = \{b_j : j \in X\}$ and $b_{X^c} = \{b_1, \dots, b_{k'}\} \setminus b_X$, it follows that

$$\sum_{b_1, \dots, b_{k'}} F(b_X) G(b_1, \dots, b_{k'}) \leq \Big(\sum_{b_X} F(b_X) \Big) \Big(\max_{b_X} \sum_{b_{X^c}} G(b_1, \dots, b_{k'}) \Big)$$

by taking the maximum over all indices in b_X in the factors that occur in G.

Let $n \geq 1$ be the number of factors in F. Since every cumulant included in F is of order four or higher, each factor involves at least eight indices as arguments. As every index belongs to two edges and, hence, has to appear exactly twice in the product of F and G, there are at least $4n$ distinct indices in F. By definition, the total number of distinct indices occurring in F is equal to $|X|$ such that $|X| \geq 4n$. Introducing additional summation labels $b'_1, \dots, b'_l$ to sum over all indices involved in the respective factors and applying Lemma 5.3.4 thus implies

$$\sum_{b_X} F(b_X) \leq \sum_{b_X, b'_1, \dots, b'_l} F(b_X, b'_1, \dots, b'_l) \leq C(N^2)^n \leq C N^{|X|/2}. \tag{5.40}$$

For the summation involving G, we follow Step 3 from the proof of Lemma 5.3.5. This yields

$$\max_{b_X} \sum_{b_X^c} G(b_1, \dots, b_{k'}) \leq C N^{(k - |X|)/2 + 1/2} \tag{5.41}$$

from applying the recursive summation procedure, as all terms in the sum follow (CR). Hence, we obtain a contribution of order $\sqrt{N}$ for every summation, including the ones that were carried out along the rewriting and sum-in step, and possibly an additional factor $\sqrt{N}$ from an estimate similar to (5.8). Combining (5.40) and (5.41) thus yields

$$N^{-k/2-1} \Big(\sum_{b_X} F(b_X) \Big) \Big(\max_{b_X} \sum_{b_X^c} G(b_1, \dots, b_{k'}) \Big) \leq C N^{-1/2}.$$

In the previously excluded cases, there either is a term similar to the terms of type 1, 2 or 3 with the 3-cumulant replaced by a cumulant of order four or higher, or a matrix element for which introducing vectors as in (5.21) would require summing it into a higher-order cumulant. In both cases, we estimate the matrix element trivially by the norm of the corresponding matrix, which yields a factor of $\sqrt{N}$ more than prescribed by (CR). However, the occurrence of the matrix element implies that the j-cumulant must involve at least $j+1$ distinct indices. Hence, one more summation than in the minimal case is carried out when the summation over all indices in the cumulant is evaluated, i.e., the additional factor $\sqrt{N}$ is balanced out. This implies that the estimates again follow (CR), giving the claim in the general case. In particular, we obtain a sub-leading contribution to (5.5) whenever the term on the left-hand side of (5.23) involves cumulants of order four or higher. □

6 Dynamics of a Rank-One Multiplicative Perturbation of a Unitary Matrix

We provide a dynamical study of a model of multiplicative perturbation of a unitary matrix introduced by Fyodorov. In particular, we identify a flow of deterministic domains that bound the spectrum with high probability, separating the outlier from the typical eigenvalues at all sub-critical timescales. These results are obtained under generic assumptions on U that hold for a variety of unitary random matrix models.

Preprint of G. Dubach and J. Reker. Dynamics of a rank-one multiplicative perturbation of a unitary matrix. *To appear in: Random Matrices: Theory Appl., online ready available at https://doi.org/10.1142/S2010326324500072*, 2024.

6.1 Introduction

The recent survey [71] offers an overview of a variety of exact results concerning rank-one perturbations of random matrices. Among those are

(i) Hermitian perturbations of Hermitian operators, where the now well-understood phenomenon of BBP transition occurs [14];

(ii) Generic perturbations of non-Hermitian operators, exemplified by the milestone work of Tao [137];

as well as another, intermediate, class of models, namely,

(iii) Perturbations of a normal (Hermitian or Unitary) operator resulting in a loss of normality.

Interest in these particular models has its origin in physics, as they are relevant to scattering theory (see for instance the general introduction [77]), and as such, they have been the subject of several studies in recent years, both from the mathematical and the physics perspective. The model for which the most precise estimates are available to date is the rank-one anti-Hermitian perturbation of a Hermitian matrix, namely

$$M(t) := H + \mathrm{i}tvv^*, \qquad t \geq 0, \tag{6.1}$$

where H is a matrix with Gaussian entries independent up to the condition of Hermitianity (the so-called *Gaussian Unitary Ensemble*, or GUE) and v a unit vector, sampled uniformly on the sphere, and independent from H. This model was first studied in the pioneering works [78, 79, 75]. Recently, the more general case of a Wigner matrix was considered in [57]. It was shown that a unique outlier becomes separated for all times $t > 1 + N^{-\frac{1}{3}+\epsilon}$ for any $\epsilon > 0$. Shortly thereafter, it was confirmed in [76], using

exact expressions available in the integrable GUE case, that $1 + \mathcal{O}(N^{-\frac{1}{3}})$ is indeed the critical timescale for the emergence of the outlier.

Another relevant setting is the so-called UA model, or rank-one multiplicative perturbation of a unitary matrix, first introduced by Fyodorov [73] (see also the review [80]). It is defined by

$$W(a) := UA, \qquad A = \mathrm{diag}(a, 1, \dots, 1), \quad a \in [0, 1], \tag{6.2}$$

with the matrix U being Haar-distributed on the unitary group (the so-called *Circular Unitary Ensemble*, or CUE). This model was studied in [72] at the timescale $a = \mathcal{O}(N^{-\frac{1}{2}})$, which happens to be the critical timescale for the emergence of an outlier at the origin. The purpose of this paper is to conduct a dynamical study (replacing $a \in [0, 1]$ by $t \in \mathbb{R}$, considered as a time parameter) of such UA models under fairly generic assumptions on the matrix U.

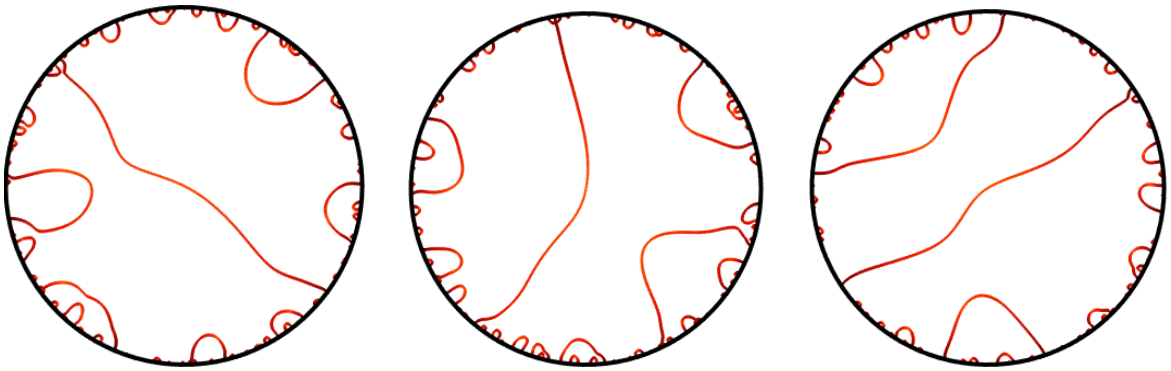

Fig. 6.1. Trajectories of the eigenvalues of the UA model of size 100×100. The evolution of the time parameter $t \in (-1, 1)$ is represented by shades of red.

6.1.1 Definitions and Notations

Throughout the paper, $\mathrm{tr}(\cdot)$ denotes the (non-normalized) trace and $\mathrm{Sp}(\cdot)$ denotes the spectrum of a matrix. We denote the open unit disc and an open disk of center z_0 and radius r by $\mathbb{D}$ and $D(z_0, r)$, respectively. The boundary of a set S is denoted by ∂S and its complement by S^c.

We say that an event (more precisely, a sequence of events) $\Omega = \Omega^{(N)}$ occurs *with high probability (w.h.p.)* if

$$\mathbb{P}\left(\Omega^{(N)}\right) \xrightarrow[N\to\infty]{} 1, \tag{6.3}$$

and that it occurs *with overwhelming probability (w.o.p.)* if for any $A > 0$,

$$1 - \mathbb{P}\left(\Omega^{(N)}\right) < N^{-A} \tag{6.4}$$

for large enough N.

An important tool that we need is the notion of *uniform stochastic domination*, which is defined as follows. Let

$$X = \left\{X^{(N)}(z) : \ N \in \mathbb{N}, z \in \mathbb{D}\right\} \quad \text{and} \quad Y = \left\{Y^{(N)}(z) : \ N \in \mathbb{N}, z \in \mathbb{D}\right\}$$

be two families of complex random variables that are indexed by $N \in \mathbb{N}$ and $z \in \mathbb{D}$. We say that X is *stochastically dominated* by Y, uniformly for $z \in \mathscr{S}^{(N)} \subset \mathbb{D}$ if, for all $\varepsilon > 0$ we have that

$$\forall z \in \mathscr{S}^{(N)} \qquad |X^{(N)}(z)| \leq N^{\epsilon} |Y^{(N)}(z)|$$

with overwhelming probability. We denote this by $X \prec Y$, specifying over which domain $\mathscr{S}^{(N)}$ the uniformity holds.

We say that G_N has a *strongly separated outlier towards the origin for* $t \in T \subset [-1,1]$ if there exist $\alpha_1 < \alpha_2$, $c_1, c_2 \in (0,1)$ such that with high probability, for every $t \in T$,

(i) $D\left(0, c_1 \frac{N^{\alpha_1}}{1+N^{\alpha_1}}\right)$ contains exactly one eigenvalue of $G_N(t)$;

(ii) $\mathbb{D} \backslash D\left(0, c_2 \frac{N^{\alpha_2}}{1+N^{\alpha_2}}\right)$ contains $N-1$ eigenvalues of $G_N(t)$.

Whenever $\alpha_1 < 0$, we say that G_N has a strongly separated outlier *at the origin*.

Definition of the Model and Statement of Results

Let $U \in \mathbb{U}_N(\mathbb{C})$ be a unitary matrix and define

$$A(t) := I_N - (1-t)vv^* \in M_N(\mathbb{C}) \tag{6.5}$$

with $t \in \mathbb{R}$ and a unit vector $v \in \mathbb{C}^N$. We consider the multiplicative rank one perturbation of U that is given by

$$G = G(t) = UA(t). \tag{6.6}$$

Note that this model coincides with (6.2) if $v = e_1$ and $|t| < 1$, taking $a = t$. It is readily checked that $G(t)$ is unitary for $t = \pm 1$, so that its spectrum is in $\partial\mathbb{D}$. The eigenvalue trajectories of $G(t)$ for $|t| \leq 1$ lie in $\overline{\mathbb{D}}$ by Weyl's inequality (as the smallest singular value of $G(t)$ is equal to $|t|$ while all others have modulus 1 by direct computation). For $t = 0$, the spectrum of $G(0)$ coincides with that of a truncated unitary matrix, with an extra eigenvalue at 0. Observing that, for $t \neq 0$, G is invertible with $G(t)^{-1} = G(t^{-1})^*$, the trajectories for $|t| \geq 1$ lie in the complement of $\mathbb{D}$, and their properties are readily obtained from the $|t| \leq 1$ case. We will hence focus our study of the spectrum of $G(t)$ to $t \in (-1,1)$, as represented on Fig. 6.1.

For U a unitary matrix and v a unit vector, we define the quantity

$$\omega_1^{(N)} := \sqrt{N} v^* U^* v \tag{6.7}$$

and the functions

$$\mathscr{W}(z) := v^* \frac{1}{I_N - zU^*} v, \qquad \mathscr{W}_2(z) := v^* \frac{(zU^*)^2}{I_N - zU^*} v, \qquad s_1(z) := 1 + \frac{\omega_1^{(N)}}{\sqrt{N}} z. \tag{6.8}$$

Note that $\mathscr{W}$ is named by analogy with the 'weighted resolvent' in [57, Def. 2], as it plays a similar role in the analysis below. However, there are important differences in the behavior and scaling of the two objects, which we address in more detail in the appendix. It is immediate to check that

$$\mathscr{W}(z) = s_1(z) + \mathscr{W}_2(z). \tag{6.9}$$

Our main results are stated using the following assumptions.

Assumption 6.1.1 (Simplicity)**.** *The spectrum of the random unitary matrix U is almost surely simple; the unit vector v is independent of U and distributed with a non-vanishing density on the sphere $\mathbb{S}_{N-1}(\mathbb{C})$.*

Note that the simplicity of the spectrum holds for all typical models of random matrices. Whenever Assumption 6.1.1 is used, we denote the randomness in v by $\mathbb{P}_v$. In Section 6.2.2, we will work under the following regularity assumptions on U and v.

Assumption 6.1.2 (Regularity for $\mathscr{W}_2$)**.** *For any $\delta > 0$,*

$$\mathscr{W}_2(z) \prec \frac{|z|^2}{\sqrt{N}(1-|z|)} \tag{6.10}$$

holds uniformly for $z \in \mathbb{D}_\delta := D(0, 1 - N^{-\delta})$.

Assumption 6.1.3 (Regularity for ω_1)**.** *For any $\epsilon > 0$,*

$$N^{\varepsilon} > |\omega_1^{(N)}| > N^{-\epsilon} \tag{6.11}$$

holds with high probability.

This is verified in particular for a Haar-distributed unitary matrix and a uniform unit vector v independent from U, as will be proved in Theorem 6.3.1. Note that one of our results (Theorem 6.2.7) does not rely on Assumption 6.1.3, but only on the deterministic bound

$$|\omega_1^{(N)}| \leq N^{\frac{1}{2}}, \tag{6.12}$$

which follows from the definition of $\omega_1^{(N)}$, and U being unitary.

Our main result is a description of the different scales of these dynamics, under the regularity assumption. In particular, the emergence of an outlier at the origin is connected to the critical timescale $t = \mathcal{O}(N^{-\frac{1}{2}})$, in a way that can be described as follows:

Theorem 6.1.4. *Under Assumptions 6.1.2 and 6.1.3, for any $\alpha > \varepsilon > 0$, the following holds:*

- *There is a strongly separated outlier at the origin for $|t| < N^{-\frac{1}{2}-\alpha}$.*
- *With high probability, for all $|t| > N^{-\frac{1}{2}+\alpha}$, all eigenvalues are in $\mathbb{D} \backslash D(0, 1 - N^{-\alpha+\varepsilon})$.*

More precise statements are given as Theorems 6.2.7 and 6.2.8 in Section 6.2.2.

An Overview of the Method: Rouché's Theorem and Isotropic Local Law.

The method used in [57] to study the model (6.1) essentially relies on two key facts:

(i) The characterization of the spectrum at time t as the preimage of i/t by a random meromorphic function $\mathscr{W}$;

(ii) A uniform isotropic local law that allows to approximate $\mathscr{W}$ by a deterministic function on a contour.

These facts together allow to locate the eigenvalues, and to isolate the outlier from the rest of the spectrum via Rouché's theorem, establishing that it is strongly separated[1] at the proper timescales. The present case shares some important similarities. In

[1] with the appropriate notion of 'strong separation' in this setting.

particular, the analog of (i) (stated below as Lemma 6.2.1) holds with the appropriate function $\mathscr{W}$, and some of the resulting properties, such as the almost sure non-intersection of trajectories, are analogous to the weakly non-Hermitian case as well. Apart from these formal analogies between the two models, the UA model (6.6) also involves some specific features. Especially, the isotropic local law (ii) is replaced by the regularity assumption on $\mathscr{W}_2$ (Assumption 6.1.2), which plays an equivalent role, but holds in fact more generally (see discussion in the Appendix). Another particular feature is the role of the parameter $\omega_1^{(N)}$ and the corresponding Assumption 6.1.3, which do not have an analog in the Hermitian setting.

Plan of the Paper.

Section 6.2 contains our results concerning the trajectories of the UA model (6.6). We first derive general properties of the trajectories, only relying on Assumption 6.1.1, in Section 6.2.1, and then establish precise estimates on these trajectories under Assumptions 6.1.2 and 6.1.3, in Section 6.2.2 – from which Theorem 6.1.4 follows. Finally, in Section 6.3, we focus on the particular CUE case, proving that it verifies all relevant assumptions, and also providing a simple proof that the timescales put forward in Theorem 6.1.4 are optimal. The relevance and generality of Assumption 6.1.2 are discussed in more detail in the Appendix.

Acknowledgments

The authors would like to thank Paul Bourgade, Djalil Chafaï, László Erdős, Peter Forrester, Yan Fyodorov, and Boris Khoruzhenko for numerous discussions and comments. G.D. gratefully acknowledges support from the Fondation des Sciences Mathématiques de Paris. The work of J.R. is partially supported by ERC Advanced Grant "RMTBeyond" No. 101020331.

6.2 Study of the Trajectories

6.2.1 Properties of the Trajectories Under the Simplicity Assumption

Lemma 6.2.1. *Let $z \in \mathbb{D}$ and consider the UA model (6.6) with $t \in (-1, 1)$. Then*

$$z \in \mathrm{Sp}(G(t)) \;\Leftrightarrow\; \mathscr{W}(z) = \frac{1}{1-t}.$$

Proof. By definition, z is in the spectrum of $G(t)$ if and only if

$$0 = \det(G(t) - z) = \det(U - z - (1-t)Uvv^*).$$

As $|z| < 1$, this is equivalent to

$$\det(I_N - (1-t)(U-z)^{-1}Uvv^*) = 0.$$

Using Sylvester's identity and defining $\mathscr{W}(z) = v^*(U-z)^{-1}Uv$, we find that z is in the spectrum of $G(t)$ if and only if

$$1 - (1-t)\mathscr{W}(z) = 0$$

which is the claim. □

This strong property is due to the perturbation having rank one. In particular, it implies that the trajectories of the eigenvalues are given by the zero level lines of the imaginary part of $\mathscr{W}$, i.e.,

$$\bigcup_{j=1}^{N}\{\lambda_j(t) : t \in (-1,1)\} = \{z \in \mathbb{D} : \Im\mathscr{W}(z) = 0\}. \tag{6.13}$$

Another important property that follows from the simplicity assumption is the non-intersection of the trajectories.

Theorem 6.2.2 (Non-intersection of trajectories)**.** *Under Assumption 6.1.1, for $N \geq 5$, the trajectories of the eigenvalues almost surely do not intersect, i.e.,*

$$\mathbb{P}\left(\{\exists s,t \in (-1,1),\ \exists i,j \in [\![1,N]\!]\ :\ (s,i) \neq (t,j),\ \lambda_i(s) = \lambda_j(t)\}\right) = 0. \tag{6.14}$$

Proof. The proof is analogous to that of [57, Thm. 2] with a few adjustments to account for the different geometry. We prove it for any fixed U with a simple spectrum, only relying on the randomness of v (i.e., according to the probability measure $\mathbb{P}_v$). If $\lambda_i(s) = \lambda_j(t)$ for some $i,j \in [\![1,N]\!]$ and $s,t \geq (-1,1)$, Lemma 6.2.1 implies that

$$\mathscr{W}(z) = \frac{1}{1-s} = \frac{1}{1-t} \tag{6.15}$$

and thus $t = s$. So we only need to consider the possible intersection of two trajectories λ_i and λ_j $(i \neq j)$ at the same time t. This implies that the function $\mathscr{W}(z) - \frac{1}{1-t}$ vanishes at $z \in \mathbb{D}$ with multiplicity at least two. Hence, any intersection point z_0 lies in the set

$$\mathcal{S} := \{z \in \mathbb{D} : \Im\mathscr{W}(z) = 0\} \cap \{z \in \mathbb{D} : \mathscr{W}'(z) = 0\}.$$

We show that almost surely, there is no z_0 that satisfies both conditions. First, note that

$$\mathscr{W}'(z) = \sum_{j=1}^{N} |\langle u_j|v\rangle|^2 \frac{\mathrm{e}^{-\mathrm{i}\theta_j}}{(1-z\mathrm{e}^{-\mathrm{i}\theta_j})^2},$$

where $\mathrm{e}^{\mathrm{i}\theta_1}, \dots, \mathrm{e}^{\mathrm{i}\theta_N}$ and $u_1, \dots, u_N$ denote the eigenvalues of U and the associated unit eigenvectors, respectively. Hence, the condition $\mathscr{W}'(z_0) = 0$ is equivalent to z_0 being the root of a complex polynomial of degree $2N-2$ ($\mathbb{P}_v$-almost surely) that does not vanish on the unit circle. In particular, there is a finite number of such points. Allowing for multiplicity, we denote them by $(Z_k)_{k=1}^{2N-2}$.

Next, note that $0 \notin \mathcal{S}$ by inspection. We may thus assume $z \neq 0$ and replace the condition $\mathscr{W}'(z) = 0$ by $z\mathscr{W}'(z) = 0$. For any $z \notin \partial\mathbb{D}$, define the real vectors

$$Y_1(z) := \left(\Im\frac{1}{1-z\mathrm{e}^{-\mathrm{i}\theta_j}}\right)_{j=1}^{N},\ Y_2(z) := \left(\Re\frac{z\mathrm{e}^{-\mathrm{i}\theta_j}}{(1-z\mathrm{e}^{-\mathrm{i}\theta_j})^2}\right)_{j=1}^{N},\ Y_3(z) := \left(\Im\frac{z\mathrm{e}^{-\mathrm{i}\theta_j}}{(1-z\mathrm{e}^{-\mathrm{i}\theta_j})^2}\right)_{j=1}^{N}$$

as well as

$$X := (|\langle u_j|v\rangle|^2)_{j=1}^{N}.$$

With this notation, the condition $\Im\mathscr{W}(z) = 0$ can be written as $\langle X|Y_1(z)\rangle$ while $z\mathscr{W}'(z) = 0$ translates to $\langle X|Y_2(z)\rangle = \langle X|Y_3(z)\rangle = 0$. The remainder of the proof relies on the following (deterministic) observation.

Lemma 6.2.3. *Let $N \geq 5$. For any $z \in \mathbb{D}\backslash\{0\}$, it holds that $Y_1(z) \notin \mathrm{Span}_{\mathbb{R}}(Y_2(z), Y_3(z))$.*

Proof. For $z \neq 0$ with $|z| < 1$ denote

$$a_j + \mathrm{i}b_j = \frac{1}{1 - z\mathrm{e}^{-\mathrm{i}\theta_j}}$$

which can be rearranged to give

$$|z| = |z\mathrm{e}^{-\mathrm{i}\theta_j}| = \left|1 - \frac{1}{a_j + \mathrm{i}b_j}\right|$$

i.e., $(a_j + \mathrm{i}b_j)^{-1}$ lies on a circle with center 1 and radius $|z|$. Hence, by inversion, $a_j + \mathrm{i}b_j$ lies on a circle with center $(1 - |z|^2)^{-1}$ and radius $|z|(1 - |z|^2)^{-1}$. We denote this circle by $\mathscr{C}_{|z|}$. As the phases θ_j are distinct by assumption, (a_j, b_j) are N distinct points on $\mathscr{C}_{|z|}$. Assume now that $Y_1(z) = \alpha Y_2(z) + \beta Y_2(z)$ for some $\alpha, \beta \in \mathbb{R}$. Since

$$\frac{z\mathrm{e}^{-\mathrm{i}\theta_j}}{(1 - z\mathrm{e}^{-\mathrm{i}\theta_j)^2}} = \frac{1 - (1 - z\mathrm{e}^{-\mathrm{i}\theta_j})}{(1 - z\mathrm{e}^{-\mathrm{i}\theta_j)^2}} = (a_j + \mathrm{i}b_j)^2 - (a_j + \mathrm{i}b_j),$$

and $Y_1(z) = b_j$, we can rewrite the linear dependence condition as

$$b_j = \alpha(a_j^2 - b_j^2 - a_j) + \beta(2a_jb_j - b_j).$$

Note that this describes a hyperbola $\mathscr{H}_{\alpha,\beta}$ (in the degenerate case, we obtain a union of two lines). Bézout's theorem for curves gives a bound as to the number of intersection points between this curve and a circle:

$$N \leq |\mathscr{C}_{|z|} \cap \mathscr{H}_{\alpha_\beta}| \leq 4,$$

which is the desired contradiction. □

Lemma 6.2.3 implies that

$$\forall z \in \mathbb{D}, \quad \mathbb{P}_v(\Im \mathscr{W}(z) = 0 \mid \mathscr{W}'(z) = 0) = 0,$$

as Y_1 is linearly independent of $Y_2(z), Y_3(z)$ and X is distributed with a non-vanishing density on the simplex

$$\left\{(x_1, \ldots, x_N) \ : \ x_j \geq 0, \sum_{j=1}^{N} x_j = 1\right\}.$$

by Assumption 6.1.1. We aim to deduce that

$$\mathbb{P}_v(\exists z \in \mathbb{D} : \Im \mathscr{W}(z) = 0 \ \wedge \ \mathscr{W}'(z) = 0) = 0$$

to conclude that trajectories originating from two distinct eigenvalues $\mathbb{P}_v$-a.s. do not intersect. Let $\pi \in \mathfrak{S}_{2N}$ be a uniformly random permutation and let $\mathbb{P}$ denote the probability with respect to π as well as v. Moreover, define

$$N_z := |\{k \ : \ Z_k = z\}|$$

and recall that the points Z_k were defined as the zeros of $\mathscr{W}'$, i.e., $\mathscr{W}'(z) = 0$ if and only if $N_z \geq 1$. The randomness in π ensures that

$$\mathbb{P}(Z_{\pi(k)} = z \mid N_z \geq 1) = \sum_{m=1}^{2N-2} \mathbb{P}(Z_{\pi(k)} = z \mid N_z = m) \geq \frac{1}{2N}$$

so that we can write

$$0 = \mathbb{P}(\Im\mathscr{W}(z) = 0 \mid N_z \geq 1) \geq \mathbb{P}(\Im\mathscr{W}(z) = 0 \mid Z_{\pi(k)} = z)\mathbb{P}(Z_{\pi(k)} = z \mid N_z \geq 1)$$
$$\geq \frac{1}{2N}\mathbb{P}(\Im\mathscr{W}(Z_{\pi(k)}) = 0 \mid Z_{\pi(k)} = z)$$

for fixed $k = 1, \ldots, 2N-2$. This implies $\mathbb{P}(\Im\mathscr{W}(Z_k) = 0 \mid Z_{\pi(k)} = z) = 0$ and it follows that

$$\mathbb{P}(\Im\mathscr{W}(Z_k) = 0) = 0$$

by integrating out the conditioning. As there are only finitely many choices for k, a union bound yields

$$\mathbb{P}_v(\exists k : \Im\mathscr{W}(Z_k) = 0) = \mathbb{P}(\exists k : \Im\mathscr{W}(Z_{\pi(k)}) = 0) \leq \sum_{k=1}^{2N-2} \mathbb{P}(\Im\mathscr{W}(Z_{\pi(k)}) = 0) = 0,$$

implying $\mathbb{P}_v(\mathcal{S}) = 0$. We conclude that

$$\mathbb{P}_v(\exists t \in (-1,1), i \neq j : \lambda_j(t) = \lambda_j(t)) = 0$$

which completes the proof. □

Non-intersection of trajectories is an essential input, as it allows to identify each $\langle_k(t)$ as the unique continuous trajectory linking the eigenvalue $\mathrm{e}^{\mathrm{i}\theta_k}$ at time $t = 1$ to one of the eigenvalues at any time t. In particular, we may consider the trajectory of the eigenvalue that eventually becomes an outlier (for t approaching 0) and give meaning to concepts such as the 'origin' of the outlier, the shape, and length of individual trajectories, etc. A realization with two trajectories getting very close to one another is given in Fig. 6.2.

Remarkably, the evolution of the eigenvalues of the UA model is described by a closed system of differential equations that governs the behavior of the trajectories. We provide these equations below for $t \in (0,1]$ (the equations for $t \in [-1,0)$ can be derived similarly). Note that this result is not used as input for our main theorems. However, it is an important feature of these dynamics, and may be of interest for future studies of this model.

Theorem 6.2.4 (Differential Equations)**.** *Assume that U satisfies Assumption 6.1.1. We denote its eigenvalues by $\mathrm{e}^{\mathrm{i}\theta_1}, \ldots, \mathrm{e}^{\mathrm{i}\theta_N}$ with $\theta_j \in [0, 2\pi)$ and the associated unit eigenvectors by $u_1, \ldots, u_N$. Let v be a unit vector and $t \in (0,1]$. Then the evolution of the eigenvalues $\lambda_1(t), \ldots, \lambda_N(t)$ of $G(t) = U(I_N - (1-t)vv^*)$ is governed by the differential equations*

$$\lambda_j'(t) = \frac{1 - |\lambda_j(t)|^2}{t(t^2-1)} \left(\prod_{k=1}^{N} \lambda_k(t)\right) \prod_{k \neq j} \frac{\lambda_j\overline{\lambda_k(t)} - 1}{\lambda_j(t) - \lambda_k(t)}, \quad j = 1, \ldots, N, \tag{6.16}$$

with initial conditions $\lambda_j(1) = \mathrm{e}^{\mathrm{i}\theta_j}$ and

$$\lambda_j'(1) = \mathrm{e}^{\mathrm{i}\theta_j}|\langle u_j|v\rangle|^2, \quad j = 1, \ldots, N. \tag{6.17}$$

Proof. Let $L_1, \ldots, L_N, R_1, \ldots, R_N$ be a bi-orthogonal system of left and right eigenvectors associated with $G(t)$. We start with the basic perturbative identity

$$\lambda_j'(t) = \langle L_j|G'(t)|R_j\rangle = \langle L_j|U|v\rangle\langle v|R_j\rangle \tag{6.18}$$

to obtain a dynamic description of the eigenvalues (cf. proof of [57, Thm. 4]). Using that $\langle L_j|(UA) = \lambda_j\langle L_j|$ and that $A^{-1} = I - (1-t^{-1})vv^*$, the relation (6.18) can be rewritten as

$$\lambda_j'(t) = \frac{\lambda_j(t)}{t}\langle L_j|v\rangle\langle v|R_j\rangle. \tag{6.19}$$

For $t = 1$, we have $G(t) = U$ and the eigenvalues are given by $\lambda_j(1) = \mathrm{e}^{\mathrm{i}\theta_j}$. Moreover, $L_j = R_j = u_j$, which yields the initial condition stated in (6.17). For $t \in (0,1)$, note that the quantity $\langle L_j|v\rangle\langle v|R_j\rangle$ is invariant under a unitary change of basis. It is thus sufficient to compute it for a Schur form T of G. W.l.o.g. let $\lambda_1, \dots, \lambda_N$ be the order in which the eigenvalues appear on the diagonal of T, i.e., $\lambda_j = T_{jj}$. By exchangeability, it is sufficient to show (6.16) for $j = 1$. Writing $G = STS^*$ with a suitable unitary matrix S, note that

$$TT^* = SGG^*S^* = SAA^*S = I - (t^2-1)Svv^*S^*$$

i.e., TT^* is a rank-one perturbation of identity. Let $\widetilde{v} := Sv$. By [56, Lem. 4], we can express the columns of T as well as the entries of $\sqrt{t^2-1}\widetilde{v}$ in terms of the eigenvalues $\lambda_1, \dots, \lambda_N$. In particular, the k-th column of T is of the form

$$(T_{kj})_{j=1}^N = \begin{pmatrix} \tau_k \\ \lambda_k \\ 0 \end{pmatrix}, \quad \tau_k = \frac{|\lambda_k|^2 - 1}{\overline{\lambda_k}} \frac{1}{\widetilde{v}_k} \widetilde{v}^{(k-1)}$$

with $\widetilde{v}^{(k-1)}$ being the column vector containing the first $k-1$ entries of $\widetilde{v}$. As T is (upper) triangular, we further have $R_1 = e_1$ and thus $\langle \widetilde{v}|R_1\rangle = \overline{\widetilde{v}_1}$. For the following argument, denote $L_1 = (1, \ell_2, \dots, \ell_N)$ and $L_1^{(d)} = (1, \ell_2, \dots, \ell_d)$ for $d = 1, \dots, N$. This vector satisfies the recursion

$$\ell_{k+1} = \frac{1}{\lambda_1 - \lambda_{k+1}}\langle L_1^{(k)}|\tau_{k+1}\rangle = \frac{1}{\lambda_1 - \lambda_{k+1}} \frac{|\lambda_{k+1}|^2 - 1}{\overline{\lambda_{k+1}}\widetilde{v}_{k+1}} \langle L_1^{(k)}|\widetilde{v}^{(k)}\rangle, \tag{6.20}$$

where τ_{k+1} denotes the first k entries of the $(k+1)$-th column of T. We use this connection to compute $\langle L_1|\widetilde{v}\rangle$ recursively. The initial condition is given by

$$\langle L_1^{(1)}|\widetilde{v}^{(1)}\rangle = \widetilde{v}_1$$

and we iterate the recursion by

$$\begin{aligned}
\langle L_1^{(k+1)}|\widetilde{v}^{(k+1)}\rangle &= \langle L_1^{(k)}|\widetilde{v}^{(k)}\rangle + \ell_{k+1}\widetilde{v}_{k+1} \\
&= \left(1 + \frac{1}{\lambda_1 - \lambda_{k+1}} \frac{|\lambda_{k+1}|^2 - 1}{\overline{\lambda_{k+1}}}\right) \langle L_1^{(k)}|\widetilde{v}^{(k)}\rangle \\
&= \frac{\lambda_1\overline{\lambda_{k+1}} - 1}{\overline{\lambda_{k+1}}(\lambda_1 - \lambda_{k+1})} \langle L_1^{(k)}|\widetilde{v}^{(k)}\rangle
\end{aligned}$$

using (6.20). This yields

$$\langle L_1|\widetilde{v}\rangle = \langle L_1^{(N)}|\widetilde{v}^{(N)}\rangle = \widetilde{v}_1 \prod_{k=2}^N \left(\frac{\lambda_1\overline{\lambda_{k+1}} - 1}{\overline{\lambda_{k+1}}(\lambda_1 - \lambda_{k+1})}\right).$$

Hence, by (6.19), we obtain for any $t \in (0,1)$ that

$$\begin{aligned}
\lambda_1'(t) &= \frac{\lambda_1(t)}{t}\langle L_1|v\rangle\langle v|R_1\rangle = \frac{\lambda_1(t)}{t(t^2-1)}\langle L_1|\widetilde{v}\rangle\langle \widetilde{v}|R_1\rangle \\
&= \frac{\lambda_1(t)|\widetilde{v}_1|^2}{t(t^2-1)} \prod_{k=2}^N \left(\frac{\lambda_1(t)\overline{\lambda_k(t)} - 1}{\overline{\lambda_k(t)}(\lambda_1(t) - \lambda_k(t))}\right).
\end{aligned}$$

Moreover, replacing $|\widetilde{v}_1|^2$ by

$$|\widetilde{v}_1|^2 = (1 - |\lambda_1|^2) \prod_{k=2}^{N} |\lambda_k|^2$$

using [56, Lem. 4], we obtain (6.16), which is a function of t and the eigenvalues only. □

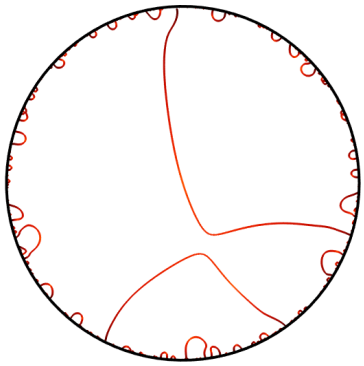

Fig. 6.2. Trajectories of the eigenvalues of the UA model of size 500×500. The evolution of the time parameter $t \in (-1, 1)$ is represented by shades of red. On this particular realization, one extreme trajectory gets close to the trajectory of the outlier, and appears to be repelled away from it with high velocity, according to the singularity of the ODE (6.16).

6.2.2 Estimates on the trajectories under the Regularity Assumption

We now state and prove the estimates on the trajectories, under Assumptions 6.1.2 and 6.1.3, that constitute our main results. The method consists in comparing the function $\mathscr{W}$ to its affine part s_1 – see definitions (6.7) and (6.8). As the eigenvalues are characterized by $\mathscr{W}(z) = \frac{1}{1-t}$ (by Lemma 6.2.1), we denote by z_t the unique solution of the equation $s_1(z) = \frac{1}{1-t}$, that is

$$z_t := \frac{\sqrt{N}}{\omega_1^{(N)}} \frac{t}{1-t}, \tag{6.21}$$

so that, in particular, z_t is the expected location of the outlier in the sub-critical regime; but also, $|z_t|$ will appear as a relevant scale parameter in all regimes.

In order to deduce the location of eigenvalues from the comparison of $\mathscr{W}$ and s_1, we rely on the classical tool of complex analysis known as Rouché's theorem, which we recall:

Theorem 6.2.5 (Rouché's Theorem)**.** *Let f and g be two holomorphic functions on a domain $\Omega \subset \mathbb{C}$ with closed and simple boundary $\partial\Omega$. If $|f - g| < |g|$ on $\partial\Omega$, then f and g have the same number of zeros (counted with multiplicity) inside Ω.*

The next proposition identifies a domain on which the Rouché inequality holds for the functions at stake here.

Proposition 6.2.6. *Under Assumption 6.1.2, for every $\eta, \delta > 0$, it holds with overwhelming probability that for every $t \in (-1, 1)$, the functions*

$$f_t(z) = \mathscr{W}(z) - \frac{1}{1-t} \quad \text{and} \quad g_t(z) = s_1(z) - \frac{1}{1-t} \tag{6.22}$$

verify the inequality $|f_t - g_t| < |g_t|$ on the domain $\mathscr{R}_{t,\eta} \cap \mathbb{D}_\delta$, with

$$\mathscr{R}_{t,\eta} := \left\{ z \ : \ \frac{N^\eta |z|^2}{1 - |z|} < \left|\omega_1^{(N)}\right| |z - z_t| \right\}. \tag{6.23}$$

Note that (6.23) defines a random domain, depending on $\omega_1^{(N)}$, and that Proposition 6.2.6 only relies on Assumption 6.1.2.
The proof proceeds by direct computation, unpacking the definitions of the objects and using the relevant assumptions.

Proof. For any $\eta > \epsilon > 0$ and $z \in \mathscr{R}_{t,\eta}$ the following holds:

$$\begin{aligned}
|f_t(z) - g_t(z)| &= |\mathscr{W}_2(z)| && \text{(by definition of } f_t, g_t, \mathscr{W}_2) \\
&< \frac{N^\epsilon |z|^2}{\sqrt{N}(1-|z|)} && \text{(w.o.p., by Assumption 6.1.2)} \\
&< N^{\epsilon-\eta} \frac{|\omega_1^{(N)}|}{\sqrt{N}} |z - z_t| && \text{(by definition of } \mathscr{R}_{t,\eta}^{\omega_1}) \\
&= N^{\epsilon-\eta} |g_t(z)| && \text{(by definition of } g_t)
\end{aligned}$$

which yields the expected inequality as $\epsilon < \eta$. □

We can now establish more precise versions of our main result (Theorem 6.1.4).

Theorem 6.2.7. *Under Assumption 6.1.2, for any $\alpha > \varepsilon > 0$ and $\delta > \frac{1}{2}$, the following holds with overwhelming probability:*

(i) For $|t| < |\omega_1^{(N)}| N^{-\frac{1}{2}-\alpha}$, exactly one eigenvalue is in D_1, while all other eigenvalues are in $\mathbb{D} \backslash D_2$, where D_1, D_2 are defined as

$$D_1 := D\left(z_t, \rho_1 = |z_t|^2 \frac{N^\epsilon}{|\omega_1|}\right), \qquad D_2 := D\left(0, \rho_2 = \left(1 + \frac{N^\epsilon}{|\omega_1|}\right)^{-1}\right). \tag{6.24}$$

(ii) If $|\omega_1^{(N)}| < \frac{1}{2} N^{\frac{1}{2}}$, then for $|t| \in (2|\omega_1^{(N)}| N^{-\frac{1}{2}}, 1)$, all eigenvalues are in $\mathbb{D} \backslash (D_3 \cap \mathbb{D}_\delta)$, where

$$D_3 := D\left(0, \rho_3 = \left(1 + \frac{N^\varepsilon}{|\omega_1|\,|z_t|}\right)^{-1}\right). \tag{6.25}$$

The proof proceeds by a straightforward inspection of the Rouché domain $\mathscr{R}_{t,\eta}$ (with $\eta < \varepsilon$) in the different regimes.

Proof. For $|t| < |\omega_1^{(N)}| N^{-\frac{1}{2}-\alpha}$, by definition of z_t and using the deterministic bound (6.12), we find

$$|z_t| < N^{-\alpha}, \qquad \frac{|z_t|^2}{|\omega_1^{(N)}|} < N^{-\frac{1}{2}-2\alpha}, \qquad |z_t| + \rho_1 < \rho_2 < 1 - N^{-\delta}, \tag{6.26}$$

so that in particular $D_1 \subset D_2 \subset \mathbb{D}_\delta$. We now check that $\partial D_1, \partial D_2$ are in the domain $\mathscr{R}_{t,\eta}$ with any $\eta < \epsilon$.

1. For $z \in \partial D_1$, by inspection $1 - |z| \asymp 1$, $|z| \asymp |z_t|$. Hence, the left-hand side of the inequality in (6.23) is of order $|z_t|^2 N^\eta$, while the right-hand side is

   $$|\omega_1||z - z_t| = |z_t|^2 N^\epsilon$$

   by definition of D_1. The inequality follows as we chose $\eta < \varepsilon$. By Rouché's theorem, D_1 contains a unique eigenvalue.

2. For $z \in \partial D_2$, the left-hand side of the inequality in (6.23) is $N^{\eta}\rho_2^2(1-\rho_2)^{-1}$ and by inspection the right-hand side is of order $|\omega_1|\rho_2$, so that the inequality holds if

$$\frac{\rho_2}{1-\rho_2} = |\omega_1|N^{-\varepsilon} \ll |\omega_1|N^{-\eta}$$

which, again, is true as $\eta < \epsilon$. By Rouché's theorem, D_2 contains a unique eigenvalue, which was already isolated (in D_1 by the above argument), so that all other eigenvalues are in $\mathbb{D}\backslash D_2$.

Assuming $|\omega_1^{(N)}| < \frac{1}{2}N^{\frac{1}{2}}$, and $|t| > 2|\omega_1^{(N)}|N^{-\frac{1}{2}}$, by definition of z_t, we have $|z_t| > 2$, so that $|z - z_t| \asymp |z_t|$. Consider $z \in \partial D_3$. As $\rho_3 < 1$, we have $|z|^2 < |z|$, and so

$$\frac{|z|^2}{1-|z|} < \frac{|z|}{1-|z|} = \frac{\rho_3}{1-\rho_3} = N^{-\varepsilon}|\omega_1||z_t| < N^{-\eta}|\omega_1||z_t|$$

as $\eta < \varepsilon$, proving that ∂D_3 is in the domain $\mathscr{R}_{t,\eta}$. As $s_1(z) - \frac{1}{1-t}$ does not vanish inside D_3, we conclude by Rouché's theorem that all eigenvalues are outside $D_3 \cap \mathbb{D}_\delta$. □

Under Assumption 6.1.3, the dependence on $\omega_1^{(N)}$ can be removed from the domains considered. Note the result now only holds with high probability.

Theorem 6.2.8. *Under Assumptions 6.1.2 and 6.1.3, for any $\alpha > \varepsilon > 0$, the following holds with high probability:*

(i) For $|t| < N^{-\frac{1}{2}-\alpha}$, exactly one eigenvalue is in $D\left(z_t, |z_t|^2N^{\varepsilon}\right)$, while all other eigenvalues are in $\mathbb{D}\backslash D\left(0, N^{-\varepsilon}\right)$.

(ii) For $|t| > N^{-\frac{1}{2}+\alpha}$, all eigenvalues are in $\mathbb{D}\backslash D\left(0, 1-|z_t|^{-1}N^{\varepsilon}\right)$.

Proof. This follows from Theorem 6.2.7 by choosing the parameters carefully: $\delta > 0$ from Assumption 6.1.2 large enough and $\varepsilon > 0$ from Assumption 6.1.3 small enough. □

6.3 Particulars of the CUE Case

Of particular interest is the situation where U is a Haar-distributed Unitary matrix – the so-called Circular Unitary Ensemble (CUE). This is the case considered in particular by the papers [73, 74, 80, 72, 71]. We first prove that our general regularity assumption holds in this case; we then rely on the integrability of the model to give an argument as to the optimality of the main theorem.

6.3.1 Proof of the Regularity Assumption for CUE

This section is devoted to the proof of the following theorem, which establishes that our main results (Theorems 6.1.4, 6.2.7, 6.2.8) apply in particular to a CUE matrix.

Theorem 6.3.1. *Assumptions 6.1.2 and 6.1.3 hold for U a* CUE(N) *matrix and v a unit vector independent of U. More precisely, if U is an $N \times N$ Haar-distributed unitary matrix, then*

$$\left|\omega_1^{(N)}\right|^2 := N|v^*U^*v|^2 \stackrel{d}{=} N\beta_{1,N} \asymp 1, \tag{6.27}$$

and for any $\delta > 0$, the stochastic domination

$$\mathscr{W}_2(z) := v^* \frac{(zU^*)^2}{I_N - zU^*} v \prec \frac{|z|^2}{\sqrt{N}(1-|z|)} \tag{6.28}$$

holds uniformly in z with $1 - |z| > N^{-\delta}$.

By unitary invariance of CUE matrices, and v being independent from U, we can assume $v = e_1$ in all proofs below. Also, throughout the Section, we denote the eigenvalues of U by $\mathrm{e}^{\mathrm{i}\theta_1}, \dots, \mathrm{e}^{\mathrm{i}\theta_N}$ and the corresponding eigenvectors by $r_1, \dots, r_N$. To illustrate the method, we first give a precise estimation of the variance of $\mathscr{W}_2(z)$.

Proposition 6.3.2. *For any $z \in \mathbb{D}, N \geq 1$,*

$$\mathbb{E}\mathscr{W}_2(z) = 0, \qquad \mathrm{Var}\mathscr{W}_2(z) \asymp \frac{|z|^4}{N(1-|z|)}, \tag{6.29}$$

where $\asymp$ means that it is bounded above and below, up to an absolute multiplicative constant.

Proof. For any fixed z, the spectral decomposition yields

$$\mathscr{W}_2(z) = e_1^* \frac{(zU^*)^2}{I_N - zU^*} e_1 = \sum_{k=1}^N |r_{k1}|^2 \frac{z^2 \mathrm{e}^{-2\mathrm{i}\theta_k}}{1 - z\mathrm{e}^{-\mathrm{i}\theta_k}} = \sum_{k=1}^N |r_{k1}|^2 \sum_{\ell \geq 2} z^l \mathrm{e}^{-\mathrm{i}\ell\theta_k} \tag{6.30}$$

The expectation being zero follows directly by unitary invariance. For the variance, we write

$$\begin{aligned}
\mathrm{Var}\mathscr{W}_2(z) &= \mathbb{E} \sum_{k_1,k_2=1}^N |r_{k_1 1}|^2 |r_{k_2 1}|^2 \sum_{\ell_1,\ell_2 \geq 2} z^{\ell_1} \overline{z}^{\ell_2} \mathrm{e}^{-\mathrm{i}\ell_1\theta_{k_1}} \mathrm{e}^{\mathrm{i}\ell_2\theta_{k_2}} \\
&= \sum_{k_1,k_2=1}^N \mathbb{E}\left(|r_{k_1 1}|^2 |r_{k_2 1}|^2\right) \sum_{\ell \geq 2} |z|^{2\ell} \mathbb{E}\left(\mathrm{e}^{-\mathrm{i}\ell(\theta_{k_1} - \theta_{k_2})}\right)
\end{aligned}$$

where we have used the independence of eigenvectors and eigenvalues to split the expectation, and unitary invariance to cancel the terms with $\ell_1 \neq \ell_2$. The expectation of the eigenvector term is obtained explicitly from Weingarten calculus, namely

$$\mathbb{E}\left(|r_{k_1 1}|^2 |r_{k_2 1}|^2\right) = \frac{1 + \delta_{k_1,k_2}}{N(N+1)}. \tag{6.31}$$

For the eigenvalues we invoke the fact (Proposition 3.11 in [110]) that

$$\mathbb{E}|\operatorname{tr}(U^\ell)|^2 = \min(\ell, N) \tag{6.32}$$

from which it follows in particular that

$$\mathbb{E}\left(\mathrm{e}^{\mathrm{i}\ell(\theta_1 - \theta_2)}\right) = \frac{\min(\ell, N) - N}{N(N-1)} = -\frac{(N-\ell)_+}{N(N-1)}. \tag{6.33}$$

Finally, by splitting the sum according to $k_1 = k_2$ and $k_1 \neq k_2$, we deduce the following exact expression of the variance of $\mathscr{W}_2$

$$\begin{aligned}
\mathrm{Var}\mathscr{W}_2(z) &= \sum_{k=1}^N \frac{2}{N(N+1)} \sum_{\ell \geq 2} |z|^{2\ell} - \sum_{k_1 \neq k_2} \frac{1}{N(N+1)} \sum_{\ell=2}^N |z|^{2\ell} \frac{N-\ell}{N(N-1)} \\
&= \frac{2|z|^4}{(N+1)(1-|z|^2)} - \frac{1}{N(N+1)} \sum_{\ell=2}^N (N-\ell)|z|^{2\ell} \\
&< \frac{2|z|^4}{(N+1)(1-|z|^2)},
\end{aligned}$$

which proves the upper bound; but also, from the same identity,

$$\begin{aligned}\mathrm{Var}\mathscr{W}_2(z) &= \frac{|z|^4}{(N+1)(1-|z|^2)} + \frac{|z|^{2N+2}}{(N+1)(1-|z|^2)} + \frac{1}{N(N+1)}\sum_{\ell=2}^{N}\ell|z|^{2\ell} \\ &> \frac{|z|^4}{(N+1)(1-|z|^2)},\end{aligned}$$

which proves the lower bound on the variance. □

To generalize the above argument to all moments, we need the following lemma, which generalizes the overall bound of order N^{-1} in (6.33).

Lemma 6.3.3. *For any $m \geq 2$, distinct $k_1, \dots, k_m$ and non-zero integer coefficients $a_1, \dots, a_m$, we have*

$$\left|\mathbb{E}\left(\mathrm{e}^{\mathrm{i}\sum_{l=1}^m a_l\theta_{k_l}}\right)\right| < m!N^{-m+1}. \tag{6.34}$$

Proof. This follows from classical results on CUE as a determinantal point process (see for instance Proposition 3.7 and Lemma 3.8 in [110]). The correlation functions of CUE(N) are given by

$$\rho_m(\theta_1, \dots, \theta_m) = \det\left(K_N(\theta_j, \theta_k)\right)_{j,k=1}^m \tag{6.35}$$

with the kernel

$$K_N(x,y) = \sum_{\ell=0}^{N-1}\mathrm{e}^{\mathrm{i}\ell(x-y)}. \tag{6.36}$$

Note in particular that $K_N(x,x) = N$ whereas for $x \neq y$ it is a sum of exponentials, not all of which will contribute – this is the reason behind Lemma 8 in a nutshell. It follows in particular that

$$\begin{aligned}\mathbb{E}\left(\mathrm{e}^{\mathrm{i}\sum_{l=1}^m a_l\theta_{k_l}}\right) &= \frac{(N-m)!}{N!}\int \mathrm{e}^{\mathrm{i}\sum_{l=1}^m a_l\theta_l}\det\left(K_N(\theta_j,\theta_k)\right)_{j,k=1}^m \mathrm{d}\theta_1\cdots\mathrm{d}\theta_m \\ &= \frac{(N-m)!}{N!}\int \mathrm{e}^{\mathrm{i}\sum_{l=1}^m a_l\theta_l}\sum_{\sigma\in\mathfrak{S}_m}\epsilon(\sigma)\prod_{j=1}^m K_N(\theta_j,\theta_{\sigma(j)})\mathrm{d}\theta_1\cdots\mathrm{d}\theta_m \\ &= \frac{(N-m)!}{N!}\sum_{\sigma\in\mathfrak{S}_m}\sum_{\ell_1,\dots\ell_m=0}^{N-1}\epsilon(\sigma)\int \mathrm{e}^{\mathrm{i}\sum_{l=1}^m a_l\theta_l}\mathrm{e}^{\mathrm{i}\sum_{j=1}^m \ell_j(\theta_j-\theta_{\sigma(j)})}\mathrm{d}\theta_1\cdots\mathrm{d}\theta_m\end{aligned}$$

as each term gives ± 1 if and only if the sums in the exponentials exactly match, and zero otherwise. We bound it by

$$\begin{aligned}\left|\mathbb{E}\left(\mathrm{e}^{\mathrm{i}\sum_{l=1}^m a_l\theta_{k_l}}\right)\right| &\leq \frac{(N-m)!}{N!}\left|\left\{\begin{array}{c}\sigma\in\mathfrak{S}_m, \\ 0\leq\ell_1,\dots,\ell_m\leq N-1\end{array} : \sum_{l=1}^m a_l\theta_l \equiv \sum_{j=1}^m \ell_j(\theta_{\sigma(j)}-\theta_j)\right\}\right| \\ &\leq \frac{(N-m)!}{N!}m!N,\end{aligned}$$

where $\equiv$ stands for equality as polynomials in $\theta_1, \dots, \theta_m$. Indeed, there are less than $m!$ choice of σ, N choices of ℓ_1, and at most one choice of $(\ell_2, \dots, \ell_m)$ that works once σ and ℓ_1 are fixed[2]. The claim follows. □

[2]One could also note that the only choice of σ that can possibly work is a large cycle, resulting in the whole sum being of the sign of $\epsilon(\sigma) = (-1)^{m+1}$; but we do not use this particular fact.

Lemma 6.3.4. *For any $M \geq 1$, indices $k_1, k_1', \ldots, k_M, k_M'$ (not necessarily distinct) and positive integer coefficients $\ell_1, \ell_1' \ldots, \ell_M, \ell_M' \geq 1$,*

$$\left| \mathbb{E}\left(\mathrm{e}^{\mathrm{i}\sum_{j=1}^{M} \ell_j \theta_{k_j} - \ell_j' \theta_{k_j'}} \right) \right| < (2M)! N^{M - |\{k_1, k_1', \ldots, k_M, k_M'\}|}. \tag{6.37}$$

Proof. Let $\mathfrak{m} = \mathfrak{m}(\ell_j, \ell_j', k_j, k_j')_{j=1,\ldots,M}$ stand for the number of different θ_k terms with non-zero coefficient in the linear combination $\sum_{j=1}^{M} \ell_j \theta_{k_j} - \ell_j' \theta_{k_j'}$. The key combinatorial fact will be the following inequality, that holds for any choice of k, k', ℓ, ℓ':

$$|\{k_1, k_1', \ldots, k_M, k_M'\}| \leq \mathfrak{m} + M \quad \text{with equality only if} \quad \mathfrak{m} = 0. \tag{6.38}$$

Indeed, the difference between the cardinal $|\{k_1, k_1', \ldots, k_M, k_M'\}|$ and $\mathfrak{m}$ is exactly the number of θ_k terms that exactly vanish due to a correspondence between coefficients ℓ, ℓ'. Due to the positive sign of these coefficients, this always implies that (at least) one term θ_k and (at least) one term $\theta_{k'}$ have canceled one another out, which can be the case for at most M distinct coefficients, hence the inequality. Moreover, if such cancellations do concern M distinct coefficients (equality case), then M distinct θ_k terms vanish with M distinct $\theta_{k'}$ terms, with an exact correspondence between the two sides – in particular, in this case, the whole linear combination vanishes, and $\mathfrak{m} = 0$.

We now consider three possible cases: either $\mathfrak{m} = 0$, $\mathfrak{m} = 1$ or $\mathfrak{m} \geq 2$.

(i) If $\mathfrak{m} = 0$, the the l.h.s. of (6.37) is 1 whereas the r.h.s is at least $(2M)!$ by inequality (6.38).

(ii) If $\mathfrak{m} = 1$, the l.h.s is 0 by unitary invariance.

(iii) If $\mathfrak{m} \geq 2$, the linear combination has $\mathfrak{m}$ distinct terms with non-zero integer coefficients, and Lemma 6.3.3 gives

$$\left| \mathbb{E}\left(\mathrm{e}^{\mathrm{i}\sum_{j=1}^{M} \ell_j \theta_{k_j} - \ell_j' \theta_{k_j'}} \right) \right| < \mathfrak{m}!\ N^{-\mathfrak{m}+1} \leq (2M)!\ N^{-\mathfrak{m}+1}.$$

Moreover, as $\mathfrak{m} \neq 0$ we are not in the equality case of (6.38) and so

$$-\mathfrak{m} + 1 \leq M - |\{k_1, k_1', \ldots, k_M, k_M'\}|.$$

The claim follows. □

Proposition 6.3.5. *For any $M \geq 1$, $z \in \mathbb{D}$, for large enough N,*

$$\mathbb{E}|\mathscr{W}_2(z)|^{2M} < \frac{C_M |z|^{4M}}{N^M (1 - |z|)^{2M-1}} \tag{6.39}$$

where C_M is a constant depending only on M and not on N, z.

Proof. We start with the spectral decomposition, expand, and split the expectation as in the proof of Proposition 6.3.2. This yields

$$\begin{aligned}
&\mathbb{E}|\mathscr{W}_2(z)|^{2M} \\
&= \sum_{k_1, k_1', \ldots, k_M, k_M' = 1}^{N} \mathbb{E}\left(\prod_{j=1}^{M} |r_{1,k_j}|^2 |r_{1,k_j'}|^2 \right) \sum_{\ell_1, \ell_1', \ldots, \ell_M, \ell_M' \geq 2} z^{\sum_{j=1}^{M} \ell_j} \overline{z}^{\sum_{j=1}^{M} \ell_j'} \mathbb{E}\left(\mathrm{e}^{\sum_{j=1}^{M} \ell_j \theta_{k_j} - \ell_j' \theta_{k_j'}} \right) \\
&\leq |z|^{4M} \sum_{k_1, k_1', \ldots, k_M, k_M' = 1}^{N} \underbrace{\mathbb{E}\left(\prod_{j=1}^{M} |r_{1,k_j}|^2 |r_{1,k_j'}|^2 \right)}_{\text{(I)}} \sum_{\substack{\ell_1, \ell_1', \ldots, \ell_M, \ell_M' \geq 0 \\ \sum \ell_j = \sum \ell_j'}} |z|^{2\sum_{j=1}^{M} \ell_j} \underbrace{\left| \mathbb{E}\left(\mathrm{e}^{\sum_{j=1}^{M} \ell_j \theta_{k_j} - \ell_j' \theta_{k_j'}} \right) \right|}_{\text{(II)}},
\end{aligned}$$

where we used that, by unitary invariance, only the terms with $\sum_{j=1}^M \ell_j = \sum_{j=1}^M \ell'_j$ contribute to the second sum. We now proceed in successive bounds.

The eigenvector term (I) is bounded by a general Weingarten bound

$$\mathbb{E}\left(\prod_{j=1}^{M} |r_{1,k_j}|^2 |r_{1,k'_j}|^2\right) < C_M N^{-2M} \tag{6.40}$$

with a constant C_M that depends only on M. This can be obtained for instance from Theorem 1.1 in [41], provided $N > 2\sqrt{3}M^{7/4}$ (note that this requirement is the only reason for the provision of N being "large enough" in the statement). Indeed, this theorem gives a uniform bound on all Weingarten function $\mathrm{Wg}(\sigma\tau^{-1}, N)$ with $(\sigma, \tau) \in \mathfrak{S}_M^2$. By the fundamental theorem of Weingarten calculus, one can write the expectation (I) as a sum of such Weingarten functions with admissible pairs (σ, τ). Moreover, this sum has a maximum of $(M!)^2$ terms – a number which only depends on M and is therefore absorbed into the constant C_M.

The eigenvalue term (II) is directly bounded by Lemma 6.3.4. The factor $(2M)!$ is absorbed into the constant C_M, so that together, these first two bounds yield

$$\begin{aligned}\mathbb{E}|\mathscr{W}_2(z)|^{2M} &< C_M|z|^{4M} \sum_{k_1,k'_1,\dots,k_M,k'_M=1}^{N} N^{-2M} \sum_{\substack{\ell_1,\ell'_1,\dots,\ell_M,\ell'_M \geq 0 \\ \sum \ell_j = \sum \ell'_j}} N^{M-|\{k_1,k'_1,\dots,k_M,k'_M\}|} |z|^{2\sum_j \ell_j} \\ &= \frac{C_M|z|^{4M}}{N^M} \underbrace{\sum_{k_1,k'_1,\dots,k_M,k'_M=1}^{N} N^{-|\{k_1,k'_1,\dots,k_M,k'_M\}|}}_{\text{(III)}} \underbrace{\sum_{\ell_1,\ell'_1,\dots,\ell_M,\ell'_M \geq 0} \mathbf{1}_{\sum \ell_j = \sum \ell'_j} |z|^{2\sum_{j=1}^M \ell_j}}_{\text{(IV)}}\end{aligned}$$

The sum (III) can be written with a first sum over configurations of $k_1, k'_1, \dots k_M, k'_M$, i.e. partitions π of $[\![1, 2M]\!]$, and then a sum over choices of indices coherent with π:

$$\text{(III)} = \sum_{\pi \in \mathscr{P}([\![1,2M]\!])} \sum_{(k_j,k'_j)_{j=1}^M \sim \pi} N^{-|\{k_1,k'_1,\dots,k_M,k'_M\}|}, \tag{6.41}$$

where by $(a_j)_j \sim \pi$ we mean that $a_i = a_j$ if and only if i, j are in the same block of π. For a given partition π, the number of possible choices of k_j, k'_j is exactly

$$N(N-1)\cdots(N-|\pi|+1) \leq N^{|\pi|} = N^{|\{k_1,k'_1,\dots,k_M,k'_M\}|}$$

so that (III) is bounded by $|\mathscr{P}([\![1, 2M]\!])|$. Note that this constant depends only on M and we can absorb it into C_M.

The last term (IV) can be computed exactly. Recall that the number of partitions of an integer $L \geq 0$ in $M \geq 1$ parts is given by

$$\left|\left\{\ell_1, \dots, \ell_M \geq 0 \ : \ \sum_{i=1}^{M} \ell_i = L\right\}\right| = \binom{L+M-1}{M-1}. \tag{6.42}$$

The expression (IV) can hence be identified as a hypergeometric function of $|z|^2$, i.e.,

$$\sum_{\ell_1,\ell'_1,\dots,\ell_M,\ell'_M \geq 0} |z|^{2\sum_j \ell_j} \mathbf{1}_{\sum_j \ell_j = \sum_j \ell'_j} = \sum_{L \geq 0} \binom{L+M-1}{M-1}^2 |z|^{2L} = {}_2F_1\left(M, M; 1; |z|^2\right). \tag{6.43}$$

We remind the reader of the definition of the hypergeometric functions

$$ {}_2F_1\left(a,b;c;z\right) = \sum_{n\geq 0} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!}, \tag{6.44} $$

where $(\cdot)_n$ is the rising Pochhammer symbol, defined for any $q \in \mathbb{C}$ by

$$ (q)_n = \begin{cases} 1 & \text{if } n = 0 \\ q(q+1)\cdots(q+n-1) & \text{if } n \geq 1. \end{cases} \tag{6.45} $$

One of the most celebrated identities of hypergeometric functions is Euler's identity

$$ {}_2F_1\left(a,b;c;z\right) = (1-z)^{c-a-b}\, {}_2F_1\left(c-a,c-b;c;z\right). \tag{6.46} $$

Applying (6.46) to the parameters from (6.43) yields

$$ \begin{aligned} {}_2F_1\left(M,M;1;|z|^2\right) &= (1-|z|^2)^{1-2M}\, {}_2F_1\left(1-M,1-M;1;|z|^2\right) \\ &= \frac{1}{(1-|z|^2)^{2M-1}} \sum_{n=0}^{M-1} \frac{((1-M)_n)^2}{n!^2} |z|^{2n} \end{aligned} $$

where the last equality follows from the definitions (6.44) and (6.45). The last term is a polynomial with positive terms and hence bounded by a constant depending only on M (for instance, its value at $|z| = 1$). This concludes the proof. □

We can now give the proof of Theorem 6.3.1.

Proof of Theorem 6.3.1. For CUE, it is well known that $\omega_1 = U_{11}^* \stackrel{d}{=} \mathrm{e}^{\mathrm{i}\theta}\sqrt{\beta_{1,N}}$ with uniform θ and an independent beta distribution (see, e.g., [110, Ch. 2] or [29]). This implies (6.27).

It remains to establish (6.28). Observing that the desired stochastic domination bound trivially holds at the origin, let $z \neq 0$ and consider the function $\widetilde{\mathscr{W}_2}(z) := z^{-2}\mathscr{W}_2(z)$. Fix $\ell > 2\delta$ and let $\mathbb{L} \subseteq \mathbb{D}$ be a lattice with $|\mathbb{L}| \simeq N^\ell$ such that for every $z \in \mathbb{D}_\delta = D(0, 1-N^{-\delta})$ there is some $z_0 \in \mathbb{L}$ with $|z - z_0| \leq C_{\mathbb{L}} N^{-\ell/2}$. We start by establishing a stochastic domination bound for $\widetilde{\mathscr{W}_2}$ that holds point-wise at $z_0 \in \mathbb{L} \cap \mathbb{D}_\delta$. Fix $\varepsilon, A > 0$ and let $M \in \mathbb{N}$ such that $2\varepsilon M > A + \ell$. Using Markov's inequality and the moment bound in Proposition 6.3.5 yields

$$ \mathbb{P}\left(\widetilde{\mathscr{W}_2}(z_0) > N^\varepsilon \frac{1}{\sqrt{N}(1-|z_0|)}\right) \leq \frac{C_M}{N^{2M\varepsilon}} \leq N^{-A-\ell}. \tag{6.47} $$

In particular, $\mathscr{W}_2(z_0) \prec \frac{|z_0|^2}{\sqrt{N}(1-|z_0|)}$ for any lattice point z_0 with $|z_0| < 1 - N^{-\delta}$. By a union bound, (6.47) extends to

$$ \mathbb{P}\left(\exists z_0 \in \mathbb{L} \cap \mathbb{D}_\delta : \widetilde{\mathscr{W}_2}(z_0) > N^\varepsilon \frac{1}{\sqrt{N}(1-|z_0|)}\right) < N^{-A}. $$

Next, let $z \in \mathbb{D}_\delta$. Recalling that we may assume $v = e_1$, compute

$$ \widetilde{\mathscr{W}_2}'(z) = \left(\frac{2zU^*(I_N - zU^*) - z^2(U^*)^3}{z^2(I_N - zU^*)^2} - \frac{(zU^*)^2}{z^3(I_N - zU^*)}\right)_{11} = \left(\frac{-(U^*)^3}{(I_N - zU^*)^2}\right)_{11} $$

We estimate the derivative by

$$ |\widetilde{\mathscr{W}_2}'(z)| \leq \left\| \frac{-(U^*)^3}{(I_N - zU^*)^2} \right\| \leq \frac{1}{(1-|z|)^2}, $$

where the last inequality follows from the spectral theorem and the fact that all eigenvalues of U lie on the unit circle. Applying the mean value theorem for the analytic function $\widetilde{\mathscr{W}_2}(z)$ yields

$$\begin{aligned}\left|\widetilde{\mathscr{W}_2}(z)-\widetilde{\mathscr{W}_2}(z_0)\right| &\le |z-z_0|\max\left\{\left|\widetilde{\mathscr{W}_2}'(\xi)\right| \;:\; \xi\in[z_0,z]\right\} \le C_{\mathbb{L}}N^{-\ell/2}\frac{1}{(1-|z|)^2}\\ &\le C_{\mathbb{L}}N^{-\ell/2+\delta}\frac{1}{1-|z|}\end{aligned}$$

where we w.l.o.g. picked $z_0\in\mathbb{L}\cap\mathbb{D}_\delta$ that is closer to the center of the disc than z. Hence, the stochastic domination bound extends to all of $\mathbb{D}_\delta$ and we obtain

$$\widetilde{\mathscr{W}_2}(z)\prec\frac{1}{\sqrt{N}(1-|z|)}$$

uniformly for $z\in D(0,1-\delta)\setminus\{0\}$. As multiplying both sides with $|z|^2$ does not change the stochastic domination bound, (6.28) follows. □

6.3.2 Optimality of the Critical Timescale for CUE

The estimates derived above from Assumption 6.1.2 (see Proposition 6.2.6) do not give sufficient precision at the critical timescale. What holds in the CUE case (and is expected more generally) is that, typically, for $t=\mu N^{-\frac{1}{2}}$, several eigenvalues are at an order one distance from both the origin and the unit circle, implying in particular that the outlier is not strongly separated at the critical timescale.

Theorem 6.3.6. *If U is Haar distributed, and v is independent from U, there is no strongly separated outlier at time $t=\alpha N^{-\frac{1}{2}}$.*

We give a short proof inspired by [80, Sec. 3.3] that relies on an exact formula from [73] and a Poissonization trick. A more complete treatment of these statistics is provided in [74].

Proof. Let $t\in[0,1]$. It is known from [73] that the joint density of the eigenvalues of the UA model at time t is proportional to[3]

$$(1-t^2)^{1-N}\delta\left(t^2-\prod_{l=1}^N|z_l|^2\right)\prod_{1\le j<k\le N}|z_j-z_k|^2 \tag{6.48}$$

with a normalizing constant that does not depend on t. We now assume the parameter t to be random such that t^2 has distribution $F(x)\mathrm{d}x$ for some suitable density F. The eigenvalues of the UA model at this random time t have the joint density

$$\left(1-\prod_{l=1}^N|z_l|^2\right)^{1-N}F\left(\prod_{l=1}^N|z_l|^2\right)\prod_{1\le j<k\le N}|z_j-z_k|^2. \tag{6.49}$$

For any $k\ge1$, we choose $F(x)=x^{k-1}(1-x)^{N-1}$, which is to say that t^2 follows a $\beta_{k,N}$ distribution. The joint density of eigenvalues now simplifies to

$$\prod|z_l|^{2(k-1)}\prod_{1\le j<k\le N}|z_j-z_k|^2, \tag{6.50}$$

[3]Note that this density would be easy to handle, if not for the delta function. This condition is the symptom of a dimensional mismatch, which we resolve by taking a random t ('Poissonization'), thus adding a degree of freedom that we can later remove ('de-Poissonization') to prove the claim.

i.e., they form a determinantal point process with radial symmetry. In this setting, Kostlan's theorem gives a particularly nice description of the distribution of the squared norm of the radii, namely

$$\{|\langle_1|^2, \dots, |\langle_N|^2\} \stackrel{d}{=} \{\beta_{k,1}, \dots, \beta_{k+N-1,1}\} \tag{6.51}$$

with the β variables on the right-hand side being independent (see [98] for Kostlan's original paper in the complex Ginibre case, and [55, 90] for generalizations of this counter-intuitive property and various comments). The relevant probabilities can now be computed exactly. Assuming $k = 1$ for simplicity, the random time parameter is of order $N^{-\frac{1}{2}}$. More precisely, for any $b > a > 0$ we have

$$\begin{aligned}\mathbb{P}\left(t \in \left(\frac{a}{\sqrt{N}}, \frac{b}{\sqrt{N}}\right)\right) = \mathbb{P}\left(\beta_{1,N} \in \left(\frac{a^2}{N}, \frac{b^2}{N}\right)\right) &= N\int_{a^2/N}^{b^2/N}(1-x)^{N-1}\mathrm{d}x \\ &= \left(1-\frac{a^2}{N}\right)^N - \left(1-\frac{b^2}{N}\right)^N \to \mathrm{e}^{-a^2} - \mathrm{e}^{-b^2} > 0,\end{aligned}$$

using that t^2 is $\beta_{1,N}$-distributed. The probability of having either none or several eigenvalues at some distance from the origin or the unit circle can be given explicitly using (6.51). For any $a \in (0,1)$, it follows that

$$\mathbb{P}\left(\forall j, |\langle_j|^2 > a\right) = \prod_{j=1}^{N}\mathbb{P}\left(\beta_{j,1} > a\right) = \prod_{j=1}^{N}\int_a^1 jx^{j-1}\mathrm{d}x = \prod_{j=1}^{N}(1-a^j) > \prod_{j=1}^{\infty}(1-a^j).$$

This last product is the inverse of the generating series of partitions of integers, which is known to be convergent for $|a| < 1$. This implies that the above probability is non-zero (i.e., there is a non-zero probability that a disk of order one is empty of eigenvalues), and that this probability tends to 1 when $a \to 0$.

Let $E_{2,b}$ be the event that more than one eigenvalue is in $D(0,b)$. Similar to the above computations, we deduce the lower bound

$$\mathbb{P}\left(E_{2,b}\right) > \mathbb{P}\left(\beta_{1,1} < b^2, \beta_{2,1} < b^2\right) = b^6 \xrightarrow[b\to 1^-]{} 1.$$

The claim now follows by contradiction. Assume there is a strongly separated outlier for all times $t_0 > \mu N^{-1/2}$. Denote by $p_\mu > 0$ the probability of the random time t being in this regime, recalling that t^2 follows a $\beta_{1,N}$ distribution. In the given setting, we can choose a radius $a > 0$ such that the disk $D(0,a)$ is free of eigenvalue with probability larger than $1 - \frac{p_\mu}{2}$. These two observations result in a positive probability of both events (empty disk and $t > \mu N^{-\frac{1}{2}}$) occurring simultaneously. Hence, the strong separation must hold with $\alpha_1 \geq 0$.

Similarly, we can choose a radius $b > 0$ such that there is a probability larger than $1 - \frac{p_\mu}{2}$ of having two eigenvalues in $D(0,b)$. By a similar argument, the strong separation must hold with $\alpha_2 \leq 0$. This is a contradiction, as the definition of strong separation requires that $\alpha_1 < \alpha_2$. □

6.A On the Regularity Assumption

We would like to close this work with a few remarks about the regularity assumption for $\mathscr{W}_2$ (Assumption 6.1.2) and how it differs from the isotropic local law used in [57].

First, note that this assumption holds for models beyond the CUE case and does not rely on level repulsion. Indeed, the argument provided for CUE above applies, for instance, to any unitary invariant matrix (i.e. $U \overset{d}{=} QUQ^*$ for any unitary Q) with eigenvalues distributed according to a generic α-determinantal processes with $\alpha \in [-1, 1]$ (see [43, 109] for a definition and history of these processes). In particular, we obtain an analog of Lemma 6.3.3 with an α-determinant, provided only that $|\alpha| \leq 1$. This covers the special case of i.i.d. points ($\alpha = 0$) and that of a permanental point process ($\alpha = 1$), for which repulsion is replaced by some amount of attraction between points. A natural question, then, would be to find optimal conditions under which the regularity assumption holds.

Secondly, we contrast Assumption 6.1.2 with the stochastic domination bound for CUE:

$$\operatorname{tr} \frac{(zU^*)^2}{1 - zU^*} \prec \frac{|z|^2}{N(1 - |z|)^2} \tag{6.52}$$

which holds uniformly in $\mathbb{D}_\delta$. The proof of (6.52) is straightforward from Johansson's lemma, which we briefly recall for the convenience of the reader.

Lemma 6.A.1 (Lemma 2.9 in [92])**.** *Let U be a CUE matrix and f be a real function defined on the unit disk. Further, define*

$$\|f\|_H^2 = \sum_{k \in \mathbb{Z}} |k||\hat{f}_k|^2, \qquad \hat{f}_k = \frac{1}{2\pi} \int_0^{2\pi} f(\mathrm{e}^{\mathrm{i}\theta}) \mathrm{e}^{-\mathrm{i}k\theta} \mathrm{d}\theta.$$

If $\|f\|_H < \infty$, then

$$\mathbb{E}\left(\mathrm{e}^{\operatorname{tr} f(U)}\right) \leq \mathrm{e}^{N\hat{f}_0 + \frac{1}{2}\|f\|_H^2}.$$

Note that (6.52) may be interpreted (up to normalization) as an averaged local law for the same object as Assumption 6.1.2. The bound (6.52) is more precise and relies on level repulsion (but not on any assumption concerning the eigenvectors, which do not play any role here). It also follows from Lemma 6.A.1 that, in the CUE case, the bound in the regularity assumption can be strengthened to $|z|^2(N(1 - |z|))^{-1/2}$, an improvement that relies on level repulsion and so holds with less generality. Moreover, this does not affect the estimates related to the emergence of an outlier, only the bound on all eigenvalues close to the unit circle in the supercritical regimes.

References

[1] D. Abanin, W. De Roeck, W. W. Ho, and F.Huveneers. A rigorous theory of many-body prethermalization for periodically driven and closed quantum systems. *Commun. Math. Phys.*, 354:809–827, 2017.

[2] A. Adhikari and Z. Che. Edge universality of correlated Gaussians. *Electron. J. Probab.*, 24, 2019.

[3] A. Aggarwal, P. Lopatto, and J. Marcinek. Eigenvector statistics of Lévy matrices. *Ann. Probab.*, 49(4):1778–1846, 2021.

[4] O. H. Ajanki, L. Erdős, and T. Krüger. Stability of the matrix Dyson equation and random matrices with correlated entries. *Probab. Theory Relat. Fields*, 173:293–373, 2019.

[5] G. Akemann, J. Baik, and P. Di Francesco (eds.). *The Oxford Handbook of Random Matrix Theory*. Oxford University Press, 2015.

[6] J. Alt, L. Erdős, and T. Krüger. The Dyson equation with linear self-energy: Spectral bands, edges and cusps. *Doc. Math.*, 25:1421–1539, 2020.

[7] J. Alt, L. Erdős, T. Krüger, and D. Schröder. Correlated random matrices: Band rigidity and edge universality. *Ann. Probab.*, 48(2):963–1001, 2020.

[8] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, 2009.

[9] G. W. Anderson and O. Zeitouni. A law of large numbers for finite-range dependent random matrices. *Commun. Pure Appl. Math.*, 61(8):1118–1154, 2008.

[10] G. Ben Arous and A. Guionnet. The spectrum of heavy tailed random matrices. *Commun. Math. Phys.*, 278:pages 715–751, 2008.

[11] 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.

[12] Z. D. Bai and J. Yao. On the convergence of the spectral empirical process of Wigner matrices. *Bernoulli*, 11:1059–1092, 2005.

[13] Z. D. Bai and Y. Q. Yin. Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix. *Ann. Prob.*, 16(4):1729–1749, 1988.

[14] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697, 2005.

[15] B. N. Balz and P. Reimann. Typical relaxation of isolated many-body systems which do not thermalize. *Phys. Rev. Lett.*, 118:190601, 2017.

[16] M. Banna, F. Merlevède, and M. Peligrad. On the limiting spectral distribution for a large class of symmetric random matrices with correlated entries. *Stochastic Process. Appl.*, 125(7):2700–2726, 2015.

[17] Z. Bao and Y. He. Quantitative CLT for linear eigenvalue statistics of Wigner

matrices. *Ann. Appl. Probab.*, 33(6), 2023.

[18] 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.

[19] R. Bauerschmidt, A. Knowles, J. Huang, and H.-T. Yau. Bulk eigenvalue statistics for random regular graphs. *Ann. Prob.*, 45(6), 2017.

[20] F. Bekerman, T. Leblé, and S. Serfaty. CLT for fluctuations of β-ensembles with general potential. *Electron. J.Probab.*, 23:1–31, 2018.

[21] F. Bekerman and A. Lodhia. Mesoscopic central limit theorem for general β-ensembles. *Ann. Inst. H. Poincaré Probab. Stat.*, 54(4):1917–1938, 2018.

[22] F. Benaych-Georges, C. Bordenave, and A. Knowles. Spectral radii of sparse random matrices. *Ann. Inst. H. Poincaré Probab. Stat.*, 56(3):2141–2161, 2020.

[23] J. Berges, S. Borsányi, and C. Wetterich. Prethermalization. *Phys. Rev. Lett.*, 93(14):142002, 2004.

[24] G. Biroli, C. Kollath, and A. M. Läuchli. Effect of rare fluctuations on the thermalization of isolated quantum systems. *Phys. Rev. Lett.*, 105(25):250401, 2010.

[25] F. Borgonovi, F. M. Izrailev, and L. F. Santosand V. G. Zelevinsky. Quantum chaos and thermalization in isolated systems of interacting particles. *Phys. Rep.*, 626:1–58, 2016.

[26] G. Borot, S. Charbonnier, E. Garcia-Failde, F. Leid, and S. Shadrin. Functional relations for higher-order free cumulants. *Preprint, arXiv:2112.12184*, 2021.

[27] G. Borot and A. Guionnet. Asymptotic expansion of β matrix models in the one-cut regime. *Commun. Math. Phys.*, 317:447–483, 2013.

[28] P. Bourgade. Random band matrices. In *Proc. Int. Cong. of Math.*, volume 4, pages 2777–2802, 2018.

[29] P. Bourgade, C. P. Hughes, A. Nikeghbali, and M. Yor. The characteristic polynomial of a random unitary matrix: A probabilistic approach. *Duke Math. J.*, 145(1):45–69, 2008.

[30] P. Bourgade, K. Mody, and M. Pain. Optimal local law and central limit theorem for β-ensembles. *Commun. Math. Phys.*, 390:1017–1079, 2022.

[31] G. Cipolloni, L. Erdős, and J. Henheik. Eigenstate thermalisation at the edge for Wigner matrices. *Preprint, arXiv:2309.05488*, 2023.

[32] G. Cipolloni, L. Erdős, and J. Henheik. Out-of-time-ordered correlators for Wigner matrices. *Preprint, arXiv:2402.17609*, 2024.

[33] G. Cipolloni, L. Erdős, J. Henheik, and O. Kolupaiev. Gaussian fluctuations in the equipartition principle for Wigner matrices. *Forum Math. Sigma*, 11:E74, 2023.

[34] G. Cipolloni, L. Erdős, J. Henheik, and D. Schröder. Optimal lower bound on eigenvector overlaps for non-Hermitian random matrices. *J. Funct. Anal.*, 287(4):110495, 2024.

[35] 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.

[36] G. Cipolloni, L. Erdős, and D. Schröder. Eigenstate thermalization hypothesis for Wigner matrices. *Commun. Math. Phys.*, 388(2):1005–1048, 2021.

[37] G. Cipolloni, L. Erdős, and D. Schröder. Optimal multi-resolvent local laws for Wigner matrices. *Electron. J. Probab.*, 27:1–38, 2022.

[38] G. Cipolloni, L. Erdős, and D. Schröder. Rank-uniform local law for Wigner matrices. *Forum Math. Sigma*, 10:E96, 2022.

[39] G. Cipolloni, L. Erdős, and D. Schröder. Thermalization for Wigner matrices. *J. Funct. Anal.*, 282(8):109394, 2022.

[40] G. Cipolloni, L. Erdős, and D. Schröder. Functional central limit theorems for Wigner matrices. *Ann. Appl. Probab.*, 33(1):447–489, 2023.

[41] B. Collins and S. Matsumoto. Weingarten calculus via orthogonality relations: new applications. *ALEA*, 14(1):631–656, 2017.

[42] 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.

[43] F. D. Cunden, S. N. Majumdar, and N. O'Connell. Free fermions and α-determinantal processes. *J. Phys. A*, 52(16):165202, 2019.

[44] L. Dabelow. *Predicting quantum many-body dynamics out of equilibrium.* PhD thesis, Universität Bielefeld, 2020.

[45] L. Dabelow and P. Reimann. Typicality of prethermalization. *Phys. Rev. Let.*, 122(14):140601, 2019.

[46] L. Dabelow and P. Reimann. Predicting imperfect echo dynamics in many-body quantum systems. *Z. Naturforsch. A*, 75(5):403–411, 2020.

[47] L. Dabelow and P. Reimann. Relaxation theory for perturbed many-body quantum systems versus numerics and experiment. *Phys. Rev. Let.*, 124:120602, 2020.

[48] L. Dabelow and P. Reimann. Typical relaxation of perturbed quantum many-body systems. *J. Stat. Mech.: Theory Exp.*, page 013106, 2021.

[49] E. B. Davies. The functional calculus. *J. London Math. Soc.*, 52(1):166–176, 1995.

[50] A. Boutet de Monvel, A. Khorunzhy, and V. Vasilchuk. Limiting eigenvalue distribution of random matrices with correlated entries. *Markov Process. Related Fields*, 2(4):607–636, 1996.

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

[52] M. Diaz, A. Jaramillo, and J. C. Pardo. Fluctuations for matrix-valued Gaussian processes. *Ann. Henri Poincaré*, 58(4):2216–2249, 2022.

[53] 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.*, page 2250044, 2022.

[54] J. Dieplinger and S. Bera. Finite-size prethermalization at the chaos-to-integrable crossover. *Phys. Rev. B*, 107:224207, 2023.

[55] G. Dubach. Powers of Ginibre eigenvalues. *Electron. J. Probab.*, 23:1–31, 2018.

[56] G. Dubach. Explicit formulas concerning eigenvectors of weakly non-unitary matrices. *Electron. Commun. Probab.*, 28:1–11, 2023.

[57] G. Dubach and L. Erdős. Dynamics of a rank-one perturbation of a Hermitian matrix. *Electron. Commun. Probab.*, 28:1–13, 2023.

[58] G. Dubach and J. Reker. Dynamics of a rank-one multiplicative perturbation of a unitary matrix. *To appear in: Random Matrices: Theory Appl., online ready available at https://doi.org/10.1142/S2010326324500072*, 2024.

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

[60] L. Erdős, J. Henheik, J. Reker, and V. Riabov. Prethermalization for deformed Wigner matrices. *Preprint, arXiv:2310.06677*, 2023.

[61] L. Erdős and H. C. Ji. Functional CLT for non-Hermitian random matrices. *Ann. Inst. H. Poincaré Probab. Stat.*, 59(4):2083–2105, 2023.

[62] 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(59):1–58, 2013.

[63] L. Erdős and P. Mühlbacher. Bounds on the norm of Wigner-type random matrices. *Random Matrices: Theory Appl.*, 8(3):1950009, 2019.

[64] L. Erdős, B. Schlein, and H.-T. Yau. Wegner estimate and level repulsion for Wigner random matrices. *Int. Math. Res. Not. IMRN*, 2010(3):436–479, 2010.

[65] L. Erdős, B. Schlein, and H.-T. Yau. Universality of random matrices and local relaxation flow. *Inv. Math.*, 185:75–119, 2011.

[66] L. Erdős, B. Schlein, H.-T. Yau, and J. Yin. The local relaxation flow approach to universality of the local statistics for random matrices. *Ann. Inst. H. Poincaré Probab. Statist.*, 48(1):1–46, 2012.

[67] L. Erdős and H.-T. Yau. A dynamical approach to random matrix theory. Courant Lecture Notes in Mathematics 28. 2017.

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

[69] L. Erdős, T. Krüger, and D. Schröder. Random matrices with slow correlation decay. *Forum Math. Sigma*, 7:E8, 2019.

[70] R. Rashidi Far, T. Oraby, W. Bryc, and R. Speicher. On slow-fading MIMO systems with nonseparable correlation. *IEEE Trans. Inform. Theory*, 54(2):544–553, 2008.

[71] P. J. Forrester. Rank 1 perturbations in random matrix theory – a review of exact results. *Random Matrices: Theory Appl.*, 12(4):2330001, 2023.

[72] P. J. Forrester and J. R. Ipsen. A generalization of the relation between zeros of the complex Kac polynomial and eigenvalues of truncated unitary matrices. *Prob. Theory Relat. Fields*, 31:833–847, 2019.

[73] Y. V. Fyodorov. Spectra of random matrices close to unitary and scattering theory for discrete-time systems. In *Disordered and complex systems*, volume 553 of *AIP Conference Proceedings*, pages 191–196. Amer. Inst. Phys., Melville, NY, 2001.

[74] Y. V. Fyodorov and B. A. Khoruzhenko. Extreme eigenvalues of random sub-unitary matrices: from Fréchet to Gumbel. *in prep.*

[75] Y. V. Fyodorov and B. A. Khoruzhenko. Systematic analytical approach to correlation functions of resonances in quantum chaotic scattering. *Phys. Rev. Lett.*, 83:65–68, 1999.

[76] Y. V. Fyodorov, B. A. Khoruzhenko, and M. Poplavskyi. Extreme eigenvalues and the emerging outlier in rank-one non-Hermitian deformations of the Gaussian Unitary Ensemble. *Entropy*, 25(1):74, 2023.

[77] Y. V. Fyodorov and D. V. Savin. Resonance scattering of waves in chaotic systems. In *The Oxford Handbook of Random Matrix Theory*, Oxford Handbooks in Mathematics, pages 703–722. Oxford University Press, Oxford, 2011.

[78] Y. V. Fyodorov and H. J. Sommers. Statistics of S-matrix poles in few-channel chaotic scattering: Crossover from isolated to overlapping resonances. *JETP Lett.*, 63(12):1026–1030, 1996.

[79] Y. V. Fyodorov and H. J. Sommers. Statistics of resonance poles, phase shifts and time delays in quantum chaotic scattering: Random matrix approach for systems with broken time-reversal invariance. *J. Math. Phys.*, 38(4):1918–1981, 1997.

[80] Y. V. Fyodorov and H. J. Sommers. Random matrices close to Hermitian or unitary: overview of methods and results. *J. Phys. A Math.*, 36(12):3303, 2003.

[81] Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1:233–241, 1981.

[82] M. Ledoux G. Lambert and C. Webb. Quantitative normal approximation of linear statistics of β-ensembles. *Ann. Probab.*, 47(5):2619–2685, 2019.

[83] M. Gallone and B. Langella. Prethermalization and conservation laws in quasi-periodically driven quantum systems. *Preprint, arXiv:2306.14022*, 2023.

[84] C. Gogolin and J. Eisert. Equilibration, thermalisation, and the emergence of statistical mechanics in closed quantum systems. *Rep. Prog. Phys.*, 79(5):056001, 2016.

[85] 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.

[86] W. Hachem, P. Loubaton, and J. Najim. The empirical eigenvalue distribution of a gram matrix: from independence to stationarity. *Markov Process. Related Fields*, 11(4):629–648, 2005.

[87] Y. He. Mesoscopic linear statistics of Wigner matrices of mixed symmetry class. *J. Stat. Phys.*, 175:932–959, 2019.

[88] Y. He and A. Knowles. Mesoscopic eigenvalue statistics of Wigner matrices. *Ann. Appl. Probab.*, 27(3):1510–1550, 2017.

[89] Y. He and A. Knowles. Mesoscopic eigenvalue density correlations of Wigner matrices. *Probab. Theory Relat. Fields*, 177:147–216, 2020.

[90] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal processes and independence. *Probab. Surv.*, 3:206–229, 2006.

[91] 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.

[92] K. Johansson. On random matrices from the compact classical groups. *Ann. Math.*, 145(3):519–545, 1997.

[93] K. Johansson. On fluctuations of eigenvalues of random Hermitian matrices. *Duke Math. J.*, 91(1):151–204, 1998.

[94] F. Juhász. On the spectrum of a random graph. In *Algebraic methods in graph theory. Coll. Math. Soc. J. Bolyai*, volume 25, pages 313–316. North Holland, 1981.

[95] A. M. Khorunzhi, 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.

[96] A. M. Khorunzhi, B. A. Khoruzhenko, and L. A. Pastur. Asymptotic properties of large random matrices with independent entries. *J. Math. Phys.*, 37:5033–5060, 1996.

[97] M. Kollar, F. A. Wolf, and M. Eckstein. Generalized Gibbs ensemble prediction of prethermalization plateaus and their relation to nonthermal steady states in integrable systems. *Phys. Rev. B*, 84:054304, 2011.

[98] E. Kostlan. On the spectra of Gaussian matrices. *Linear Algebra Appl.*, 162, 1992.

[99] B. Landon and P. Sosoe. Almost-optimal bulk regularity conditions in the CLT for Wigner matrices. *Preprint, arXiv:2204.03419*, 2022.

[100] T. Langen, T. Gasenzer, and J. Schmiedmayer. Prethermalization and universal dynamics in near-integrable quantum systems. *J. Stat. Mech.*, 2016:064009, 2016.

[101] 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.

[102] Y. Li and Y. Xu. On fluctuations of global and mesoscopic linear statistics of generalized Wigner matrices. *Bernoulli*, 27(2):1057–1076, 2021.

[103] A. Lytova. On non-Gaussian limiting laws for certain statistics of Wigner matrices. *Zh. Mat. Fiz. Anal. Geom.*, 9:536–581, 2013.

[104] 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.

[105] A. Lytova and L. Pastur. Fluctuations of matrix elements of regular functions of Gaussian random matrices. *J. Stat. Phys.*, 134:147–159, 2009.

[106] C. Male. Freeness over the diagonal and global fluctuations of complex Wigner matrices. *J. Oper. Theory*, 85(1):153–181, 2021.

[107] 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.

[108] K. Mallayya, M. Rigol, and W. De Roeck. Prethermalization and thermalization in isolated quantum systems. *Phys. Rev. X*, 9:021027, 2019.

[109] F. Maunoury. Necessary and sufficient conditions for the existence of α-determinantal

processes. In *Séminaire de Probabilités XLVIII*, pages 423–444. Springer, 2016.

[110] E. S. Meckes. *The random matrix theory of the classical compact groups*. Cambridge University Press, 2019.

[111] J. A. Mingo and R. Speicher. *Free Probability and Random Matrices*. Vol. 35, Fields Institute Research Monographs, Springer, New York, 2017.

[112] J.A. Mingo and A. Nica. Annular noncrossing permutations and partitions, and second-order asymptotics for random matrices. *Int. Math. Res. Not.*, 2004(28):1413–1460, 2004.

[113] J.A. Mingo and R. Speicher. Second order freeness and fluctuations of random matrices I: Gaussian and Wishart random matrices and cyclic Fock spaces. *J. Funct. Anal.*, 235(1):226–270, 2006.

[114] M. Moeckel and S. Kehrein. Interaction quench in the Hubbard model. *Phys. Rev. Let.*, 100:175702, 2008.

[115] T. Mori, T.N. Ikeda, E. Kaminishi, and M. Ueda. Thermalization and prethermalization in isolated quantum systems: a theoretical overview. *J. Phys. B: At. Mol. Opt. Phys.*, 51(11):112001, 2018.

[116] C. Nation and D. Porras. Quantum chaotic fluctuation-dissipation theorem: Effective Brownian motion in closed quantum systems. *Phys. Rev. E*, 99:052139, 2019.

[117] M. Ottolini. Spectral norm of random matrices with non-identically distributed entries. Master's thesis, University Pisa, 2017.

[118] C. E. I. Redelmeier. Real second-order freeness and the asymptotic real second-order freeness of several real matrix models. *Int. Math. Res. Not.*, 2014(12):3353–3395, 2014.

[119] C. E. I. Redelmeier. Real and quaternionic second-order free cumulants and connections to matrix cumulants. *Preprint, arXiv:1808.10589v2*, 2018.

[120] P. Reimann. Typical fast thermalization processes in closed many-body systems. *Nat. Comm.*, 7:10821, 2016.

[121] J. Reker. On the operator norm of a Hermitian random matrix with correlated entries. *Random Matrices: Theory Appl.*, 11(4):2250036, 2022.

[122] J. Reker. Multi-point functional central limit theorem for Wigner matrices. *Preprint, arXiv.2307.11028*, 2023.

[123] J. Reker. Fluctuation moments for regular functions of Wigner matrices. *Math. Phys. Anal. Geom.*, 27:10, 2024.

[124] V. Riabov. Linear eigenvalue statistics at the cusp. *Preprint, arXiv:2307.07432*, 2023.

[125] V. Riabov. Mesoscopic eigenvalue statistics for Wigner-type matrices. *Preprint, arXiv:2301.01712*, 2023.

[126] Z. Rudnick and P. Sarnak. The behavior of eigenstates of arithmetic hyperbolic manifolds. *Comm. Math. Phys.*, 161:195–213, 1994.

[127] J. H. Schenker and H. Schulz-Baldes. Semicircle law and freeness for random matrices with symmetries or correlations. *Math. Res. Lett.*, 12(4):531–542, 2005.

[128] 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.

[129] M. Shcherbina. Fluctuations of linear eigenvalue statistics of β matrix models in the multi-cut regime. *J. Stat. Phys.*, 151:1004–1034, 2013.

[130] J. W. Silverstein. Weak convergence of random functions defined by the eigenvectors of sample covariance matrices. *Ann. Probab.*, 18(3):1174–1194, 1990.

[131] 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.

[132] 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.

[133] 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.

[134] 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.

[135] P. Sosoe and P. Wong. Regularity conditions in the CLT for linear eigenvalue statistics of Wigner matrices. *Adv. Math.*, 249:37–87, 2013.

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

[137] T. Tao. Outliers in the spectrum of iid matrices with bounded rank perturbations. *Probab. Theory Relat. Fields*, 155:231—263, 2013.

[138] M. Ueda. Quantum equilibration, thermalization and prethermalization in ultracold atoms. *Nat. Rev. Phys.*, 2:669–681, 2020.

[139] V. H. Vu. Spectral norm of random matrices. *Combinatorica*, 27(6):721–736, 2007.

[140] E. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Ann. Math.*, 62:548–564, 1955.

[141] J. Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20, 1928.