Institute of Science and Technology Austria
Klosterneuburg, Austria

DOCTOR OF PHILOSOPHY

From Dyson to Pearcey: Universal Statistics in Random Matrix Theory

Dominik Schröder

March 2019

The thesis of Dominik Schröder, titled *From Dyson to Pearcey: Universal Statistics in Random Matrix Theory*, is approved by:

Supervisor: László Erdős, *IST Austria*

Committee Member: Jan Maas, *IST Austria*

Committee Member: Gerald Teschl, *University of Vienna*

Defence chair: Krishnendu Chatterjee, *IST Austria*

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.

Dominik Schröder
March 2019

Abstract

In the first part of this thesis we consider large random matrices with arbitrary expectation and a general slowly decaying correlation among its entries. We prove universality of the local eigenvalue statistics and optimal local laws for the resolvent in the bulk and edge regime. The main novel tool is a systematic diagrammatic control of a multivariate cumulant expansion.

In the second part we consider Wigner-type matrices and show that at any cusp singularity of the limiting eigenvalue distribution the local eigenvalue statistics are universal and form a Pearcey process. Since the density of states typically exhibits only square root or cubic root cusp singularities, our work complements previous results on the bulk and edge universality and it thus completes the resolution of the Wigner-Dyson-Mehta universality conjecture for the last remaining universality type. Our analysis holds not only for exact cusps, but approximate cusps as well, where an extended Pearcey process emerges. As a main technical ingredient we prove an optimal local law at the cusp, and extend the fast relaxation to equilibrium of the Dyson Brownian motion to the cusp regime.

In the third and final part we explore the entrywise linear statistics of Wigner matrices and identify the fluctuations for a large class of test functions with little regularity. This enables us to study the rectangular Young diagram obtained from the interlacing eigenvalues of the random matrix and its minor, and we find that, despite having the same limit, the fluctuations differ from those of the algebraic Young tableaux equipped with the Plancharel measure.

Acknowledgment

First of all, I would like to express my sincere gratitude to my advisor László Erdős. I cannot imagine a more encouraging and productive research environment than the one under László's supervision at IST Austria. I am thankful for his patience, his tireless support and his breadth of knowledge.

I would like to thank the members of my thesis committee, Gerald Teschl and Jan Maas, for their support and encouragement throughout the last four years of my PhD studies.

I also would like to express my gratitude to my colleagues for creating such a supportive and stimulating research environment at IST Austria. In particular, I would like to thank my fellow PhD students Johannes Alt and Giorgio Cipolloni and former group member Torben Krüger for the productive discussions and fruitful collaborations.

I gratefully acknowledge the financial support from the *IST Austria Excellence Scholarship* and my advisor's *ERC Advanced Grant No. 338804*.

Last but not least, I would like to thank my friends and family, in particular my parents Gabriele Walker-Schröder and Joachim Schröder, and Viola Vetter. I am immensely grateful for their support in my studies, and my life, in general.

About the Author

After studying at ETH Zurich, LMU Munich and the University of Cambridge, Dominik Schröder received his B.S. in *Mathematics* in 2013, his M.S. in *Theoretical in Mathematical Physics* in 2014, and his M.A.St. degree in *Mathematics* in 2015. In September 2015, he then joined IST Austria for his PhD studies in the research group of László Erdős. His research interests lie in probability theory, with a focus on the analysis of universal properties of random matrices.

Included Publications

[DS1] L. Erdős and D. Schröder, *Fluctuations of rectangular Young diagrams of interlacing Wigner eigenvalues*, Int. Math. Res. Not. IMRN, 3255–3298 (2016), MR3805203.

[DS2] L. Erdős and D. Schröder, *Fluctuations of functions of Wigner matrices*, Electron. Commun. Probab. **21**, Paper no. 86, 15 (2016), MR3600514.

[DS3] L. Erdős, T. Krüger, and D. Schröder, *Random matrices with slow correlation decay*, to appear in Forum Math. Sigma (2017), arXiv:1705.10661.

[DS4] J. Alt, L. Erdős, T. Krüger, and D. Schröder, *Correlated random matrices: Band rigidity and edge universality*, preprint (2018), arXiv:1804.07744.

[DS5] L. Erdős, T. Krüger, and D. Schröder, *Cusp universality for random matrices I: Local law and the complex Hermitian case*, preprint (2018), arXiv:1809.03971.

[DS6] G. Cipolloni, L. Erdős, T. Krüger, and D. Schröder, *Cusp universality for random matrices II: The real symmetric case.* preprint (2018), arXiv:1811.04055.

Contents

List of Figures

Introduction 1

Random matrices, i.e. matrices with random entries, have first been proposed as models in mathematical statistics by [178]. More intensive studies of random matrices and their spectral properties began with [176], where Wigner suggested that the energy levels of heavy nuclei are distributed like the eigenvalues of large Hermitian random matrices. Later, the study of the spectral properties of random matrices gained importance also in other areas of physics and mathematics, and now is a very active, flourishing field. Examples include quantum chaos [32], disordered quantum systems [69], wireless communications [59], the error analysis of numerical algorithms [68], the zeros of the Riemann zeta function [113] and random neural networks [150].

1.1 Hermitian random matrix models of increasing generality

Within this thesis we exclusively work on Hermitian random matrix models which arise as generalizations of the *Wigner matrices* introduced in [176]. Other extensively studied random matrix models include *invariant ensembles*, *sample-covariance matrices* and *non-Hermitian random matrices*. To clarify the terminology we briefly outline the *mean field* ensembles considered within this thesis, which, in the random matrix context, means that all matrix entries have variances of comparable sizes.

GOE (Gaussian orthogonal ensemble): Matrices $W = W^* \in \mathbb{R}^{N\times N}$ such that the upper-triangular entries are independent zero mean Gaussian random variables satisfying $\mathbf{E}\, w_{aa}^2 = 2/N$, and $\mathbf{E}\, w_{ab}^2 = 1/N$ for $a \neq b$.

GUE (Gaussian unitary ensemble): Matrices $W = W^* \in \mathbb{C}^{N\times N}$ such that the upper-triangular entries are independent zero mean Gaussians such that $w_{aa} \in \mathbb{R}$ with $\mathbf{E}\, w_{aa}^2 = 1/N$ and $w_{ab} \in \mathbb{C}$ with $\mathbf{E}\, w_{ab}^2 = 0$ and $\mathbf{E}\, |w_{ab}|^2 = 1/N$ for $a \neq b$.

Wigner matrices: Matrices $W = W^* \in \mathbb{C}^{N\times N}$ such that the upper-triangular entries $\{\, w_{ab} \mid a \leq b \,\}$ are independent, the off-diagonal entries $\{\, w_{ab} \mid a < b \,\}$ are identically distributed with $\mathbf{E}\, w_{ab} = 0$, $\mathbf{E}\, |w_{ab}|^2 = 1/N$, and the diagonal entries w_{aa} are identically distributed with $\mathbf{E}\, w_{aa} = 0$, $c/N \leq \mathbf{E}\, |w_{aa}|^2 \leq C/N$ for some positive N-independent constants c, C.

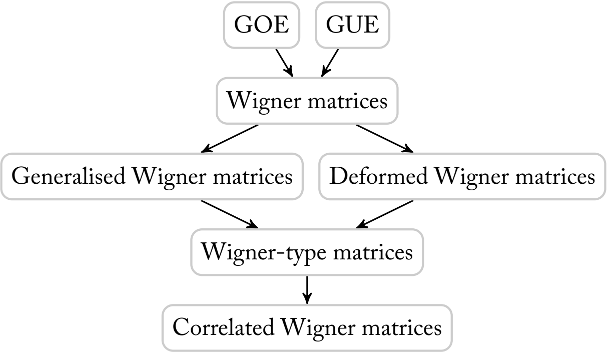


Figure 1.1: Hierarchy of increasing generality of random matrix models. Arrows from A to B indicate that A is a special case of B.

Generalised Wigner matrices: Matrices $W = W^* \in \mathbb{C}^{N\times N}$ such that the upper-triangular entries $\{ w_{ab} \mid a \le b \}$ are independent and satisfy $\mathbf{E}\, w_{ab} = 0, \sum_b s_{ab} = 1 + \mathcal{O}\left(N^{-1}\right)$ and $c/N \le s_{ab} \le C/N$, where $s_{ab} := \mathbf{E}\, |w_{ab}|^2$.

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

Wigner-type matrices: Matrices $H = A + W \in \mathbb{C}^{N\times N}$ such that $A = A^* = \mathbf{E}\, H$ is diagonal and the upper-triangular entries $\{ w_{ab} \mid a \le b \}$ of $W = W^*$ are independent and satisfy $\mathbf{E}\, w_{ab} = 0$ and $c/N \le s_{ab} \le C/N$.

Correlated Wigner matrices: Matrices of the form $H = A + W = H^* \in \mathbb{C}^{N\times N}$, where $A = A^* = \mathbf{E}\, H$ and the *covariance operator* $\mathcal{S}[R] := WRW$ satisfies

$$cN^{-1} \operatorname{Tr} R \le \mathcal{S}[R] \le CN^{-1} \operatorname{Tr} R$$

for any positive semidefinite matrices R in the sense of quadratic forms.

One readily checks that these models form a hierarchy of increasing generality as indicated in Figure 1.1, in particular all models except for GOE/GUE allow for both the complex Hermitian and real symmetric symmetry classes. We stress that correlated Wigner matrices not only allow for general correlation structures but also general expectations. Note that the above models are scaled in such a way that the spectrum remains bounded as N grows since the average expected squared eigenvalue modulus is given by

$$\frac{1}{N} \mathbf{E} \sum_i \lambda_i^2 = \frac{1}{N} \mathbf{E} \operatorname{Tr} H^2 = \frac{1}{N} \sum_{i,j} \mathbf{E}\, |h_{ij}|^2 \sim 1.$$

We study the spectral properties of these random matrix models on three different scales, which we will now describe, with a focus on the very different mathematical techniques their analysis requires.

1.2 Global scale

Given a $N \times N$ random matrix $H = H^* = (h_{ij})_{i,j=1}^N$ and its eigenvalues $\lambda_1, \dots, \lambda_N$, a central object in the study of the spectral properties is the *empirical spectral distribution*

(ESD)

$$\mu_N := \frac{1}{N}\sum_{n=1}^N \delta_{\lambda_n}.$$

The weak convergence of μ_N to some deterministic measure μ is called a *global law* since it allows to predict the approximate proportion of eigenvalues in intervals of small, but N-independent size. For example, Wigner's *semicircle law* [177] states that for Wigner matrices and a wide range of N-independent test functions f, it holds that almost surely that

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

where $\rho_{\mathrm{sc}}(x) := \sqrt{(4-x^2)_+}/2\pi$ is the density of the semicircular distribution.

There are at least three different approaches for determining the limiting spectral measure μ for a given random matrix ensemble, the moment method and two derivations based on the analysis of resolvents. Those methods are increasingly more powerful and general and were developed parallel to the study of the increasingly general models.

1.2.1 Moment method

In [177] Wigner used a simple tree counting argument to show that under mild additional high-moment assumptions it holds that

$$\mathbf{E}\int x^k\,\mathrm{d}\mu_N(x) = \mathbf{E}\frac{1}{N}\operatorname{Tr} H^k = \begin{cases} C_{k/2} & k \text{ even}, \\ 0 & k \text{ odd}\end{cases} + \mathcal{O}\left(\frac{1}{N}\right), \tag{1.1}$$

where

$$C_n := \frac{1}{n+1}\binom{2n}{n} = \binom{2n}{n} - \binom{2n}{n+1},$$

is commonly known as the n-th *Catalan number*. To conclude from (1.1), and an additional bound on the variance, that the ESD follows the semircircular distribution μ_{sc} we could simply compute that

$$\int_{-2}^2 x^k\rho_{\mathrm{sc}}(x)\,\mathrm{d}x = \int_{-2}^2 \frac{x^k\sqrt{4-x^2}}{2\pi}\,\mathrm{d}x = \begin{cases} C_{k/2} & \text{if } k \text{ is even}, \\ 0 & \text{else.}\end{cases} \tag{1.2}$$

But the method of identifying μ_{sc} through its moments via (1.2) has the disadvantage that it requires solving the *moment problem*, or somehow guessing that the semicircular distribution might have the Catalan numbers as its moments.

In order to identify the measure μ from the moments more directly, we can alternatively compute the *Stieltjes transform*

$$m_\mu(z) := \int_{\mathbb{R}} \frac{1}{x-z}\,\mathrm{d}\mu(x)$$

for $z\in\mathbb{H}$ in the upper half plane $\mathbb{H} := \{\, z\in\mathbb{C} \mid \Im z > 0 \,\}$ as

$$\begin{aligned} m_\mu(z) &= -\frac{1}{z}\int_{\mathbb{R}} \frac{1}{1-x/z}\,\mathrm{d}\mu(x) = -\sum_{k=0}^\infty \frac{1}{z^{k+1}}\int_{\mathbb{R}} x^k\,\mathrm{d}\mu(x) \\ &= -\sum_{k=0}^\infty \frac{(2k)!}{k!(k+1)!z^{2k+1}} = \frac{\sqrt{z^2-4}-z}{2}, \end{aligned} \tag{1.3}$$

where the correct branch of the square root has to be chosen such that $\Im m_\mu(z) > 0$ whenever $\Im z > 0$ and $m_\mu(z) \to 0$ as $|z| \to \infty$. From (1.3) we quickly obtain a formula for the density ρ associated with μ by means of *Stieltjes inversion*. The imaginary part of the Stieltjes transform

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

at $z = x + \mathrm{i}\eta$ yields, up to a factor of π, the convolution of μ with the Poisson-kernel of width η. Therefore we can conclude from (1.3) that

$$\frac{\mathrm{d}\mu(x)}{\mathrm{d}x} = \lim_{\eta \searrow 0} \Im \frac{\sqrt{(x+\mathrm{i}\eta)^2 - 4} - x - \mathrm{i}\eta}{2\pi} = \frac{\sqrt{(4-x^2)_+}}{2\pi} = \rho_{\mathrm{sc}}(x).$$

An alternative interpretation of (1.3) which is more suitable for the subsequent generalisations is that the Stieltjes transform m_{sc} of ρ_{sc} is the unique function solving the equation

$$m = -(z+m)^{-1}, \quad \Im m > 0. \tag{1.5}$$

1.2.2 Resolvent method

Instead of computing the Stieltjes transform indirectly via the moments, we find from spectral calculus that

$$m_{\mu_N}(z) = \frac{1}{N} \sum_n (\lambda_n - z)^{-1} = \frac{1}{N} \operatorname{Tr}(H - z)^{-1} = \frac{1}{N} \operatorname{Tr} G(z) =: \langle G(z) \rangle\,,$$

i.e. that the Stieltjes transform of the ESD is given by the normalised trace of the resolvent $G(z) := (H - z)^{-1}$. Given the particularly simple formula (1.3), it is reasonable to hope that we can see more directly, without computing any moments, that the normalised trace $\langle G \rangle$ of the resolvent approximately is given by (1.3).

Schur complement formula

It follows from the well known *Schur complement formula* that for Wigner-type matrices $H = W + \operatorname{diag}(\boldsymbol{a})$ and $i \in [N] := \{1, \dots, N\}$,

$$G_{ii} = \frac{1}{a_i + w_{ii} - z - \langle w_i, G^{(i)} w_i \rangle}, \tag{1.6}$$

where $w_i \in \mathbb{C}^{N-1}$ denotes the i-th column of W with the entry w_{ii} removed, and $G^{(i)} := (H^{(i)} - z)^{-1}$ is the resolvent of the matrix $H^{(i)}$ obtained by removing the i-th row and column from H. Since w_i is independent from $G^{(i)}$ it follows that

$$\langle w_i, G^{(i)} w_i \rangle = \sum_{a,b} \overline{w_{ai}} G^{(i)}_{ab} w_{bi} \approx \sum_a (\mathbf{E}\,|w_{ai}|^2) G^{(i)}_{aa} = \sum_a s_{ia} G^{(i)}_{aa}$$

by a quadratic large deviation estimate conditional on $H^{(i)}$. To show that this approximation holds with high probability, some additional moment conditions have to be imposed on W.

Assuming that the resolvent G is stable in the sense that $G_{aa}^{(i)}$ is comparable with G_{aa} for $a \neq i$, it follows that G approximately satisfies

$$G_{ii} \approx \frac{1}{a_i - z - \sum_a s_{ia} G_{aa}}, \qquad \text{or, more compactly,} \quad \boldsymbol{g} \approx -(z - \boldsymbol{a} + S\boldsymbol{g})^{-1}, \tag{1.7}$$

where[1] $\boldsymbol{g} := \operatorname{diag}(G)$ and the inversion should be understood entrywise. The relation (1.7) suggests that we can find a deterministic approximation for G by solving the *quadratic vector equation (QVE)*

$$\boldsymbol{m} = -(z - \boldsymbol{a} + S\boldsymbol{m})^{-1}, \quad \Im \boldsymbol{m} > 0 \tag{1.8}$$

for $\Im z > 0$. The side condition $\Im m_i > 0$ ensures the existence of a unique solution and is also fulfilled by the resolvent itself, $\Im G_{ii} > 0$ due to the self-adjointness of H. To make the approximation $G \approx \operatorname{diag}(\mathbf{m})$ (also implying that the off-diagonal entries of G are small) precise, one has to analyse the *stability* of (1.8) with respect to small perturbations. The density obtained from $\langle \boldsymbol{m} \rangle$ via (1.4) is commonly referred to as the *self-consistent density* since it is obtained via solving a self-consistent equation for $\boldsymbol{m}$. We note that in the previously considered case of Wigner matrices (1.8) simplifies to the scalar equation (1.5). Compared to the moment method, the resolvent approach is more robust and in some sense also more canonical as it does not require computing high moments, and no analytic identity for computing the moment generating function is needed.

Cumulant expansion

Once correlations are present in W, the Schur complement formula (1.6) becomes less useful since analysing $\langle w_i, G^{(i)} w_i \rangle$ becomes more involved. There is, however, another approach of deriving (1.8) and its analogue for correlated matrices relying on the simple integration by parts identity

$$\mathbf{E}\, x f(x) = \widetilde{\mathbf{E}}\, \mathbf{E}\, \widetilde{x}^2 f'(x) \tag{1.9}$$

for zero mean Gaussian random variables x and differentiable functions f, where $\widetilde{x}$ denotes an independent copy of x with expectation $\widetilde{\mathbf{E}}$. The identity (1.9) generalizes to the non-commutative setting of Gaussian matrices (with arbitrary covariances) as

$$\mathbf{E}\, W f(W) = \widetilde{\mathbf{E}}\, \mathbf{E}\, \widetilde{W} (\partial_{\widetilde{W}} f)(W), \tag{1.10}$$

where $\partial_{\widetilde{W}}$ denotes the directional derivative in direction $\widetilde{W}$. Using (1.10) and the identity $(H - z)G = 1$ for $H = A + W$, we find

$$1 = \mathbf{E}\, WG + (A - z)\, \mathbf{E}\, G = -\mathbf{E}\Big[(\widetilde{\mathbf{E}} \widetilde{W} G \widetilde{W}) + (z - A) \Big] G = -\mathbf{E}[z - A + \mathcal{S}[G]] G, \tag{1.11}$$

suggesting that the solution to the *matrix Dyson equation (MDE)*

$$-M^{-1} = z - A + \mathcal{S}[M], \quad \Im M = \frac{M - M^*}{2\mathrm{i}} > 0 \tag{1.12}$$

is a good deterministic approximation for the resolvent G. Note that (1.12) generalises (1.8) in that if $\boldsymbol{m}$ solves (1.8), then $M = \operatorname{diag}(\boldsymbol{m})$ solves (1.12) since for Wigner-type matrices

[1] As a slight abuse of notation we denote both the diagonal vector of a matrix A by $\operatorname{diag}(A)$, as well as the diagonal matrix obtained from a vector $\boldsymbol{a}$ by $\operatorname{diag}(\boldsymbol{a})$.

the covariance operator $\mathcal{S}$ simplifies to $\mathcal{S}[\mathrm{diag}(\boldsymbol{m})] = \mathrm{diag}(S\boldsymbol{m})$. Making the approximation $G \approx M$ rigorous requires a careful stability analysis of (1.12), as well as establishing an approximate but high-probability version of (1.11) for general non-Gaussian random matrices by replacing (1.10) by a suitable *cumulant expansion*. In the scalar setting the standard cumulant expansion formula is

$$\mathbf{E}\, x f(x) = \sum_k \frac{\kappa_{k+1}}{k!}\, \mathbf{E}\, f^{(k)}(x), \tag{1.13}$$

where the cumulants κ_k of the random variable x can, for example, be defined as the coefficients of the cumulant generating function

$$\log \mathbf{E}\, e^{\mathrm{i}tx} = \sum_k \kappa_k \frac{(\mathrm{i}t)^k}{k!}.$$

For an alternative combinatorial definition, and a matrix-valued version of (1.13), we refer the reader to Chapter 2.

1.2.3 Classification of self-consistent densities

Given the self-consistent equation (1.12), it is a natural question, which self-consistent densities can arise from the solutions to (1.12) via Stieltjes-inversion

$$\rho(x) := \pi^{-1} \lim_{\eta \searrow 0} \langle \Im M(x + \mathrm{i}\eta) \rangle .$$

The recent analysis in [12] provides a complete classification of the singularity structure of possible solutions to (1.12), also in the more general setting of von Neumann algebras. In [12] it is shown that M is necessarily $1/3$-Hölder continuous in z, and that ρ has the following properties:

(i) $\operatorname{supp} \rho$ consists of finitely many compact intervals,

(ii) ρ is analytic whenever $\rho > 0$,

(iii) if $\mathfrak{e} \in \partial \operatorname{supp} \rho$ is an *edge point*, then $\rho(\mathfrak{e} \pm x) = c\sqrt{x} + \mathcal{O}(\sqrt{x})$ and $\rho(\mathfrak{e} \mp x) = 0$ for $x \ll 1$ and some constant $c > 0$,

(iv) if $\mathfrak{c} \in \operatorname{supp} \rho$ with $\rho(\mathfrak{c}) = 0$ is a *cusp point*, then $\rho(\mathfrak{c} + x) = c\,|x|^{1/3} + \mathcal{O}(|x|^{1/3})$ for some constant $c > 0$,

(v) no other singularities can occur in ρ.

In this sense, the self-consistent density depicted in Figure 1.2, is a typical example including all possible singularities. We commonly refer to the spectral regimes corresponding to (ii), (iii) and (iv) as the *bulk*, *edge* and *cusp regime*.

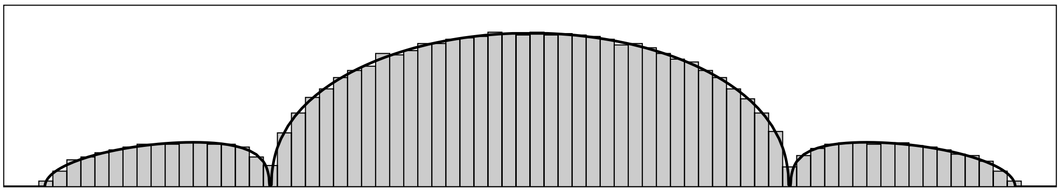


FIGURE 1.2: Typical spectral density of Wigner-type matrices featuring the bulk, cusp and edge regimes. The shaded area shows the histogram of the eigenvalues of a single 2000×2000 random matrix which very closely matches the predicted self-consistent density.

1.3 Mesoscopic scale

In Section 1.2 we sketched three approaches of identifying a deterministic approximation $M = M(z)$ to the resolvent G. According to (1.4), a global law for a random matrix ensemble means effectively establishing a bound on $\Im \langle G - M \rangle$ for $z = x + \mathrm{i}\eta$ for some N-independent $\eta \ll 1$. On this scale, however, $\Im \langle G \rangle$ still involves $\approx \eta N$ eigenvalues and a law of large numbers type result is reasonable to expect. Therefore it is natural to ask for which N-dependent η we can still establish that $\langle G \rangle$ is well approximated by $\langle M \rangle$. Since in the bulk any neighbouring eigenvalues should have an average distance of $1/N$ we can hope for such a *local law* on *mesoscopic scales* $\eta \gg 1/N$. In the most general correlated setting, see Chapters 2 and 3, we will, for example, prove that

$$|\langle \mathbf{x}, (G - M)\mathbf{y} \rangle| \prec \frac{\|\mathbf{x}\| \, \|\mathbf{y}\|}{\sqrt{N\eta}}, \qquad |\langle B(G - M) \rangle| \prec \frac{\|B\|}{N\eta}, \tag{1.14}$$

where $\prec$ is a suitable notion of high probability bound up to N^ϵ-factors, $\eta = \Im z$ and $\mathbf{x}, \mathbf{y}, B$ are arbitrary deterministic vectors and matrices. The bounds in (1.14) exhibit a *fluctuation averaging* feature in the sense that the bound on the average $\langle G - M \rangle$ is an order better than the average bound on individual entries $(G - M)_{aa}$.

The method behind the proof of (1.14) consists of two largely separate arguments, a deterministic and a probabilistic one. Using (1.12) we write

$$\begin{aligned} \mathcal{B}[G - M] &= -MD + M\mathcal{S}[G - M](G - M), \\ \mathcal{B} &:= 1 - M\mathcal{S}[\cdot]M, \quad D := WG + \mathcal{S}[G]G, \end{aligned} \tag{1.15}$$

where we call $\mathcal{B}$ the *stability operator* and D the *error matrix* measuring the deviation of WG from its leading order approximation $-\mathcal{S}[G]G$ coming from the second moment calculation (1.11). The deterministic step in the proof of (1.14) is the analysis of the non-selfadjoint stability operator $\mathcal{B}$, and in particular its smallest eigenvalue β which poses difficulties when solving the quadratic equation (1.15) for $G - M$. This analysis, which in the most general setting is presented in [12], roughly speaking shows that in the spectral bulk we have $|\beta| \sim 1$, close to spectral edges $|\beta| \sim \rho$ and close to cusps $|\beta| \sim \rho^2$. Proofs of mesoscopic local laws in the corresponding spectral regimes pose increasing technical difficulties for this very reason. The smallness of $|\beta|$ needs to be balanced by increasingly stronger estimates on the error matrix D.

The probabilistic part of the argument consists of establishing high-probability bounds on the error matrix D in an isotropic and averaged form

$$|\langle \mathbf{x}, D\mathbf{y}\rangle| \prec \|\mathbf{x}\| \, \|\mathbf{y}\| \sqrt{\frac{\rho}{N\eta}}, \qquad |\langle BD\rangle| \prec \|B\| \, \frac{\rho}{N\eta}. \tag{1.16}$$

On a technical level, (1.16) is proven in a high moment sense which requires identifying the cancellation effect between WG and $\mathcal{S}[G]G$ to high powers by iterated cumulant expansions. In the cusp regime, not even (1.16) is sufficient for establishing (1.14) due to the presence of the eigenvalue $|\beta| \sim \rho^2$. Instead, another input is required, namely, that we have an improved bound of the form $|\langle PMD\rangle| \prec \rho^2/N\eta$, when D is averaged against the eigenmatrix P corresponding to β. This exploits a delicate structural property of D related to the local symmetry of the density ρ around the cusp.

Rigidity, delocalization and absence of eigenvalues outside of the spectrum

An optimal mesoscopic local law as in (1.14) has three important consequences which we briefly mention. From the spectral decomposition of G in terms of the eigenvalues λ_i and ℓ^2-normalised eigenvectors $\boldsymbol{u}_i$ of H we find from the local law (1.14) and the boundedness of M that

$$1 \succ |M_{ii}| + |(G-M)_{ii}| \ge (\Im G)_{ii} = \sum_k \frac{\eta\, |u_k(i)|^2}{(x-\lambda_k)^2 + \eta^2} \gtrsim \frac{|u_k(i)|^2}{\eta}$$

for $z = x + \mathrm{i}\eta$ and x close to λ_k at a distance of η. By choosing $\eta = N^{-1+\epsilon}$ in the spectral bulk it follows that $|u_k(i)| \prec N^{-1/2}$ which means that the ℓ^2-normalised vector $\boldsymbol{u}_k$ is completely *delocalised*.

By a standard argument using a Cauchy-integral formula, we can also conclude from (1.14) that the eigenvalues λ_i are *rigid* in the sense that they satisfy

$$|\lambda_i - \gamma_i| \prec \eta_{\mathrm{f}}(\gamma_i),$$

where the quantiles γ_i and the fluctuation scale η_{f} of ρ are defined as

$$\int_{-\infty}^{\gamma_i} \rho(x)\,\mathrm{d}x = \frac{i}{N}, \quad \int_{x-\eta_{\mathrm{f}}(x)}^{x+\eta_{\mathrm{f}}(x)} \rho(y)\,\mathrm{d}y = \frac{1}{N}.$$

The fluctuation scale is thus comparable with the difference of consecutive quantiles. The fact that eigenvalues fluctuate only on this scale is non-trivial and somewhat unusual. It implies the existence of strong correlations among the eigenvalues. In the bulk, edge and cusp regimes the fluctuation scale is given by $\eta_{\mathrm{f}} \sim N^{-1}$, $N^{-2/3}$ and $N^{-3/4}$, respectively.

Finally, using (1.14) we can exclude the existence of eigenvalues well outside $\operatorname{supp}\rho$ with very high probability. The relevant measure also here is the fluctuation scale as, for example, the extreme eigenvalues can fluctuate only on a scale of $N^{-2/3}$ beyond the support of ρ. More precisely, if we denote two neighbouring support edges by $x, y \in \partial \operatorname{supp}\rho$ with $\rho|_{[x,y]} \equiv 0$, then we can, similarly to the rigidity conclude that, with overwhelming probability, we find no eigenvalues in $[x + N^\epsilon \eta_{\mathrm{f}}(x), y - N^\epsilon \eta_{\mathrm{f}}(y)]$. Additionally, we can also conclude a strong notion of *band rigidity* in the sense that the number of eigenvalues close to any support interval of ρ is deterministic with overwhelming probability as long as the support interval is separated by at least $N^{-3/4+\epsilon}$ from the neighbouring support intervals.

1.4 Microscopic scale

On the scale of the eigenvalue spacing the fluctuation of individual eigenvalues becomes relevant. It has first been conjectured by Wigner in the 1950's, and subsequently formalized as the *Wigner-Dyson-Mehta (WDM) conjecture* [135], that the local statistics of eigenvalues is universal in the sense that their distribution is independent of any model specifics. The emerging distributions should be viewed as the *random matrix analogue of the central limit theorem* and the normal distribution. They depend only on the symmetry class of the matrix and the local singularity type of the density, but on no other model specifics. The classification from Section 1.2.3 indicates that, up to scaling by a constant, the density around single eigenvalues can only exhibit three different behaviours. Within the spectral bulk the density ρ is positive and real analytic and therefore, on the scale $1/N$ of individual eigenvalues, is essentially constant. Around edges and cusps the density is simply given by $x^{1/2}$ or $x^{1/3}$, indicating that the typical distance between consecutive eigenvalues is $N^{-2/3}$ and $N^{-3/4}$. The WDM conjecture roughly states that eigenvalue fluctuations depend only on the local behaviour of the density, and not on far away effects or any other characteristics of the ensemble.

The dependence on the symmetry class is also natural since the *eigenvalue repulsion* is stronger in the complex Hermitian symmetry class than in the real symmetric one. This effect is already visible for 2×2 matrices H since there the eigenvalue difference is given by

$$|\lambda_1 - \lambda_2| = \sqrt{|h_{12}|^2 + (h_{11} - h_{22})^2}$$

and therefore in the complex Hermitian symmetry class $\mathbf{P}(|\lambda_1 - \lambda_2| \le \epsilon) \sim \epsilon^3$, while in the real symmetric symmetry class $\mathbf{P}(|\lambda_1 - \lambda_2| \le \epsilon) \sim \epsilon^2$ for continuously distributed h_{ij}.

In order to formulate the universality of local eigenvalue statistics we define the *k-point function* $p_k^{(N)}$ of H implicitly by the relation

$$\int_{\mathbb{R}^k} f(\boldsymbol{x}) p_k^{(N)}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \binom{N}{k}^{-1} \sum_{\{i_1,\dots,i_k\}\subset[N]} \mathbf{E}\, f(\lambda_{i_1},\dots,\lambda_{i_k})$$

for any smooth compactly supported test function f, where the summation is over all k-element subsets of $[N]$.

We now formulate the universal statistics that are expected to hold for very general ensembles. Later we give precise conditions where we can prove the emergence of those statistics.

Conjecture (WDM conjecture for the Hermitian symmetry class)**.** *Assume that* $\mathfrak{b}$, $\mathfrak{e}$ *and* $\mathfrak{c}$ *are bulk, edge and cusp points of some density* ρ *with parameters* $\gamma_{\mathfrak{e}}, \gamma_{\mathfrak{c}}$ *defined in such a way that*

$$\rho(\mathfrak{e} \pm x) = \gamma_{\mathfrak{e}}^{3/2} x^{1/2}/\pi + \mathcal{O}(x^{1/2}), \quad \rho(\mathfrak{c} + x) = \sqrt{3}\gamma_{\mathfrak{c}}^{4/3} |x|^{1/3}/2\pi + \mathcal{O}(|x|^{1/3}).$$

Then the universal correlation functions are given by

$$\frac{1}{\rho(\mathfrak{b})^k} p_k^{(N)}\Big(\mathfrak{b} + \frac{\boldsymbol{x}}{\rho(\mathfrak{b})N}\Big) \approx \det\Big(\frac{\sin\pi(x_i - x_j)}{\pi(x_i - x_j)}\Big)_{i,j\in[k]}, \tag{Bulk}$$

$$\frac{N^{k/3}}{\gamma_{\mathfrak{e}}^k} p_k^{(N)}\Big(\mathfrak{e} + \frac{\boldsymbol{x}}{\gamma_{\mathfrak{e}} N^{2/3}}\Big) \approx \det\Big(K_{\mathrm{Airy}}(x_i, x_j)\Big)_{i,j\in[k]}, \tag{Edge}$$

$$\frac{N^{k/4}}{\gamma_{\mathfrak{c}}^k} p_k^{(N)}\Big(\mathfrak{c} + \frac{\boldsymbol{x}}{\gamma_{\mathfrak{c}} N^{3/4}}\Big) \approx \det\Big(K_{\mathrm{Pearcey}}(x_i, x_j)\Big)_{i,j\in[k]}, \tag{Cusp}$$

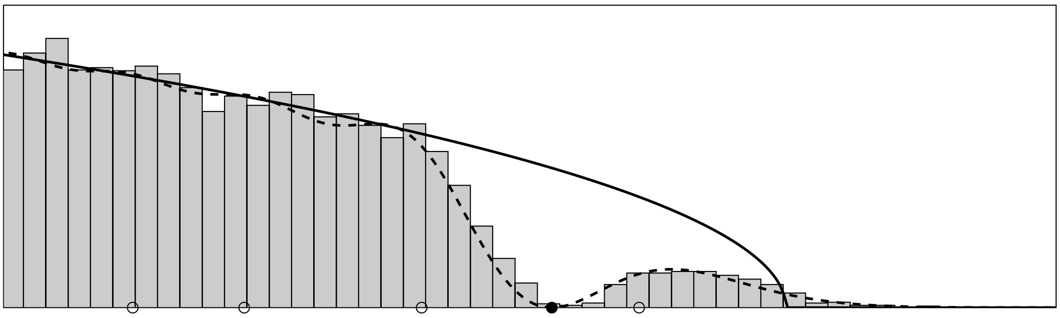


Figure 1.3: Conditional eigenvalue distribution with Airy-2-point function. The continuous line represents the self-consistent density of a Wigner-type random matrix. The circles (both filled and non-filled) represent the eigenvalues of a specific realisation of the matrix ensemble. The histogram shows the empirical eigenvalue distribution of 1000 other realizations which all also happen to have an eigenvalue very close to the filled circle. The dashed line shows the Airy-2-point function $K_{\mathrm{Airy}}(x,x)K_{\mathrm{Airy}}(y,y) - K_{\mathrm{Airy}}(x,y)K_{\mathrm{Airy}}(y,x)$ where the location of the filled circle is set as one of the two arguments.

where the approximation is meant up to an error of $N^{-c(k)}$ when integrated against smooth compactly supported test functions in $\boldsymbol{x} = (x_1, \dots, x_k)$.

Point processes where the k-point function is determinantal are commonly referred to as *determinantal processes*. The kernel in the bulk case is known as the *sine kernel*. The *Airy kernel* for the edge case is given by

$$K_{\mathrm{Airy}}(x,y) := \frac{\mathrm{Ai}(x)\,\mathrm{Ai}'(y) - \mathrm{Ai}'(x)\,\mathrm{Ai}(y)}{x-y}, \quad \mathrm{Ai}(x) := \frac{1}{\pi}\int_0^\infty \cos\Big(\frac{t^3}{3} + tx\Big)\,\mathrm{d}t$$

in terms of the *Airy function* $\mathrm{Ai}(x)$, see Figure 1.3 for a plot of the corresponding 2-point function. The Pearcey kernel K_{Pearcey} has a representation as a two-dimensional contour integral which can be found in (4.5) in Chapter 4. We note that analogous statements also hold for matrices in the real symmetric symmetry class, but the corresponding correlation functions have, while still being determinantal, more complicated kernels. In the cusp case it is not even known whether the universal real symmetric k-point function is determinantal.

The explicit kernels in the WDM conjecture were all computed first for some specific Gaussian model. In the bulk [136] and edge case [85] this reference model was the *Gaussian unitary ensemble (GUE)* and the computation essentially reduces to an asymptotic analysis of Hermite polynomials. In the cusp case [50] the reference model was a deformed GUE matrix with expectation $\mathrm{diag}(1, \dots, 1, -1, \dots, -1)$ and the computation was based on the saddle point analysis of an explicit contour integral formula obtained via the *Harish-Chandra-Itzykson-Zuber* integral over the unitary group.

Three step strategy

Proving the WDM conjecture beyond the Gaussian ensembles turned out to be a difficult task. Even for the simplest model of Wigner matrices this was only achieved in [161] in 1999 at regular edges, and later in a series of papers [74, 75, 167, 81] in 2010 also in the spectral bulk. In the vicinity of cusps universality was only achieved very recently in [DS5, DS6] which are enclosed as Chapters 4–5 of this thesis. While the first universality proof at the regular edges in [161] essentially was an ingenious but laborious extensions of the classical moment method, it turned out that the bulk statistics are inaccessible via moments. Instead, the *three-step strategy* was developed, see [78] for a pedagogical introduction. The first step consists of proving the local law as in (1.14) on optimal mesoscopic scales. We now briefly outline the remaining two steps.

Addition of a small Gaussian component via Green function comparison

The goal of the second step is the addition of a small Gaussian component to the matrix $H = A + W$ while preserving the leading order term of the k-point function. We consider the *Ornstein-Uhlenbeck (OU)* process

$$\mathrm{d}H_t = -\frac{1}{2}(H_t - A)\,\mathrm{d}t + \Sigma^{1/2}[\mathrm{d}B_t], \quad H_0 = H, \quad \Sigma[R] := \mathbf{E}\, W \operatorname{Tr} WR \tag{1.17}$$

with B_t being a standard Hermitian matrix valued Brownian motion. The SDE (1.17) has the solution

$$H_t = A + e^{-t/2}W + \int_0^t e^{(s-t)/2}\Sigma^{1/2}[\mathrm{d}B_t] \tag{1.18}$$

from which it follows that H_t preserves expectation and covariances. Since the self-consistent density ρ only depends on the first and second moments of H, it also is invariant under the OU flow. Therefore it is reasonable to expect that the leading order of the k-point function should also remain unchanged, and indeed, a simple continuity argument on the time-evolved resolvents $G_t = (H_t - z)^{-1}$ evaluated with $\Im z = N^{-\epsilon}\eta_{\mathrm{f}}(\Re z)$ shows exactly that, as long as time t is not too long. The threshold times for this simple continuity argument turn out to be

$$t \ll \begin{cases} N^{-1/2} & \text{bulk,} \\ N^{-1/6} & \text{edge,} \\ N^{-1/4} & \text{cusp.} \end{cases} \tag{1.19}$$

Using technically more involved arguments this result can be extended to larger times, but for the universality proof following the three step strategy with an optimal local law, these time thresholds are sufficient.

Proving universality for the evolved matrices becomes feasible due to the effective Gaussian component in (1.18). It follows from the assumed lower bound $\mathbf{E}\, WRW \ge c\,\langle R\rangle$ that, in distribution, the Gaussian component in (1.18) can be decomposed as

$$\int_0^t e^{(s-t)/2}\Sigma^{1/2}[\mathrm{d}B_t] \overset{\mathrm{d}}{=} U_t' + \sqrt{ct}U$$

where U is a GUE/GOE matrix, and U_t' is a Gaussian matrix independent from U. Thus H_t can be written as

$$H_t \overset{\mathrm{d}}{=} \widetilde{H}_t + \sqrt{ct}U, \tag{1.20}$$

where $\widetilde{H}_t := A + e^{-t/2}W + U_t'$ and U is independent of $\widetilde{H}_t$. It is easy to check that the local law (1.14) also applies to the perturbed matrix $\widetilde{H}_t$ from which we conclude that $\widetilde{H}_t$ satisfies optimal eigenvalue rigidity. In summary, the Green function comparison step achieves reducing universality of general random matrices to a universality of random matrices with an explicit GUE/GOE component of a certain maximal size.

Universality for matrices with a small GUE/GOE component

The final step towards universality complements the previous step in the sense that it proves universality for random matrices with an explicit GUE/GOE component of a certain minimal size, like the rhs. of (1.20). This argument is conditional on the randomness in $\widetilde{H}_t$ and only relies on the randomness from U. In the Hermitian symmetry class there are two techniques available for this step; a saddle point analysis of the explicit formula for the correlation kernel from the Harish-Chandra-Itzykson-Zuber integral, and the fast relaxation to equilibrium of *Dyson Brownian motion (DBM)*. The saddle point analysis for matrices with sizeable Gaussian component was introduced in [109] and later extended to small components in [74], while the DBM method was introduced in [75]. In the real symmetric symmetry class, however, only the latter is feasible[2]. We now give a short sketch of the DBM method; for an exemplary saddle point analysis in the cusp regime the reader is referred to Chapter 4.

Consider the SDE

$$\mathrm{d}X_s = \frac{\mathrm{d}B_s}{\sqrt{N}}, \quad X_0 = X \quad \text{with solution} \quad X_s = X + \frac{B_s}{\sqrt{N}}, \tag{1.21}$$

where B_s is again a Hermitian matrix valued Brownian motion and X is any Hermitian matrix. Here we denote the *time variable* by s to avoid confusions with the time variable t in the previously considered OU flow. In distribution we thus have $X_s \overset{\mathrm{d}}{=} X + \sqrt{s}U$ for fixed s with U being a GUE/GOE matrix. For the universality proof we will choose $X_0 = \widetilde{H}_t$ such that $X_{ct} \overset{\mathrm{d}}{=} H_t$ according to (1.20). In parallel we consider a second flow X_s' also evolved according to (1.21) with the same Brownian motion, but a different initial condition $X_0' = \widetilde{U}_t$ with $\widetilde{U}_t$ being some appropriate Gaussian comparison model. In practice $\widetilde{U}_t$ is chosen in such a way that the self-consistent density of $\widetilde{U}_t + \sqrt{ct}U$ matches the one of H_t and thereby ρ around the respective expansion points. Our goal is now to prove that after sufficiently long times s the eigenvalues $\lambda_i(s)$ and $\lambda_i'(s)$ of X_s and X_s' are with high probability very close to each other. If this *relaxation to equilibrium* happens already at the time $s = ct$, then this proves universality for the matrix H_t and therefore by the Green function comparison argument also for H. In this sense *universality* means that the local spectral distribution of any given matrix model agrees with the one of some fixed Gaussian comparison model. To obtain the exact formulae of the correlation kernels an explicit calculation about the correlation kernels of Gaussian matrices is needed.

Freeman Dyson made the important and somewhat surprising observation [67] that the flow (1.21) induces a flow purely on eigenvalues $\lambda_i(s)$ which does not involve eigenvectors. The SDE of the eigenvalue flow can readily be found from standard eigenvalue perturbation formulae. If $A = A(s)$ is a smooth matrix-valued function of s with simple eigenvalues λ_i and eigenvectors $\boldsymbol{v}_i$, then by differentiating the equations $A\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$ and $\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle = 1$ twice, one quickly obtains

$$\dot{\lambda}_i = \langle \boldsymbol{v}_i, \dot{A}\boldsymbol{v}_i \rangle, \quad \ddot{\lambda}_i = \langle \boldsymbol{v}_i, \ddot{A}\boldsymbol{v}_i \rangle + 2\sum_{j\neq i} \frac{|\langle \boldsymbol{v}_j, \dot{A}\boldsymbol{v}_i \rangle|^2}{\lambda_i - \lambda_j}$$

[2] The Harish-Chandra-Itzykson-Zuber integral over the orthogonal group allows to compute the correlation kernel of real anti-symmetric but not real symmetric matrices.

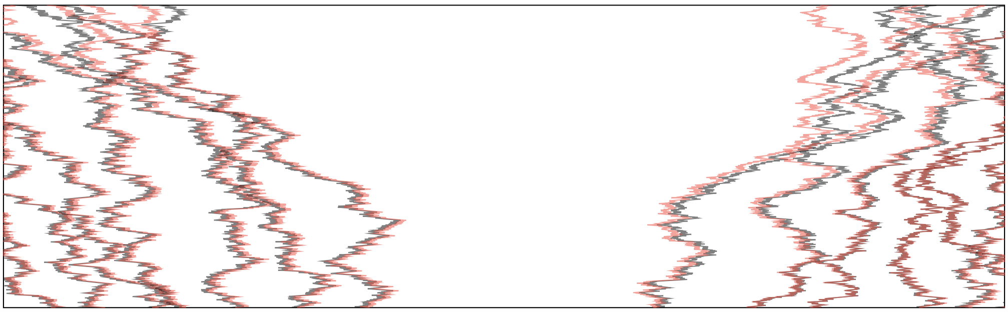


Figure 1.4: Relaxation to equilibrium of Dyson Brownian motion in the cusp regime. The red an black paths show simulated coupled Brownian motions (with time flowing from top to bottom) in the cusp regime with different initial conditions. Along the flow the paths become increasingly closer, and perform an overall movement towards the centre corresponding to the gap closure in the semicircular flow.

from which we conclude that

$$\lambda_i(s+\epsilon) \overset{\mathrm{d}}{=} \lambda_i(X_s+\sqrt{\epsilon}U) \approx \lambda_i(s) + \sqrt{\epsilon}\,\langle \boldsymbol{v}_i, U\boldsymbol{v}_i\rangle + \epsilon \sum_{j\neq i} \frac{|\langle \boldsymbol{v}_j, U\boldsymbol{v}_i\rangle|^2}{\lambda_i-\lambda_j}.$$

Since the GUE/GOE matrix U is invariant under unitary/orthogonal transformations, and is independent from $(\boldsymbol{v}_i)_{i\in[N]}$, it follows from orthonormality of the latter that $\langle \boldsymbol{v}_i, U\boldsymbol{v}_j\rangle \overset{\mathrm{d}}{=} u_{ij}$. Since $|u_{ij}|^2$ concentrates around $1/N$ we conclude that

$$\mathrm{d}\lambda_i = \frac{1}{N}\sum_{j\neq i}\frac{1}{\lambda_i-\lambda_j}\,\mathrm{d}s + \sqrt{\frac{2}{\beta N}}\,\mathrm{d}b_i, \quad \lambda_i(0)=\lambda_i(X)$$

where $(b_i)_{i\in[N]}$ is a standard vector valued Brownian motion, and $\beta = 1,2$ in the real and complex symmetry classes.

By considering the processes λ_i, λ_i' with different initial conditions $\lambda_i(X_0), \lambda_i(X_0')$ but coupling them using the same Brownian motions, it follows that their difference $\delta_i := \lambda_i - \lambda_i'$ satisfies

$$\frac{\mathrm{d}\delta_i}{\mathrm{d}s} = \sum_{j\neq i} B_{ji}(\delta_j-\delta_i), \quad B_{ji} := \frac{1}{N}\frac{1}{(\lambda_i-\lambda_j)(\lambda_i'-\lambda_j')}, \tag{1.22}$$

which can be viewed as a discretised integral equation of *parabolic type* with time-dependent random coefficients B_{ji}. In order to analyse (1.22) it is convenient to localise the processes using a *short range approximation* which leads to a parabolic equation whose short time heat kernel has rapid off-diagonal decay. In this short range approximation we replace particles beyond a certain distance by a forcing term which only involves their deterministic density, i.e. we consider

$$\mathrm{d}\widehat{\lambda}_i = \frac{1}{N}\sum_{\substack{j\neq i\\ |j-i|\leq L}}\frac{1}{\widehat{\lambda}_i-\widehat{\lambda}_j}\,\mathrm{d}s + \sum_i\left(\int_{-\infty}^{\gamma_{i-L}} + \int_{\gamma_{i+L}}^{\infty}\right)\frac{\rho(y)}{\widehat{\lambda}_i-y}\,\mathrm{d}y\,\mathrm{d}s + \sqrt{\frac{2}{\beta N}}\,\mathrm{d}b_i,$$

with initial condition $\widehat{\lambda}_i(0) = \lambda_i(X)$, where $\gamma_i = \gamma_i(s)$ and $\rho = \rho_s$ denote the time-evolved quantiles and density of X_s, and L is a suitable cut-off length scale. In parallel we consider

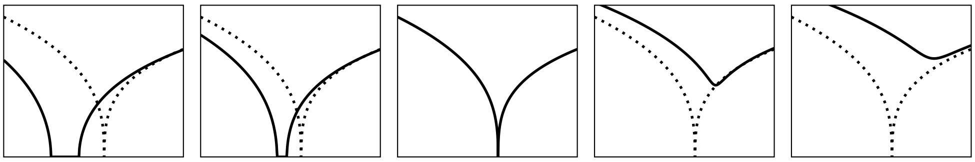


Figure 1.5: Semicircular evolution ρ_s in the cusp regime where s is increasing from left to right. A density (continuous line) with two support intervals gets continuously transformed into a density with a non-zero local minimum via a cusp singularity (dotted line). This transformation happens on three different time scales; the density performs an overall movement which is linear in time s, while the gap size shrinks as $|s-s_*|^{3/2}$ and the local minimum grows as $|s-s_*|^{1/2}$, where s_* is the time of the cusp formation.

the short-range approximation $\widehat{\lambda}_i'$ defined analogously but with initial condition $\widehat{\lambda}_i'(0) = \lambda_i(X')$. These short-range approximations are useful since the comparison ensemble cannot be chosen in such a way which matches the density of the original matrix globally but only locally. The closeness of λ_i and $\widehat{\lambda}_i$ and the one of λ_i' and $\widehat{\lambda}_i'$ are achieved via *finite speed of propagation estimates*.

By *heat kernel decay estimates* for the analogue of (1.22) for the difference $\widehat{\delta}_i = \widehat{\lambda}_i - \widehat{\lambda}_i'$ one can show that after a sufficiently long time $|\widehat{\delta}_i(s)|$ is, with high probability, smaller than the fluctuation scale of individual eigenvalues, thus $\widehat{\lambda}_i(s)$ and $\widehat{\lambda}_i'(s)$ and thereby also $\lambda_i(s)$ and $\lambda_i'(s)$ are very close to each other, irrespective of the different initial conditions. This effect is referred to as the *fast relaxation to equilibrium* of the DBM. The time-scale of this relaxation is $s \gg N^{-1}$, $s \gg N^{-1/3}$ and $s \gg N^{-1/2}$ in the bulk, edge and cusp case, respectively. These relaxation times leave sufficient room to choose t and $s = ct$ in such a way that the times are short enough to retain the validity of the Green function continuity argument, cf. (1.19), and long enough to ensure relaxation.

The analysis of the DBM requires a precise understanding of the evolution of the self consistent density along the flow (1.21). From (1.12) it follows that the resolvent $G_s = (X_s - z)^{-1}$ is approximated by the solution $M_s \approx G_s$ to the MDE

$$-M_s^{-1} = z - X_0 + \frac{1}{N}\mathbf{E}\, B_s M_s B_s = z - X_0 + s\,\langle M_s\rangle$$

which is solved by $M_s = m_s I$, where m_s is the solution of the scalar equation

$$m_s(z) = \langle (X_0 - z - s m_s(z))^{-1}\rangle = \langle G_0(z + s m_s(z))\rangle\,. \tag{1.23}$$

The flow m_s is known as the *semicircular flow* [31] and (1.23) is the defining equation for the *free additive convolution* of the ESD of X_0 and a semicircular distribution of variance s. The analysis of m_s is important mainly in the edge and even more so in the cusp regimes since there the self-consistent density ρ_s corresponding to m_s changes qualitatively on the time scale of the DBM analysis, see Figure 1.5 for a representative example. This change has to be tracked very accurately and is used to match the contribution of the long range parts of the dynamics.

1.5 Random matrix models beyond the scope of this thesis

Next to the aforementioned Hermitian mean field random matrix models, there are extensive ongoing research efforts on other random matrix ensembles, some of which we want to mention here for completeness.

1.5.1 Non mean-field Hermitian matrices

For a given bandwidth $1 \ll W \ll N$ we define the 1-dimensional band matrix $H = H^*$ such that

$$\mathbf{E}\,|h_{ij}|^2 = \begin{cases} W^{-1} & \text{if } |i-j| \le W/2, \\ 0 & \text{else.} \end{cases}$$

Due to the normalisation it follows that the Dyson equation for $\langle (H-z)^{-1} \rangle$ is the same as for Wigner matrices and therefore the self-consistent density of ρ is semicircular. On the mesoscopic and microscopic level, however, numerical evidence [53, 54] suggests a phase transition at $W \sim \sqrt{N}$. For $W \gg \sqrt{N}$ it is expected that the eigenvalue statistics lie in the Wigner-Dyson-Mehta universality class with completely delocalised eigenvectors, while for $W \ll \sqrt{N}$ it is expected that the eigenvalues form a Poisson point process, and the eigenvectors are localised. At the moment of writing the delocalised phase has rigorously been established for $W \gg N^{3/4}$ in [46], while the localised phase has been established for $W \ll N^{1/8}$ in [152]. Closing this gap further or even rigorously studying this phase-transition is one of the *major open problems* in random matrix theory. Understanding the phase transition of this band matrix model is also of physical interest since it is expected [164] to share spectral properties with the *random Schrödinger* operator $H = \Delta + \lambda V$ on $[-N, N]$ with periodic boundary conditions for $\lambda \sim W^{-1}$ if the potential $(V(i),\ i \in \mathbb{Z})$ forms an i.i.d. family of centred unit variance random variables.

1.5.2 Sparse random matrices

Let $G = G(N, p)$ be the Erdős-Rényi random graph on $[N]$ vertices for which any given edge is present independently with a probability of p. The adjacency matrix A of G then is a sparse random matrix undergoing a phase transition. It is known that an optimal bulk local law and bulk eigenvector delocalisation [99] hold true whenever $pN \ge C \log N$, while bulk universality [103] is known for $pN \ge N^{\epsilon}$. On the contrary, for $pN \le (1-\epsilon)\log N$ the graph G has, with high probability, isolated vertices and thereby A also exhibits localised eigenvectors. The edge regime of Erdős-Rényi graphs has a phase transition already for much larger values of p. Indeed, for $pN \gg N^{1/3}$ Tracy-Widom universality has been proven [124], while for $N^{2/9} \ll pN \ll N^{1/3}$ the top eigenvalue is approximately Gaussian [104]. Random d-regular graphs are another commonly studied model for sparse random matrices [27].

1.5.3 Invariant ensembles

Invariant ensembles are another natural way of endowing the set of Hermitian matrices with a probability measure, other than the entry-wise approach suggested by Wigner. For sufficiently fast growing potentials $V : \mathbb{R} \to \mathbb{R}$ we define a probability measure on Hermitian matrices such that

$$\mathbf{P}(H) = C \exp\Big(-\frac{\beta}{2} N \operatorname{Tr} V(H)\Big), \tag{1.24}$$

where C is a normalisation constant, $\beta = 1, 2$ in the real symmetric and complex Hermitian case, and $V(H)$ is defined through a functional calculus. By integrating out the orthogonal or unitary matrices for diagonalisation of H one readily finds that the measure $\mathbf{P}$ induces a measure involving only the eigenvalues λ_i of H of the form

$$\mathbf{P}\big((\lambda_i)_{i\in[N]}\big) = C \exp\Big(-\frac{\beta}{2} N \sum_i V(\lambda_i)\Big) \prod_{i<j} |\lambda_i - \lambda_j|^\beta \,. \tag{1.25}$$

This probability measure can be interpreted as the *Gibbs measure* of N particles with logarithmic interaction in the confining potential V, i.e.

$$\mathbf{P}\big((\lambda_i)_{i\in[N]}\big) = C \exp(-\beta N \mathcal{H}), \qquad \mathcal{H} := \frac{1}{2} \sum_i V(\lambda_i) - \frac{1}{N} \sum_{i<j} \log |\lambda_i - \lambda_j| \,. \tag{1.26}$$

It is interesting to note that under the measure (1.24), the entries of H can only be stochastically independent if V is a quadratic polynomial [60], in which case it is called a *Gaussian β-ensemble*. The potential $V(x) = x^2/2$ induces the previously introduced GOE/GUE ensembles. Universality for a wide range of invariant ensembles has been settled via orthogonal polynomial methods generalising the GOE/GUE computations; for a review the reader is referred to [63]. While the probability measures (1.24) give rise to the particle measure (1.26) only for $\beta = 1, 2$ (and $\beta = 4$ for symplectic matrices, albeit with a different normalisation factor), the universality phenomenon has also been established for general $\beta \geq 1$ for (1.26) with DBM methods [42], and, alternatively, by using optimal transportation ideas [28] for $\beta > 0$.

1.5.4 Non-Hermitian random matrices

Non-Hermitian random matrices are in general harder to analyse than their Hermitian counterparts since the complex eigenvalues are unstable even with respect to tiny perturbations. The circular law states that the empirical spectral density of an i.i.d. random matrix $X \in \mathbb{C}^{N\times N}$ with zero mean and unit variance $\mathbf{E}\,|x_{ab}|^2 = N^{-1}$ converges almost surely weakly to the uniform measure on the disk $\{\, z \in \mathbb{C} \mid |z| \leq 1 \,\}$. Under optimal moment conditions this global law was only obtained in [166] after a series of intermediate results with additional moment assumptions, much later than Wigner's semicircular law, the Hermitian analogue. An optimal mesoscopic local law above the average eigenvalue spacing of $N^{-1/2}$ was obtained in [45, 179]. On the microscopic scale universality of single eigenvalue statistics is still a *major open problem* and has only been settled in the perturbative regime of four matching moments [169].

On a technical level the main reason for the difficulty posed by non-Hermitian random matrix models is that the resolvent $(X - z)^{-1}$ is unstable. Beyond the Gaussian case, the only known technique for extracting mesoscopic and microscopic statistics of X is a Hermitization trick due to Girko [90]. For generic functions f the linear statistic of the eigenvalues σ_i of X can be written as

$$\begin{aligned}\sum_i f(\sigma_i) &= \frac{1}{4\pi}\int_{\mathbb{C}} \Delta f(z) \log|\det H^z| \,\mathrm{d}z \\ &= -\frac{1}{4\pi}\int_{\mathbb{C}} \Delta f(z)\Big(\log|\det G^z(\mathrm{i}T)| - 2\int_0^T \Im \operatorname{Tr} G^z(\mathrm{i}\eta)\,\mathrm{d}\eta\Big)\,\mathrm{d}z,\end{aligned} \tag{1.27}$$

for general $T>0$, where we introduced the Hermitian matrix H^z with resolvent G^z as

$$H^z = \begin{pmatrix} 0 & X-z \\ X^*-\overline{z} & 0\end{pmatrix}, \qquad G^z(w) := (H^z-w)^{-1}.$$

The resolvent G^z can, in principle, be analysed by similar methods as the resolvent of the previously considered mean-field random matrices. However, there are three additional difficulties: Firstly, (1.27) requires an additional estimate on the smallest singular value of $X-z$ in order to study the regime $0 \le \eta \ll \eta_{\mathrm{f}}(0)$. Second, the stability operator $\mathcal{B}$ corresponding to G^z has a second unstable direction due to the 0 blocks of H^z. And finally, studying the edge regime of X requires studying H^z for $|z|=1$ for which the self-consistent density ρ^z of H^z has a cusp singularity in 0.

1.6 Overview of results

The remainder of this thesis consists of largely unmodified versions of the published or submitted papers [DS1]–[DS6] in thematic rather than chronological order published in course of the author's PhD studies. Among those [DS3]–[DS6] are very closely related and represent a long term project on universality. The chronologically first two publications [DS1, DS2] report on a separate set of results. To guide the reader we now briefly summarise the findings and put the key results into context.

1.6.1 Correlated Wigner matrices

In [DS3] we proved an optimal local law and universality in the bulk regime for a general class of correlated Wigner matrices with polynomially decaying correlations and arbitrary expectations. Previously, such results were only available for general matrices with exponentially decaying correlations and expectations [8] or matrices with special translational invariant correlation structures [55]. The key technical achievement of [DS3] was the development of a flexible diagrammatic cumulant expansion replacing the traditional Schur complement analysis which is inherently limited when it comes to correlated matrices since row expansions are considerably less efficient than entrywise expansions. The cumulant expansion allows to gain multiplicatively from several cancellations in the computation of high moments of the error matrix D, enabling the analysis of polynomially decaying correlation structures.

In [DS4] we extended the results of [DS3] to the edge regime. The necessary probabilistic estimates could be imported from [DS3], while for the shape and stability analysis we could refer to [12]. The main technical novelty of [DS4] was a particularly strong notion of *band rigidity* which means that eigenvalues cannot have fluctuations bridging gaps of size $N^{-3/4+\epsilon}$ in the support of ρ. The absence of such eigenvalues which are delocalised between two neighbouring support bands differentiates Wigner-type matrices from other random matrix models such as β-ensembles. As a consequence of the band rigidity we could prove a strong fixed label version of edge universality, also for internal edges.

1.6.2 Cusp universality

In [DS5] we proved an optimal local law in the cusp regime of Wigner-type matrices. Together with a saddle-point and semicircular flow analysis this enabled us to achieve the first Pearcey-class universality proof for random matrices. The main technical achievement of [DS5] was establishing a second order cancellation in the error matrix analysis essentially due to the cusp symmetry. Within the framework of the diagrammatic cumulant expansion originally developed in [DS3] this cancellation becomes accessible via a resummation of certain subgraphs.

In [DS6] we extended the result of [DS5] to the real symmetric symmetry class where the saddle point analysis of [DS5] had to be replaced by a DBM analysis in the cusp regime. Besides adapting the corresponding edge analysis from [122], the main challenge in [DS5] was to establish near-optimal rigidity for interpolated ensembles for which local laws were not available.

1.6.3 Entrywise linear statistics

In [DS1] we studied the fluctuations of rectangular Young diagrams obtained by interlacing eigenvalues, after it was previously shown [51] that their deterministic limit agrees with that of Young diagrams equipped with the Plancherel measure. We found that this correspondence does not carry over to the level of fluctuations even though both are Gaussian.

Later we realised in [DS2] that the methods developed in [DS1] allow to identify the entrywise fluctuations of $f(H)_{ij}$ for functions of Wigner matrices for much rougher functions f than previously. The main reason for this improvement was the use of optimal local laws and an inversion formula due to Pleijel [148] rather than the more commonly used Helffer-Sjöstrand representation.

Random Matrices with Slow Correlation Decay 2

We consider large random matrices with a general slowly decaying correlation among its entries. We prove universality of the local eigenvalue statistics and optimal local laws for the resolvent away from the spectral edges, generalizing the recent result of [8] to allow slow correlation decay and arbitrary expectation. The main novel tool is a systematic diagrammatic control of a multivariate cumulant expansion.

Published as L. Erdős, T. Krüger, and D. Schröder, *Random matrices with slow correlation decay*, to appear in Forum Math. Sigma (2017), `arXiv:1705.10661`.

2.1 Introduction

In recent years it has been proven for increasingly general random matrix ensembles that their spectral measure converges to a deterministic measure up to the scale of individual eigenvalues as the size of the matrix tends to infinity, and that the fluctuation of the individual eigenvalues follows a universal distribution, independent of the specifics of the random matrix itself. The former is commonly called a *local law*, whereas the latter is known as the *Wigner-Dyson-Mehta (WDM) universality conjecture*, first envisioned by Wigner in the 1950's and formalized later by Dyson and Mehta in the 1960's [135]. In fact, the conjecture extends beyond the customary random matrix ensembles in probability theory and is believed to hold for any random operator in the delocalization regime of the Anderson metal-insulator phase transition. Given this profound universality conjecture for general disordered quantum systems, the ultimate goal of local spectral analysis of large random matrices is to prove the WDM conjecture for the largest possible class of matrix ensembles. In the current paper we complete this program for random matrices with a general, slow correlation decay among its matrix elements. Previous works covered only correlations with such a fast decay that, in a certain sense, they could be treated as a perturbation of the independent model. Here we present a new method that goes well beyond the perturbative regime. It relies on a novel multi-scale version of the cumulant expansion and its rigorous

Feynman diagrammatic representation that can be useful for other problems as well. To put our work in context, we now explain the previous results.

In the last ten years a powerful new approach, the *three-step strategy* has been developed to resolve WDM universality problems, see [78] for a summary. In particular, the WDM conjecture in its classical form, stated for Wigner matrices with a general distribution of the entries, has been proven with this strategy in [74, 75, 81]; an independent proof for the Hermitian symmetry class was given in [167]. Recent advances have crystallized that the only model dependent step in this strategy is the first one, the local law. The other two steps, the fast relaxation to equilibrium of the Dyson Brownian motion and the approximation by Gaussian divisible ensembles, have been formulated as very general "black-box" tools whose only input is the local law [120, 77, 121]. Thus the proof of the WDM universality, at least for mean field ensembles, is automatically reduced to obtaining a local law.

Both local law and universality have first been established for *Wigner matrices*, which are real symmetric or complex Hermitian $N \times N$ matrices with mean-zero entries which are independent and identically distributed (i.i.d.) up to symmetry [75, 76]. For Wigner matrices it has long been known that the *limiting*, or *self-consistent* density is the *Wigner semicircle law*. In subsequent work the condition on the i.i.d. entries has been relaxed in several steps. First, it was proven in [81], that for *generalized Wigner ensembles*, i.e., for matrices with stochastic variance profile and uniform upper and lower bound on the variance of the matrix entries, the local law and universality also hold, with the self-consistent density still given by the semicircle law. Next, the condition of stochasticity was removed by introducing the *Wigner-type* ensemble [9], in which case the self-consistent density is, generally, not semicircular any more. Finally, the independence condition was dropped and in [8] both a local law on the optimal local scale and bulk universality were obtained for matrices with correlated entries with fast decaying general correlations. Special correlation structures were also considered before in [6, 55] on a local scale. We also mention that there exists an extensive literature on the global law for random matrices with correlated entries [47, 91, 94, 153, 151, 16, 23]. These results, however, either concern Gaussian random matrices or more specific correlation structures than considered in the present work. In a parallel development the zero-mean condition on the matrix elements has also been relaxed. First this was achieved for the *deformed Wigner ensembles* that have diagonal deterministic shifts in [138, 126] and more recently for i.i.d. Wigner matrices shifted by an arbitrary deterministic matrix in [100].

In this paper we prove a local law and bulk universality for random matrices with a slowly decaying correlation structure and arbitrary expectation, generalizing both [100, 8]. The main point is to considerably relax the condition on the decay of correlations compared to [8]: We allow for a polynomial decay of order two in a neighbourhood of size $\ll \sqrt{N}$ around every entry and we only have to assume a polynomial decay of a certain finite order outside these neighbourhoods. Another novelty is that our new concept of neighbourhoods is completely general, it is not induced by the product structure of the index set labelling the matrix elements. In particular, the improved correlation condition also includes many other matrix models of interest, for example, general block matrix type models, that have not been covered by [8].

Regarding strategy of proving the local law, the starting point is to find the deterministic approximation of the resolvent $G(z) = (H - z)^{-1}$ of the random matrix H with a complex spectral parameter z in the upper half plane $\mathbb{H} = \{ z \in \mathbb{C} \mid \Im z \geq 0 \}$. This approximation is given as the solution $M = M(z)$ to the *Matrix Dyson Equation (MDE)*

$$1 + (z - A + \mathcal{S}[M])M = 0,$$

where the expectation matrix $A := \mathbf{E}\, H$ and the linear map $\mathcal{S}[V] := \mathbf{E}(H-A)V(H-A)$ on the space of matrices R encode the first two moments of the random matrix. The resolvent approximately satisfies the MDE with an additive perturbation term

$$D := (H - A)G + \mathcal{S}[G]G.$$

The smallness of D and stability of the MDE against small perturbations imply that G is indeed close to M. The necessary stability properties of the MDE have already been established in [8], so the main focus in this paper is to bound D in appropriate norms that can then be fed into the stability analysis. Most proofs of the previous local laws loosely follow a strategy of first reducing the problem to a smaller number of relevant variables, such as the diagonal entries of G. Instead, correlated ensembles require to carry out the analysis genuinely on the matrix level since G is not even approximately diagonal. This key feature distinguishes the current paper as well as [8] from all previous works, where the Dyson equation was only a scalar equation for the trace of the resolvent or a vector equation for its diagonal elements. Although adding a general expectation matrix A to a Wigner matrix already induces a non-diagonal resolvent, diagonalization of A reduced the analysis to the scalar level in [100]. A similar algebraic reduction is not possible for general correlations even if they decay as fast as in [8]. However, in [8] every matrix quantity, such as G or M, still had a very fast off-diagonal decay and thus it was sufficient to focus only on matrix elements very close to the diagonal; the rest was treated as an irrelevant error. For the slow correlation decay considered in this paper such direct perturbative treatment for the off-diagonal elements is not possible. In fact, with our new method we can even handle the essentially optimal integrable correlation decay on a scale $\sqrt{N}$ near the diagonal.

To obtain a probabilistic bound on D, essentially two approaches are available. When G is approximately diagonal and when the columns of H are independent, one may use resolvent expansion formulas involving minors that lead to standard linear and quadratic large deviation bounds – a natural idea that first arose in the works of Girko and Pastur [143, 89], as well as in the works of Bai et. al., e.g. [19]. For correlated models the natural extension of this method requires a somewhat involved successive expansion of minors; this was the main technical tool in [8]. This approach is thus restricted to very fast correlation decay since it is essentially a perturbation around nearly diagonal matrices. The alternative method relies on the cumulant expansion of the form $\mathbf{E}\, hf(h) = \sum_k(\kappa_{k+1}/k!)\,\mathbf{E}\, f^{(k)}$, where κ_k is the k-th order cumulant of the random variable h. The power of this expansion in studying resolvents of random matrices was first recognized in [116] and it has been revived in several recent papers, e.g. [123, DS1, 98]. It gives more flexibility than the minor expansion on two accounts. First, it can handle the stochastic effect of individual matrix elements instead of treating an entire column at the same time. This observation was essential in [100] to handle deformations of Wigner matrices with an arbitrary expectation matrix. Single entry expansions, as opposed to expansion by entire columns, also appeared in the proof of a version of the *fluctuation averaging theorem* [83], but in this context it did not have any major advantage over the row expansions. Secondly, a multivariate version of the cumulant expansion is inherently well suited to correlated models; it automatically keeps track of the correlation structure without artificial cut-offs and strong restrictions on the off-diagonal decay. This is the method we use to bound D in the current work to handle the slow correlation decay effectively.

After presenting our main results in Section 2.2, in Section 2.3 we first give a multivariate cumulant expansion formula with an explicit error term that is especially well suited for mean field random matrix models. The main ingredient is a novel *pre-cumulant decoupling identity*, Lemma 2.3.1. We were not able to find these formulas in the literature; related formulas, however, have probably been known. They are reminiscent to the Wick polynomials, their relationship is explained in Appendix 2.B. Some consequences are collected in Section 2.3.3 via a toy model. When applying it to our problem, in order to bookkeep the numerous terms, we develop a graphical language which allows us to actually compute $\mathbf{E}\,|\Lambda(D)|^p$ up to a tiny error for arbitrary linear functionals Λ. The structure of D contains an essential cancellation: the term $(H-A)G$ is compensated by $\mathcal{S}[G]G$ that acts as a counter term or *self-energy renormalization* in the physics terminology. Our cumulant expansion automatically exploits this cancellation to all orders and the diagrammatic representation in Sections 2.4.1–2.4.4 conveniently visualizes this mechanism. Section 2.4 contains the main novel part of this paper, in Section 2.5 we combine the bounds on D with the stability argument for the MDE to prove the local law. Section 2.6 is devoted to the short proofs of bulk universality and other natural corollaries of the local law.

Acknowledgements. T.K. gratefully acknowledges private communications with Antti Knowles on the preliminary version of [100]. D.S. would like to thank Nikolaos Zygouras for raising the question how our novel pre-cumulants are related to Wick polynomials.

2.2 Main results

For a Hermitian $N\times N$ random matrix $H=H^{(N)}$ we denote its resolvent by

$$G(z)=G^{(N)}(z)=(H-z)^{-1},$$

where the spectral parameter z is assumed to be in the upper half plane $\mathbb{H}$. The first two moments of H determine the limiting behaviour of $G(z)$ in the large N limit. More specifically, let

$$A:=\mathbf{E}\,H,\qquad H=:A+\frac{1}{\sqrt{N}}W,\qquad \mathcal{S}[V]:=\frac{1}{N}\mathbf{E}\,WVW,$$

where $\mathcal{S}$ is a linear map on the space of $N\times N$ matrices and W is a random matrix with zero expectation. Then the unique, deterministic solution $M=M(z)$ to the matrix Dyson equation (MDE)

$$1+(z-A+\mathcal{S}[M])M=0\quad\text{under the constraint}\quad \Im M:=\frac{1}{2\mathrm{i}}[M-M^*]>0,\qquad(2.1)$$

approximates the random matrix $G(z)$ increasingly well as N tends to ∞. Here $\Im M>0$ indicates that the matrix $\Im M$ is positive definite. The properties of (2.1) and its solution have been comprehensively studied in [8]. In particular, it has been shown that

$$\frac{1}{N}\operatorname{Tr}M(z)=\int_{\mathbb{R}}\frac{1}{x-z}\,\mathrm{d}\mu(x)$$

is the Stieltjes transform of a measure μ on $\mathbb{R}$, which we call the *self-consistent density of states*, and whose support $\operatorname{supp}\mu$ we call the *self-consistent spectrum*. Under an additional flatness Assumption (see Assumption (2.E) later) it has also been shown that μ is absolutely continuous with compactly supported Hölder continuous probability density

$$\mathrm{d}\mu(x) = \rho(x)\,\mathrm{d}x \qquad \text{and that} \qquad \rho(z) := \frac{1}{\pi N}\Im \operatorname{Tr} M(z)$$

is the harmonic extension of $\rho\colon \mathbb{R} \to [0,\infty)$. Moreover, (2.1) is stable with respect to small additive perturbations and therefore it is sufficient to show that the error matrix $D = D(z)$ defined by

$$D := 1 + (z - A + \mathcal{S}[G])G = (H - A + \mathcal{S}[G])G = \frac{W}{\sqrt{N}}G + \mathcal{S}[G]G \tag{2.2}$$

is small.

Choosing the correct norm to measure smallness of the error terms is a key technical ingredient. Similarly to the resolvent G, the error matrix D is very large in the usual induced $\ell^p \to \ell^q$ matrix norms, but its quadratic form $\langle \mathbf{x}, D\mathbf{y}\rangle$ is under control with very high probability for any fixed deterministic vectors $\mathbf{x}, \mathbf{y}$. Furthermore, to improve precision, we will distinguish two different concepts of measuring the size of D. We will show that D can be bounded in *isotropic sense* as $|\langle \mathbf{x}, D\mathbf{y}\rangle| \lesssim \|\mathbf{x}\|\,\|\mathbf{y}\|/\sqrt{N\Im z}$ for fixed deterministic vectors $\mathbf{x}, \mathbf{y}$ as well as in an *averaged sense* as $N^{-1}|\operatorname{Tr} BD| \lesssim \|B\|/N\Im z$ for fixed deterministic matrices B. Here $\|\mathbf{x}\|, \|\mathbf{y}\|, \|B\|$ denote the standard (Euclidean) vector norm $\|\mathbf{x}\|^2 = \sum_a |x_a|^2$ and (matrix) operator norm $\|B\| := \sup_{\|\mathbf{x}\|,\|\mathbf{y}\|\le 1}|\langle \mathbf{x}, B\mathbf{y}\rangle|$. The second step of the proof will be to show that because D is small, and (2.1) is stable under small additive perturbations, also $G - M$ is small in an appropriate sense.

2.2.1 Notations and conventions

An inequality with a subscript indicates that we allow for a constant in the bound depending only on the quantities in the subscript. For example, $A(N,\epsilon) \le_\epsilon B(N,\epsilon)$ means that there exists a constant $C = C(\epsilon)$, independent of N, such that $A(N,\epsilon) \le C(\epsilon)B(N,\epsilon)$ holds for all N and $\epsilon > 0$. In many statements we will implicitly assume that N is sufficiently large, depending on any other parameters of the model. Moreover, we will write $f \sim g$ if $f = \mathcal{O}(g)$ and $g = \mathcal{O}(f)$, if it is clear from the context in which regime we claim this comparability and how the implicit constant may depend on parameters.

An abstract index set J of size N labels the rows and columns of our matrix (generally one can think of $J = [N] := \{1, \dots, N\}$ but there is no need for having a (partial) order or a notion of distance on J). The elements of J will be denoted by letters $a, b, \dots$ and $i, j, \dots$ from the beginning of the alphabet. We will use boldfaced letters $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v}, \dots$ from the end of the alphabet to denote J-vectors with entries $\mathbf{x} = (x_a)_{a\in J}$. We will denote the set of ordered pairs of indices by $I := J \times J$ and will often call the elements of I *labels* to avoid confusion with other types of indices, and will denote them by Greek letters $\alpha = (a,b) \in I$. The matrix element w_{ab} will thus often be denoted by w_α. Summations of the form $\sum_a$ and $\sum_\alpha$ are always understood to sum over all $a \in J$ and $\alpha \in I$.

For indices $a, b \in J$ and vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^J$ we shall use the notations

$$A_{\mathbf{x}\mathbf{y}} := \langle \mathbf{x}, A\mathbf{y}\rangle\,, \qquad A_{\mathbf{x}a} := \langle \mathbf{x}, Ae_a\rangle\,, \qquad A_{a\mathbf{x}} := \langle e_a, A\mathbf{x}\rangle\,,$$

where e_a is the a-th standard basis vector. We will frequently write $\Delta^{ab} = e_a e_b^t$ for the matrix of all zeros except a one in the (a,b) entry. The normalized trace of an $N \times N$ matrix is denoted by $\langle A\rangle := N^{-1}\operatorname{Tr} A$. Sometimes we will also use the notation $\langle z\rangle := 1+|z|$ for the complex number z, but this should not create confusions as it will only be used for z. We will furthermore use the maximum norm and the normalized Hilbert-Schmidt norm

$$\|A\|_{\max} := \max_{a,b} |A_{ab}|\,, \quad \|A\|_{\text{hs}} := \Big[\frac{1}{N}\sum_{a,b} |A_{ab}|^2\Big]^{1/2}$$

for an $N\times N$ matrix A.

2.2.2 Assumptions

We now formulate our main assumptions on W and A.

Assumption (2.A) (Bounded expectation). *There exists some constant C such that $\|A\|\le C$ for all N.*

Assumption (2.B) (Finite moments). *For all $q\in\mathbb{N}$ there exists a constant μ_q such that $\mathbf{E}\,|w_\alpha|^q \le \mu_q$ for all α.*

Next, we formulate our conditions on the correlation decay conveniently phrased in terms of the multivariate cumulants κ of random variables of $\{\, w_\alpha \mid \alpha\in I\,\}$. In Appendix 2.A we recall the definition and some basic properties of multivariate cumulants. First we present a simple condition in terms of a tree type ρ-mixing decay of the cumulants with respect to the standard Euclidean metric on $[N]^2$. Later, in Section 2.2.5, we formulate weaker and more general conditions which we actually use for the proof of our results but their formulation is quite involved, so for the sake of clarity we first rather state simpler but more restrictive assumptions.

Consider $J=[N]$, $I=[N]^2$ equipped with the standard Euclidean distance modulo the Hermitian symmetry, i.e., for $\alpha,\beta\in I$ we set $d(\alpha,\beta) := \min\{|\alpha-\beta|\,,|\alpha^t-\beta|\}$ where $\alpha^t := (b,a)$ for $\alpha=(a,b)$. This distance naturally extends to subsets of I, i.e., $d(A,B)=\min\{\, d(\alpha,\beta)\mid \alpha\in A,\beta\in B\,\}$ for any $A,B\subset I$.

Assumption (2.CD) (Polynomially decaying metric correlation structure). *For the $k=2$ point correlation we assume a decay of the type*

$$|\kappa(f_1(W),f_2(W))| \le \frac{C}{1+d(\operatorname{supp} f_1,\operatorname{supp} f_2)^s}\,\|f_1\|_2\,\|f_2\|_2\,, \tag{2.3a}$$

for some $s>12$ and all square integrable functions f_1,f_2 on $N\times N$ matrices. For $k\ge 3$ we assume a decay condition of the form

$$|\kappa(f_1(W),\dots,f_k(W))| \le_k \prod_{e\in E(T_{\min})} |\kappa(e)|\,, \tag{2.3b}$$

where $T_{\min}$ is the minimal spanning tree in the complete graph on the vertices $1,\dots,k$ with respect to the edge length $d(\{i,j\}) = d(\operatorname{supp} f_i,\operatorname{supp} f_j)$, i.e. the tree for which the sum of the lengths $d(e)$ is minimal, and $\kappa(\{i,j\})=\kappa(f_i,f_j)$.

A correlation decay of type (2.3b) is typical for various statistical physics models, see, e.g. [65]. Besides the assumptions on the decay of correlations we also impose a *flatness condition* to guarantee the stability of the Dyson equation:

Assumption (2.E) (Flatness). *There exist constants* $0 < c < C$ *such that*

$$c \langle T \rangle \leq \mathcal{S}[T] \leq C \langle T \rangle$$

for any positive semi-definite matrix T.

Flatness is a certain *mean field* condition on the random matrix W. In particular, choosing T to be the diagonal matrix with a single nonzero entry in the (i, i) element, flatness implies that the variances of the matrix elements $\mathbf{E}\,|w_{ij}|^2$ are comparable for all $i, j = 1, \ldots, N$.

2.2.3 Local law

We now formulate our main theorem on the isotropic and averaged local laws. They compare the resolvent G with the (unique) solution to the MDE in (2.1) away from the spectral edges. To specify the range of spectral parameters z we define two spectral domains specified via any given parameters $\delta, \gamma > 0$. Outside of the self-consistent spectrum we will work on

$$\mathbb{D}^{\delta}_{\text{out}} := \left\{ z \in \mathbb{H} \;\middle|\; |z| \leq N^{C_0}, \operatorname{dist}(z, \operatorname{supp} \mu) \geq N^{-\delta} \right\}$$

for some arbitrary fixed $C_0 \geq 100$. Under Assumption (2.E), which guarantees the existence of a density ρ, we consider the spectral domains

$$\mathbb{D}^{\delta}_{\gamma} := \left\{ z \in \mathbb{H} \;\middle|\; |z| \leq N^{C_0},\; \Im z \geq N^{-1+\gamma},\; \rho(\Re z) + \operatorname{dist}(\Re z, \operatorname{supp} \mu) \geq N^{-\delta} \right\}$$

that will be used away from the edges of the self-consistent spectrum.

Theorem 2.2.1 (Local law outside of the spectrum and global law). *Under Assumptions (2.A), (2.B) and (2.CD), the following statements hold: For any* $\epsilon > 0$ *there exists* $\delta > 0$ *such that for all* $D > 0$ *we have the isotropic law away from the spectrum,*

$$\mathbf{P}\left(|\langle \mathbf{x}, (G - M)\mathbf{y} \rangle| \leq \|\mathbf{x}\| \, \|\mathbf{y}\| \, \frac{N^{\epsilon}}{\langle z \rangle^2 \sqrt{N}} \quad \textit{in} \quad \mathbb{D}^{\delta}_{\textit{out}} \right) \geq 1 - CN^{-D} \tag{2.4a}$$

for all deterministic vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$ *and we have the averaged law away from the spectrum,*

$$\mathbf{P}\left(|\langle B(G - M) \rangle| \leq \|B\| \, \frac{N^{\epsilon}}{\langle z \rangle^2 N} \quad \textit{in} \quad \mathbb{D}^{\delta}_{\textit{out}} \right) \geq 1 - CN^{-D} \tag{2.4b}$$

for all deterministic matrices $B \in \mathbb{C}^{N \times N}$. *In fact, for small* ϵ, δ *can be chosen such that* $\delta = c\epsilon$ *for some absolute constant* $c > 0$. *Here* $G = G(z)$, $M = M(z)$ *and* $C = C(D, \epsilon)$ *is some constant, depending only on its arguments and the constants in Assumptions (2.A)–(2.CD). Moreover, instead of Assumption (2.CD) it is sufficient to assume the more general Assumptions (2.C) (or (2.C)' for complex Hermitian matrices) and (2.D), as stated in Section 2.2.5.*

If we additionally assume flatness in the form of Assumption (2.E), then we also obtain an optimal local law away from the spectral edges, especially in the bulk,

Theorem 2.2.2 (Local law in the bulk of the spectrum). *Under Assumptions (2.A), (2.B), (2.CD) and (2.E), the following statements hold: For any $\gamma, \epsilon > 0$ there exists $\delta > 0$ such that for all $D > 0$ we have the isotropic law in the bulk,*

$$\mathbf{P}\left(|\langle \mathbf{x}, (G-M)\mathbf{y}\rangle| \le \|\mathbf{x}\|\,\|\mathbf{y}\|\,\frac{N^\epsilon}{\sqrt{N\Im z}} \quad in \quad \mathbb{D}_\gamma^\delta\right) \ge 1 - CN^{-D} \tag{2.5a}$$

for all deterministic vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$ and we have the averaged law in the bulk,

$$\mathbf{P}\left(|\langle B(G-M)\rangle| \le \|B\|\,\frac{N^\epsilon}{N\Im z} \quad in \quad \mathbb{D}_\gamma^\delta\right) \ge 1 - CN^{-D} \tag{2.5b}$$

for all deterministic matrices $B \in \mathbb{C}^{N\times N}$. In fact, δ can be chosen such that $\delta = c\min\{\epsilon, \gamma\}$ for some absolute constant $c > 0$. Here $C = C(D, \epsilon, \gamma)$ is some constant, depending only on its arguments and the constants in Assumptions (2.A)–(2.E). Moreover, as in the previous theorem, instead of Assumption (2.CD) it is sufficient to assume the more general Assumptions (2.C) (or (2.C)' for complex Hermitian matrices) and (2.D), as stated in Section 2.2.5.

Note that both theorems cover the regime where z is far away from the spectrum; in this case the estimates in Theorem 2.2.1 are stronger and require less conditions. Theorem 2.2.2 is really relevant when $\Re z$ is inside the bulk of the spectrum and $\Im z$ is very small; this is why we called it local law in the bulk. In the literature this regime is typically characterized by $\rho(\Re z) \ge \delta$ for some $\delta > 0$, but in Theorem 2.2.2 it is extended to $\rho(\Re z) \ge N^{-\delta}$ for some sufficiently small $\delta > 0$.

2.2.4 Delocalization, rigidity and universality

The local law is the main input for eigenvector delocalization, eigenvalue rigidity and universality, as stated below. We formulate them as corollaries since they follow from a general theory that has been developed recently. We explain how to adapt the general arguments to prove these corollaries in Sections 2.5.4 and 2.6.

Corollary 2.2.3 (No eigenvalues outside the support of the self-consistent density). *Under the assumptions of Theorem 2.2.1 there exists a $\delta > 0$ such that for any $D > 0$,*

$$\mathbf{P}\left(\operatorname{Spec} H \not\subset (-N^{-\delta}, N^{-\delta}) + \operatorname{supp}\mu\right) \le_D N^{-D},$$

where $\operatorname{supp}\mu \subset \mathbb{R}$ *is the support of the self-consistent density of states μ.*

Corollary 2.2.4 (Bulk delocalization). *Under the assumptions of Theorem 2.2.2 it holds for an ℓ^2-normalized eigenvector $\mathbf{u}$ corresponding to a bulk eigenvalue λ of H that*

$$\mathbf{P}\left(\max_{a\in J}|u_a| \ge \frac{N^\epsilon}{\sqrt{N}},\ H\mathbf{u} = \lambda\mathbf{u},\ \rho(\lambda) \ge \delta\right) \le_{\epsilon,\delta,D} N^{-D}$$

for any $\epsilon, \delta, D > 0$.

Corollary 2.2.5 (Bulk rigidity). *Under the assumptions of Theorem 2.2.2 the following holds. Let $\lambda_1 \le \cdots \le \lambda_N$ be the ordered eigenvalues of H and denote the classical position of the eigenvalue close to energy $E \in \mathbb{R}$ by*

$$k(E) := \left\lceil N\int_{-\infty}^{E} \rho(x)\,\mathrm{d}x\right\rceil,$$

where $\lceil\cdot\rceil$ *denotes the ceiling function. It then holds that*

$$\mathbf{P}\left(\sup\left\{\,\left|\lambda_{k(E)}-E\right|\;\middle|\;E\in\mathbb{R},\,\rho(E)\ge\delta\,\right\}\ge\frac{N^{\epsilon}}{N}\right)\le_{\epsilon,\delta,D}N^{-D}$$

for any $\epsilon,\delta,D>0$.

For proving the bulk universality we replace the lower bound from Assumption (2.E) by the following, stronger, assumption:

Assumption (2.F) (Fullness). *There exists a constant* $\lambda>0$ *such that*

$$\mathbf{E}\left|\operatorname{Tr}BW\right|^{2}\ge\lambda\operatorname{Tr}B^{2}$$

for any deterministic matrix B *of the same symmetry class as* H.

Fullness is a technical condition which ensures that the covariance matrix of W is bounded from below by that of a full GUE or GOE matrix with variance λ. Note this is the only condition that induces the difference between the complex Hermitian and real symmetric symmetry classes in the following universality statement.

Corollary 2.2.6 (Bulk universality). *Under the assumptions of Theorem 2.2.2 and additionally Assumption (2.F) the following holds: Let* $k\in\mathbb{N}$, $\delta>0, E\in\mathbb{R}$ *with* $\rho(E)\ge\delta$ *and let* $\Phi\colon\mathbb{R}^{k}\to\mathbb{R}$ *be a compactly supported smooth test function. Denote the* k*-point correlation function of the eigenvalues of* H *by* ρ_k *and denote the corresponding* k*-point correlation function of the GOE/GUE-point process by* Υ_k. *Then there exists a positive constant* $c=c(\delta,k)>0$ *such that*

$$\left|\int_{\mathbb{R}^k}\Phi(\boldsymbol{t})\left[\frac{1}{\rho(E)}\rho_k\Big(E\mathbf{1}+\frac{\boldsymbol{t}}{N\rho(E)}\Big)-\Upsilon_k(\boldsymbol{t})\right]\mathrm{d}\boldsymbol{t}\right|\le_{\Phi,\delta,k}N^{-c},$$

$$\begin{aligned}\Bigg|\mathbf{E}\,\Phi\Big(&\big(N\rho(\lambda_{k(E)})[\lambda_{k(E)+j}-\lambda_{k(E)}]\big)_{j=1}^{k}\Big)\\&-\mathbf{E}_{\mathrm{GOE/GUE}}\,\Phi\Big(\big(N\rho_{sc}(0)[\lambda_{\lceil N/2\rceil+j}-\lambda_{\lceil N/2\rceil}]\big)_{j=1}^{k}\Big)\Bigg|\le_{\Phi,\delta,k}N^{-c},\end{aligned}$$

where $\mathbf{1}$ *is the vector of* k *ones,* $\mathbf{1}=(1,\dots,1)$, *the expectation* $\mathbf{E}_{\mathrm{GOE/GUE}}$ *is taken with respect to the Gaussian matrix ensemble in the same symmetry class as* H, *and* ρ_{sc} *denotes the semicircular density.*

Remark 2.2.7. *We chose the standard Euclidean distance on* J *in the formulation of Assumption (2.CD) merely for convenience. In the context of [8] a similar key assumption was formulated in terms of a pseudometric* δ *on* J *which has sub-*P *dimensional volume, i.e.,*

$$\max_{a\in J}\left|\{\,b\in J\mid\delta(a,b)\le\tau\,\}\right|\le\tau^{P}$$

for all $\tau>1$ *and some* $P>0$. *This pseudometric naturally extends to* I *as a product metric modulo the symmetry,*

$$\delta_2((a,b),(c,d)):=\min\{\max\{\delta(a,c),\delta(b,d)\},\max\{\delta(a,d),\delta(b,c)\}\}$$

and to any two subsets A,B *of* I *as* $\delta_2(A,B):=\min\{\,\delta_2(\alpha,\beta)\mid\alpha\in A,\beta\in B\,\}$. *All our results hold in this more general setup as well if* d *is replaced by* δ_2 *in Assumption (2.CD) and we require that* $s>12P$. *We do not pursue the pseudometric formulation further in the present work since the relaxed decay conditions formulated in Section 2.2.5 are more general as they allow for further symmetries in the matrix, for which (2.CD) is not satisfied irrespective of the pseudometric. A typical example for such an additional symmetry is the fourfold model (see [11]).*

2.2.5 Relaxed assumption on correlation decay

We now state the more general conditions on the correlation structure which are actually used in the proof of Theorem 2.2.2 and its corollaries, and are implied by Assumption (2.CD). For the more general conditions we split the correlation into two regimes. In the short range regime we express the correlation decay as a condition on cumulants, while in the long range regime, beyond neighbourhoods of size $\sqrt{N}$, we impose a mixing condition.

In the short range regime we assume the boundedness of certain norms on cumulants $\kappa(\alpha_1,\dots,\alpha_k) := \kappa(w_{\alpha_1},\dots,w_{\alpha_k})$ of matrix entries w_α, which are modifications of the usual ℓ^1-summability condition

$$\frac{1}{N^2}\sum_{\alpha_1,\dots,\alpha_k} |\kappa(\alpha_1,\dots,\alpha_k)| < \infty.$$

Cumulant norms

In order to formulate the conditions on the cumulants concisely, we from now on assume that W is real symmetric. We refer the reader to Appendix 2.C for the necessary modifications for the complex Hermitian case. In Appendix 2.A we will recall the equivalent analytical and combinatorial definitions of κ for the reader's convenience (see also [163]). We note that κ is invariant under any permutation of its arguments. Here we recall one central property of cumulants (which is also proved in the appendix): If $w_{\alpha_1},\dots,w_{\alpha_j}$ are independent from $w_{\alpha_{j+1}},\dots,w_{\alpha_k}$ for some $1\le j\le k-1$, then $\kappa(\alpha_1,\dots,\alpha_k)$ vanishes. Intuitively, the k-th order cumulant $\kappa(\alpha_1,\dots,\alpha_k)$ measures the part of the correlation of $w_{\alpha_1},\dots,w_{\alpha_k}$, which is truly of k-body type. For our results, cumulants of order four and higher require simple ℓ^1-type bounds, while the second and third order cumulants are controlled in specific, somewhat stronger norms. Finiteness of these norms imply a decay of correlation in a certain combinatorial sense even without a distance on the index set I. The isotropic and the averaged bound on D require slightly different norms, so we define two sets of norms distinguished by appropriate superscripts and we also define their sums without superscript.

We first introduce some custom notations which keep the definition of the cumulant norms relatively compact. If, in place of an index $a\in J$, we write a dot $(\cdot)$ in a scalar quantity then we consider the quantity as a vector indexed by the coordinate at the place of the dot. For example $\kappa(a_1\cdot, a_2b_2)$ is a J-vector, the i-th entry of which is $\kappa(a_1 i, a_2 b_2)$, and $\|\kappa(a_1\cdot,a_2b_2)\|$ is its (Euclidean) vector norm. Similarly, $\|A(*,*)\|$ refers to the operator norm of the matrix with matrix elements $A(i,j)$. We also define a combination of these conventions, in particular $\|\,\|\kappa(\mathbf{x}*,*\cdot)\|\,\|$ will denote the operator norm $\|A\|$ of the matrix A with matrix elements $A(i,j) = \|\kappa(\mathbf{x}i,j\cdot)\| = \|\sum_a x_a\kappa(ai,j\cdot)\|$. Since $\|A\| = \|A^t\|$ this does not introduce ambiguities with respect of the order of i,j. Notice that we use dot $(\cdot)$ for the dummy variable related to the inner norm and star $(*)$ for the outer norm.

For k-th order cumulants we set

$$|||\kappa|||_k := |||\kappa|||_k^{\mathrm{av}} + |||\kappa|||_k^{\mathrm{iso}}, \qquad |||\kappa|||^{\mathrm{av/iso}} = |||\kappa|||^{\mathrm{av/iso}}_{\le R} := \max_{2\le k\le R} |||\kappa|||_k^{\mathrm{av/iso}}, \tag{2.6a}$$

where the averaged norms are given by

$$\begin{aligned} |||\kappa|||_2^{\mathrm{av}} &:= \big\| \, |\kappa(*,*)| \, \big\|, \qquad |||\kappa|||_k^{\mathrm{av}} := N^{-2} \sum_{\alpha_1,\dots,\alpha_k} |\kappa(\alpha_1,\dots,\alpha_k)|, \qquad k \geq 4, \\ |||\kappa|||_3^{\mathrm{av}} &:= \Big\| \sum_{\alpha_1} |\kappa(\alpha_1,*,*)| \Big\| \\ &\quad + \inf_{\kappa=\kappa_{dd}+\kappa_{dc}+\kappa_{cd}+\kappa_{cc}} \Big(|||\kappa_{dd}|||_{dd} + |||\kappa_{dc}|||_{dc} + |||\kappa_{cd}|||_{cd} + |||\kappa_{cc}|||_{cc} \Big) \end{aligned} \tag{2.6b}$$

and the infimum is taken over all decompositions of κ in four symmetric functions κ_{dd}, κ_{cd}, etc. The letters d and c refer to "direct" and "cross", see Remark 2.2.8 below. The corresponding norms are given by

$$\begin{aligned} |||\kappa|||_{cc} = |||\kappa|||_{dd} &:= N^{-1} \sqrt{ \sum_{b_2,a_3} \Big(\sum_{a_2,b_3} \sum_{\alpha_1} |\kappa(\alpha_1, a_2b_2, a_3b_3)| \Big)^2 }, \\ |||\kappa|||_{cd} &:= N^{-1} \sqrt{ \sum_{b_3,a_1} \Big(\sum_{a_3,b_1} \sum_{\alpha_2} |\kappa(a_1b_1, \alpha_2, a_3b_3)| \Big)^2 }, \\ |||\kappa|||_{dc} &:= N^{-1} \sqrt{ \sum_{b_1,a_2} \Big(\sum_{a_1,b_2} \sum_{\alpha_3} |\kappa(a_1b_1, a_2b_2, \alpha_3)| \Big)^2 }. \end{aligned} \tag{2.6c}$$

For the isotropic bound we define

$$\begin{aligned} |||\kappa|||_2^{\mathrm{iso}} &:= \inf_{\kappa=\kappa_d+\kappa_c} \big(|||\kappa_d|||_d + |||\kappa_c|||_c \big) \\ |||\kappa|||_d &:= \sup_{\|\mathbf{x}\|\leq 1} \big\| \, \|\kappa(\mathbf{x}*, \cdot *)\| \, \big\|, \qquad |||\kappa|||_c := \sup_{\|\mathbf{x}\|\leq 1} \big\| \, \|\kappa(\mathbf{x}*, *\cdot)\| \, \big\|, \\ |||\kappa|||_k^{\mathrm{iso}} &:= \Big\| \sum_{\alpha_1,\dots,\alpha_{k-2}} |\kappa(\alpha_1,\dots,\alpha_{k-2},*,*)| \Big\|, \qquad k \geq 3, \end{aligned} \tag{2.6d}$$

where the inner norms in (2.6d) indicate vector norms and the outer norms operator norms, and the infimum is taken over all decomposition of κ into the sum of symmetric κ_c and κ_d.

Remark 2.2.8. *We remark that the particular form of the norms $|||\kappa|||_2^{iso}$ and $|||\kappa|||_3^{av}$ on κ is chosen to conform with the Hermitian symmetry. For example, in the case of Wigner matrices we have*

$$\kappa(a_1b_1, a_2b_2) = \delta_{a_1,a_2}\delta_{b_1,b_2} + \delta_{a_1,b_2}\delta_{b_1,a_2} =: \kappa_d(a_1b_1,a_2b_2) + \kappa_c(a_1b_1,a_2b_2), \tag{2.7}$$

i.e., the cumulant naturally splits into a direct and a cross part κ_d and κ_c. In general, the splitting $\kappa = \kappa_c + \kappa_d$ may not be unique but for the sharpest bound we can consider the most optimal splitting; this is reflected in the infimum in the definition of $|||\kappa|||_2^{iso}$. Note that in the example (2.7) *$|||\kappa_d|||_d$ and $|||\kappa_c|||_c$ are bounded, but $|||\kappa_c|||_d$ would not be. A similar rationale stands behind the definition of $|||\kappa|||_3^{av}$.*

We also remark that only the conditions on $|||\kappa|||_2^{iso}$ and $|||\kappa|||_3^{av}$ use the product structure $I = J \times J$. All other decay conditions are inherently conditions on index pairs $\alpha \in I$.

Assumption (2.C) (κ–correlation decay). *There exists a constant C such that for all $R \in \mathbb{N}$ and $\epsilon > 0$*

$$|||\kappa|||_2^{iso} \leq C, \qquad |||\kappa||| = |||\kappa|||_{\leq R} := \max_{2\leq k\leq R} |||\kappa|||_k \leq_{\epsilon,R} N^{\epsilon}$$

where the norms $|||\cdot|||_k$ and $|||\cdot|||_2^{iso}$ on k-th order cumulants were defined in (2.6). *If the matrix W is complex Hermitian we use Assumption (2.C)', as stated in Appendix 2.C instead of Assumption (2.C).*

Furthermore, in the long range regime beyond certain neighbourhoods of size $\ll \sqrt{N}$ we assume a finite polynomial decay of correlations that is reminiscent of the standard ρ-mixing condition in statistical physics (see, e.g. [48] for an overview of various mixing conditions). We will need this decay in a certain iterated sense that we now formulate precisely.

Assumption (2.D) (Higher order correlation decay). *There exists $\mu > 0$ such that the following holds: For every $\alpha \in I$ and $q, R \in \mathbb{N}$ there exists a sequence of nested sets $\mathcal{N}_k = \mathcal{N}_k(\alpha)$ such that $\alpha \in \mathcal{N}_1 \subset \mathcal{N}_2 \subset \cdots \subset \mathcal{N}_R = \mathcal{N} \subset I$, $|\mathcal{N}| \leq N^{1/2-\mu}$ and*

$$\kappa\Big(f(W_{I\setminus\bigcup_j \mathcal{N}_{n_j+1}(\alpha_j)}), g_1(W_{\mathcal{N}_{n_1}(\alpha_1)\setminus\bigcup_{j\neq 1}\mathcal{N}(\alpha_j)}), \dots, g_q(W_{\mathcal{N}_{n_q}(\alpha_q)\setminus\bigcup_{j\neq q}\mathcal{N}(\alpha_j)})\Big)$$
$$\leq_{R,q,\mu} N^{-3q} \left\|f\right\|_{q+1} \prod_{j=1}^{q} \left\|g_j\right\|_{q+1}$$

for any $n_1, \dots, n_q < R$, $\alpha_1, \dots, \alpha_q \in I$ and functions $f, g_1, \dots, g_q$. We will refer to these sets as "neighbourhoods" of α, although we do not assume any topological structure on I. For any $\mathcal{N} \subset I$, here $W_{\mathcal{N}}$ denotes the set of w_α indexed by $\alpha \in \mathcal{N}$.

Remark 2.2.9. *For the proof of Theorem 2.2.2 we need Assumptions (2.B), (2.C) and (2.D) only for finitely many values of q, R up to some threshold, depending only on the parameters D, γ, ϵ in the statement and μ from Assumption (2.D). This follows from the fact that the high moment bound from Theorem 2.4.1 is only needed for a finite value of p which relates to certain threshold on q, R.*

2.2.6 Some examples

We end this section by providing examples of correlated matrix models satisfying Assumptions (2.C)–(2.D). Our main example is the one already advertised in Assumption (2.CD). In Example 2.2.10 we check that Assumption (2.CD) indeed implies (2.C)–(2.D).

Example 2.2.10 (Polynomially decaying model). *Recall the metric setting of Assumption (2.CD). Simple calculations show that Assumption (2.C) is satisfied even if we only request $s \geq 2$ in* (2.3)*, independent of the chosen neighbourhood systems. As for Assumption (2.D), we define the neighbourhoods $\mathcal{N}_k(\alpha) := \{\, \beta \in I \mid d(\alpha, \beta) \leq k\, N^{1/4-\mu} \,\}$ so that $d(\mathcal{N}_k(\alpha), \mathcal{N}_{k+1}(\alpha)^c) = N^{1/4-\mu}$. To ensure that*

$$\Big|\kappa(f_1(W_{\mathcal{N}_n(\alpha)}), f_2(W_{\mathcal{N}_{n+1}(\alpha)^c}))\Big| \leq \frac{\left\|f_1\right\|_2 \left\|f_2\right\|_2}{N^3},$$

we thus have to choose $s \geq 12/(1-4\mu)$. The tree decay structure (2.3b) *then ensures that Assumption (2.D) is satisfied for all q.*

Example 2.2.11 (Block matrix). *For $n, M, N \in \mathbb{N}$ with $nM = N$ we set $J = [N]$ and consider an $n \times n$-block matrix with identical copies of an $M \times M$ Wigner matrix in each block. We introduce an equivalence relation on $I = J \times J$ in such a way that we first identify $a \sim b \in J$ if $a = b\,(mod\, M)$, and then $(a, b) \sim (c, d) \in I$ if $(a, b) = (c, d)$ or $(a, b) = (d, c)$ according* *to the Hermitian symmetry. Then the correlation structure is such that $\kappa(\alpha_1, \ldots, \alpha_k) = \mathcal{O}(1)$ if $\alpha_1, \ldots, \alpha_k$ all belong to the same equivalence class and $\kappa(\alpha_1, \ldots, \alpha_k) = 0$ otherwise. Since every entry is correlated with at most $\mathcal{O}(n^2)$ other entries, Assumptions (2.C), (2.D) are clearly satisfied as long as n is bounded.*

The same correlation structure is obtained if the blocks contain possibly different random matrices with independent entries (respecting only the overall Hermitian symmetry, but possibly without symmetry within each block), see e.g. the ensemble discussed in [13]. Furthermore, one may combine the block matrix model with a polynomially decaying model from Example 2.2.10 to construct yet another example for which Theorem 2.2.2 is applicable. In this general model the matrices in each block should merely exhibit a polynomially decaying correlation instead of strictly independent elements.

Example 2.2.12 (Correlated Gaussian matrix models). *Since all higher order cumulants for Gaussian random variables vanish, our method allows to prove the local law (and its corollaries) for correlated Gaussian random matrix models under even weaker conditions. In fact, besides Assumptions (2.A) and (2.E) (or (2.F) for universality) we only have to assume that*

$$|||\kappa|||_2^{av} + |||\kappa|||_2^{iso} \leq_\epsilon N^\epsilon$$

for all $\epsilon > 0$. In particular, this includes the polynomially decaying model from Example 2.2.10 for $s \geq 2$. These statements can be directly proved by following our general proof, setting all higher order cumulants to zero and using neighbourhoods $\mathcal{N}(\alpha) = I$ for all α. The details are left to the reader.

Example 2.2.13 (Fourfold symmetry). *A Wigner matrix W with fourfold symmetry is a matrix of independent entries w_α of unit variance up to the symmetries $w_{a,b} = w_{b,a} = w_{-a,-b} = w_{-b,-a}$ for all $a, b \in \mathbb{Z}/N\mathbb{Z}$. From the explicit formula*

$$\kappa(ab, cd) = \kappa_d(ab, cd) + \kappa_c(ab, cd) := (\delta_{a,c}\delta_{b,d} + \delta_{a,-c}\delta_{b,-d}) + (\delta_{a,d}\delta_{b,c} + \delta_{a,-d}\delta_{b,-c}),$$

and a similar one for the third order cumulants, Assumption (2.C) is straightforward to verify. By choosing the neighbourhoods $\mathcal{N}(\alpha)$ to contain the three other companions of α from the symmetry, it is obvious that also Assumption (2.D) is fulfilled. Strictly speaking, the flatness condition (2.E) is violated by the fourfold symmetry, but as the resulting M is diagonal, there is an easy replacement for the flatness. For more details on the random matrix model with a fourfold symmetry we refer the reader to [11].

A similar argument shows that Assumptions (2.C)–(2.D) are also satisfied for other symmetries which naturally split in such a way that $w_{a,b}$ is identified with $w_{f_1(a),f_2(b)}$ and $w_{g_1(b),g_2(a)}$ for a finite collection of functions f_i, g_i. The appropriate replacement for the flatness condition (2.E), however, has to be checked on a case-by-case basis.

2.3 General multivariate cumulant expansion

The goal of this section is the derivation of a finite-order multivariate cumulant expansion with a precise control on the approximation error.

2.3.1 Precumulants: Definition and relation to cumulants

We begin by introducing the concept of *pre-cumulants* and establishing some of their important properties. For any collection of random variables $X, Y_1, \dots, Y_m$ we define the quantities

$$\begin{aligned} K(X) &:= X \\ K_{t_1,\dots,t_m}(X;\boldsymbol{Y}) &= K_{t_1,\dots,t_m}(X;Y_1,\dots,Y_m) \\ &:= Y_m(\mathbb{1}_{t_m \le t_{m-1}} - \mathbf{E})Y_{m-1}(\mathbb{1}_{t_{m-1}\le t_{m-2}} - \mathbf{E})Y_{m-2}\dots Y_1(\mathbb{1}_{t_1\le 1} - \mathbf{E})X \end{aligned}$$

for $m \ge 1$, that depend on real parameters $t_1, \dots, t_m \in [0,1]$. We will call them *time ordered pre-cumulants*. We moreover introduce the integrated *symmetrized pre-cumulants*

$$K(X;\boldsymbol{Y}) := \sum_{\sigma\in S_{|\boldsymbol{Y}|}} \iiint_0^1 K_{t_1,\dots,t_{|\boldsymbol{Y}|}}(X;\sigma(\boldsymbol{Y}))\,\mathrm{d}\boldsymbol{t},$$

where $S_{|\boldsymbol{Y}|}$ is the group of permutations on a $|\boldsymbol{Y}|$-element set and $\mathrm{d}\boldsymbol{t} = \mathrm{d}t_1\dots\mathrm{d}t_m$ indicates integration over $[0,1]^{|\boldsymbol{Y}|}$. Note that the first variable X of $K(X;\boldsymbol{Y})$ plays a special role. Moreover, $K(X;\boldsymbol{Y})$ is invariant under permutations of the components of the vector $\boldsymbol{Y}$. These pre-cumulants are – other than the actual cumulants – random variables, but their expectations turn out to produce the traditional cumulants, justifying their name. While they appear to be very natural objects in the study of cumulants, we are not aware whether the pre-cumulants K have been previously studied, and whether the result of the following lemma is already known.

Lemma 2.3.1 (Pre-cumulant Lemma). *Let X be a random variable and let $\boldsymbol{Y}$, $\boldsymbol{Z}$ be random vectors. Then we have*

$$\mathbf{E}\,K(X;\boldsymbol{Y}) = \kappa(X,\boldsymbol{Y}), \tag{2.8a}$$

$$K(X;\boldsymbol{Y}) = \kappa(X,\boldsymbol{Y}) + X(\Pi\boldsymbol{Y}) - \sum_{\boldsymbol{Y}'\subset\boldsymbol{Y}}(\Pi\boldsymbol{Y}')\kappa(X,\boldsymbol{Y}\setminus\boldsymbol{Y}'), \tag{2.8b}$$

and the pre-cumulant decoupling identity

$$\begin{aligned} K(X;\boldsymbol{Y}\sqcup\boldsymbol{Z}) - \kappa(X,\boldsymbol{Y}\sqcup\boldsymbol{Z}) &= (\Pi\boldsymbol{Z})[K(X;\boldsymbol{Y}) - \kappa(X,\boldsymbol{Y})] \\ &\quad - \sum_{\substack{\boldsymbol{Y}'\subset\boldsymbol{Y}\\ \boldsymbol{Z}'\subsetneq\boldsymbol{Z}}}(\Pi\boldsymbol{Y}')(\Pi\boldsymbol{Z}')\kappa(X,(\boldsymbol{Y}\setminus\boldsymbol{Y}')\sqcup(\boldsymbol{Z}\setminus\boldsymbol{Z}')), \end{aligned} \tag{2.8c}$$

where $\boldsymbol{Y}'\subset\boldsymbol{Y}$ indicates that $\boldsymbol{Y}'$ is a sub-vector of $\boldsymbol{Y}$ (with $\boldsymbol{Y}'=\emptyset$ and $\boldsymbol{Y}'=\boldsymbol{Y}$ allowed) and $\boldsymbol{Y}\setminus\boldsymbol{Y}'$ is the vector of the remaining entries. By $\boldsymbol{Z}'\subsetneq\boldsymbol{Z}$ we denote all proper sub-vectors of $\boldsymbol{Z}$, i.e., not including $\boldsymbol{Z}$. By $\Pi\boldsymbol{Z}$ we mean the product of all entries of the vector $\boldsymbol{Z}$, while by $\boldsymbol{Z}\cup\boldsymbol{Y}$ we mean the concatenation of the two vectors $\boldsymbol{Z},\boldsymbol{Y}$. The order of the vector is of no importance as $K(X;\boldsymbol{Y})$ is symmetric with respect to the vector $\boldsymbol{Y}$ and κ is overall symmetric.

We note that (2.8c) is intentionally not symmetric in $\boldsymbol{Y}$, $\boldsymbol{Z}$, although an analogous formula holds with $\boldsymbol{Y}$ and $\boldsymbol{Z}$ interchanged. The relation (2.8c) should be interpreted as a refined version of the fact that centred precumulants factor independent random variables.

Indeed, if $\boldsymbol{Z}$ was independent of $X, \boldsymbol{Y}$, then the sum in (2.8c) would vanish by independence properties of the cumulant and (2.8c) would simplify to

$$K(X;\boldsymbol{Y}\sqcup\boldsymbol{Z})-\kappa(X,\boldsymbol{Y}\sqcup\boldsymbol{Z})=(\Pi\boldsymbol{Z})\big[K(X;\boldsymbol{Y})-\kappa(X,\boldsymbol{Y})\big].$$

In our applications $\boldsymbol{Z}$ will depend only very weakly on X and $\boldsymbol{Y}$, hence the sum in (2.8c) will be a small error term.

Proof. By the definition of the pre-cumulants, we have for $\boldsymbol{Y}=(Y_1,\dots,Y_m)$

$$\begin{aligned}K(X;\boldsymbol{Y})=\sum_{\sigma\in S_m}\iiint_0^1 Y_{\sigma(m)}(\mathbb{1}_{t_m\le t_{m-1}}-\mathbf{E})Y_{\sigma(m-1)}(\mathbb{1}_{t_{m-1}\le t_{m-2}}-\mathbf{E})\dots\\ \times(\mathbb{1}_{t_2\le t_1}-\mathbf{E})Y_{\sigma(1)}(\mathbb{1}_{t_1\le 1}-\mathbf{E})X\,\mathrm{d}\boldsymbol{t}.\end{aligned}$$

Multiplying out the brackets and pulling the characteristic functions involving the t-variables out of the expectations, each term is a product of moments of $(X,\boldsymbol{Y})$-monomials. We rearrange the sum according to the number of moments in the form that $K(X;\boldsymbol{Y})=\sum_{b=0}^m\phi_b$, where ϕ_b contains exactly b moments. These terms are given by

$$\begin{aligned}\phi_b=(-1)^b\sum_{1\le j_1<\dots<j_b\le m}\sum_{\sigma\in S_m}\iiint_0^1\mathbb{1}_{t_m\le\dots\le t_{j_b}}\mathbb{1}_{t_{j_b-1}\le\dots\le t_{j_{b-1}}}\dots\mathbb{1}_{t_{j_1-1}\le\dots\le t_1}\,\mathrm{d}\boldsymbol{t}\\ \times Y_{\sigma(m)}\dots Y_{\sigma(j_b)}(\mathbf{E}\,Y_{\sigma(j_b-1)}\dots Y_{\sigma(j_{b-1})})\dots\\ \times(\mathbf{E}\,Y_{\sigma(j_2-1)}\dots Y_{\sigma(j_1)})(\mathbf{E}\,Y_{\sigma(j_1-1)}\dots Y_{\sigma(1)}X),\end{aligned}\tag{2.9}$$

for $b\ge 1$, and the integral in (2.9) can be computed to give

$$\iiint_0^1[\cdots]\,\mathrm{d}\boldsymbol{t}=\frac{1}{(m-j_b+1)!}\frac{1}{(j_b-j_{b-1})!}\cdots\frac{1}{(j_2-j_1)!}\frac{1}{(j_1-1)!}=:V.$$

Here we introduced an additional variable $t_0=1$ for notational convenience and follow the convention that the last factor in (2.9) for $j_1=1$ reads $\mathbf{E}\,X$. For $b=0$ the analogue of (2.9) is given by

$$\phi_0=\Big(\sum_{\sigma\in S_m}\iiint_0^1\mathbb{1}_{t_m\le\dots\le t_1}\,\mathrm{d}\boldsymbol{t}\Big)Y_1\dots Y_mX=Y_1\dots Y_mX.$$

Let the summation indices $1\le j_1<\dots<j_b\le m$ be fixed and fix a labelled partition of $[m]=\pi_1\sqcup\dots\sqcup\pi_{b+1}$ into subsets of sizes $|\pi_1|=j_1-1,|\pi_2|=j_2-j_1,\dots,|\pi_b|=j_b-j_{b-1}$ and $|\pi_{b+1}|=m-j_b+1$. Those permutations σ in (2.9) for which $\sigma([1,j_1-1])=\pi_1,\sigma([j_1,j_2-1])=\pi_2,\dots,\sigma([j_{b-1},j_b-1])=\pi_b$ and $\sigma([j_b,m])=\pi_{m+1}$ all produce the same term $(-1)^bV\Pi\boldsymbol{Y}_{\pi_{b+1}}\dots(\mathbf{E}\,\Pi\boldsymbol{Y}_{\pi_2})(\mathbf{E}\,X\Pi\boldsymbol{Y}_{\pi_1})$, where $\boldsymbol{Y}_\pi=(\,Y_k\mid k\in\pi\,)$. We note that π_1 plays a special role since it is explicitly allowed to be the empty set, in which the last factor is just X. The combinatorial factor V is precisely cancelled by the number of such permutations, i.e., $1/V$. Thus (2.9) can be rewritten as

$$\phi_b=(-1)^b\sum_{\substack{\pi_1\sqcup\dots\sqcup\pi_{b+1}=[m]\\|\pi_j|\ge1\text{ for }j\ge2}}\Pi\boldsymbol{Y}_{\pi_{b+1}}(\mathbf{E}\,\Pi\boldsymbol{Y}_{\pi_b})\dots(\mathbf{E}\,\Pi\boldsymbol{Y}_{\pi_2})(\mathbf{E}\,X\Pi\boldsymbol{Y}_{\pi_1}),\tag{2.10a}$$

and therefore

$$K(X;\boldsymbol{Y}) = \sum_{b=0}^{m}(-1)^b \sum_{\substack{\pi_1\sqcup\dots\sqcup\pi_{b+1}=[m] \\ |\pi_j|\geq 1 \text{ for } j\geq 2}} \Pi\boldsymbol{Y}_{\pi_{b+1}}(\mathbf{E}\,\Pi\boldsymbol{Y}_{\pi_b})\dots(\mathbf{E}\,X\Pi\boldsymbol{Y}_{\pi_1}). \tag{2.10b}$$

We recognize the expectation of (2.10a) as the sum over all unlabelled partitions $\mathcal{P} \vdash (X,\boldsymbol{Y})$ with $|\mathcal{P}| = b+1$ blocks, under-counting by a factor of $b!$ as the first b factors on the rhs. of(2.10a) after taking the expectation are interchangeable (the last factor is special due to X). We can thus conclude that $\mathbf{E}\,K(X;\boldsymbol{Y})$ reads

$$\mathbf{E}\,K(X;\boldsymbol{Y}) = \sum_{b=0}^{m}(-1)^b b! \sum_{\substack{\mathcal{P}\vdash(X,\boldsymbol{Y}) \\ |\mathcal{P}|=b+1}} \prod_{A\in\mathcal{P}} \mathbf{E}\,\Pi(X,\boldsymbol{Y})_A = \kappa(X,\boldsymbol{Y}), \tag{2.11}$$

where we used (2.89) in the ultimate step, an identity that is equivalent to the analytical definition of the cumulant, see Appendix 2.A for more details. This completes the proof of (2.8a). Now (2.8b) follows from first separating $b=0$ to produce the $X(\Pi\boldsymbol{Y})$ term and then separating the π_{b+1} summation in (2.10b) so that $\boldsymbol{Y}_{\pi_{b+1}}$ plays the role of $\boldsymbol{Y}'$ for $\boldsymbol{Y}' \neq \emptyset$. The sum over the remaining moments is exactly the cumulant $\kappa(X,\boldsymbol{Y}\setminus\boldsymbol{Y}')$, see (2.11). Finally, the term $\boldsymbol{Y}' = \emptyset$ in (2.8b) cancels the first $\kappa(X,\boldsymbol{Y})$ term, completing the proof of (2.8b). The identity (2.8c) follows from (2.8b) where $\boldsymbol{Y}$ plays the role of $\boldsymbol{Y}\sqcup\boldsymbol{Z}$. The $\boldsymbol{Z}'=\boldsymbol{Z}$ term is considered separately, and then the identity (2.8b) is used again, this time for X and $\boldsymbol{Y}$. □

2.3.2 Precumulant expansion formula

We consider a random vector $\boldsymbol{w}\in\mathbb{R}^{\mathcal{I}}$, indexed by an abstract set $\mathcal{I}$, and a sufficiently often differentiable function $f\colon\mathbb{R}^{\mathcal{I}}\to\mathbb{C}$. The goal is to derive an expansion for $\mathbf{E}\,w_{i_0}f(w)$ in the variables indexed by a fixed subset $\mathcal{N}\subset\mathcal{I}$ that contains a distinguished element $i_0\in\mathcal{N}$. The expansion will be in terms of cumulants $\kappa(w_{i_1},\dots,w_{i_m})$ and expectations $\mathbf{E}\,\partial_{\boldsymbol{i}}f$ of derivatives $\partial_{\boldsymbol{i}}f := \partial_{i_1}\dots\partial_{i_m}f$, where we identify $\partial_i = \partial_{w_i}$ and $\boldsymbol{i} = \{i_1,\dots,i_m\}$. To state the expansion formula compactly we first introduce some notations and definitions. We recall that a multiset is an unordered set with possible multiple appearances of the same element. For a given tuple $\boldsymbol{i} = (i_1,\dots,i_m)\in\mathcal{N}^m$ we define the multisets

$$\underline{w}_{\boldsymbol{i}} := \{w_{i_1},\dots,w_{i_m}\} \quad \text{and the augmented multiset} \quad \underline{w}_{i_0\boldsymbol{i}} := \{w_{i_0}\}\sqcup\underline{w}_{\boldsymbol{i}},$$

where we consider $\sqcup$ as a disjoint union in the sense that $\underline{w}_{i_0\boldsymbol{i}}$ has $m+1$ elements (counting repetitions), regardless of whether $i_0 = i_k$ for some $k\in[m]$. Similarly, we write $\underline{w}_*\subset\underline{w}$ to indicate that $\underline{w}_*$ is a sub-multiset of a multiset $\underline{w}$. As cumulants are invariant under permutations of their entries we will write $\kappa(\underline{w})$ for multisets $\underline{w}$ of random variables. We will also write $\Pi\underline{w} := \prod_{j=1}^m w_{i_j}$ for the product of elements of a multiset $\underline{w} = \{\, w_{i_j} \mid j\in[m]\,\}$.

Equipped with Lemma 2.3.1 we can now state and prove the version of the multivariate cumulant expansion with a remainder that is best suitable for our application. Recall from (2.8a) that $\mathbf{E}\,K(w_{i_0};\underline{w}_{\boldsymbol{i}}) = \kappa(\underline{w}_{i_0\boldsymbol{i}})$.

Proposition 2.3.2 (Multivariate cumulant expansion). *Let $f\colon \mathbb{R}^{\mathcal{I}} \to \mathbb{C}$ be R times differentiable with bounded derivatives and let $\boldsymbol{w} \in \mathbb{R}^{\mathcal{I}}$ be a random vector with finite moments up to order R. Fix a subset $\mathcal{N} \subset \mathcal{I}$ and an element $i_0 \in \mathcal{N}$, then it holds that*

$$\mathbf{E}\, w_{i_0} f(w) = \sum_{m=0}^{R-1} \sum_{\boldsymbol{i}\in\mathcal{N}^m} \left[\mathbf{E}\, \frac{\kappa(\underline{w}_{i_0\boldsymbol{i}})}{m!} \partial_{\boldsymbol{i}} f + \mathbf{E}\, \frac{K(w_{i_0}; \underline{w}_{\boldsymbol{i}}) - \kappa(\underline{w}_{i_0\boldsymbol{i}})}{m!} \partial_{\boldsymbol{i}} f\big|_{\boldsymbol{w}_{\mathcal{N}}=0}\right] + \Omega, \tag{2.12a}$$

where

$$\Omega(f, i_0, \mathcal{N}) := \sum_{\boldsymbol{i}\in\mathcal{N}^R} \mathbf{E} \mathop{\iiint}\limits_0^1 K_{t_1,\ldots,t_R}(w_{i_0}, \ldots, w_{i_R})\, \mathrm{d}t_1 \ldots \mathrm{d}t_{R-1} \int_0^1 (\partial_{\boldsymbol{i}} f)(t_R \boldsymbol{w}', \boldsymbol{w}'')\, \mathrm{d}t_R, \tag{2.12b}$$

and where for $m = 0$ the derivative should be considered as the 0-th derivative, i.e. as the function itself. Here we introduced a decomposition $\boldsymbol{w} = (\boldsymbol{w}', \boldsymbol{w}'')$ of all random variables $\boldsymbol{w} = \boldsymbol{w}_{\mathcal{I}}$ such that $\boldsymbol{w}' = \boldsymbol{w}_{\mathcal{N}} = (w_i \,|\, i \in \mathcal{N})$ and $\boldsymbol{w}'' = \boldsymbol{w}_{\mathcal{N}^c} = (w_i \,|\, i \in \mathcal{I} \setminus \mathcal{N})$ and we write $f(\boldsymbol{w}) = f(\boldsymbol{w}', \boldsymbol{w}'')$. Moreover, if $\mathbf{E}\, |w_i|^{2R} \le \mu_{2R}$ for all $i \in \mathcal{I}$, then

$$|\Omega(f, i_0, \mathcal{N})| \le_R \mu_{2R}^{1/2} \sum_{\boldsymbol{i}\in\mathcal{N}^R} \int_0^1 \Big(\mathbf{E}\, |(\partial_{\boldsymbol{i}} f)(t_R \boldsymbol{w}', \boldsymbol{w}'')|^2\Big)^{1/2} \mathrm{d}t_R. \tag{2.13}$$

Proof. For functions $f = f(\boldsymbol{w})$, $g = g(\boldsymbol{w})$ a Taylor expansion yields, for any $s \ge 0$,

$$\begin{aligned}\mathbf{E}\, g(\boldsymbol{w}) f(s\boldsymbol{w}', \boldsymbol{w}'') &= (\mathbf{E}\, g)(\mathbf{E}\, f(0, \boldsymbol{w}'')) + \mathbf{Cov}(g, f(0, \boldsymbol{w}'')) \\ &\quad + \sum_{i\in\mathcal{N}} \int_0^s \mathbf{E}\, g(\boldsymbol{w}) w_i (\partial_i f)(t\boldsymbol{w}', \boldsymbol{w}'')\, \mathrm{d}t\end{aligned}$$

and after another Taylor expansion to restore $f(\boldsymbol{w}', \boldsymbol{w}'')$ in the first term we find

$$\begin{aligned}\mathbf{E}\, g(\boldsymbol{w}) f(s\boldsymbol{w}', \boldsymbol{w}'') &= (\mathbf{E}\, g)(\mathbf{E}\, f) + \mathbf{Cov}(g, f(0, \boldsymbol{w}'')) \\ &\quad + \sum_{i\in\mathcal{N}} \int_0^1 \mathbf{E}\, w_i [\mathbb{1}_{t\le s} g - (\mathbf{E}\, g)] (\partial_i f)(t\boldsymbol{w}', \boldsymbol{w}'')\, \mathrm{d}t.\end{aligned} \tag{2.14}$$

Here we follow the convention that if no argument is written, then $\mathbf{E}\, g = \mathbf{E}\, g(\boldsymbol{w})$. Starting with $g(\boldsymbol{w}) = w_{i_0}$, the last term in (2.14) requires to compute $\mathbf{E}\, K_t(w_{i_0}; w_i)(\partial_i f)(t\boldsymbol{w}', \boldsymbol{w}'')$ with $t = t_1$, $i = i_1$. So this has the structure $\mathbf{E}\, \widetilde{g}\widetilde{f}(t\boldsymbol{w}', \boldsymbol{w}'')$ with $\widetilde{g} = K_{t_1}$ and $\widetilde{f} = \partial_{i_1} f$ and we can use (2.14) again. Iterating this procedure with

$$(g(\boldsymbol{w}), s, i, t) = (K_{t_1,\ldots,t_{m-1}}(w_{i_0}; w_{i_1} \ldots, w_{i_{m-1}}), t_{m-1}, i_m, t_m)$$

for $m = 1, \ldots, R$, we arrive at

$$\begin{aligned}\mathbf{E}\, w_{i_0} f &= \sum_{m=0}^{R-1} \sum_{i_1,\ldots,i_m\in\mathcal{N}} \Big(\mathbf{E} \mathop{\iiint}\limits_0^1 K_{t_1,\ldots,t_m}\, \mathrm{d}\boldsymbol{t}\Big) (\mathbf{E}\, \partial_{\boldsymbol{i}} f) \\ &\quad + \sum_{m=0}^{R-1} \sum_{i_1,\ldots,i_l\in\mathcal{N}} \mathbf{E} \Big(\mathop{\iiint}\limits_0^1 K_{t_1,\ldots,t_m}\, \mathrm{d}\boldsymbol{t} - \mathbf{E} \mathop{\iiint}\limits_0^1 K_{t_1,\ldots,t_m}\, \mathrm{d}\boldsymbol{t}\Big) (\partial_{\boldsymbol{i}} f)(0, \boldsymbol{w}'') \\ &\quad + \sum_{i_1,\ldots,i_R\in\mathcal{N}} \mathbf{E} \mathop{\iiint}\limits_0^1 K_{t_1,\ldots,t_R}\, \mathrm{d}t_1 \ldots \mathrm{d}t_{R-1} \int_0^1 (\partial_{\boldsymbol{i}} f)(t_R \boldsymbol{w}', \boldsymbol{w}'')\, \mathrm{d}t_R,\end{aligned} \tag{2.15}$$

where $K_{t_1,\dots,t_m} = K_{t_1,\dots,t_m}(w_{i_0},\dots,w_{i_m})$ and $\mathrm{d}\boldsymbol{t} = \mathrm{d}t_1 \dots \mathrm{d}t_m$. We note that (2.15) does not include the sum over permutations, but since the summation over all $i_1,\dots,i_m$ is taken we can artificially insert the permutation as in

$$\sum_{i_1,\dots,i_m} \phi(i_1,\dots,i_m) = \frac{1}{m!} \sum_{i_1,\dots,i_m} \sum_{\sigma\in S_m} \phi(i_{\sigma(1)},\dots,i_{\sigma(m)}).$$

Now (2.12a) follows from combining (2.15) with (2.8a). Finally, (2.13) follows directly from a simple application of the Hölder inequality. □

2.3.3 Toy model

Proposition 2.3.2 will be the main ingredient for the probabilistic part of the proofs of Theorems 2.2.1 and 2.2.2. For pedagocial reasons we first demonstrate the multiplicative cancellation effect of *self-energy renormalization* through iterated cumulant expansion in a toy model.

Let f and $\boldsymbol{w}$ be as in Proposition 2.3.2 and let us suppose that $\mathcal{I}$ is equipped with a metric d. We furthermore assume that $\mathbf{E}\,\boldsymbol{w} = 0$ and that the multivariate cumulants of $\boldsymbol{w}$ follow a tree-like mixing decay structure as in Example 2.2.10, i.e.,

$$\kappa(f_1(\boldsymbol{w}),\dots,f_k(\boldsymbol{w})) \lesssim \prod_{\{i,j\}\in E(T_{\min})} \frac{1}{1 + d(\operatorname{supp} f_i, \operatorname{supp} f_j)^s} \tag{2.16}$$

for some $s > 0$, where $T_{\min}$ is the tree such that the sum of $d(\operatorname{supp} f_i, \operatorname{supp} f_j)$ along its edges $\{i,j\} \in E(T_{\min})$ is minimal. Fix now a finite positive integer parameter R and a large length scale $l > 0$. Around every $i \in \mathcal{I}$ we use the metric d to define neighbourhoods $\mathcal{N}(i) := \{\, j \in \mathcal{I} \mid d(i,j) \leq lR \,\}$ and $\mathcal{N}_k(i) := \{\, j \in I \mid d(i,j) \leq lk \,\}$, as in Assumption (2.D). For definiteness we furthermore assume that $\mathcal{I}$ has dimension two in the sense that $|\mathcal{N}| \sim l^2R^2$ as for the standard labelling of a matrix where $\mathcal{I} = [N]^2$. We now assume that f does not depend strongly on any single w_i, more specifically, for an multi-index $\boldsymbol{i}$ we assume

$$|\partial_{\boldsymbol{i}} f| \lesssim |\mathcal{N}|^{-(1+\epsilon)|\boldsymbol{i}|}, \qquad \boldsymbol{i} = (i_1,\dots,i_p), \qquad |\boldsymbol{i}| = p. \tag{2.17}$$

This bound ensures that the size of the derivative in the Taylor expansion in the neighbourhood $\mathcal{N}$ compensates for the combinatorics.

2.3.3.1 Expansion of a weakly dependent function

For this setup we want to study the size of the expression

$$\mathbf{E}\, w_{i_1} \dots w_{i_p} f(\boldsymbol{w})$$

where $i_1,\dots,i_p$ are in general position in the sense that their $\mathcal{N}(i_k)$ neighbourhoods do not intersect. If f were constant we could use the following lemma:

Lemma 2.3.3. *Assume that $\boldsymbol{w}$ has a tree-like correlation decay as in* (2.16) *and assume that the random variables $g_0(\boldsymbol{w}),\dots,g_p(\boldsymbol{w})$ have mutually l-separated supports, i.e. that*

$$d(\operatorname{supp} g_i, \operatorname{supp} g_j) \gtrsim l$$

for all $i \neq j$. If furthermore $\mathbf{E}\, g_k = 0$ for $k = 1, \dots, p$, then it holds that

$$|\mathbf{E}\, g_0 \dots g_p| \lesssim l^{-s\lceil p/2\rceil}.$$

Proof. Due to a basic identity on cumulants, see (2.87), we have that

$$\mathbf{E}\, g_0 \dots g_p = \sum_{A_1 \sqcup \dots \sqcup A_k = [0,p]} \kappa(g_{A_1}) \dots \kappa(g_{A_k}),$$

where the sum goes over all partitions $[0, p]$ and $g_A = \{\, g_k \mid k \in A \,\}$. From (2.16) it follows that

$$|\kappa(g_{A_k})| \lesssim l^{-s(|A_k|-1)}$$

and due to the assumption of zero mean $\mathbf{E}\, g_k = 0$ for $k \in [p]$ we have that $\kappa(g_A) = 0$ whenever $A = \{k\}$ for some $k \in [p]$. It follows that the worst case is given by pair partitions with $|A_k| = 2$ for all A_k not containing 0 which completes the proof. □

From this lemma with $g_0 = 1$ and $g_k = w_{i_k}$ for $k = 1, \dots, p$ we conclude that for constant f we have the asymptotic bound $|f\,\mathbf{E}\, w_{i_1} \dots w_{i_p}| \lesssim l^{-s\lceil p/2\rceil}$ by the zero mean assumption $\kappa(w_i) = \mathbf{E}\, w_i = 0$. We now want to argue that for weakly dependent f as in (2.17) a similar bound still holds true although f depends on all variables. Note that the weak dependence renders the minimal spanning tree distance trivial and a direct application of (2.16) would not give any decay. For brevity, we introduce the notations

$$\kappa(i, \boldsymbol{j}) := \kappa(w_i, w_{\boldsymbol{j}}), \qquad K(i; \boldsymbol{j}) := K(w_i; w_{\boldsymbol{j}}),$$

i.e. we identify cumulants and precumulants as functions of indices rather than random variables. We begin by expanding the first w_{i_1} to obtain from (2.12a)

$$\begin{aligned} \mathbf{E}\, w_{i_1} \dots w_{i_p} f = \sum_{\boldsymbol{j}_1}^{\mathcal{N}(i_1)} \mathbf{E} \left[\frac{\kappa(i_1, \boldsymbol{j}_1)}{|\boldsymbol{j}_1|!} + \frac{K(i_1; \boldsymbol{j}_1) - \mathbf{E}\, K(i_1; \boldsymbol{j}_1)}{|\boldsymbol{j}_1|!} \Big|^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}(i_1)}=0} \right] \\ \times w_{i_2} \dots w_{i_p} \partial_{\boldsymbol{j}_1} f + \mathcal{O}\left(l^{-2\epsilon R}\right), \end{aligned} \tag{2.18}$$

where we set $\sum_{\boldsymbol{j}}^{\mathcal{N}} := \sum_{0 \le m < R} \sum_{\boldsymbol{j} \in \mathcal{N}^m}$ and the parameter R, the maximal order of the expansion, is omitted for brevity. The notation $|^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}}=0}$ means that in all expressions to the right, the argument $\boldsymbol{w}$ is set to zero in the set $\mathcal{N}$, i.e. $\boldsymbol{w}_{\mathcal{N}} = 0$. This effect includes expectation values and cumulants. Note that $|^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}_1}=0} |^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}_2}=0} = |^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}_1 \cup \mathcal{N}_2}=0}$, i.e. the effects of multiple $|^{\rightarrow}$ operators accumulate. For example,

$$f(w_1, w_2)|^{\rightarrow}_{w_1=0}\, g(w_1, w_2)|^{\rightarrow}_{w_2=0}\, h(w_1, w_2) = f(w_1, w_2) g(0, w_2) h(0, 0). \tag{2.19}$$

However, the order of $|^{\rightarrow}_{w_1=0}$ and $|^{\rightarrow}_{w_2=0}$ matters as long as there is a nontrivial function in between, clearly

$$g(w_1, w_2)|^{\rightarrow}_{w_2=0}\, f(w_1, w_2)|^{\rightarrow}_{w_1=0}\, h(w_1, w_2) = g(w_1, w_2) f(0, w_2) h(0, 0),$$

which is different from (2.19). Finally, the error term in (2.18) was estimated using (2.13), and by comparing the combinatorics $|\mathcal{N}|^R$ of the summation to the size of the R-th derivative,

$|\partial_{i_1} \dots \partial_{i_R} f| \le |\mathcal{N}|^{-(1+\epsilon)R}$. We will choose $R \approx ps/4\epsilon$ large, so that the error term is negligible.

Iterating this procedure, we find

$$\mathbf{E}\, w_{i_1} \dots w_{i_p} f = \Big(\prod_{k\in[p]} \sum_{\boldsymbol{j}_k}^{\mathcal{N}(i_k)} \Big) \mathbf{E} \overset{\rightarrow}{\prod_{k\in[p]}} \left[\frac{\kappa(i_k, \boldsymbol{j}_k)}{|\boldsymbol{j}_k|!} + \frac{K(i_k; \boldsymbol{j}_k) - \mathbf{E}\, K(i_k; \boldsymbol{j}_k)}{|\boldsymbol{j}_k|!} \Big|^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}(i_k)}=0} \right] \\ \times \partial_{\boldsymbol{j}_1} \dots \partial_{\boldsymbol{j}_p} f + \mathcal{O}\left(l^{-sp/2} \right) \tag{2.20}$$

where $\prod^{\rightarrow}_{k\in[p]} a_k$ indicates that the order of the factors a_k is taken to be increasing in k, i.e., as $a_1 \dots a_p$. This is important due to the noncommutativity of the effect of the $|^{\rightarrow}$ operation on subsequent factors. We now open the bracket in (2.20) and first consider the extreme case, where we take the product all the first terms from each bracket, i.e., the product of p factors with κ. In this case the summation is of order 1 as the cumulant assumption (2.16) implies that $\sum_{\boldsymbol{j}\in\mathcal{I}^k} |\kappa(i, \boldsymbol{j})| \lesssim 1$ for any fixed i_1 if $s \ge 2$. Therefore the worst case is when the least total number of derivatives is taken, i.e., when $|\boldsymbol{j}_l| = 1$ for all l, in which case $\left|\partial_{\boldsymbol{j}_1} \dots \partial_{\boldsymbol{j}_p} f\right| \lesssim |\mathcal{N}|^{-(1+\epsilon)p} \lesssim l^{-2p}$. Now we consider the other extreme case where all the $(K - \mathbf{E}\, K) = (K - \kappa)$ factors are multiplied. There we a priori do not see the smallness as the summation size $|\mathcal{N}|^{|\boldsymbol{j}_1|+\dots+|\boldsymbol{j}_p|}$ roughly cancels the derivative size $|\mathcal{N}|^{-(1+\epsilon)(|\boldsymbol{j}_1|+\dots+|\boldsymbol{j}_p|)}$. The desired smallness thus has to come from the correlation decay (2.16). We can, however not directly apply the tree-like decay structure since there does not have to be a "security distance" between the supports of $\boldsymbol{w}_{\boldsymbol{j}_k}$ and f. For those k with such a security we can apply the tree-like decay immediately, and for those k where there is no such security distance we instead use (2.8c) to write $K - \kappa$ approximately as the product of two functions whose supports are separated by a security distance of scale l. Indeed, if $\boldsymbol{j}_k$ is not separated from $\operatorname{supp} f$ at least by l, then by the pigeon hole principle of placing less than R labels into R nested layers, it splits into two groups $\boldsymbol{j}_k^{(i)}$ and $\boldsymbol{j}_k^{(o)}$ of "inside" and "outside" indices such that $\operatorname{dist}(\boldsymbol{j}_k^{(i)}, \boldsymbol{j}_k^{(o)}) \gtrsim l$. Now by (2.8c) we have that

$$K(i_k; \boldsymbol{j}_k) - \kappa(i_k; \boldsymbol{j}_k) = (\Pi \boldsymbol{j}_k^{(o)}) [K(i_k; \boldsymbol{j}_k^{(i)}) - \kappa(i_k, \boldsymbol{j}_k^{(i)})] \\ - \sum_{\boldsymbol{n}_k^{(o)} \subsetneq \boldsymbol{j}_k^{(o)}} \sum_{\boldsymbol{n}_k^{(i)} \subset \boldsymbol{j}_k^{(i)}} (\Pi \boldsymbol{n}_k^{(i)})(\Pi \boldsymbol{n}_k^{(o)}) \kappa(i_k, \boldsymbol{j}_k^{(i)} \setminus \boldsymbol{n}_k^{(i)}, \boldsymbol{j}_k^{(o)} \setminus \boldsymbol{n}_k^{(o)}), \tag{2.21}$$

where $\Pi \boldsymbol{j} := \Pi \boldsymbol{w}_{\boldsymbol{j}}$. When multiplying (2.21) for all k, in the product of the second terms we (multiplicatively) collect p decay factors l^{-s}, resulting in l^{-sp}. For the product of the first terms we have to estimate a term of the type $\mathbf{E}\, g_1 \dots g_p \widetilde{f}$ with g_k being zero mean random variables such that all factors have mutually l-separated support. Here we set $g_k := K(i_k; \boldsymbol{j}_k^{(i)}) - \kappa(i_k, \boldsymbol{j}_k^{(i)})$ and absorbed the $\Pi \boldsymbol{j}_k^{(o)}$ factors into $\widetilde{f}$. It follows that

$$|\mathbf{E}\, g_1 \dots g_p \widetilde{f}| \lesssim l^{-s\lceil p/2 \rceil},$$

from Lemma 2.3.3. In this argument we only considered the two extreme cases when we opened the bracket in (2.20) and even in the product $\Pi(K - \kappa)$, after using (2.21) for each factor we only considered the two extreme cases. There are many mixed terms in both steps but they can be estimated similarly and altogether we have

$$\left|\mathbf{E}\, w_{i_1} \dots w_{i_p} f\right| \lesssim l^{-2p} + l^{-sp/2},$$

i.e. a power law decay whose exponent is proportional to the number of factors.

2.3.3.2 Expansion of a product of weakly dependent functions and self-energy renormalization

Now we generalize the expansion from Section 2.3.3.1 and consider another simple example: the iterated expansion of multipole weakly dependent functions. In particular, we will introduce the concept of *self-energy renormalization*.

Let $f_1, \dots, f_p$ be some functions of $\boldsymbol{w}$ which also depend weakly on each single w_i in such a way that $|\partial_{\boldsymbol{j}} f| \lesssim |\mathcal{N}|^{-(1+\epsilon)|\boldsymbol{j}|}$, and let $i_1, \dots, i_p$ be in general position as in the previous example. We want to study

$$\mathbf{E} \prod_{k\in[p]} w_{i_k} f_k,$$

which, by (2.20) with f replaced by $\prod f_k$, can be expanded to

$$\mathbf{E} \prod_{k\in[p]} w_{i_k} f_k = \prod_{k\in[p]} \Bigg(\sum_{\boldsymbol{j}_k}^{\mathcal{N}(i_k)} \sum_{(\boldsymbol{j}_k^l)_{l\in[p]}=\boldsymbol{j}_k} \Bigg) \mathbf{E} \overset{\rightarrow}{\prod_{k\in[p]}} \\ \left[\frac{\kappa(i_k, \boldsymbol{j}_k)}{|\boldsymbol{j}_k|!} + \frac{K(i_k; \boldsymbol{j}_k) - \mathbf{E}\, K(i_k; \boldsymbol{j}_k)}{|\boldsymbol{j}_k|!} \Bigg|^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}(i_k)}=0} \right] \prod_{n\in[p]} (\partial_{\boldsymbol{j}^n} f_n) + \mathcal{O}\left(l^{-sp/2}\right).$$

Here the second sum is the sum over all partitions $\boldsymbol{j}_k^1 \sqcup \dots \sqcup \boldsymbol{j}_k^p = \boldsymbol{j}_k$ of the multi-index $\boldsymbol{j}_k$, the multi-index $\boldsymbol{j}^n$ is given by the disjoint union $\boldsymbol{j}^n = \boldsymbol{j}_1^n \sqcup \dots \sqcup \boldsymbol{j}_p^n$, and we choose $R \approx ps/4\epsilon$, as in the previous example (recall that R is the maximal order of expansion, i.e. $|\boldsymbol{j}_k| \le R$). Thus $\boldsymbol{j}_k^n$ encodes those derivatives hitting f_n which originate from the expansion according to w_{i_k}. By expanding the product we can rewrite this expression as

$$\mathbf{E} \prod_{k\in[p]} w_{i_k} f_k = \sum_{L_1 \sqcup L_2 = [p]} \mathbf{E} \prod_{k\in L_1} \left[\sum_{\boldsymbol{j}_k}^{\mathcal{N}(i_k)} \frac{\kappa(i_k, \boldsymbol{j}_k)}{|\boldsymbol{j}_k|!} \sum_{(\boldsymbol{j}_k^n)_{n\in[p]}=\boldsymbol{j}_k} \right] \\ \times \overset{\rightarrow}{\prod_{k\in L_2}} \left[\sum_{\boldsymbol{j}_k}^{\mathcal{N}(i_k)} \frac{K(i_k; \boldsymbol{j}_k) - \mathbf{E}\, K(i_k; \boldsymbol{j}_k)}{|\boldsymbol{j}_k|!} \Bigg|^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}(i_k)}=0} \sum_{(\boldsymbol{j}_k^n)_{n\in[p]}=\boldsymbol{j}_k} \right] \\ \times \prod_{n\in[p]} (\partial_{\boldsymbol{j}^n} f_n) + \mathcal{O}\left(l^{-sp/2}\right).$$

It turns out that in many relevant cases, in particular after the summation over $i_1, \dots, i_k$, the leading contribution comes from those $k \in L_1$ for which $|\boldsymbol{j}_k| = 1$ and $\left|\boldsymbol{j}_k^k\right| = 1$. To counteract these leading terms we subtract this contribution from each factor $w_{i_k} f_k$ and instead compute

$$\mathbf{E} \prod_{k\in[p]} [w_{i_k} f_k - \sum_{j\in\mathcal{N}(i_k)} \kappa(i_k, j) \partial_j f_k] \\ = \sum_{L_1 \sqcup L_2 = [p]} \mathbf{E} \prod_{k\in L_1} \left[\sum_{\boldsymbol{j}_k}^{\mathcal{N}(i_k)} \frac{\kappa(i_k, \boldsymbol{j}_k)}{|\boldsymbol{j}_k|!} \sum_{(\boldsymbol{j}_k^n)_{n\in[p]}=\boldsymbol{j}_k} \mathbb{1}(|\boldsymbol{j}_k^k| = 0 \text{ if } |\boldsymbol{j}_k| = 1) \right] \tag{2.22} \\ \times \overset{\rightarrow}{\prod_{k\in L_2}} \left[\sum_{\boldsymbol{j}_k}^{\mathcal{N}(i_k)} \frac{K(i_k; \boldsymbol{j}_k) - \mathbf{E}\, K(i_k; \boldsymbol{j}_k)}{|\boldsymbol{j}_k|!} \Bigg|^{\rightarrow}_{\boldsymbol{w}_{\mathcal{N}(i_k)}=0} \sum_{(\boldsymbol{j}_k^n)_{n\in[p]}=\boldsymbol{j}_k} \right] \prod_{n\in[p]} (\partial_{\boldsymbol{j}^n} f_n) + \mathcal{O}\left(l^{-sp/2}\right).$$

We note that this substraction or *self-energy renormalization* does not affect the power counting bound of $l^{-2p}+l^{-sp/2}$ because it does not change the order of the terms but only excludes certain allocations of derivatives. However, beyond power counting, this exclusion can still reduce the effective size of the term considerably, see Section 2.4 where f is the resolvent of a random matrix.

2.4 Bound on the error matrix D through a multivariate cumulant expansion

In this section we prove an isotropic and averaged bound on the error matrix D defined in (2.2), in the form of high-moment estimates using the multivariate cumulant expansion. To formalize the bounds, we define the high-moment norms for random variables X and random matrices A by

$$\|X\|_p := (\mathbf{E}\,|X|^p)^{1/p}, \quad \|A\|_p := \sup_{\|\mathbf{x}\|,\|\mathbf{y}\|\le 1} \|\langle \mathbf{x}, A\mathbf{y}\rangle\|_p = \Big[\sup_{\|\mathbf{x}\|,\|\mathbf{y}\|\le 1} \mathbf{E}\,|\langle \mathbf{x}, A\mathbf{y}\rangle|^p\Big]^{1/p},$$

where the supremum is taken over deterministic vectors $\mathbf{x}, \mathbf{y}$.

Theorem 2.4.1 (Bound on the Error). *Under Assumptions (2.A), (2.B) and (2.D), there exist a constant C_* such that for any $p\ge 1$, $\epsilon>0$, z with $\Im z \ge N^{-1}$, $B\in\mathbb{C}^{N\times N}$ and $\mathbf{x},\mathbf{y}\in\mathbb{C}^N$ it holds that*

$$\frac{\|\langle \mathbf{x}, D\mathbf{y}\rangle\|_p}{\|\mathbf{x}\|\,\|\mathbf{y}\|} \le_{\epsilon,p} (1+|||\mathcal{S}|||+|||\kappa|||^{iso}_{\le R})N^\epsilon\sqrt{\frac{\|\Im G\|_q}{N\Im z}} \times \Big(1+\langle z\rangle\|G\|_q\Big)^{\frac{C_*}{\mu}}\Big(1+\frac{\langle z\rangle\|G\|_q}{N^\mu}\Big)^{\frac{C_* p}{\mu}} \tag{2.23a}$$

$$\frac{\|\langle BD\rangle\|_p}{\|B\|} \le_{\epsilon,p} (1+|||\mathcal{S}|||+|||\kappa|||^{av}_{\le R})N^\epsilon\frac{\|\Im G\|_q}{N\Im z} \times \Big(1+\langle z\rangle\|G\|_q\Big)^{\frac{C_*}{\mu}}\Big(1+\frac{\langle z\rangle\|G\|_q}{N^\mu}\Big)^{\frac{C_* p}{\mu}}, \tag{2.23b}$$

where $q=C_*p^4/\mu\epsilon$, $R=4p/\mu$, and for convenience we separately defined

$$|||\mathcal{S}||| := |||\kappa|||^{iso}_2. \tag{2.24}$$

Remark 2.4.2. *We remark that the size of $\mathcal{S}$ can be effectively controlled by $|||\kappa|||^{iso}_2$, justifying the definition of $|||\mathcal{S}|||$. To see this we note that due to*

$$\mathcal{S}[V] = \frac{1}{N}\sum_{\alpha_1,\alpha_2}\kappa(\alpha_1,\alpha_2)\Delta^{\alpha_1}V\Delta^{\alpha_2}$$

an arbitrary partition of $\kappa=\kappa_c+\kappa_d$ naturally induces a partition $\mathcal{S}=\mathcal{S}_c+\mathcal{S}_d$. Furthermore, it is easy to see that $\|\mathcal{S}_c[V]T\|_p \le |||\kappa_c|||_c\,\|V\|_{2p}\,\|T\|_{2p}$ and $\|\mathcal{S}_d[V]T\|_p\le |||\kappa_d|||_d\,\|V\|_{2p}\,\|T\|_{2p}$, c.f. Lemma 2.D.2, thus

$$\|\mathcal{S}[V]T\|_p \le |||\kappa|||^{iso}_2\,\|V\|_{2p}\,\|T\|_{2p}.$$

Here we recall that the double-index α stands for a pair $\alpha = (a, b)$ of single indices, and that the matrix Δ^α is a matrix of 0's except for a 1 in the (a, b)-entry.

Remark 2.4.3. *We point out an additional feature of the estimates* (2.23a)–(2.23b)*: they not only provide the optimal power of* $\|\Im G\|_q / (N\Im z)$*, but the power of* $\|G\|_q$*, without an extra smallness factor* $N^{-\mu}$*, is independent of* p*. This will be essential in the second part of the proof of the local law, see* (2.76) *later.*

The main tool for proving Theorem 2.4.1 is the multivariate cumulant expansion from Proposition 2.3.2. To connect to the toy model considered in Section 2.3.3, we note that the *self-energy renormalization* of $N^{-1/2}WG$ is $-\mathcal{S}[G]G$, up to an irrelevant contribution from indices $j \notin \mathcal{N}(i_k)$ in (2.22). In this sense the error term $D = N^{-1/2}WG + \mathcal{S}[G]G$ is the difference of $N^{-1/2}WG$ and its self-energy renormalization. As already noted in the context of the toy model we recall that this substraction does not change the power counting of the resulting terms. It does, however, exclude certain allocations of derivatives which in the case of $N^{-1/2}WG$ means that the main contributions coming from the diagonal elements of the form G_{aa} are absent. Off-diagonal elements G_{ab} are smaller on average, in fact the main gain comes from the key formula about resolvents of Hermitian matrices

$$\sum_b |G_{ab}|^2 = \frac{\Im G_{bb}}{\eta},$$

where $\eta = \Im z$. This identity follows directly from the spectral theorem. In the physics literature it is often called *Ward identity* and we will refer to it with this name. Notice that a sum of order N is reduced to a $1/\eta$ factor, so the Ward identity effectively gains a factor of $1/(N\eta)$ over the naive power counting. However, this effect is available only if off-diagonal elements of the resolvent are summed up, the same reduction would not take place in the sum $\sum_a |G_{aa}|^2$ which remains of order N. So the precise index structure is important. The next calculation shows this effect in the simplest case.

Exemplary gain through self-energy renormalization

We now give a short calculation to demonstrate the role of self-energy renormalization term $\mathcal{S}[G]G$ while computing $\mathbf{E}\,\langle D\rangle^2$. Notice that

$$\begin{aligned}\langle D\rangle &= \frac{1}{N}\sum_a [\sum_b \frac{w_{ab}}{\sqrt{N}} G_{ba} + (\mathcal{S}[G]G)_{aa}] \\ &= \frac{1}{N}\sum_{a,b} [\frac{w_{ab}}{\sqrt{N}} G_{ba} - \sum_{c,d} \frac{\kappa(ab, cd)}{N} \partial_{cd} G_{ba}]\end{aligned} \tag{2.25}$$

is the sum of terms of the form $w_i f$ plus their self-energy renormalization

$$-N^{-1}\sum_{c,d} \kappa(ab, cd)\partial_{cd} G_{ba}$$

where $i = (a, b)$ and $f = G_{ab}$. We note that (2.25) is the direct analogue of the self-energy renormalization in the toy-model discussed in Section 2.3.3, see (2.22). In (2.25) we expanded $\mathcal{S}[V] = \sum_{\alpha,\beta} N^{-1}\kappa(\alpha, \beta)\Delta^\alpha V \Delta^\beta$ and used the fact that the resolvent derivative reads $\Delta_\alpha G = -G\Delta^\alpha G$. Thus one should think of $\mathcal{S}[G]G$ as being the matrix self-energy renormalization of $N^{-1/2}WG$. To present this example in the simplest form, we assume that W is a Gaussian random matrix which automatically makes all higher order cumulants vanish. We find

$$\begin{aligned}\mathbf{E}\,\langle D\rangle^2 = N^{-1}&\sum_{\alpha_1,\beta_1}\kappa(\alpha_1,\beta_1)\,\mathbf{E}\,\langle\Delta^{\alpha_1}G\rangle\,\langle\Delta^{\beta_1}G\rangle\\ &+N^{-2}\sum_{\alpha_1,\beta_1}\kappa(\alpha_1,\beta_1)\sum_{\alpha_2,\beta_2}\kappa(\alpha_2,\beta_2)\,\mathbf{E}\,\langle\Delta^{\alpha_1}G\Delta^{\beta_2}G\rangle\,\langle\Delta^{\alpha_2}G\Delta^{\beta_1}G\rangle\,,\end{aligned}$$

the first term of which can be further bounded by

$$\begin{aligned}N^{-1}\sum_{\alpha_1,\beta_1}\left|\kappa(\alpha_1,\beta_1)\,\langle\Delta^{\alpha_1}G\rangle\,\langle\Delta^{\beta_1}G\rangle\right| &\le \frac{|||\kappa|||_2^{\mathrm{av}}}{N}\sum_{\alpha}|\langle\Delta^{\alpha}G\rangle|^2\\ &=\frac{|||\kappa|||_2^{\mathrm{av}}}{N^3}\sum_{a,b}|G_{ba}|^2=\frac{|||\kappa|||_2^{\mathrm{av}}}{N^2}\frac{\langle\Im G\rangle}{\eta}.\end{aligned}$$

For the second term we instead compute

$$\begin{aligned}&\sum_{\alpha_1,\beta_1}\sum_{\alpha_2,\beta_2}\left|\frac{\kappa(\alpha_1,\beta_1)\kappa(\alpha_2,\beta_2)}{N^2}\,\langle\Delta^{\alpha_1}G\Delta^{\beta_2}G\rangle\,\langle\Delta^{\alpha_2}G\Delta^{\beta_1}G\rangle\right|\\ &\qquad\le\frac{(|||\kappa|||_2^{\mathrm{av}})^2}{N^2}\sum_{\alpha_1,\alpha_2}|\langle\Delta^{\alpha_2}G\Delta^{\alpha_1}G\rangle|^2\\ &\qquad=\frac{(|||\kappa|||_2^{\mathrm{av}})^2}{N^4}\sum_{a_1,b_1,a_2,b_2}|G_{b_2a_1}|^2\,|G_{b_1a_2}|^2=(|||\kappa|||_2^{\mathrm{av}})^2\frac{\langle\Im G\rangle^2}{(N\eta)^2}\end{aligned}$$

and we conclude that

$$\mathbf{E}\,|\langle D\rangle|^2\le\frac{1}{N^2}\,\mathbf{E}\left[\frac{|||\kappa|||_2^{\mathrm{av}}\,\langle\Im G\rangle}{\eta}+\left(\frac{|||\kappa|||_2^{\mathrm{av}}\,\langle\Im G\rangle}{\eta}\right)^2\right],$$

which is small if $\eta\gg 1/N$. Without self-energy renormalization, however, i.e. for

$$\mathbf{E}\,\langle N^{-1/2}WG\rangle^2$$

we, for example, also encounter a term of the type

$$N^{-2}\sum_{\alpha_1,\beta_1}\kappa(\alpha_1,\beta_1)\sum_{\alpha_2,\beta_2}\kappa(\alpha_2,\beta_2)\,\mathbf{E}\,\langle\Delta^{\alpha_1}G\Delta^{\beta_1}G\rangle\,\langle\Delta^{\alpha_2}G\Delta^{\beta_2}G\rangle\,,$$

which is of order 1 because it lacks the gain from the Ward identity.

2.4.1 Computation of high moments of D through cancellation identities

Before going into the proof of Theorem 2.4.1, we sketch the strategy. For arbitrary linear (or conjugate linear in the sense that $\Lambda(\lambda\cdot)=\overline{\lambda}\Lambda(\cdot)$ for $\lambda\in\mathbb{C}$) functionals $\Lambda^{(1)},\dots,\Lambda^{(k)}$ we derive an explicit expansion for

$$\mathbf{E}\,\Lambda^{(1)}(D)\dots\Lambda^{(k)}(D)\tag{2.26}$$

in terms of joint cumulants $\kappa(\alpha_1,\dots,\alpha_k)$ of the entries of W and expectations of products of factors of the form

$$\Lambda(\Delta^{\alpha_1}G\Delta^{\alpha_2}G\dots G\Delta^{\alpha_k}G).$$

In other words, we express (2.26) solely in terms of matrix elements of G, which allows for a very systematic estimate. For the main part of the expansion we will then specialize to $\Lambda^{(k)}(D) = \langle BD\rangle$, $\Lambda^{(k)}(D) = \langle \mathbf{x}, D\mathbf{y}\rangle$ or their complex conjugates, and develop a graphical representation of the expansion. In this framework both the averaged and the isotropic bound on D reduce to a sophisticated power counting argument which – with the help of Ward estimates – directly gives the desired size of the averaged and isotropic error.

Equipped with the cumulant expansion from Proposition 2.3.2, we now aim at expressing $\mathbf{E}\,\Lambda^{(1)}(D)\dots\Lambda^{(p)}(D)$ for linear and conjugate linear functions $\Lambda^{(j)}$, purely in terms of the expectation of products of G's in the form

$$\Lambda_{\alpha_1,\dots,\alpha_k} := -(-1)^k N^{-k/2}\begin{cases}\Lambda(\Delta^{\alpha_1}G\dots\Delta^{\alpha_k}G) & \text{if } \Lambda \text{ is linear}\\ \Lambda(\Delta^{\alpha_1^t}G\dots\Delta^{\alpha_k^t}G) & \text{if } \Lambda \text{ is conjugate linear}\end{cases} \tag{2.27}$$

for double indices $\alpha_1,\dots,\alpha_k \in I = J\times J$, where we recall that for $\alpha = (a,b)$ the transpose α^t denotes $\alpha^t = (b,a)$. The sign choice will make the subsequent expansion sign-free. The reason for the $N^{-k/2}$ pre-factor is that the $\Lambda_{\alpha_1,\dots,\alpha_k}$ terms appear through k derivatives of G's each of which carries a $N^{-1/2}$ from the scaling $H = A + N^{-1/2}W$. Since the derivatives of G naturally come with many permutations from the Leibniz rule, we will also use the notations

$$\begin{aligned}\Lambda_{\{\alpha_1,\dots,\alpha_m\}} &:= \sum_{\sigma\in S_m}\Lambda_{\alpha_{\sigma(1)},\dots,\alpha_{\sigma(m)}}, \qquad \Lambda_{\alpha,\{\alpha_1,\dots,\alpha_m\}} := \sum_{\sigma\in S_m}\Lambda_{\alpha,\alpha_{\sigma(1)},\dots,\alpha_{\sigma(m)}},\\ \Lambda_{\underline{\alpha},\underline{\beta}} &:= \sum_{\alpha\in\underline{\alpha}}\Lambda_{\alpha,\underline{\alpha}\cup\underline{\beta}\setminus\{\alpha\}}\end{aligned} \tag{2.28}$$

for multisets $\{\alpha_1,\dots,\alpha_m\}$, $\underline{\alpha}$, $\underline{\beta}$. We will follow the convention that underlined Greek letters denote multisets of labels from I, while non-underlined Greek letters still denote single labels from I. By convention we set $\Lambda_\emptyset = \Lambda_{\emptyset,\underline{\beta}} = 0$. The last two definitions in (2.28) reflect the fact that the first index of Λ will often play a special role since derivatives of $\Lambda_{\alpha_1,\dots,\alpha_k}$ will all keep α_1 as their first index. With these notations, we note that

$$\Lambda_{\underline{\alpha}} = -\mathbb{1}(|\alpha|>0)\Lambda(G^{-1}\partial_{\underline{\alpha}}G), \qquad \Lambda_{\underline{\alpha},\underline{\beta}} = \partial_{\underline{\beta}}\Lambda_{\underline{\alpha}}$$

hold for arbitrary multisets $\underline{\alpha}$, where $|\underline{\alpha}|$ denotes the number of elements (counting multiplicity) in the multiset.

Expansion of a single factor of D

We now use Proposition 2.3.2 to compute $\mathbf{E}\,\Lambda(D)f$ for any random variable f (later f will be the product of the other Λ's). In the remainder of Section 2.4 the neighbourhoods $\mathcal{N} = \mathcal{N}(\alpha)$ are those from Assumption (2.D). The analogue of the length scale l from Section 2.3.3 is thus $N^{1/4-\mu/2}$, while the parameter R is still a large integer, depending only on p and μ. We expand

$$\mathbf{E}\,\Lambda(D)f = \mathbf{E}\,\frac{1}{\sqrt{N}}\sum_\alpha w_\alpha\Lambda(\Delta^\alpha G)f + \mathbf{E}\,\Lambda(\mathcal{S}[G]G)f = \mathbf{E}\sum_\alpha w_\alpha\Lambda_\alpha f + \mathbf{E}\,\Lambda(\mathcal{S}[G]G)f$$

and from (2.12a) we obtain

$$\mathbf{E}\,\Lambda(D)f = \sum_{\alpha}\sum_{0\le m<R}\sum_{\boldsymbol{\beta}\in\mathcal{N}^m}\mathbf{E}\left[\frac{\kappa(\alpha,\underline{\beta})}{m!}+\frac{K(\alpha;\underline{\beta})-\kappa(\alpha,\underline{\beta})}{m!}\bigg|_{W_{\mathcal{N}}=0}^{\rightarrow}\right]\partial_{\underline{\beta}}\Lambda_\alpha f$$
$$+\mathbf{E}\,\Lambda(\mathcal{S}[G]G)f+\sum_{\alpha}\Omega(\Lambda_\alpha f,\alpha,\mathcal{N}). \tag{2.29}$$

Here we follow the convention that $\boldsymbol{\beta}$ is the tuple with elements $(\beta_1,\dots,\beta_m)$ and $\underline{\beta}$ is the multiset obtained from the entries $\underline{\beta}=\{\beta_1,\dots,\beta_m\}$, and we recall that for $\mathcal{I}=I$ we denote $\kappa(w_{\alpha_1},\dots,w_{\alpha_k})$ and $K(w_{\alpha_1};w_{\alpha_2},\dots,w_{\alpha_k})$ by $\kappa(\alpha_1,\dots,\alpha_k)$ and $K(\alpha_1;\alpha_2,\dots,\alpha_k)$ (in contrast to the general setting of Section 2.3 where κ was viewed as a function of the random variables). For $m=0$ the first term in the first bracket of (2.29) vanishes due to $\kappa(\alpha)=\mathbf{E}\,w_\alpha=0$; for $m=1$ its contribution is given by

$$\sum_{\alpha\in I,\beta\in\mathcal{N}}\kappa(\alpha,\beta)\partial_\beta(\Lambda_\alpha f)=\sum_{\alpha\in I,\beta\in\mathcal{N}}\kappa(\alpha,\beta)\Lambda_{\alpha,\beta}f+\sum_{\alpha\in I,\beta\in\mathcal{N}}\kappa(\alpha,\beta)\Lambda_\alpha\partial_\beta f,$$

where we observe that the first term almost cancels the

$$\mathbf{E}\,\Lambda(\mathcal{S}[G]G)f=-\sum_{\alpha,\beta\in I}\kappa(\alpha,\beta)\Lambda_{\alpha,\beta}f$$

term except for the small contribution from $\beta\notin\mathcal{N}$. We thus rewrite (2.29) in the form

$$\mathbf{E}\,\Lambda(D)f=\mathbf{E}\sum_{\alpha\in I,\beta\in\mathcal{N}}\kappa(\alpha,\beta)\Lambda_\alpha\partial_\beta f \tag{2.30a}$$
$$+\mathbf{E}\sum_{\alpha\in I}\sum_{m<R}\sum_{\boldsymbol{\beta}\in\mathcal{N}^m}\left[\frac{\kappa(\alpha,\underline{\beta})}{l!}\mathbb{1}_{m\ge2}+\frac{K(\alpha;\underline{\beta})-\kappa(\alpha,\underline{\beta})}{m!}\bigg|_{W_{\mathcal{N}}=0}^{\rightarrow}\right]\partial_{\underline{\beta}}(\Lambda_\alpha f)$$
$$+\mathbf{E}\bigg(-\sum_{\alpha,\beta\in I}\kappa(\alpha,\beta)+\sum_{\alpha\in I,\beta\in\mathcal{N}}\kappa(\alpha,\beta)\bigg)\Lambda_{\alpha,\beta}f+\sum_{\alpha}\Omega(\Lambda_\alpha f,\alpha,\mathcal{N}).$$

In the above derivation of (2.30) we used directly that Λ is linear. In the case of conjugate linear we replace $\Lambda(D)$ by $\Lambda(D^*)$ which is linear again. This replacement is remedied by the fact that in the definition of $\Lambda_{\alpha_1,\dots,\alpha_k}$ in (2.27) we consider transposed double indices. More generally, following the same computation, we have

$$\mathbf{E}\,\Lambda(\partial_{\underline{\gamma}}D)f=\mathbf{E}\,\Lambda_{\underline{\gamma}}f+\mathbf{E}\sum_{\alpha\in I,\beta\in\mathcal{N}}\kappa(\alpha,\beta)\Lambda_{\alpha,\underline{\gamma}}\partial_\beta f \tag{2.30b}$$
$$+\mathbf{E}\sum_{\alpha\in I}\sum_{m<R}\sum_{\boldsymbol{\beta}\in\mathcal{N}^m}\left[\frac{\kappa(\alpha,\underline{\beta})}{m!}\mathbb{1}_{m\ge2}+\frac{K(\alpha;\underline{\beta})-\kappa(\alpha,\underline{\beta})}{m!}\bigg|_{W_{\mathcal{N}}=0}^{\rightarrow}\right]\partial_{\underline{\beta}}(\Lambda_{\alpha,\underline{\gamma}}f)$$
$$+\mathbf{E}\bigg(-\sum_{\alpha,\beta\in I}\kappa(\alpha,\beta)+\sum_{\alpha\in I,\beta\in\mathcal{N}}\kappa(\alpha,\beta)\bigg)\Lambda_{\alpha,\{\beta\}\sqcup\underline{\gamma}}f+\sum_{\alpha}\Omega(\Lambda_{\alpha,\underline{\gamma}}f,\alpha,\mathcal{N}).$$

We think of the first two terms and the first term of the square bracket in the third term (2.30b) as the leading order terms. The second summand in the third term will be small due to the structure of the pre-cumulants and the fact that the subsequent function $\partial\Lambda f$ has the $\mathcal{N}$-randomness removed. The fourth term is small because the two sums in the parenthesis almost cancel; and finally the fifth term will be small by choosing R sufficiently large. We call (2.30) (approximate) *cancellation identities* as they exhibit the cancellation of second order statistics due to the definition of $\mathcal{S}$ and D.

Iterated expansion of multiple factors of D

We now use (2.30b) repeatedly to compute $\mathbf{E}\prod_{k\in[p]}\Lambda^{(k)}(D)$. As a first step we expand the D in the $\Lambda^{(1)}$ factor, for which the special case (2.30a) is sufficient and we find

$$\begin{aligned}
&\mathbf{E}\,\Lambda^{(1)}(D)\prod_{k\geq 2}\Lambda^{(k)}(D)=\sum_{\alpha_1\in I}\Omega\Big(\Lambda^{(1)}_{\alpha_1}\prod_{k\geq 2}\Lambda^{(k)}(D),\alpha_1,\mathcal{N}(\alpha_1)\Big)\\
&+\mathbf{E}\sum_{\substack{\alpha_1\in I\\ \beta_1\in\mathcal{N}(\alpha_1)}}\kappa(\alpha_1,\beta_1)\Lambda^{(1)}_{\alpha_1}\partial_{\beta_1}\Big(\prod_{k\geq 2}\Lambda^{(k)}(D)\Big)\\
&+\mathbf{E}\Big(-\sum_{\alpha_1,\beta_1\in I}\kappa(\alpha_1,\beta_1)+\sum_{\substack{\alpha_1\in I\\ \beta_1\in\mathcal{N}(\alpha_1)}}\kappa(\alpha_1,\beta_1)\Big)\Lambda^{(1)}_{\alpha_1,\beta_1}\prod_{k\geq 2}\Lambda^{(k)}(D)\\
&+\mathbf{E}\sum_{\alpha_1\in I}\sum_{m<R}\sum_{\boldsymbol{\beta}_1\in\mathcal{N}(\alpha_1)^m}\left[\frac{\kappa(\alpha_1,\underline{\beta}_1)}{m!}\mathbb{1}_{m\geq 2}+\frac{K(\alpha_1;\underline{\beta}_1)-\kappa(\alpha_1,\underline{\beta}_1)}{m!}\Big|^{\rightarrow}_{W_{\mathcal{N}(\alpha_1)}=0}\right]\\
&\qquad\qquad\times\partial_{\underline{\beta}_1}\Big(\Lambda^{(1)}_{\alpha_1}\prod_{k\geq 2}\Lambda^{(k)}(D)\Big).
\end{aligned}\tag{2.31}$$

We now distribute the $\underline{\beta}_1$-derivatives in the last term among the $\Lambda^{(1)}_{\alpha_1}$ and $\Lambda^{(k)}(D)$ factors according to the Leibniz rule. We handle the ∂_{β_1} derivative in the second term similarly but observe that this is slightly different in the sense that the ∂_{β_1} derivative does not hit the $\Lambda^{(1)}_{\alpha_1}$ factor. In other words, terms involving second order cumulants ($m=1$) come with the restriction that $\partial_{\beta_1}\Lambda^{(1)}_{\alpha_1}$ derivative is absent. This is precisely the effect we already encountered in Section 2.3.3; the self-energy normalization does not cancel all second order terms, it merely puts a restriction on the index-allocations in such a way that gains through Ward estimates are guaranteed in all remaining terms. In order to write (2.31) more concisely we introduce the notations

$$\begin{aligned}
\sum_{\alpha_l,\boldsymbol{\beta}_l}^{\sim(l)}&:=\sum_{\alpha_l\in I}\sum_{1\leq m<R}\sum_{\boldsymbol{\beta}_l\in\mathcal{N}(\alpha_l)^m}\frac{\kappa(\alpha_l,\underline{\beta}_l)}{m!}\sum_{\underline{\beta}_l^1\sqcup\dots\sqcup\underline{\beta}_l^p=\underline{\beta}_l}\mathbb{1}\left(|\underline{\beta}_l^l|=0\text{ if }|\underline{\beta}_l|=1\right),\\
\sum_{\alpha_l,\boldsymbol{\beta}_l}^{*}&:=\sum_{\alpha_l\in I}\sum_{0\leq m<R}\sum_{\boldsymbol{\beta}_l\in\mathcal{N}(\alpha_l)^m}\sum_{\underline{\beta}_l^1\sqcup\dots\sqcup\underline{\beta}_l^p=\underline{\beta}_l}\frac{K(\alpha_l;\underline{\beta}_l)-\kappa(\alpha_l,\underline{\beta}_l)}{m!},\\
\sum_{\alpha_l,\beta_l^l}^{\#}&:=\Big[-\sum_{\alpha_l,\beta_l^l\in I}\kappa_{\mathcal{S}}(\alpha_l,\beta_l^l)+\sum_{\alpha_l\in I}\sum_{\beta_l^l\in\mathcal{N}(\alpha_l)}\kappa(\alpha_l,\beta_l^l)\Big],
\end{aligned}\tag{2.32}$$

where $\kappa_{\mathcal{S}}(\alpha_1,\dots,\alpha_k):=\kappa(\widetilde{w}_{\alpha_1},\dots,\widetilde{w}_{\alpha_k})$ and where $\widetilde{W}=(\widetilde{w}_\alpha)_{\alpha\in I}$ is an identical copy of W. The reason for introducing this identical copy will become apparent in the next step. We furthermore follow the convention that $\underline{\beta}_l^k=\emptyset$ if $\underline{\beta}_l^k$ does not appear in the summation (which is the case for all $k\neq l$ in $\sum^{\#}_{\alpha_l,\beta_l^l}$ in (2.33)). Using these notations we can write (2.31) as

$$
\begin{aligned}
\mathbf{E}\prod_{k\in[p]}\Lambda^{(k)}(D) &= \Omega \\
&+ \mathbf{E}\Bigg(\sum_{\alpha_1,\boldsymbol{\beta}_1}^{\sim(1)} + \sum_{\alpha_1,\boldsymbol{\beta}_1}^{*}\Big|_{W_{\mathcal{N}(\alpha_1)}=0}^{\rightarrow} + \sum_{\alpha_1,\beta_1^1}^{\#}\Bigg)\Lambda_{\alpha_1,\underline{\beta}_{1}^{1}}\prod_{k=2}^{p}\Lambda^{(k)}(\partial_{\underline{\beta}_{1}^{k}}D) + \Omega,
\end{aligned}
\tag{2.33}
$$

where the error term Ω collects all other terms and is defined in (2.34) below. We point out that the notations introduced in (2.32) implicitly depend on the parameter R determining the order of expansion.

Estimate of error term Ω

It remains to estimate the error term Ω which is bounded by

$$
\begin{aligned}
\Omega &:= \sum_{\alpha_1\in I}\Omega\Big(\Lambda_{\alpha_1}^{(1)}\prod_{k\geq 2}\Lambda^{(k)}(D),\alpha_1,\mathcal{N}(\alpha_1)\Big) \\
&\leq_R \sum_{\alpha_1,\boldsymbol{\beta}_1\in\mathcal{N}(\alpha_1)^R}\Big\|\partial_{\underline{\beta}_1}\Big(\Lambda_{\alpha_1}^{(1)}\prod_{k\geq 2}\Lambda^{(k)}(D)\Big)\Big|_{\widehat{W}_t}\Big\|_2
\end{aligned}
\tag{2.34}
$$

for some $t\in[0,1]$, where $\widehat{W}_t=\widehat{W}_t^{(\alpha_1)}=tW_{\mathcal{N}(\alpha_1)}+W_{\mathcal{N}(\alpha_1)^c}$, where we recall the definition of $\Omega(\Lambda,\alpha,f)$ in (2.12a) and its bound in (2.13). To further estimate this expression, we first distribute the $\partial_{\underline{\beta}_1}$ derivative to the p factors involving $\Lambda^{(1)},\dots,\Lambda^{(p)}$ following the Leibniz rule, and then separate those factors by a simple application of Hölder inequality into p factors of $\|\cdot\|_{2p}$ norms. Each of these factors can be written as a sum of terms of the type $\|\Lambda^{(k)}(\partial_{\underline{\gamma}}G|_{\widehat{W}_t})\|_{2p}$ or $\|\Lambda^{(k)}(\partial_{\underline{\gamma}}D|_{\widehat{W}_t})\|_{2p}$ for some derivative operator $\partial_{\underline{\gamma}}$. We can then estimate these norms using $\|\Lambda(R)\|_q\leq\|\Lambda\|\,\|R\|_q$ and

$$
\|\partial_{\underline{\gamma}}G|_{\widehat{W}_t}\|_q+\|\partial_{\underline{\gamma}}D|_{\widehat{W}_t}\|_q\leq_{|\underline{\gamma}|}N^{-|\underline{\gamma}|/2}(1+|||S|||)(1+\langle z\rangle\,\|G\|_{Cq|\underline{\gamma}|})^{|\underline{\gamma}|+5},
\tag{2.35}
$$

where the second inequality follows from Lemma 2.D.3, and we note that $Cp|\underline{\gamma}|\leq CRp^2$. We now count the total number of derivatives: There are $R+1$ derivatives from $|\underline{\beta}_1|$ and α_1, each providing a factor of $N^{-1/2}$. It remains to account for the $\alpha_1,\boldsymbol{\beta}_1$-sums which is at most of size $\sum_{\alpha_1}|\mathcal{N}(\alpha_1)|^R\leq N^{2+R/2-\mu R}$. We now choose R large enough so that

$$
N^{2-(R+1)/2+R/2-\mu R}\leq N^{-p},
$$

which is satisfied if we choose $R\geq 3p/\mu$. Combining these rough bounds we have shown that, up to irrelevant combinatorial factors,

$$
\Omega\leq_{p,\mu}N^{-p}\Big[\prod_{k=1}^{p}\|\Lambda^{(k)}\|\Big]\Big(1+|||\mathcal{S}|||\Big)^p\Big(1+\langle z\rangle\,\|G\|_{Cp^3/\mu}\Big)^{Cp/\mu}.
\tag{2.36}
$$

Main expansion formula for multiple factors of D

Formula (2.33) with the bound (2.36) on the error term is the first step where the cumulant expansion was used in the $\Lambda^{(1)}(D)$ factor. Now we iterate this procedure for the $\Lambda^{(2)}(D),\Lambda^{(3)}(D),\dots$ inductively. We arrive at the following proposition modulo the claimed bound on the overall error which we will prove after an extensive explanation.

Proposition 2.4.4. *Let $\Lambda^{(1)}, \dots, \Lambda^{(p)}$ be linear (or conjugate linear) functionals and let $p \in \mathbb{N}$ be given. Then we have*

$$\mathbf{E} \prod_{k\in[p]} \Lambda^{(k)}(D) = \Omega \tag{2.37}$$

$$+\, \mathbf{E} \overset{\rightarrow}{\prod_{l\in[p]}} \left(1 + \sum_{\alpha_l,\beta_l}^{\sim(l)} + \sum_{\alpha_l,\beta_l}^{*} \Big|_{W_{\mathcal{N}(\alpha_l)}=0}^{\rightarrow} + \sum_{\alpha_l,\beta_l^l}^{\#} \right) \prod_{k\in[p]} \begin{cases} \Lambda^{(k)}_{\alpha_k, \bigsqcup_{l\in[p]} \underline{\beta}_l^k} & \textit{if } \sum_{\alpha_k} \\ \Lambda^{(k)}_{\bigsqcup_{l<k} \underline{\beta}_l^k, \bigsqcup_{l>k} \underline{\beta}_l^k} & \textit{else,} \end{cases}$$

where "if $\sum_{\alpha_k}$" means cases where after multiplying out the first product $\prod_l$ the summation over the index α_k is performed. Under Assumptions (2.A), (2.B) and (2.D), the error term Ω is bounded by

$$|\Omega| \le_{p,\mu} N^{-p} \left[\prod_{k=1}^{p} \|\Lambda^{(k)}\| \right] (1 + |||\mathcal{S}|||)^p \Big(1 + \langle z \rangle \|G\|_q \Big)^{\frac{Cp}{\mu}} \left(1 + \frac{\langle z \rangle \|G\|_q}{N^\mu} \right)^{\frac{Cp^2}{\mu}}, \tag{2.38}$$

if we choose $R = 4p/\mu$ to be order of expansion in the summations, see (2.32). *Furthermore, we set $q := Cp^3/\mu$ for some constant C, and $\|\Lambda^{(k)}\|$ denotes the operator norm of the linear functional $\Lambda^{(k)}$.*

For (2.37) we recall the convention that $\underline{\beta}_l^k = \emptyset$ whenever $\underline{\beta}_l^k$ is not summed, i.e., for the contribution from the 1 in the l-th factor, or the contribution from $\sum^{\#}$ in the l-th factor for $k \neq l$. Moreover, we remind the reader that the custom notation $|_{W_{\mathcal{N}}=0}^{\rightarrow}$ was introduced right after (2.20). We also note that the terms with a 1 from the first factor vanish as they contain $\Lambda^{(1)}_{\emptyset, \sqcup_{l>1} \underline{\beta}_l^1} = 0$. Moreover, we can now explain why we introduced the identical copy $\widetilde{W}$ of W in the definition of $\kappa_{\mathcal{S}}$ in (2.32). The cumulants in the representation of the term $\mathcal{S}[G]G = -\sum_{\alpha,\beta\in I} \kappa_{\mathcal{S}}(\alpha,\beta)\Lambda_{\alpha,\beta}$ should not be affected by the restriction imposed by the operation $|_{W_{\mathcal{N}}=0}^{\rightarrow}$. Changing W to $\widetilde{W}$ within the definition of $\kappa_{\mathcal{S}}$ protects it from the action of $|_{W_{\mathcal{N}}=0}^{\rightarrow}$ that turns all subsequent W variables zero. This non-restriction of the particular sum is formally achieved by writing $\mathcal{S}$ in terms of $\kappa_{\mathcal{S}}$ instead of κ. This is only a notational pedantry, in the next step where we multiply (2.37) out, it will disappear. We remark that because of the effect of $|_{W_{\mathcal{N}}=0}^{\rightarrow}$ the order in which the product in (36) is performed matters. It starts with $l = 1$ and ends with $l = p$.

We point out that the estimate (2.38) not only provides the necessary N^{-p} factor, but it also involves at most $O(p)$ power of $\|G\|_q$ without an extra smallness factor $N^{-\mu}$, see Remark 2.4.3. While from the perspective of an N-power counting, any factor $\|G\|_q$ is neutral, of order one, we need to track that its power is not too big. Factors of $\|G\|_q$ that come with a factor $N^{-\mu}$ can be handled much easier and are not subject to the restriction of their power.

Reformulation of the main expansion formula

We now derive an alternative, less compact formula (2.39) for (2.37) which avoids the provisional $|^{\rightarrow}$ notation. By expanding the first product in (2.37) we can rearrange (2.37) according to partitions $[p] = L_1 \sqcup \dots \sqcup L_4$, where L_i contains those indices l for which the l-th factor in the product contributes with its i-th term. In particular $L := L_2 \sqcup L_3 \sqcup L_4 \subset [p]$ contains those indices l, for which $\alpha_l, \boldsymbol{\beta}_l$ are summed. We shall use the nomenclature that labels α_l and the elements of $\underline{\beta}_l$ are *type-l* labels. These labels have been generated in the l-th application of the cancellation identities (2.30). The partition $\underline{\beta}_l^1 \sqcup \cdots \sqcup \underline{\beta}_l^p = \underline{\beta}_l$ encodes how these labels have been distributed among the p factors via the Leibniz rule. Thus labels $\underline{\beta}_l^k$ have been generated on $\Lambda^{(k)}$ at the l-th application of (2.30). Thus L encodes the types of labels present in the different parts of the expansion. To specify the number of type–l labels we introduce the notations

$$M_l := |\underline{\beta}_l|, \quad M_l^k := |\underline{\beta}_l^k|.$$

Thus the number of labels of *type l* is $M_l + 1$ and the number of type l-labels in $\Lambda^{(k)}$ is $M_l^k + \delta_{lk}$. We observe that in all non-zero terms of (2.37) the labels α_l, $\underline{\beta}_l$ for $l \in L$ are distributed to the $\Lambda^{(1)}, \dots, \Lambda^{(p)}$ in such a way that

(a) there are p factors $\Lambda^{(1)}, \dots, \Lambda^{(p)}$,

(b) every $\Lambda^{(k)}$ carries at least one label (that is for all k, $\sum_{l \in L}(M_l^k + \delta_{kl}) \geq 1$),

(c) for every $l \in L$, there exist at least two and at most $R - 1$ *type-l* labels (that is for all $l \in L$, $M_l \geq 1$), for $l \in L_4$ there exist exactly two *type-l* labels in such a way that $M_l = M_l^l = 1$,

(d) if for some $l \in L_2$ there are exactly two *type-l* labels, then these two labels must occur in distinct $\Lambda's$ (that is, if $l \in L_2$ and $M_l = 1$, then $M_l^l = 0$).

(e) for every $l \in L$, the first index of $\Lambda^{(l)}$ is α_l.

We now reformulate (2.37) in such a way that we first sum up over the partitions $L_1 \sqcup L_2 \sqcup L_3 \sqcup L_4 = [p]$, the collection of multiplicities $M = (\, M_l^k \mid l \in L, k \in [p]\,)$ and the permutations of indices, and only then perform the actual summation over the labels from I. As the first three sums carry no N, they are irrelevant for the N-power counting. From (2.37) we find

$$\begin{aligned}\mathbf{E} \prod_{k \in [p]} \Lambda^{(k)}(D) = \mathbf{E} \sum_{\bigsqcup L_i = [p]} \sum_{M}^{\sim(L)} C_M \sum_{\sigma}^{\sim(M)} \left[\prod_{l \in L_3} \sum_{\alpha_l, \boldsymbol{\beta}_l \notin \mathcal{N}_{L_3}^{<l}}^{(M,l)} \frac{K(\alpha_l; \boldsymbol{\beta}_l) - \kappa(\alpha_l, \boldsymbol{\beta}_l)}{|\boldsymbol{\beta}_l|!} \right] \mathcal{M}' \\ + \mathcal{O}_{p,\mu}(N^{-p}),\end{aligned} \tag{2.39}$$

where

$$\mathcal{M}' := \left[\prod_{l \in L_4} \left(- \sum_{\alpha_l, \beta_l^l \in I} + \sum_{\alpha_l \in I \setminus \mathcal{N}_{L_3}^{<l}} \sum_{\beta_l^l \in \mathcal{N}(\alpha_l) \setminus \mathcal{N}_{L_3}^{<l}} \right) \frac{\kappa(\alpha_l, \beta_l^l)}{1!} \right] \mathcal{M},$$

$$\mathcal{M} := \left[\prod_{l \in L_2} \sum_{\alpha_l, \boldsymbol{\beta}_l \notin \mathcal{N}_{L_3}^{<l}}^{(M,l)} \frac{\kappa(\alpha_l, \boldsymbol{\beta}_l)}{|\boldsymbol{\beta}_l|!} \right] \left[\left(\prod_{k \in L} \Lambda^{(k)}_{\alpha_k, \sigma_k(\boldsymbol{\beta}^k)} \right) \left(\prod_{k \notin L} \Lambda^{(k)}_{\sigma_k(\boldsymbol{\beta}^k)} \right) \right] \Bigg|_{W_{\mathcal{N}_{L_3}} = 0},$$

and where $\sum_M^{\sim(L)}$ is the sum over all arrays M fulfilling (a)–(e) above and C_M are purely combinatorial constants bounded by a function of p, R; $C_M \leq C(p, R)$, in which we also absorbed the (-1)'s from the L_4 terms. Moreover, $\sum_{\sigma}^{\sim(M)}$ is the sum over all permutations σ_1,..., σ_p in the permutation groups $S_{M^1}, \dots, S_{M^p}$ (where $M^k := \sum_{l \in L} M_l^k$) such that for $k \notin L$ the first element of $\sigma_k(\boldsymbol{\beta}^k)$ is from $(\,\boldsymbol{\beta}_l^k \mid l \in L \cap [k])$. Furthermore, for any $\mathcal{N} \subset I$ we set

$$\sum_{\alpha_l, \boldsymbol{\beta}_l \notin \mathcal{N}}^{(M,l)} := \sum_{\alpha_l \in I \setminus \mathcal{N}} \prod_{k \in [p]} \sum_{\boldsymbol{\beta}_l^k \in (\mathcal{N}(\alpha_l) \setminus \mathcal{N})^{M_l^k}} .$$

Finally, we introduced the notations $\mathcal{N}_{L_3}^{<l} := \bigcup_{l > k \in L_3} \mathcal{N}(\alpha_k)$, and $\mathcal{N}_{L_3} := \bigcup_{k \in L_3} \mathcal{N}(\alpha_k)$. Here the $\boldsymbol{\beta}_l^k$ are actual (ordered) tuples and not multisets, which is why we denote them by boldfaced Greek letters to avoid possible confusion with the previously used $\underline{\beta}_l^k$. In (2.39) we furthermore used the short-hand notation $\boldsymbol{\beta}^k = (\boldsymbol{\beta}_l^k)_{l \in L}$ for the tuple (ordered according to the natural order on $L \subset [p] \subset \mathbb{N}$) of $\boldsymbol{\beta}_l^k$. We note that the artificial κ_S from (2.37) has been removed in (2.39) since we "pushed" the $|^{\rightarrow}$-operator all the way to the end. In the following we will establish bounds on (2.39) for fixed L and M and fixed permutations $\sigma_1, \dots, \sigma_p$. Since the number of possible choices for M, L and permutations is finite, depending on R and p only, this will be sufficient for bounding $\mathbf{E} \prod \Lambda^{(k)}(D)$. We also stress that the (multi)labels $\boldsymbol{\beta}_l^k$ themselves are not important, but only their type l.

Proof of the error bound in Proposition 2.4.4

We now turn to the proof of the claimed error bound (2.38). So far this was only done for the error from the first cumulant expansion in (2.36).

Proof of the error bound in Proposition 2.4.4. The error Ω in (2.37) is a sum over p terms, where the j-th term is the error from the expansion of $\Lambda^{(j)}(D)$. Recalling the definition of $\Omega(f, i, \mathcal{N})$ from (2.12b), this j-th expansion error is given by

$$\Omega_j := \sum_{\alpha_j} \Omega\left(\prod_{l<j} \Big(1 + \sum_{\alpha_l, \boldsymbol{\beta}_l}^{\sim(l)} + \sum_{\alpha_l, \boldsymbol{\beta}_l}^{*} \Big|_{W_{\mathcal{N}(\alpha_l)}=0}^{\rightarrow} + \sum_{\alpha_l, \beta_l^l}^{\#} \Big) \prod_{k=1}^{p} \widetilde{\Lambda}_k, \alpha_j, \mathcal{N}(\alpha_j) \right),$$

where

$$\widetilde{\Lambda}_k := \left\{ \begin{array}{ll} \Lambda^{(k)}_{\alpha_k, \bigsqcup_{l \in [p]} \underline{\beta}_l^k} & \text{if } k = j \text{ or } (k < j, \sum_{\alpha_k}) \\ \Lambda^{(k)}(\partial_{\bigsqcup_{l<k} \underline{\beta}_l^k} D) & \text{if } k > j \\ \Lambda^{(k)}_{\bigsqcup_{l<k} \underline{\beta}_l^k, \bigsqcup_{l>k} \underline{\beta}_l^k} & \text{else,} \end{array} \right\}$$

and where "if $(k < j, \sum_{\alpha_k})$" means "if $k < j$ and α_k is summed". This j-th error Ω_j can be estimated through (2.13) and Assumption (2.B) by the sum of

$$\begin{aligned} & \Bigg[\prod_{l \in L_2 \sqcup L_3} \sum_{\alpha_l, \boldsymbol{\beta}_l}^{(M,l)} \Bigg] \Bigg[\prod_{l \in L_4} \sum_{\alpha_l, \beta_l^l \in I} \Bigg] \sum_{\alpha_j} \sum_{\boldsymbol{\beta}_j \in \mathcal{N}(\alpha_j)^R} \\ & \times \left\| \Big(\prod_{k \in L} \Lambda^{(k)}_{\alpha_k, \sigma_k(\boldsymbol{\beta}^k)} \Big) \Big(\prod_{k \in [j] \setminus L} \Lambda^{(k)}_{\sigma_k(\boldsymbol{\beta}^k)} \Big) \Big(\prod_{k>j} \Lambda^{(k)}(\partial_{\sigma_k(\boldsymbol{\beta}^k)} D) \Big|_{\widehat{W}} \Big) \right\|_2 , \end{aligned} \tag{2.40}$$

over partitions $L = L_2 \sqcup L_3 \sqcup L_4 \subset [j-1]$, arrays M fulfilling (a)–(e) above and partitions σ_k. In all terms $\widehat{W}$ is a modification of W which differs from W in at most $C\sqrt{N}$ entries.

The previously studied error from (2.34) for example corresponds to $j = 1, L_2 = L_3 = L_4 = \emptyset$. The combinatorics of all these summations are independent of N, hence can be neglected. So we can focus on a single term of the form (2.40). The norm in (2.40) will first be estimated by Hölder and then by (2.35) to reduce it to many factor of $\|G\|_q$. We now have to count the size of the sums, the number of $N^{-1/2}$ factors from the derivatives, and the number of $\|G\|_q$'s we collect in the bound. We start with the sums which are at most of size

$$\begin{aligned} &N^{2|L_2\sqcup L_3|}(N^{1/2-\mu})^{M_{L_2\sqcup L_3}}(N^2\cdot N^2)^{|L_4|}N^2(N^{1/2-\mu})^R \\ &\quad = N^{2|L_2\sqcup L_3|+(M_{L_2\sqcup L_3}+R)(1/2-\mu)+4|L_4|+2}. \end{aligned} \tag{2.41}$$

Here the first factor comes from the α_l summations for $l \in L_2 \sqcup L_3$, while the second term comes from the corresponding $\boldsymbol{\beta}_l$ summations. The third factor comes from the α_l, β_l^l-summations for $l \in L_4$, and finally the fifth and sixth factor correspond to the α_j and $\boldsymbol{\beta}_j$ summations. Next, we count the total number of derivatives. Every index α_l and β_l^k accounts for a derivative, and each derivative contributes a factor of $N^{-1/2}$. So we have

$$(N^{-1/2})^{|L_2\sqcup L_3|+M_{L_2\sqcup L_3}+2|L_4|+(R+1)} = N^{-|L_2\sqcup L_3|/2-M_{L_2\sqcup L_3}/2-|L_4|-(R+1)/2}, \tag{2.42}$$

so that altogether from (2.41) and (2.42) we have an N-power of

$$N^{3/2(|L_2\sqcup L_3|+1)+3|L_4|-R\mu}N^{-\mu M_{L_2\sqcup L_3}} \le N^{-p}N^{-\mu M_{L_2\sqcup L_3}}.$$

It remains to count the number of $\|G\|_{CRp^2} = \|G\|_q$ coming from the application of (2.35), which in total provides

$$\begin{aligned} &\sum_{k\in L}(1+|\boldsymbol{\beta}^k|+5) + \sum_{k\in[j]\setminus L}(|\boldsymbol{\beta}^k|+5) + \sum_{k>j}(|\boldsymbol{\beta}^k|+5) \\ &\qquad = 5p + |L_2\sqcup L_3| + M_{L_2\sqcup L_3} + 2\,|L_4| + R + 1 \le Cp/\mu + M_{L_2\sqcup L_3} \end{aligned}$$

factors of $\|G\|_q$. The claim (2.38) now follows from the trivial estimate $M_{L_2\sqcup L_3} \le Rp \le Cp^2/\mu$. □

Subsequently we establish a bound on the rhs. of (2.39), by first estimating it in terms of $\|\mathcal{M}'\|_p$, then estimating $\|\mathcal{M}'\|_p$ in terms of $\|\mathcal{M}\|_p$ and finally bounding the leading contribution $\mathcal{M}$. We consider the first two steps in this procedure as errors stemming from the neighbourhood structure of the expansion, while the third step is concerned with the leading order contribution from the expansions. In Section 2.4.2 we consider the errors stemming from the neighbourhood structure, while in Sections 2.4.3 and 2.4.4 we derive bounds on $\|\mathcal{M}\|_p$ for the averaged and isotropic case, separately. For simplicity we first carry out the technically most involved argument from Sections 2.4.3–2.4.4 in the extreme case $L_3 = L_4 = \emptyset$ where the neighbourhood errors are absent. Finally, we explain the necessary modifications for the general case in Section 2.4.5.

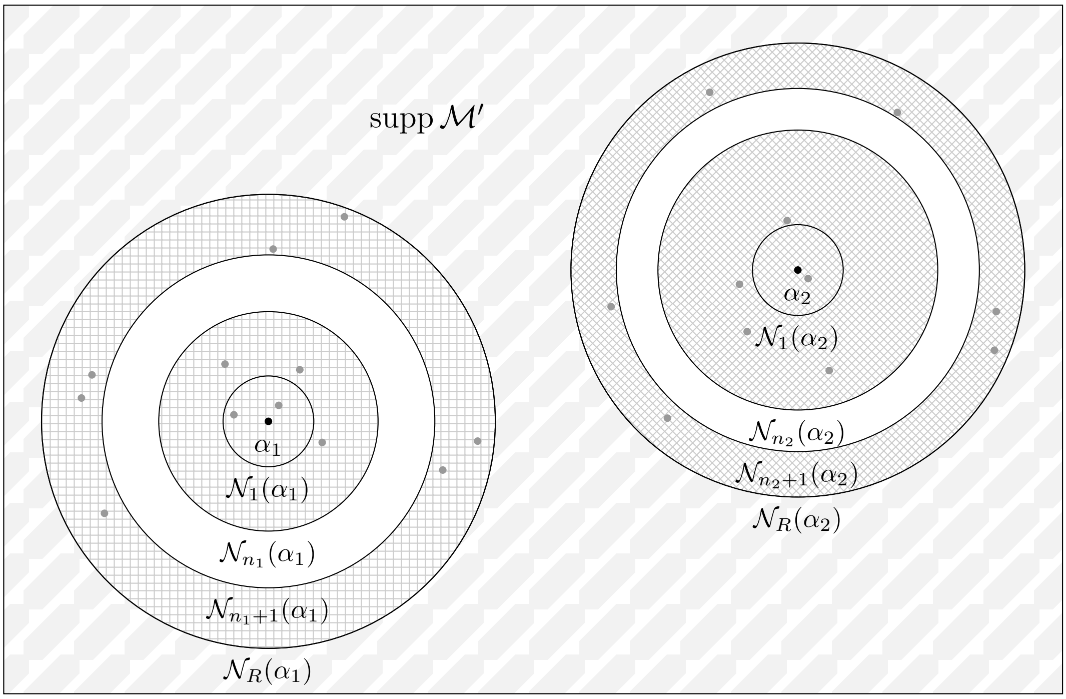

supp $\mathcal{M}'$
α_1
$\mathcal{N}_1(\alpha_1)$
$\mathcal{N}_{n_1}(\alpha_1)$
$\mathcal{N}_{n_1+1}(\alpha_1)$
$\mathcal{N}_R(\alpha_1)$
α_2
$\mathcal{N}_1(\alpha_2)$
$\mathcal{N}_{n_2}(\alpha_2)$
$\mathcal{N}_{n_2+1}(\alpha_2)$
$\mathcal{N}_R(\alpha_2)$

FIGURE 2.1: Illustration for the bound on $\mathcal{E}$ in (2.43). Gray dots ● denote the β_1, β_2 labels. Since there are $|\beta_i| < R$ labels and R rings, there is always one empty ring by the pigeonhole principle.

2.4.2 Bound on neighbourhood errors

We start with the bound on the L_3-factors in (2.39). Neglecting the irrelevant combinatorial factors $|\boldsymbol{\beta}_l|!$ and the summations over L_i, M and σ, we have to estimate

$$\begin{aligned}\mathcal{E} &:= \left[\prod_{l\in L_3} \sum_{\alpha_l,\boldsymbol{\beta}_l\notin\mathcal{N}^{<l}_{L_3}}^{(M,l)}\right]\mathcal{E}(\alpha_{L_3},\boldsymbol{\beta}_{L_3}) \\ &:= \left[\prod_{l\in L_3} \sum_{\alpha_l,\boldsymbol{\beta}_l\notin\mathcal{N}^{<l}_{L_3}}^{(M,l)}\right]\mathbf{E}\,\mathcal{M}'\prod_{l\in L_3}\left[K(\alpha_l;\boldsymbol{\beta}_l)-\kappa(\alpha_l,\boldsymbol{\beta}_l)\right].\end{aligned} \tag{2.43}$$

By the pigeon hole-principle we find that for every $l \in L_3$ and any assignment of $\alpha_l, \boldsymbol{\beta}_l$ there exist some $n_l < R$ such that we have a partition $\underline{\beta}_l = \underline{\beta}_l^{(i)} \sqcup \underline{\beta}_l^{(o)}$ into *inside* and *outside* elements with $\underline{\beta}_l^{(i)} \subset \mathcal{N}_{n_l}(\alpha_l)$ and $\underline{\beta}_l^{(o)} \subset \mathcal{N}_{n_l+1}(\alpha_l)^c$ since $|\underline{\beta}_l| = M_l < R$ (see rule (c)). We recall the nested structure of the neighbourhoods as stated in Assumption (2.D), and provide an illustration of the "security layers" in Figure 2.1. According to (2.8c) we can then write (L_3' collects those indices where we took the middle term of (2.8c) in the l factor)

$$\mathcal{E}(\alpha_{L_3},\boldsymbol{\beta}_{L_3}) = \sum_{L_3=L_3'\sqcup L_3''} (-1)^{\left|L_3''\right|} \prod_{l\in L_3''}\Bigg[\sum_{\gamma_l^{(i)}\subset\beta_l^{(i)}}\ \sum_{\gamma_l^{(o)}\subsetneq\beta_l^{(o)}} \kappa(\alpha_l, \underline{\beta}_l^{(i)}\setminus\underline{\gamma}_l^{(i)}, \underline{\beta}_l^{(o)}\setminus\underline{\gamma}_l^{(o)})\Bigg] \\ \times \mathbf{E}\, f \prod_{l\in L_3'} [K(\alpha;\underline{\beta}_l^{(i)}) - \kappa(\alpha,\underline{\beta}_l^{(i)})],$$

where

$$f := \mathcal{M}' \prod_{l\in L_3'} (\Pi\underline{\beta}_l^{(o)}) \prod_{l\in L_3''}\left[(\Pi\underline{\gamma}_l^{(i)})(\Pi\underline{\gamma}_l^{(o)})\right]$$

is a random variable supported in $\bigcap_{l\in L_3'}\mathcal{N}_{n_l+1}(\alpha_l)^c$, i.e., well separated from the variables $K(\alpha_l;\underline{\beta}_l^{(i)})$ for $l\in L_3'$. It remains to estimate a quantity of the type $\mathbf{E}\, fg_1\dots g_k$, where $f,g_1,\dots,g_k$ are random variables whose supports are pairwise separated by "security layers" and where each g_i is of the form $K-\kappa$ with $\mathbf{E}\, g_i = 0$. Here $k=|L_3'|$ and from Lemma 2.3.3 and Assumption (2.D) it follows that $\mathbf{E}\, fg_1\dots g_k \le_k \|f\|_{k+1} N^{-3\lceil k/2\rceil}$. According to Lemma 2.A.1 the $\kappa(\alpha_l,\underline{\beta}_l^{(i)}\setminus\underline{\gamma}_l^{(i)},\underline{\beta}_l^{(o)}\setminus\underline{\gamma}_l^{(o)})$ factors are also at least N^{-3} small and we can conclude that

$$|\mathcal{E}(\alpha_{L_3},\boldsymbol{\beta}_{L_3})| \le_{p,R} N^{-3\lceil |L_3|/2\rceil} \left\|\mathcal{M}'\right\|_p. \tag{2.44}$$

Next, we use the triangle inequality to pull the L_4 summation out of $\left\|\mathcal{M}'\right\|_p$ to achieve a bound in terms of $\left\|\mathcal{M}\right\|_p$. We have

$$\begin{aligned} &\left|\Bigg(-\sum_{\alpha_l,\beta_l^l\in I} + \sum_{\alpha_l\in I\setminus\mathcal{N}_{L_3}^{<l}}\ \sum_{\beta_l^l\in\mathcal{N}(\alpha_l)\setminus\mathcal{N}_{L_3}^{<l}}\Bigg)\kappa(\alpha_l,\beta_l^l)\right| \\ &\quad\le \Bigg(\sum_{\alpha_l\in I\setminus\mathcal{N}_{L_3}^{<l}}\ \sum_{\beta_l^l\in\mathcal{N}_{L_3}^{<l}} + \sum_{\alpha_l\in I\setminus\mathcal{N}_{L_3}^{<l}}\ \sum_{\beta_l^l\in I\setminus\mathcal{N}(\alpha_l)} + \sum_{\alpha_l\in\mathcal{N}_{L_3}^{<l}}\ \sum_{\beta_l^l\in I}\Bigg)|\kappa(\alpha_l,\beta_l^l)| \\ &\quad\le \Bigg(\sum_{\beta_l^l\in\mathcal{N}_{L_3}^{<l}}\ \sum_{\alpha_l\in\mathcal{N}(\beta_l^l)} + \sum_{\beta_l^l\in\mathcal{N}_{L_3}^{<l}}\ \sum_{\alpha_l\in I\setminus\mathcal{N}(\beta_l^l)} + \sum_{\alpha_l\in I}\ \sum_{\beta_l^l\in I\setminus\mathcal{N}(\alpha_l)} \\ &\qquad\qquad + \sum_{\alpha_l\in\mathcal{N}_{L_3}^{<l}}\ \sum_{\beta_l^l\in\mathcal{N}(\alpha_l)} + \sum_{\alpha_l\in\mathcal{N}_{L_3}^{<l}}\ \sum_{\beta_l^l\in I\setminus\mathcal{N}(\alpha_l)}\Bigg)|\kappa(\alpha_l,\beta_l^l)| \le CN, \end{aligned}$$

where we estimated the first and the fourth term with two small summations purely by size $(CN^{1/2-\mu})^2 \le CN$ and the other terms using the fact that $|\kappa(\alpha,\beta)| \lesssim N^{-3}$ for $\beta\in I\setminus\mathcal{N}(\alpha)$. Summarizing, we thus have that

$$\begin{aligned} &\left|\mathbf{E}\prod\Lambda^{(k)}(D)\right| \le_{p,\mu} N^{-p} \\ &\quad + \sum_{\bigsqcup L_i=[p]}\ \sideset{}{^{\sim(L)}}\sum_{M} \frac{N^{|L_4|}}{N^{3\lceil |L_3|/2\rceil}} \sideset{}{^{\sim(M)}}\sum_{\sigma} \Bigg[\prod_{l\in L_3}\ \sideset{}{^{(M,l)}}\sum_{\alpha_l,\boldsymbol{\beta}_l\notin\mathcal{N}_{L_3}^{<l}}\Bigg]\Bigg[\prod_{l\in L_4}\max_{\alpha_l,\beta_l^l\in I}\Bigg] \left\|\mathcal{M}\right\|_p, \end{aligned} \tag{2.45}$$

and it only remains to estimate the leading order term $\mathcal{M}$, as defined in (2.39). This has to be done separately for averaged and isotropic bound and should be considered as the main part of the proof. To simplify notations we will first prove the bound on $\mathcal{M}$ for the case that $L_3 = L_4 = \emptyset$ and $\mathcal{N}(\alpha) = I$. In particular $L_3 = \emptyset$ implies that $\mathcal{N}_{L_3} = \emptyset$ and therefore in the next two Sections 2.4.3 and 2.4.4 we now aim at deriving a bound on $\|\mathcal{M}((\Lambda^{(k)})_{k\in[p]}; L, M, \sigma)\|_p$, where

$$\mathcal{M}((\Lambda^{(k)})_{k\in[p]}; L, M, \sigma) := \left[\prod_{l\in L} \sum_{\alpha_l,\boldsymbol{\beta}_l}^{(M,l)} \frac{\kappa(\alpha_l,\boldsymbol{\beta}_l)}{|\boldsymbol{\beta}_l|!}\right] \left(\prod_{k\in L} \Lambda^{(k)}_{\alpha_k,\sigma_k(\boldsymbol{\beta}^k)}\right) \left(\prod_{k\notin L} \Lambda^{(k)}_{\sigma_k(\boldsymbol{\beta}^k)}\right),$$
$$\sum_{\alpha_l,\boldsymbol{\beta}_l}^{(M,l)} := \sum_{\alpha_l\in I} \prod_{k\in[p]} \sum_{\boldsymbol{\beta}_l^k\in I^{M_l^k}} . \qquad (2.46)$$

The definition of $\mathcal{M}$ in (2.46) agrees with the one in (2.39) in the special case $L_3 = L_4 = \emptyset$, except for a tiny contribution from $\boldsymbol{\beta}_l \not\subset \mathcal{N}(\alpha_l)$. The reason for extending the sum here to the whole index set is twofold: First, we do not have to keep track of the summation ranges of individual indices, and, second, we demonstrate that for the main terms separating the contribution outside of the neighbourhoods $\mathcal{N}$ is not necessary, all estimates on $\mathcal{M}$ would also hold for the unrestricted sum. In particular, the neighbourhood decay condition is not necessary for the main terms, they are used only for bounding $\mathcal{M}'$ in terms of $\mathcal{M}$ in Section 2.4.2. This fact was already advertised in Example 2.2.12 where we claimed that in the Gaussian case we can considerably relax our decay conditions. Later, in Section 2.4.5 we will explain how to elevate the proof for the special case $L_3 = L_4 = \emptyset$ with extended index sets to the general case.

2.4.3 Averaged bound on D

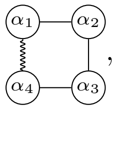
To treat (2.46) systematically, we introduce a graphical representation for any M, L and permutations σ in (2.46). For the averaged local law we need averaged estimates on D, so we set

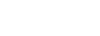
$$\Lambda^{(k)}(D) := \langle BD\rangle \qquad \text{or} \qquad \Lambda^{(k)}(D) := \overline{\langle BD\rangle},$$

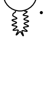
where B is a generic norm-bounded matrix, $\|B\| \lesssim 1$ and we recall that $\langle\cdot\rangle = N^{-1}\operatorname{Tr}$ denotes the normalized trace. A factor $\Lambda_{\alpha_1,\dots,\alpha_n}$ can be represented as a directed cyclic graph on the vertex set $\{\alpha_1,\dots,\alpha_n\}$. Up to sign we have

$$\begin{aligned} |\Lambda_{\alpha_1,\dots,\alpha_n}| &= N^{-n/2}\langle B\Delta^{\alpha_1}G\Delta^{\alpha_2}G\dots\Delta^{\alpha_n}G\rangle \\ &= N^{-1-n/2}G_{b_1a_2}G_{b_2a_3}\dots G_{b_{n-1}a_n}(GB)_{b_na_1}, \end{aligned} \qquad (2.47)$$

which we represent as a cyclic graph in such a way that the vertices represent labels $\alpha_i = (a_i, b_i)$ and a directed edge from $\alpha_i = (a_i, b_i)$ to $\alpha_j = (a_j, b_j)$ represents $G_{b_ia_j}$. Since we will always draw the graphs in a clockwise orientation we will not indicate the direction of the edges specifically. The specific GB factor will be denoted by a wiggly line instead of a straight line used for the G factors. As an example, we have the correspondences

$\Lambda_{\alpha_1,\alpha_2,\alpha_3,\alpha_4} \leftrightarrow$ 

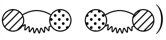
, $\Lambda_{\alpha_1,\alpha_2} \leftrightarrow$ 

and $\Lambda_{\alpha_1} \leftrightarrow$ 

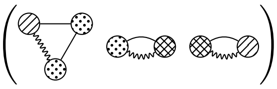
.

In (2.46) the labels of type l are connected through the $\kappa(\alpha_l, \boldsymbol{\beta}_l)$ factor which strongly links those labels due to the decay properties of the cumulants. We represent this fact graphically as a vertex colouring of the graph in which label types correspond to colours. The set of colours representing the label types L will be denoted by C. The $M_l + 1$ vertices of a given type l will be denoted by V_c, where c is the colour corresponding to l.

We define $\mathrm{Val}(\Gamma)$, the *value* of a graph Γ, as summation over all labels consistent with the colouring, such that equally coloured labels are linked through a cumulant, of the product of the corresponding Λ's, just as in (2.46). For example, we have

$$\sum_{\alpha_1,\beta_1^2(1)} \kappa(\alpha_1, \beta_1^2(1)) \sum_{\alpha_2,\beta_2^1(1)} \kappa(\alpha_2, \beta_2^1(1))\Lambda_{\alpha_1,\beta_2^1(1)}\Lambda_{\alpha_2,\beta_1^2(1)} = \mathrm{Val}\,($$

$$) \qquad (2.48)$$

or

$$\sum_{\alpha_1,\beta_1^2(1)} \kappa(\alpha_1, \beta_1^2(1)) \sum_{\alpha_2,\beta_2^1(1),\beta_2^1(2)} \frac{\kappa(\alpha_2, \beta_2^1(1), \beta_2^1(2))}{2!}$$
$$\times \sum_{\alpha_3,\beta_3^2(1)} \kappa(\alpha_3, \beta_3^2(1))\Lambda_{\alpha_1,\beta_2^1(2),\beta_2^1(1)}\Lambda_{\alpha_2,\beta_1^2(1)}\Lambda_{\alpha_3,\beta_1^3(1)} = \mathrm{Val}\left(\right.$$

$$\left.\right),$$

where we choose the variable names for the labels in accordance with (2.46) following the convention that the elements of the tuple $\boldsymbol{\beta}_l^k$ are denoted by $(\beta_l^k(1), \beta_l^k(2), \dots)$. We warn the reader that $\mathrm{Val}(\Gamma)$, the value of a diagram itself is a random variable unlike in customary Feynman diagrammatic expansion theory. In the following we will derive bounds on the value of diagrams. To separate the conceptual from the technical difficulties we first derive those bounds in a vague $\lesssim$ sense which ignores a technical subtlety: The entries G_{ab} of the resolvent are bounded with overwhelming probability, but usually not almost surely. In the first conceptual step we will tacitly assume such an almost sure bound and write $|G_{ab}| \lesssim 1$. Later in Section 2.4.3.1 we will make the bounds rigorous in a high-moment sense. We note that if $\Lambda(D) = \overline{\langle BD \rangle}$, then the edges would represent G^* and $(GB)^*$ instead of G and GB and the order would be reversed (recall that the double indices are transposed in (2.27)) but the counting argument is not sensitive to these nuances, so we omit these distinctions in our graphs.

We now rephrase the rules on M in this graphical representation. They dictate that we need to consider the set of all vertex coloured graphs Γ with cyclic components such that

(a) there exist p connected components, all of which are cycles,

(b) each connected component contains at least one vertex,

(c) each colour colours at least two vertices,

(d) if a colour colours exactly two vertices, then these vertices are in different components.

(e) for each colour there exists a component in which the vertex after the wiggled edge (in clockwise orientation) is of that colour.

We note that these rules, compared to (2.46), disregarded the restrictions on the permutations σ_k for $k \notin L$ as these are not relevant for the averaged bound. The set of graphs satisfying (a)–(e) will be denoted by $\mathcal{G}^{\mathrm{av}}(p,R)$ and for each L, M, σ the main term $\mathcal{M}$ from (2.46) is given by the value of some graph $\Gamma \in \mathcal{G}^{\mathrm{av}(p,R)}$.

$$\mathcal{M}\Big(\big(\langle B\cdot\rangle^{[p/2]}, \overline{\langle B\cdot\rangle}^{[p/2]}\big); L, M, \sigma\Big) = \mathrm{Val}(\Gamma), \qquad \Gamma = \Gamma(L,M,\sigma) \in \mathcal{G}^{\mathrm{av}(p,R)} \quad (2.49)$$

where $\langle B\cdot\rangle^{[p/2]}$ denotes the tuple of $p/2$ functionals mapping $D \mapsto \langle BD\rangle$ and similarly for $\overline{\langle B\cdot\rangle}$. As the number of such graphs is finite for given p, R it follows that it is sufficient to prove the required bound for every single graph.

As for any fixed colour $\sum_{\oslash} \le N^2 \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \kappa \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert^{\mathrm{av}}$, the naive size of the value $\mathrm{Val}(\Gamma)$ is bounded by

$$\mathrm{Val}(\Gamma) \lesssim N^{-p} \prod_{c\in C} N^{2-|V_c|/2} \le 1 \quad (2.50)$$

since according to (2.47) every component contributes a factor N^{-1} and every label contributes a factor $N^{-1/2}$, and where the ultimate inequality followed from $|V_c| \ge 2$ and $|C| \le p$. We now demonstrate that using Ward identities of the form

$$\sum_a |G_{ab}|^2 = \frac{(\Im G)_{bb}}{\eta}$$

we can improve upon this naive size by a factor of ψ^{2p}, where $\psi \approx 1/\sqrt{N\eta}$ and $\eta := \Im z$. We will often use the Ward identity in the form

$$\sum_b |G_{ab}| \le \sqrt{N}\sqrt{\sum_b |G_{ab}|^2} = N\sqrt{\frac{(\Im G)_{aa}}{N\eta}} \lesssim N\psi, \quad \sum_b |(GB)_{ab}| \lesssim \|B\|\, N\psi \quad (2.51a)$$

which explicitly exhibits a gain of a factor ψ over the trivial bound of order N. Together with the previous bound

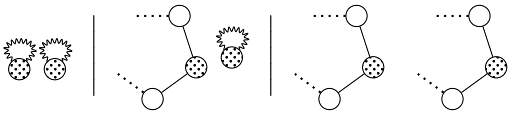
$$\sum_b |G_{ab}|^2 \le N\psi^2, \qquad \sum_b |(GB)_{ab}|^2 \lesssim \|B\|^2 N\psi^2 \quad (2.51b)$$

we will call (2.51a)–(2.51b) *Ward estimates*. Here we used the trivial bound $|G| \lesssim 1$ and we set $\psi := \sqrt{\Im G/N\eta}$ (where $\Im G$ is meant in an isotropic sense which we will define rigorously later).

We consider the subset of colours $C' := \{\, c \in C \mid |V_c| \le 3 \,\} \subset C$ which colour either two or three vertices and we intend to use Ward identities only when summing up vertices with those colours. However, one may not use Ward estimates for every such summation, e.g. even if both a and b were indices of eligible labels, one cannot gain from both of them in the sum $\sum_{a,b} |G_{ab}|$. We thus need a systematic procedure to identify sufficiently many labels so that each summation over them can be performed by using Ward estimates. In the following, we first describe a procedure how to *mark* those edges we can potentially use for Ward estimates. Secondly, we will show that for sufficiently many marked edges the Ward estimates can be used in parallel.

Procedure for colours appearing twice in Γ

If a colour ⊛ appears twice, then it appears in two different components of Γ, i.e., in one of the following forms



where the white vertices can be of any colour other than ⊛ (and may even coincide), the dotted edges indicate an arbitrary continuation of the component and some additional edges may be wiggled. The picture only shows those two components with colour ⊛, the other components of Γ are not drawn. Vertical lines separate different cases. When summing up the ⊛-coloured labels, we can use the Ward estimates on all edges adjacent to ⊛ using the operator norm $|||\kappa|||_2^{\text{av}} = \| \, |\kappa(*,*)| \, \|$ on κ. To see this we note that

$$\sum_{\alpha_1,\alpha_2} |\kappa(\alpha_1,\alpha_2) A_{\alpha_1} B_{\alpha_2}| \leq |||\kappa|||_2^{\text{av}} \sqrt{\sum_{\alpha_1} |A_{\alpha_1}|^2} \sqrt{\sum_{\alpha_2} |B_{\alpha_2}|^2}, \tag{2.52}$$

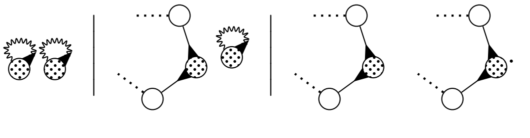
after which (2.51b) with

$$A_{\alpha_1}, B_{\alpha_1} \in \{ \, G_{b_1 a_1}, (GB)_{b_1 a_1}, G_{c a_1} G_{b_1 d}, (GB)_{c a_1} G_{b_1 d}, G_{c a_1} (GB)_{b_1 d} \, \}$$

and arbitrary fixed indices c, d is applicable.

Remark 2.4.5. *In the sequel we will not write up separate estimates for edges representing GB instead of G as the same Ward estimates* (2.51a)–(2.51b) *hold true and the bound is automatic in the sense that there are in total p wiggly edges in Γ, each of which will contribute a factor of $\|B\|$ to the final estimate, regardless of whether the corresponding edge has been bounded trivially $|(GB)_\alpha| \lesssim \|B\|$ or by* (2.51a)–(2.51b).

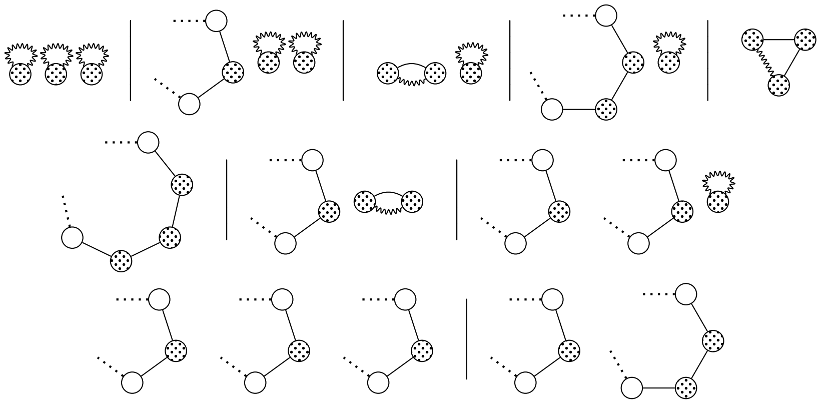
We find that for every edge connected to ⊛ we can gain a factor ψ compared to the naive size of the ⊛-sum, using only the trivial bound $|G| \lesssim 1$. We will indicate visually that an edge has potential for a gain of ψ through some colour by putting a mark (a small arrow) pointing from the vertex towards the edge. Thus in the case where ⊛ appears twice we mark all edges adjacent to ⊛ to obtain the following marked graphs



We note that these marks indicate that we can use a Ward estimate for every marked edge, when performing the ⊛-summation, while keeping all other labels fixed. When simultaneously summing over labels from different colours it is not guaranteed any more that we can perform a Ward estimate for every marked edge. We will later resolve this possible issue by introducing the concept of *effective* and *ineffective* marks.

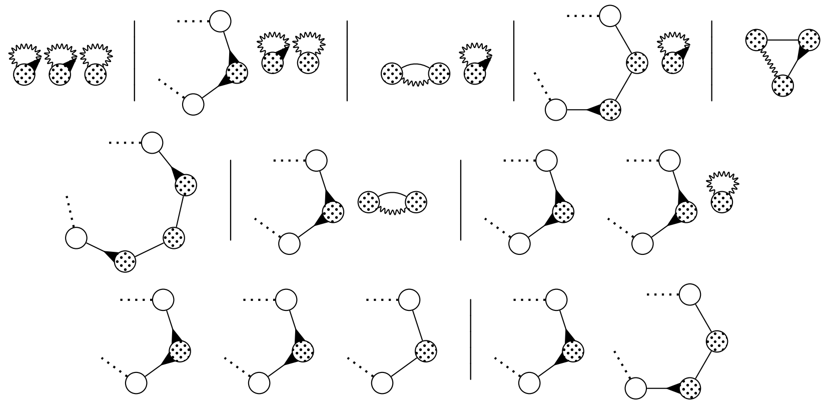
Procedure for colours appearing three times in Γ

If a colour ⊛ appears three times, then the following ten setups are possible



(2.53)

where we explicitly allow components with open continuations to be connected (unlike in the previous case, where rule (d) applied). We now mark the edges adjacent to ⊛ as follows and observe that at most two remain unmarked. Explicitly, we choose the markings



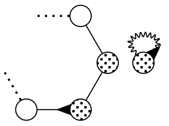
and observe that in all but the fifth graph we can gain a factor of ψ for every marked edge using the first term in the norm $|||\kappa|||_3^{\mathrm{av}}$. For example, in the second graph this follows from

$$\sum_{\alpha_1,\alpha_2,\alpha_3} |\kappa(\alpha_1,\alpha_2,\alpha_3)G_{ca_3}G_{b_3d}G_{b_1a_1}G_{b_2a_2}| \lesssim |||\kappa|||_3^{\mathrm{av}} \sqrt{\sum_{\alpha_2}|G_{b_2a_2}|^2}\sqrt{\sum_{\alpha_3}|G_{ca_3}G_{b_3d}|^2}$$
$$\lesssim |||\kappa|||_3^{\mathrm{av}}\, N^2\psi^3$$

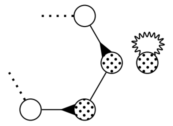
and in third graph from

$$\sum_{\alpha_1,\alpha_2,\alpha_3} |\kappa(\alpha_1,\alpha_2,\alpha_3)G_{b_1a_2}G_{b_2a_1}G_{b_3a_3}| \lesssim \sum_{\alpha_2,\alpha_3}|G_{b_3a_3}|\sum_{\alpha_1}|\kappa(\alpha_1,\alpha_2,\alpha_3)| \lesssim |||\kappa|||_3^{\mathrm{av}}\, N^2\psi,$$

where c and d are the connected indices from the white vertices in the graph. The computations for the other graphs are identical. We note that the markings we chose above are not the only ones possible. For example we could have replaced



by



. (2.54)

For the fifth graph in (2.53) the second term in the $|||\kappa|||_3^{\text{av}}$ is necessary. The norms in (2.6c) ensure that we can perform at least one Ward estimate and we have

$$\sum_{\alpha_1,\alpha_2,\alpha_3} |\kappa(\alpha_1,\alpha_2,\alpha_3) G_{b_1 a_2} G_{b_2 a_3} G_{b_3 a_1}| \lesssim |||\kappa|||_3^{\text{av}} N^2 \psi.$$

Indeed, for example

$$\begin{aligned}\sum_{\alpha_1,\alpha_2,\alpha_3} |\kappa_{cd}(\alpha_1,\alpha_2,\alpha_3) G_{b_1 a_2} G_{b_2 a_3} G_{b_3 a_1}| &\lesssim \sum_{\alpha_1,\alpha_2,\alpha_3} |\kappa_{cd}(\alpha_1,\alpha_2,\alpha_3) G_{b_3 a_1}| \\ &\le |||\kappa_{cd}|||_{cd} N \sqrt{\sum_{b_3,a_1} |G_{b_3 a_1}|^2} \lesssim |||\kappa_{cd}|||_{cd} N^2 \psi\end{aligned}$$

and the other three cases are similar.

Procedure for all other colours in Γ

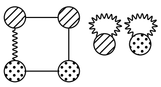
For colours in $C \setminus C'$, i.e., those which appear four times or more, we do not intend to use any Ward estimates and therefore we do not place any additional markings. Thus we only have to control the size of the summation over any fixed colour, as is guaranteed by the finiteness of $|||\kappa|||_k^{\text{av}}$.

Counting of markings

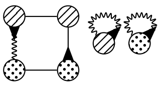
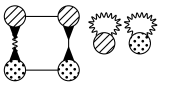
After we have chosen all markings, we select the "useful" ones. We call an edge *ineffectively marked* if it only carries one mark and joins two distinctly C'-coloured vertices. All other marked edges we call *effectively marked* because the parallel gain through a Ward estimate is guaranteed for all those edges. In total, there are at least $\sum_{c\in C'} |V_c|$ edges adjacent to C' (i.e., adjacent to a C'-coloured vertex). After the above marking procedure there are at most $2\sum_{c\in C'}(|V_c|-2)$ unmarked or ineffectively marked edges adjacent to C'. To see this we note that edges between two C'-colours with only one marking are counted as unmarked from the perspective of exactly one of the two colours. Thus we find that there are at least

$$\sum_{c\in C'} |V_c| - 2\sum_{c\in C'}(|V_c|-2) = \sum_{c\in C'}(4-|V_c|) \tag{2.55}$$

effectively marked edges adjacent to C' after the marking procedure. We illustrate this counting in an example. In the graph



we have $V_{\oslash} = V_{\odot} = 3$ and there are six edges adjacent to $C' = \{\oslash, \odot\}$. After the marking procedure we could for example obtain the graphs

 or  ,

where the second graph would result from the replaced marking in (2.54). In both cases there are two effectively marked edges, in accordance with (2.55); in the first example there are also two ineffectively marked edges.

Power counting estimate

The strategy now is that we iteratively perform the Ward estimates colour by colour in C' in no particular order. In each step we thus remove all the edges adjacent to some given colour, either through Ward estimates (if the edge was marked in that colour), or through the trivial bound $|G_\alpha| \lesssim 1$. If some edge is missing because it already was removed in a previous step, then the corresponding G is replaced by 1 in that estimate (e.g. in (2.52)). This might reduce the number of available Ward estimates in some steps, but the concept of effective markings ensures that whenever an effectively marked edge is removed, then a gain through a Ward estimate is guaranteed. After the summation over all colours from C' we have thus performed Ward estimates in all the effectively marked edges, which amounts to at least

$$\sum_{c\in C'} (4 - |V_c|)$$

gains of the factor ψ. We note that ineffectively marked edges may not be estimated by a Ward estimates, as it might be necessary to bound the corresponding G trivially while performing the sum over another colour. Using only the gains from the effective marks, we can improve on the naive power counting (2.50) to conclude that the value of Γ is bounded by

$$\mathrm{Val}(\Gamma) \lesssim N^{-p} \prod_{c\in C\setminus C'} N^{2-|V_c|/2} \prod_{c\in C'} (N\psi^2)^{2-|V_c|/2} \leq \psi^{2p} N^{2|C\setminus C'|-|V_{C\setminus C'}|/2}, \quad (2.56)$$

where we used that $|C'| \leq |C| \leq p$, $|V_c| = 2, 3$ for $c \in C'$ and $|V_c| \geq 4$ otherwise, and that $N\psi^2 \geq 1$.

2.4.3.1 Detailed bound

The argument above tacitly assumed bounds of the form $|G_\alpha| \lesssim 1$ and $\sum_\alpha |G_\alpha|^2 \lesssim N^2\psi^2$. Apart from unspecified and irrelevant constants, these bounds are not available almost surely, they hold only in the sense of high moments, e.g. $\mathbf{E}\,|G_\alpha|^q \leq_q 1$. Secondly, the definition of ψ intentionally left the role of $\Im G$ in it vague. The precise definition of ψ will involve high L^q norms of $\Im G$. Moreover, different G-factors in the monomials Λ are not independent. All these difficulties can be handled by the following general Hölder inequality. Suppose, we aim at estimating

$$\mathbf{E} \sum_A X_A \sum_B Y_{A,B}$$

for random variables $X_A, Y_{A,B}$, then we use the Hölder inequality to estimate

$$\Big\| \sum_A X_A \sum_B Y_{A,B} \Big\|_q \leq \Big(\sum_A \Big)^{\epsilon} \Big\| \sum_A X_A \Big\|_{2q} \max_A \Big\| \sum_B Y_{A,B} \Big\|_{1/\epsilon} \quad (2.57)$$

for $0 < \epsilon \leq 1/2q$. In our procedure (2.57) enables us to iteratively bound the graphs colour by colour at the expense of an additional factor $N^{2pR\epsilon}$ in every colour step of the bound, as the total sum is at most of size N^{2pR}. To estimate a G or an $\Im G$ directly we use the Hölder inequality and note that there are at most $|V| = \sum_c |V_c| \leq pR$ factors of the form G or GB, so that we can estimate those terms isotropically by $\|G\|_{pR/\epsilon}$, $\|B\|\,\|G\|_{pR/\epsilon}$ and $\|\Im G\|_{pR/\epsilon}$. We use (2.57) at most with $q \in \{1, 2, 4, \dots, 2^{p-1}\}$ and thus have a restriction of $0 < \epsilon \le 2^{-p}$. Thus, combining the power counting above with the iterated application of the Hölder inequality, we have shown that

$$\|\mathrm{Val}(\Gamma)\|_p \le_{p,R,\epsilon} N^{2p^2R\epsilon}\Big(1+|||\kappa|||^{\mathrm{av}}\Big)^p \|B\|^p\,(1+\|G\|_{pR/\epsilon}^{|V|})\psi_{pR/\epsilon}^{2p} N^{2|C\setminus C'|-|V_{C\setminus C'}|/2},$$

where $\psi_q := \sqrt{\frac{\|\Im G\|_q}{N\eta}}$, for all $\Gamma \in \mathcal{G}^{\mathrm{av}}(p,R)$ and $0 < \epsilon \le 1/2^p$. Therefore, together with (2.49) we conclude the bound

$$\begin{aligned}&\Big\|\mathcal{M}\Big(\big(\langle B\cdot\rangle^{[p/2]}, \overline{\langle B\cdot\rangle}^{[p/2]}\big); L, M, \sigma\Big)\Big\|_p \\ &\le_{p,R,\epsilon} N^{2p^2R\epsilon}\Big(1+|||\kappa|||^{\mathrm{av}}\Big)^p \|B\|^p\,(1+\|G\|_{pR/\epsilon}^{|V|})\psi_{pR/\epsilon}^{2p} N^{2|C\setminus C'|-|V_{C\setminus C'}|/2}\end{aligned} \tag{2.58}$$

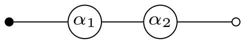
on (2.46).

2.4.4 Isotropic bound on D

We turn to the isotropic bound on D, i.e. we give bounds on (2.46) with functionals Λ of the following type. We consider fixed vectors $\mathbf{x}, \mathbf{y}$ and set $\Lambda(D) = D_{\mathbf{xy}}$ or $\Lambda(D) = \overline{D_{\mathbf{xy}}}$. Up to sign we then have

$$|\Lambda_{\alpha_1,\dots,\alpha_n}| = N^{-n/2}(\Delta^{\alpha_1}G \dots \Delta^{\alpha_n}G)_{\mathbf{xy}} = N^{-n/2}\mathbf{x}_{a_1}G_{b_1a_2}\dots G_{b_{n-1}a_n}G_{b_n\mathbf{y}}.$$

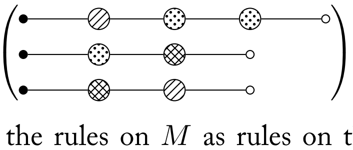
The graph component representing $\Lambda_{\alpha_1,\dots,\alpha_n}$ is a chain in contrast to the cycles in the averaged case. We also have additional *edges* representing the first $\mathbf{x}_{a_1}$ and last $G_{b_n\mathbf{y}}$ factor which we will picture as •—— and ——∘, respectively. These are special edges that are adjacent to one vertex only (the dots • and ∘ are not considered as vertices). We will call them *initial* and *final* edge. Due to these special edges we should, strictly speaking, talk about a special class of hypergraphs consisting of a union of chains each of them starting and ending with such a special edge, but for simplicity we continue to use the term *graph*. For example we have the correspondence

$\Lambda_{\alpha_1,\alpha_2} \leftrightarrow$ 

.

For $\Lambda(D) = \overline{D_{\mathbf{xy}}}$ the edges represent $\overline{x_{a_1}}, G^*_{b_k\mathbf{y}}$ and $G^*_{b_ka_{k+1}}$ but we do not indicate complex and Hermitian conjugate visually as they have no consequences on the argument. We follow the same convention regarding the colouring, as we did in the averaged case and for example have the representation

$$\begin{aligned}&\sum_{\alpha_1,\beta_1^2(1)} \kappa(\alpha_1,\beta_1^2(1)) \sum_{\alpha_2,\beta_2^1(1),\beta_2^1(2)} \frac{\kappa(\alpha_2,\beta_2^1(1),\beta_2^1(2))}{2!} \\ &\times \sum_{\alpha_3,\beta_3^2(1)} \kappa(\alpha_3,\beta_3^2(1))\Lambda_{\alpha_1,\beta_2^1(2),\beta_2^1(1)}\Lambda_{\alpha_2,\beta_3^2(1)}\Lambda_{\alpha_3,\beta_1^3(1)}\end{aligned}$$

$= \mathrm{Val}\Big($  $\Big)$.

We again rephrase the rules on M as rules on the graph Γ. We consider all vertex coloured graphs Γ such that the connected components are chains with an initial edge of type •—— and a final edge of type ——∘ such that

(a) there exist p connected components, all of which are chains,

(b) every component contains at least one vertex,

(c) every colour occurs at least once on a vertex adjacent to •——,

(d) every colour occurs at least twice,

(e) if a colour occurs exactly twice, then it occurs in two different chains.

The set of graphs satisfying (a)–(e) will be denoted by $\mathcal{G}^{\mathrm{iso}}(p,R)$ and for each L, M, σ in (2.46) we can write the main term $\mathcal{M}$ as

$$\mathcal{M}\Big((\langle \mathbf{x}, \cdot \mathbf{y}\rangle^{[p/2]}, \overline{\langle \mathbf{x}, \cdot \mathbf{y}\rangle}^{[p/2]}); L, M, \sigma) = \mathrm{Val}(\Gamma), \quad \Gamma \in \mathcal{G}^{\mathrm{iso}(p,R)} \tag{2.59}$$

where $\langle \mathbf{x}, \cdot \mathbf{y}\rangle^{[p/2]}$ denotes the tuple of $p/2$ functionals mapping $D \mapsto \langle \mathbf{x}, D\mathbf{y}\rangle$ and similarly for $\overline{\langle \mathbf{x}, \cdot \mathbf{y}\rangle}$. As the number of such graphs is finite for given p, R it follows that it sufficient to prove the required bound for every single graph.

In contrast to the averaged case, where each Λ carried a factor $1/N$ from the definition of $\Lambda(D) = N^{-1} \operatorname{Tr} BD$, now the naive size of the sum over Γ is not of order 1, but of order

$$\mathrm{Val}(\Gamma) \lesssim \prod_{c \in C} N^{2-|V_c|/2} = N^{2|C|-|V|/2}, \tag{2.60}$$

which can be large. Consequently we have to be more careful in our bound and first make use of a cancellation.

Step 1: Improved naive size

We first observe that we can reduce the naive size (2.60) to order 1, without using any Ward estimates, yet. The improvement comes from the fact that sums of the type

$$\sum_a v_a G_{ab} = G_{\mathbf{v}b}$$

can be directly bounded via the right hand side by $|G_{\mathbf{v}b}| \lesssim \|\mathbf{v}\|$ using the isotropic bound. Note that the naive estimate on the left hand side would be

$$\left| \sum_a v_a G_{ab} \right| \lesssim \sum_a |v_a| \le \sqrt{N}\, \|\mathbf{v}\|$$

and even with a Ward estimate it can only be improved to

$$\left| \sum_a v_a G_{ab} \right| \le \|\mathbf{v}\| \sqrt{\sum_a |G_{ab}|^2} \le \sqrt{N}\psi\, \|\mathbf{v}\|$$

So the procedure "summing up a vector $\mathbf{v}$ into the argument of G" is much more efficient than a Ward estimate. The limitation of this idea is that only deterministic vectors $\mathbf{v}$ can be summed up, since isotropic bounds on $G_{\mathbf{uv}}$ hold only for fixed vectors $\mathbf{u}, \mathbf{v}$.

Improvement for colours occurring twice in Γ

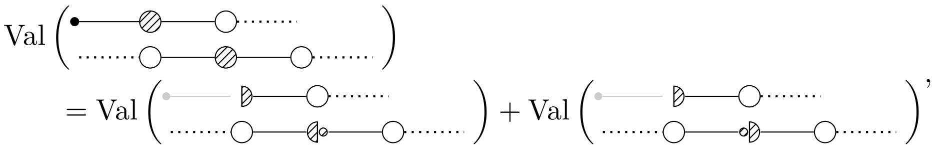
For colours which occur exactly twice we can sum up the $\mathbf{x}$ into a G factor without paying the price of an N factor from this summation. To do so, we consider an arbitrary partition of $\kappa = \kappa_c + \kappa_d$, where one should think of that $\kappa_d(\alpha_1, \alpha_2)$ forces $\alpha_1 = (a_1, b_1)$ to be close to $\alpha_2 = (a_2, b_2)$, whereas $\kappa_d(\alpha_1, \alpha_2)$ forces (a_1, b_1) to be close to (b_2, a_2). In both cases we can, according to rule (b), perform two single index summations as follows. First, we sum up the index a_1 of $\mathbf{x}$ as

$$\sum_{a_1} \kappa(a_1 b_1, a_2 b_2) x_{a_1} = \kappa(\mathbf{x} b_1, a_2 b_2).$$

Then we sum up its companion b_2 or a_2, depending on whether we consider the cross or direct term:

$$\sum_{b_2} \kappa_c(\mathbf{x} b_1, a_2 b_2) G_{b_2 \mathbf{v}} = G_{\kappa_c(\mathbf{x} b_1, a_2 \cdot)\mathbf{v}} \quad \text{or} \quad \sum_{a_2} \kappa_d(\mathbf{x} b_1, a_2 b_2) G_{\mathbf{v} a_2} = G_{\mathbf{v}\kappa_d(\mathbf{x} b_1, \cdot b_2)},$$

where $\mathbf{v}$ can be any vector or index. Thus we effectively performed a single label (two index) summation into a single G factor that will be estimated by a constant in the isotropic norm. We indicate this summation graphically by introducing half-vertices ◁ and ▷ representing the single leftover indices a and b corresponding to a label $\alpha = (a, b)$ and new (half)edges ∘—— and ——∘ representing the $G_{\kappa_c(\mathbf{x} b_1, a_2 \cdot)\mathbf{v}}$ and $G_{\mathbf{v}\kappa_d(\mathbf{x} b_1, \cdot b_2)}$ factors. To indicate that the half-edges representing $\mathbf{x}$ have been summed, we grey them out. This partial summation can thus be graphically represented as



since

$$\begin{aligned}\sum_{a_1, b_1, a_2, b_2} & \kappa(a_1 b_1, a_2 b_2)\, (\mathbf{x}_{a_1} G_{b_1 \mathbf{u}})\, (G_{\mathbf{v} a_2} G_{b_2 \mathbf{w}}) \\ &= \sum_{b_1, a_2} (G_{b_1 \mathbf{u}}) \left(G_{\mathbf{v} a_2} G_{\kappa_c(\mathbf{x} b_1, a_2 \cdot)\mathbf{w}} \right) + \sum_{b_1, b_2} (G_{b_1 \mathbf{u}}) \left(G_{\mathbf{v}\kappa_d(\mathbf{x} b_1, \cdot b_2)} G_{b_2 \mathbf{w}} \right)\end{aligned}$$

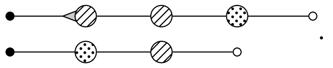
where $\mathbf{u}, \mathbf{v}, \mathbf{w}$ are the connecting indices from the white vertices.

Improvement for colours occurring three times in Γ

For colours which appear exactly three times we cannot perform the summation of $\mathbf{x}$ directly. We can, however use a Cauchy-Schwarz in the vertex adjacent to the $\mathbf{x}$–edge to improve the naive size of the ⊘–sum to $N^{3/2}$ from N^2. Explicitly, for any index or vector $\mathbf{v}$ we use that

$$\sum_{a_i, b_i} |\mathbf{x}_{a_i} G_{b_i \mathbf{v}}| \lesssim \sum_{a_i, b_i} |\mathbf{x}_{a_i}| \le N^{3/2} \left(\sum_{a_i} |\mathbf{x}_{a_i}|^2 \right)^{1/2} = N^{3/2} \|\mathbf{x}\|.$$

To indicate the intend to use the Cauchy-Schwarz improvement on a specific $\mathbf{x}$ edge, we mark the corresponding edge with a marking originating in the adjacent vertex, very much similar to the marking procedure in the averaged case. To differentiate this marking from those indicating the potential for a Ward estimate we use a grey marking •——◁○. As an example we would indicate



After these two improvements over (2.60) the naive size (naive in the sense without any Ward estimates, yet) of the summed graph is

$$\mathrm{Val}(\Gamma) \lesssim \Big(\prod_{\substack{c\in C,\\ |V_c|=2}} N^{1-|V_c|/2}\Big)\Big(\prod_{\substack{c\in C,\\ |V_c|=3}} N^{3/2-|V_c|/2}\Big)\Big(\prod_{\substack{c\in C,\\ |V_c|\geq 4}} N^{2-|V_c|/2}\Big) \leq 1. \tag{2.61}$$

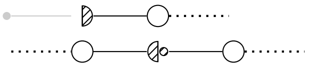
Notice that the first two factors give 1, so the improved power counting for colours with two or three occurrences is neutral. We thus restored the order 1 bound and can now focus on the counting of Ward estimates, with which we can further improve the bound.

Step 2: Further improvements through Ward estimates

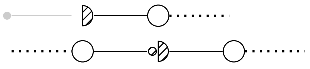
The counting procedure is very similar to what we used in the averaged law in the sense that we mark potential edges for Ward estimates colour by colour. To be consistent with the improved naive bound we count the grey initial edges (those from the summation of colours occurring twice) and the initial edges with a grey arrow (those from the summation of colours appearing three times) as unmarked, since they will not be available for Ward estimates.

Marking procedure for colours occurring twice

Colours occurring twice can, after Step 1, only occur in the reduced forms

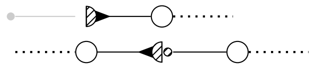


and

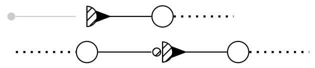


,

where we allow ⋯⋯○—— and ——○⋯⋯ to stand for an arbitrary continuation of the graph, as well as the initial •—— and final edge ——∘. In both cases we mark the edges adjacent to the remaining two half-vertices to obtain:



and



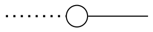
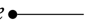
Thus for colours appearing twice we always leave two edges unmarked (which includes the greyed out initial edge). Using the $|||\kappa|||_2^{\mathrm{iso}}$ norm we indeed find that the solid edges in the two graphs above can be bounded by

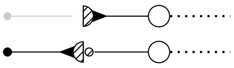
$$\begin{aligned}\sum_{b_1,a_2} \Big|G_{b_1\mathbf{u}}G_{\mathbf{v}a_2}G_{\kappa_c(\mathbf{x}b_1,a_2\cdot)\mathbf{w}}\Big| &\lesssim \|\mathbf{w}\| \sum_{b_1,a_2} |G_{b_1\mathbf{u}}G_{\mathbf{v}a_2}\, \|\kappa_c(\mathbf{x}b_1,a_2\cdot)\|| \\ &\lesssim N^2\psi^2\, |||\kappa|||_2^{\mathrm{iso}}\, \|\mathbf{x}\|\, \|\mathbf{u}\|\, \|\mathbf{v}\|\, \|\mathbf{w}\|\end{aligned} \tag{2.62}$$

and

$$\begin{aligned}\sum_{b_1,b_2} \Big|G_{b_1\mathbf{u}}G_{\mathbf{v}\kappa_d(\mathbf{x}b_1,\cdot b_2)}G_{b_2\mathbf{w}}\Big| &\lesssim \|\mathbf{v}\| \sum_{b_1,b_2} |G_{b_1\mathbf{u}}G_{b_2\mathbf{w}}\, \|\kappa_d(\mathbf{x}b_1,\cdot b_2)\|| \\ &\lesssim N^2\psi^2\, |||\kappa|||_2^{\mathrm{iso}}\, \|\mathbf{x}\|\, \|\mathbf{u}\|\, \|\mathbf{v}\|\, \|\mathbf{w}\|\end{aligned}$$

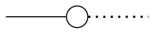
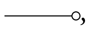
where the vectors (which are allowed to be indices, as well) $\mathbf{u}, \mathbf{v}, \mathbf{w}$ are the endpoints of the edges in the three white vertices.

Remark 2.4.6. *In the case that*  *stands for the initial edge*  *, we cannot use the Ward estimate, but use instead that* $\mathbf{x}$ *is a vector of finite norm, providing a gain of* $N^{-1/2} \ll \psi$. *For example, we could bound the graph*



by $$\sum_{b_1,a_2}\left|G_{b_1\mathbf{u}}\mathbf{x}_{a_2}G_{\kappa_c(\mathbf{x}b_1,a_2\cdot)\mathbf{w}}\right| \lesssim N^2\frac{\psi}{\sqrt{N}}\,|||\kappa|||_2^{iso}\,\|\mathbf{x}\|^2\,\|\mathbf{u}\|\,\|\mathbf{w}\|,$$

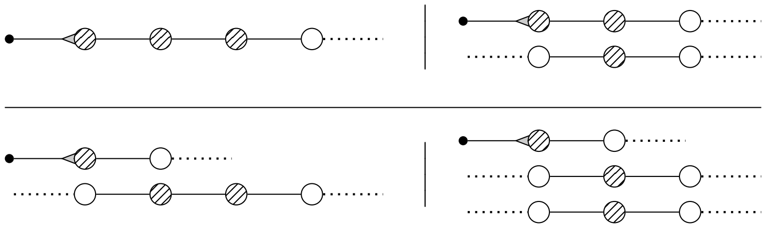
which is better than (2.62) *as* $\sqrt{N}\psi \geq 1$. *In the sequel we will not specifically distinguish this case when instead of a Ward estimate we have to use the finite norm of* $\mathbf{x}$, *as the procedure is identical and the resulting bound is always smaller in the latter case.*

We also will not separately consider the case when  *stands for the final edge* *, as we can use the same Ward estimate as before, with the difference that* $\mathbf{u}$ *and/or* $\mathbf{w}$ *are replaced by* $\mathbf{y}$.

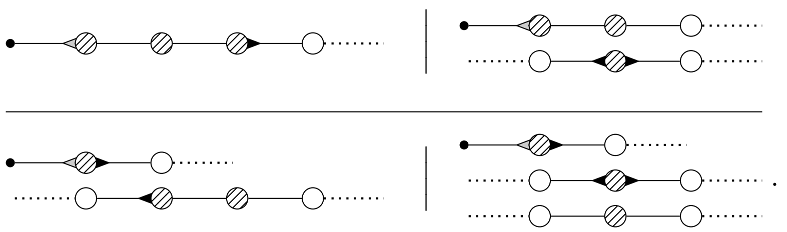
We will not manually keep track of the number of $\|\mathbf{x}\|$, $\|\mathbf{y}\|$ *in the bound as it is automatic in the sense that there are* p *initial and* p *final edges in* Γ, *each contributing a factor of* $\|\mathbf{x}\|$, $\|\mathbf{y}\|$ *to the final estimate.*

Marking procedure for colours occurring three times

Colours appearing three times occur in one of the following four forms



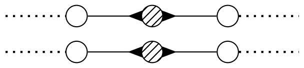
and in all cases we mark the edges adjacent to ⊘ in such a way that at most three edges (including the initial edge with the grey mark) remain unmarked. Indeed, we mark the edges as follows.



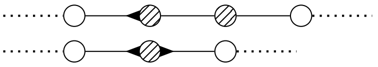
Very similar to the bound using $|||\kappa|||_3^{\mathrm{av}}$, we find that using the norm $|||\kappa|||_3^{\mathrm{iso}}$ we can perform Ward estimates on all marked edges.

Marking procedure for colours occurring more than three times

For any colour c occurring more than three times we claim that we can always mark edges in such a way that at most $2|V_c| - 4$ edges adjacent to V_c remain unmarked. Indeed, if we call an edge connected to two c–coloured vertices *c–internal* and denote their set E_c^{int}, then there are $2\,|V_c| - |E_c^{\text{int}}|$ edges adjacent to c. Out of this set of all c-adjacent edges, we mark any two and thus the claim is trivially fulfilled if $|E_c^{\text{int}}| \geq 2$. If $|E_c^{\text{int}}| = 0$, then the graph contains two single vertices of colour c, for which we mark all four adjacent edges, i.e.


,

also confirming the claim in this case. Finally, if $|E_c^{\text{int}}| = 1$, then the graph has to contain


,

for which we mark the three indicated edges, confirming the claim also in this final case. We note (again similarly to $|||\kappa|||_3^{\text{av}}$) that the norm $|||\kappa|||_k^{\text{iso}}$ allows to perform Ward estimates on all marked edges.

Counting of markings

In contrast to the averaged case, we now call an edge *ineffectively marked* if it only carries one mark and connects any two distinctly coloured vertices (in the averaged case the analogous definition was restricted to C'-coloured vertices). All other marked edges we call *effectively marked*. In particular the initial and final edge are always effectively marked, once they are marked. By construction, all effectively marked edges can be summed up by Ward estimates. In total, there are exactly $p + \sum_{c\in C} |V_c|$ edges in Γ. After the marking procedure there are at most

$$\sum_{c\in C, |V_c|=2} 2 + \sum_{c\in C, |V_c|=3} 3 + \sum_{c\in C, |V_c|\geq 4} (2\,|V_c| - 4)$$

unmarked or ineffectively marked edges in Γ. Thus there are at least

$$\begin{aligned}\Big(p + \sum_{c\in C} |V_c|\Big) - \Big(\sum_{c\in C, |V_c|=2} 2 + \sum_{c\in C, |V_c|=3} 3 + \sum_{c\in C, |V_c|\geq 4} (2\,|V_c| - 4)\Big) \\ = p + \sum_{c\in C, |V_c|\geq 4} (4 - |V_c|)\end{aligned} \tag{2.63}$$

effectively marked edges in Γ, which can be negative, but it turns out that in this case the (improved) naive size already is sufficiently small.

Power counting estimate

The strategy for performing the Ward estimates is identical to that in the averaged case; we perform them colour by colour in an arbitrary order. According to the improved naive bound from Step 1, and recalling that the power counting for $|V_c| = 2$ and $|V_c| = 3$ gives 1, i.e. is neutral, and the counting of additional effective markings we find that the summed value of Γ is bounded by

$$\mathrm{Val}(\Gamma) \lesssim N^{2|C\setminus C'| - |V_{C\setminus C'}|/2} \psi^{(p + 4|C\setminus C'| - |V_{C\setminus C'}|)_+},$$

where C' are those colours c with $|V_c| = 2, 3$.

Detailed estimate

Finally, this power counting is performed with the procedure of iterated Hölder inequalities, exactly as in the averaged case to obtain

$$\begin{aligned}\|\mathrm{Val}(\Gamma)\|_p \le_{\epsilon,R,p} N^{2p^2R\epsilon}\Big(1+|||\kappa|||^{\mathrm{iso}}\Big)^p \|\mathbf{x}\|^p \|\mathbf{y}\|^p \,(1+\|G\|^{|V|}_{pR/\epsilon})\\ \times\, \psi_{pR/\epsilon}^{(p+4|C\setminus C'|-|V_{C\setminus C'}|)_+} N^{2|C\setminus C'|-|V_{C\setminus C'}|/2}\end{aligned}$$

for all $\Gamma \in \mathcal{G}^{\mathrm{iso}}(p,R)$ and $0<\epsilon\le 1/2^p$. Therefore we conclude together with (2.59) that

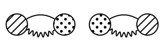
$$\begin{aligned}\Big\|\mathcal{M}\Big((\langle \mathbf{x},\cdot\mathbf{y}\rangle^{[p/2]},\overline{\langle \mathbf{x},\cdot\mathbf{y}\rangle}^{[p/2]});L,M,\sigma\Big)\Big\|_p \le_{\epsilon,R,p} N^{2p^2R\epsilon}\Big(1+|||\kappa|||^{\mathrm{iso}}\Big)^p \|\mathbf{x}\|^p \|\mathbf{y}\|^p\\ \times\,(1+\|G\|^{|V|}_{pR/\epsilon})\psi_{pR/\epsilon}^{(p+4|C\setminus C'|-|V_{C\setminus C'}|)_+} N^{2|C\setminus C'|-|V_{C\setminus C'}|/2}. \qquad (2.64)\end{aligned}$$

2.4.5 Modifications for general case

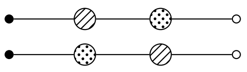
In the previous Sections 2.4.3 and 2.4.4 we estimated $\mathcal{M}$ defined in (2.46) under the simplifying assumptions $L_3=L_4=\emptyset$ and $\mathcal{N}(\alpha_l)=I$. For the final bound in (2.45) we need to treat all other cases. In this section we now demonstrate that these simplifying assumptions are not substantial and that the results from (2.55) and (2.63) on the number of available Ward estimates remain valid in the more general setting. By definition, $\mathcal{M}$ depends on the labels of types L_3 and L_4, which are considered fixed in the subsequent discussion. The graphs we introduced to systematically bound $\mathcal{M}$ do not change in their form for the general case, but only have additional *fixed* vertices α_l,β_l for $l\in L_3\cup L_4$, which we consider as uncoloured. Thus we enlarge the set graphs $\mathcal{G}^{\mathrm{av}}$ and $\mathcal{G}^{\mathrm{iso}}$ to $\widetilde{\mathcal{G}}^{\mathrm{av}}$ and $\widetilde{\mathcal{G}}^{\mathrm{iso}}$, which are defined by the previously stated rules (a)-(e) with the addition of

(f) certain vertices may be uncoloured.

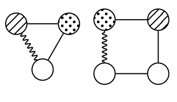
These uncoloured vertices represent exactly those labels of types L_3 and L_4, which are parameters of $\mathcal{M}$, as defined in (2.39). For example, the previously studied graphs



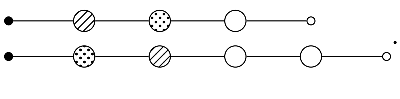
and



can be extended to

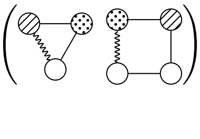


and

.

The definition of the value naturally extends to these larger classes of graphs, but without a summation over the uncoloured vertices. In the above example (2.48) is then replaced by

$$\sum_{\alpha_1,\beta_1^2(1)} \kappa(\alpha_1,\beta_1^2(1)) \sum_{\alpha_2,\beta_2^1(1)} \kappa(\alpha_2,\beta_2^1(1))\Lambda_{\alpha_1,\beta_2^1(1),\gamma(1)}\Lambda_{\alpha_2,\beta_1^2(1),\gamma(2),\gamma(3)}$$

$= \mathrm{Val}\Bigg($  $\Bigg)$, where $\gamma(1), \gamma(2), \gamma(3)$ are the fixed labels and the value of the graph depends on them. The isotropic case is analogous.

The argument in Sections 2.4.3 and 2.4.4, however, only concern those labels which are actually summed over, i.e., those of type l for $l \in L_2$. In other words, we only aim at improving the L_2-summation by Ward estimates. The presence of additional fixed labels do neither change the naive bounds, the improvement through Ward estimates, nor the counting of those Ward estimates.

Next, the restricted summations due to the neighbourhood sets $\mathcal{N}(\alpha) \subset I$ do also not change the argument. In fact, Ward estimates stay true for restricted summations since

$$\sum_{a \in \mathcal{J}} |G_{a\mathbf{x}}|^2 \le \sum_{a \in J} |G_{a\mathbf{x}}|^2 = \frac{\Im G_{\mathbf{x}\mathbf{x}}}{\eta}$$

for arbitrary $\mathcal{J} \subset J$. Also the procedure for improving the naive size in Section 2.4.4 holds true if only summed over subsets, i.e.,

$$\sum_{a_1 \in \mathcal{J}} \kappa(a_1 b_1, a_2 b_2) x_{a_1} = \kappa(\mathbf{x}_{\mathcal{J}} b_1, a_2 b_2),$$

where the sub-vector $\mathbf{x}_{\mathcal{J}}$ has bounded norm $\|\mathbf{x}_{\mathcal{J}}\| \le \|\mathbf{x}\|$.

Finally, the modification of $\mathcal{M}$ by setting $W_{\mathcal{N}_{L_3}} = 0$ also does not change the substance of the argument as the bound verbatim also covers this modified W, and the final bounds can be rephrased in terms of $\|G\|$ as $\|\widehat{G}\|_q \le_q 1 + \|G\|_{6q}^3$, as demonstrated in Lemma 2.D.3.

2.4.6 Proof of Theorem 2.4.1

We now have all the ingredients to complete the proof of Theorem 2.4.1 starting from (2.45), where we recall that $\mathcal{M}$ was defined in (2.39).

Proof of the averaged bound

We recall from (2.50) that for the averaged bound the naive size of $\mathcal{M}$ is given by

$$\mathcal{M} \lesssim N^{-p} N^{-|L|/2 - M_L/2} N^{2|L_2|},$$

where the first factor comes from the normalized trace, the second from the derivatives and the third from the L_2 summations. We demonstrated in Section 2.4.3 (see (2.56) and the counting estimate (2.55)) that through Ward estimates we can improve the naive size $N^{2|L_2|}$ of the L_2 summation to

$$\begin{aligned}\mathcal{M} &\lesssim N^{-p} N^{-|L_3 \sqcup L_4|/2 - M_{L_3 \sqcup L_4}/2} \prod_{\substack{l \in L_2 \\ M_l \ge 3}} N^{3/2 - M_l/2} \prod_{\substack{l \in L_2 \\ M_l \le 2}} (N\psi^2)^{3/2 - M_l/2} \\ &\le N^{-|L_1|} N^{-3|L_3|/2 - M_{L_3}/2} N^{-2|L_4|} \psi^{2|L_2|} \prod_{\substack{l \in L_2 \\ M_l \ge 3}} N^{(3-M_l)/2},\end{aligned}$$

where we used that $N\psi^2 \geq 1$ and that $M_{L_4} = |L_4|$ and we recall that $p = |L_1| + |L_2| + |L_3| + |L_4|$. Consequently we have from (2.45) that

$$\begin{aligned}
\mathbf{E}\,|\langle BD\rangle|^p &\lesssim_{p,\mu} N^{-p} + \sum_{\bigsqcup L_i=[p]} N^{-|L_1|}\psi^{2|L_2|}\Bigg[\prod_{\substack{l\in L_2\\ M_l\geq 3}} N^{(3-M_l)/2}\Bigg] N^{-|L_3|-\mu M_{L_3}} N^{-|L_4|},\\
&\lesssim_{p,\mu} N^{-p} + \psi^{2p} \sum_{\bigsqcup L_i=[p]} \Bigg[\prod_{\substack{l\in L_2\\ M_l\geq 3}} N^{(3-M_l)/2}\Bigg] N^{-\mu M_{L_3}} \qquad (2.65)\\
&\lesssim_{p,\mu} \psi^{2p} \sum_{\bigsqcup L_i=[p]} N^{-\frac{1}{2}(M_{L_2}-3p)_+ -\mu M_{L_3}},
\end{aligned}$$

where we bounded the L_3-summation in (2.45) by

$$N^{2|L_3|}\big(N^{1/2-\mu}\big)^{M_{L_3}} = N^{2|L_3|+M_{L_3}/2} N^{-\mu M_{L_3}}$$

in the first line, and used $N^{-1} \leq \psi^2$ in the second. To conclude the moment bound (2.23b) from (2.65) we have to count the number of $\|G\|_q$'s just as in the proof of (2.58). The key point is to collect enough $N^{-\mu}$ factors so that all but maybe $O(p)$ factors $\|G\|_q$ could be compensated by an $N^{-\mu}$. Since all $|L_i|$ and $M_{L_4} = |L_4|$ are of order p, the only way of collecting more than Cp factors of $\|G\|_q$ is having M_{L_2} or M_{L_3} bigger than a constant times p. But in this case we collect the same order of factors of the type $N^{-1/2}$ or $N^{-\mu}$ from (2.65) and the claim follows since $N^{-1/2} \leq N^{-\mu}$.

Proof of the isotropic bound

We recall from (2.61) that for the isotropic law the improved naive size of $\mathcal{M}$ is given by

$$\mathcal{M} \lesssim N^{-|L_3\sqcup L_4|/2 - M_{L_3\sqcup L_4}/2} \prod_{\substack{l\in L_2\\ M_l\geq 3}} N^{3/2-M_l/2}$$

and from (2.63) that we can always perform at least $\big(p + \sum_{l\in L_2,\, M_l\geq 3}(3 - M_l)\big)_+$ Ward estimates. Consequently, with Proposition 2.4.4 and (2.45) we obtain

$$\begin{aligned}
\mathbf{E}\,|D_{\mathbf{xy}}|^p \lesssim_{p,\mu} N^{-p} + \sum_{\bigsqcup L_i=[p]} & N^{-\mu M_{L_3}} \psi^{\left(p+\sum_{l\in L_2,\, M_l\geq 3}(3-M_l)\right)_+}\\
&\times \prod_{\substack{l\in L_2\\ M_l\geq 3}} N^{3/2-M_l/2},
\end{aligned} \qquad (2.66)$$

where we again bounded the L_3 summation in (2.45) by $N^{2|L_3|+M_{L_3}/2}N^{-\mu M_{L_3}}$. The rhs. of (2.66) is bounded by ψ^p since every missing ψ power is compensated by an $N^{-1/2} \ll \psi$. To conclude the moment bound (2.23a) from (2.66) we again have to count the number of $\|G\|_q$-factors as in the proof of (2.64) This very similar to the averaged case above and completes the proof of Theorem 2.4.1.

2.5 Proof of the stability of the MDE and proof of the local law

Before going into the proof of Theorem 2.2.2, we collect some facts from [102, 13, 8] about the deterministic MDE (2.1) and its solution.

Proposition 2.5.1 (Stability of MDE and properties of the solution). *The following hold true under Assumption (2.A).*

(i) *The MDE* (2.1) *has a unique solution $M = M(z)$ for all $z \in \mathbb{H}$ and moreover the map $z \mapsto M(z)$ is holomorphic.*

(ii) *The holomorphic function $\langle M \rangle : \mathbb{H} \to \mathbb{H}$ is the Stieltjes transform of a probability measure μ on $\mathbb{R}$.*

(iii) *There exists a constant $c > 0$ such that we have the bounds*

$$\frac{c}{\langle z \rangle + |||\mathcal{S}||| \operatorname{dist}(z, \operatorname{supp}\mu)^{-1}} \le \|M(z)\| \le \frac{1}{\operatorname{dist}(z, \operatorname{supp}\mu)}$$
$$\|\Im M\| \le \frac{\eta}{\operatorname{dist}(z, \operatorname{supp}\mu)^2},$$

where we recall the definition of $|||\mathcal{S}|||$ in (2.24).

(iv) *There exist constants $c, C > 0$ such that*

$$\left\|(1 - \mathcal{C}_{M(z)}\mathcal{S})^{-1}\right\|_{hs \to hs} \le c\Big[\frac{\langle z \rangle}{\operatorname{dist}(z, \operatorname{supp}\mu)} + \frac{|||\mathcal{S}|||}{\operatorname{dist}(z, \operatorname{supp}\mu)^2}\Big]^C,$$

where $\mathcal{C}$ is the sandwiching *operator $\mathcal{C}_R[T] := RTR$. The norm on the lhs. is the operator norm where $1 - \mathcal{C}_M\mathcal{S}$ is viewed as a linear map on the space of matrices equipped with the Hilbert-Schmidt norm.*

If, in addition, Assumption (2.E) is also satisfied, then the following statements hold true, as well.

(v) *The measure μ from* (ii) *is absolutely continuous with respect to the Lebesgue measure and has a continuous density $\rho\colon \mathbb{R} \to [0, \infty)$, called the self-consistent density of states, which is also real analytic on the open set $\{ \rho > 0 \}$.*

(vi) *There exist constants $c, C > 0$ such that we have the bounds*

$$\frac{c}{\langle z \rangle} \le \|M(z)\| \le \frac{C}{\rho(z) + \operatorname{dist}(z, \operatorname{supp}\rho)}, \;\; c\rho(z) \le \Im M(z) \le C\langle z \rangle^2 \|M(z)\|^2 \rho(z)$$

in terms of the harmonic extension $\rho(z) := \pi^{-1}\Im\langle M(z) \rangle$ of the self-consistent density of states to the upper half plane $\mathbb{H}$.

(vii) *There exist constants $c, C > 0$ such that*

$$\left\|(1 - \mathcal{C}_{M(z)}\mathcal{S})^{-1}\right\|_{hs \to hs} \le c\left(1 + [\rho(z) + \operatorname{dist}(z, \operatorname{supp}\rho)]^{-C}\right).$$

Proof. Parts (i)–(ii) follow from [102, Thm. 2.1]. Parts (iii)–(iv) follow from [13, Section 3] and $\|M\| \ge \|M^{-1}\|^{-1}$. Finally, parts (v)–(vii) follow from [8, Prop. 2.2, 4.2, 4.4]. □

Due to Assumption (2.C), (2.24) and (2.73) below we have $|||\mathcal{S}||| \leq C$. Therefore parts (iii),(iv),(vi) and (vii) show that we have

$$\langle z \rangle \, \|M(z)\| \leq_\epsilon N^\epsilon \quad \text{and} \quad \left\| (1 - \mathcal{C}_M \mathcal{S})^{-1} \right\|_{\mathrm{hs} \to \mathrm{hs}} \leq_\epsilon N^\epsilon \quad \text{in} \quad \mathbb{D}^\delta_{\mathrm{out}} \tag{2.67}$$

and also in $\mathbb{D}^\delta_0$ under Assump. (2.E), for some $\delta = \delta(\epsilon) > 0$. Similarly to (2.67), we will often state estimates that hold both in the spectral domain $\mathbb{D}^\delta_{\mathrm{out}}$ without Assumption (2.E) as well as in the spectral domain $\mathbb{D}^\delta_\gamma$ but under Assumption (2.E). We recall that according to our convention about $\leq_\epsilon$, (2.67) implies the existence of a constant $C(\epsilon)$ such that the inequalities hold true with that constant for all z in the given ϵ-dependent domains.

2.5.1 Definition of an isotropic norm suitable for the stability analysis

For a fixed $z \in \mathbb{H}$ define the map

$$\mathcal{J}_z[G, D] \coloneqq 1 + (z - A + \mathcal{S}[G])G - D$$

on arbitrary matrices G and D. From the definition of $D = D(z)$ (2.2) and the solution $M = M(z)$ of the MDE (2.1) it follows that $\mathcal{J}_z[M(z), 0] = 0$ and $\mathcal{J}_z[G(z), D(z)] = 0$. Throughout this discussion we will fix z and we omit it from the notation, i.e. $\mathcal{J} = \mathcal{J}_z$. We will consider G as a function $G = G(D)$ of an arbitrary error matrix D satisfying $\mathcal{J}[G(D), D] = 0$. Via the implicit function theorem, this relation defines a unique function $G(D)$ for sufficiently small D and $G(D)$ will be analytic as long as $\mathcal{J}$ is stable. The stability will be formulated in a specific norm that takes into account that the smallness of D can only be established in isotropic sense, i.e. in the sense of high moment bound on $D_{\mathbf{xy}}$ for any fixed deterministic vectors $\mathbf{x}, \mathbf{y}$. To define this special norm, we fix vectors $\mathbf{x}, \mathbf{y}$ and define sets of vectors containing the standard basis vectors e_a, $a \in J$, recursively by

$$\begin{aligned} I_0 &\coloneqq \{\, \mathbf{x}, \mathbf{y} \,\} \cup \{\, e_a \mid a \in J \,\}, \\ I_{k+1} &\coloneqq I_k \cup \{\, M\mathbf{u} \mid \mathbf{u} \in I_k \,\} \cup \{\, \kappa_c((M\mathbf{u})a, b\cdot), \kappa_d((M\mathbf{u})a, \cdot b) \mid \mathbf{u} \in I_k,\ a, b \in J \,\}, \end{aligned}$$

which give rise to the norm

$$\begin{aligned} \|G\|_* = \|G\|_*^{K,\mathbf{x},\mathbf{y}} &\coloneqq \sum_{0 \leq k < K} N^{-k/2K} \|G\|_{I_k} + N^{-1/2} \max_{\mathbf{u} \in I_K} \frac{\|G_{\cdot \mathbf{u}}\|}{\|\mathbf{u}\|}, \\ \|G\|_I &\coloneqq \max_{\mathbf{u}, \mathbf{v} \in I} \frac{|G_{\mathbf{uv}}|}{\|\mathbf{u}\| \, \|\mathbf{v}\|}, \end{aligned}$$

where we will choose K later.

Theorem 2.5.2. *Let $K \in \mathbb{N}$, $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$, and denote the open ball of radius δ around M in $(\mathbb{C}^{N \times N}, \|\cdot\|_*^{K,\mathbf{x},\mathbf{y}})$ by $B_\delta(M)$. Then for*

$$\epsilon_1 \coloneqq \frac{\left[1 + |||\mathcal{S}||| \, \|M\|^2 + |||\mathcal{S}|||^2 \, \|M\|^4 \, \|(1 - \mathcal{C}_M \mathcal{S})^{-1}\|_{hs \to hs}\right]^{-2}}{10 N^{1/K} \, \|M\|^2 \, |||\mathcal{S}|||}, \quad \epsilon_2 \coloneqq \sqrt{\frac{\epsilon_1}{10 \, |||\mathcal{S}|||}} \tag{2.68}$$

there exists a unique function $G\colon B_{\epsilon_1}(0) \to B_{\epsilon_2}(M)$ *with* $G(0) = M$ *that satisfies* $\mathcal{J}[G(D), D] = 0$. *Moreover, the function G is analytic and satisfies*

$$\|G(D_1) - G(D_2)\|_* \le 10\, N^{1/2K} \left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{*\to*} \|M\| \, \|D_1 - D_2\|_* \,. \tag{2.69}$$

for any $D_1, D_2 \in B_{\epsilon_1}(0)$.

Proof. First, we rewrite the equation $\mathcal{J}[G, D] = 0$ in the form $\widetilde{\mathcal{J}}[V, D] = 0$, where

$$\widetilde{\mathcal{J}}[V, D] := (1 - \mathcal{C}_M\mathcal{S})V - M\mathcal{S}[V]V + MD, \qquad V := G - M$$

and for arbitrary V and D we claim the bounds

$$\begin{aligned} \|M\mathcal{S}[V]V\|_* &\le N^{1/2K} \, |||\mathcal{S}||| \, \|M\| \, \|V\|_*^2 \,, && (2.70a)\\ \|MD\|_* &\le N^{1/2K} \, \|M\| \, \|D\|_* \,, \\ \left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{*\to*} &\le 1 + |||\mathcal{S}||| \, \|M\|^2 + |||\mathcal{S}|||^2 \, \|M\|^4 \left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{\mathrm{hs}\to\mathrm{hs}} . && (2.70b) \end{aligned}$$

We start with the proof of (2.70a). Let $\kappa = \kappa_c + \kappa_d$ be an arbitrary partition which induces a partition of $\mathcal{S} = \mathcal{S}_c + \mathcal{S}_d$ (as in Remark 2.4.2). Then for $\mathbf{u}, \mathbf{v} \in I_k$ we compute

$$\begin{aligned} \frac{|(M\mathcal{S}_c[V]V)_{\mathbf{uv}}|}{\|\mathbf{u}\| \, \|\mathbf{v}\|} &\le \frac{1}{N} \sum_{a,b} \frac{\left|V_{ab} V_{\kappa_c((M\mathbf{u})a,b\cdot)\mathbf{v}}\right|}{\|\mathbf{u}\| \, \|\mathbf{v}\|} \\ &\le |||\kappa_c|||_c \, \|V\|_{\max} \, \|M\| \min\left\{ \|V\|_{I_{k+1}}, \frac{\|V_{\cdot\mathbf{v}}\|}{\|\mathbf{v}\|} \right\}, \\ \frac{|(M\mathcal{S}_d[V]V)_{\mathbf{uv}}|}{\|\mathbf{u}\| \, \|\mathbf{v}\|} &\le \frac{1}{N} \sum_{a,b} \frac{\left|V_{a\kappa_d((M\mathbf{u})a,\cdot b)} V_{b\mathbf{v}}\right|}{\|\mathbf{u}\| \, \|\mathbf{v}\|} \\ &\le |||\kappa_d|||_d \, \|M\| \min\left\{ \|V\|_{I_{k+1}} \frac{\|V_{\cdot\mathbf{v}}\|}{\sqrt{N} \, \|\mathbf{v}\|}, \|V\|_{\max} \frac{\|V_{\cdot\mathbf{v}}\|}{\|\mathbf{v}\|} \right\}, \end{aligned} \tag{2.71}$$

where we used $|V_{a\mathbf{w}}| \le \sqrt{N} \, \|V\|_{\max} \, \|\mathbf{w}\|$ in the second bound of (2.71), so that

$$\begin{aligned} \|M\mathcal{S}_e[V]V\|_* &= \sum_{0\le k<K} \frac{\|M\mathcal{S}_e[V]V\|_{I_k}}{N^{k/2K}} + \max_{\mathbf{u}\in I_K} \frac{\|(M\mathcal{S}_e[V]V)_{\cdot\mathbf{u}}\|}{\sqrt{N} \, \|\mathbf{u}\|} \\ &\le N^{1/2K} \, |||\kappa_e|||_e \, \|M\| \, \|V\|_*^2 \end{aligned}$$

for $e \in \{c, d\}$ and (2.70a) follows immediately, recalling (2.24).

$$\frac{|(MD)_{\mathbf{uv}}|}{\|\mathbf{u}\| \, \|\mathbf{v}\|} \le \|M\| \min\left\{ \|D\|_{I_{k+1}}, \frac{\|D_{\cdot\mathbf{v}}\|}{\|\mathbf{v}\|} \right\}.$$

Finally, we show (2.70b). We use a three term geometric expansion to obtain

$$\begin{aligned} \left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{*\to*} &\le 1 + \|\mathcal{C}_M\mathcal{S}\|_{*\to*} + \|\mathcal{C}_M\mathcal{S}\|_{*\to\mathrm{hs}} \left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{\mathrm{hs}\to\mathrm{hs}} \|\mathcal{C}_M\mathcal{S}\|_{\mathrm{hs}\to*} \\ &\le 1 + \|M\|^2 \, \|\mathcal{S}\|_{\max\to\|\cdot\|} + \|M\|^4 \, \|\mathcal{S}\|_{\max\to\|\cdot\|} \left\|(1 - \mathcal{C}_M\mathcal{S})^{-1}\right\|_{\mathrm{hs}\to\mathrm{hs}} \|\mathcal{S}\|_{\mathrm{hs}\to\|\cdot\|} \end{aligned} \tag{2.72}$$

and it only remains to derive bounds on $\|\mathcal{S}\|_{\max\to\|\cdot\|}$ and $\|\mathcal{S}\|_{\mathrm{hs}\to\|\cdot\|}$. We begin to compute for the cross part κ_c and arbitrary normalized vectors $\mathbf{v},\mathbf{u}\in\mathbb{C}^N$ that

$$\begin{aligned}|\mathcal{S}_c[V]_{\mathbf{vu}}| &= \Big|\frac{1}{N}\sum_{b_1,a_2}\langle\kappa_c(\mathbf{v}b_1,a_2\cdot),\mathbf{u}\rangle\, V_{b_1a_2}\Big|\\ &\le \frac{\|V\|_{\max}}{N}\sum_{b_1,a_2}\|\kappa_c(\mathbf{v}b_1,a_2\cdot)\| \le |||\kappa_c|||_c\,\|V\|_{\max}\,,\end{aligned}$$

and

$$\begin{aligned}|\mathcal{S}_c[V]_{\mathbf{vu}}| &= \Big|\frac{1}{N}\sum_{a_1,a_2,b_2} v_{a_1}\,\langle\kappa_c(a_1\cdot,a_2b_2),V_{\cdot a_2}\rangle\,u_{b_2}\Big|\\ &\le \frac{1}{N}\sum_{a_1,a_2,b_2}|v_{a_1}|\,\|\kappa_c(a_1\cdot,a_2b_2)\|\,|u_{b_2}|\,\|V_{\cdot a_2}\| \le \frac{|||\kappa_c|||_c}{N}\sum_{a_2}\|V_{\cdot a_2}\|\\ &\le |||\kappa_c|||_c\sqrt{\frac{1}{N}\sum_{b_1,a_2}|V_{b_1a_2}|^2} = |||\kappa_c|||_c\,\|V\|_{\mathrm{hs}}\,.\end{aligned}$$

Next, we estimate for the direct part κ_d that

$$\begin{aligned}|\mathcal{S}_d[V]_{\mathbf{vu}}| &= \Big|\frac{1}{N}\sum_{b_1,b_2}\langle\kappa_d(\mathbf{v}b_1,\cdot b_2),V_{b_1\cdot}\rangle\,u_{b_2}\Big|\\ &\le \frac{1}{N}\sum_{b_1,b_2}\|V_{b_1\cdot}\|\,\|\kappa_d(\mathbf{v}b_1,\cdot b_2)\|\,|u_{b_2}| \le \frac{|||\kappa_d|||_d}{N}\sqrt{\sum_{b_1}\|V_{b_1\cdot}\|^2}\\ &\le \frac{|||\kappa_d|||_d}{N}\sqrt{\sum_{b_1,a_2}|V_{b_1a_2}|^2} \le |||\kappa_d|||_d\min\left\{\frac{\|V\|_{\mathrm{hs}}}{\sqrt{N}},\|V\|_{\max}\right\},\end{aligned}$$

so that it follows that, using (2.24),

$$\begin{aligned}\|\mathcal{S}[V]\| &= \sup_{\|\mathbf{v}\|,\|\mathbf{u}\|\le 1}|\mathcal{S}[V]_{\mathbf{vu}}| \le |||\mathcal{S}|||\min\{\|V\|_{\mathrm{hs}},\|V\|_{\max}\}\,,\\ \max\Big\{\|S\|_{\max\to\|\cdot\|},\|S\|_{\mathrm{hs}\to\|\cdot\|}\Big\} &\le |||\mathcal{S}|||\end{aligned} \tag{2.73}$$

and therefore (2.70b) follows from (2.72) with (2.73). Now the statement (2.69) follows from the implicit function theorem as formulated in Lemma 2.D.1 applied to the equation $\widetilde{\mathcal{J}}[G-M,D]=0$ written in the form

$$(1-\mathcal{C}_M\mathcal{S})V - M\mathcal{S}[V]V = -MD$$

with $A=1-\mathcal{C}_M\mathcal{S}$, $B=M$ and $d=D$ in the notation of Lemma 2.D.1. □

This general stability result will be used in the following form

$$\|G-M\|_* \le_\epsilon N^{\epsilon+1/2K}\frac{\|D\|_*}{\langle z\rangle}\quad\text{in } \mathbb{D}^\delta_{\mathrm{out}} \text{ and in } \mathbb{D}^\delta_0 \text{ under A. (2.E)} \tag{2.74}$$

for some $\delta=\delta(\epsilon)>0$, as long as $\|D\|_*\le N^{-1/2K}\langle z\rangle^2$ by applying it to $D_1=0, D_2=D(z)$ and using (2.67) and (2.70b).

2.5.2 Stochastic domination and relation to high moment bounds

In order to keep the notation compact, we now introduce a commonly used (see, e.g. [73]) notion of high-probability bound.

Definition 2.5.3 (Stochastic Domination). *If*

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

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

$$\sup_{u \in U^{(N)}} \mathbf{P}\left[X^{(N)}(u) > N^{\epsilon} Y^{(N)}(u)\right] \le N^{-D}$$

for large enough $N \ge N_0(\epsilon, D)$. In this case we use the notation $X \prec Y$.

It can be checked (see [73, Lemma 4.4]) that $\prec$ satisfies the usual arithmetic properties, e.g. if $X_1 \prec Y_1$ and $X_2 \prec Y_2$, then also $X_1 + X_2 \prec Y_1 + Y_2$ and $X_1 X_2 \prec Y_1 Y_2$. We will say that a (sequence of) events $A = A^{(N)}$ holds with *overwhelming probability* if $\mathbf{P}(A^{(N)}) \ge 1 - N^{-D}$ for any $D > 0$ and $N \ge N_0(D)$. In particular, under Assumption (2.B), we have $w_{ij} \prec 1$.

In the following lemma we establish that a control of the $\|\cdot\|_*^{K,\mathbf{x},\mathbf{y}}$-norm for all $\mathbf{x}, \mathbf{y}$ in a high probability sense is essentially equivalent to a control of the $\|\cdot\|_p$-norm for all p.

Lemma 2.5.4. *Let R be a random matrix and Φ a deterministic control parameter. Then the following implications hold:*

(i) If $\Phi \ge N^{-C}$, $\|R\| \le N^C$ and $|R_{\mathbf{x}\mathbf{y}}| \prec \Phi$ for all normalized $\mathbf{x}, \mathbf{y}$ and some C, then also $\|R\|_p \le_{p,\epsilon} N^{\epsilon}\Phi$ for all $\epsilon > 0, p \ge 1$.

(ii) Conversely, if $\|R\|_p \le_{p,\epsilon} N^{\epsilon}\Phi$ for all $\epsilon > 0, p \ge 1$, then $\|R\|_*^{K,\mathbf{x},\mathbf{y}} \prec \Phi$ for any fixed $K \in \mathbb{N}, \mathbf{x}, \mathbf{y} \in \mathbb{C}^N$.

Proof. We begin with the proof of (ii) and infer from Markov's inequality and Hölder's inequality (as in (2.57)) that

$$\begin{aligned}\mathbf{P}\left(\|R\|_* > N^{\sigma}\Phi\right) \le \mathbf{P}\left(2\,\|R\|_{I_K} > N^{\sigma}\Phi\right) &\le_p \frac{\mathbf{E}\,\|R\|_{I_K}^p}{N^{\sigma p}\Phi^p} \\ &\le_p |I_K|^{2/r}\frac{\mathbf{E}\,\|R\|_{pr}^p}{N^{\sigma p}\Phi^p} \le_{p,r,\epsilon} |I_K|^{2/r} N^{\epsilon p - \sigma p},\end{aligned}$$

and since $|I_K| \le 4^K N^{K+2}$ we conclude that $\|R\|_* \prec \Phi$ by choosing ϵ sufficiently small and p, r sufficiently large. On the other hand, (i) directly follows from

$$\|R\|_p \le N^{\epsilon}\Phi + \sup_{\|\mathbf{x}\|,\|\mathbf{y}\| \le 1}\left(|R_{\mathbf{x}\mathbf{y}}|\,\mathbf{P}[|R_{\mathbf{x}\mathbf{y}}| \ge N^{\epsilon}\Phi]^{1/p}\right). \qquad \square$$

2.5.3 Bootstrapping step

The proof of the local law follows a *bootstrapping procedure*: First, we prove the local law for $\eta \geq N$, and afterwards we iteratively show that if the local law holds for $\eta \geq N^{\gamma_0}$, then it also holds for $\eta \geq N^{\gamma_1}$ for some $\gamma_1 < \gamma_0$. We now formulate the iteration step.

Proposition 2.5.5. *The following holds true under the assumptions of Theorem 2.2.2: Let $\delta, \gamma > 0$ and $\gamma_0 > \gamma_1 \geq \gamma$ with $4(2C_*/\mu + 1)(\gamma_0 - \gamma_1) < \gamma < 1/2$ and suppose that*

$$\|G - M\|_p \leq_{\gamma,p} \frac{N^{-\gamma/6}}{\langle z\rangle} \quad \textit{in} \quad \mathbb{D}^{\delta}_{\gamma_0}, \tag{2.75}$$

holds for all $p \geq 1$, where C_* is the constant from Theorem 2.4.1. Then the same inequality (2.75) (with a possibly different implicit constant depending on γ, δ, p) holds also true in $\mathbb{D}^{\delta'}_{\gamma_1}$ for some $\delta' = \delta'(\gamma, \delta) > 0$. Furthermore, the same statement holds true under the assumptions of Theorem 2.2.1 if we replace $\mathbb{D}^{\delta}_{\gamma_0}$ and $\mathbb{D}^{\delta}_{\gamma_1}$ by $\mathbb{D}^{\delta}_{\gamma_0} \cap \mathbb{D}^{\delta}_{out}$ and $\mathbb{D}^{\delta}_{\gamma_1} \cap \mathbb{D}^{\delta}_{out}$, respectively, in the above sentence.

Proof. We first prove the assertion under the assumptions of Theorem 2.2.2. In the proof we will abbreviate the step size from γ_0 to γ_1 by $\gamma_s := \gamma_0 - \gamma_1$. We will suppress the dependence of the constants on δ, γ in our notation. In particular, (2.75) and (2.67) imply $\|G\|_p \leq_{p,\gamma} N^{\gamma_s} \langle z\rangle^{-1}$ in $\mathbb{D}^{\delta'}_{\gamma_0}$ with $\delta' = \delta'(\gamma)$. For fixed E the function $\eta \mapsto f(\eta) := \eta \|G(E + \mathrm{i}\eta)\|_p$ satisfies

$$\begin{aligned}\liminf_{\epsilon\to 0} \frac{f(\eta+\epsilon) - f(\eta)}{\epsilon} &\geq \|G(E+\mathrm{i}\eta)\|_p - \eta \left\| \lim_{\epsilon\to 0} \frac{G(E+\mathrm{i}(\eta+\epsilon)) - G(E+\mathrm{i}\eta)}{\epsilon} \right\|_p \\ &= \|G(E+\mathrm{i}\eta)\|_p - \eta \left\|G(E+\mathrm{i}\eta)^2\right\|_p \geq 0,\end{aligned}$$

where we used

$$\eta \left|\langle \mathbf{x}, G^2 \mathbf{y}\rangle\right| \leq \frac{\eta}{2}\left(\langle \mathbf{x}, |G|^2 \mathbf{x}\rangle + \langle \mathbf{y}, |G|^2 \mathbf{y}\rangle\right) \leq \frac{1}{2}\left(\langle \mathbf{x}, \Im G \mathbf{x}\rangle + \langle \mathbf{y}, \Im G \mathbf{y}\rangle\right)$$

in the last step. We thus know that $\eta \mapsto \eta \|G(E+\mathrm{i}\eta)\|_p$ is monotone and we can conclude that $\langle z\rangle \|G\|_p \leq_{p,\gamma} N^{2\gamma_s}$ in $\mathbb{D}^{\delta'}_{\gamma_1}$. From (2.23a) and $\gamma_s < \mu$ it thus follows that

$$\|D\|_p \leq_{p,\gamma,\epsilon} N^{\epsilon + 2(C_*/\mu + 1/2)\gamma_s - \gamma/2} \leq N^{\epsilon - \gamma/4} \quad \text{in} \quad \mathbb{D}^{\delta'}_{\gamma_1}. \tag{2.76}$$

Note that the exponent in the right hand side is independent of p; this was possible because the power of $\|G\|_q$ in (2.23a) was linear in p.

We now relate these high moment bounds to high probability bounds in the $\|\cdot\|_*$ norm, as defined before Theorem 2.5.2 and find for any fixed $\mathbf{x}, \mathbf{y}$ and K that $\|D\|_* \prec N^{-\gamma/4}$ from Lemma 2.5.4(ii) (we recall that the $\|\cdot\|_*$ implicitly depends on $\mathbf{x}, \mathbf{y}$ and K). Next, we apply (2.74) to obtain

$$\|G - M\|_* \chi(\|G - M\|_* \leq N^{-\gamma/9}) \prec \frac{N^{-\gamma/5}}{\langle z\rangle} \quad \text{in} \quad \mathbb{D}^{\delta'}_{\gamma_1}, \tag{2.77}$$

provided $K \geq 10/\gamma$. The bound (2.77) shows that there is a gap in the set of possible values for $\|G - M\|_*$. The extension of (2.75) to $\mathbb{D}^{\delta'}_{\gamma_1}$ then follows from a standard continuity argument using a fine grid of intermediate values of η: Suppose that (2.77) were true as a deterministic inequality. Since $\eta \mapsto \|(G-M)(E+\mathrm{i}\eta)\|_*$ is continuous, and for $\eta = N^{-1+\gamma_0}$ we know that $\|(G-M)(E+\mathrm{i}\eta)\|_* \le N^{-\gamma/6}$ by (2.75) and Lemma 2.5.4(ii), we would conclude the same bound for $\eta = N^{-1+\gamma_1}$. Going back to the $\|\cdot\|_p$-norm by Lemma 2.5.4(i) we could conclude (2.75) in $\mathbb{D}^\delta_{\gamma_1}$. Since (2.77) may not control $\|G-M\|_*$ on a set of very small probability, and we cannot exclude a "bad" set for every $\eta \in [N^{-1+\gamma_1}, N^{-1+\gamma_0}]$, we use a fine N^{-3}-grid. The relation (2.77) is only used for a discrete set of η's and intermediate values are controlled by the η^{-1}-Lipschitz continuity of $\|G-M\|_*$ in the continuity argument above. This completes the proof of Proposition 2.5.5 in the setup of Theorem 2.2.2. The proof in the setup of Theorem 2.2.1 is identical except for the fact that the inequalities (2.67) and (2.74) only hold true in the restricted set $\mathbb{D}^\delta_{\mathrm{out}}$ without Assumption (2.E). □

2.5.4 Proof of the local law and the absence of eigenvalues outside of the support

We now have all the ingredients to complete the proof of Theorems 2.2.1 and 2.2.2.

Proof of Theorems 2.2.1, 2.2.2 and Corollary 2.2.3. We will first prove Theorem 2.2.2 and then remark in the end how to adapt it to prove Theorem 2.2.1. The proof involves five steps. In the first step we derive a weak initial isotropic bound, which we improve in the second step to obtain the isotropic local law. In the third step we use the isotropic local law to obtain the averaged local law in the bulk, which we use in the fourth step to establish that with very high probability there are no eigenvalues outside the support of ρ, also proving Corollary 2.2.3. Finally, in the fifth step we use the fact that there are no eigenvalues outside the support of ρ to improve the isotropic and averaged law outside the support.

Step 1: Initial isotropic bound.

We claim the initial bound

$$\|G-M\|_p \le_{p,\gamma} \frac{N^{-\gamma/6}}{\langle z\rangle} \quad \text{in} \quad \mathbb{D}^\delta_\gamma \tag{2.78}$$

for some $\delta = \delta(\gamma)$. First, we aim at proving (2.78) for large $\eta \ge N$, i.e., in $\mathbb{D}^\delta_{\gamma=2} = \mathbb{D}^\delta_2$ for arbitrary δ. We use that

$$\|H\| = \max_k |\lambda_k| \le \sqrt{\operatorname{Tr}|H|^2} \le \sqrt{\operatorname{Tr}|A|^2} + \sqrt{N^{-1}\operatorname{Tr}|W|^2} \prec \sqrt{N},$$

as follows from Assumptions (2.A) and (2.B). Since $|z| \ge N$ and $\|H\| \prec \sqrt{N}$, we have $\|G\|_p \le_p \langle z\rangle^{-1}$ and $\|\Im G\|_p \le_p \langle z\rangle^{-2}\eta$ and thus from Theorem 2.4.1 it follows that that

$$\|D\|_p \le_{p,\epsilon} \frac{N^\epsilon}{\langle z\rangle\sqrt{N}} \quad \text{in} \quad \mathbb{D}^\delta_2.$$

We now fix normalized vectors $\mathbf{x}, \mathbf{y}$ and any $K \ge 10/\gamma$ in the norm $\|\cdot\|_* = \|\cdot\|_*^{K,\mathbf{x},\mathbf{y}}$ and translate these p norm bounds into high-probability bounds using Lemma 2.5.4 to infer

$\|D\|_* \prec \langle z\rangle^{-1}/\sqrt{N}$ and $\|G\|_* \prec \langle z\rangle^{-1}$. Using the stability in the form of (2.74) and absorbing N^ϵ factors into $\prec$ we conclude

$$\|G-M\|_* \prec \frac{N^{1/2K}}{\langle z\rangle^2\sqrt{N}} \quad \text{in} \quad \mathbb{D}_2^\delta.$$

Now (2.78) in $\mathbb{D}_2^\delta$ follows from 2.5.4(i) since $\mathbf{x},\mathbf{y}$ and K were arbitrary. By applying Proposition 2.5.5 iteratively starting from $\gamma_0 = 2$ and (possibly) reducing δ in every step we can then conclude that (2.78) holds in all of $\mathbb{D}_\gamma^\delta$ for some $\delta = \delta(\gamma) > 0$.

Step 2: Iterative improvement of the isotropic bound.

We now iteratively improve the initial bound (2.78) until we reach the intermediate bound

$$\|G-M\|_p \leq_{p,\epsilon} \frac{N^\epsilon}{\langle z\rangle}\Bigg(\sqrt{\frac{\|\Im M\|}{N\eta}} + \frac{1}{\langle z\rangle}\frac{1}{N\eta}\Bigg) \quad \text{in} \quad \mathbb{D}_\gamma^\delta \tag{2.79}$$

for $\delta = \delta(\epsilon) > 0$. From (2.78) and the bound on $\langle z\rangle\|M\|$ from (2.67) we conclude that $\langle z\rangle\|G\|_p$ is N^ϵ-bounded in $\mathbb{D}_\gamma^\delta$ for some $\delta = \delta(\epsilon) > 0$. Then from Theorem 2.4.1 and (2.78), again, it follows that

$$\|D\|_p \leq_{p,\epsilon} N^\epsilon\sqrt{\frac{\|\Im G\|_q}{N\eta}} \quad \text{and} \quad \|G-M\|_* + \|D\|_* \prec N^{-\gamma/6} \quad \text{in} \quad \mathbb{D}_\gamma^\delta. \tag{2.80}$$

From now on all claimed bounds hold true uniformly in all of $\mathbb{D}_\gamma^\delta$; we will therefore suppress this qualifier in the following steps. In order to prove (2.79), we show inductively

$$\|G-M\|_p \leq_{p,\epsilon} N^\epsilon\Psi_l, \tag{2.81}$$

where we define successively improving control parameters $(\Psi_l)_{l=0}^L$ through $\Psi_0 := 1$ and $\Psi_{l+1} := N^{-\sigma}\Psi_l = N^{-(l+1)\sigma}$, where $\sigma \in (0,1)$ is arbitrary. The final iteration step L is chosen to be the largest integer such that

$$\Psi_L \geq \frac{N^\sigma}{\langle z\rangle}\Bigg(\sqrt{\frac{\|\Im M\|}{N\eta}} + \frac{1}{\langle z\rangle}\frac{N^\sigma}{N\eta}\Bigg). \tag{2.82}$$

For the induction step from l to $l+1$, we write $\Im G = \Im M + \Im(G-M)$ and we continue from (2.80) and (2.81) and estimates that

$$\|D\|_p \leq_{p,\epsilon} N^\epsilon\Bigg(\sqrt{\frac{\|\Im M\|}{N\eta}} + \sqrt{\frac{\Psi_l}{N\eta}}\Bigg) \leq_{p,\epsilon} N^\epsilon\Bigg(\sqrt{\frac{\|\Im M\|}{N\eta}} + \frac{1}{\langle z\rangle}\frac{N^\sigma}{N\eta} + \langle z\rangle N^{-\sigma}\Psi_l\Bigg).$$

Thus we also have, for any normalized $\mathbf{x},\mathbf{y}$,

$$\|D\|_* = \|D\|_*^{K,\mathbf{x},\mathbf{y}} \prec \sqrt{\frac{\|\Im M\|}{N\eta}} + \frac{1}{\langle z\rangle}\frac{N^\sigma}{N\eta} + \langle z\rangle N^{-\sigma}\Psi_l$$

and from (2.74) we conclude

$$\|G-M\|_* \prec \frac{N^{1/2K}}{\langle z\rangle}\Bigg(\sqrt{\frac{\|\Im M\|}{N\eta}} + \frac{1}{\langle z\rangle}\frac{N^\sigma}{N\eta}\Bigg) + N^{1/2K-\sigma}\Psi_l$$

provided $K \geq 7/\gamma$ (c.f. the bound on $\|D\|_*$ from (2.80) and the definition of ϵ-neighbourhoods in (2.68)). In particular, since K can be chosen arbitrarily large, we find, for any normalized $\mathbf{x}, \mathbf{y}$ that

$$|(G-M)_{\mathbf{x}\mathbf{y}}| \prec \frac{1}{\langle z\rangle}\left(\sqrt{\frac{\|\Im M\|}{N\eta}} + \frac{1}{\langle z\rangle}\frac{N^{\sigma}}{N\eta}\right) + N^{-\sigma}\Psi_l \leq 2N^{-\sigma}\Psi_l,$$

where we used $l < L$ and (2.82) in the last step. By the definition of Ψ_{l+1} we infer

$$\|G-M\|_p \leq_{p,\epsilon} N^{\epsilon}\Psi_{l+1},$$

completing the induction step, and thereby the proof of (2.79).

Finally, in order to obtain (2.5a) from (2.79), we recall

$$\|\Im M\| \leq \|M\| \leq_{\epsilon} N^{\epsilon} \tag{2.83}$$

from Proposition 2.5.1(vi) and (2.5a) follows.

Step 3: Averaged bound.

First, it follows from (2.1) and (2.2) or equivalently from $\widetilde{\mathcal{J}}[G-M,D]=0$ that $G-M$ satisfies the following quadratic relation

$$G-M = (1-\mathcal{C}_M\mathcal{S})^{-1}\big[-MD + M\mathcal{S}[G-M](G-M)\big]$$

and therefore

$$\begin{aligned}\|\langle B(G-M)\rangle\|_p \leq{}& \left\|\langle B(1-\mathcal{C}_M\mathcal{S})^{-1}[MD]\rangle\right\|_p \\ &+ \left\|\langle B(1-\mathcal{C}_M\mathcal{S})^{-1}[M\mathcal{S}[G-M](G-M)]\rangle\right\|_p.\end{aligned}$$

By geometric expansion, as in (2.72), it follows that

$$\left\|(1-\mathcal{C}_M\mathcal{S})^{-1}\right\|_{\|\cdot\|\to\|\cdot\|} \leq 1 + \|M\|^2\,|||\mathcal{S}||| + \|M\|^4\,|||\mathcal{S}|||^2\left\|(1-\mathcal{C}_M\mathcal{S})^{-1}\right\|_{\mathrm{hs}\to\mathrm{hs}}$$

and thus that $\left\|((1-\mathcal{C}_M S)^{-1})^*[B^*]\right\| \leq_{\epsilon} N^{\epsilon}\|B\|$ by (2.67). Using (2.23b), where $((1-\mathcal{C}_M S)^{-1})^*[B^*]$ plays the role of B, and writing $\|\Im G\|_q \leq \|\Im M\| + \|G-M\|_q$ and using (2.79) we can conclude that

$$\|\langle B(G-M)\rangle\|_p \leq_{p,\epsilon,\gamma} \frac{\|B\|\,N^{\epsilon}}{\langle z\rangle}\left[\frac{\|\Im M\|}{N\eta} + \sqrt{\frac{\|\Im M\|}{N\eta}}\frac{1}{N\eta} + \frac{1}{(N\eta)^2}\right] \tag{2.84}$$

from Lemma 2.D.2. Now (2.5b) follows directly from (2.84) and (2.83).

The proof of Theorem 2.2.2 is now complete. For the proof of Theorem 2.2.1 the first three steps are identical except that we only work in the resticted domains $\mathbb{D}_{\gamma}^{\delta}\cap\mathbb{D}_{\mathrm{out}}^{\delta}$. Due to (2.67) and (2.74), it then follows that in $\mathbb{D}_{\mathrm{out}}^{\delta}$ the only place where the above proof used Assumption (2.E) is (2.83). In the absence of Assumption (2.E) we replace (2.83) by the bound $\|\Im M\| \leq \eta \operatorname{dist}(z, \operatorname{supp}\mu)^{-2}$ from Proposition 2.5.1 in (2.79) and (2.84), which only adds another negligible N^ϵ factor. This proves

$$\begin{aligned} \|G-M\|_p &\leq_{p,\epsilon} \frac{N^\epsilon}{\langle z\rangle}\Big(\sqrt{\frac{1}{N}}+\frac{1}{\langle z\rangle}\frac{1}{N\eta}\Big), \\ \|\langle B(G-M)\rangle\|_p &\leq_{p,\epsilon,\gamma} \frac{\|B\|\,N^\epsilon}{\langle z\rangle}\left[\frac{1}{N}+\sqrt{\frac{1}{N}}\frac{1}{N\eta}+\frac{1}{(N\eta)^2}\right] \end{aligned} \tag{2.85}$$

in the restricted domain $\mathbb{D}_\gamma^\delta \cap \mathbb{D}_{\text{out}}^\delta$. We now need two additional steps to prove Theorem 2.2.1 in all of $\mathbb{D}_{\text{out}}^\delta$.

Step 4: Absence of eigenvalues outside of the support.

For $B = 1$ it follows from (2.85) and a spectral decomposition of H that with very high probability in the sense of Corollary 2.2.3 there are no eigenvalues outside the support of μ. Indeed, if there is an eigenvalue λ with $\operatorname{dist}(\lambda, \operatorname{supp}\mu) \geq N^{-\delta}$, then $|\langle G(\lambda+i\eta)\rangle| \geq |\langle \Im G(\lambda+i\eta)\rangle| \geq 1/N\eta$. From (2.85) with $\epsilon = 1/4$ and $\gamma = 1/2$ we have

$$\begin{aligned} \mathbf{P}\left(\exists\lambda \text{ with } \operatorname{dist}(\lambda,\operatorname{supp}\mu)\geq N^{-\delta}\right) &\leq \mathbf{P}\left(|\langle G-M\rangle| \geq c/N\eta \text{ in } \mathbb{D}_{\text{out}}^\delta \cap \mathbb{D}_{1/2}^\delta\right) \\ &\lesssim \inf_{\eta \geq N^{-1/2}} \left(N^\epsilon\left[\eta+\frac{1}{\sqrt{N}}+\frac{1}{N\eta}\right]\right)^p \lesssim N^{-p/4}. \end{aligned}$$

Now Corollary 2.2.3 follows from the remark about the dependence of δ on ϵ in Theorem 2.2.1.

Step 5: Improved bounds outside of the support.

Now we fix z such that $\operatorname{dist}(z,\operatorname{supp}\rho) \geq N^{-\delta}$ and $\eta \geq N^{-1+\gamma}$. Then we have $\|\Im G\| \prec \eta\langle z\rangle^{-2}$ and $\|G\| \prec \langle z\rangle^{-1}$ and also $\|\Im G\|_p \leq_{p,\epsilon} N^\epsilon \eta \langle z\rangle^{-2}$ and $\|G\|_p \leq_{p,\epsilon} N^\epsilon \langle z\rangle^{-1}$ and we infer from Theorem 2.4.1 that

$$\|D\|_p \leq_{p,\epsilon} \frac{N^\epsilon}{\langle z\rangle\sqrt{N}} \quad \text{and therefore} \quad \|D\|_* \prec \frac{1}{\langle z\rangle\sqrt{N}}.$$

Again using stability in the form of (2.74) we find

$$\|G-M\|_* \prec \frac{N^{1/2K}}{\langle z\rangle^2\sqrt{N}}$$

and since K was arbitrary we also have

$$\|G-M\|_p \leq_{p,\epsilon} \frac{N^\epsilon}{\langle z\rangle^2\sqrt{N}}.$$

By Lipschitz-continuity of G and M with Lipschitz constant of order one we can extend the regime of validity of this bound from $\eta \geq N^{-1+\gamma}$ to $\eta \geq 0$ to conclude (2.4a). The improvement on the averaged law outside of the support of the ρ then follows immediately from the improved isotropic law and the fact that with very high probability there are no eigenvalues outside of the support of ρ. □

2.6 Delocalization, rigidity and universality

In this section we infer eigenvector delocalization, eigenvalue rigidity and universality in the bulk from the local law in Theorem 2.2.2. These proofs are largely independent of the correlation structure of the random matrix, so arguments that have been developed for Wigner matrices over the last few years can be applied with minimal modifications. Especially the *three step strategy* for proving bulk universality (see [80] for a short summary) has been streamlined recently [120, 77, 121] so that the only model-dependent input is the local law. The small modifications required for the correlated setup have been presented in detail in [8] and we will not repeat them. Here we only explain why the proofs in [8] work under the more general conditions imposed in the current paper. In fact, the proof of the eigenvector delocalization and eigenvector rigidity from [8] holds *verbatim* in the current setup as well. The proof of the bulk universality in [8] used that the correlation length was N^ϵ at a technical step that can be easily modified for our weaker assumptions. In the following we will highlight which arguments of [8] have to be modified in the current, more general, setup.

Proof of Corollary 2.2.4 on bulk eigenvector delocalization. As usual, delocalization of eigenvectors corresponding to eigenvalues in the bulk is an immediate corollary of the local law since for the eigenvectors $\boldsymbol{u}_k = \big(u_k(i)\big)_{i\in J}$ and eigenvalues λ_k of H and $i \in J$ we find from the spectral decomposition

$$C \gtrsim \Im G_{ii} = \eta \sum_k \frac{|u_k(i)|^2}{(E-\lambda_k)^2+\eta^2} \ge \frac{|u_k(i)|^2}{\eta} \quad \text{for} \quad z = E + \mathrm{i}\eta,$$

where the first inequality is meant in a high-probability sense and follows from the boundedness of M and Theorem 2.2.2, and the last inequality followed assuming that E is η-close to λ_k. □

Proof of Corollary 2.2.5 on bulk eigenvalue rigidity. Rigidity of bulk eigenvalues follows, verbatim as in [8, Corollary 2.9], from the improved local law away from the spectrum and [9, Lemma 5.1]. □

Proof of Corollary 2.2.6 on bulk universality. Bulk universality follows from the *three step strategy*, out of which only the third step requires a minor modification, compared to [8]. Since in [8] arbitrarily high polynomial decay outside of N^ϵ neighbourhoods was assumed, we have to replace to three term Taylor expansion in [8, Lemma 7.5] by an $2/\mu$-term cumulant expansion to accommodate for neighbourhoods of sizes $N^{1/2-\mu}$.

The key input for the universality proof through Dyson Brownian motion is the Ornstein Uhlenbeck (OU) process, which creates a family $H(t)$ of interpolating matrices between the original matrix $H = H(0)$ and a matrix with sizeable Gaussian component, for which universality is known from the second step of the three step strategy. The OU process is defined via

$$\mathrm{d}H(t) = -\frac{1}{2}(H(t)-A)\,\mathrm{d}t + \Sigma^{1/2}[\mathrm{d}B(t)], \qquad \text{where} \quad \Sigma[R] := \mathbf{E}\,\langle W^* R\rangle\, W,$$

where $B(t)$ is a matrix of independent (real, or complex according to the symmetry class of H) Brownian motions. It is designed in a way which preserves mean and covariances along the flow, i.e., $H(t) = A + N^{-1/2}W(t)$ and it is easy to check that $\mathbf{E}\,W(t) = 0$ and

$\mathbf{Cov}(w_\alpha(t), w_\beta(t)) = \mathbf{Cov}(w_\alpha(0), w_\beta(0))$, where $W(t) = (w_\alpha(t))_{\alpha\in I}$. Furthermore, Assumptions (2.C), (2.D) hold also, uniformly in t, for $W(t)$. Indeed, adding an independent Gaussian vector $\boldsymbol{g} = (g_{\alpha_1}, \dots, g_{\alpha_k})$ to $(w_{\alpha_1}, \dots, w_{\alpha_k})$ leaves the cumulant invariant by additivity

$$\kappa(w_{\alpha_1} + g_{\alpha_1}, \dots, w_{\alpha_k} + g_{\alpha_k}) = \kappa(w_{\alpha_1}, \dots, w_{\alpha_k}) + \kappa(g_{\alpha_1}, \dots, g_{\alpha_k})$$

and the fact that cumulants of Gaussian vectors vanish for $k \geq 3$ (for $k \geq 2$ we already noticed that, by design, the expectation and the covariance is invariant under t). We now estimate

$$\mathbf{E}\, f(N^{-1/2}W(t)) - \mathbf{E}\, f(N^{-1/2}W(0))$$

for smooth functions f. For notational purposes we set $v_\alpha(t) = N^{-1/2}w_\alpha(t)$ and $V(t) = N^{-1/2}W(t)$ and will often suppress the t-dependence. It follows from Ito's formula that

$$2\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\, f(V) = -\mathbf{E}\sum_\alpha v_\alpha(\partial_\alpha f)(V) + \sum_{\alpha,\beta}\mathbf{Cov}(v_\alpha, v_\beta)\,\mathbf{E}(\partial_\alpha\partial_\beta f)(V).$$

We now apply Proposition 2.3.2 to the first term and obtain

$$\begin{aligned}2\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\, f = &-\sum_{2\leq m<R}\sum_\alpha\sum_{\boldsymbol{\beta}\in\mathcal{N}^m}\frac{\kappa(v_\alpha, v_{\boldsymbol{\beta}})}{m!}(\mathbf{E}\,\partial_\alpha\partial_{\boldsymbol{\beta}} f)\\ &-\sum_{m<R}\sum_\alpha\sum_{\boldsymbol{\beta}\in\mathcal{N}^m}\mathbf{E}\,\frac{K(v_\alpha; v_{\boldsymbol{\beta}}) - \kappa(v_\alpha, v_{\boldsymbol{\beta}})}{m!}\partial_\alpha\partial_{\boldsymbol{\beta}} f|_{W_{\mathcal{N}}=0}\\ &-\sum_\alpha\Omega(\partial_\alpha f, \alpha, \mathcal{N}) + \sum_\alpha\sum_{\beta\in I\setminus\mathcal{N}}\kappa(v_\alpha, v_\beta)\,\mathbf{E}\,\partial_\alpha\partial_\beta f,\end{aligned}$$

where we used a cancellation for the $m = 1$ term in $\beta \in \mathcal{N}$ and the fact that $\kappa(v_\alpha) = \mathbf{E}\, v_\alpha = 0$ for the $m = 0$ term. We now estimate the four terms separately. The sum in the last term is of size N^4, the derivative contributes an N^{-1} and the covariance is assumed to be N^{-3} small, i.e., the last term is of order 1. The first term for fixed m is of size $|||\kappa|||^{\mathrm{av}}\, N^{2-(m+1)/2}$ and therefore altogether of size $|||\kappa|||^{\mathrm{av}}\sqrt{N}$. Estimating the sums by their size, and the derivative by its prefactor $N^{-(R+1)/2}$, we find from (2.13) that the third term is of size

$$N^2\,|\mathcal{N}|^R\,N^{-(R+1)/2} \leq N^{3/2-\mu R},$$

which can be made smaller than $\sqrt{N}$ by choosing $R = 2/\mu$. Finally, the second term is naively of size $N^{3/2}$, but using (2.8c), the security layers and the pigeon-hole principle as in (2.21) or in (2.44), this can be improved to $N^{-3/2}$. We can conclude that

$$\left|\mathbf{E}\,\frac{\mathrm{d}}{\mathrm{d}t}f(V(t))\right| \lesssim \sqrt{N} \quad\text{and therefore}\quad |\mathbf{E}\, f(V(t)) - \mathbf{E}\, f(V(0))| \lesssim t\sqrt{N}.$$

The remaining argument of [8, Section 7.2] can be, assuming fullness as in Assumption (2.F), followed verbatim to conclude bulk universality. □

2.A Cumulants

In this section we provide some results on cumulants which we refer to in the main part of the proof. The section largely follows the approach of [163, 134], but our application requires a more quantitative version of the independence property exhibited by cumulants, which we work out here.

Cumulants $\kappa_{\boldsymbol{m}}$ of a random vector $\boldsymbol{w} = (w_1, \dots, w_l)$ are traditionally defined as the coefficients of log-characteristic function

$$\log \mathbf{E}\, e^{i\boldsymbol{t}\cdot \boldsymbol{w}} = \sum_{\boldsymbol{m}} \kappa_{\boldsymbol{m}} \frac{(i\boldsymbol{t})^{\boldsymbol{m}}}{\boldsymbol{m}!},$$

while the (mixed) moments of $\boldsymbol{w}$ are the coefficients of the characteristic function

$$\mathbf{E}\, e^{i\boldsymbol{t}\cdot \boldsymbol{w}} = \sum_{\boldsymbol{m}} (\mathbf{E}\, \boldsymbol{w}^{\boldsymbol{m}}) \frac{(i\boldsymbol{t})^{\boldsymbol{m}}}{\boldsymbol{m}!},$$

where $\sum_{\boldsymbol{m}}$ is the sum over all multi-indices $\boldsymbol{m} = (m_1, \dots, m_l)$. Thus

$$\exp\left(\sum_{\boldsymbol{m}} \kappa_{\boldsymbol{m}} \frac{(i\boldsymbol{t})^{\boldsymbol{m}}}{\boldsymbol{m}!}\right) = \sum_{\boldsymbol{m}} (\mathbf{E}\, \boldsymbol{w}^{\boldsymbol{m}}) \frac{(i\boldsymbol{t})^{\boldsymbol{m}}}{\boldsymbol{m}!}. \tag{2.86}$$

It is easy to check that for a set $A \subset [l]$ the coefficient of $\prod_{a\in A} t_a$ in (2.86) is given by

$$\mathbf{E}\, \Pi \underline{w}_A = \Big(\prod_{a\in A} \partial_{t_a}\Big) \exp\left(\sum_{\boldsymbol{m}} \kappa_{\boldsymbol{m}} \frac{\boldsymbol{t}^{\boldsymbol{m}}}{\boldsymbol{m}!}\right)\Bigg|_{\boldsymbol{t}=0} = \sum_{\mathcal{P} \vdash A} \kappa^{\mathcal{P}},$$

where $\mathcal{P} \vdash A$ indicates the summation over all partitions of the (multi)set A, and where for partitions $\mathcal{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_b\}$ of A we defined $\kappa^{\mathcal{P}} = \prod_{k=1}^{b} \kappa_{\chi(\mathcal{P}_k)}$ with $\chi(\mathcal{P}_k)$ being the characteristic multi-index of the set $\mathcal{P}_k$. Thus for a partition $\mathcal{Q}$ of $[l]$ it follows that

$$M^{\mathcal{Q}} := \prod_{\mathcal{Q}_i \in \mathcal{Q}} \mathbf{E}\, \Pi \underline{w}_{\mathcal{Q}_i} = \prod_{\mathcal{Q}_i \in \mathcal{Q}} \sum_{\mathcal{P} \vdash \mathcal{Q}_i} \kappa^{\mathcal{P}} = \sum_{\mathcal{P} \leq \mathcal{Q}} \kappa^{\mathcal{P}}, \tag{2.87}$$

where $\mathcal{P} \leq \mathcal{Q}$ indicates that $\mathcal{P}$ is a finer partition than $\mathcal{Q}$.

Now we establish the inverse of the relation (2.87), i.e., express cumulants in terms of products of moments. To do so, we notice that the set of partitions $\mathcal{P}$ on $[l]$ (or, in fact, any finite set) is a partially ordered set with respect to the relation $\leq$. It is, in fact, also a lattice, as any two partitions $\mathcal{P}, \mathcal{Q}$ have both a unique greatest lower bound $\mathcal{P} \wedge \mathcal{Q}$ and a unique least upper bound $\mathcal{P} \vee \mathcal{Q}$. One then defines the *incidence algebra* as the algebra of scalar functions f mapping intervals $[\mathcal{P}, \mathcal{Q}] = \{\, \mathcal{R} \mid \mathcal{P} \leq \mathcal{R} \leq \mathcal{Q} \,\}$ to scalars $f(\mathcal{P}, \mathcal{Q})$ equipped with point-wise addition and scalar multiplication and the product $*$

$$(f * g)(\mathcal{P}, \mathcal{Q}) = \sum_{\mathcal{P} \leq \mathcal{R} \leq \mathcal{Q}} f(\mathcal{P}, \mathcal{R}) g(\mathcal{R}, \mathcal{Q}).$$

There are three special elements in the incidence algebra; the δ function mapping $[\mathcal{P}, \mathcal{Q}]$ to $\delta(\mathcal{P}, \mathcal{Q}) = 1$ if $\mathcal{P} = \mathcal{Q}$ and $\delta(\mathcal{P}, \mathcal{Q}) = 0$ otherwise, the ζ function mapping all intervals $[\mathcal{P}, \mathcal{Q}]$ to $\zeta(\mathcal{P}, \mathcal{Q}) = 1$, and finally the Möbius function defined inductively via

$$\mu(\mathcal{P}, \mathcal{Q}) = \begin{cases} 1, & \text{if } \mathcal{P} = \mathcal{Q}, \\ -\sum_{\mathcal{P} \leq \mathcal{R} < \mathcal{Q}} \mu(\mathcal{P}, \mathcal{R}), & \text{if } \mathcal{P} < \mathcal{Q}. \end{cases}$$

The δ function is the unit element of the incidence algebra. It is well known (and easy to check) that the multiplicative inverse of the zeta function is the Möbius function, and vice versa, i.e., that $\mu * \zeta = \zeta * \mu = \delta$. Thus it follows that for any functions F and G on the partitions, we have

$$F(\mathcal{P}) = \sum_{\mathcal{Q}\le\mathcal{P}} G(\mathcal{Q}) \qquad \text{if and only if} \qquad G(\mathcal{Q}) = \sum_{\mathcal{P}\le\mathcal{Q}} \mu(\mathcal{P},\mathcal{Q})F(\mathcal{P}).$$

Applying this equivalence to (2.87) yields

$$\kappa^{\mathcal{P}} = \sum_{\mathcal{Q}\le\mathcal{P}} \mu(\mathcal{Q},\mathcal{P})M^{\mathcal{Q}} \tag{2.88}$$

and thus it only remains to identify μ. One can check that for $\mathcal{P}\le\mathcal{Q}$, $\mu(\mathcal{P},\mathcal{Q})$ is given by

$$\mu(\mathcal{P},\mathcal{Q}) = (-1)^{n-r}0!^{r_1}1!^{r_2}\dots(n-1)!^{r_n},$$

where n is the number of blocks of $\mathcal{P}$, r is the number of blocks of $\mathcal{Q}$ and r_i is the number of blocks of $\mathcal{Q}$ which contain exactly i blocks of $\mathcal{P}$. For the particular choice of the trivial partition $\{[l]\}$ of $[l]$ it follows that

$$\begin{aligned}\kappa(w_1,\dots,w_l) := \kappa_{(1,\dots,1)} = \kappa^{\{[l]\}} &= \sum_{\mathcal{P}}(-1)^{|\mathcal{P}|-1}(|\mathcal{P}|-1)!M^{\mathcal{P}}\\ &= \sum_{\mathcal{P}}(-1)^{|\mathcal{P}|-1}(|\mathcal{P}|-1)!\prod_{\mathcal{P}_i\in\mathcal{P}}\mathbf{E}\,\Pi\underline{w}_{\mathcal{P}_i},\end{aligned} \tag{2.89}$$

providing an alternative (purely combinatorial) definition of cumulants.

Lemma 2.A.1. *If for a partition of the index set $[n] = A \sqcup B$ with $|A|, |B| > 0$ the random variables $\underline{w}_A$ and $\underline{w}_B$ are independent, then $\kappa(\underline{w}_{[n]}) = \kappa(\underline{w}_A,\underline{w}_B) = 0$. If, instead of independence, we merely assume that*

$$\mathbf{Cov}(f(w_i \mid i\in A), g(w_j \mid j \in B)) \le \epsilon\,\|f\|_2\,\|g\|_2 \tag{2.90}$$

for all f, g, and that the random variables w_i have finite $2n$-th moments $\max_i \mathbf{E}\,|w_i|^{2n} \le \mu_{2n}$, then we still have

$$|\kappa(\underline{w}_{[n]})| \le \epsilon\, C(n,\mu_{2n}).$$

Proof. We first recall the well known proof, based on the relations (2.87)–(2.88), that the cumulant of independent $\underline{w}_A$, $\underline{w}_B$ vanishes. Let $\mathcal{P}$ be a partition on $[n]$, $\mathcal{Q}$ a partition on A and $\mathcal{R}$ a partition on B. $\mathcal{P}$ naturally induces partitions $\mathcal{P}\cap A$ and $\mathcal{P}\cap B$ on A and B; conversely $\mathcal{Q}$ and $\mathcal{R}$ naturally induce a partition $\mathcal{Q}\cup\mathcal{R}$ on $[n]$. We observe that $\mathcal{Q}\le\mathcal{P}\cap A$ and $\mathcal{R}\le\mathcal{P}\cap B$ if and only if $\mathcal{Q}\cup\mathcal{R}\le\mathcal{P}$. We then compute

$$\begin{aligned}\kappa(\underline{w}_{[n]}) &= \sum_{\mathcal{P}}\mu(\mathcal{P},\{[n]\})M^{\mathcal{P}} = \sum_{\mathcal{P}}\mu(\mathcal{P},\{[n]\})M^{\mathcal{P}\cap A}M^{\mathcal{P}\cap B}\\ &= \sum_{\mathcal{P}}\mu(\mathcal{P},\{[n]\})\Big(\sum_{\mathcal{Q}\le\mathcal{P}\cap A}\kappa^{\mathcal{Q}}\Big)\Big(\sum_{\mathcal{R}\le\mathcal{P}\cap B}\kappa^{\mathcal{R}}\Big)\\ &= \sum_{\mathcal{Q}\vdash A}\sum_{\mathcal{R}\vdash B}\sum_{\mathcal{P}\vdash[n]}\zeta(\mathcal{Q}\cup\mathcal{R},\mathcal{P})\mu(\mathcal{P},\{[n]\})\kappa^{\mathcal{Q}}\kappa^{\mathcal{R}}\\ &= \sum_{\mathcal{Q}\vdash A}\sum_{\mathcal{R}\vdash B}\delta(\mathcal{Q}\cup\mathcal{R},\{[n]\})\kappa^{\mathcal{Q}}\kappa^{\mathcal{R}} = 0,\end{aligned}$$

where the first equality followed from (2.88), the second equality from independence, the third equality from (2.87), the fourth equality from the previous observation, the fifth equality from $\delta = \zeta * \mu$ and the ultimate equality from the fact that the trivial partition cannot be decomposed into two partitions on smaller sets, using that $|A|, |B| > 0$.

If $\underline{w}_A$ and $\underline{w}_B$ are not independent but merely (2.90) holds, then there is an additional covariance term in the second step in the above equation. We write

$$M^{\mathcal{P}} = \prod_{\mathcal{P}_i \in \mathcal{P}} \mathbf{E}\, \Pi \underline{w}_{\mathcal{P}_i} = \prod_{\mathcal{P}_i \in \mathcal{P}} \Big[(\mathbf{E}\, \Pi \underline{w}_{\mathcal{P}_i \cap A})(\mathbf{E}\, \Pi \underline{w}_{\mathcal{P}_i \cap B}) + \mathbf{Cov}(\Pi \underline{w}_{\mathcal{P}_i \cap A}, \Pi \underline{w}_{\mathcal{P}_i \cap B}) \Big]$$

and thus the claim follows from (2.90). □

2.B Precumulants and Wick polynomials

The precumulants defined in Section 2.3 are structurally similar to the well known *Wick polynomials* (which are also known as *Appell polynomials*). We first recall some basic definitions and facts about Wick polynomials from [88]. For a random vector $\boldsymbol{X}$ of length $|\boldsymbol{X}|$ we can define the Wick polynomial $:\boldsymbol{X}:$ as the derivative

$$:\boldsymbol{X}: := \partial_{t_1} \dots \partial_{t_{|\boldsymbol{X}|}} \frac{e^{\boldsymbol{t}\cdot\boldsymbol{X}}}{\mathbf{E}\, e^{\boldsymbol{t}\cdot\boldsymbol{X}}}\Big|_{\boldsymbol{t}=0}.$$

Alternatively, we can define $:\boldsymbol{X}:$ combinatorially as

$$:\boldsymbol{X}: = \sum_{\boldsymbol{X}' \subset \boldsymbol{X}} (\Pi \boldsymbol{X}') \sum_{\mathcal{P} \vdash \boldsymbol{X} \setminus \boldsymbol{X}'} (-1)^{|\mathcal{P}|} \prod_{\mathcal{P}_i \in \mathcal{P}} \kappa(\mathcal{P}_i).$$

or indirectly via

$$\Pi \boldsymbol{X} = \sum_{\boldsymbol{X}' \subset \boldsymbol{X}} :\boldsymbol{X}': (\mathbf{E}\, \Pi(\boldsymbol{X} \setminus \boldsymbol{X}')). \tag{2.91}$$

One useful property of Wick polynomials is that for any random variable Y we have

$$\mathbf{E}\, Y :\boldsymbol{X}_1 \sqcup \boldsymbol{X}_2: = 0 \qquad \text{whenever} \quad \boldsymbol{X}_1 \quad \text{is independent of} \quad \{\boldsymbol{X}_2, Y\} \tag{2.92}$$

and $\boldsymbol{X}_1$ is not empty. Eq. (2.92) follows, for example, immediately from the analytical definition since

$$\mathbf{E}\, Y :\boldsymbol{X}_1 \sqcup \boldsymbol{X}_2: = \partial_{\boldsymbol{t}} \frac{\mathbf{E}\, Y e^{\boldsymbol{t}_1 \cdot \boldsymbol{X}_1 + \boldsymbol{t}_2 \cdot \boldsymbol{X}_2}}{\mathbf{E}\, e^{\boldsymbol{t}_1 \cdot \boldsymbol{X}_1 + \boldsymbol{t}_2 \cdot \boldsymbol{X}_2}}\Big|_{\boldsymbol{t}=0} = \partial_{\boldsymbol{t}} \frac{\mathbf{E}\, Y e^{\boldsymbol{t}_2 \cdot \boldsymbol{X}_2}}{\mathbf{E}\, e^{\boldsymbol{t}_2 \cdot \boldsymbol{X}_2}}\Big|_{\boldsymbol{t}=0}$$

by independence and the remaining derivative vanishes as the function is constant with respect to $\boldsymbol{t}_1$.

Our pre-cumulants $K(X; \boldsymbol{Y})$ and their centered versions $K(X; \boldsymbol{Y}) - \kappa(X, \boldsymbol{Y})$ are inherently non-symmetric functions due to the special role of X. After symmetrization, however, we can express them through Wick polynomials as

$$\sum_{X \in \boldsymbol{X}} [K(X; \boldsymbol{X} \setminus \{X\}) - \kappa(\boldsymbol{X})] = |\boldsymbol{X}|\, \Pi \boldsymbol{X} - \sum_{\boldsymbol{X}' \subset \boldsymbol{X}} |\boldsymbol{X}'|\, (\mathbf{E}\, \Pi \boldsymbol{X}') :\boldsymbol{X} \setminus \boldsymbol{X}':. \tag{2.93}$$

In order to prove (2.93) we start from (2.8b) and compute

$$\begin{aligned}\sum_{X\in\boldsymbol{X}}\left[K(X;\boldsymbol{X}\setminus\{X\})-\kappa(\boldsymbol{X})\right]&=|\boldsymbol{X}|\,\Pi\boldsymbol{X}-\sum_{\boldsymbol{X}'\subset\boldsymbol{X}}|\boldsymbol{X}\setminus\boldsymbol{X}'|\,(\Pi\boldsymbol{X}')\kappa(\boldsymbol{X}\setminus\boldsymbol{X}')\\&=|\boldsymbol{X}|\,\Pi\boldsymbol{X}-\sum_{\boldsymbol{X}''\subset\boldsymbol{X}'\subset\boldsymbol{X}}|\boldsymbol{X}\setminus\boldsymbol{X}'|:\boldsymbol{X}''\colon(\mathbf{E}\,\Pi(\boldsymbol{X}'\setminus\boldsymbol{X}''))\kappa(\boldsymbol{X}\setminus\boldsymbol{X}'),\end{aligned}$$

where the second inequality followed from (2.91). We now relabel the summation indices to obtain

$$\begin{aligned}&\sum_{X\in\boldsymbol{X}}\left[K(X;\boldsymbol{X}\setminus\{X\})-\kappa(\boldsymbol{X})\right]\\&\qquad=|\boldsymbol{X}|\,\Pi\boldsymbol{X}-\sum_{\boldsymbol{X}''\subset\boldsymbol{X}'\subset\boldsymbol{X}}|\boldsymbol{X}''|:\boldsymbol{X}\setminus\boldsymbol{X}'\colon(\mathbf{E}\,\Pi(\boldsymbol{X}'\setminus\boldsymbol{X}''))\kappa(\boldsymbol{X}''),\end{aligned}$$

from which (2.93) follows using the well known cumulant identity

$$|\boldsymbol{X}'|\,\mathbf{E}\,\Pi\boldsymbol{X}'=\sum_{\boldsymbol{X}''\subset\boldsymbol{X}'}|\boldsymbol{X}''|\,(\mathbf{E}\,\Pi(\boldsymbol{X}'\setminus\boldsymbol{X}''))\kappa(\boldsymbol{X}''). \tag{2.94}$$

In order to prove (2.94), we use (2.87) on the rhs. to obtain

$$\begin{aligned}\sum_{\boldsymbol{X}''\subset\boldsymbol{X}'}|\boldsymbol{X}''|\,(\mathbf{E}\,\Pi(\boldsymbol{X}'\setminus\boldsymbol{X}''))\kappa(\boldsymbol{X}'')&=\sum_{\boldsymbol{X}''\subset\boldsymbol{X}'}|\boldsymbol{X}''|\,\kappa(\boldsymbol{X}'')\sum_{\mathcal{P}\vdash\boldsymbol{X}'\setminus\boldsymbol{X}''}\kappa^{\mathcal{P}}\\&=\sum_{\mathcal{P}\vdash\boldsymbol{X}'}\kappa^{\mathcal{P}}\sum_{\substack{\boldsymbol{X}''\subset\boldsymbol{X}'\\\boldsymbol{X}''\in\mathcal{P}}}|\boldsymbol{X}''|=|\boldsymbol{X}'|\sum_{\mathcal{P}\vdash\boldsymbol{X}'}\kappa^{\mathcal{P}},\end{aligned}$$

from which (2.94) follows by another application of (2.87).

Finally we remark that a quantitative variant of (2.92) for the pre-cumulants was centrally used in our proof in Section 2.4.2. Qualitatively the analogue of (2.92) for pre-cumulants reads

$$\mathbf{E}\,Y[K(X;\boldsymbol{X}_1,\boldsymbol{X}_2)-\kappa(X,\boldsymbol{X}_1,\boldsymbol{X}_2)]=0\quad\text{if}\quad\{X,\boldsymbol{X}_1\}\quad\text{is independent of}\quad\{\boldsymbol{X}_2,Y\}$$

and $\boldsymbol{X}_2$ is non-empty. Indeed, from the pre-cumulant decoupling identity (2.8c) we have that

$$\begin{aligned}\mathbf{E}\,Y[K(X;\boldsymbol{X}_1,\boldsymbol{X}_2)-\kappa(X,\boldsymbol{X}_1,\boldsymbol{X}_2)]&=\mathbf{E}\,Y(\Pi\boldsymbol{X}_2)[K(X;\boldsymbol{X}_1)-\kappa(X,\boldsymbol{X}_1)]\\&\quad-\sum_{\substack{\boldsymbol{X}_1'\subset\boldsymbol{X}_1\\\boldsymbol{X}_2'\subsetneq\boldsymbol{X}_2}}\mathbf{E}\,Y(\Pi\boldsymbol{X}_1')(\Pi\boldsymbol{X}_2')\kappa(X,\boldsymbol{X}_1\setminus\boldsymbol{X}_1',\boldsymbol{X}_2\setminus\boldsymbol{X}_2')\end{aligned}$$

and the first term vanishes due to independence and (2.8c), and the second term vanishes due to Lemma 2.A.1 because the argument of κ splits into two independent groups.

2.C Modifications for complex Hermitian W

Our main arguments were carried out for the real symmetric case. We now explain how to modify our proofs if W is complex Hermitian. A quick inspection of the proofs shows that the only modification concerns Proposition 2.3.2 where we have to replace the cumulant expansion by its complex variant. We reduce the problem to the real case by considering real and imaginary parts of each variable separately. Another option would have been to consider w and $\overline{w}$ independent variables, but our choice seems to require the least modifications. In order to compute $\mathbf{E}\, w_{i_0} f(\boldsymbol{w})$ for a random vector $\boldsymbol{w} \in \mathbb{C}^{\mathcal{I}}$, $w_{i_0} \in \mathbb{C}$ and a function $f\colon \mathbb{C}^{\mathcal{I}} \to \mathbb{C}$, we can define $\widetilde{f}\colon \mathbb{R}^{\mathcal{I}\sqcup\mathcal{I}} \to \mathbb{C}$ by mapping $(\boldsymbol{w}^{\Re}, \boldsymbol{w}^{\Im}) \mapsto f(\boldsymbol{w}^{\Re} + i\boldsymbol{w}^{\Im})$, where the new index set $\mathcal{I} \sqcup \mathcal{I}$ should be understood as two copies of $\mathcal{I}$ in the sense that $\mathcal{I} \sqcup \mathcal{I} = \{\, (i,\Re), (i,\Im) \mid i \in \mathcal{I} \,\}$. If we want to expand $w_{i_0} f(\boldsymbol{w})$ in the variables of some fixed index set $\mathcal{N} \subset \mathcal{I}$, we separately apply Proposition 2.3.2 to $\mathbf{E}\, \widetilde{w}_{(i_0,\Re)} \widetilde{f}(\widetilde{\boldsymbol{w}})$ and $\mathbf{E}\, \widetilde{w}_{(i_0,\Im)} \widetilde{f}(\widetilde{\boldsymbol{w}})$ in $\mathcal{N} \sqcup \mathcal{N}$, where $\widetilde{\boldsymbol{w}} = (\Re \boldsymbol{w}, \Im \boldsymbol{w})$ and $\widetilde{w}_{(i,\Re)} = \Re w_i$, $\widetilde{w}_{(i,\Im)} = \Im w_i$. It follows that

$$\mathbf{E}\, w_{i_0} \widetilde{f}(\widetilde{\boldsymbol{w}}) = \sum_{l>0} \sum_{\widetilde{\boldsymbol{i}} \in (\mathcal{N}\sqcup\mathcal{N})^l} \frac{\kappa(\widetilde{w}_{(i_0,\Re)}, \widetilde{\boldsymbol{w}}_{\widetilde{\boldsymbol{i}}}) + \kappa(\mathrm{i}\widetilde{w}_{(i_0,\Im)}, \widetilde{\boldsymbol{w}}_{\widetilde{\boldsymbol{i}}})}{l!} \partial_{\widetilde{\boldsymbol{i}}} (\mathbf{E}\, \widetilde{f}) + \widetilde{\Omega}^1 + \widetilde{\Omega}^2, \quad (2.95)$$

where the error terms are those from two applications of (2.12a). We note that we can make sense of κ with complex arguments directly through Definition (2.89). We now want to go back to a summation over our initial index set $\mathcal{N}$ and therefore regroup the terms in (2.95) according to the first indices of $\widetilde{\boldsymbol{i}}$. To formulate the result compactly we introduce the tensors

$$\widetilde{\kappa}(w_{i_0}, \dots, w_{i_l}) := \kappa\Big[\begin{pmatrix} \Re w_{i_0} \\ \mathrm{i}\Im w_{i_0} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} \Re w_{i_l} \\ \mathrm{i}\Im w_{i_l} \end{pmatrix} \Big] \in (\mathbb{R} \times \mathrm{i}\mathbb{R})^{\otimes(l+1)}$$

and

$$\widetilde{\partial}_{\boldsymbol{i}} := \begin{pmatrix} \partial_{\Re w_{i_1}} \\ \partial_{\Im w_{i_1}} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} \partial_{\Re w_{i_l}} \\ \partial_{\Im w_{i_l}} \end{pmatrix},$$

where the application of κ is understood in an entrywise sense and the derivative tensor has dimension $(\mathbb{C}^2)^{\otimes l}$. By saying that κ is understood in an entrywise sense, we mean, by slight abuse of notation that, for example,

$$\kappa\left(\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \otimes \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \right) = \kappa\Big(\sum_{i,j=1}^{2} v_i w_j \, e_i \otimes e_j \Big) := \sum_{i,j=1}^{2} \kappa(v_i, w_j) \, e_i \otimes e_j,$$

where e_1, e_2 is the standard basis of $\mathbb{R} \times \mathrm{i}\mathbb{R}$. Due to the special nature of the index i_0 we see from (2.95) that $\Re w_{i_0}$ and $\mathrm{i}\Im w_{i_0}$ always occur in a sum of two and the rhs. of (2.95) can be expressed in terms of the partial trace $\operatorname{Tr}_1 \widetilde{\kappa}(w_{i_0}, \dots, w_{i_l}) \in (\mathbb{R} \times i\mathbb{R})^{\otimes l}$ along the first dimension, which corresponds to i_0. Thus we can compactly write (2.95) as

$$\mathbf{E}\, w_{i_0} f(\boldsymbol{w}) = \sum_{0 \le l < R} \sum_{\boldsymbol{i} \in \mathcal{N}^l} \frac{\langle \operatorname{Tr}_1 \widetilde{\kappa}(w_{i_0}, \boldsymbol{w}_{\boldsymbol{i}}), \mathbf{E}(\widetilde{\partial}_{\boldsymbol{i}} f) \rangle}{l!} + \widetilde{\Omega}^1 + \widetilde{\Omega}^2, \qquad (2.96)$$

where the scalar product is taken between two tensors of size 2^l. For example, the $l = 1$ term from (2.96) reads

$$\sum_{i_1 \in \mathcal{N}} \Big(\frac{\kappa(\Re w_{i_0}, \Re w_{i_1}) + \kappa(\mathrm{i}\Im w_{i_0}, \Re w_{i_1})}{1!} (\mathbf{E}\, \partial_{\Re w_{i_1}} f) \\ + \frac{\kappa(\Re w_{i_0}, \mathrm{i}\Im w_{i_1}) + \kappa(\mathrm{i}\Im w_{i_0}, \mathrm{i}\Im w_{i_1})}{1!} (\mathbf{E}\, \partial_{\Im w_{i_1}} f) \Big).$$

The rest of the argument in Section 2.4 can be carried out verbatim for any specific choice of distribution of $\Re$, $\Im$ to the entries of κ. We only have to replace the norms $|||\kappa|||^{\mathrm{av}}$ and $|||\kappa|||^{\mathrm{iso}}$ in Assumption (2.C) by applying them entrywise to $\widetilde{\kappa}$, i.e.,

$$\begin{aligned}|||\widetilde{\kappa}(w_{\alpha_1},\dots,w_{\alpha_k})|||^{\mathrm{av}} &:= \sum_{\mathfrak{X}_1,\dots,\mathfrak{X}_k\in\{\Re,\Im\}} |||\kappa(\mathfrak{X}_1 w_{\alpha_1},\dots,\mathfrak{X}_k w_{\alpha_k})|||^{\mathrm{av}},\\ |||\widetilde{\kappa}(w_{\alpha_1},\dots,w_{\alpha_k})|||^{\mathrm{iso}} &:= \sum_{\mathfrak{X}_1,\dots,\mathfrak{X}_k\in\{\Re,\Im\}} |||\kappa(\mathfrak{X}_1 w_{\alpha_1},\dots,\mathfrak{X}_k w_{\alpha_k})|||^{\mathrm{iso}}.\end{aligned}$$

Assumption (2.C)' (Hermitian κ-correlation decay). *We assume that for all $R\in\mathbb{N}$ and $\epsilon>0$*

$$|||\widetilde{\kappa}|||^{av} \le_{\epsilon,R} N^{\epsilon} \quad \textit{and} \quad |||\widetilde{\kappa}|||^{iso} \le_{\epsilon,R} N^{\epsilon}.$$

Since there are at most 2^R such choices this change has no impact on any of the claimed bounds which always implicitly allow for an R–dependent constant.

2.D Proofs of auxiliary results

Lemma 2.D.1 (Quadratic Implicit Function Theorem). *Let $\|\cdot\|$ be a norm on $\mathbb{C}^d$, $A,B\in\mathbb{C}^d$ and $Q\colon \mathbb{C}^d\times\mathbb{C}^d\to\mathbb{C}^d$ a bounded $\mathbb{C}^d$-valued quadratic form, i.e.,*

$$\|Q\| = \sup_{x,y}\frac{\|Q(x,y)\|}{\|x\|\,\|y\|} < \infty.$$

Suppose that A is invertible. Then for $\epsilon_2 := \big[2\,\|A^{-1}\|\,\|Q\|\,\big]^{-1}$ and $\epsilon_1 := \epsilon_2\big[2\,\|A^{-1}\|\,\|B\|\,\big]^{-1}$ there is a unique function $X\colon B_{\epsilon_1}\to B_{\epsilon_2}$ such that

$$AX(d) + Q(X(d),X(d)) = Bd,$$

where B_ϵ denotes the open ϵ-ball around 0. Moreover, the function X is analytic and satisfies

$$\|X(d_1)-X(d_2)\| \le 2\left\|A^{-1}\right\|\|B\|\,\|d_1-d_2\| \quad \textit{for all} \quad d_1,d_2\in B_{\epsilon_1/2}.$$

Proof. A simple application of the Banach fixed point theorem. □

Lemma 2.D.2. *For random matrices R,T and $p\ge 1$ it holds that*

$$\|\mathcal{S}[V]T\|_p \le |||\mathcal{S}|||\,\|V\|_{2p}\,\|T\|_{2p}\,.$$

Proof. Let $\kappa=\kappa_c+\kappa_d$ be an arbitrary partition, which induces a partition of $\mathcal{S}$ since

$$\mathcal{S}[V] = \frac{1}{N}\sum_{\alpha_1,\alpha_2}\kappa(\alpha_1,\alpha_2)\Delta^{\alpha_1}V\Delta^{\alpha_2}.$$

For vectors $\mathbf{x},\mathbf{y}$ with $\|\mathbf{x}\|,\|\mathbf{y}\|\le 1$ we compute

$$\begin{aligned}\|(\mathcal{S}[V]T)\mathbf{x}\mathbf{y}\|_p &= \Big\|\frac{1}{N}\sum_{b_1,a_2,b_2}\kappa(\mathbf{x}b_1,a_2b_2)V_{b_1a_2}T_{b_2\mathbf{y}}\Big\|_p\\ &\le \Big\|\frac{1}{N}\sum_{b_1,b_2}V_{b_1\kappa_c(\mathbf{x}b_1,\cdot b_2)}T_{b_2\mathbf{y}}\Big\|_p + \Big\|\frac{1}{N}\sum_{b_1,a_2}R_{b_1a_2}T_{\kappa_d(\mathbf{x}b_1,a_2\cdot)\mathbf{y}}\Big\|_p\\ &\le \frac{\|V\|_{2p}\,\|T\|_{2p}}{N}\Big[\sum_{b_1,b_2}\|\kappa_d(\mathbf{x}b_1,\cdot b_2)\| + \sum_{b_1,a_2}\|\kappa_c(\mathbf{x}b_1,a_2\cdot)\|\,\Big]\\ &\le \Big[\,|||\kappa_d|||_d + |||\kappa_c|||_c\Big]\,\|V\|_{2p}\,\|T\|_{2p}\end{aligned}$$

and the result follows from optimizing over the decompositions of κ and recalling the definition (2.24). □

Lemma 2.D.3. *For any $t \in [0,1]$, $q \geq 1$ and multi-set $\underline{\beta} \subset I$ we have under Assumption (2.A) that*

$$\begin{aligned}
\left\|\partial_{\underline{\beta}} G|_{\widehat{W}}\right\|_q &\leq_{|\underline{\beta}|} N^{-|\underline{\beta}|/2}\Big(1 + \|G\|_{6q(|\underline{\beta}|+1)}^{|\underline{\beta}|+1}\Big) \\
\left\|\partial_{\underline{\beta}} D|_{\widehat{W}}\right\|_q &\leq_{|\underline{\beta}|} N^{-|\underline{\beta}|/2}(1 + |||\mathcal{S}|||)(1 + |z|\, \|G\|_{12q|\underline{\beta}|+2}^{3})\Big(1 + \|G\|_{12q(|\underline{\beta}|+2)}^{|\underline{\beta}|+2}\Big),
\end{aligned} \tag{2.97}$$

where $\widehat{W}_\alpha = tw_\alpha$ for $\alpha \in \mathcal{N}$ and $\widehat{W}_\alpha = w_\alpha$ otherwise for a set $\mathcal{N} \subset I$ of size $|\mathcal{N}| \leq N^{1/2}$.

Proof. We write $\underline{\beta} = \{\beta_1, \dots, \beta_n\}$ and its easy to see inductively that

$$\partial_{\underline{\beta}} G|_{\widehat{W}} = \frac{(-1)^n}{N^{n/2}} \sum_{\sigma \in S_n} \widehat{G}\Delta^{\beta_{\sigma(1)}}\widehat{G}\Delta^{\beta_{\sigma(2)}}\widehat{G}\dots\widehat{G}\Delta^{\beta_{\sigma(n)}}\widehat{G},$$

where $\widehat{G} = G(\widehat{W})$. From the resolvent identity it follows that

$$\begin{aligned}
\widehat{G} - G = {}& \frac{1}{\sqrt{N}} G(W - \widehat{W})G + \frac{1}{N} G(W - \widehat{W})G(W - \widehat{W})G \\
&+ \frac{1}{N^{3/2}} \widehat{G}(W - \widehat{W})G(W - \widehat{W})G(W - \widehat{W})G
\end{aligned}$$

and therefore by the trivial bound $\|\widehat{G}\| \leq 1/\eta$ and Assumption (2.B) it follows that

$$\begin{aligned}
\left\|\widehat{G} - G\right\|_q \leq {}& \frac{|\mathcal{N}|\, \|G\|_{3q}^2 \max_\alpha \|w_\alpha\|_{3q}}{\sqrt{N}} + \frac{|\mathcal{N}|^2\, \|G\|_{5q}^3 \max_\alpha \|w_\alpha\|_{5q}^2}{N} \\
&+ \frac{|\mathcal{N}|^3\, \|G\|_{6q}^3 \max_\alpha \|w_\alpha\|_{6q}^3}{N^{3/2}\eta} \leq_q (1 + \|G\|_{6q}^3)
\end{aligned}$$

and therefore also $\|\widehat{G}\|_q \leq_q (1 + \|G\|_{6q}^3)$, from which the first inequality in (2.97) follows immediately.

Similarly, the second inequality in (2.97) follows from the easily verifiable identity

$$\begin{aligned}
\partial_{\underline{\beta}} D|_{\widehat{W}} = \frac{(-1)^n}{N^{n/2}} \sum_{\sigma \in S_n} \Big[& \widehat{D}\Delta^{\beta_{\sigma(1)}}\widehat{G}\dots\Delta^{\beta_{\sigma(n)}}\widehat{G} \\
&+ \sum_{k=1}^n \mathcal{S}[\widehat{G}\Delta^{\beta_{\sigma(1)}}\widehat{G}\dots\Delta^{\beta_{\sigma(k)}}\widehat{G}]\widehat{G}\Delta^{\beta_{\sigma(k+1)}}\widehat{G}\dots\Delta^{\beta_{\sigma(n)}}\widehat{G}\Big]
\end{aligned}$$

and

$$\|\widehat{D}\|_q \leq C(1 + |z|\, \|G\|_{6q}^3) + \|\mathcal{S}[\widehat{G}]\widehat{G}\|_q \tag{2.98}$$

together with Lemma 2.D.2. To see why (2.98) holds we write $D = (1+z-A)G + \mathcal{S}[G]G$, so that $\|\widehat{D}\|_q \leq (1+|z|\,\|\widehat{G}\|_q) + \|\mathcal{S}[\widehat{G}]\widehat{G}\|_q$ holds uniformly for $\eta \geq N^{-1}$ for some constant C. □

Correlated Random Matrices: Band Rigidity and Edge Universality 3

We prove edge universality for a general class of correlated real symmetric or complex Hermitian Wigner matrices with arbitrary expectation. Our theorem also applies to internal edges of the self-consistent density of states. In particular, we establish a strong form of band rigidity which excludes mismatches between location and label of eigenvalues close to internal edges in these general models.

Published as J. Alt, L. Erdős, T. Krüger, and D. Schröder, *Correlated random matrices: Band rigidity and edge universality*, preprint (2018), arXiv:1804.07744.

3.1 Introduction

Spectral statistics of large random matrices exhibit a remarkably robust universality pattern; the local distribution of eigenvalues is independent of details of the matrix ensemble up to symmetry type. In the bulk of the spectrum this was first observed by Wigner and formalized by Dyson and Mehta [135] who also computed the correlation functions of the Gaussian ensembles in the 1960's. At the spectral edges the correct statistics was identified by Tracy and Widom both in the GUE and GOE ensembles [170, 171] in the mid 1990's.

Beyond Gaussian ensembles, the first actual proofs of universality for Wigner matrices took different paths in the bulk and at the edge. While in the bulk only limited progress was made until a decade ago, the first fairly general edge universality proof by Soshnikov [161] appeared shortly after [170, 171]. The main reason is that edge statistics is accessible via an ingenious but laborious extension of the classical moment method of Wigner. In contrast, the bulk universality required fundamentally new tools based on resolvents and the analysis of the Dyson Brownian motion developed in a series of work [74, 75, 82, 72, 71, 79]. This method, called the *three-step strategy*, is summarized in [78]. In certain cases parallel results [168, 167] were obtained via the *four moment comparison theorem*.

Despite its initial success [161], the moment method for edge universality seems limited when it comes to generalisations beyond Wigner matrices with i.i.d. entries; the resolvent approach is much more flexible. Its primary goal is to establish *local laws*, i.e. proving that the local eigenvalue density on scales slightly above the eigenvalue spacing becomes deterministic as the dimension of the matrix tends to infinity. Refined versions of the local law even identify resolvent matrix elements with a spectral parameter very close to the real axis. In contrast to the bulk, at the spectral edge this information can be boosted to detect individual eigenvalue statistics by comparison with the Gaussian ensemble. These ideas have led to the proof of the Tracy-Widom edge universality for Wigner matrices with high moment conditions [82], see also [168] with vanishing third moment. Finally, a necessary and sufficient condition on the entry distributions was found in [127] following an almost optimal necessary condition in [17]. Direct resolvent comparison methods have been used to prove Tracy-Widom universality for *deformed Wigner matrices*, i.e. matrices with a deterministic diagonal expectation, [123], even in a certain sparse regime [124]. The extension of this approach to sample covariance matrices with a diagonal population covariance matrix at extreme edges [125] has resolved a long standing conjecture in the statistics literature. Tracy-Widom universality for general population covariance matrices, including internal edges, was established in [117].

The next level of generality is to depart from the i.i.d. case. While the resolvent method for proving local laws can handle *generalized Wigner ensemble*, i.e. matrices $H = (h_{ab})$ with merely stochastic variance profile $\sum_b \mathbf{Var}\, h_{ab} = 1$, varying variances cannot be simultaneously matched with a GUE/GOE ensemble so the direct comparison does not work. The problem was resolved in [41] with a general approach that also covered invariant β-ensembles. While Dyson Brownian motion did not play a direct role in [41], the proof used the addition of a small Gaussian component and the concept of local ergodicity of the Gibbs state; ideas developed originally in [75, 76] in the context of bulk universality.

A fully dynamical approach to edge universality, following an earlier development in the bulk based on the *three-step strategy*, has recently been given in [122]. In general, the first step within any three-step strategy is the local law providing a priori bounds. The second step is the fast relaxation to equilibrium of the Dyson Brownian motion that proves universality for Gaussian divisible ensembles. The third step is a perturbative comparison argument to remove the small Gaussian component. Recent advances in the bulk have crystallized that the only model dependent step in this strategy is the first one. The other two steps have been formulated as very general "black-box" tools whose only input is the local law see [120, 122, 77, 121]. Using the three-step approach and [122], edge universality for sparse matrices was proved in [104] and for correlated Gaussian matrices with a quite specific two-scale correlation structure in [1]. All these edge universality results only cover the *extremal edges* of the spectrum, while the self-consistent (deterministic) density of states ϱ may be supported on several intervals.

Multiple interval support becomes ubiquitous for *Wigner-type* matrices [9], i.e. matrices with independent entries and general expectation and variance profile. A prerequisite for Tracy-Widom universality, the square root singularity in the density, even at the *internal edges*, is a universal phenomenon for a very large class of random matrices since it is inherent to the underlying *Dyson equation*. This was demonstrated for Wigner-type matrices in [7] and here we extend it for correlated random matrices with a general correlation structure. We remark that a second singularity type, the *cubic root cusp*, is also possible; the corresponding analysis of the Dyson equation is given in [12], while the optimal local law and the universal spectral statistics are proven in [DS5, DS6].

In the current paper we show that the eigenvalue statistics at the spectral edges of ϱ

follow the Tracy-Widom distribution, assuming only a mild decay of correlation between entries, but otherwise no special structure. We can handle any internal edge as well. In the literature internal edge universality for matrices of Wigner-type has first been established for deformed GUE ensembles [158] which critically relied on contour integral methods, only available for Gaussian models in the Hermitian symmetry class. A similar method handled extreme eigenvalues of deformed GUE [107, 52]. A more general approach for internal edges has been given in [117] that could handle any deformed Wigner matrices with general expectation, as long as the variance profile is constant, by comparing it with the corresponding Gaussian model. Our method requires neither constant variance nor independence of the matrix elements.

The proof of our general form of edge universality at all internal edges follows the three-step strategy and uses the recent paper [122] for the second step and well established canonical arguments for the third step that will be summarized. The backbone of the work is thus the first step, an optimal local law at the spectral edges, the proof of which has two well separated components; a probabilistic and a deterministic one. The probabilistic component is insenstive to the location in the spectrum and follows directly from [DS3]. Here we present a compact and practically self-contained proof of the deterministic component of the local law that can be followed without consulting previous works; we only rely on some general results from functional analysis proven in [8] and some minor technicality on the Dyson equation from [12]. First, we develop a detailed shape analysis of the self-consistent density ϱ near the regular edges, generalizing the previous bulk result from [8] and the singularity analysis in the independent case from [7]. Second, we prove a strong version of the local law that excludes eigenvalues in the internal gaps. Third, we establish a topological rigidity phenomenon for the *bands*, the connected components that constitute the support of ϱ.

Band rigidity is a new phenomenon for the Dyson equation and it asserts that the number of eigenvalues within each band exactly matches the mass that ϱ predicts for that band. The topological nature of band rigidity guarantees that this mass remains constant along the deformations of the model as long as the gaps between the bands remain open. A similar rigidity (also called "exact separation of eigenvalues") has first been established for sample covariance matrices in [18] and it also played a key role in Tracy-Widom universality proof at internal edges in [117]. Note that band rigidity is a much stronger concept than the customary rigidity in random matrix theory [82] that allows for an uncertainty in the location of N^{ϵ} eigenvalues. In other words, there is no mismatch whatsoever between location and label of the eigenvalues near the internal edges along the matrix Dyson Brownian motion, the label of the eigenvalue uniquely determines to which spectral band it belongs.

Our result highlights a key difference between Wigner-type matrix models and invariant β-ensembles. For self-consistent densities with multiple support intervals (the so called *multi-cut* regime), the number of particles (eigenvalues) close to some support interval fluctuates for invariant ensembles with general potentials [39]. As a consequence internal edge universality results (see e.g. [145, 29]) require a stochastic relabelling of eigenvalues.

Our setup is a general $N \times N$ random matrix $H = H^*$ with a slowly decaying correlation structure and arbitrary expectation, under the very same general conditions as the recent bulk universality result from [DS3]. The starting point is to find the deterministic approximation of the resolvent $G(z) = (H - z)^{-1}$ with a complex spectral parameter z in the upper half plane. This approximation is given by the solution $M = M(z)$ to the *Matrix Dyson Equation (MDE)*, see (3.1) below. The resolvent $G(z)$ approximately satisfies the MDE with an additive perturbation term which was already shown to be sufficiently small in [DS3]. This fact, combined with a careful stability and shape analysis of the MDE in Section 3.4 imply that G is indeed close to M. In order to prove edge universality we use a correlated Ornstein-Uhlenbeck process H_t which adds a small Gaussian component of size t to the original matrix model, while preserving expectation and covariance. We prove that the resolvent satisfies the optimal local law uniformly along the flow and appeal to the recent result from [122] to prove edge universality for H_t whenever $t \gg N^{-1/3}$. In the final step we perform a resolvent comparison together with our band rigidity to show that the eigenvalue correlation functions of H_t matches those of H as long as $t \ll N^{-1/6}$ which yields the desired edge universality.

After presenting our main results in Section 3.2, we then prove the optimal local law in Section 3.3. Section 3.4 contains the analysis of the MDE. Both types of rigidity are shown in Section 3.5. Section 3.6 is devoted to the proof of edge universality.

Notations

If for some constants $c, C > 0$ it holds that $f \le Cg$ or $cg \le f \le Cg$, then we write $f \lesssim g$ and $f \sim g$, respectively. These constants c, C may depend on some basic parameters which we call model parameters later. We denote vectors by bold-faced lower case Roman letters $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$, and matrices by upper case Roman letters $A, B \in \mathbb{C}^{N\times N}$. The standard scalar product and Euclidean norm on $\mathbb{C}^N$ will be written as $\langle \mathbf{x}, \mathbf{y}\rangle$ and $\|\mathbf{x}\|$, while we also write $\langle A, B\rangle := N^{-1} \operatorname{Tr} A^* B$ for the scalar product of matrices, and $\langle A\rangle := N^{-1} \operatorname{Tr} A$. The usual operator norm induced by the vector norm $\|\cdot\|$ will be denoted by $\|A\|$, while the Hilbert-Schmidt (or Frobenius) norm will be denoted by $\|A\|_{\text{hs}} := \sqrt{\langle A, A\rangle}$. The operator norms induced on linear maps $\mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ by $\|\cdot\|_{\text{hs}}$ and $\|\cdot\|$ are denoted by $\|\cdot\|_{\text{sp}}$ and $\|\cdot\|$, respectively. The identity matrix in $\mathbb{C}^{N\times N}$ is indicated by I and the identity mapping on $\mathbb{C}^{N\times N}$ by Id. For random variables $X, Y, \dots$ we denote the joint cumulant by $\kappa(X, Y, \dots)$. For integers n we define $[n] := \{1, \dots, n\}$.

3.2 Main results

We consider correlated real symmetric and complex Hermitian random matrices of the form

$$H = A + W, \qquad \mathbf{E}\, W = 0$$

with deterministic $A \in \mathbb{C}^{N\times N}$ and sufficiently fast decaying correlations among the matrix elements of W. The matrix entries $w_{ab} = w_\alpha$ are often labelled by double indices $\alpha = (a, b) \in [N]^2$. The randomness W is scaled in such a way that $\sqrt{N} w_\alpha$ are random variables of order one[1]. This requirement ensures that the size of the spectrum of H is kept of order 1, as N tends to infinity. Our first aim is to prove that the resolvent $G = G(z) = (H - z)^{-1}$ is well approximated by the solution $M = M(z)$ to the *Matrix Dyson equation (MDE)*

$$\begin{aligned} I + (z - A + \mathcal{S}[M])M = 0, \quad \Im M := \frac{M - M^*}{2\mathrm{i}} > 0, \\ \mathcal{S}[R] := \mathbf{E}\, W R W, \quad z \in \mathbb{H} := \{\, z \in \mathbb{C} \mid \Im z > 0 \,\} \end{aligned} \tag{3.1}$$

in a neighbourhood around the edges of the spectrum. We suppress the dependence of G and M, and similarly of many other quantities, on the spectral parameter z in our notation.

[1] In some previous works, as in [DS3], the convention $H = A + W/\sqrt{N}$ with order one w_α was used.

Estimates on z-dependent quantities are always meant uniformly for z in some specified domain. From the solution M we define $\varrho\colon \mathbb{H} \to \mathbb{R}$ and extend it to the real line

$$\varrho(z) := \frac{1}{\pi}\Im \langle M(z)\rangle, \quad z \in \mathbb{H}, \qquad \varrho(\tau) := \lim_{\eta \searrow 0} \varrho(\tau + \mathrm{i}\eta), \quad \tau \in \mathbb{R}. \tag{3.2}$$

By [8, Proposition 2.2] the limit in (3.2) exists and ϱ is a Hölder continuous function on $\mathbb{H} \cup \mathbb{R}$ under Assumptions (3.A) and (3.E) below. The *self-consistent density of states* is the restriction of ϱ to $\mathbb{R}$ which approximates the density of eigenvalues of H increasingly well as N tends to infinity. Its support, $\operatorname{supp}\varrho \subset \mathbb{R}$, is called the *self-consistent spectrum*. We remark that ϱ on $\mathbb{H}$ is the harmonic extension of $\varrho|_{\mathbb{R}}$. We now list our main assumptions, which are identical to those from [DS3], apart from the additional Assumption (3.G), which was automatically satisfied in [DS3], i.e. in the bulk regime (cf. Remark 3.2.3 below). All constants in Assumptions (3.A)–(3.G) and Definition 3.2.4 are called *model parameters*.

Assumption (3.A) (Bounded expectation). *There exists some constant C such that $\|A\| \le C$ for all N.*

Assumption (3.B) (Finite moments). *For all $q \in \mathbb{N}$ there exists a constant μ_q such that*

$$\mathbf{E}|\sqrt{N}w_\alpha|^q \le \mu_q$$

for all α.

Assumption (3.CD) (Polynomially decaying metric correlation structure). *For the $k = 2$ point correlation we assume*

$$\left|\kappa\Big(f_1(\sqrt{N}W), f_2(\sqrt{N}W)\Big)\right| \le C_2 \frac{\sqrt{\mathbf{E}\,|f_1(\sqrt{N}W)|^2}\sqrt{\mathbf{E}\,|f_2(\sqrt{N}W)|^2}}{1 + d(\operatorname{supp} f_1, \operatorname{supp} f_2)^s}, \tag{3.3}$$

for some $s > 12$ and all square integrable functions f_1, f_2. For $k \ge 3$ we assume a decay condition of the form

$$\left|\kappa\Big(f_1(\sqrt{N}W), \dots, f_k(\sqrt{N}W)\Big)\right| \le C_k \prod_{e \in E(T_{min})} |\kappa(e)|,$$

where T_{min} is the minimal spanning tree in the complete graph on the vertices $1, \dots, k$ with respect to the edge length $\operatorname{dist}(\{i,j\}) = d(\operatorname{supp} f_i, \operatorname{supp} f_j)$, i.e. the tree for which the sum of the lengths $\operatorname{dist}(e)$ is minimal, and $\kappa(\{i,j\}) = \kappa(f_i, f_j)$. Here d is the standard Euclidean metric on the index space $[N]^2$ and $\operatorname{supp} f \subset [N]^2$ denotes the set indexing all entries in $\sqrt{N}W$ that f genuinely depends on, and $C_k < \infty$ are some absolute constants.

Remark 3.2.1. *All results in this paper and their proofs hold verbatim if Assumption (3.CD) is replaced by the more general Assumptions (2.C)–(2.D). In particular, the metric structure imposed on the index space $[N]^2$ is not essential. For details the reader is referred to Section 2.2.5.*

Assumption (3.E) (Flatness). *There exist constants $0 < c < C$ such that $c\langle T\rangle \le \mathcal{S}[T] \le C\langle T\rangle$ for any positive semi-definite matrix T.*

Assumption (3.F) (Fullness). *There exists a constant* $\lambda > 0$ *such that* $N\,\mathbf{E}\,|\mathrm{Tr}\,BW|^2 \geq \lambda\,\mathrm{Tr}\,B^2$ *for any deterministic matrix* B *of the same symmetry class (either real symmetric or complex Hermitian) as* H.

Assumption (3.G) (Bounded self-consistent Green function). *There exist constants* $\omega_*, M_* > 0$ *such that*

$$\sup_z \|M(z)\| \leq M_*,$$

where the supremum is taken over all $z \in \mathbb{H}$ *with* $|\Re z - \tau_0| \leq \omega_*$ *and* $0 < \Im z \leq 1$.

Remark 3.2.2. *Assumption (3.E) is an effective mean field condition that provides upper and lower bounds on the variances of the entries of* W. *In fact it is equivalent to* $\mathbf{E}\,|\langle \mathbf{x}, W\mathbf{y}\rangle|^2 \sim 1/N$ *for all normalised* $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$. *Assumption (3.F) is equivalent to* $\mathcal{S} - \lambda\mathcal{S}_{\mathrm{G}}$ *remaining positivity preserving, where* $\mathcal{S}_{\mathrm{G}}$ *is the self-energy operator of a full GUE/GOE matrix.*

Remark 3.2.3. *The boundedness of* $\|M\|$ *is automatically satisfied in the spectral bulk. At the edges, however, the boundedness cannot be guaranteed under Assumptions (3.A)–(3.E) but has to be verified for each concrete model (see [12, Section 9] for a large class of models for which* $\|M\|$ *is guaranteed to be bounded).*

Our main technical result is an optimal local law at *regular edges* $\tau_0 \in \partial\,\mathrm{supp}\,\varrho$ asserting that $G(z) = (H-z)^{-1}$ is well approximated by $M(z)$ in the $N \to \infty$ limit. Around such an edge we consider the domain of spectral parameters $z = \tau + \mathrm{i}\eta$ whose imaginary part $\Im z = \eta$ is slightly larger than $1/N$, i.e. in the spectral domain

$$\mathbb{D}^\delta_\gamma := \left\{\, z \in \mathbb{D}^\delta \;\middle|\; \Im z \geq N^{-1+\gamma} \,\right\} \text{ with } \mathbb{D}^\delta := \{\, \tau + \mathrm{i}\eta \mid |\tau - \tau_0| \leq \delta\,, 0 < \eta \leq 1 \,\}$$

for any $\gamma, \delta > 0$.

Definition 3.2.4 (Regular edge). *We call an edge* $\tau_0 \in \partial\,\mathrm{supp}\,\varrho$ regular *if the limit*

$$\lim_{\mathrm{supp}\,\varrho \ni \tau \to \tau_0} \frac{\varrho(\tau)}{\sqrt{|\tau - \tau_0|}} = \frac{\gamma_{\mathrm{edge}}^{3/2}}{\pi} \tag{3.4}$$

exists for some slope parameter γ_{edge} *that satisfies* $0 < c_* \leq \gamma_{\mathrm{edge}} \leq c^* < \infty$ *for some constants* c_*, c^*.

Remark 3.2.5. *We remark that there are several equivalent characterisations of* regular edges. *We chose* (3.4) *here because it highlights that the essential prerequisite for Tracy-Widom universality is a local square-root singularity. According to the classification result from [12] it follows that* (3.4) *is equivalent*[2] *to assuming that the gap in* $\mathrm{supp}\,\varrho$ *adjacent to* τ_0 *is of size* $\gtrsim 1$.

Theorem 3.2.6 (Edge local law). *Let Assumptions (3.A)–(3.E) and (3.G) be satisfied for some regular edge* $\tau_0 \in \partial\,\mathrm{supp}\,\varrho$. *Then for any* $D, \gamma, \epsilon > 0$ *and sufficiently small* $\delta > 0$, *there exists*

[2] In fact, in [12, Section 7.6] it is proven that if the self-consistent spectrum $\mathrm{supp}\,\varrho$ has a macroscopic gap next to some $\tau_0 \in \partial\,\mathrm{supp}\,\varrho$, then ϱ has a square root behaviour at τ_0. Together with Theorem 3.4.1 later, this shows that regular edges in the sense of (3.4) are precisely those $\tau_0 \in \partial\,\mathrm{supp}\,\varrho$ which are adjacent to macroscopic gaps.

some $C < \infty$ depending only on these and the model parameters such that with $G = G(z)$ and $M = M(z)$ we have the isotropic local law,

$$\mathbf{P}\left(|\langle \mathbf{x}, (G-M)\mathbf{y}\rangle| \leq N^\epsilon \|\mathbf{x}\| \|\mathbf{y}\| \left(\sqrt{\frac{\varrho}{N\Im z}} + \frac{1}{N\Im z}\right) \quad in \quad \mathbb{D}^\delta_\gamma \right) \geq 1 - CN^{-D} \tag{3.5a}$$

for all deterministic vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$ and the averaged local law,

$$\mathbf{P}\left(|\langle B(G-M)\rangle| \leq N^\epsilon \frac{\|B\|}{N\Im z} \quad in \quad \mathbb{D}^\delta_\gamma \right) \geq 1 - CN^{-D} \tag{3.5b}$$

for all deterministic matrices $B \in \mathbb{C}^{N\times N}$. Moreover, at a distance at least $N^{-2/3+\epsilon}$ away from the self-consistent spectrum we have the improved averaged local law for any $\epsilon > 0$

$$\begin{aligned}\mathbf{P}\Bigg(|\langle B(G-M)\rangle| \leq \frac{N^\epsilon \|B\|}{N \operatorname{dist}(z, \operatorname{supp}\varrho)} \quad in \quad \left\{ z \in \mathbb{D}^\delta \;\middle|\; \frac{\operatorname{dist}(z,\operatorname{supp}\varrho)}{N^{-2/3+\epsilon}} \geq 1 \right\}\Bigg) \\ \geq 1 - CN^{-D}\end{aligned} \tag{3.5c}$$

with C also depending on ϵ.

Corollary 3.2.7 (No eigenvalues outside the support of the self-consistent density). *Under the assumptions of Theorem 3.2.6 we have for any $\epsilon, D > 0$ and sufficiently small $\delta > 0$*

$$\mathbf{P}\left(\exists \lambda \in \operatorname{Spec} H \mid |\tau_0 - \lambda| \leq \delta\,,\ \operatorname{dist}(\lambda, \operatorname{supp}\varrho) \geq N^{-2/3+\epsilon} \right) \leq_{\epsilon,D} N^{-D},$$

where $\leq_{\epsilon,D}$ means a bound up to some multiplicative constant $C = C(\epsilon, D)$.

Corollary 3.2.8 (Delocalisation). *Under the assumptions of Theorem 3.2.6 it holds for an ℓ^2-normalized eigenvector $\mathbf{u}$ corresponding to an eigenvalue λ of H close to the edge τ_0 that*

$$\sup_{\|\mathbf{x}\|=1} \mathbf{P}\left(|\langle \mathbf{x}, \mathbf{u}\rangle| \geq \frac{N^\epsilon}{\sqrt{N}} \mid H\mathbf{u} = \lambda\mathbf{u},\ \|\mathbf{u}\| = 1,\ |\tau_0 - \lambda| \leq \delta \right) \leq_{\epsilon,D} N^{-D}$$

for any $\epsilon, D > 0$ and sufficiently small $\delta > 0$.

Corollary 3.2.9 (Band rigidity and eigenvalue rigidity). *Under the assumptions of Theorem 3.2.6 the following holds. For any $\epsilon, D > 0$ there exists some $C < \infty$ such that for any $\tau \in \mathbb{R} \setminus$ $\operatorname{supp}\varrho$ with $\operatorname{dist}(\tau, \operatorname{supp}\varrho) \geq \epsilon$ the number of eigenvalues less than τ is with high probability deterministic, i.e. that*

$$\mathbf{P}\left(|\operatorname{Spec} H \cap (-\infty, \tau)| = N \int_{-\infty}^{\tau} \varrho(x)\,\mathrm{d}x \right) \geq 1 - CN^{-D}. \tag{3.6a}$$

We also have the following strong form of eigenvalue rigidity in a neighbourhood of a regular edge τ_0. Let $\lambda_1 \leq \cdots \leq \lambda_N$ be the ordered eigenvalues of H and denote the index of the N-quantile close to energy $\tau \in \operatorname{int}(\operatorname{supp}\varrho)$ by $k(\tau) := \lceil N \int_{-\infty}^{\tau} \varrho(x)\,\mathrm{d}x\rceil$. It then holds that

$$\mathbf{P}\left(\sup_\tau \left| \lambda_{k(\tau)} - \tau \right| \geq \min\left\{ \frac{N^\epsilon}{N|\tau - \tau_0|^{1/2}}, \frac{N^\epsilon}{N^{2/3}} \right\} \right) \leq_{\epsilon,D} N^{-D} \tag{3.6b}$$

for any $\epsilon, D > 0$ and sufficiently small $\delta > 0$, where the supremum is taken over all $\tau \in \operatorname{supp}\varrho$ such that $|\tau - \tau_0| \leq \delta$.

Remark 3.2.10 (Integer mass). *Note that* (3.6a) *entails the non trivial fact that for* $\tau \notin \operatorname{supp} \varrho$, $N \int_{-\infty}^{\tau} \varrho(x) \, \mathrm{d}x$ *is always an integer, see Proposition 3.5.1 below. Moreover, it then trivially implies that* $N \int_a^b \varrho(x) \, \mathrm{d}x$ *is an integer for each* spectral band $[a, b]$, *i.e. connected component of* $\operatorname{supp} \varrho$. *Finally,* (3.6a) *also shows that the number of eigenvalues in each band is given by this integer with overwhelming probability. This is in sharp contrast to invariant* β*-ensembles where no such mechanism is present. For example, for an odd number of particles in a symmetric double-well potential,* $N \int_{-\infty}^{0} \varrho(x) \, \mathrm{d}x = N/2$ *is a half integer.*

The main application of the optimal local law from Theorem 3.2.6 is edge universality, as stated in the following theorem, generalising several previous edge universality results listed in the introduction. For definiteness we only state and prove the result for regular right edges. The corresponding statement for left edges can be proven along the same lines.

Theorem 3.2.11 (Edge universality). *Under the Assumptions (3.A)–(3.G) the following statement holds true. Assume that* $\tau_0 \in \partial \operatorname{supp} \varrho$ *is a right regular edge of* ϱ *with* slope parameter γ_{edge} *as in Definition 3.2.4. The integer (see Remark 3.2.10)* $i_0 := N \int_{-\infty}^{\tau_0} \varrho(x) \, \mathrm{d}x$ *labels the largest eigenvalue* λ_{i_0} *close to the band edge* τ_0 *with high probability. Furthermore, for test functions* $F \colon \mathbb{R}^{k+1} \to \mathbb{R}$ *such that* $\|F\|_\infty + \|\nabla F\|_\infty \le C < \infty$ *we have*

$$\Big| \mathbf{E}\, F\Big(\gamma_{\mathrm{edge}} N^{2/3}(\lambda_{i_0} - \tau_0), \dots, \gamma_{\mathrm{edge}} N^{2/3}(\lambda_{i_0-k} - \tau_0)\Big) \\ - \mathbf{E}\, F\Big(N^{2/3}(\mu_N - 2), \dots, N^{2/3}(\mu_{N-k} - 2)\Big) \Big| \lesssim N^{-c}$$

for some $c = c_k > 0$. *Here* $\mu_1, \dots, \mu_N$ *are the eigenvalues of a standard GUE/GOE matrix, depending on the symmetry class of* H.

From Theorem 3.2.11 we can immediately conclude that the eigenvalues of H near the regular edges follow the Tracy-Widom distribution. We remark that the direct analogue of Theorem 3.2.11 does not hold true for invariant β-ensembles with a *multi-cut* density. This is due to the fact that the number of particles close to a band of the self-consistent density, commonly known as the *filling fraction*, is known to be a fluctuating quantity for general classes of potentials. We refer the reader to [33] for a description of this phenomenon, to [142, 157] for non-Gaussian linear statistics in the multi-cut regime and to [39] for results on the fluctuations of filling fractions. Variants of Theorem 3.2.11 which allow for a relabelling of eigenvalues for invariant β-ensembles can be found in [145, 29].

3.3 Proof of the local law

The proof of a local law consists of three largely separate arguments. The first part concerns the analysis of the *stability operator*

$$\mathcal{B}[R] := R - M\mathcal{S}[R]M \tag{3.7}$$

for $R \in \mathbb{C}^{N \times N}$, and shape analysis of the solution M to (3.1). The second part is proving that the resolvent G is indeed an approximate solution to (3.1) in the sense that

$$D := I + (z - A + \mathcal{S}[G])G = WG + \mathcal{S}[G]G \tag{3.8}$$

is small. Finally, the third part consists of a bootstrap argument starting in the domain $\mathbb{D}_1^\delta$ and iteratively increasing the domain to $\mathbb{D}_\gamma^\delta$ while maintaining the desired bound on $G - M$.

3.3.1 Stability

From (3.1) and (3.8), we see that the difference between G and M is described by the relation

$$\mathcal{B}[G-M] = -MD + M\mathcal{S}[G-M](G-M). \tag{3.9}$$

To prove estimates on $G-M$ we need to analyse $\mathcal{B}$, the stability operator. Near the edge we will demonstrate that $\mathcal{B}$ has a very small (in absolute value) simple eigenvalue, that we will denote by β, and it turns out that β is well separated away from the rest of the spectrum of $\mathcal{B}$. Let P and B denote the corresponding left and right eigenvectors of $\mathcal{B}$, i.e. $\mathcal{B}^*[P] = \bar{\beta}P$ and $\mathcal{B}[B] = \beta B$, and we will specify their normalisation later. Note that $\mathcal{B}$ is typically not self-adjoint, so $P \neq B$. Since β is small, $\mathcal{B}^{-1}$ is unstable in the direction of the eigenspace of β. We therefore separate this unstable direction by writing $G - M = \Theta B + \mathit{Error}$ where

$$\Theta := \frac{\langle P, G-M\rangle}{\langle P, B\rangle} \tag{3.10}$$

is the key quantity and the error term lies in spectral subspace complementary to B. We will then establish bounds in terms of Θ and D from (3.9). We note that this separation is not necessary in the bulk regime studied in [DS3], where the stability operator is bounded in every direction, which explains the additional complexity of the proof of Theorem 3.2.6 compared to the bulk local law in [DS3].

The reader should not be confused by the term "eigenvector" in the context of operators $\mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ as eigenvectors are in fact matrices in this setting, e.g. the eigenvectors P and B of $\mathcal{B}$ above are actually matrices in $\mathbb{C}^{N\times N}$.

We begin by collecting some qualitative and quantitative information about the MDE and its stability operator, which will be proven in Section 3.4.5 below. We note that (i) was first obtained in [102] and (ii) goes back to [8].

Proposition 3.3.1 (Stability of MDE and properties of the solution). *The following hold true under Assumption (3.A), (3.E) and (3.G) for some $\tau_0 \in \mathbb{R}$.*

(i) The MDE (3.1) *has a unique solution $M = M(z)$ for all $z \in \mathbb{H}$ and moreover the map $z \mapsto M(z)$ is holomorphic.*

(ii) The holomorphic function $\langle M\rangle : \mathbb{H} \to \mathbb{H}$ is the Stieltjes transform of a compactly supported probability measure with continuous density $\varrho\colon \mathbb{R} \to [0,\infty)$ given by (3.2). *Moreover, ϱ is real analytic on the open set $\{\, \varrho > 0 \,\}$.*

If $\tau_0 \in \partial \operatorname{supp} \varrho$ is a regular edge then there is $\delta_* \sim 1$ such that, for all $z \in \mathbb{H}$ satisfying $|z-\tau_0| \le \delta_*$, we have

(iii) The harmonic extension of the self-consistent density of states scales like

$$\varrho(z) \sim \begin{cases} \sqrt{\kappa+\eta}, & \text{if } \tau \in \operatorname{supp}\varrho, \\ \eta/\sqrt{\kappa+\eta}, & \text{if } \tau \notin \operatorname{supp}\varrho, \end{cases}$$

where $\tau = \Re z$, $\eta = \Im z$ and $\kappa := |\tau - \tau_0|$.

(iv) There exist $P, B \in \mathbb{C}^{N\times N}$ left and right eigenvectors of $\mathcal{B}$ such that

$$\begin{aligned} \|\mathcal{B}^{-1}\|_{\mathrm{sp}} &\lesssim (\kappa+\eta)^{-1/2}, & \|\mathcal{B}^{-1}\mathcal{Q}\|_{\mathrm{sp}} + \|B\| + \|P\| &\lesssim 1, \\ |\beta| &\sim \sqrt{\kappa+\eta}, & |\langle P, M\mathcal{S}[B]B\rangle| \sim 1, \qquad |\langle P, B\rangle| &\sim 1, \end{aligned}$$

where $\mathcal{Q} := 1 - \mathcal{P}$ and $\mathcal{P} := \langle P, \cdot\rangle B / \langle P, B\rangle$ are spectral projections of $\mathcal{B}$.

We now design a suitable norm following [DS3]. For cumulants of matrix elements $\kappa(w_{ab}, w_{cd})$ we use the short-hand notation $\kappa(ab, cd)$. We also use the short-hand notation $\kappa(\mathbf{x}b, cd)$ for the $\mathbf{x} = (x_a)_{a\in[N]}$-weighted linear combination $\sum_a x_a\kappa(ab, cd)$ of such cumulants. We use the notation that replacing an index in a scalar quantity by a dot $(\cdot)$ refers to the corresponding vector, e.g. $A_{a\cdot}$ is a short-hand notation for the vector $(A_{ab})_{b\in[N]}$. We fix two vectors $\mathbf{x}, \mathbf{y}$ and some large integer K and define the sets

$$\begin{aligned} I_0 &:= \{\,\mathbf{x}, \mathbf{y}\,\} \cup \{\, e_a, P^*_{a\cdot} \mid a \in [N]\,\}, \\ I_{k+1} &:= I_k \cup \{\, M\mathbf{u} \mid \mathbf{u} \in I_k\,\} \cup \{\, \kappa_c((M\mathbf{u})a, b\cdot), \kappa_d((M\mathbf{u})a, \cdot b) \mid \mathbf{u} \in I_k,\ a, b \in [N]\,\}, \end{aligned}$$

where $\kappa_c + \kappa_d = \kappa$ is a decomposition of κ according to the Hermitian symmetry[3]. Due to (3.3) such a decomposition exists in a way that the operator norms of the matrices $\|\kappa_d(\mathbf{x}a, \cdot b)\|$ and $\|\kappa_c(\mathbf{x}a, b\cdot)\|$, indexed by (a, b), are bounded uniformly in $\mathbf{x}$ with $\|\mathbf{x}\| \le 1$. We now define the norm

$$\begin{aligned} \|R\|_* = \|R\|_*^{K,\mathbf{x},\mathbf{y}} &:= \sum_{0\le k<K} N^{-k/2K} \|R\|_{I_k} + N^{-1/2} \max_{\mathbf{u}\in I_K} \frac{\|R_{\cdot\mathbf{u}}\|}{\|\mathbf{u}\|}, \\ \|R\|_I &:= \max_{\mathbf{u},\mathbf{v}\in I} \frac{|R_{\mathbf{u}\mathbf{v}}|}{\|\mathbf{u}\|\,\|\mathbf{v}\|}. \end{aligned}$$

We note that the sets I_k and thereby also the norm $\|\cdot\|_*$ depend implicitly on the spectral parameter z via M and P.

Remark 3.3.2. *Compared to [DS3], the sets I_k contain some additional vectors generated by the vectors of the form $P^*_{a\cdot}$ in I_0. This addition is necessary to control the spectral projection $\mathcal{P}$ in the $\|\cdot\|_*$-norm. We note, however, that the precise form of the sets I_k were not important for the proofs in [DS3]. It was only used that these sets contain deterministic vectors, and that their cardinality grows at most as some finite power $|I_k| \lesssim N^{C_k}$ of N.*

In terms of this norm we obtain the following easy estimate on $G - M$ in terms of its projection Θ onto the unstable direction of the stability operator $\mathcal{B}$.

Proposition 3.3.3. *For sufficiently small δ and fixed z such that $\|G - M\|_* \lesssim N^{-3/K}$ there are deterministic matrices R_1, R_2 with norm $\lesssim 1$ such that*

$$G - M = \Theta B - \mathcal{B}^{-1}\mathcal{Q}[MD] + \mathcal{E}, \qquad \|\mathcal{E}\|_* \lesssim N^{2/K}(|\Theta|^2 + \|D\|_*^2), \tag{3.11a}$$

with an error term $\mathcal{E}$, where Θ, defined in (3.10), satisfies the approximate quadratic equation

$$\xi_1\Theta + \xi_2\Theta^2 = \mathcal{O}\left(N^{2/K}\,\|D\|_*^2 + |\langle R_1 D\rangle| + |\langle R_2 D\rangle|\right) \tag{3.11b}$$

[3] If h_{ab} is strongly correlated with h_{cd} then, by Hermitian symmetry, it is also strongly correlated with $h_{dc} = \overline{h_{cd}}$. Therefore it is natural to split the covariance into a *direct* and *cross* contribution. The precise splitting $\kappa = \kappa_c + \kappa_d$ is chosen via an optimisation problem; the precise definition is irrelevant for the current proof, see Remark 2.2.8 for more details.

with

$$|\xi_1| \sim \sqrt{\eta + \kappa}, \qquad |\xi_2| \sim 1$$

and any implied constants are uniform in $\mathbf{x}, \mathbf{y}$ *and* $z \in \mathbb{D}^\delta$.

Proof. We begin with an auxiliary lemma about the $\|\cdot\|_*$-norm of some important quantities, the proof of which we defer to the appendix.

Lemma 3.3.4. *Depending only on the model parameters we have the estimates for any* $R \in \mathbb{C}^{N\times N}$,

$$\|M\mathcal{S}[R]R\|_* \lesssim N^{1/2K}\|R\|_*^2, \qquad \|MR\|_* \lesssim N^{1/2K}\|R\|_*,$$
$$\|\mathcal{Q}\|_{*\to*} \lesssim 1, \qquad \|\mathcal{B}^{-1}\mathcal{Q}\|_{*\to*} \lesssim 1.$$

Decomposing $G - M = \mathcal{P}[G-M] + \mathcal{Q}[G-M]$ and inverting $\mathcal{B}$ in (3.9) on the range of $\mathcal{Q}$ yields

$$\begin{aligned} G - M &= \Theta B + \mathcal{Q}[G-M] = \Theta B - \mathcal{B}^{-1}\mathcal{Q}[MD] + \mathcal{O}\left(N^{1/2K}\|G-M\|_*^2\right) \\ &= \Theta B - \mathcal{B}^{-1}\mathcal{Q}[MD] + \mathcal{O}\left(N^{3/2K}(|\Theta|^2 + \|D\|_*^2)\right), \end{aligned}$$

where $\mathcal{O}(\cdot)$ is meant with respect to the $\|\cdot\|_*$-norm and the second equality followed by iteration, Lemma 3.3.4 and the assumption on $\|G-M\|_*$. Going back to the original equation (3.9) we find

$$\begin{aligned} \beta\Theta B + \mathcal{B}\mathcal{Q}[G-M] = -MD &+ M\mathcal{S}[\Theta B - \mathcal{B}^{-1}\mathcal{Q}[MD]](\Theta B - \mathcal{B}^{-1}\mathcal{Q}[MD]) \\ &+ \mathcal{O}\left(N^{2/K}(|\Theta|^3 + \|D\|_*^3)\right) \end{aligned}$$

and thus by projecting with $\mathcal{P}$ we arrive at the quadratic equation

$$\begin{aligned} \mu_0 - \mu_1\Theta + \mu_2\Theta^2 &= \mathcal{O}\left(N^{2/K}(|\Theta|^3 + \|D\|_*^3)\right), \\ \mu_0 &= \langle P, M\mathcal{S}[\mathcal{B}^{-1}\mathcal{Q}[MD]]\mathcal{B}^{-1}\mathcal{Q}[MD] - MD\rangle, \\ \mu_1 &= \langle P, M\mathcal{S}[B]\mathcal{B}^{-1}\mathcal{Q}[MD] + M\mathcal{S}[\mathcal{B}^{-1}\mathcal{Q}[MD]]B\rangle + \beta\langle P, B\rangle, \\ \mu_2 &= \langle P, M\mathcal{S}[B]B\rangle. \end{aligned}$$

We now proceed by analysing the coefficients in this quadratic equation. We estimate the quadratic term in μ_0 directly by $N^{2/K}\|D\|_*^2$, while we write the linear term as $\langle R_1 D\rangle$ for the deterministic $R_1 := -M^*P$ with $\|R_1\| \lesssim 1$. For the linear coefficient μ_1 we similarly find a deterministic matrix R_2 such that $\|R_2\| \lesssim 1$ and $\mu_1 = \langle R_2 D\rangle + \beta\langle P, B\rangle$. Finally, we find from Proposition 3.3.1(iv) that $|\mu_2| \sim 1$ and $|\beta\langle P, B\rangle| \sim \sqrt{\kappa + \eta}$. By incorporating the $|\Theta| N^{2/K}$ term into ξ_2 we obtain (3.11b). Here δ has to be chosen sufficiently small such that Proposition 3.3.1 is applicable. □

3.3.2 Probabilistic bound

We now collect the averaged and isotropic bound on D from [DS3]. We first introduce a commonly used (see, e.g. [73]) notion of high-probability bound.

Definition 3.3.5 (Stochastic Domination). *If*

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

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

$$\sup_{u\in U^{(N)}}\mathbf{P}\left[X^{(N)}(u)>N^{\epsilon}Y^{(N)}(u)\right]\le N^{-D}$$

for large enough $N\ge N_0(\epsilon,D)$. In this case we use the notation $X\prec Y$.

It can be checked (see [73, Lemma 4.4]) that $\prec$ satisfies the usual arithmetic properties, e.g. if $X_1\prec Y_1$ and $X_2\prec Y_2$, then also $X_1+X_2\prec Y_1+Y_2$ and $X_1X_2\prec Y_1Y_2$. To formulate the result compactly we also introduce the notations

$$\begin{aligned}|R|\prec\Lambda\text{ in }\mathbb{D}\quad&\iff\quad\|R\|_*^{K,\mathbf{x},\mathbf{y}}\prec\Lambda\text{ unif. in }\mathbf{x},\mathbf{y}\text{ and }z\in\mathbb{D},\\ |R|_{\mathrm{av}}\prec\Lambda\text{ in }\mathbb{D}\quad&\iff\quad\frac{|\langle BR\rangle|}{\|B\|}\prec\Lambda\text{ unif. in }B\text{ and }z\in\mathbb{D}\end{aligned}$$

for random matrices $R=R(z)$ and a deterministic control parameter $\Lambda=\Lambda(z)$, where $B,\mathbf{x},\mathbf{y}$ are deterministic matrices and vectors. We also define an isotropic high-moment norm, already used in [DS3], for $p\ge1$ and a random matrix R,

$$\|R\|_p:=\sup_{\mathbf{x},\mathbf{y}}\frac{\left(\mathbf{E}\,|\langle\mathbf{x},R\mathbf{y}\rangle|^p\right)^{1/p}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}.$$

Proposition 3.3.6 (Bound on the Error). *Under the Assumptions (3.A)–(3.E) there exists a constant C such that for any fixed vectors $\mathbf{x},\mathbf{y}$ and matrices B and spectral parameters $z\in\mathbb{D}^\delta$, and any $p\ge1$, $\epsilon>0$,*

$$\begin{aligned}\frac{\|\langle\mathbf{x},D\mathbf{y}\rangle\|_p}{\|\mathbf{x}\|\,\|\mathbf{y}\|}&\le_{\epsilon,p}N^{\epsilon}\sqrt{\frac{\|\Im G\|_q}{N\Im z}}\left(1+\|G\|_q\right)^C\left(1+\frac{\|G\|_q}{N^\mu}\right)^{Cp}\\ \frac{\|\langle BD\rangle\|_p}{\|B\|}&\le_{\epsilon,p}N^{\epsilon}\frac{\|\Im G\|_q}{N\Im z}\left(1+\|G\|_q\right)^C\left(1+\frac{\|G\|_q}{N^\mu}\right)^{Cp},\end{aligned}$$

where $q:=Cp^4/\epsilon$. Here $\mu>0$ depends on s in Assumption (3.CD). In particular, if $|G-M|\prec\Lambda\lesssim1$, then

$$|D|\prec\sqrt{\frac{\varrho+\Lambda}{N\eta}},\qquad|D|_{\mathrm{av}}\prec\frac{\varrho+\Lambda}{N\eta}.$$

Proof. This follows from combining Theorem 2.4.1, the following lemma[4] from Lemma 2.5.4 and $\|M\|\le M_*$. □

Lemma 3.3.7. *Let R be a random matrix and Φ a deterministic control parameter. Then the following implications hold:*

[4] Cf. Remark 3.3.2, where we argue that the proof of [DS3] about $\|\cdot\|_*$ hold true verbatim in the present case despite the slightly larger sets I_k.

(i) If $\Phi \geq N^{-C}$, $\|R\| \leq N^C$ *and* $|R_{\mathbf{x}\mathbf{y}}| \prec \Phi \|\mathbf{x}\| \|\mathbf{y}\|$ *for all* $\mathbf{x}, \mathbf{y}$ *and some* C*, then* $\|R\|_p \leq_{p,\epsilon} N^\epsilon \Phi$ *for all* $\epsilon > 0, p \geq 1$.

(ii) Conversely, if $\|R\|_p \leq_{p,\epsilon} N^\epsilon \Phi$ *for all* $\epsilon > 0, p \geq 1$*, then* $\|R\|_*^{K,\mathbf{x},\mathbf{y}} \prec \Phi$ *for any fixed* $K \in \mathbb{N}, \mathbf{x}, \mathbf{y} \in \mathbb{C}^N$.

3.3.3 Bootstrapping

We now fix $\gamma > 0$ and start with the proof of Theorem 3.2.6. Phrased in terms of the $\|\cdot\|_*$-norm we will prove

$$
\begin{aligned}
|G - M| &\prec N^{2/K} \left(\sqrt{\frac{\varrho}{N\eta}} + \frac{1}{N\eta} \right) \quad \text{in} \quad \mathbb{D}, \\
|G - M|_{\text{av}} &\prec N^{2/K} \begin{cases} \frac{1}{N\eta} & \Re z \in \operatorname{supp} \varrho \\ \frac{1}{N(\kappa+\eta)} + \frac{N^{2/K}}{(N\eta)^2\sqrt{\kappa+\eta}} & \Re z \notin \operatorname{supp} \varrho \end{cases} \quad \text{in} \quad \mathbb{D},
\end{aligned}
\tag{3.12}
$$

for $\mathbb{D} = \mathbb{D}_\gamma^\delta$ and $K \gg 1/\gamma$, i.e. for $K\gamma$ sufficiently large. In order to prove (3.12) we use the following iteration procedure.

Proposition 3.3.8. *There exists a constant* $\gamma_s > 0$ *depending only on* K *and* γ *such that* (3.12) *for* $\mathbb{D} = \mathbb{D}_{\gamma_0}^\delta$ *with* $\gamma_0 > \gamma$ *implies* (3.12) *also for* $\mathbb{D} = \mathbb{D}_{\gamma_1}^\delta$ *with* $\gamma_1 := \max\{\gamma, \gamma_0 - \gamma_s\}$.

Proof of (3.12) *for* $\mathbb{D} = \mathbb{D}_\gamma^\delta$*, assuming Proposition 3.3.8.* For $\mathbb{D} = \mathbb{D}_\gamma^\delta$ with $\gamma \geq 1$ we have (3.12) by Theorem 2.2.1. For $\gamma < 1$ we iteratively apply Proposition 3.3.8 starting from[5] $\mathbb{D}_1^\delta$ finitely many times until we have shown (3.12) for $\mathbb{D} = \mathbb{D}_\gamma^\delta$. □

Proof of Proposition 3.3.8. We now suppose that (3.12) has been proven for some $\mathbb{D} = \mathbb{D}_{\gamma_0}^\delta$ and aim at proving (3.12) for $\mathbb{D} = \mathbb{D}_{\gamma_1}^\delta$ for some $\gamma_1 = \gamma_0 - \gamma_s$, $0 < \gamma_s \ll \gamma$. The proof has two stages. Firstly, we will establish the rough bounds

$$
|\Theta| \prec N^{-5/K} \quad \text{and} \quad |G - M| \prec N^{-5/K} \quad \text{in} \quad \mathbb{D}_{\gamma_1}^\delta, \tag{3.13}
$$

and then in the second stage improve upon this bound iteratively until we reach (3.12) for $\mathbb{D} = \mathbb{D}_{\gamma_1}^\delta$.

Rough bound.

By (3.12), Lemma 3.3.7 and monotonicity of the map $\eta \mapsto \eta \|G(\tau + i\eta)\|_p$ (see e.g. the proof of Proposition 2.5.5) we find $\|G\|_p \leq_{\epsilon,p} N^{\epsilon+\gamma_s} \leq N^{2\gamma_s}$ in $\mathbb{D}_{\gamma_1}^\delta$. As long as $2\gamma_s < \mu$ we thus have

$$
\begin{aligned}
\|D\|_p &\leq_{\epsilon,p} \frac{N^{\epsilon+2C\gamma_s+\gamma_s}}{\sqrt{N\eta}} \leq \frac{N^{\gamma_s(2+2C)}}{\sqrt{N\eta}}, \\
\|\langle BD \rangle\|_p &\leq_{\epsilon,p} \|B\| \frac{N^{\epsilon+2\gamma_s+2\gamma_s C}}{N\eta} \leq \|B\| \frac{N^{\gamma_s(3+2C)}}{N\eta}.
\end{aligned}
$$

[5] Strictly speaking, in the very first step we start from $\mathbb{D}^\delta \cap \{\Im z \geq \delta/2\}$ instead of $\mathbb{D}_1^\delta$ since, depending on the value of δ, the latter might be empty.

We now fix $\mathbf{x},\mathbf{y}$ and it follows from (3.11b) that

$$\left|\xi_1\Theta+\xi_2\Theta^2\right|\prec\frac{N^{2\gamma_s(3+2C)+2/K}}{N\eta}\quad\text{in}\quad\mathbb{D}^\delta_{\gamma_1}$$

and consequently by Lipschitz continuity of the lhs. with a Lipschitz constant of $\eta^{-2}\le N^2$, and choosing K,γ_s large and respectively small enough depending on γ we find that with high probability $|\xi_1\Theta+\xi_2\Theta^2|\le N^{-10/K}$ in all of $\mathbb{D}^\delta_{\gamma_1}$. The following lemma translates the bound on $|\xi_1\Theta+\xi_2\Theta^2|$ into a bound on $|\Theta|$.

Lemma 3.3.9. *Let $d=d(\eta)$ be a monotonically decreasing function in $\eta\ge 1/N$ and assume $0\le d\lesssim N^{-\epsilon}$ for some $\epsilon>0$. Suppose that*

$$\left|\xi_1\Theta+\xi_2\Theta^2\right|\lesssim d\quad\textit{for all } z\in\mathbb{D}^\delta,\qquad\textit{and}\qquad|\Theta|\lesssim\min\left\{\frac{d}{\sqrt{\kappa+\eta}},\sqrt{d}\right\}\quad\textit{for some } z_0,$$

then also $|\Theta|\lesssim\min\{d/\sqrt{\kappa+\eta},\sqrt{d}\}$ for all $z'\in\mathbb{D}^\delta$ with $\Re z'=\Re z_0$ and $\Im z'<\Im z_0$.

Proof. This proof is basically identical to the analysis of the solutions to the same approximate quadratic equation, as appeared in various previous works, see e.g. [78, Section 9]. In the spectral bulk this is trivial since then $|\xi_1|\sim\sqrt{\kappa+\eta}\sim 1$. Near a spectral edge we observe that $(\kappa+\eta)/d$ is monotonically increasing in η. First suppose that $(\kappa+\eta)/d\gg 1$ from which it follows that $|\Theta|\lesssim d/\sqrt{\kappa+\eta}\lesssim\sqrt{d}$ in the relevant branch determined by the given estimate on Θ at z_0. Now suppose that below some η-threshold we have $(\kappa+\eta)/d\lesssim 1$. Then we find $|\Theta|\lesssim\sqrt{\kappa+\eta}+\sqrt{d}\lesssim\sqrt{d}\lesssim d/\sqrt{\kappa+\eta}$ and the claim follows also in this regime. □

Since (3.13) holds in $\mathbb{D}^\delta_{\gamma_0}$ and $1/N\eta\le N^{-100/K}$, we know $|\Theta|\le\min\{N^{-10/K}/\sqrt{\kappa+\eta}, N^{-5/K}\}$ and therefore can conclude the rough bound $|\Theta|\prec N^{-5/K}$ in all of $\mathbb{D}^\delta_{\gamma_1}$ by Lemma 3.3.9 with $d=N^{-10/K}$. Consequently we have also that

$$\|G-M\|_*\mathbf{1}(\|G-M\|_*\le N^{-3/K})\prec N^{-5/K}\quad\text{in}\quad\mathbb{D}^\delta_{\gamma_1}.$$

Due to this gap in the possible values for $\|G-M\|_*$ it follows from a standard continuity argument that $\|G-M\|_*\prec N^{-5/K}$ and therefore since $\mathbf{x},\mathbf{y}$ were arbitrary, $|\Theta|\prec N^{-5/K}$ and $|G-M|\prec N^{-5/K}$ in all of $\mathbb{D}^\delta_{\gamma_1}$.

Strong bound.

All of the following bounds hold uniformly in the domain $\mathbb{D}^\delta_{\gamma_1}$ which is why we suppress this qualifier. By combining Propositions 3.3.3 and 3.3.6 we find for deterministic $0\le\theta\le\Lambda\le N^{-3/K}$ under the assumptions $|\Theta|\prec\theta$, $|G-M|\prec\Lambda$, that

$$|G-M|\prec\theta+N^{2/K}\sqrt{\frac{\varrho+\Lambda}{N\eta}},\qquad\left|\xi_1\Theta+\xi_2\Theta^2\right|\prec N^{2/K}\frac{\varrho+\Lambda}{N\eta}.\tag{3.14}$$

The bound on $|G-M|$ in (3.14) is a self-improving bound and we find after iteration that

$$|G-M|\prec\theta+N^{2/K}\left(\frac{1}{N\eta}+\sqrt{\frac{\varrho+\theta}{N\eta}}\right),$$

hence

$$\left|\xi_1\Theta + \xi_2\Theta^2\right| \prec N^{2/K}\frac{\varrho+\theta}{N\eta} + N^{4/K}\frac{1}{(N\eta)^2}.$$

We now distinguish whether $\Re z$ is inside or outside the spectrum. Inside we have $\varrho \sim \sqrt{\kappa+\eta}$, so we fix θ and use Lemma 3.3.9 with $d = N^{2/K}(\sqrt{\kappa+\eta}+\theta)/(N\eta)+N^{4/K}/(N\eta)^2$ to conclude $|\Theta| \prec \min\{d/\sqrt{\kappa+\eta}, \sqrt{d}\}$ from the input assumption $|\Theta| \prec N^{2/K}/N\eta$ in $\mathbb{D}_{\gamma_0}$. Iterating this bound, we obtain

$$|\Theta| \prec N^{2/K}\frac{1}{N\eta}, \quad \text{hence} \quad |G-M| \prec N^{2/K}\left(\sqrt{\frac{\varrho}{N\eta}}+\frac{1}{N\eta}\right).$$

By an analogous argument, outside of the spectrum we have an improved bound on Θ

$$|\Theta| \prec N^{2/K}\frac{1}{N(\kappa+\eta)} + N^{4/K}\frac{1}{(N\eta)^2\sqrt{\kappa+\eta}},$$

because $\varrho \sim \eta/\sqrt{\kappa+\eta}$. Finally, for the claimed bound on $|G-M|_{\mathrm{av}}$ we use (3.11a) in order to obtain a bound on $|G-M|_{\mathrm{av}}$ in terms of a bound on Θ. □

Due to (3.12), we now have all the ingredients to prove the local law, as well as delocalisation of eigenvectors, and the absence of eigenvalues away from the support of ϱ.

Proof of Theorem 3.2.6, Corollary 3.2.7 and Corollary 3.2.8. The local law inside the spectrum (3.5a)–(3.5b) follows immediately from (3.12). Now we prove Corollary 3.2.7. If there exists an eigenvalue λ with $\operatorname{dist}(\lambda, \operatorname{supp}\varrho) > N^{-2/3+\omega}$, then at, say, $z = \lambda + \mathrm{i}N^{-4/5}$ we have $|\langle G-M\rangle| \ge cN^{-1/5}$. On the other hand we know from the improved local law (3.12) that with high probability $|\langle G-M\rangle| \le N^{-1/4}$ and we obtain the claim.

We now turn to the proof of Corollary 3.2.8. For the eigenvectors $\mathbf{u}_k$ and eigenvalues λ_k of H we find from the spectral decomposition and the local law with high probability

$$1 \gtrsim \Im\langle \mathbf{x}, G\mathbf{x}\rangle = \eta\sum_k \frac{|\langle \mathbf{x}, \mathbf{u}_k\rangle|^2}{(\tau-\lambda_k)^2+\eta^2} \ge \frac{|\langle \mathbf{x}, \mathbf{u}_k\rangle|^2}{\eta} \quad \text{for} \quad z = \tau + \mathrm{i}\eta$$

for any normalised $\mathbf{x} \in \mathbb{C}^N$, where the last inequality followed assuming that τ is chosen η-close to λ_k. With the choice $\eta = N^{-1+\gamma}$ for arbitrarily small $\gamma > 0$ the claim follows. Note that for this proof only (3.5a) of Theorem 3.2.6 was used.

Finally, we establish (3.5c) and consider $z \in \mathbb{D}^\delta$ with $\operatorname{dist}(\Re z, \operatorname{supp}\varrho) \ge N^{-2/3+\omega}$ and $\mathbf{x}, \mathbf{y}, B$ fixed. As in the proof of [9, Corollary 1.11], the optimal local law (3.12) implies rigidity up to the edge as formulated in Corollary 3.2.9. The only difference is that this standard argument proves (3.6b) only if the supremum is restricted to $\tau \in \operatorname{supp}\varrho$ with $\operatorname{dist}(\tau, \partial\operatorname{supp}\varrho) \ge N^{-2/3+\epsilon}$. The cause for this restriction is a possible mismatch of the labelling of the edge eigenvalues, in other words the precise location of N^ϵ eigenvalues near an internal gap is not established yet; they may belong to either band adjacent to this gap. This shortcoming will be remedied by the band rigidity in the proof of Corollary 3.2.9 in Section 3.5 below. However, for the current argument, the imprecise location of N^ϵ eigenvalues does not matter. In fact, already from this version of rigidity, together with the delocalisation of eigenvectors (Corollary 3.2.8) and the absence of eigenvalues outside of the spectrum by Corollary 3.2.7 we have, at $z = \tau + \mathrm{i}\eta$ (recall that we consider $z \in \mathbb{D}^\delta$ with $\mathrm{dist}(\Re z, \operatorname{supp} \varrho) \geq N^{-2/3+\omega}$),

$$\Im \langle \mathbf{x}, G(z)\mathbf{x}\rangle = \eta \sum_k \frac{|\langle \mathbf{x}, \mathbf{u}_k\rangle|^2}{(\tau-\lambda_k)^2+\eta^2} \prec \frac{1}{N}\sum_k \frac{\eta}{(\tau-\lambda_k)^2+\eta^2} \prec \int_{\mathbb{R}} \frac{\eta\, \varrho(x)\,\mathrm{d}x}{|\tau-x|^2+\eta^2}$$

for any normalised vector $\mathbf{x}$. From the square root behaviour of ϱ at the edge and $\kappa(z) \geq N^{-2/3+\omega}$ we can easily infer $\|\Im G\|_* \prec \eta/\sqrt{\kappa+\eta}$. Therefore it follows from Proposition 3.3.6 that $\|D\|_*^2 + |\langle RD\rangle| \prec 1/(N\sqrt{\kappa+\eta})$ and from (3.11b) and Lemma 3.3.9 that $|\Theta| \prec N^{2/K-1}/(\kappa+\eta)$. Finally, we thus obtain,

$$|G-M|_{\mathrm{av}} \prec \frac{N^{2/K}}{N(\kappa+\eta)} + \frac{N^{2/K}}{N\sqrt{\kappa+\eta}} \lesssim N^{2/K}\frac{1}{N(\kappa+\eta)}$$

from (3.11a) and (3.5c) follows. □

3.4 Analysis of the Matrix Dyson equation

The essential prerequisite for edge universality is the regularity of the edge, i.e. the local square root behavior of the self consistent density ϱ as imposed in Definition 3.2.4. For the proof of universality via [122], however, it is necessary to first establish that the square-root behaviour and the adjacent gap persist in a macroscopic interval. This is achieved in the following main theorem whose proof will be given in Section 3.4.4 after several preparatory results. In particular, as a second main result of this section, in Theorem 3.4.2, we will give a sharp estimate on the inverse of the stability operator $\mathcal{B} = \mathrm{Id} - M\mathcal{S}[\,\cdot\,]M$ which also plays a central role in the proof of the local law in Section 3.3.

Theorem 3.4.1 (Behaviour of ϱ close to a square root edge)**.** *Let (3.A), (3.E) and (3.G) be satisfied for some $\tau_0 \in \mathbb{R}$. If $\tau_0 \in \partial \operatorname{supp} \varrho$ is a regular edge then there are $c \sim 1$ and $\delta_* \sim 1$ such that*

$$\varrho(\tau_0+\omega) = \begin{cases} c\,|\omega|^{1/2} + \mathcal{O}(|\omega|), & \text{if } \omega \in [-\delta_*, 0], \\ 0, & \text{if } \omega \in [0, \delta_*]. \end{cases}$$

In this section and, in particular, the previous theorem, the comparison relation $\sim$ is understood with respect to the constants in (3.A), (3.E) and (3.G) as well as in (3.4).

We now outline the strategy for the proof of Theorem 3.4.1. First, we will extend M to the real line by showing that it is $1/2$-Hölder continuous in the vicinity of τ_0 (see Corollary 3.4.3 below). The Hölder continuity also yields an a-priori bound on $\Delta := M(\tau_0+\omega) - M(\tau_0)$, hence on $\varrho(\tau_0+\omega) = \pi^{-1}\langle \Im M(\tau_0+\omega)\rangle = \pi^{-1}\langle \Im \Delta\rangle$ as well, with small $\omega \in \mathbb{R}$. Second, by using this bound, we will verify that Δ is governed by a scalar quantity analogous to Θ from (3.10) which satisfies a quadratic equation (see Proposition 3.4.12 below). The fact that $\Im \Delta \geq 0$ will select the correct solution to this quadratic equation and Theorem 3.4.1 will follow from analysing the stability of this solution.

The equation for Δ can be obtained from subtracting the MDE at $\tau_0 + \omega$ and τ_0. It reads as

$$\mathcal{B}[\Delta] = M\mathcal{S}[\Delta]\Delta + \omega M^2 + \omega M\Delta, \qquad M = M(\tau_0). \tag{3.15}$$

To express Δ from (3.15) it is therefore essential to understand the instabilities of $\mathcal{B}^{-1}$ very precisely. The main difficulty is that near the edge $\mathcal{B}$ has a small eigenvalue that is very sensitive to a delicate balance between $\mathcal{S}$ and M. An additional complication is that $\mathcal{B}$ is non-selfadjoint. Both obstacles are overcome by representing $\mathcal{B}$ in the form $\mathcal{B} = \mathcal{V}(\mathcal{U} - \mathcal{F})\mathcal{V}^{-1}$, where $\mathcal{U}$ is unitary, $\mathcal{V}$ is bounded invertible, $\mathcal{F}$ is self-adjoint and it preserves the cone of positive matrices. Thus a Perron-Frobenius argument can be applied to $\mathcal{F}$, i.e. its norm can be obtained simply by finding its top eigenvector. In this way we can very precisely determine the size of $M\mathcal{S}[\cdot]M$ and estimate its top eigenvalue without explicitly solving the MDE. This representation of $\mathcal{B}$ (cf. (3.23) below) with the Perron-Frobenius argument is one of the main results of [8] and the analysis of $\mathcal{F}$ will partly be imported from [8]. We will see that $\mathcal{B}^{-1}$ has precisely one unstable direction and we will obtain the quadratic equation for Θ, the projection of Δ, onto this direction. The sharp estimate on the eigenvalue of the unstable direction will give rise to the following bound on $\mathcal{B}^{-1}$.

Theorem 3.4.2 (Sharp bound on $\mathcal{B}^{-1}$ near a regular edge)**.** *Let (3.A), (3.E) and (3.G) be satisfied for a regular edge $\tau_0 \in \partial \operatorname{supp} \varrho$. Then there is $\delta_* \sim 1$ such that we have*

$$\|\mathcal{B}(z)^{-1}\|_{\mathrm{sp}} + \left\|\mathcal{B}(z)^{-1}\right\| \lesssim \frac{1}{\varrho(z) + \eta\varrho(z)^{-1}},$$

for all $z \in \mathbb{H}$ satisfying $|z - \tau_0| \leq \delta_*$, where $\eta = \Im z$.

From the previous theorem, we will immediately conclude the $1/2$-Hölder continuity stated in the following corollary. The proofs of both statements will be given in Section 3.4.2 below.

Corollary 3.4.3 (Hölder-continuity of M)**.** *Let (3.A), (3.E) and (3.G) be satisfied for a regular edge $\tau_0 \in \mathbb{R}$. Then M is uniformly $1/2$-Hölder continuous around τ_0 in the sense that there is $\delta_* \sim 1$ such that*

$$\|M(z_1) - M(z_2)\| \lesssim |z_1 - z_2|^{1/2}$$

for all $z_1, z_2 \in \{\tau + \mathrm{i}\eta\colon\ |\tau - \tau_0| \leq \delta_*,\ 0 < \eta < \infty\}$. In particular, M has a unique extension to $[\tau_0 - \delta_*, \tau_0 + \delta_*]$.

3.4.1 Analysis of the stability operator

In this section, we will always assume that (3.A), (3.E) and (3.G) are satisfied for some $\tau_0 \in \mathbb{R}$. The main result of this section is the bound on the inverse of the stability operator $\mathcal{B}$ in Proposition 3.4.4 below. We introduce the *balanced polar decomposition*

$$M = Q^* U Q, \tag{3.16}$$

where we define

$$W := (\Im M)^{-1/2}(\Re M)(\Im M)^{-1/2} + \mathrm{i}I, \quad Q := |W|^{1/2}\,(\Im M)^{1/2}, \quad U := \frac{W}{|W|}. \tag{3.17}$$

We remark that W is normal, $|W| := (W^*W)^{1/2}$, U is unitary and $\Im U$ is positive definite. In this context, the balanced polar decomposition first appeared in [8]. We also define

$$S := \operatorname{sign} \Re U, \qquad F_U := \varrho^{-1}\Im U, \qquad \sigma := \langle S F_U^3 \rangle. \tag{3.18}$$

The quantities $\mathcal{B}$, W, Q, U, S, F_U and σ introduced above all depend on z through the z-dependence of M. In the following, we will mostly omit this dependence from our notation.

Proposition 3.4.4 (General bound on $\mathcal{B}^{-1}$). *If (3.A), (3.E) and (3.G) are satisfied for some $\tau_0 \in \mathbb{R}$ then, uniformly for all $z \in \mathbb{D}^{\omega_*}$, we have*

$$\|\mathcal{B}(z)^{-1}\|_{\mathrm{sp}} + \left\|\mathcal{B}(z)^{-1}\right\| \lesssim \frac{1}{\varrho(z)(\varrho(z)+|\sigma(z)|)+\eta\varrho(z)^{-1}}, \qquad \eta = \Im z. \tag{3.19}$$

This proposition will be shown at the end of the present section. Now, we apply it to show that M is $1/3$-Hölder continuous.

Corollary 3.4.5 ($1/3$-Hölder continuity of M). *Let (3.A), (3.E) and (3.G) be satisfied for some $\tau_0 \in \mathbb{R}$. Then the solution M of the MDE, (3.1), is uniformly $1/3$-Hölder continuous around τ_0 in the sense that, for each $\theta \in (0, \omega_*)$, we have*

$$\|M(z_1) - M(z_2)\| \lesssim_\theta |z_1 - z_2|^{1/3}$$

for all $z_1, z_2 \in \{\tau + \mathrm{i}\eta\colon\ |\tau - \tau_0| \leq \omega_* - \theta,\ 0 < \eta < \infty\}$.

Before we prove the previous corollary, we state and prove the following lemma. It collects a few basic properties of M, Q and U which will often be used in the following.

Lemma 3.4.6 (Properties of M, Q and U). *Uniformly for $z \in \mathbb{D}^{\omega_*}$, we have*

$$\left\|M(z)^{-1}\right\| \sim \|M(z)\| \sim 1, \tag{3.20a}$$

$$\Im M(z) \sim \langle \Im M(z)\rangle, \tag{3.20b}$$

$$\|Q(z)\| \sim \left\|Q(z)^{-1}\right\| \sim 1, \tag{3.20c}$$

$$\Im U(z) \sim \langle \Im U(z)\rangle \sim \varrho(z), \tag{3.20d}$$

where $A \lesssim B$ and $A \sim B$ for matrices A, B indicate that $A \leq CB$ and $cB \leq A \leq CB$ for some constants c, C in the sense of quadratic forms.

Proof of Lemma 3.4.6. The bounds in (3.20a) and (3.20b) follow easily from the bound on $\|M\|$ on $\mathbb{D}^{\omega_*}$ as well as the flatness of $\mathcal{S}$ (see e.g. the proof of Proposition 4.2 in [8]).

For the proof of (3.20c), we use the monotonicity of the square root and (3.20b) to obtain

$$\begin{aligned} Q^*Q &= (\Im M)^{1/2}(1 + (\Im M)^{-1/2}(\Re M)(\Im M)^{-1}(\Re M)(\Im M)^{-1/2})^{1/2}(\Im M)^{1/2} \\ &\sim \langle \Im M\rangle^{-1/2}(\Im M)^{1/2}\Big((\Im M)^{-1/2}((\Im M)^2 + (\Re M)^2)(\Im M)^{-1/2}\Big)^{1/2}(\Im M)^{1/2}. \end{aligned}$$

Thus, employing $(\Re M)^2 + (\Im M)^2 \sim 1$ by (3.20a) yields (3.20c) due to (3.20b).

Owing to (3.20c), (3.20d) is a direct consequence of (3.20b). This completes the proof of Lemma 3.4.6. □

In the following, we will use the derivative of M with respect to z several times. For $z \in \mathbb{H}$, we take the derivative of (3.1) with respect to z. Owing to the invertibility of $\mathcal{B} = \mathcal{B}(z)$, this yields

$$\partial_z M(z) = \mathcal{B}^{-1}[M(z)^2] \tag{3.21}$$

for $z \in \mathbb{D}^{\omega_*}$.

Proof of Corollary 3.4.5. As $\partial_z \Im M(z) = (2\mathrm{i})^{-1}\partial_z M(z)$ due to the analyticity of M, we conclude from (3.21) and (3.19) and (3.20b) that

$$\|\partial_z \Im M(z)\| \lesssim \varrho(z)^{-2} \sim \|\Im M(z)\|^{-2} \,.$$

This implies that $z \mapsto (\Im M(z))^3$ is Lipschitz-continuous on $\mathbb{D}^{\omega_*}$. Therefore, $\Im M(z)$ is $1/3$-Hölder continuous on $\mathbb{D}^{\omega_*}$ (see e.g. Theorem X.1.1 in [30]) and, thus, M is uniformly $1/3$-Hölder continuous on $\{\tau + \mathrm{i}\eta\colon\, |\tau - \tau_0| \leq \omega_* - \theta,\ 0 < \eta < \infty\}$ for all $\theta \in (0, \omega_*)$ (see e.g. Lemma A.7 in [7] as well as Lemma A.1 in [12] for a slightly more general formulation). □

For the analysis of the stability operator $\mathcal{B}$ defined in (3.7), we now introduce the Hermitian operator $\mathcal{F}\colon \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ defined through

$$\mathcal{F} := \mathcal{C}_{Q,Q^*}\mathcal{S}\mathcal{C}_{Q^*,Q}. \tag{3.22}$$

Here, we used the following notation for operators on $\mathbb{C}^{N\times N}$. For $T_1, T_2 \in \mathbb{C}^{N\times N}$, we define the operator $\mathcal{C}_{T_1,T_2}\colon \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ through

$$\mathcal{C}_{T_1,T_2}[R] = T_1 R T_2$$

for all $R \in \mathbb{C}^{N\times N}$. We also set $\mathcal{C}_T := \mathcal{C}_{T,T}$. The importance of $\mathcal{F}$ for the analysis of $\mathcal{B}$ and its inverse comes from the following consequence of the balanced polar decomposition (3.16):

$$\mathcal{B} = \mathrm{Id} - \mathcal{C}_M\mathcal{S} = \mathcal{C}_{Q^*,Q}\mathcal{C}_U(\mathcal{C}_U^* - \mathcal{F})\mathcal{C}_{Q^*,Q}^{-1}. \tag{3.23}$$

When $\varrho = \varrho(z)$ is small, we will view $\mathcal{B}$ as a perturbation of the operator $\mathcal{B}_0$, which we introduce now. We define

$$\begin{aligned} \mathcal{B}_0 &:= \mathcal{C}_{Q^*,Q}(\mathrm{Id} - \mathcal{C}_S\mathcal{F})\mathcal{C}_{Q^*,Q}^{-1}, \\ \mathcal{E} &:= (\mathcal{C}_{Q^*SQ} - \mathcal{C}_M)\mathcal{S} = \mathcal{C}_{Q^*,Q}(\mathcal{C}_S - \mathcal{C}_U)\mathcal{F}\mathcal{C}_{Q^*,Q}^{-1}, \end{aligned} \tag{3.24}$$

with U and Q defined in (3.17), S defined in (3.18) and $\mathcal{F}$ defined in (3.22). Note $\mathcal{B}_0 = \mathrm{Id} - \mathcal{C}_{Q^*SQ}\mathcal{S}$, i.e. in the definition of $\mathcal{B}$, the unitary matrix U in $M = Q^*UQ$ is replaced by S. Thus, we have $\mathcal{B} = \mathcal{B}_0 + \mathcal{E}$.

In the following, we will often use (3.20c) and (3.20d). In particular, since $I - |\Re U| = I - \sqrt{I - (\Im U)^2} \leq (\Im U)^2 \lesssim \varrho^2$, we also obtain

$$\Re U = S + \mathcal{O}(\varrho^2), \qquad \Im U \,=\, \mathcal{O}(\varrho)\,, \qquad \Re M = Q^*SQ + \mathcal{O}(\varrho^2) \tag{3.25}$$

and with $\mathcal{C}_S - \mathcal{C}_U = \mathcal{O}(\|S - U\|) = \mathcal{O}(\varrho)$ we get

$$\mathcal{E} \,=\, \mathcal{O}(\varrho)\,. \tag{3.26}$$

Here, we use the notation $\mathcal{R} = \mathcal{T} + \mathcal{O}(\alpha)$ for operators $\mathcal{R}$ and $\mathcal{T}$ on $\mathbb{C}^{N\times N}$ and $\alpha > 0$ if $\|\mathcal{R} - \mathcal{T}\| \lesssim \alpha$. By the functional calculus, the normal matrices $U, \Re U, S$ and F_U commute. Hence, $\mathcal{C}_S[F_U] = F_U$.

The MDE, (3.1), the balanced polar decomposition, $M = Q^*UQ$, and the definition of $\mathcal{F}$ in (3.22) yield

$$-\,U^* = Q(z - A)Q^* + \mathcal{F}[U]. \tag{3.27}$$

We take the imaginary part of (3.27) and use (3.20c) as well as (3.20d) to conclude that

$$(\mathrm{Id} - \mathcal{F})[F_U] = \eta \varrho^{-1} Q Q^* = \mathcal{O}(\eta \varrho^{-1}). \tag{3.28}$$

We also introduce the operator $\mathcal{B}_*$, and view it as a perturbation of $\mathcal{B}_0$, via

$$\mathcal{B}_* := \mathrm{Id} - \mathcal{C}_{M^*,M}\mathcal{S}, \qquad \mathcal{E}_* := (\mathcal{C}_{Q^*SQ} - \mathcal{C}_{M^*,M})\mathcal{S} = \mathcal{C}_{Q^*,Q}(\mathcal{C}_S - \mathcal{C}_{U^*,U})\mathcal{F}\mathcal{C}_{Q^*,Q}^{-1}.$$

Hence, we have $\mathcal{B}_* = \mathcal{B}_0 + \mathcal{E}_*$. Analogously to (3.26), we conclude from (3.25) that

$$\mathcal{E}_* = \mathcal{O}(\varrho). \tag{3.29}$$

In the following, for $z \in \mathbb{C}$ and $\varepsilon > 0$, we denote by $D_\varepsilon(z) := \{w \in \mathbb{C}\colon\ |z - w| < \varepsilon\}$ the disk in $\mathbb{C}$ of radius ε around z.

Lemma 3.4.7 (Spectral properties of stability operator for small density). *Let* $\mathcal{T} \in \{\mathrm{Id} - \mathcal{F}, \mathrm{Id} - \mathcal{C}_S\mathcal{F}, \mathcal{B}_0, \mathcal{B}, \mathcal{B}_*\}$. *Then there are* $\varrho_* \sim 1$ *and* $\varepsilon \sim 1$ *such that*

$$\|(\mathcal{T} - \omega\mathrm{Id})^{-1}\|_{\mathrm{sp}} + \left\|(\mathcal{T} - \omega\mathrm{Id})^{-1}\right\| + \left\|(\mathcal{T}^* - \omega\mathrm{Id})^{-1}\right\| \lesssim 1 \tag{3.30}$$

uniformly for all $z \in \mathbb{D}^{\omega_*}$ *satisfying* $\varrho(z) + \eta\varrho(z)^{-1} \le \varrho_*$ *and for all* $\omega \in \mathbb{C}$ *with* $\omega \notin D_\varepsilon(0) \cup D_{1-2\varepsilon}(1)$. *Furthermore, there is a single simple (algebraic multiplicity* 1*) eigenvalue* λ *in the disk around* 0, *i.e.*

$$\mathrm{Spec}(\mathcal{T}) \cap D_\varepsilon(0) \;=\; \{\lambda\} \quad \textit{and} \quad \mathrm{rank}\,\mathcal{P}_\mathcal{T} = 1, \tag{3.31}$$

where

$$\mathcal{P}_\mathcal{T} := -\frac{1}{2\pi\mathrm{i}}\int_{\partial D_\varepsilon(0)} (\mathcal{T} - \omega\mathrm{Id})^{-1}\mathrm{d}\omega.$$

Proof. First, we introduce the bounded operators $\mathcal{V}_t\colon \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ for $t \in [0,1]$ interpolating between Id and $\mathcal{C}_S$ by

$$\mathcal{V}_t := (1-t)\mathrm{Id} + t\mathcal{C}_S\,.$$

We will perform the proof one by one for the choices $\mathcal{T} = \mathrm{Id} - \mathcal{F}, \mathrm{Id} - \mathcal{V}_t\mathcal{F}, \mathcal{B}_0, \mathcal{B}, \mathcal{B}_*$ in that order. We will first show that the operator $\mathrm{Id} - \mathcal{F}$ has a spectral gap above the single eigenvalue around 0, so for this choice the statements are easy. Then we perform two approximations. First, we interpolate between $\mathrm{Id} - \mathcal{F}$ and $\mathrm{Id} - \mathcal{C}_S\mathcal{F}$ via $\mathrm{Id} - \mathcal{V}_t\mathcal{F}$. This gives Lemma 3.4.7 for $\mathcal{T} = \mathcal{B}_0$. Then we use perturbation theory to get the results for $\mathcal{T} = \mathcal{B} = \mathcal{B}_0 + \mathcal{O}(\varrho)$ and for $\mathcal{T} = \mathcal{B}_* = \mathcal{B}_0 + \mathcal{O}(\varrho)$. Note that for all these choices of $\mathcal{T}$ the bound $\|\mathrm{Id} - \mathcal{T}\|_{\mathrm{hs}\to\|\cdot\|} \lesssim 1$ holds due to $\|\mathcal{S}\|_{\mathrm{hs}\to\|\cdot\|} \lesssim 1$, $\|M\| \lesssim 1$ and (3.20c). Hence, the invertibility of $\mathcal{T} - \omega\mathrm{Id}$ as an operator on $(\mathbb{C}^{N\times N}, \|\cdot\|)$ and on $(\mathbb{C}^{N\times N}, \|\cdot\|_{\mathrm{hs}})$ are therefore closely related as

$$\left\|(\mathcal{T} - \omega\mathrm{Id})^{-1}\right\| \le |1-\omega|^{-1}\left(1 + \|\mathrm{Id} - \mathcal{T}\|_{\mathrm{hs}\to\|\cdot\|}\|(\mathcal{T} - \omega\mathrm{Id})^{-1}\|_{\mathrm{sp}}\right).$$

The proof of this bound is elementary, see e.g. Lemma B.2 (ii) of [12]. In particular, it suffices to show (3.31) and the $\|\cdot\|_{\mathrm{sp}}$-norm bound

$$\|(\mathcal{T} - \omega\mathrm{Id})^{-1}\|_{\mathrm{sp}} \lesssim 1\,, \tag{3.32}$$

for $\omega \notin D_\varepsilon(0) \cup D_{1-2\varepsilon}(1)$ in (3.30) to establish the lemma. For $\mathcal{T} = \mathrm{Id} - \mathcal{F}$ both of these assertions are true due to the following facts about the operator $\mathcal{F}$ that have been the backbone of the analysis of [8]:

(a) The norm $\|\mathcal{F}\|_{\mathrm{sp}}$ of the Hermitian operator $\mathcal{F}\colon \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ is a simple eigenvalue of $\mathcal{F}$. Moreover, there is a unique, positive definite eigenvector $F \in \mathbb{C}^{N\times N}$ such that $\mathcal{F}[F] = \|\mathcal{F}\|_{\mathrm{sp}} F$ and $\|F\|_{\mathrm{hs}} = 1$. This eigenvector satisfies

$$1 - \|\mathcal{F}\|_{\mathrm{sp}} = (\Im z)\frac{\langle F\,, QQ^*\rangle}{\langle F\,, \Im U\rangle}. \tag{3.33a}$$

In particular, $\|\mathcal{F}\|_{\mathrm{sp}} \le 1$.

Furthermore, uniformly for all $z \in \mathbb{D}^{\omega_*}$, the following properties hold true:

(b) The eigenvector F is bounded from above and below, i.e.

$$F \sim 1. \tag{3.33b}$$

(c) The operator $\mathcal{F}$ has a spectral gap $\vartheta \sim 1$, i.e.

$$\operatorname{Spec}(\mathcal{F}/\|\mathcal{F}\|_{\mathrm{sp}}) \subset [-1+\vartheta, 1-\vartheta] \cup \{1\}. \tag{3.33c}$$

(d) The eigenvector $F, \mathcal{F}F = \|\mathcal{F}\|_{\mathrm{sp}} F$, satisfies

$$F = \|F_U\|_{\mathrm{hs}}^{-1} F_U + \mathcal{O}(\eta\varrho^{-1})\,, \tag{3.33d}$$

These facts are proven as Lemma 4.7 in [8] using Lemma 3.4.6 instead of (4.11) and (4.23) in the proof of (4.33) in [8]. Moreover, the proof of (3.33d) follows from (3.28) and $\|\mathcal{F}\|_{\mathrm{sp}} = 1 + \mathcal{O}(\eta\varrho^{-1})$ (cf. (3.33a)) by straightforward perturbation theory of the simple isolated eigenvalue $\|\mathcal{F}\|_{\mathrm{sp}}$.

Now we consider the choice $\mathcal{T} = \mathcal{T}_t = \mathrm{Id} - \mathcal{V}_t\mathcal{F}$. Once (3.32), and with it (3.30), is established for $\mathcal{T}_t$, the statement about the single isolated eigenvalue (3.31) follows. Indeed, assuming (3.30) for $\mathcal{T} = \mathcal{T}_t$, we obtain that $\mathcal{T}_t$ and, hence, the rank of $\mathcal{P}_{\mathcal{T}_t}$ is a continuous function of t on $[0,1]$. Hence, the rank of $\mathcal{P}_{\mathcal{T}_t}$ is constant along this interpolation. On the other hand, rank $\mathcal{P}_{\mathcal{T}_0} = 1$ by Fact (a) above. Therefore, for each $t \in [0,1]$, $\operatorname{Spec}(\mathcal{T}_t) \cap D_\varepsilon(0)$ consists of precisely one simple eigenvalue. We are thus left with establishing (3.32) for $\mathcal{T}_t$. As $\|\mathcal{V}_t\|_{\mathrm{sp}} \le 1$ and $\|\mathcal{F}\|_{\mathrm{sp}} \le 1$ the bound (3.32) is certainly satisfied for $|\omega| \ge 3$. Thus, we now assume $|\omega| \le 3$. In order to conclude (3.32), we now show a lower bound on $\|((1-\omega)\mathrm{Id} - \mathcal{V}_t\mathcal{F})[R]\|_{\mathrm{hs}}$ for all normalized, $\|R\|_{\mathrm{hs}} = 1$, elements $R \in \mathbb{C}^{N\times N}$. We decompose R as $R = \alpha F + R^\perp$, where $R^\perp \perp F$ with respect to the Hilbert-Schmidt scalar product on $\mathbb{C}^{N\times N}$ and $\alpha \in \mathbb{C}$. Then

$$\begin{aligned}&\|((1-\omega)\mathrm{Id} - \mathcal{V}_t\mathcal{F})[R]\|_{\mathrm{hs}}^2\\ &\quad= |\alpha|^2\,|\omega|^2 + \|((1-\omega)\mathrm{Id} - \mathcal{V}_t\mathcal{F})[R^\perp]\|_{\mathrm{hs}}^2 + \mathcal{O}(\eta\varrho^{-1}),\end{aligned} \tag{3.34}$$

because of $\|\mathcal{F}\|_{\mathrm{sp}} = 1 + \mathcal{O}(\eta\varrho^{-1})$, $\mathcal{V}_t[F_U] = F_U$ together with (3.33d), and because the mixed terms are negligible due to

$$\langle F\,, \mathcal{V}_t\mathcal{F}[R^\perp]\rangle \;=\; \langle \mathcal{F}\mathcal{V}_t[F]\,, R^\perp\rangle \;=\; \mathcal{O}(\|R^\perp\|_{\mathrm{hs}}\eta\varrho^{-1})\,.$$

Using the spectral gap $\vartheta \sim 1$ of $\mathcal{F}$ from (3.33c) and $R^\perp \perp F$ we infer (3.32) from (3.34) by estimating

$$\|((1-\omega)\mathrm{Id} - \mathcal{V}_t\mathcal{F})[R^\perp]\|_{\mathrm{hs}}^2 \;\ge\; \operatorname{dist}(\omega, D_{1-\vartheta}(1))^2\|R^\perp\|_{\mathrm{hs}}^2 \;\ge\; (\vartheta - 2\varepsilon)^2(1-|\alpha|^2),$$

optimizing in α and choosing $\varepsilon \le \vartheta/3$. This shows the lemma for $\mathcal{T} = \mathrm{Id} - \mathcal{V}_t\mathcal{F}$.

Since $\mathcal{B}_0$ is related by the similarity transform (3.24) to $\mathrm{Id} - \mathcal{V}_1\mathcal{F} = \mathrm{Id} - \mathcal{C}_S\mathcal{F}$ and $\|Q\|\,\|Q^{-1}\| \lesssim 1$ (cf. (3.20c)), the operator $\mathcal{B}_0$ inherits the properties listed in the lemma from $\mathrm{Id} - \mathcal{C}_S\mathcal{F}$. Finally, we can perform analytic perturbation theory for the simple isolated eigenvalue in $D_\varepsilon(0)$ of $\mathcal{B}_0$ to verify the lemma for $\mathcal{T} = \mathcal{B} = \mathcal{B}_0 + \mathcal{E}$ with $\mathcal{E} = \mathcal{O}(\varrho)$ (cf. (3.26)) and $\mathcal{T} = \mathcal{B}_* = \mathcal{B}_0 + \mathcal{E}_*$ with $\mathcal{E}_* = \mathcal{O}(\varrho)$ (cf. (3.29)) if ϱ_* is sufficiently small. This completes the proof of Lemma 3.4.7. □

In the following corollary, we use the concepts of *left* and *right eigenvector* of an operator $\mathcal{T}\colon \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$. We say $V_l \in \mathbb{C}^{N\times N}$ ($V_r \in \mathbb{C}^{N\times N}$) is a left (right) eigenvector of $\mathcal{T}$ corresponding to the eigenvalue $\lambda \in \mathbb{C}$ of $\mathcal{T}$ if $\mathcal{T}^*[V_l] = \bar{\lambda} V_l$ ($\mathcal{T}[V_r] = \lambda V_r$).

Corollary 3.4.8. *Let $z \in \mathbb{D}^{\omega_*}$ satisfy $\varrho(z) + \eta\varrho(z)^{-1} \le \varrho_*$ for $\varrho_* \sim 1$ from Lemma 3.4.7.*

Let β_0 and β be the isolated eigenvalues in $D_\varepsilon(0)$ of $\mathcal{B}_0$ and $\mathcal{B}$, respectively, from Lemma 3.4.7. Furthermore, let $\mathcal{P}_0 = \mathcal{P}_{\mathcal{B}_0}$ and $\mathcal{P} = \mathcal{P}_{\mathcal{B}}$ be the spectral projections corresponding to the isolated eigenvalue of $\mathcal{B}_0$ and $\mathcal{B}$, respectively (see (3.31)). Then with $\mathcal{Q}_0 := \mathrm{Id} - \mathcal{P}_0$ and $\mathcal{Q} := \mathrm{Id} - \mathcal{P}$ we have

$$\left\|\mathcal{B}^{-1}\mathcal{Q}\right\| + \|\mathcal{B}^{-1}\mathcal{Q}\|_{\mathrm{sp}} + \left\|\mathcal{B}_0^{-1}\mathcal{Q}_0\right\| \lesssim 1. \tag{3.35}$$

We define $B_0 := \mathcal{P}_0\mathcal{C}_{Q^*,Q}[F_U]$ and $P_0 := \mathcal{P}_0^*\mathcal{C}_{Q,Q^*}^{-1}[F_U]$. Then B_0 and P_0 are right and left eigenvector of $\mathcal{B}_0$ corresponding to β_0 and we have

$$B_0 = \mathcal{C}_{Q^*,Q}[F_U] + \mathcal{O}(\eta\varrho^{-1}), \qquad P_0 = \mathcal{C}_{Q,Q^*}^{-1}[F_U] + \mathcal{O}(\eta\varrho^{-1}), \tag{3.36a}$$

$$\beta_0 = \frac{\eta}{\varrho}\frac{\pi}{\langle F_U^2\rangle} + \mathcal{O}(\eta^2\varrho^{-2}) = \mathcal{O}(\eta\varrho^{-1})\,. \tag{3.36b}$$

We also define $B := \mathcal{P}[B_0]$ and $P := \mathcal{P}^*[P_0]$. This yields right and left eigenvectors of $\mathcal{B}$ corresponding to β which satisfy

$$B = B_0 + \mathcal{O}(\varrho)\,, \tag{3.37a}$$

$$P = P_0 + \mathcal{O}(\varrho)\,, \tag{3.37b}$$

$$\beta\langle P, B\rangle = \pi\eta\varrho^{-1} - 2\mathrm{i}\varrho\sigma + \mathcal{O}(\varrho^2 + \eta + \eta^2\varrho^{-2})\,. \tag{3.37c}$$

Moreover, we have

$$\|B\| \lesssim 1, \qquad \|P\| \lesssim 1, \qquad |\langle P, B\rangle| \sim 1. \tag{3.38}$$

The following identity will be used a few times

$$\langle F_U QQ^*\rangle = \varrho^{-1}\langle \Im M\rangle = \pi. \tag{3.39}$$

It is obtained by a direct computation starting from the definition of F_U in (3.18), the balanced polar decomposition, $M = Q^*UQ$, and $\varrho(z) = \pi^{-1}\langle\Im M(z)\rangle$.

Proof. The bounds in (3.35) are a direct consequence of Lemma 3.4.7. Using (3.28) and $\mathcal{C}_S[F_U] = F_U$, we see that

$$\mathcal{B}_0^*\mathcal{C}_{Q,Q^*}^{-1}[F_U] = \eta\varrho^{-1}I\,, \qquad \mathcal{B}_0\mathcal{C}_{Q^*,Q}[F_U] = \mathcal{O}(\eta\varrho^{-1})\,. \tag{3.40}$$

The representations of B_0 and P_0 in (3.36a) follow by simple perturbation theory because β_0 is a nondegenerate isolated eigenvalue. The expression for β_0 in (3.36b) is seen by taking the scalar product with B_0 in the first identity of (3.40) as well as using (3.36a) and (3.39).

The expansions (3.37) follow by first order analytic perturbation theory. Indeed, $B = B_0 + \mathcal{O}(\varrho)$ and $P = P_0 + \mathcal{O}(\varrho)$ as $\mathcal{E} = \mathcal{B} - \mathcal{B}_0 = \mathcal{O}(\varrho)$ due to (3.26). For the proof of (3.37c), we first compute $\mathcal{E}[B_0]$. From (3.36a), we obtain the first equality below:

$$\mathcal{E}[B_0] = \mathcal{C}_{Q^*,Q}(\mathcal{C}_S - \mathcal{C}_U)\mathcal{F}[F_U] + \mathcal{O}(\eta) = -2\mathrm{i}\varrho\mathcal{C}_{Q^*,Q}[SF_U^2] + \mathcal{O}(\varrho^2 + \eta), \qquad (3.41)$$

For the second equality in (3.41), we used (3.28), $\|\mathcal{C}_S - \mathcal{C}_U\| = \mathcal{O}(\varrho)$ and $(\mathcal{C}_S - \mathcal{C}_U)[F_U] = 2(\Im U - \mathrm{i}\Re U)(\Im U)F_U = -2\mathrm{i}\varrho SF_U^2 + \mathcal{O}(\varrho^2)$ due to (3.25). For the proof of (3.37c), we start from $\mathcal{B}[B] = \beta B, \mathcal{B} = \mathcal{B}_0 + \mathcal{E}$, use (3.37a), (3.37b) as well as $\mathcal{E} = \mathcal{O}(\varrho)$ and obtain

$$\beta\langle P\,, B\rangle = \beta_0\langle P_0\,, B_0\rangle + \langle P_0\,, \mathcal{E}[B_0]\rangle + \mathcal{O}(\varrho^2).$$

Together with the following two expansions, this yields (3.37c). We have

$$\begin{aligned}\beta_0\langle P_0\,, B_0\rangle &= \pi\eta\varrho^{-1} + \mathcal{O}(\eta^2\varrho^{-2}),\\ \langle P_0\,, \mathcal{E}[B_0]\rangle &= -2\mathrm{i}\varrho\langle SF_U^3\rangle + \mathcal{O}(\varrho^2 + \eta) = -2\mathrm{i}\varrho\sigma + \mathcal{O}(\varrho^2 + \eta).\end{aligned}$$

The first expansion is a consequence of $\langle P_0\,, B_0\rangle = \langle F_U^2\rangle + \mathcal{O}(\eta\varrho^{-1})$ due to (3.36a) and (3.36b). The second expansion follows from (3.36a) and (3.41).

The first two bounds in (3.38) follow directly from (3.37a) and (3.37b) as well as (3.36a), (3.20c) and (3.20d). Moreover, (3.36a), (3.37a) and (3.37b) imply $|\langle P\,, B\rangle| \sim \langle F_U^2\rangle \sim 1$ by (3.20d). This completes the proof of Corollary 3.4.8. □

Proof of Proposition 3.4.4. As in the proof of Lemma 3.4.7, it suffices to show the bound on $\|\mathcal{B}^{-1}\|_{\mathrm{sp}}$ in (3.19).

From (3.23), by using Lemma 3.4.6, we conclude that

$$\begin{aligned}\|\mathcal{B}^{-1}\|_{\mathrm{sp}} \lesssim \|(\mathcal{C}_U^* - \mathcal{F})^{-1}\|_{\mathrm{sp}} &\lesssim |1 - \|\mathcal{F}\|_{\mathrm{sp}}\langle F\,, \mathcal{C}_U^*[F]\rangle|^{-1}\\ &\lesssim \Big(1 - \|\mathcal{F}\|_{\mathrm{sp}} + |1 - \langle F\,, \mathcal{C}_U^*[F]\rangle|\Big)^{-1}.\end{aligned}$$

Here, we applied the Rotation-Inversion Lemma, Lemma 4.9 in [8], with $\mathcal{T} = \mathcal{F}$ and $\mathcal{U} = \mathcal{C}_U^*$ in the second step. Its conditions are met due to Fact (a) and Fact (c) about $\mathcal{F}$ from the proof of Lemma 3.4.7.

Owing to (3.33a) as well as (3.20c) and (3.20d), we have $1 - \|\mathcal{F}\|_{\mathrm{sp}} \sim \eta\varrho^{-1}$. Therefore, it suffices to show that

$$|1 - \langle F\,, \mathcal{C}_U^*[F]\rangle| \gtrsim \varrho(\varrho + |\sigma|) \qquad (3.42)$$

when $\eta\varrho^{-1}$ is small. As $1 \geq \langle F\Re UF\Re U\rangle$ due to $\|F\|_{\mathrm{hs}} = 1$, we estimate

$$|1 - \langle F\,, \mathcal{C}_U^*[F]\rangle| = |1 - \langle FU^*FU^*\rangle| \gtrsim \langle F\Im UF\Im U\rangle + |\langle F\Im UF\Re U\rangle|\,.$$

Since $\Im U \sim \varrho$ by (3.20d), the first term on the right-hand side scales like $\sim \varrho^2$. This proves (3.42) when $\varrho \geq \varrho_*$ for any $\varrho_* \sim 1$ as $|\sigma| \lesssim 1$. If ϱ_* is sufficiently small and $\varrho + \eta\varrho^{-1} \leq \varrho_*$ then we use $\langle F\Im UF\Re U\rangle = \varrho\|F_U\|_{\mathrm{hs}}^{-2}\langle F_U^3 S\rangle + \mathcal{O}(\varrho^3 + \eta)$ by (3.33d) and (3.25) to conclude (3.42) and, thus, (3.19) in the missing regime. This completes the proof of Proposition 3.4.4. □

3.4.2 Sharp bound on $\mathcal{B}^{-1}$ and $1/2$-Hölder continuity of M

In this section, we will prove Theorem 3.4.2 and Corollary 3.4.3. They will be proven directly after the following proposition, the main result of the present section. It shows that σ introduced in (3.18) is of order one close to regular edges $\tau_0 \in \partial \operatorname{supp} \varrho$. For the formulation of this proposition, we define

$$\mathcal{A}[R,T] := \frac{1}{2}\Big(M\mathcal{S}[R]T + T\mathcal{S}[R]M\Big) \tag{3.43}$$

with $R, T \in \mathbb{C}^{N\times N}$.

Proposition 3.4.9. *Let (3.A), (3.E) and (3.G) be satisfied for some $\tau_0 \in \mathbb{R}$. If $\tau_0 \in \partial \operatorname{supp} \varrho$ is a regular edge then the following statements hold true*

(i) At $z = \tau_0$, for P and B defined as in Corollary 3.4.8, we have

$$|\langle P, \mathcal{A}[B,B]\rangle| \sim 1.$$

(ii) There is $\delta_* \sim 1$ such that

$$|\sigma(z)| \sim 1$$

for all $z \in \mathbb{H}$ satisfying $|z - \tau_0| \le \delta_*$.

Proposition 3.4.9 immediately implies Theorem 3.4.2 and Corollary 3.4.3.

Proof of Theorem 3.4.2. By Proposition 3.4.9 (ii), there is $\delta_* \sim 1$ such that $|\sigma(z)| \sim 1$ for all $z \in \mathbb{H}$ satisfying $|z - \tau_0| \le \delta_*$. Therefore, Theorem 3.4.2 follows directly from Proposition 3.4.4. □

Proof of Corollary 3.4.3. We proceed exactly as in the proof of Corollary 3.4.5 but use Theorem 3.4.2 instead of (3.19) for all $z \in \mathbb{H}$ such that $|z - \tau_0| \le \delta_*$, where δ_* is chosen as in Theorem 3.4.2. □

The proof of Proposition 3.4.9 requires two auxiliary lemmas whose proofs are postponed until the end of the section. Some statements in these lemmas will be stated for more general $\tau_0 \in \mathbb{R}$ not only when τ_0 is a regular edge, although we will eventually use them in this case.

We now choose $\theta = \omega_*/2$ in Corollary 3.4.5 and work on the set $\mathbb{D}^{\omega_*/2}$ in the following. Note that $\mathbb{D}^{\omega_*/2} \subset \mathbb{D}^{\omega_*}$. By Hölder-continuity we can then extend M to $\overline{\mathbb{D}^{\omega_*/2}}$, and we denote the extension by M as well. Moreover, the operators $\mathcal{B}$ and $\mathcal{B}_*$ are defined for all $z \in \overline{\mathbb{D}^{\omega_*/2}}$ and the results about $\mathcal{B}$ and $\mathcal{B}_*$ in Lemma 3.4.7 hold true on $\overline{\{z \in \mathbb{D}^{\omega_*/2} \colon \varrho(z) + \eta\varrho(z)^{-1} \le \varrho_*\}}$, where the closure is taken with respect to the Euclidean topology on $\mathbb{C}$. Lemma 3.4.10 below shows that this set contains a neighbourhood around any point $\tau_0 \in \partial \operatorname{supp} \varrho$.

Lemma 3.4.10. *Let (3.A), (3.E) and (3.G) hold true for some $\tau_0 \in \mathbb{R}$. Then the following holds true:*

(i) There is $\varrho_* \sim 1$ such that, for the eigenvalue β_* of $\mathcal{B}_* = \mathrm{Id} - \mathcal{C}_{M^*,M}\mathcal{S}$ in $D_\varepsilon(0)$ (cf. Lemma 3.4.7), we have

$$|\beta_*| \sim \eta/\varrho \tag{3.44}$$

uniformly for $z \in \mathbb{D}^{\omega_*}$ satisfying $\varrho(z) + \eta\varrho(z)^{-1} \le \varrho_*$.

(ii) If $\tau_0 \in \partial \operatorname{supp} \varrho$ *and* $\varrho_* \sim 1$ *then there is* $\delta_* \sim 1$ *such that* $\varrho(z) + \eta\varrho(z)^{-1} \leq \varrho_*$ *for all* $z \in \mathbb{H}$ *satisfying* $|z - \tau_0| \leq \delta_*$.

Moreover, we have

$$\lim_{\eta \downarrow 0} \eta \varrho(\tau_0 + \mathrm{i}\eta)^{-1} = 0. \tag{3.45}$$

Lemma 3.4.11. *Let (3.A), (3.E) and (3.G) be satisfied for some* $\tau_0 \in \mathbb{R}$. *Then there is* $\varrho_* \sim 1$ *such that, uniformly for all* $z \in \mathbb{D}^{\omega_*}$ *satisfying* $\varrho(z) + \eta\varrho(z)^{-1} \leq \varrho_*$, *we have*

$$\langle P, \mathcal{A}[B, B]\rangle = \sigma + \mathcal{O}(\varrho + \eta\varrho^{-1}), \tag{3.46a}$$
$$\langle P, M\mathcal{S}[B]B\rangle = \sigma + \mathcal{O}(\varrho + \eta\varrho^{-1}). \tag{3.46b}$$

We remark that (3.46b) will be used in the next section.

Proof of Proposition 3.4.9. In this proof, we will analyse M and $\mathcal{B} = \mathrm{Id} - \mathcal{C}_M\mathcal{S}$ on the real line outside the self-consistent spectrum, i.e. we will consider spectral parameters $z = \tau + \mathrm{i}\eta$ such that $\tau \in [\tau_0 - \omega_*/2, \tau_0 + \omega_*/2] \setminus \operatorname{supp} \varrho$ and $\eta = 0$. In particular, $\varrho(\tau) = 0$ and thus $M = M^*$ by (3.20b). Owing to the continuity of M (Corollary 3.4.5), M satisfies the MDE, (3.1), also for these spectral parameter z. Moreover, $\varrho(\tau + \mathrm{i}\eta) \lesssim \eta / \operatorname{dist}(\tau + \mathrm{i}\eta, \operatorname{supp} \varrho)^2$ as $\langle M \rangle$ is the Stieltjes transform of the measure μ on $\mathbb{R}$ (compare (3.64)). Thus, $\mathcal{B}$ is invertible at $\tau \notin \operatorname{supp} \varrho$ due to Proposition 3.4.4 as the term $\eta\varrho^{-1}$ has a uniform lower bound for $z = \tau + \mathrm{i}\eta$ with $\eta > 0$. In particular, M and β are differentiable with respect to $\omega = \tau - \tau_0$ for $\tau \notin \operatorname{supp} \varrho$. First order perturbation theory of the isolated eigenvalue β of the non-selfadjoint operator $\mathcal{B}$ yields

$$\begin{aligned} \partial_\omega \beta &= -\frac{\langle P, \mathcal{C}_{\partial_\omega M, M}\mathcal{S}[B]\rangle}{\langle P, B\rangle} - \frac{\langle P, \mathcal{C}_{M, \partial_\omega M}\mathcal{S}[B]\rangle}{\langle P, B\rangle} \\ &= -\frac{\langle P, (\partial_\omega M)\mathcal{S}[B]M + M\mathcal{S}[B](\partial_\omega M)\rangle}{\langle P, B\rangle}. \end{aligned} \tag{3.47}$$

For definiteness, we assume in the following that τ_0 is a right edge. Hence, $\omega > 0$. The argument for a left edge works completely analogously.

Owing to the invertibility of $\mathcal{B}$, the MDE, (3.1), is differentiable at τ with respect to ω. Similarly to (3.21), we obtain

$$\partial_\omega M = \mathcal{B}^{-1}[M^2] = \frac{\langle P, M^2\rangle}{\beta\langle P, B\rangle} B + \mathcal{B}^{-1}\mathcal{Q}[M^2].$$

In the second step, we inserted $\mathcal{P} + \mathcal{Q} = \mathrm{Id}$ and employed the definition of $\mathcal{P} = \mathcal{P}_\mathcal{B}$ in Corollary 3.4.8. We insert this into (3.47) and get from Lemma 3.4.7 and (3.35) that

$$\begin{aligned} \partial_\omega \beta &= -\frac{\langle P, M^2\rangle}{\beta\langle P, B\rangle^2}\langle P, B\mathcal{S}[B]M + M\mathcal{S}[B]B\rangle + \mathcal{O}(1) \\ &= \frac{2\langle P, M^2\rangle}{\beta\langle P, B\rangle^2}\langle P, \mathcal{A}[B, B]\rangle + \mathcal{O}(1). \end{aligned}$$

The bounds in (3.38) of Corollary 3.4.8 yield $\|P\| \lesssim 1$ and, hence, $|\langle P, M^2\rangle| \lesssim 1$ by Assumption (3.G). By (3.38), we have $|\langle P, B\rangle| \sim 1$ if $\eta > 0$. Thus, as a consequence of the continuity of M by Corollary 3.4.5 and, hence, of P and B, the derivative of β^2 is bounded by $|\partial_\omega(\beta^2)| \lesssim |\langle P, \mathcal{A}[B,B]\rangle| + |\beta|$. This implies

$$|\beta|^2 \lesssim |\langle P, \mathcal{A}[B,B]\rangle|\, \omega + \omega^2. \tag{3.48}$$

On the other hand, from (3.44) and the continuity of β_*, and $\beta_* = \beta$ for $\eta = 0$ (as $M = M^*$) we get

$$|\beta(\tau_0 + \omega)| \sim \lim_{\eta \downarrow 0} \frac{\eta}{\varrho(\tau_0 + \omega + \mathrm{i}\eta)} \sim \bigg(\int_0^\delta \frac{\varrho(\tau_0 - \omega')}{(\omega' + \omega)^2} \mathrm{d}\omega' \bigg)^{-1},$$

for some $\delta \sim 1$. From this and (3.4), we conclude that

$$\liminf_{\omega \downarrow 0} \frac{|\beta(\tau_0 + \omega)|}{\sqrt{\omega}} \sim \limsup_{\omega \downarrow 0} \frac{|\beta(\tau_0 + \omega)|}{\sqrt{\omega}} \sim 1\,,$$

i.e. $|\beta|^2 \sim \omega$ as $\omega \downarrow 0$. Therefore, we find $|\langle P, \mathcal{A}[B,B]\rangle| \gtrsim 1$ at $z = \tau_0$ due to (3.48). The upper bound follows from $\|P\| \lesssim 1$ and $\|B\| \lesssim 1$ by Corollary 3.4.8. This completes the proof of (i).

For the proof of (ii), we conclude that $\langle P, \mathcal{A}[B,B]\rangle$ is a uniformly $1/3$-Hölder continuous function of z on $\{w \in \mathbb{H} \cup \mathbb{R} \colon |w - \tau_0| \le \delta_*\}$ for some $\delta_* \sim 1$ due to Corollary 3.4.5 and Lemma 3.4.10 (ii). By possibly shrinking $\delta_* \sim 1$, we can thus assume that $|\langle P, \mathcal{A}[B,B]\rangle| \sim 1$ for all $z \in \mathbb{H}$ satisfying $|z - \tau_0| \le \delta_*$. From Lemma 3.4.10 (ii) and (3.46a), we conclude that $|\sigma(z)| \sim 1$ for all $z \in \mathbb{H}$ such that $|z - \tau_0| \le \delta_*$ for some sufficiently small $\delta_* \sim 1$. Hence, we have completed the proof of Proposition 3.4.9. □

Proof of Lemma 3.4.10. Similarly to the proof of Corollary 3.4.8, we find a left eigenvector P_* of $\mathcal{B}_*$ corresponding to β_*, i.e. $(\mathcal{B}_*)^*[P_*] = \overline{\beta_*} P_*$, such that

$$P_* = Q^{-1} F_U (Q^*)^{-1} + \mathcal{O}(\varrho + \eta\varrho^{-1}) \tag{3.49}$$

provided that $z \in \mathbb{D}^{\omega_*}$ satisfies $\varrho(z) + \eta\varrho(z)^{-1} \le \varrho_*$. We take the imaginary part of (3.1) and compute the scalar product with P_*. This yields

$$\beta_* = \frac{\eta}{\varrho} \frac{\langle P_*, M^* M\rangle}{\langle P_*, \varrho^{-1} \Im M\rangle}. \tag{3.50}$$

Using (3.49) and the balanced polar decomposition, $M = Q^* U Q$, we obtain

$$\begin{aligned} \langle P_*, M^* M\rangle &= \langle F_U Q Q^*\rangle + \mathcal{O}(\varrho + \eta\varrho^{-1}) = \pi + \mathcal{O}(\varrho + \eta\varrho^{-1}), \\ \langle P_*, \varrho^{-1} \Im M\rangle &= \langle F_U^2\rangle + \mathcal{O}(\varrho + \eta\varrho^{-1}). \end{aligned}$$

Here, we used that U and F_U commute and (3.39) in order to compute $\langle P_*, M^* M\rangle$. We thus deduce that $|\langle P_*, M^* M\rangle| \sim 1$ and $|\langle P_*, \varrho^{-1} \Im M\rangle| \sim 1$ for all $z \in \mathbb{D}^{\omega_*}$ satisfying $\varrho(z) + \eta\varrho(z)^{-1} \le \varrho_*$ for some sufficiently small $\varrho_* \sim 1$ due to (3.20d). Therefore, taking the absolute value in (3.50) and using these scaling relations complete the proof of (3.44).

For the proof of (ii), we remark that, owing to the continuity of ϱ, we have

$$\lim_{\eta \downarrow 0} \varrho(\tau + \mathrm{i}\eta)^{-1} \eta = 0$$

for all $\tau \in \mathbb{R}$ satisfying $\varrho(\tau) > 0$. From (3.44), we thus conclude that $\beta_*(\tau) = 0$ if $\varrho(\tau) > 0$ for all $\tau \in [\tau_0 - \omega_*/2, \tau_0 + \omega_*/2]$. The continuity of M from Corollary 3.4.5 implies that $\mathcal{B}_*$

is also $1/3$-Hölder continuous. Consequently, β_* is also $1/3$-Hölder continuous as it is an isolated eigenvalue of $\mathcal{B}_*$. Owing to the continuity of ϱ, we find a sequence $(\tau_n)_n$ such that $\tau_n \to \tau_0 \in \partial \operatorname{supp} \varrho$ and $\varrho(\tau_n) > 0$ for all n. Thus, the continuity of β_* yields $\beta_*(\tau_0) = 0$. Therefore, we have $|\beta_*| + \varrho = 0$ at $z = \tau_0$. Hence, the $1/3$-Hölder continuity of $|\beta_*| + \varrho$ implies that there is $\delta_* \sim 1$ such that $\varrho(z) + \eta\varrho(z)^{-1} \leq \varrho_*$ since $\varrho + \eta\varrho^{-1} \sim \varrho + |\beta_*|$ by (3.44). From $\beta_*(\tau_0) = 0$ and (3.44), we directly conclude (3.45). This completes the proof of Lemma 3.4.10. □

Proof of Lemma 3.4.11. First, we use the balanced polar decomposition, $M = Q^*UQ$, (3.22) and the definition of $\mathcal{A}$ in (3.43) to obtain

$$\mathcal{A}[R,T] = \frac{1}{2}\mathcal{C}_{Q^*,Q}\Big[U(\mathcal{F}\mathcal{C}_{Q^*,Q}^{-1}[R])\mathcal{C}_{Q^*,Q}^{-1}[T] + \mathcal{C}_{Q^*,Q}^{-1}[T](\mathcal{F}\mathcal{C}_{Q^*,Q}^{-1}[R])U\Big] \tag{3.51}$$

for $R, T \in \mathbb{C}^{N\times N}$.

We choose $\varrho_* \sim 1$ small enough such that Lemma 3.4.7 is applicable. By using $U = S + \mathcal{O}(\varrho)$ due to (3.25) as well as (3.18), (3.28) and (3.36a) in (3.51), we get

$$\mathcal{A}[B_0, B_0] = \mathcal{C}_{Q^*,Q}[SF_U^2] + \mathcal{O}(\varrho + \eta\varrho^{-1}).$$

In order to show (3.46a), we use (3.37a) as well as (3.37b) and obtain

$$\langle P, \mathcal{A}[B,B]\rangle = \langle P_0, \mathcal{A}[B_0,B_0]\rangle + \mathcal{O}(\varrho) = \langle SF_U^3\rangle + \mathcal{O}(\varrho+\eta\varrho^{-1}) = \sigma + \mathcal{O}(\varrho+\eta\varrho^{-1}).$$

This completes the proof of (3.46a). A similar computation yields (3.46b). □

3.4.3 Derivation of the quadratic equation

In this section, we expand $M(\tau_0 + \omega)$ around $M(\tau_0)$ for a regular edge $\tau_0 \in \partial \operatorname{supp} \varrho$. We show that this approximation is to leading order dominated by a scalar-valued quantity, Θ, which satisfies a quadratic equation. That is the content of the following proposition which is the main result of this section.

Proposition 3.4.12 (Quadratic equation for shape analysis)**.** *Let (3.A), (3.E) as well as (3.G) be satisfied for some regular edge $\tau_0 \in \partial \operatorname{supp} \varrho$. Then there is $\delta_* \sim 1$ such that the following hold true:*

(a) For all $\omega \in [-\delta_*, \delta_*]$, we have

$$M(\tau_0 + \omega) - M(\tau_0) = \Theta(\omega)B + R(\omega), \tag{3.52}$$

where $\Theta\colon [-\delta_*, \delta_*] \to \mathbb{C}$ and $R\colon [-\delta_*, \delta_*] \to \mathbb{C}^{N\times N}$ are defined by

$$\Theta(\omega) := \left\langle \frac{P}{\langle B, P\rangle}, M(\tau_0+\omega) - M(\tau_0)\right\rangle, \quad R(\omega) := \mathcal{Q}[M(\tau_0+\omega) - M(\tau_0)]. \tag{3.53}$$

Here, $P = P(\tau_0)$, $B = B(\tau_0)$ and $\mathcal{Q} = \mathcal{Q}(\tau_0)$ are the eigenvectors and spectral projection of $\mathcal{B}(\tau_0)$ introduced in Corollary 3.4.8. We have $B = B^*$ and $P = P^*$ as well as $B \sim 1$ and $P \sim 1$. Moreover, $\Theta(\omega)$ and $R(\omega)$ are bounded by

$$|\Theta(\omega)| \lesssim |\omega|^{1/2}, \qquad \Im\Theta(\omega) \geq 0, \qquad \|\Im R(\omega)\| \lesssim |\omega|^{1/2}\,\Im\Theta(\omega) \tag{3.54}$$

uniformly for all $\omega \in [-\delta_*, \delta_*]$.

(b) The function Θ satisfies the quadratic equation

$$\sigma\Theta^2(\omega) + \omega\Xi(\omega) = 0, \qquad\qquad \Xi(\omega) = \pi(1+\nu(\omega)), \tag{3.55}$$

for all $\omega \in [-\delta_*, \delta_*]$, where $\sigma = \langle P, M\mathcal{S}[B]B\rangle$, $M = M(\tau_0)$, and the error term $\nu(\omega)$ satisfies

$$|\nu(\omega)| \lesssim |\omega|^{1/2}, \qquad |\Im\nu(\omega)| \lesssim \Im\Theta(\omega) \tag{3.56}$$

for all $\omega \in [-\delta_*, \delta_*]$.

The definition $\sigma = \langle P, M\mathcal{S}[B]B\rangle$ for $\tau_0 \in \partial\operatorname{supp}\varrho$ extends the definition of σ in (3.18) on $\mathbb{H}$ owing to (3.46b), (3.45) as well as the continuity of M and, thus, P, B and ϱ.

We warn the reader that, in this section, functions of z like M, B, P, U, Q, etc. without argument are understood to be evaluated at τ_0 instead of the generic spectral parameter z which is the convention in most of the other parts of this work.

Proof. The first bound in (3.54) follows directly from Corollary 3.4.3.

From (3.36a), (3.37a), (3.37b), $\varrho(\tau_0) = 0$ and (3.45), we conclude that B and P are the limits of Hermitian, positive-definite matrices which are ~ 1 due to Lemma 3.4.6. Thus, $B = B^* \sim 1$ and $P = P^* \sim 1$. This also implies that $\Im\Theta(\omega) \geq 0$ in (3.54) as $\Im M(\tau_0 + \omega)$ is always positive semidefinite and $\Im M(\tau_0) = 0$.

In the following lemma whose proof we postpone till the end of this section we establish a quadratic equation for Θ.

Lemma 3.4.13 (Derivation of the quadratic equation). *Let $\Theta(\omega)$ and $R(\omega)$ be defined as in (3.53) and $\mathcal{A}$ be defined as in (3.43). Then there is $\delta_* \sim 1$ such that, for all $\omega \in [-\delta_*, \delta_*]$, $\Theta = \Theta(\omega)$ satisfies the quadratic equation*

$$\mu_2\Theta^2 + \mu_1\Theta + \mu_0 = e(\omega)$$

with some error term $e(\omega) = \mathcal{O}(|\omega|^{3/2})$ and with coefficients

$$\mu_2 = \langle P, \mathcal{A}[B,B]\rangle, \qquad \mu_1 = -\beta\langle P, B\rangle, \qquad \mu_0 = \omega\langle P, M^2\rangle. \tag{3.57}$$

Moreover, for all $\omega \in [-\delta_*, \delta_*]$, we have

$$|\Im e(\omega)| \lesssim |\omega|\,\Im\Theta(\omega), \qquad\qquad \|\Im R(\omega)\| \lesssim |\omega|^{1/2}\,\Im\Theta(\omega). \tag{3.58}$$

We now compute the coefficients defined in (3.57) precisely. This will yield the quadratic equation in (3.55).

Owing to (3.46a), (3.46b), (3.45), $\varrho(\tau_0) = 0$ and the continuity of M and, thus, P, B and ϱ, we have $\mu_2 = \sigma$ as defined in Proposition 3.4.12 (b).

The expansion in (3.37c) implies $\mu_1 = 0$ at τ_0 by (3.45). We now compute μ_0. At $z \in \mathbb{H}$ satisfying $\varrho(z) + \varrho(z)^{-1}\Im z \leq \varrho_*$, we conclude from (3.37b), (3.36a) and the balanced polar decomposition, $M = Q^*UQ$, from (3.16) that

$$\begin{aligned}\langle P, M^2\rangle &= \langle Q^{-1}F_U(Q^*)^{-1}, Q^*UQQ^*UQ\rangle + \mathcal{O}(\varrho + \eta\varrho^{-1}) \\ &= \langle F_U QQ^*\rangle + \mathcal{O}(\varrho + \eta\varrho^{-1}) = \pi + \mathcal{O}(\varrho + \eta\varrho^{-1}).\end{aligned}$$

Here, we also employed that $U = S + \mathcal{O}(\varrho)$ by (3.25) and F_U and S commute in the second step and (3.39) in the last step. Thus, we have $\mu_0 = \omega\pi$ at τ_0 by (3.45).

We set $\nu(\omega) := -(\pi\omega)^{-1}e(\omega)$ with $e(\omega)$ as introduced in Lemma 3.4.13. This immediately implies the first bound in (3.56). From (3.58), we conclude the second estimate in (3.56) and the third estimate in (3.54). This completes the proof of Proposition 3.4.12. □

Proof of Lemma 3.4.13. Owing to the Hölder-continuity of M, we conclude that $M(z)$ is invertible and satisfies (3.1) for all $z \in \overline{\mathbb{D}^{\omega_*/2}}$. Hence, evaluating (3.1) at $z = \tau_0 + \omega$ and $z = \tau_0$, computing their difference and introducing $M := M(\tau_0)$ as well as $\Delta := M(\tau_0 + \omega) - M$, we obtain

$$\mathcal{B}[\Delta] = \mathcal{A}[\Delta, \Delta] + \omega M^2 + \omega \mathcal{K}[\Delta], \qquad \mathcal{K}[\Delta] := \frac{1}{2}(M\Delta + \Delta M). \tag{3.59}$$

In order to compute $R = \mathcal{Q}[\Delta]$, we apply $\mathcal{B}^{-1}\mathcal{Q}$ to (3.59), use $\Delta = \Theta B + R$ and, owing to the Hölder-continuity of M, $\|\Delta(\omega)\| \lesssim |\omega|^{1/2}$ and $|\Theta(\omega)| \lesssim |\omega|^{1/2}$, find $\delta_* \sim 1$ such that

$$\|R(\omega)\| \lesssim |\omega|, \qquad \|\Im R(\omega)\| \lesssim |\omega|^{1/2} \Im \Theta(\omega) \tag{3.60}$$

for all $\omega \in [-\delta_*, \delta_*]$. Here, in order to estimate $\Im R$, we used that $M = M^*$ and, hence, $\mathcal{A}[B,B] = \mathcal{A}[B,B]^*$ as $\tau_0 \in \partial \operatorname{supp} \varrho$. This shows the second estimate in (3.58).

We apply $\langle P, \cdot \rangle$ to (3.59) and use the decomposition $\Delta = \Theta B + R$ as well as $\mathcal{B}[B] = \beta B$ which yield

$$\begin{aligned} \Theta\beta\langle P, B\rangle &= \omega\langle P, M^2\rangle + \Theta^2\langle P, \mathcal{A}[B,B]\rangle + e, \\ e &:= \langle P, \Theta(\mathcal{A}[B,R] + \mathcal{A}[R,B]) + \mathcal{A}[R,R]\rangle + \omega\langle P, \mathcal{K}[\Delta]\rangle. \end{aligned}$$

From (3.60), we conclude

$$|e(\omega)| \lesssim |\omega|^{3/2}, \qquad |\Im e(\omega)| \lesssim |\omega| \, \Im\Theta(\omega).$$

This establishes the quadratic equation as well as the missing bounds on e and, thus, completes the proof of Lemma 3.4.13. □

3.4.4 Shape analysis

In this section, we conclude Theorem 3.4.1 from Proposition 3.4.12.

Proof of Theorem 3.4.1. We recall that $\sigma = \mu_2 = \langle P, M\mathcal{S}[B]B + B\mathcal{S}[B]M\rangle/2$ as in the proof of Proposition 3.4.12 and $|\sigma| \sim 1$ by Proposition 3.4.9 (i). We will show that there is $\delta_* \sim 1$ such that

$$\varrho(\tau_0 + \omega) = \begin{cases} \dfrac{\pi^{1/2}}{|\sigma|^{1/2}} |\omega|^{1/2} + \mathcal{O}(|\omega|), & \text{if } \operatorname{sign}\omega = \operatorname{sign}\sigma, \\ 0, & \text{if } \operatorname{sign}\omega = -\operatorname{sign}\sigma, \end{cases} \tag{3.61}$$

for all $\omega \in [-\delta_*, \delta_*]$. This directly implies Theorem 3.4.1 with $c = \sqrt{\pi/|\sigma|}$ as we conclude $\sigma < 0$ from (3.4) and (3.61).

We now compute $\Theta(\omega)$ in (3.52) by identifying the correct solution of (3.55). The general quadratic equation $\Omega(\zeta)^2 + \zeta = 0$ with $\zeta \in \mathbb{C}$ has two solutions:

$$\Omega_\pm(\zeta) = \pm \begin{cases} \mathrm{i}\zeta^{1/2}, & \text{if } \Re\zeta \geq 0, \\ -(-\zeta)^{1/2}, & \text{if } \Re\zeta < 0, \end{cases}$$

where $\zeta^{1/2}$ denotes the standard branch of the square root with the branch cut $(-\infty, 0)$.

Since $\Theta(\omega)$ is a continuous function of ω and $|\nu(\omega)| < 1$ for all $\omega \in [-\delta_*, \delta_*]$ for $\delta_* \sim 1$ sufficiently small due to the first bound in (3.56), we conclude from (3.55) that there are $p, q \in \{+, -\}$ such that

$$\begin{aligned}\Theta(\omega) &= \Omega_p(\Lambda(\omega))\mathbf{1}(\omega/\sigma < 0) + \Omega_q(\Lambda(\omega))\mathbf{1}(\omega/\sigma \geq 0),\\ \Lambda(\omega) &:= \frac{\pi\omega}{\sigma}(1 + \nu(\omega))\end{aligned} \tag{3.62}$$

for all $\omega \in [-\delta_*, \delta_*]$.

We now show that $q = +$ by a proof by contradiction. We assume $q = -$. For $\omega/\sigma \geq 0$, we have

$$\Im\Omega_-(\Lambda(\omega)) = -\left(\frac{\pi\omega}{\sigma}\right)^{1/2} + \mathcal{O}\Big(|\nu(\omega)|\,|\omega|^{1/2}\Big).$$

For sufficiently small ω we thus obtain $\Im\Omega_-(\Lambda(\omega)) < 0$ in contradicition to $\Im\Theta(\omega) \geq 0$ from (3.54). This implies $q = +$.

Next, we prove that $\Im\Theta(\omega) = 0$ for all $\omega \in I_{\delta_*}$ with $\delta_* \sim 1$ sufficiently small, where $I_{\delta_*} := \{\omega \in \mathbb{R}\colon\ \operatorname{sign}\omega = -\operatorname{sign}\sigma,\ \ |\omega| \leq \delta_*\}$. We will not determine p in (3.62) but rather show that $\Im\Theta = 0$ on I_{δ_*} for either choice of p (In fact, $p = +$ can be shown [12, Proposition 7.10 (ii)]). By possibly shrinking $\delta_* \sim 1$, we get

$$|\Re\Omega_\pm(\Lambda(\omega))| \sim |\omega|^{1/2}$$

as $\sigma \in \mathbb{R}$ and $|\sigma| \sim 1$. Therefore, taking the imaginary part of (3.55) and using the second bound in (3.56), (3.62) and $\sigma \in \mathbb{R}$ yield

$$|\omega|^{1/2}\,\Im\Theta(\omega) \lesssim |\omega|\,\Im\Theta(\omega)$$

for all $\omega \in I_{\delta_*}$. If $\delta_* \sim 1$ is sufficiently small then we obtain $\Im\Theta(\omega) = 0$ for all $\omega \in I_{\delta_*}$.

We now take the imaginary part of (3.52) and apply $\langle\,\cdot\,\rangle$. Hence, we obtain

$$\varrho(\tau_0 + \omega) = \Im\Theta(\omega)\pi^{-1}\langle B\rangle + \pi^{-1}\langle\Im R(\omega)\rangle = \Im\Theta(\omega) + \mathcal{O}\Big(|\omega|^{1/2}\,\Im\Theta(\omega)\Big) \tag{3.63}$$

for all $\omega \in [-\delta_*, \delta_*]$. Here, we used $B = B^*$ in the first step and $\langle B\rangle = \pi$ by (3.37a), (3.36a), (3.39) and (3.45) as well as the third bound in (3.54) in the second step.

Since $q = +$ in (3.62), we can bound $\Im\Theta(\omega) = \Im\Omega_+(\Lambda(\omega))$ directly in (3.63) to obtain the first case in (3.61). Since $\Im\Theta(\omega) = 0$ for all $\omega \in I_{\delta_*}$, (3.63) implies the second case in (3.61). This completes the proof of (3.61) and, thus, the one of Theorem 3.4.1. □

3.4.5 Proof of Proposition 3.3.1

We have now established all results which are necessary for the proof of Proposition 3.3.1.

Proof of Proposition 3.3.1. Claims (i) and (ii) follow directly from [102] and [8].

Part (iii) is a direct consequence of Theorem 3.4.1 and the Stieltjes transform representation of $\langle M(z)\rangle$, i.e.

$$\langle M(z)\rangle = \int_{\mathbb{R}} \frac{\varrho(\tau)}{\tau - z}\,\mathrm{d}\tau \tag{3.64}$$

for $z \in \mathbb{H}$ (this simple calculation can be found, e.g. in Corollary A.1 in [7]).

For the proof of (iv), we first remark that (iii) implies $\varrho(z)+\eta\varrho(z)^{-1} \sim \sqrt{|\tau-\tau_0|+\eta}$ for all $z \in \mathbb{H}$ satisfying $|z-\tau_0| \leq \delta_*$. Thus, Theorem 3.4.2 yields the first bound in (iv). Owing to (3.35), we have $\|\mathcal{B}^{-1}\mathcal{Q}\|_{\mathrm{sp}} \lesssim 1$. Moreover, we choose P and B as in Corollary 3.4.8. This completes the proof of the second bound in (iv) due to (3.38).

Moreover, $|\sigma| \sim 1$ by Proposition 3.4.9. Hence, we conclude $|\langle P, M\mathcal{S}[B]B\rangle| \sim 1$ from Lemma 3.4.11. Furthermore, owing to (3.38), we have $|\langle P, B\rangle| \sim 1$. Thus, since $\sigma \in \mathbb{R}$ and $|\sigma| \sim 1$ we get from (3.37c) that

$$|\beta| \sim |\beta\langle P, B\rangle| \sim \varrho + \eta\varrho^{-1} \sim \sqrt{|\tau-\tau_0|+\eta}.$$

This completes the proof of Proposition 3.3.1. □

3.5 Band rigidity

Within this section we establish band rigidity for correlated random matrices H. This topological rigidity phenomenon asserts that the number of eigenvalues of H within a spectral band, i.e. a connected component of $\operatorname{supp}\varrho$, does not fluctuate and is accurately predicted by the self-consistent density of states with high probability. On the level of the MDE this phenomenon is reflected by the *band mass formula* (3.65) below, guaranteeing that $N\varrho$ assigns only integer values to each band. In particular, small continuous deformations of the data $(A, \mathcal{S})$ of the MDE cannot change these values.

Proposition 3.5.1 (Band mass formula). *For $\tau \in \mathbb{R} \setminus \operatorname{supp}\varrho$ the integrated self-consistent density of states satisfies*

$$\int_{-\infty}^{\tau} \varrho(x)\mathrm{d}x = \frac{1}{N}\left|\operatorname{Spec}(M(\tau)) \cap (-\infty, 0)\right|. \tag{3.65}$$

In particular, $N\int_{-\infty}^{\tau}\varrho(x)\mathrm{d}x$ is an integer.

Before we prove Proposition 3.5.1 we show how it is used to establish band rigidity for H.

Proof of Corollary 3.2.9. We begin with the proof of (3.6a) and consider a flow that interpolates between $H = H_0$ and a deterministic matrix H_1. We fix $\tau \notin \operatorname{supp}\varrho$ with $\epsilon := \operatorname{dist}(\tau, \operatorname{supp}\varrho) > 0$ and set

$$H_t := \sqrt{1-t}W + A_t, \quad A_t := A - t\mathcal{S}[M(\tau)], \quad \mathcal{S}_t := (1-t)\mathcal{S}, \quad t \in [0,1]. \tag{3.66}$$

The MDE corresponding to H_t is

$$I + (z - A_t + \mathcal{S}_t[M_t(z)])M_t(z) = 0 \tag{3.67}$$

with data $(A_t, \mathcal{S}_t)$, solution $M_t(z)$ and self-consistent density of states ϱ_t. We refer to this t-dependent MDE as MDE_t. It is designed in such a way that $M_t(\tau)$ at the fixed spectral parameter $z = \tau$ is kept constant at t varies. Moreover, by the following lemma, whose proof we postpone, τ stays away from the self-consistent spectrum along the flow.

Lemma 3.5.2. *Let $\epsilon := \operatorname{dist}(\tau, \operatorname{supp}\varrho) > 0$ and M_t be the solution to* MDE_t (3.67). *Then* $\operatorname{dist}(\tau, \operatorname{supp}\varrho_t) \gtrsim_\epsilon 1$ *and* $\lim_{\eta\downarrow 0} M_t(\tau + \mathrm{i}\eta) = M(\tau)$ *for all* $t \in [0,1]$.

We will now show that along the flow, with overwhelming probability, no eigenvalue crosses the spectral parameter τ. More precisely we claim that

$$\mathbf{P}\Big(\tau \in \operatorname{Spec} H_t \text{ for some } t \in [0,1]\Big) \leq_\epsilon N^{-D} \tag{3.68}$$

for any $D > 0$. Since $H_0 = H$ and $H_1 = A - \mathcal{S}[M(\tau)]$, (3.68) implies that with overwhelming probability

$$|\operatorname{Spec} H \cap (-\infty, \tau)| = |\operatorname{Spec}(A - \mathcal{S}[M(\tau)] - \tau) \cap (-\infty, 0)| = N \left\langle \mathbf{1}_{(-\infty,0)}(M(\tau)) \right\rangle,$$

where the last identity used the the MDE (3.1) at $z = \tau$. Now (3.6a) follows from the band mass formula (3.65), i.e. from $\langle \mathbf{1}_{(-\infty,0)}(M(\tau)) \rangle = \int_{-\infty}^{\tau} \varrho(\lambda)\, \mathrm{d}\lambda$.

It remains to show (3.68). We first consider the regime of values t close to 1. Since τ is separated away from $\operatorname{supp} \varrho$, and $M(\tau)$ is bounded we conclude from (3.1) at $z = \tau$ that the spectrum of $A - \mathcal{S}[M(\tau)]$ is also separated away from τ. Moreover, applying Corollary 2.2.3 to $H = W$ yields $\|W\| \leq C$ with overwhelming probability as the corresponding self-consistent density of states has compact support by Proposition 3.3.1(ii). Since therefore H_t is a small perturbation of $A - \mathcal{S}[M(\tau)]$ as long as t is close to 1, we conclude that the spectrum of H_t is bounded away from τ as well for every fixed $t \geq 1 - c$ for some small enough constant $c > 0$. We are thus left with the regime $t \leq 1 - c$, where the flatness condition from Assumption (3.E) for H_t is satisfied. In this regime we use Corollary 2.2.3 again. Since $\operatorname{dist}(\tau, \operatorname{supp} \varrho_t) \geq_\epsilon 1$ this corollary implies that the spectrum of H_t is bounded away from τ with overwhelming probability for every fixed $t \leq 1 - c$. Applying a discrete union bound in t together with the Lipschitz continuity of the eigenvalues in t for the flow (3.66) on the set $\|W\| \leq C$ yields (3.68).

Finally, (3.6b) follows from the optimal local law as in the proof of Theorem 3.2.6 and Corollary 3.2.7 above. This time, however, (3.6a) ensures that there is no mismatch between location and label of eigenvalues close to internal edges. In the spectral bulk this potential discrepancy between label and location does not matter as (3.6b) allows for an N^ϵ-uncertainty. At the spectral edge, however, neighbouring eigenvalues can lie on opposite sides of a spectral gap and we need (3.6a) to make sure that each eigenvalue has, with high probability, a definite location with respect to the spectral gap. □

Proof of Lemma 3.5.2. Note that $M(z)$ is analytic and bounded away from the self-consistent spectrum because it admits a Stieltjes transform representation (cf. Proposition 2.1 of [8]). We consider MDE_t (3.67) at a spectral parameter $\tau + \zeta$ with some $\zeta \in \mathbb{H}$ such that $|\zeta| \ll 1$ and subtract it from MDE_t at spectral parameter τ. Properly symmetrised the resulting quadratic equation for $\Delta = \Delta(\zeta) = M_t(\tau + \zeta) - M(\tau)$ takes the form

$$\mathcal{B}_t[\Delta] = \zeta M^2 + \frac{\zeta}{2}(M\Delta + \Delta M) + (1-t)\mathcal{A}[\Delta, \Delta], \tag{3.69}$$

where $M = M(\tau)$, $\mathcal{A}$ is as in (3.43) and $\mathcal{B}_t = \mathrm{Id} - (1-t)\mathcal{C}_M \mathcal{S}$ is the stability operator. We will see that equation (3.69) is linearly stable in the sense that $\left\|\mathcal{B}_t^{-1}\right\| \leq_\epsilon 1$ uniformly in t. Note that the terms containing Δ on the right hand side are lower order. Thus we may apply the implicit function theorem to show that $\Delta(\zeta)$ is an analytic function for sufficiently small ζ with $\Delta(\zeta) = \zeta \mathcal{B}_t^{-1}[M^2] + \mathcal{O}\left(|\zeta|^2\right)$. In particular, it extends to small $\zeta \in \mathbb{C}$. Since $M = M(\tau)$ is self-adjoint and $\mathcal{B}_t^{-1}$ preserves the cone of positive definite matrices, $M + \Delta(\zeta)$

coincides for any small $\zeta \in \mathbb{H}$ with the unique solution to MDE_t with positive definite imaginary part. But since $\Delta(\zeta)$ is analytic in ζ for any small enough ζ, even with negative imaginary part, $M_t(z)$ can be analytically extended to a t-independent neighbourhood of τ in $\mathbb{C}$. Furthermore, since $\mathcal{B}_t$ and $R \mapsto \mathcal{A}[R, R]$ preserve the space of self-adjoint matrices, this extension takes self-adjoint values on the real line. Thus for every t the density $\varrho_t = \frac{1}{\pi}\langle \Im M_t\rangle$ vanishes in a neighbourhood of τ, i.e. $\operatorname{dist}(\tau, \operatorname{supp}\varrho_t) \geq_\epsilon 1$.

To show the bound on $\mathcal{B}_t^{-1}$ we use the symmetrisation (3.23) with the self energy operator $\mathcal{S}_t = (1-t)\mathcal{S}$ to see that

$$\left\|\mathcal{B}_t^{-1}\right\|_{\mathrm{sp}} \leq_\epsilon \left\|(\mathcal{C}_U^* - \mathcal{F}_t)^{-1}\right\|_{\mathrm{sp}} \lesssim \frac{1}{1-(1-t)\left\|\mathcal{F}\right\|_{\mathrm{sp}}}, \tag{3.70}$$

where U is unitary and $\mathcal{F}_t = (1-t)\mathcal{F}$ with the self-adjoint operator $\mathcal{F}$ from (3.22). Exactly as in the proof of Lemma 3.4.7 the boundedness of $\mathcal{B}_t^{-1}$ in the $\|\cdot\|_{\mathrm{sp}}$-norm also implies $\left\|\mathcal{B}_t^{-1}\right\| \leq_\epsilon 1$. Thus it remains to show that the right hand side of (3.70) is bounded. For this purpose we apply the lower bound on $1 - \|\mathcal{F}\|_{\mathrm{sp}} \geq_\epsilon 1$ from [13, Lemma 3.6], finishing the proof of the lemma. □

Proof of Proposition 3.5.1. Let $\epsilon := \operatorname{dist}(\tau, \operatorname{supp}\varrho) > 0$. Again we make use of MDE_t (3.67). Recall that $M(\tau)$ solves MDE_t at spectral parameter τ, which stays away from the self-consistent spectrum by Lemma 3.5.2.

Since $M_t(z)$ is the Stieltjes transform of a matrix valued measure on $\operatorname{supp}\varrho_t$ it can be analytically extended to $\mathbb{C}\setminus\operatorname{supp}\varrho_t$, a set that contains the spectral parameter τ for which $M_t(\tau) = M(\tau)$ by the lemma. When ϱ and $M(\tau)$ are replaced by ϱ_t and $M_t(\tau)$, respectively, in (3.65) then clearly this identity holds at time $t = 1$ since $M_1(z) = (A - \mathcal{S}[M(\tau)] - z)^{-1} = (\tau + M(\tau)^{-1} - z)^{-1}$ is the resolvent of the self-adjoint matrix $\tau + M(\tau)^{-1}$. As $M_t(\tau) = M(\tau)$, it suffices to establish that the left-hand side of (3.65) with ϱ replaced by ϱ_t does not change along the flow.

To show that the left hand side is independent of t, we differentiate the contour integral representation

$$\int_{-\infty}^{\tau} \varrho_t(x)\mathrm{d}x = -\oint \frac{\mathrm{d}z}{2\pi\mathrm{i}}\langle M_t(z)\rangle,$$

where the contour encircles $[\min\operatorname{supp}\varrho_t, \tau)$ counterclockwise, passing through the real line only at τ and to the left of $\inf_t \min\operatorname{supp}\varrho_t$. With $M_t = M_t(z)$ we find

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\oint \langle M_t\rangle\,\mathrm{d}z &= \oint \langle(\mathcal{C}_{M_t^*}^{-1} - \mathcal{S}_t)^{-1}[I], \mathcal{S}[M(\tau) - M_t]\rangle \\ &= \oint \partial_z\Big(\langle M_t\mathcal{S}[M(\tau)]\rangle - \frac{1}{2}\langle M_t\mathcal{S}[M_t]\rangle\Big)\mathrm{d}z = 0,\end{aligned}$$

where the formula $(\mathcal{C}_{M_t}^{-1} - \mathcal{S}_t)[\partial_t M_t] = \mathcal{S}[M(\tau) - M_t]$, used in the first identity, is obtained by differentiating MDE_t with data (3.66) with respect to t and the formula $(\mathcal{C}_{M_t}^{-1} - \mathcal{S}_t)[\partial_z M_t] = I$, used in the second identity, follows from differentiating (3.67) with respect to z. □

3.6 Proof of Universality

In order to prove Theorem 3.2.11, we define the Ornstein Uhlenbeck (OU) process starting from $H = H_0$ by

$$\mathrm{d}H_t = -\frac{1}{2}(H_t - A)\,\mathrm{d}t + \Sigma^{1/2}[\mathrm{d}B_t], \qquad \Sigma[R] := \mathbf{E}\, W\,\mathrm{Tr}(WR), \tag{3.71}$$

where B_t is a matrix of, up to symmetry, independent (real or complex, depending on the symmetry class of H) Brownian motions and $\Sigma^{1/2}$ is the square root of the positive definite operator $\Sigma\colon \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$. We note that the same process has already been used in [DS3, 55, 8] to prove bulk universality. The proof now has two steps: Firstly, we will prove edge universality for H_t if $t \gg N^{-1/3}$ and then we will prove that for $t \ll N^{-1/6}$, the eigenvalues of H_t have the same k-point correlation functions as those of $H = H_0$.

3.6.1 Dyson Brownian Motion

The process (3.71) can be integrated, and we have

$$H_t - A = e^{-t/2}(H_0 - A) + \int_0^t e^{(s-t)/2}\Sigma^{1/2}[\mathrm{d}B_s],$$

$$\int_0^t e^{(s-t)/2}\Sigma^{1/2}[\mathrm{d}B_s] \sim \mathcal{N}(0, (1-e^{-t})\Sigma).$$

The process is designed in such a way that it preserves expectation $\mathbf{E}\, H_t = A$ and covariances $\mathbf{Cov}(h^t_{ab}, h^t_{cd}) = \mathbf{Cov}(h_{ab}, h_{cd})$ along the flow. Due to the fullness Assumption (3.F) there exists a constant $c > 0$ such that $(1-e^{-t})\Sigma - ct\Sigma^{\mathrm{GUE/GOE}} \geq 0$ for $t \leq 1$, where $\Sigma^{\mathrm{GOE/GUE}}$ denotes the covariance operator of the GOE/GUE ensembles. It follows that we can write

$$H_t = \widetilde{H}_t + \sqrt{ct}U, \qquad \kappa_t = \kappa - ct\kappa^{\mathrm{GOE/GUE}}, \qquad \mathbf{E}\,\widetilde{H}_t = A, \qquad U \sim \text{GOE/GUE},$$

where κ_t here denotes the cumulants of $\widetilde{H}_t$, and U is chosen to be independent of $\widetilde{H}_t$. Due to the fact that Gaussian cumulants of degree more than 2 vanish, it is easy to check that $H_t, \widetilde{H}_t$ satisfy the assumptions of Theorem 3.2.6 uniformly in, say, $t \leq N^{-1/10}$. From now on we fix $t = N^{-1/3+\epsilon}$ with some small $\epsilon > 0$.

Since the MDE is purely determined by the first two moments of the corresponding random matrix, it follows that $G_t := (H_t - z)^{-1}$ is close to the same M in the sense of a local law for all t. For $\widetilde{G}_t := (\widetilde{H}_t - z)^{-1}$ we have the MDE

$$I + (z - A + \mathcal{S}_t[M_t])M_t = 0, \qquad \mathcal{S}_t := \mathcal{S} - ct\mathcal{S}^{\mathrm{GOE/GUE}} \tag{3.72}$$

that can be viewed as a perturbation of the original MDE with $t = 0$. The corresponding self-consistent density of states is $\varrho_t(\tau) := \lim_{\eta\searrow 0} \Im \langle M_t(\tau + \mathrm{i}\eta)\rangle /\pi$. The fact that M_t remains bounded uniformly in $t \leq N^{-1/10}$ follows from a similar (but much simpler) argument as those leading to the local law in Section 3.3. The analogue of (3.9) with G replaced by $M_t(z)$ is obtained by subtracting (3.1) from (3.72) and the analogue of the error term D is trivially controlled by t. The details are presented in the MDE perturbation result in [12, Proposition 10.1] with $S = \mathcal{S}$, $S_t = \mathcal{S}_t$ and $a_t = A$ as the condition on S_t in [12, Eq. (10.1)] is obviously satisfied for this choice of S_t due to $\left\lVert \mathcal{S}^{\mathrm{GOE/GUE}}[R] \right\rVert \lesssim \langle R \rangle$ for all positive semidefinite matrices R. In particular the shape analysis from Section 3.4 also applies to M_t.

The Stieltjes transforms of the free convolutions of the empirical spectral density of $\widetilde{H}_t$ and ϱ_t with the semicircular distribution generated by $\sqrt{ct}U$ are given implicitly as the unique solutions to the equations

$$\widetilde{m}_{\mathrm{fc}}^t(z) = \langle \widetilde{G}_t(z + ct\widetilde{m}_{\mathrm{fc}}^t(z))\rangle\,, \qquad m_{\mathrm{fc}}^t(z) = \langle M_t(z + ctm_{\mathrm{fc}}^t(z))\rangle\,.$$

We denote the corresponding right-edges close to τ_0 by $\widetilde{\tau}_t$ and τ_t. By differentiating the defining equations for m_{fc}^t and $\widetilde{m}_{\mathrm{fc}}^t$ we find

$$\begin{aligned} \frac{(m_{\mathrm{fc}}^t)'(z)}{1+ct(m_{\mathrm{fc}}^t)'(z)} &= \langle M_t'(\xi_t(z))\rangle\,, \qquad \frac{(\widetilde{m}_{\mathrm{fc}}^t)'(z)}{1+ct(\widetilde{m}_{\mathrm{fc}}^t)'(z)} = \langle \widetilde{G}_t'(\widetilde{\xi}_t(z))\rangle\,, \\ \frac{(m_{\mathrm{fc}}^t)''(z)}{(1+ct(m_{\mathrm{fc}}^t)'(z))^3} &= \langle M_t''(\xi_t(z))\rangle\,, \end{aligned} \tag{3.73a}$$

where $\xi_t(z) := z + ctm_{\mathrm{fc}}^t(z)$ and $\widetilde{\xi}_t(z) := z + ct\widetilde{m}_{\mathrm{fc}}^t(z)$. From the first two equalities in (3.73a) we conclude

$$1 = ct\,\langle M_t'(\xi_t(\tau_t))\rangle\,, \qquad 1 = ct\,\langle \widetilde{G}_t'(\widetilde{\xi}_t(\widetilde{\tau}_t))\rangle\,, \tag{3.73b}$$

by considering the $z \to \tau_t$ and $z \to \widetilde{\tau}_t$ limits and that $(m_{\mathrm{fc}}^t)'$, $(\widetilde{m}_{\mathrm{fc}}^t)'$ blow up at the edge due to the well known square root behaviour of the density along the semicircular flow. We now compare the edge location and edge slope of the densities ϱ_{fc}^t and $\widetilde{\varrho}_{\mathrm{fc}}^t$ corresponding to m_{fc}^t and $\widetilde{m}_{\mathrm{fc}}^t$ with that of M. Very similar estimates for deformed Wigner ensembles have been used in [104]. We split the analysis into four claims.

Claim 1

$|\tau_t - \tau_0| \lesssim t/N$. Using that $\mathcal{S}^{\mathrm{GUE}}[R] = \langle R\rangle$, $\mathcal{S}^{\mathrm{GOE}}[R] = \langle R\rangle + R^t/N$ and (3.72) evaluated at $\xi_t(z)$, we find using the boundedness of M_t,

$$\begin{aligned} &I + (z - A + \mathcal{S}[M_t(\xi_t(z))])M_t(\xi_t(z)) \\ &\quad = ct\Big(\mathcal{S}^{\mathrm{GOE/GUE}}[M_t(\xi_t(z))] - \langle M_t(\xi_t(z))\rangle\Big)M_t(\xi_t(z)) = \mathcal{O}\left(\frac{t}{N}\right). \end{aligned}$$

It thus follows that $M_t(\xi_t(z))$ approximately satisfies the MDE for M at z. By using the first bound in Proposition 3.3.1(iv) expressing the stability of the MDE against small additive perturbations it follows that

$$\begin{aligned} \Big|m_{\mathrm{fc}}^t(z) - \langle M(z)\rangle\Big| = |\langle M_t(\xi_t(z)) - M(z)\rangle| &\lesssim \frac{t}{N\sqrt{\eta + \mathrm{dist}(\Re z, \partial\,\mathrm{supp}\,\varrho)}} \\ &\le \frac{t}{N\sqrt{\mathrm{dist}(\Re z, \partial\,\mathrm{supp}\,\varrho)}}. \end{aligned} \tag{3.74}$$

Suppose first that $\tau_0 = \tau_t + \delta$ for some positive $\delta > 0$. Then $\sqrt{\delta} \lesssim \Im\langle M(\tau_t + \delta/2)\rangle \lesssim t/N\sqrt{\delta}$, where the first bound follows from the square root behaviour of ϱ at the edge τ_0, while the second bound comes from (3.74) at $z = \tau_t + \delta/2$ and $\Im m_{\mathrm{fc}}^t(\tau_t + \delta/2) = 0$. We thus conclude $\delta \lesssim t/N$. If on the contrary $\tau_0 = \tau_t - \delta$ for some $\delta > 0$, then with a similar argument $\sqrt{\delta} \lesssim \Im m_{\mathrm{fc}}^t(\tau_0 + \delta/2) \lesssim t/N$ and we have $\delta \lesssim t/N$ also in this case and the claim follows.

Claim 2

$|\gamma_t - \gamma| \lesssim (t/N)^{1/4}$, where $\gamma = \gamma_{\mathrm{edge}}$ from Definition 3.2.4. From the third equality in (3.73a) we can relate the edge-slope of m_{fc}^t to M_t''. Indeed, if $\gamma_t^{3/2}$ denotes the slope, i.e. $\varrho_{\mathrm{fc}}^t(x) = \gamma_t^{3/2}\sqrt{(\tau_t - x)_+}/\pi + o(\tau_t - x)$, then using the elementary integrals

$$\lim_{\eta\to 0} \eta^{1/2} \int_0^\infty \frac{\sqrt{x}/\pi}{(x - \mathrm{i}\eta)^2}\,\mathrm{d}x = \frac{\mathrm{i}^{1/2}}{2}, \qquad \lim_{\eta\to 0} \eta^{3/2} \int_0^\infty \frac{\sqrt{x}/\pi}{(x - \mathrm{i}\eta)^3}\,\mathrm{d}x = \frac{\mathrm{i}^{3/2}}{8}$$

we obtain the precise divergence asymptotics of the derivatives $(m_{\mathrm{fc}}^t)'(z)$ and $(m_{\mathrm{fc}}^t)''(z)$ as $z = \tau_t + \mathrm{i}\eta \to \tau_t$ and conclude

$$\frac{2}{\gamma_t^3} = \lim_{z\to\tau_t} \frac{(ct)^3 (m_{\mathrm{fc}}^t)''(z)}{(1 + ct(m_{\mathrm{fc}}^t)'(z))^3} = (ct)^3 \left\langle M_t''(\xi_t(\tau_t))\right\rangle, \quad \text{i.e.} \quad \gamma_t = \frac{\left(\left\langle M_t''(\xi_t(\tau_t))\right\rangle/2\right)^{-1/3}}{ct}.$$

We now use (3.74) at, say, $z = x := \tau_0 - \sqrt{t/N}$. By Claim 1 we have $\tau_t - x \sim \sqrt{t/N}$ and thus

$$\begin{aligned} \gamma_t^{3/2} &= \frac{\Im m_{\mathrm{fc}}^t(x)}{\sqrt{\tau_t - x}} + \mathcal{O}\left((t/N)^{1/4}\right) = \frac{\Im\left\langle M(x)\right\rangle}{\sqrt{\tau_t - x}} + \mathcal{O}\left((t/N)^{1/4}\right) \\ &= \frac{\Im\left\langle M(x)\right\rangle}{\sqrt{\tau_0 - x}} + \mathcal{O}\left((t/N)^{1/4}\right) = \gamma^{3/2} + \mathcal{O}\left((t/N)^{1/4}\right), \end{aligned}$$

where we used Claim 1 again in the third equality. This completes the proof of the claim.

Claim 3

$|\widetilde{\tau}_t - \tau_t| \prec 1/Nt$. Since M_t has a square root edge at some $\widehat{\tau}_t$, it follows from the first equality in (3.73b) that $\xi_t(\tau_t) - \widehat{\tau}_t \sim t^2$. Using rigidity in the form of Corollary 3.2.9 for the matrix $\widetilde{H}_t$ to estimate $\widetilde{G}_t'$ from below at a spectral parameter outside of the support, we have the bound

$$ct = \left|\left\langle \widetilde{G}_t'(\widetilde{\xi}_t(\widetilde{\tau}_t))\right\rangle\right|^{-1} \prec \left|\widetilde{\xi}_t(\widetilde{\tau}_t) - \widehat{\tau}_t\right|^{1/2}.$$

Consequently using the local law in the form of Lemma 3.A.1 it follows that

$$\left|\left\langle M_t'(\widetilde{\xi}_t(\widetilde{\tau}_t))\right\rangle\right| = 1/ct + \mathcal{O}_\prec(1/Nt^4) \sim 1/t,$$

whence $\widetilde{\xi}_t(\widetilde{\tau}_t) - \widehat{\tau}_t \sim t^2$ where we again used the square root singularity of $\langle M_t\rangle$ at $\widehat{\tau}_t$. We can conclude, starting from (3.73b), that

$$\begin{aligned} 0 &= \langle M_t'(\xi_t(\tau_t))\rangle - \langle \widetilde{G}_t'(\widetilde{\xi}_t(\widetilde{\tau}_t))\rangle = \langle M_t'(\xi_t(\tau_t))\rangle - \langle M_t'(\widetilde{\xi}_t(\widetilde{\tau}_t))\rangle + \langle (M_t' - \widetilde{G}_t')(\widetilde{\xi}_t(\widetilde{\tau}_t))\rangle \\ &\sim |\xi_t(\tau_t) - \widetilde{\xi}_t(\widetilde{\tau}_t)|/t^3 + \mathcal{O}_\prec(1/Nt^4), \end{aligned}$$

where we used that $|\langle M_t''(\widehat{\tau}_t + rt^2)\rangle| \sim t^{-3}$ for $c < r < C$ and the improved local law $\langle G' - M'\rangle \prec 1/N\kappa^2$ at a distance $\kappa \sim t^2$ away from the spectrum, as stated in Lemma 3.A.1. We thus find that $|\xi_t(\tau_t) - \widetilde{\xi}_t(\widetilde{\tau}_t)| \prec 1/Nt$. It remains to relate this to an estimate on $|\tau_t - \widetilde{\tau}_t|$. We have

$$|\tau_t - \widetilde{\tau}_t| \lesssim |\xi_t(\tau_t) - \widetilde{\xi}_t(\widetilde{\tau}_t)| + t|m_{\mathrm{fc}}^t(\tau_t) - m_{\mathrm{fc}}^t(\widetilde{\tau}_t)| + t|(m_{\mathrm{fc}}^t - \widetilde{m}_{\mathrm{fc}}^t)(\widetilde{\tau}_t)|,$$

where we bounded the second term by $t|\langle M_t(\xi_t(\tau_t)) - M_t(\widetilde{\xi}_t(\widetilde{\tau}_t))\rangle| \prec 1/Nt$ using

$$|\langle M_t'(\widehat{\tau}_t + rt^2)\rangle| \sim 1/t$$

and the third term by $t|\langle (M_t - \widetilde{G}_t)(\widetilde{\xi}_t(\widetilde{\tau}_t))\rangle| \prec 1/Nt$ using the local law t^2 away from $\operatorname{supp}\varrho_t$. Thus we can conclude that $|\tau_t - \widetilde{\tau}_t| \prec 1/Nt$.

Claim 4

$|\gamma_t - \widetilde{\gamma}_t| \prec 1/Nt^3$. We first note that $\gamma_t \sim 1$ follows from $|\langle M_t''(\xi_t(\tau_t))\rangle| \sim t^{-3}$. Therefore it suffices to estimate

$$\begin{aligned} t^3|\langle M_t''(\xi_t(\tau_t)) - \widetilde{G}_t''(\widetilde{\xi}_t(\widetilde{\tau}_t))\rangle| &\le t^3|\langle M_t''(\xi_t(\tau_t)) - M_t''(\widetilde{\xi}_t(\widetilde{\tau}_t))\rangle| \\ &\quad + t^3|\langle M_t''(\widetilde{\xi}_t(\widetilde{\tau}_t)) - \widetilde{G}_t''(\widetilde{\xi}_t(\widetilde{\tau}_t))\rangle| \prec \frac{1}{Nt^3}, \end{aligned}$$

as follows from $\langle M_t'''(\widetilde{\tau}_t + rt^2)\rangle \sim t^{-5}$ for $c < r < C$ and the local law from Lemma 3.A.1 at a distance of $\kappa \sim t^2$ away from the spectrum. Thus we have $|\gamma_t - \widetilde{\gamma}_t| \prec 1/Nt^3$.

We now check that $\widetilde{H}_t$ is η_*-regular in the sense of [122, Definition 2.1] for $\eta_* := N^{-2/3+\epsilon}$. It follows from the local law that $c\varrho_t(z) \prec \Im\langle \widetilde{G}_t(z)\rangle \prec C\varrho_t(z)$ for some constants c, C, whenever $\Im z \ge \eta_*$. Now (2.4)–(2.5) in [122] follow in high probability from the assumption that ϱ_t has a regular edge at τ_t . Furthermore, the absence of eigenvalues in the interval $[\tau_t + \eta_*, \tau_t + c/2]$ with high probability follows directly from Corollary 3.2.7. Finally, $\|\widetilde{H}_t\| \le N$ with high probability follows directly from $\|\widetilde{H}_t\| \le (\mathrm{Tr}|\widetilde{H}_t|^2)^{1/2}$. We can thus conclude that with high probability, $\widetilde{H}_t$ is $\eta_* = N^{-2/3+\epsilon}$ regular for any positive $\epsilon > 0$.

We denote the eigenvalues of $H_t = \widetilde{H}_t + c\sqrt{t}U$ by $\lambda_1^t \le \cdots \le \lambda_N^t$. Then it follows from [122, Theorem 2.2] that for $N^{-\epsilon} \ge t \ge N^{-2/3+\epsilon}$ with high probability for test functions $F\colon \mathbb{R}^{k+1} \to \mathbb{R}$ with $\|F\|_\infty + \|\nabla F\|_\infty \lesssim 1$ there exists some $c > 0$ such that

$$\begin{aligned} \Big|\mathbf{E}\Big[F\Big(\widetilde{\gamma}_t N^{2/3}(\lambda_{i_0}^t - \widetilde{\tau}_t), \dots, \widetilde{\gamma}_t N^{2/3}(\lambda_{i_0-k}^t - \widetilde{\tau}_t)\Big)\big|\widetilde{H}_t\Big] \\ - \mathbf{E}\, F\Big(N^{2/3}(\mu_N - 2), \dots, N^{2/3}(\mu_{N-k} - 2)\Big)\Big| \le N^{-c}. \end{aligned} \tag{3.75}$$

By combining (3.75) with $|\tau_0 - \widetilde{\tau}_t| \prec N^{-2/3-\epsilon}$, $|\gamma - \widetilde{\gamma}_t| \prec N^{-\epsilon}$ from Claims 1–4, we obtain

$$\begin{aligned} \Big|\mathbf{E}\Big[F\Big(\gamma N^{2/3}(\lambda_{i_0}^t - \tau_0), \dots, \gamma N^{2/3}(\lambda_{i_0-k}^t - \tau_0)\Big)\Big] \\ - \mathbf{E}\, F\Big(N^{2/3}(\mu_N - 2), \dots, N^{2/3}(\mu_{N-k} - 2)\Big)\Big| \lesssim N^{-c} + N^{-\epsilon} \end{aligned} \tag{3.76}$$

for our choice of $t = N^{-1/3+\epsilon}$.

3.6.2 Green's Function Comparison

It remains to prove that the local correlation functions of H_t agree with those of H. We want to prove that for any fixed $x_i \in \mathbb{R}$,

$$\lim_{N\to\infty} \mathbf{P}\left(N^{2/3}(\lambda_{i_0-i}^t - \tau_0) \ge x_i,\, i = 0, \dots, k\right)$$

is independent of t as long as, say, $t \le N^{-1/3+\epsilon}$. We first note that the local law holds uniformly in t also for H_t. This follows easily from the fact that the assumptions stay uniformly satisfied along the flow because expectation and covariance are preserved while higher order cumulants also remain unchanged up to a multiplication with a t-dependent constant. For $l = N^{-2/3-\epsilon/3}$, $\eta = N^{-2/3-\epsilon}$, and smooth monotonous cut-off functions K_i with

$K_i(x)=0$ for $x\le i-1$ and $K_i(x)=1$ for $x\ge i$ we have

$$\begin{aligned}\mathbf{E}\prod_{i=0}^{k}K_{i_0-i}&\left(\frac{\Im}{\pi}\int_{x_iN^{-2/3}+l}^{N^{-2/3+\epsilon}}\operatorname{Tr}G_t(x+\tau_0+\mathrm{i}\eta)\,\mathrm{d}x\right)-\mathcal{O}\left(N^{-\epsilon/9}\right)\\ &\le\mathbf{P}\left(N^{2/3}(\lambda_{i_0-i}^t-\tau_0)\ge x_i,\,i=0,\dots,k\right)\\ &\le\mathbf{E}\prod_{i=0}^{k}K_{i_0-i}\left(\frac{\Im}{\pi}\int_{x_iN^{-2/3}-l}^{N^{-2/3+\epsilon}}\operatorname{Tr}G_t(x+\tau_0+\mathrm{i}\eta)\,\mathrm{d}x\right)+\mathcal{O}\left(N^{-\epsilon/9}\right).\end{aligned}\tag{3.77}$$

We note that the strategy of expressing k-point correlation functions of edge-eigenvalues through a regularized expression involving the resolvent was already used in [82, 118, 123, 104] for proving edge universality. The precise formula (3.77) has been already used, for example, in [104, Eq. (4.8)].

In order to compare the expectations in (3.77) at times $t=0$ and $t=N^{-1/3+\epsilon}$, we claim that we have the bound

$$X_y:=\Im\int_{yN^{-2/3}\pm l}^{N^{-2/3+\epsilon}}\operatorname{Tr}G_t(\tau_0+x+\mathrm{i}\eta)\,\mathrm{d}x,\quad\left|\mathbf{E}\,g(X_{x_0},\dots,X_{x_k})\frac{\mathrm{d}X_{x_j}}{\mathrm{d}t}\right|\lesssim N^{1/6+3\epsilon}\tag{3.78}$$

for any $0\le j\le k$ and smooth function g. Assuming (3.78), it follows for the smooth functions K_j and by Taylor expansion that that for $t\lesssim N^{-1/3+\epsilon}$,

$$\begin{aligned}&\Bigg|\mathbf{E}\prod_{i=0}^{k}K_{i_0-i}\left(\frac{\Im}{\pi}\int_{x_iN^{-2/3}\pm l}^{N^{-2/3+\epsilon}}\operatorname{Tr}G_t(x+\tau_0+\mathrm{i}\eta)\,\mathrm{d}x\right)\\ &\quad-\mathbf{E}\prod_{i=0}^{k}K_{i_0-i}\left(\frac{\Im}{\pi}\int_{x_iN^{-2/3}\pm l}^{N^{-2/3+\epsilon}}\operatorname{Tr}G_0(x+\tau_0+\mathrm{i}\eta)\,\mathrm{d}x\right)\Bigg|\lesssim\frac{1}{N^{1/6-4\epsilon}}.\end{aligned}$$

Together with (3.77) we obtain for any k,x_i

$$\begin{aligned}&\mathbf{P}\left(N^{2/3}(\lambda_{i_0-i}^t-\tau_0)\ge x_i,\,i=0,\dots,k\right)\\ &\qquad=\mathbf{P}\left(N^{2/3}(\lambda_{i_0-i}^0-\tau_0)\ge x_i,\,i=0,\dots,k\right)+\mathcal{O}\left(N^{-\epsilon/9}\right).\end{aligned}\tag{3.79}$$

Eq. (3.78) for $g\equiv1$ follows from Itô's lemma in the form

$$\mathbf{E}\,\frac{\mathrm{d}f(H)}{\mathrm{d}t}=\mathbf{E}\left[-\frac{1}{2}\sum_\alpha w_\alpha(\partial_\alpha f)(H)+\frac{1}{2}\sum_{\alpha,\beta}\kappa(\alpha,\beta)(\partial_\alpha\partial_\beta f)(H)\right]$$

and the general neighbourhood cumulant expansion involving *pre-cumulants*, as introduced in Proposition 2.3.2. This expansion formula was a key input to the Green's function comparison argument in the spectral bulk in Corollary 2.2.6 for correlated matrix models under Assumptions (3.CD). Given the local law, Theorem 3.2.6, the extension of this proof to the edge is a routine power counting argument even for $g\not\equiv1$ and is left to the reader.

Proof of Theorem 3.2.11. The theorem follows directly from (3.76) and (3.79). □

3.A Auxiliary results

Proof of Lemma 3.3.4. From (2.70a)–(2.70b) we have[6]

$$\|M\mathcal{S}[R]R\|_* \lesssim N^{1/2K}\,\|R\|_*^2, \qquad \|MR\|_* \lesssim N^{1/2K}\,\|R\|_*$$

and furthermore by a three term geometric expansion also

$$\begin{aligned}\left\|\mathcal{B}^{-1}\mathcal{Q}\right\|_{*\to*} &\le (1+\|\mathcal{Q}\|_{*\to*})\\ &\quad\times\Big(1+\|\mathcal{C}_M\mathcal{S}\|_{*\to*}+\|\mathcal{C}_M\mathcal{S}\|_{*\to\mathrm{hs}}\,\|\mathcal{B}^{-1}\mathcal{Q}\|_{\mathrm{sp}}\,\|\mathcal{C}_M\mathcal{S}\|_{\mathrm{hs}\to*}\Big).\end{aligned}$$

Since

$$\begin{aligned}\|\mathcal{P}[R]\|_* = \frac{|\langle P,R\rangle|}{|\langle P,B\rangle|}\,\|B\|_* &\le \frac{\|B\|}{|\langle P,B\rangle|\,N}\sum_a |R_{P^*_{a\cdot}a}|\\ &\le \frac{\|B\|\,\|R\|_*}{|\langle P,B\rangle|\,N}\sum_a \|P^*_{a\cdot}\| \le \frac{\|P\|\,\|B\|}{|\langle P,B\rangle|}\,\|R\|_*\end{aligned}$$

it follows that $\|\mathcal{P}\|_{*\to*}\lesssim 1$ and therefore also $\|\mathcal{Q}\|_{*\to*}\lesssim 1$. Now, since $\|R\|_{\max}\le\|R\|_*\le\|R\|$ and according to (2.73) also $\max\{\|\mathcal{S}\|_{\max\to\|\cdot\|},\|\mathcal{S}\|_{\mathrm{hs}\to\|\cdot\|}\}\lesssim 1$, the lemma follows together with $\|\mathcal{B}^{-1}\mathcal{Q}\|_{\mathrm{sp}}\lesssim 1$ from Proposition 3.3.1(iv). □

Lemma 3.A.1. *Fix any $\epsilon,\delta>0$ and an integer $k\ge 0$. Under the assumptions of Theorem 3.2.6, for the k-th derivatives of M and G we have the bound*

$$\left|\langle G^{(k)}(z)-M^{(k)}(z)\rangle\right| \prec \frac{1}{N\kappa^{k+1}}. \tag{3.80}$$

uniformly in $z\in\mathbb{D}^\delta$ with $\kappa=\operatorname{dist}(z,\operatorname{supp}\varrho)\ge N^{-2/3+\epsilon}$.

Proof. We will fix $z=x+\mathrm{i}\eta$ throughout the proof. Let $\chi\colon\mathbb{R}\to\mathbb{R}$ be a smooth cut-off function such that $\chi(x')=1$ for $\kappa'=\operatorname{dist}(x',\operatorname{supp}\varrho)\le\kappa/3$ and $\chi(x')=0$ for $\kappa'\ge 2\kappa/3$ and let $\widetilde{\chi}$ be a cut-off function such that $\widetilde{\chi}(\eta')=1$ for $\eta'\le 1$ and $\widetilde{\chi}(\eta')=0$ for $\eta'\ge 2$. We also assume that the cut-off functions have bounded derivatives in the sense $\|\chi'\|_\infty\lesssim 1/\kappa$, $\|\chi''\|_\infty\lesssim 1/\kappa^2$ and $\|\widetilde{\chi}'\|_\infty\lesssim 1$. We now define $f(x'):=(x'-z)^{-k}\chi(x')$ and the almost analytic extension

$$\begin{aligned}f^{\mathbb{C}}(z')=f^{\mathbb{C}}(x'+\mathrm{i}\eta')&:=\widetilde{\chi}(\eta')\Big[f(x')+\mathrm{i}\eta' f'(x')\Big],\\ \partial_{\overline{z}}f^{\mathbb{C}}(z')&=\frac{\mathrm{i}\eta'}{2}\widetilde{\chi}(\eta')f''(x')+\frac{\mathrm{i}}{2}\widetilde{\chi}'(\eta')\Big[f(x')+\mathrm{i}\eta' f'(x')\Big].\end{aligned}$$

It follows from the Cauchy Theorem and the absence of eigenvalues outside $\{\,\chi=1\,\}$ in the sense of Corollary 3.2.7 that with high probability

$$\langle G^{(k)}(z)-M^{(k)}(z)\rangle=\frac{2}{\pi}\Re\int_{\mathbb{R}}\int_{\mathbb{R}_+}\partial_{\overline{z}}f^{\mathbb{C}}(z')\,\langle G(z')-M(z')\rangle\,\mathrm{d}\eta'\,\mathrm{d}x'.$$

[6] C.f. Remark 3.3.2 for the applicability of these bounds in the present setup.

Due to the fact that $\widetilde{\chi}' = 0$ for $\eta' \le 1$ the second term in $\partial_{\overline{z}} f^{\mathbb{C}}$ only gives a contribution of $1/N\kappa^{k+1}$ even by the local law and the $\|\cdot\|_\infty$ bound for $\partial_{\overline{z}} f^{\mathbb{C}}$ and we now concentrate on the first term. First, we exclude the integration regime $\eta' \lesssim N^{-1+\gamma}$ in which we cannot use the local law but only the trivial bound $\langle G - M\rangle \lesssim 1/\eta'$. For the contribution of this regime to (3.80) we thus have to estimate

$$\begin{aligned} &N^{-1+\gamma} \int_{\mathbb{R}} |f''(x')| \,\mathrm{d}x' \\ &\lesssim \frac{1}{N} \int_{|x-x'|\ge 2\kappa/3} \left[\frac{1}{\kappa^2 |x-x'|^k} + \frac{1}{\kappa |x-x'|^{k+1}} + \frac{1}{|x-x'|^{k+2}} \right] \mathrm{d}x' \lesssim \frac{N^\gamma}{N\kappa^{k+1}} \end{aligned}$$

and we have shown that

$$\begin{aligned} &\left| \langle G^{(k)}(z) - M^{(k)}(z) \rangle \right| \prec \frac{N^\gamma}{N\kappa^{k+1}} \\ &+ \int_{\mathbb{R}} \int_{N^{-1+\gamma}}^{2} \eta' \left[\frac{\chi(x')}{|x'-z|^{k+2}} + \frac{\chi'(x')}{|x'-z|^{k+1}} + \frac{\chi''(x')}{|x'-z|^{k}} \right] |\langle G(z') - M(z') \rangle| \,\mathrm{d}\eta' \,\mathrm{d}x'. \end{aligned}$$

We now use the local law of the form $|\langle G - M\rangle| \prec 1/N(\kappa + \eta')$ and that in the second and third term the integration regime is only of order κ to obtain the final bound of $N^\gamma/N\kappa^{k+1}$ for any $\gamma > 0$. □

Cusp Universality for Random Matrices I: Local Law and the Complex Hermitian Case 4

For complex Wigner-type matrices, i.e. Hermitian random matrices with independent, not necessarily identically distributed entries above the diagonal, we show that at any cusp singularity of the limiting eigenvalue distribution the local eigenvalue statistics are universal and form a Pearcey process. Since the density of states typically exhibits only square root or cubic root cusp singularities, our work complements previous results on the bulk and edge universality and it thus completes the resolution of the Wigner-Dyson-Mehta universality conjecture for the last remaining universality type in the complex Hermitian class. Our analysis holds not only for exact cusps, but approximate cusps as well, where an extended Pearcey process emerges. As a main technical ingredient we prove an optimal local law at the cusp for both symmetry classes. This result is also used in the companion paper [DS6] where the cusp universality for real symmetric Wigner-type matrices is proven.

Published as L. Erdős, T. Krüger, and D. Schröder, *Cusp universality for random matrices I: Local law and the complex Hermitian case*, preprint (2018), `arXiv:1809.03971`.

4.1 Introduction

The celebrated Wigner-Dyson-Mehta (WDM) conjecture asserts that local eigenvalue statistics of large random matrices are universal: they only depend on the symmetry type of the matrix and are otherwise independent of the details of the distribution of the matrix ensemble. This remarkable spectral robustness was first observed by Wigner in the bulk of the spectrum. The correlation functions are determinantal and they were computed in terms the *sine kernel* via explicit Gaussian calculations by Dyson, Gaudin and Mehta [135]. Wigner's vision continues to hold at the spectral edges, where the correct statistics was identified by Tracy and Widom for both symmetry types in terms of the *Airy kernel* [170, 171]. These universality results have been originally formulated and proven [161, 74, 168, 75, 167, 41] for traditional *Wigner matrices*, i.e. Hermitian random matrices with independent, identically distributed (i.i.d.) entries and their diagonal [123, 126] and non-diagonal [117] deformations. More recently they have been extended to *Wigner-type ensembles*, where the identical distribution is not required, and even to a large class of matrices with general correlated entries [9, DS4, 8]. In different directions of generalization, sparse matrices [72, 2, 103, 124], adjacency matrices of regular graphs [26] and band matrices [160, 44, 46] have also been considered. In parallel developments bulk and edge universal statistics have been proven for invariant β-ensembles [62, 145, 64, 144, 156, 174, 155, 42, 41, 28, 119, 56] and even for their discrete analogues [38, 108, 21, 93] but often with very different methods.

A precondition for the Tracy-Widom distribution in all these generalizations of Wigner's original ensemble is that the density of states vanishes as a square root near the spectral edges. The recent classification of the singularities of the solution to the underlying Dyson equation indeed revealed that at the edges only square root singularities appear [10, 12]. The density of states may also form a cusp-like singularity in the interior of the asymptotic spectrum, i.e. single points of vanishing density with a cubic root growth behaviour on either side. Under very general conditions, no other type of singularity may occur. At the cusp a new local eigenvalue process emerges: the correlation functions are still determinantal but the *Pearcey kernel* replaces the sine- or the Airy kernel.

The Pearcey process was first established by Brézin and Hikami for the eigenvalues close to a cusp singularity of a deformed complex Gaussian Wigner (GUE) matrix. They considered the model of a GUE matrix plus a deterministic matrix ("external source") having eigenvalues ± 1 with equal multiplicity [50, 49]. The name *Pearcey kernel* and the corresponding *Pearcey process* have been coined by [172] in reference to related functions introduced by Pearcey in the context of electromagnetic fields [146]. Similarly to the universal sine and Airy processes, it has later been observed that also the Pearcey process universality extends beyond the realm of random matrices. Pearcey statistics have been established for non-intersecting Brownian bridges [4] and in skew plane partitions [140], always at criticality. We remark, however, that critical cusp-like singularity does not always induce a Pearcey kernel, see e.g. [66].

In random matrix theory there are still only a handful of rather specific models for which the emergence of the Pearcey process has been proven. This has been achieved for deformed GUE matrices [5, 3, 52] and for Gaussian sample covariance matrices [95, 97, 96] by a contour integration method based upon the Brézin-Hikami formula. Beyond linear deformations, the Riemann-Hilbert method has been used for proving Pearcey statistics for a certain *two-matrix model* with a special quartic potential with appropriately tuned coefficients [87]. All these previous results concern only specific ensembles with a matrix integral representation. In particular, Wigner-type matrices are out of the scope of this approach.

The main result of the current paper is the proof of the Pearcey universality at the cusps for complex Hermitian Wigner-type matrices under very general conditions. Since the classification theorem excludes any other singularity, this is the third and last universal statistics that emerges from natural generalizations of Wigner's ensemble.

This third universality class has received somewhat less attention than the other two, presumably because cusps are not present in the classical Wigner ensemble. We also note that the most common invariant β-ensembles do not exhibit the Pearcey statistics as their densities do not feature cubic root cusps but are instead $1/2$-Hölder continuous for somewhat regular potentials [61]. The density vanishes either as $2k$-th or $(2k+\frac{1}{2})$-th power with their own local statistics (see [57] also for the persistence of these statistics under small additive GUE perturbations before the critical time). Cusp singularities, hence Pearcey statistics, however, naturally arise within any one-parameter family of Wigner-type ensembles whenever two spectral bands merge as the parameter varies. The classification theorem implies that cusp formation is the only possible way for bands to merge, so in that sense Pearcey universality is ubiquitous as well.

The bulk and edge universality is characterized by the symmetry type alone: up to a natural shift and rescaling there is only one bulk and one edge statistic. In contrast, the cusp universality has a much richer structure: it is naturally embedded in a one-parameter family of universal statistics within each symmetry class. In the complex Hermitian case these are given by the one-parameter family of (extended) Pearcey kernels, see (4.5) later. Thinking in terms of fine-tuning a single parameter in the space of Wigner-type ensembles, the density of states already exhibits a universal local shape right before and right after the cusp formation; it features a tiny gap or a tiny nonzero local minimum, respectively [7, 12]. When the local lengthscale ℓ of these *almost cusp* shapes is comparable with the local eigenvalue spacing δ, then the general Pearcey statistics is expected to emerge whose parameter is determined by the ratio ℓ/δ. Thus the full Pearcey universality typically appears in a *double scaling limit*.

Our proof follows the *three step strategy* that is the backbone of the recent approach to the WDM universality, see [78] for a pedagogical exposé and for detailed history of the method. The first step in this strategy is a *local law* that identifies, with very high probability, the empirical eigenvalue distribution on a scale slightly above the typical eigenvalue spacing. The second step is to prove universality for ensembles with a tiny Gaussian component. Finally, in the third step this Gaussian component is removed by perturbation theory. The local law is used for precise apriori bounds in the second and third steps.

The main novelty of the current paper is the proof of the local law at optimal scale near the cusp. To put the precision in proper context, we normalize the $N \times N$ real symmetric or complex Hermitian Wigner-type matrix H to have norm of order one. As customary, the local law is formulated in terms of the Green function $G(z) := (H - z)^{-1}$ with spectral parameter z in the upper half plane. The local law then asserts that $G(z)$ becomes deterministic in the large N limit as long as $\eta := \Im z$ is much larger than the local eigenvalue spacing around $\Re z$. The deterministic approximant $M(z)$ can be computed as the unique solution of the corresponding Dyson equation (see (4.2) and (4.9) later). Near the cusp the typical eigenvalue spacing is of order $N^{-3/4}$; compare this with the N^{-1} spacing in the bulk and $N^{-2/3}$ spacing near the edges. We remark that a local law at the cusp on the non-optimal scale $N^{-3/5}$ has already been proven in [9]. In the current paper we improve this result to the optimal scale $N^{-3/4}$ and this is essential for our universality proof at the cusp.

The main ingredient behind this improvement is an optimal estimate of the error term D (see (4.12) later) in the approximate Dyson equation that $G(z)$ satisfies. The difference $M - G$ is then roughly estimated by $\mathcal{B}^{-1}(MD)$, where $\mathcal{B}$ is the linear stability operator of the Dyson equation. Previous estimates on D (in averaged sense) were of order $\rho/N\eta$, where ρ is the local density; roughly speaking $\rho \sim 1$ in the bulk, $\rho \sim N^{-1/3}$ at the edge and $\rho \sim N^{-1/4}$ near the cusp. While this estimate cannot be improved in general, our main observation is that, to leading order, we need only the projection of MD in the single unstable direction of $\mathcal{B}$. We found that this projection carries an extra hidden cancellation due to a special local symmetry at the cusp and thus the estimate on D effectively improves to $\rho^2/N\eta$. Customary power counting is not sufficient, we need to compute this error term explicitly at least to leading order. We call this subtle mechanism *cusp fluctuation averaging* since it combines the well established fluctuation averaging procedure with the additional cancellation at the cusp. Similar estimates extend to the vicinity of the exact cusps. We identify a key quantity, denoted by $\sigma(z)$ (in (4.13a) later), that measures the distance from the cusp in a canonical way: $\sigma(z) = 0$ characterizes an exact cusp, while $|\sigma(z)| \ll 1$ indicates that z is near an almost cusp. Our final estimate on D is of order $(\rho + |\sigma|)\rho/N\eta$. Since the error term D is random and we need to control it in high moment sense, we need to lift this idea to a high moment calculation, meticulously extracting the improvement from every single term. This is performed in the technically most involved Section 4.4 where we use a Feynman diagrammatic formalism to bookkeep the contributions of all terms. Originally we have developed this language in [DS3] to handle random matrices with slow correlation decay. In the current paper we incorporate the cusp into this analysis. We identify a finite set of Feynman subdiagrams, called *σ-cells* (Definition 4.4.10) with value σ that embody the cancellation effect at the cusp. To exploit the full strength of the cusp fluctuation averaging mechanism, we need to trace the fate of the σ-cells along the high moment expansion. The key point is that σ-cells are local objects in the Feynman graphs thus their cancellation effects act simultaneously and the corresponding gains are multiplicative.

Formulated in the jargon of diagrammatic field theory, extracting the deterministic Dyson equation for M from the resolvent equation $(H - z)G(z) = 1$ corresponds to a consistent self-energy renormalization of G. One way or another, such procedure is behind every proof of the optimal local law with high probability. Our σ-cells conceptually correspond to a next order resummation of certain Feynman diagrams carrying a special cancellation.

We remark that we prove the optimal local law only for Wigner-type matrices and not yet for general correlated matrices unlike in [DS3, DS4]. In fact we use the simpler setup only for the estimate on D (Theorem 4.3.7) the rest of the proof is already formulated for the general case. This simpler setup allows us to present the cusp fluctuation averaging mechanism with the least amount of technicalities. The extension to the correlated case is based on the same mechanism but it requires considerably more involved diagrammatic manipulations which is better to develop in a separate work to contain the length of this paper.

Armed with the optimal local law we then perform the other two steps of the three step analysis. The third step, relying on the *Green function comparison theorem*, is fairly standard and previous proofs used in the bulk and at the edge need only minor adjustments. The second step, extracting universality from an ensemble with a tiny Gaussian component can be done in two ways: (i) Brézin-Hikami formula with contour integration or (ii) Dyson Brownian Motion (DBM). Both methods require the local law as an input. In the current work we follow (i) mainly because this approach directly yields the Pearcey kernel, at least for the complex Hermitian symmetry class. In the companion work [DS6] we perform the DBM analysis adapting methods of [122, 77, 121] to the cusp. The main novelty in the current work and in [DS6] is the rigidity at the cusp on the optimal scale provided below. Once this key input is given, the proof of the edge universality from [121] is modified in [DS6] to the cusp setting, proving universality for the real symmetric case as well. We remark, however, that, to our best knowledge, the analogue of the Pearcey kernel for the real symmetric case has not yet been explicitly identified.

We now explain some novelty in the contour integration method. We first note that a similar approach was initiated in the fundamental work of Johansson on the bulk universality for Wigner matrices with a large Gaussian component in [109]. This method was generalised later to Wigner matrices with a small Gaussian component in [74] as well as it inspired the proof of bulk universality via the moment matching idea [167] once the necessary local law became available. The double scaling regime has also been studied, where the density is very small but the Gaussian component compensates for it [58]. More recently, the same approach was extended to the cusp for deformed GUE matrices [52, Theorem 1.3] and for sample covariance matrices but only for large Gaussian component [95, 97, 96]. For our cusp universality, we need to perform a similar analysis but with a small Gaussian component. We represent our matrix H as $\widehat{H} + \sqrt{t}U$, where U is GUE and $\widehat{H}$ is an independent Wigner-type matrix. The contour integration analysis (Section 4.5.1) requires a Gaussian component of size at least $t \gg N^{-1/2}$.

The input of the analysis in Section 4.5.1 for the correlation kernel of H is a very precise description of the eigenvalues of $\widehat{H}$ just above $N^{-3/4}$, the scale of the typical spacing between eigenvalues — this information is provided by our optimal local law. While in the bulk and in the regime of the regular edge finding an appropriate $\widehat{H}$ is a relatively simple matter, in the vicinity of a cusp point the issue is very delicate. The main reason is that the cusp, unlike the bulk or the regular edge, is unstable under small perturbations; in fact it typically disappears and turns into a small positive local minimum if a small GUE component is added. Conversely, a cusp emerges if a small GUE component is added to an ensemble that has a density with a small gap. In particular, even if the density function $\rho(\tau)$ of H exhibits an exact cusp, the density $\widehat{\rho}(\tau)$ of $\widehat{H}$ will have a small gap: in fact ρ is given by the evolution of the semicircular flow up to time t with initial data $\widehat{\rho}$. Unlike in the bulk and edge cases, here one cannot match the density of H and $\widehat{H}$ by a simple shift and rescaling. Curiously, the contour integral analysis for the local statistics of H at the cusp relies on an optimal local law of $\widehat{H}$ with a small gap far away from the cusp.

Thus we need an additional ingredient: the precise analysis of the semicircular flow $\rho_s := \widehat{\rho} \boxplus \rho_{\mathrm{sc}}^{(s)}$ near the cusp up to a relatively long times $s \lesssim N^{-1/2+\epsilon}$; note that $\rho_t = \rho$ is the original density with the cusp. Here $\rho_{\mathrm{sc}}^{(s)}$ is the semicircular density with variance s and $\boxplus$ indicates the free convolution. In Sections 4.5.2–4.5.3 we will see that the edges of the support of the density ρ_s typically move linearly in the time s while the gap closes at a much slower rate. Already $s \gg N^{-3/4}$ is beyond the simple perturbative regime of the cusp whose natural lengthscale is $N^{-3/4}$. Thus we need a very careful tuning of the parameters: the analysis of a cusp for H requires constructing a matrix $\widehat{H}$ that is far from having a cusp but that after a relatively long time $t = N^{-1/2+\epsilon}$ will develop a cusp exactly at the right location. In the estimates we heavily rely on various properties of the solution to the Dyson equation established in the recent paper [12]. These results go well beyond the precision of the previous work [7] and they apply to a very general class of Dyson equations, including a non-commutative von-Neumann algebraic setup.

Notations. We now introduce some custom notations we use throughout the paper. For non-negative functions $f(A, B)$, $g(A, B)$ we use the notation $f \le_A g$ if there exist constants $C(A)$ such that $f(A, B) \le C(A)g(A, B)$ for all A, B. Similarly, we write $f \sim_A g$ if $f \le_A g$ and $g \le_A f$. We do not indicate the dependence of constants on basic parameters that will be called model parameters later. If the implied constants are universal, we instead write $f \lesssim g$ and $f \sim g$. Similarly we write $f \ll g$ if $f \le cg$ for some tiny absolute constant $c > 0$.

We denote vectors by bold-faced lower case Roman letters $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$, and matrices by upper case Roman letters $A, B \in \mathbb{C}^{N\times N}$. The standard scalar product and Euclidean norm on $\mathbb{C}^N$ will be denoted by $\langle \mathbf{x}, \mathbf{y} \rangle := N^{-1} \sum_{i \in [N]} \overline{x_i} y_i$ and $\|\mathbf{x}\|$, while we also write $\langle A, B \rangle := N^{-1} \operatorname{Tr} A^* B$ for the scalar product of matrices, and $\langle A \rangle := N^{-1} \operatorname{Tr} A$, $\langle \mathbf{x} \rangle := N^{-1} \sum_{a \in [N]} x_a$. We write $\operatorname{diag} R$, $\operatorname{diag} \mathbf{r}$ for the diagonal vector of a matrix R and the diagonal matrix obtained from a vector $\mathbf{r}$, and $S \odot R$ for the entrywise (Hadamard) product of matrices R, S. The usual operator norm induced by the vector norm $\|\cdot\|$ will be denoted by $\|A\|$, while the Hilbert-Schmidt (or Frobenius) norm will be denoted by $\|A\|_{\text{hs}} := \sqrt{\langle A, A \rangle}$. For integers n we define $[n] := \{1, \ldots, n\}$.

Acknowledgement. The authors are very grateful to Johannes Alt for numerous discussions on the Dyson equation and for his invaluable help in adjusting [12] to the needs of the present work.

4.2 Main results

4.2.1 The Dyson equation

Let $W = W^* \in \mathbb{C}^{N \times N}$ be a self-adjoint random matrix and $A = \operatorname{diag}(\boldsymbol{a})$ be a deterministic diagonal matrix with entries $\boldsymbol{a} = (a_i)_{i=1}^N \in \mathbb{R}^N$. We say that W is of *Wigner-type* [9] if its entries w_{ij} for $i \leq j$ are centred, $\mathbf{E}\, w_{ij} = 0$, independent random variables. We define the *variance matrix* or *self-energy matrix* $S = (s_{ij})_{i,j=1}^N$ by

$$s_{ij} := \mathbf{E}\, |w_{ij}|^2 . \tag{4.1}$$

This matrix is symmetric with non-negative entries. In [9] it was shown that as N tends to infinity, the resolvent $G(z) := (H - z)^{-1}$ of the *deformed Wigner-type matrix* $H = A + W$ entrywise approaches a diagonal matrix

$$M(z) := \operatorname{diag}(\mathbf{m}(z)).$$

The entries $\mathbf{m} = (m_1 \ldots, m_N) \colon \mathbb{H} \to \mathbb{H}^N$ of M have positive imaginary parts and solve the *Dyson equation*

$$-\frac{1}{m_i(z)} = z - a_i + \sum_{j=1}^N s_{ij} m_j(z), \qquad z \in \mathbb{H} := \{\, z \in \mathbb{C} \mid \Im z > 0 \,\}, \quad i \in [N]. \tag{4.2}$$

We call M or $\mathbf{m}$ the *self-consistent Green's function*. The normalised trace of M is the Stieltjes transform of a unique probability measure on $\mathbb{R}$ that approximates the empirical eigenvalue distribution of $A + W$ increasingly well as $N \to \infty$, motivating the following definition.

Definition 4.2.1 (Self-consistent density of states). *The unique probability measure ρ on $\mathbb{R}$, defined through*

$$\langle M(z) \rangle = \frac{1}{N} \operatorname{Tr} M(z) = \int \frac{\rho(\mathrm{d}\tau)}{\tau - z}, \qquad z \in \mathbb{H},$$

is called the self-consistent density of states (scDOS). Accordingly, its support $\operatorname{supp} \rho$ *is called self-consistent spectrum.*

4.2.2 Cusp universality

We make the following assumptions:

Assumption (4.A) (Bounded moments). *The entries of the Wigner-type matrix $\sqrt{N}W$ have bounded moments and the expectation A is bounded, i.e. there are positive C_k such that*

$$|a_i| \le C_0, \qquad \mathbf{E}\,|w_{ij}|^k \le C_k N^{-k/2}, \qquad k \in \mathbb{N}.$$

Assumption (4.B) (Fullness). *If the matrix $W = W^* \in \mathbb{C}^{N\times N}$ belongs to the complex hermitian symmetry class, then we assume*

$$\begin{pmatrix} \mathbf{E}(\Re w_{ij})^2 & \mathbf{E}(\Re w_{ij})(\Im w_{ij}) \\ \mathbf{E}(\Re w_{ij})(\Im w_{ij}) & \mathbf{E}(\Im w_{ij})^2 \end{pmatrix} \ge \frac{c}{N}\mathbb{1}_{2\times 2}, \tag{4.3}$$

as quadratic forms, for some positive constant $c > 0$. If $W = W^T \in \mathbb{R}^{N\times N}$ belongs to the real symmetric symmetry class, then we assume $\mathbf{E}\,w_{ij}^2 \ge \frac{c}{N}$.

Assumption (4.C) (Bounded self-consistent Green's function). *In a neighbourhood of some fixed spectral parameter $\tau \in \mathbb{R}$ the self-consistent Green's function is bounded, i.e. for positive C, κ we have*

$$|m_i(z)| \le C, \qquad z \in \tau + (-\kappa,\kappa) + \mathrm{i}\mathbb{R}^+.$$

We call the constants appearing in Assumptions (4.A)-(4.C) *model parameters*. All generic constants C in this paper may implicitly depend on these model parameters. Dependence on further parameters however will be indicated.

Remark 4.2.2. *The boundedness of $\mathbf{m}$ in Assumption (4.C) can be ensured by assuming some regularity of the variance matrix S. For more details we refer to [7, Chapter 6].*

From the extensive analysis in [12] we know that the self-consistent density ρ is described by explicit *shape functions* in the vicinity of local minima with small value of ρ and around small gaps in the support of ρ. The density in such *almost cusp regimes* is given by precisely one of the following three asymptotics:

(i) *Exact cusp*. There is a cusp point $\mathfrak{c} \in \mathbb{R}$ in the sense that $\rho(\mathfrak{c}) = 0$ and $\rho(\mathfrak{c}\pm\delta) > 0$ for $0 \ne \delta \ll 1$. In this case the self-consistent density is locally around $\mathfrak{c}$ given by

$$\rho(\mathfrak{c}\pm x) = \frac{\sqrt{3}\gamma^{4/3}}{2\pi} x^{1/3}\Big[1 + \mathcal{O}\left(x^{1/3}\right)\Big], \qquad x \ge 0 \tag{4.4a}$$

for some $\gamma > 0$.

(ii) *Small gap*. There is a maximal interval $[\mathfrak{e}_-, \mathfrak{e}_+]$ of size $0 < \Delta := \mathfrak{e}_+ - \mathfrak{e}_- \ll 1$ such that $\rho|_{[\mathfrak{e}_-,\mathfrak{e}_+]} \equiv 0$. In this case the density around $\mathfrak{e}_\pm$ is, for some $\gamma > 0$, locally given by

$$\rho(\mathfrak{e}_\pm \pm x) = \frac{\sqrt{3}(2\gamma)^{4/3}\Delta^{1/3}}{2\pi}\Psi_{\mathrm{edge}}(x/\Delta)\left[1 + \mathcal{O}\left(\Delta^{1/3}\Psi_{\mathrm{edge}}(x/\Delta)\right)\right] \tag{4.4b}$$

for $x \ge 0$, where the shape function around the edge is given by

$$\Psi_{\mathrm{edge}}(\lambda) := \frac{\sqrt{\lambda(1+\lambda)}}{(1+2\lambda+2\sqrt{\lambda(1+\lambda)})^{2/3} + (1+2\lambda-2\sqrt{\lambda(1+\lambda)})^{2/3} + 1} \tag{4.4c}$$

for $\lambda \ge 0$

(iii) *Non-zero local minimum.* There is a local minimum at $\mathfrak{m} \in \mathbb{R}$ of ρ such that $0 < \rho(\mathfrak{m}) \ll 1$. In this case there exists some $\gamma > 0$ such that

$$\rho(\mathfrak{m}+x) = \rho(\mathfrak{m}) + \rho(\mathfrak{m})\Psi_{\min}\left(\frac{3\sqrt{3}\gamma^4 x}{2(\pi\rho(\mathfrak{m}))^3}\right)\left[1+\mathcal{O}\left(\rho(\mathfrak{m})^{1/2}+\frac{|x|}{\rho(\mathfrak{m})^3}\right)\right], \tag{4.4d}$$

for $x \in \mathbb{R}$, where the shape function around the local minimum is given by

$$\Psi_{\min}(\lambda) := \frac{\sqrt{1+\lambda^2}}{(\sqrt{1+\lambda^2}+\lambda)^{2/3}+(\sqrt{1+\lambda^2}-\lambda)^{2/3}-1} - 1, \qquad \lambda \in \mathbb{R}. \tag{4.4e}$$

We note that the parameter γ in (4.4a) is chosen in a way which is convenient for the universality statement. We also note that the choices for γ in (4.4b)–(4.4d) are consistent with (4.4a) in the sense that in the regimes $\Delta \ll x \ll 1$ and $\rho(\mathfrak{m})^3 \ll |x| \ll 1$ the respective formulae asymptotically agree. Depending on the three cases (i)–(iii), we define the *almost cusp point* $\mathfrak{b}$ as the cusp $\mathfrak{c}$ in case (i), the midpoint $(\mathfrak{e}_- + \mathfrak{e}_+)/2$ in case (ii), and the minimum $\mathfrak{m}$ in case (iii). When the local length scale of the almost cusp shape starts to match the eigenvalue spacing, i.e. if $\Delta \lesssim N^{-3/4}$ or $\rho(\mathfrak{m}) \lesssim N^{-1/4}$, then we call the local shape a *physical cusp*. This terminology reflects the fact that the shape becomes indistinguishable from the exact cusp with $\rho(\mathfrak{c}) = 0$ when resolved with a precision above the eigenvalue spacing. In this case we call $\mathfrak{b}$ a *physical cusp point*.

The extended Pearcey kernel with a real parameter α (often denoted by τ in the literature) is given by

$$K_\alpha(x,y) = \frac{1}{(2\pi \mathrm{i})^2}\int_\Xi \mathrm{d}z \int_\Phi \mathrm{d}w \frac{\exp(-w^4/4+\alpha w^2/2 - yw + z^4/4 - \alpha z^2/2 + xz)}{w-z}, \tag{4.5}$$

where Ξ is a contour consisting of rays from $\pm\infty e^{\mathrm{i}\pi/4}$ to 0 and rays from 0 to $\pm\infty e^{-\mathrm{i}\pi/4}$, and Φ is the ray from $-\mathrm{i}\infty$ to $\mathrm{i}\infty$. The simple Pearcey kernel with parameter $\alpha = 0$ has been first observed in the context of random matrix theory by [50, 49]. We note that (4.5) is a special case of a more general extended Pearcey kernel defined in [172, Eq. (1.1)].

It is natural to express universality in terms of a rescaled k-point function $p_k^{(N)}$ which we define implicitly by

$$\binom{N}{k}^{-1} \sum_{\{i_1,\dots,i_k\}\subset[N]} f(\lambda_{i_1},\dots,\lambda_{i_k}) = \int_{\mathbb{R}^k} f(x_1,\dots,x_k)p_k^{(N)}(x_1,\dots,x_k)\,\mathrm{d}x_1\dots\mathrm{d}x_k$$

for test functions f, where the summation is over all subsets of k distinct integers from $[N]$.

Theorem 4.2.3. *Let H be a complex Hermitian Wigner matrix satisfying Assumptions (4.A)–(4.C). Assume that the self-consistent density ρ within $[\tau-\kappa,\tau+\kappa]$ from Assumption (4.C) has a physical cusp, i.e. that ρ is locally given by* (4.4) *for some $\gamma > 0$ and ρ either (i) has a cusp point $\mathfrak{c}$, or (ii) a small gap $[\mathfrak{e}_-,\mathfrak{e}_+]$ of size $\Delta := \mathfrak{e}_+ - \mathfrak{e}_- \lesssim N^{-3/4}$, or (iii) a local minimum at $\mathfrak{m}$ of size $\rho(\mathfrak{m}) \lesssim N^{-1/4}$. Then it follows that for any smooth compactly supported test function $F\colon \mathbb{R}^k \to \mathbb{R}$ it holds that*

$$\left|\int_{\mathbb{R}^k} F(x_1,\dots,x_k)\left[\frac{N^{k/4}}{\gamma^k}p_k^{(N)}\left(\mathfrak{b}+\frac{x_1}{\gamma N^{3/4}},\dots,\mathfrak{b}+\frac{x_k}{\gamma N^{3/4}}\right)\right.\right.$$
$$\left.\left. - \det(K_\alpha(x_i,x_j))_{i,j=1}^k\right]\mathrm{d}x_1\dots\mathrm{d}x_k\right| = \mathcal{O}\left(N^{-c(k)}\right),$$

where

$$\mathfrak{b} := \begin{cases} \mathfrak{c} & \textit{in case (i)} \\ (\mathfrak{e}_+ + \mathfrak{e}_-)/2 & \textit{in case (ii)} \\ \mathfrak{m} & \textit{in case (iii)} \end{cases}, \qquad \alpha := \begin{cases} 0 & \textit{in case (i)} \\ 3\,(\gamma\Delta/4)^{2/3}\,N^{1/2} & \textit{in case (ii)} \\ -\,(\pi\rho(\mathfrak{m})/\gamma)^2\,N^{1/2} & \textit{in case (iii)} \end{cases} \tag{4.6}$$

and $c(k) > 0$ *is a small constant only depending on* k*.*

4.2.3 Local law

We emphasise that the proof of Theorem 4.2.3 requires a very precise a priori control on the fluctuation of the eigenvalues even at singular points of the scDOS. This control is expressed in the form of a *local law* with an optimal convergence rate down to the typical eigenvalue spacing. We now define the scale on which the eigenvalues are predicted to fluctuate around the spectral parameter τ.

Definition 4.2.4 (Fluctuation scale). *We define the self-consistent fluctuation scale* $\eta_{\mathrm{f}} = \eta_{\mathrm{f}}(\tau)$ *through*

$$\int_{-\eta_{\mathrm{f}}}^{\eta_{\mathrm{f}}} \rho(\tau+\omega)\mathrm{d}\omega = \frac{1}{N},$$

if $\tau \in \operatorname{supp}\rho$*. If* $\tau \notin \operatorname{supp}\rho$*, then* η_{f} *is defined as the fluctuation scale at a nearby edge. More precisely, let* I *be the largest (open) interval with* $\tau \in I \subseteq \mathbb{R}\setminus\operatorname{supp}\rho$ *and set* $\Delta := \min\{|I|, 1\}$*. Then we define*

$$\eta_{\mathrm{f}} := \begin{cases} \Delta^{1/9}/N^{2/3}, & \Delta > 1/N^{3/4}, \\ 1/N^{3/4}, & \Delta \le 1/N^{3/4}. \end{cases} \tag{4.7}$$

We will see later (cf. (4.126b)) that (4.7) is the fluctuation of the edge eigenvalue adjacent to a spectral gap of length Δ as predicted by the local behaviour of the scDOS. The control on the fluctuation of eigenvalues is expressed in terms of the following local law.

Theorem 4.2.5 (Local law). *Let* H *be a deformed Wigner-type matrix of the real symmetric or complex Hermitian symmetry class. Fix any* $\tau \in \mathbb{R}$*. Assuming (4.A)–(4.C) for any* $\epsilon, \zeta > 0$ *and* $\nu \in \mathbb{N}$ *the local law holds uniformly for all* $z = \tau + \mathrm{i}\eta$ *with* $\operatorname{dist}(z, \operatorname{supp}\rho) \in [N^{\zeta}\eta_{\mathrm{f}}(\tau), N^{100}]$ *in the form*

$$\mathbf{P}\left[\,|\langle \mathbf{u}, (G(z) - M(z))\mathbf{v}\rangle| \ge N^{\epsilon}\sqrt{\frac{\rho(z)}{N\eta}}\,\|\mathbf{u}\|\,\|\mathbf{v}\|\,\right] \le \frac{C}{N^{\nu}}, \tag{4.8a}$$

for any $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$ *and*

$$\mathbf{P}\left[\,|\langle B(G(z) - M(z))\rangle| \ge \frac{N^{\epsilon}\,\|B\|}{N\operatorname{dist}(z, \operatorname{supp}\rho)}\right] \le \frac{C}{N^{\nu}}, \tag{4.8b}$$

for any $B \in \mathbb{C}^{N\times N}$*. Here* $\rho(z) := \langle \Im M(z)\rangle/\pi$ *denotes the harmonic extension of the scDOS to the complex upper half plane. The constants* $C > 0$ *in* (4.8) *only depends on* ϵ, ζ, ν *and the model parameters.*

We remark that later we will prove the local law also in a form which is uniform in $\tau \in [-N^{100}, N^{100}]$ and $\eta \in [N^{-1+\zeta}, N^{100}]$, albeit with a more complicated error term, see Proposition 4.3.11. The local law Theorem 4.2.5 implies a large deviation result for the fluctuation of eigenvalues on the optimal scale uniformly for all singularity types.

Corollary 4.2.6 (Uniform rigidity). *Let H be a deformed Wigner-type matrix of the real symmetric or complex Hermitian symmetry class satisfying Assumptions (4.A)-(4.C) for $\tau \in \operatorname{int}(\operatorname{supp}\rho)$. Then*

$$\mathbf{P}\big[\,\big|\lambda_{k(\tau)} - \tau\big| \geq N^{\epsilon}\eta_{\mathfrak{f}}(\tau)\big] \leq \frac{C}{N^{\nu}}$$

for any $\epsilon > 0$ and $\nu \in \mathbb{N}$ and some $C = C(\epsilon, \nu)$, where we defined the (self-consistent) eigenvalue index $k(\tau) := \lceil N\rho((-\infty, \tau))\rceil$, and where $\lceil x \rceil = \min\{\, k \in \mathbb{Z} \mid k \geq x \,\}$.

In particular, the fluctuation of the eigenvalue whose expected position is closest to the cusp location does not exceed $N^{-3/4+\epsilon}$ for any $\epsilon > 0$ with very high probability. The following corollary specialises Corollary 4.2.6 to the neighbourhood of a cusp.

Corollary 4.2.7 (Cusp rigidity). *Let H be a deformed Wigner-type matrix of the real symmetric or complex Hermitian symmetry class satisfying Assumptions (4.A)-(4.C) and $\tau = \mathfrak{c}$ a cusp location. Then $N\rho((-\infty, \mathfrak{c})) = k_{\mathfrak{c}}$ for some $k_{\mathfrak{c}} \in [N]$, that we call the cusp eigenvalue index. For any $\epsilon > 0$, $\nu \in \mathbb{N}$ and $k \in [N]$ with $|k - k_{\mathfrak{c}}| \leq cN$ we have*

$$\mathbf{P}\Big[\,|\lambda_k - \gamma_k| \geq \frac{N^{\epsilon}}{(1 + |k - k_{\mathfrak{c}}|)^{1/4} N^{3/4}}\Big] \leq \frac{C}{N^{\nu}},$$

where $C = C(\epsilon, \nu)$ and γ_k are the self-consistent eigenvalue locations, defined through

$$N\rho((-\infty, \gamma_k)) = k$$

.

We remark that a variant of Corollary 4.2.7 holds more generally for almost cusp points. It is another consequence of Corollary 4.2.6 that with high probability there are no eigenvalues much further than the fluctuation scale $\eta_{\mathfrak{f}}$ away from the spectrum. We note that the following corollary generalises Corollary 3.2.7 by also covering internal gaps of size $\ll 1$.

Corollary 4.2.8 (No eigenvalues outside the support of the self-consistent density). *Let $\tau \notin \operatorname{supp}\rho$. Under the assumptions of Theorem 4.2.5 we have*

$$\mathbf{P}\Big[\exists \lambda \in \operatorname{Spec} H \cap [\tau - c, \tau + c], \operatorname{dist}(\lambda, \operatorname{supp}\rho) \geq N^{\epsilon}\eta_{\mathfrak{f}}(\tau)\Big] \leq CN^{-\nu},$$

for any $\epsilon, \nu > 0$, where c and C are positive constants, depending on model parameters. The latter also depends on ϵ and ν.

Remark 4.2.9. *Theorem 4.2.5 and its consequences, Corollaries 4.2.6, 4.2.7 and 4.2.8 also hold for both symmetry classes if Assumption (4.B) is replaced by the condition that there exists an $L \in \mathbb{N}$ and $c > 0$ such that $\min_{i,j}(S^L)_{ij} \geq c/N$. A variance profile S satisfying this condition is called uniformly primitive (cf. [10, Eq. (2.5)] and [7, Eq. (2.11)]). Note that uniform primitivity is weaker than condition (4.B) on two accounts. First, it involves only the variance matrix $\mathbf{E}\,|w_{ij}|^2$ unlike (4.3) in the complex Hermitian case that also involves $\mathbf{E}\,w_{ij}^2$. Second, uniform* *primitivity allows certain matrix elements of W to vanish. In order to keep the main body of the proof conceptually simple, we will prove Theorem 4.2.5 in detail under Assumption (4.B) and we explain the necessary changes to the proof in Appendix 4.B when assuming only uniform primitivity of S.*

4.3 Local Law

In order to directly appeal to recent results on the shape of solution to Matrix Dyson Equation (MDE) from [12] and the flexible diagrammatic cumulant expansion from [DS3], we first reformulate the Dyson equation (4.2) for N-vectors $\mathbf{m}$ into a matrix equation that will approximately be satisfied by the resolvent G. This viewpoint also allows us to treat diagonal and off-diagonal elements of G on the same footing. In fact, (4.2) is a special case of

$$1 + (z - A + \mathcal{S}[M])M = 0, \tag{4.9}$$

for a matrix $M = M(z) \in \mathbb{C}^{N\times N}$ with positive definite imaginary part, $\Im M = (M - M^*)/2\mathrm{i} > 0$. The uniqueness of the solution M with $\Im M > 0$ was shown in [102]. Here the linear (*self-energy*) operator $\mathcal{S}\colon \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ is defined as $\mathcal{S}[R] := \mathbf{E}\, WRW$ and it preserves the cone of positive definite matrices. Definition 4.2.1 of the scDOS and its harmonic extension $\rho(z)$ (cf. Theorem 4.2.5) directly generalises to the solution to (4.9), see [12, Definition 2.2].

In the special case of Wigner-type matrices the self-energy operator is given by

$$\mathcal{S}[R] = \operatorname{diag}(S\mathbf{r}) + T \odot R^t, \tag{4.10}$$

where $\mathbf{r} := (r_{ii})_{i=1}^N$, S was defined in (4.1), $T = (t_{ij})_{i,j=1}^N \in \mathbb{C}^{N\times N}$ with $t_{ij} = \mathbf{E}\, w_{ij}^2 \mathbb{1}(i \neq j)$ and $\odot$ denotes the entrywise Hadamard product. The solution to (4.9) is then given by $M = \operatorname{diag}(\mathbf{m})$, where $\mathbf{m}$ solves (4.2). Note that the action of $\mathcal{S}$ on diagonal matrices is independent of T, hence the Dyson equation (4.2) for Wigner-type matrices is solely determined by the matrix S, the matrix T plays no role. However, T plays a role in analyzing the error matrix D, see (4.12) below.

The proof of the local law consists of three largely separate arguments. The first part concerns the analysis of the stability operator

$$\mathcal{B} := 1 - M\mathcal{S}[\cdot]M \tag{4.11}$$

and shape analysis of the solution M to (4.9). The second part is proving that the resolvent G is indeed an approximate solution to (4.9) in the sense that the error matrix

$$D := 1 + (z - A + \mathcal{S}[G])G = WG + \mathcal{S}[G]G \tag{4.12}$$

is small. In previous works [DS3, 9, DS4] it was sufficient to establish smallness of D in an isotropic form $\langle \mathbf{x}, D\mathbf{y}\rangle$ and averaged form $\langle BD\rangle$ with general bounded vectors/matrices $\mathbf{x}, \mathbf{y}, B$. In the vicinity of a cusp, however, it becomes necessary to establish an additional cancellation when D is averaged against the unstable direction of the stability operator $\mathcal{B}$. We call this new effect *cusp fluctuation averaging*. Finally, the third part of the proof consists of a bootstrap argument starting far away from the real axis and iteratively lowering the imaginary part $\eta = \Im z$ of the spectral parameter while maintaining the desired bound on $G - M$.

Remark 4.3.1. *We remark that the proofs of Theorem 4.2.5, and Corollaries 4.2.6, 4.2.8 use the independence assumption on the entries of W only very locally. In fact, only the proof of a specific bound on D (see* (4.23) *later), which follows directly from the main result of the diagrammatic cumulant expansion, Theorem 4.3.7, uses the vector structure and the specific form of $\mathcal{S}$ in* (4.10) *at all. Therefore, assuming* (4.23) *as an input, our proof of Theorem 4.2.5 remains valid also in the correlated setting of [DS3, DS4], as long as $\mathcal{S}$ is flat (see* (4.14) *below), and Assumption (4.C) is replaced by the corresponding assumption on the boundedness of $\|M\|$.*

For brevity we will carry out the proof of Theorem 4.2.5 only in the vicinity of almost cusps as the local law in all other regimes was already proven in [9, DS4] to optimality. Therefore, within this section we will always assume that $z = \tau + \mathrm{i}\eta = \tau_0 + \omega + \mathrm{i}\eta \in \mathbb{H}$ lies inside a small neighbourhood

$$\mathbb{D}_{\mathrm{cusp}} := \{\, z \in \mathbb{H} \mid |z - \tau_0| \le c \,\},$$

of the location τ_0 of a local minimum of the scDOS within the self-consistent spectrum $\operatorname{supp} \rho$. Here c is a sufficiently small constant depending only on the model parameters. We will further assume that either (i) $\rho(\tau_0) \ge 0$ is sufficiently small and τ_0 is the location of a cusp or internal minimum, or (ii) $\rho(\tau_0) = 0$ and τ_0 is an edge adjacent to a sufficiently small gap of length $\Delta > 0$. The results from [12] guarantee that these are the only possibilities for the shape of ρ, see (4.4). In other words, we assume that $\tau_0 \in \operatorname{supp} \rho$ is a local minimum of ρ with a shape close to a cusp (cf. (4.4)). For concreteness we will also assume that if τ_0 is an edge, then it is a right edge (with a gap of length $\Delta > 0$ to the right) and $\omega \in (-c, \frac{\Delta}{2}]$. The case when τ_0 is a left edge has the same proof.

We now introduce a quantity that will play an important role in the cusp fluctuation averaging mechanism. We define

$$\sigma(z) := \langle (\operatorname{sgn} \Re U)(\Im U/\rho)^3 \rangle\,, \quad U := \frac{(\Im M)^{-1/2}(\Re M)(\Im M)^{-1/2} + \mathrm{i}}{|(\Im M)^{-1/2}(\Re M)(\Im M)^{-1/2} + \mathrm{i}|},$$

where $\Re M := (M + M^*)/2$ is the real part of $M = M(z)$. It was proven in [12, Lemma 5.5] that $\sigma(z)$ extends to the real line as a $1/3$-Hölder continuous function wherever the scDOS ρ is smaller than some threshold $c \sim 1$, i.e. $\rho \le c$. In the specific case of $\mathcal{S}$ as in (4.10) the definition simplifies to

$$\sigma(z) := \langle \mathbf{p}\mathbf{f}^3 \rangle\,, \quad \mathbf{f} := \frac{\Im \mathbf{m}}{\rho\,|\mathbf{m}|}, \quad \mathbf{p} := \operatorname{sgn} \Re \mathbf{m}, \tag{4.13a}$$

since $M = \operatorname{diag}(\mathbf{m})$ is diagonal. When evaluated at the location τ_0 the scalar $\sigma(\tau_0)$ provides a measure of how far the shape of singularity at τ_0 is from an exact cusp. In fact, if $\sigma(\tau_0) = 0$ and $\rho(\tau_0) = 0$, then τ_0 is a cusp location. To see the relationship between the emergence of a cusp and the limit $\sigma(\tau_0) \to 0$, we refer to [12, Theorem 7.7 and Lemma 6.3]. The analogues of the quantities $\mathbf{f}, \mathbf{p}$ and σ in (4.13a) are denoted by f_u, s and σ in [12], respectively. The significance of σ for the classification of singularity types in Wigner-type ensembles was first realised in [7]. Although in this paper we will use only [12] and will not rely on [7], we remark that the definition of σ in [7, Eq. (8.11)] differs slightly from the definition (4.13a). However, both definitions equally fulfil the purpose of classifying singularity types, since the ensuing scalar quantities σ are comparable inside the self-consistent spectrum. For the interested reader, we briefly relate our notations to the respective conventions in [12] and

[7]. The quantity denoted by f in both [12] and [7] is the normalized eigendirection of the *saturated self-energy operator* F in the respective settings and is related to $\mathbf{f}$ from (4.13a) via $f = \mathbf{f}/\|\mathbf{f}\| + \mathcal{O}(\eta/\rho)$. Moreover, σ in [7] is defined as $\langle f^3 \operatorname{sgn} \Re \mathbf{m}\rangle$, justifying the comparability to σ from (4.13a).

4.3.1 Stability and shape analysis

From (4.9) and (4.12) we obtain the quadratic *stability equation*

$$\mathcal{B}[G-M] = -MD + M\mathcal{S}[G-M](G-M),$$

for the difference $G-M$. In order to apply the results of [12] to the stability operator $\mathcal{B}$, we first have to check that the flatness condition [12, Eq. (3.10)] is satisfied for the self-energy operator $\mathcal{S}$. We claim that $\mathcal{S}$ is flat, i.e.

$$\mathcal{S}[R] \sim \langle R\rangle 1 = \frac{1}{N}(\operatorname{Tr} R)1, \tag{4.14}$$

as quadratic forms for any positive semidefinite $R \in \mathbb{C}^{N\times N}$. We remark that in the earlier paper [9] in the Wigner-type case only the upper bound $s_{ij} \le C/N$ defined the concept of flatness. Here with the definition (4.14) we follow the convention of the more recent works [DS3, 12, DS4] which is more conceptual. We also warn the reader, that in the complex Hermitian Wigner-type case the condition $c/N \le s_{ij} \le C/N$ implies (4.14) only if t_{ij} is bounded away from $-s_{ij}$.

However, the flatness (4.14) is an immediate consequence of the fullness Assumption (4.B). Indeed, (4.B) is equivalent to the condition that the covariance operator Σ of all entries above and on the diagonal, defined as $\Sigma_{ab,cd} := \mathbf{E}\, w_{ab}w_{cd}$, is uniformly strictly positive definite. This implies that $\Sigma \ge c\Sigma_{\mathrm{G}}$ for some constant $c \sim 1$, where Σ_{G} is the covariance operator of a GUE or GOE matrix, depending on the symmetry class we consider. This means that $\mathcal{S}$ can be split into $\mathcal{S} = \mathcal{S}_0 + c\mathcal{S}_{\mathrm{G}}$, where $\mathcal{S}_{\mathrm{G}}$ and $\mathcal{S}_0$ are the self-energy operators corresponding to Σ_{G} and $\Sigma - c\Sigma_{\mathrm{G}}$, respectively. It is now an easy exercise to check that $\mathcal{S}_{\mathrm{G}}$ and thus $\mathcal{S}$ is flat.

In particular, [12, Proposition 3.5 and Lemma 4.8] are applicable implying that [12, Assumption 4.5] is satisfied. Thus, according to [12, Lemma 5.1] for spectral parameters z in a neighbourhood of τ_0 the operator $\mathcal{B}$ has a unique isolated eigenvalue β of smallest modulus and associated right $\mathcal{B}[V_{\mathrm{r}}] = \beta V_{\mathrm{r}}$ and left $\mathcal{B}^*[V_{\mathrm{l}}] = \overline{\beta} V_{\mathrm{l}}$ eigendirections normalised such that $\|V_{\mathrm{r}}\|_{\mathrm{hs}} = \langle V_{\mathrm{l}}, V_{\mathrm{r}}\rangle = 1$. We denote the spectral projections to V_{r} and to its complement by $\mathcal{P} := \langle V_{\mathrm{l}}, \cdot\rangle V_{\mathrm{r}}$ and $\mathcal{Q} := 1 - \mathcal{P}$. For convenience of the reader we now collect some important quantitative information about the stability operator and its unstable direction from [12].

Proposition 4.3.2 (Properties of the MDE and its solution). *The following statements hold true uniformly in $z = \tau_0 + \omega + \mathrm{i}\eta \in \mathbb{D}_{\mathrm{cusp}}$ assuming flatness as in* (4.14) *and the uniform boundedness of $\|M\|$ for $z \in \tau_0 + (-\kappa, \kappa) + \mathrm{i}\mathbb{R}_+$,*

(i) The eigendirections $V_{\mathrm{l}}, V_{\mathrm{r}}$ are norm-bounded and the operator $\mathcal{B}^{-1}$ is bounded on the complement to its unstable direction, i.e.

$$\left\|\mathcal{B}^{-1}\mathcal{Q}\right\|_{\mathrm{hs}\to\mathrm{hs}} + \|V_{\mathrm{r}}\| + \|V_{\mathrm{l}}\| \lesssim 1. \tag{4.15a}$$

(ii) *The density ρ is comparable with the explicit function $\widetilde{\rho}$ given by*

$$\begin{aligned}\rho(\tau_0+\omega+\mathrm{i}\eta) &\sim \widetilde{\rho}(\tau_0+\omega+\mathrm{i}\eta)\\ &:= \begin{cases}\rho(\tau_0)+(|\omega|+\eta)^{1/3}, & \text{if } \tau_0=\mathfrak{m},\mathfrak{c},\\ (|\omega|+\eta)^{1/2}(\Delta+|\omega|+\eta)^{-1/6}, & \text{if } \tau_0=\mathfrak{e}_-,\ \omega\in[-c,0]\\ \eta(\Delta+|\omega|+\eta)^{-1/6}(|\omega|+\eta)^{-1/2}, & \text{if } \tau_0=\mathfrak{e}_-,\ \omega\in[0,\Delta/2].\end{cases}\end{aligned} \tag{4.15b}$$

(iii) *The eigenvalue β of smallest modulus satisfies*

$$|\beta|\sim\frac{\eta}{\rho}+\rho(\rho+|\sigma|), \tag{4.15c}$$

and we have the comparison relations

$$\begin{aligned}&|\langle V_{\mathrm{l}}, M\mathcal{S}[V_{\mathrm{r}}]V_{\mathrm{r}}\rangle|\sim\rho+|\sigma|,\\ &\Big|\langle V_{\mathrm{l}}, M\mathcal{S}[V_{\mathrm{r}}]\mathcal{B}^{-1}\mathcal{Q}[M\mathcal{S}[V_{\mathrm{r}}]V_{\mathrm{r}}]+M\mathcal{S}\mathcal{B}^{-1}\mathcal{Q}[M\mathcal{S}[V_{\mathrm{r}}]V_{\mathrm{r}}]V_{\mathrm{r}}\rangle\Big|\sim 1.\end{aligned} \tag{4.15d}$$

(iv) *The quantities $\eta/\rho+\rho(\rho+|\sigma|)$ and $\rho+|\sigma|$ in* (4.15c)–(4.15d) *can be replaced by the following more explicit auxiliary quantities*

$$\begin{aligned}\widetilde{\xi}_1(\tau_0+\omega+\mathrm{i}\eta) &:= \begin{cases}(|\omega|+\eta)^{1/2}(|\omega|+\eta+\Delta)^{1/6},\\ (\rho(\tau_0)+(|\omega|+\eta)^{1/3})^2,\end{cases}\\ \widetilde{\xi}_2(\tau_0+\omega+\mathrm{i}\eta) &:= \begin{cases}(|\omega|+\eta+\Delta)^{1/3}, & \text{if } \tau_0=\mathfrak{e}_-,\\ \rho(\tau_0)+(|\omega|+\eta)^{1/3}, & \text{if } \tau_0=\mathfrak{m},\mathfrak{c}.\end{cases}\end{aligned} \tag{4.15e}$$

which are monotonically increasing in η. More precisely, it holds that $\eta/\rho+\rho(\rho+|\sigma|)\sim\widetilde{\xi}_1$ and, in the case where $\tau_0=\mathfrak{c},\mathfrak{m}$ is a cusp or a non-zero local minimum, we also have that $\rho+|\sigma|\sim\widetilde{\xi}_2$. For the case when $\tau_0=\mathfrak{e}_-$ is a right edge next to a gap of size Δ there exists a constant c_* such that $\rho+|\sigma|\sim\widetilde{\xi}_2$ in the regime $\omega\in[-c,c_*\Delta]$ and $\rho+|\sigma|\lesssim\widetilde{\xi}_2$ in the regime $\omega\in[c_*\Delta,\Delta/2]$.

Proof. We first explain how to translate the notations from the present paper to the notations in [12]: The operators $\mathcal{S}$, $\mathcal{B}$, $\mathcal{Q}$ are simply denoted by S, B, Q in [12]; the matrices V_l, V_r here are denoted by $l/\langle l, b\rangle, b$ there. The bound on $\mathcal{B}^{-1}\mathcal{Q}$ in (4.15a) follows directly from [12, Eq. (5.15)]. The bounds on $V_{\mathrm{l}}, V_{\mathrm{r}}$ in (4.15a) follow from the definition of the stability operator (4.11) together with the fact that $\|M\|\lesssim 1$ (by Assumption (4.C)) and $\|\mathcal{S}\|_{\mathrm{hs}\to\|\cdot\|}\lesssim 1$, following from the upper bound in flatness (4.14). The asymptotic expansion of ρ in (4.15b) follows from [12, Remark 7.3] and [7, Corollary A.1]. The claims in (iii) follow directly from [12, Proposition 6.1]. Finally, the claims in (iv) follow directly from [12, Remark 10.4]. □

The following lemma establishes simplified lower bounds on $\widetilde{\xi}_1, \widetilde{\xi}_2$ whenever η is much larger than the fluctuation scale η_{f}. We defer the proof of the technical lemma which differentiates various regimes to the appendix.

Lemma 4.3.3. *Under the assumptions of Proposition 4.3.2 we have uniformly in $z=\tau_0+\omega+\mathrm{i}\eta\in\mathbb{D}_{\mathrm{cusp}}$ with $\eta\geq\eta_{\mathrm{f}}$ that*

$$\widetilde{\xi}_2\gtrsim\frac{1}{N\eta}+\Big(\frac{\rho}{N\eta}\Big)^{1/2},\qquad \widetilde{\xi}_1\gtrsim\widetilde{\xi}_2\Big(\rho+\frac{1}{N\eta}\Big).$$

We now define an appropriate matrix norm in which we will measure the distance between G and M. The $\|\cdot\|_*$-norm is defined exactly as in [DS4] and similar to the one first introduced in [DS3]. It is a norm comparing matrix elements on a large but finite set of vectors with a hierarchical structure. To define this set we introduce some notations. For second order cumulants of matrix elements $\kappa(w_{ab}, w_{cd}) := \mathbf{E}\, w_{ab} w_{cd}$ we use the short-hand notation $\kappa(ab, cd)$. We also use the short-hand notation $\kappa(\mathbf{x}b, cd)$ for the $\mathbf{x} = (x_a)_{a\in[N]}$-weighted linear combination $\sum_a x_a \kappa(ab, cd)$ of such cumulants. We use the notation that replacing an index in a scalar quantity by a dot $(\cdot)$ refers to the corresponding vector, e.g. $A_{a\cdot}$ is a short-hand notation for the vector $(A_{ab})_{b\in[N]}$. Matrices $R_{\mathbf{xy}}$ with vector subscripts $\mathbf{x}, \mathbf{y}$ are understood as short-hand notations for $\langle \mathbf{x}, R\mathbf{y}\rangle$, and matrices $R_{\mathbf{x}a}$ with mixed vector and index subscripts are understood as $\langle \mathbf{x}, Re_a\rangle$ with e_a being the a-th normalized $\|e_a\| = 1$ standard basis vector. We fix two vectors $\mathbf{x}, \mathbf{y}$ and some large integer K and define the sets of vectors

$$\begin{aligned} I_0 &:= \{\mathbf{x}, \mathbf{y}\} \cup \{\, \delta_{a\cdot}, (V_1^*)_{a\cdot} \mid a \in [N] \,\}, \\ I_{k+1} &:= I_k \cup \{\, M\mathbf{u} \mid \mathbf{u} \in I_k \,\} \cup \{\, \kappa_{\mathrm{c}}((M\mathbf{u})a, b\cdot), \kappa_{\mathrm{d}}((M\mathbf{u})a, \cdot b) \mid \mathbf{u} \in I_k, a, b \in [N] \,\}. \end{aligned}$$

Here the cross and the direct part $\kappa_{\mathrm{c}}, \kappa_{\mathrm{d}}$ of the 2-cumulants $\kappa(\cdot,\cdot)$ refer to the natural splitting dictated by the Hermitian symmetry. In the specific case of (4.10) we simply have $\kappa_{\mathrm{c}}(ab, cd) = \delta_{ad}\delta_{bc} s_{ab}$ and $\kappa_{\mathrm{d}}(ab, cd) = \delta_{ac}\delta_{bd} t_{ab}$. Then the $\|\cdot\|_*$-norm is given by

$$\begin{aligned} \|R\|_* = \|R\|_*^{K,\mathbf{x},\mathbf{y}} &:= \sum_{0\le k<K} N^{-k/2K} \|R\|_{I_k} + N^{-1/2} \max_{\mathbf{u}\in I_K} \frac{\|R_{\cdot\mathbf{u}}\|}{\|\mathbf{u}\|}, \\ \|R\|_I &:= \max_{\mathbf{u},\mathbf{v}\in I} \frac{|R_{\mathbf{uv}}|}{\|\mathbf{u}\|\,\|\mathbf{v}\|}. \end{aligned}$$

We remark that the set I_k hence also $\|\cdot\|_*$ depend on z via $M = M(z)$. We omit this dependence from the notation as it plays no role in the estimates.

In terms of this norm we obtain the following estimate on $G - M$ in terms of its projection $\Theta = \langle V_{\mathrm{l}}, G - M\rangle$ onto the unstable direction of the stability operator $\mathcal{B}$. It is a direct consequence of a general expansion of approximate quadratic matrix equations whose linear stability operators have a single eigenvalue close to 0, as given in Lemma 4.A.1.

Proposition 4.3.4 (Cubic equation for Θ). *Fix $K \in \mathbb{N}$, $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$ and use $\|\cdot\|_* = \|\cdot\|_*^{K,\mathbf{x},\mathbf{y}}$. For fixed $z \in \mathbb{D}_{\mathrm{cusp}}$ and on the event that $\|G - M\|_* + \|D\|_* \lesssim N^{-10/K}$ the difference $G - M$ admits the expansion*

$$\begin{aligned} G - M &= \Theta V_{\mathrm{r}} - \mathcal{B}^{-1}\mathcal{Q}[MD] + \Theta^2 \mathcal{B}^{-1}\mathcal{Q}[M\mathcal{S}[V_{\mathrm{r}}]V_{\mathrm{r}}] + E, \\ \|E\|_* &\lesssim N^{5/K}(|\Theta|^3 + |\Theta|\,\|D\|_* + \|D\|_*^2), \end{aligned} \tag{4.16a}$$

with an error matrix E and the scalar $\Theta := \langle V_{\mathrm{l}}, G - M\rangle$ that satisfies the approximate cubic equation

$$\Theta^3 + \xi_2\Theta^2 + \xi_1\Theta = \epsilon_*. \tag{4.16b}$$

Here, the error ϵ_* satisfies the upper bound

$$\begin{aligned} |\epsilon_*| \lesssim N^{20/K}(\|D\|_*^3 + |\langle R, D\rangle|^{3/2}) + |\langle V_{\mathrm{l}}, MD\rangle| \\ + \left|\langle V_{\mathrm{l}}, M(\mathcal{S}\mathcal{B}^{-1}\mathcal{Q}[MD])(\mathcal{B}^{-1}\mathcal{Q}[MD])\rangle\right|, \end{aligned} \tag{4.16c}$$

where R is a deterministic matrix with $\|R\| \lesssim 1$ and the coefficients of the cubic equation satisfy the comparison relations

$$|\xi_1| \sim \frac{\eta}{\rho} + \rho(\rho + |\sigma|), \qquad |\xi_2| \sim \rho + |\sigma| \,. \tag{4.16d}$$

Proof. We first establish some important bounds involving the $\|\cdot\|_*$-norm. We claim that for any matrices R, R_1, R_2

$$\begin{aligned} &\|M\mathcal{S}[R_1]R_2\|_* \lesssim N^{1/2K} \|R_1\|_* \|R_2\|_* , \quad \|MR\|_* \lesssim N^{1/2K} \|R\|_* , \\ &\|\mathcal{Q}\|_{*\to*} \lesssim 1, \quad \left\|\mathcal{B}^{-1}\mathcal{Q}\right\|_{*\to*} \lesssim 1, \quad |\langle V_l, R\rangle| \lesssim \|R\|_* . \end{aligned} \tag{4.17}$$

The proof of (4.17) follows verbatim as in Lemma 3.3.4 with (4.15a) as an input. Moreover, the bound on $\langle V_l, \cdot\rangle$ follows directly from the bound on $\mathcal{Q}$. Obviously, we also have $\|\cdot\|_* \leq 2\,\|\cdot\|$.

Next, we apply Lemma 4.A.1 from the appendix with the choices

$$\mathcal{A}[R_1, R_2] := M\mathcal{S}[R_1]R_2, \qquad X := MD, \qquad Y := G - M.$$

The operator $\mathcal{B}$ in Lemma 4.A.1 is chosen as the stability operator (4.11). Then (4.119) is satisfied with $\lambda := N^{1/2K}$ according to (4.17) and (4.15a). With $\delta := N^{-25/4K}$ we verify (4.16a) directly from (4.123), where $\Theta = \langle V_l, G - M\rangle$ satisfies

$$\mu_3\Theta^3 + \mu_2\Theta^2 - \beta\Theta = -\mu_0 + \langle R, D\rangle\,\Theta + \mathcal{O}\left(N^{-1/4K}\,|\Theta|^3 + N^{20/K}\,\|D\|_*^3\right). \tag{4.18}$$

Here we used $|\Theta| \leq \|G - M\|_* \lesssim N^{-10/K}$ and $\|MD\|_* \lesssim N^{1/2K}\,\|D\|_*$. The coefficients μ_0, μ_2, μ_3 are defined through (4.122) and R is given by

$$R := M^*(\mathcal{B}^{-1}\mathcal{Q})^*[\mathcal{S}[M^*V_lV_r^*] + \mathcal{S}[V_r^*]M^*V_l].$$

Now we bound $|\langle R, D\rangle\,\Theta| \leq N^{-1/4K}\,|\Theta|^3 + N^{1/8K}\,|\langle R, D\rangle|^{3/2}$ by Young's inequality, absorb the error terms bounded by $N^{-1/4K}\,|\Theta|^3$ into the cubic term,

$$\mu_3\Theta^3 + \mathcal{O}(N^{-1/4K}\,|\Theta|^3) = \widetilde{\mu}_3\Theta^3,$$

by introducing a modified coefficient $\widetilde{\mu}_3$ and use that $|\mu_3| \sim |\widetilde{\mu}_3| \sim 1$ for any $z \in \mathbb{D}_{\mathrm{cusp}}$. Finally, we safely divide (4.18) by $\widetilde{\mu}_3$ to verify (4.16b) with $\xi_1 := -\beta/\widetilde{\mu}_3$ and $\xi_2 := \mu_2/\widetilde{\mu}_3$. For the fact $|\mu_3| \sim 1$ on $\mathbb{D}_{\mathrm{cusp}}$ and the comparison relations (4.16d) we refer to (4.15c)–(4.15d). □

4.3.2 Probabilistic bound

We now collect bounds on the error matrix D from Theorem 2.4.1 and Section 4.4. We first introduce the notion of *stochastic domination.*

Definition 4.3.5 (Stochastic domination). *Let $X = X^{(N)}, Y = Y^{(N)}$ be sequences of non-negative random variables. We say that X is stochastically dominated by Y (and use the notation $X \prec Y$) if*

$$\mathbf{P}\left[X > N^{\epsilon}Y\right] \leq C(\epsilon, \nu)N^{-\nu}, \qquad N \in \mathbb{N},$$

for any $\epsilon > 0, \nu \in \mathbb{N}$ and some family of positive constants $C(\epsilon, \nu)$ that is uniform in N and other underlying parameters (e.g. the spectral parameter z in the domain under consideration).

It can be checked (see [73, Lemma 4.4]) that $\prec$ satisfies the usual arithmetic properties, e.g. if $X_1 \prec Y_1$ and $X_2 \prec Y_2$, then also $X_1 + X_2 \prec Y_1 + Y_2$ and $X_1 X_2 \prec Y_1 Y_2$. Furthermore, to formulate bounds on a random matrix R compactly, we introduce the notations

$$\begin{aligned} |R| \prec \Lambda \quad &\iff \quad |R_{\mathbf{x}\mathbf{y}}| \prec \Lambda \,\|\mathbf{x}\|\,\|\mathbf{y}\| && \text{uniformly for all } \mathbf{x}, \mathbf{y} \in \mathbb{C}^N, \\ |R|_{\mathrm{av}} \prec \Lambda \quad &\iff \quad |\langle BR \rangle| \prec \Lambda \,\|B\| && \text{uniformly for all } B \in \mathbb{C}^{N\times N} \end{aligned}$$

for random matrices R and a deterministic control parameter $\Lambda = \Lambda(z)$. We also introduce high moment norms

$$\|X\|_p := \Big(\mathbf{E}\,|X|^p\Big)^{1/p}, \qquad \|R\|_p := \sup_{\mathbf{x},\mathbf{y}} \frac{\|\langle \mathbf{x}, R\mathbf{y}\rangle\|_p}{\|\mathbf{x}\|\,\|\mathbf{y}\|}$$

for $p \geq 1$, scalar valued random variables X and random matrices R. To translate high moment bounds into high probability bounds and vice versa we have the following easy lemma, see Lemma 3.3.7.

Lemma 4.3.6. *Let R be a random matrix, Φ a deterministic control parameter such that $\Phi \geq N^{-C}$ and $\|R\| \leq N^C$ for some $C > 0$, and let $K \in \mathbb{N}$ be a fixed integer. Then we have the equivalences*

$$\|R\|_*^{K,\mathbf{x},\mathbf{y}} \prec \Phi \textit{ uniformly in } \mathbf{x}, \mathbf{y} \iff |R| \prec \Phi \iff \|R\|_p \leq_{p,\epsilon} N^\epsilon \Phi,\ \forall \epsilon > 0,\ p \geq 1.$$

Expressed in terms of the $\|\cdot\|_p$-norm we have the following high-moment bounds on the error matrix D. The bounds (4.19a)–(4.19b) have already been established in Theorem 2.4.1; we just list them for completeness. The bounds (4.19c)–(4.19d), however, are new and they capture the additional cancellation at the cusp and are the core novelty of the present paper. The additional smallness comes from averaging against specific weights $\mathbf{p}, \mathbf{f}$ from (4.13a).

Theorem 4.3.7 (High moment bound on D with cusp fluctuation averaging)**.** *Under the assumptions of Theorem 4.2.5 for any compact set $\mathbb{D} \subset \{\, z \in \mathbb{C} \mid \Im z \geq N^{-1} \,\}$ there exists a constant C such that for any $p \geq 1, \epsilon > 0,\ z \in \mathbb{D}$ and matrices/vectors $B, \mathbf{x}, \mathbf{y}$ it holds that*

$$\|\langle \mathbf{x}, D\mathbf{y}\rangle\|_p \leq_{\epsilon,p} \|\mathbf{x}\|\,\|\mathbf{y}\|\, N^\epsilon \psi_q' \Big(1 + \|G\|_q\Big)^C \Big(1 + \frac{\|G\|_q}{\sqrt{N}}\Big)^{Cp}, \tag{4.19a}$$

$$\|\langle BD\rangle\|_p \leq_{\epsilon,p} \|B\|\, N^\epsilon \Big[\psi_q'\Big]^2 \Big(1 + \|G\|_q\Big)^C \Big(1 + \frac{\|G\|_q}{\sqrt{N}}\Big)^{Cp}. \tag{4.19b}$$

Moreover, for the specific weight matrix $B = \mathrm{diag}(\mathbf{pf})$ we have the improved bound

$$\|\langle \mathrm{diag}(\mathbf{pf}) D\rangle\|_p \leq_{\epsilon,p} N^\epsilon \sigma_q \Big[\psi + \psi_q'\Big]^2 \Big(1 + \|G\|_q\Big)^C \Big(1 + \frac{\|G\|_q}{\sqrt{N}}\Big)^{Cp}, \tag{4.19c}$$

and the improved bound on the off-diagonal component

$$\Big\|\langle \mathrm{diag}(\mathbf{pf})[T \odot G^t] G\rangle\Big\|_p \leq_{\epsilon,p} N^\epsilon \sigma_q \Big[\psi + \psi_q'\Big]^2 \Big(1 + \|G\|_q\Big)^C \Big(1 + \frac{\|G\|_q}{\sqrt{N}}\Big)^{Cp}, \tag{4.19d}$$

where we defined the following z-dependent quantities

$$\psi := \sqrt{\frac{\rho}{N\eta}}, \quad \psi'_q := \sqrt{\frac{\|\Im G\|_q}{N\eta}}, \quad \psi''_q := \|G-M\|_q,$$
$$\sigma_q := |\sigma| + \rho + \psi + \sqrt{\eta/\rho} + \psi'_q + \psi''_q$$

and $q = Cp^3/\epsilon$.

Theorem 4.3.7 will be proved in Section 4.4. We now translate the high moment bounds of Theorem 4.3.7 into high probability bounds via Lemma 4.3.6 and use those to establish bounds on $G - M$ and the error in the cubic equation for Θ. To simplify the expressions we formulate the bounds in the domain

$$\mathbb{D}_\zeta := \left\{ z \in \mathbb{D}_{\mathrm{cusp}} \,\middle|\, \Im z \geq N^{-1+\zeta} \right\}. \tag{4.20}$$

Lemma 4.3.8 (High probability error bounds)**.** *Fix $\zeta, c > 0$ sufficiently small and suppose that $|G-M| \prec \Lambda$, $|\Im(G-M)| \prec \Xi$ and $|\Theta| \prec \theta$ hold at fixed $z \in \mathbb{D}_\zeta$, and assume that the deterministic control parameters Λ, Ξ, θ satisfy $\Lambda + \Xi + \theta \lesssim N^{-c}$. Then for any sufficiently small $\epsilon > 0$ it holds that*

$$\left|\Theta^3 + \xi_2\Theta^2 + \xi_1\Theta\right| \prec N^{2\epsilon}\left(\rho + |\sigma| + \frac{\eta^{1/2}}{\rho^{1/2}} + \left(\frac{\rho+\Xi}{N\eta}\right)^{1/2}\right)\frac{\rho+\Xi}{N\eta} + N^{-\epsilon}\theta^3, \tag{4.21a}$$

as well as

$$|G-M| \prec \theta + \sqrt{\frac{\rho+\Xi}{N\eta}}, \qquad |G-M|_{\mathrm{av}} \prec \theta + \frac{\rho+\Xi}{N\eta}, \tag{4.21b}$$

where the coefficients ξ_1, ξ_2 are those from Proposition 4.3.4, and we recall that $\Theta = \langle V_l, G-M\rangle$.

Proof. We translate the high moment bounds (4.19a)–(4.19b) into high probability bounds using Lemma 4.3.6 and $|G| \prec \|M\| + \Lambda \lesssim 1$ to find

$$|D| \prec \sqrt{\frac{\rho+\Xi}{N\eta}}, \qquad |D|_{\mathrm{av}} \prec \frac{\rho+\Xi}{N\eta}. \tag{4.22}$$

In particular, these bounds together with the assumed bounds on $G - M$ guarantee the applicability of Proposition 4.3.4. Now we use (4.22) in (4.16a) to get (4.21b). Here we used (4.17), translated $\|\cdot\|_p$-bounds into $\prec$-bounds on $\|\cdot\|_*$ and vice versa via Lemma 4.3.6, and absorbed the $N^{1/K}$ factors into $\prec$ by using that K can be chosen arbitrarily large. It remains to verify (4.21a). In order to do so, we first claim that

$$\begin{aligned}&|\langle V_l, MD\rangle| + \left|\langle V_l, M(\mathcal{S}\mathcal{B}^{-1}\mathcal{Q}[MD])(\mathcal{B}^{-1}\mathcal{Q}[MD])\rangle\right| \\ &\prec N^\epsilon\left(|\sigma| + \rho + \frac{\eta^{1/2}}{\rho^{1/2}} + \Lambda + \left(\frac{\rho+\Xi}{N\eta}\right)^{1/2}\right)\frac{\rho+\Xi}{N\eta} + \theta^2\left(N^{-\epsilon}\Lambda + \left(\frac{\rho+\Xi}{N\eta}\right)^{1/2}\right)\end{aligned} \tag{4.23}$$

for any sufficiently small $\epsilon > 0$.

Proof of (4.23). We first collect two additional ingredients from [12] specific to the vector case.

(a) The imaginary part $\Im \mathbf{m}$ of the solution $\mathbf{m}$ is comparable $\Im \mathbf{m} \sim \langle \Im \mathbf{m} \rangle = \pi\rho$ to its average, and, in particular, $\mathbf{m} = \Re \mathbf{m} + \mathcal{O}(\rho)$.

(b) The eigendirections $V_{\mathrm{l}}, V_{\mathrm{r}}$ are diagonal and are approximately given by

$$V_{\mathrm{l}} = c \operatorname{diag}(\mathbf{f}/\lvert \mathbf{m} \rvert) + \mathcal{O}(\rho + \eta/\rho), \quad V_{\mathrm{r}} = c' \operatorname{diag}(\mathbf{f}\lvert \mathbf{m} \rvert) + \mathcal{O}(\rho + \eta/\rho) \tag{4.24}$$

for some constants $c, c' \sim 1$.

Indeed, (a) follows directly from [12, Proposition 3.5] and the approximations in (4.24) follow directly from [12, Corollary 5.2]. The fact that $V_{\mathrm{l}}, V_{\mathrm{r}}$ are diagonal follows from simplicity of the eigendirections in the matrix case, and the fact that $M = \operatorname{diag}(\mathbf{m})$ is diagonal and that $\mathcal{B}$ preserves the space of diagonal matrices as well as the space of off-diagonal matrices. On the latter $\mathcal{B}$ acts stably as $1 + \mathcal{O}_{\mathrm{hs}\to\mathrm{hs}}(N^{-1})$. Thus the unstable directions lie inside the space of diagonal matrices.

We now turn to the proof of (4.23) and first note that, according to (a) and (b) we have

$$M = \operatorname{diag}(\mathbf{p}\lvert \mathbf{m} \rvert) + \mathcal{O}(\rho), \qquad V_{\mathrm{l}} = c \operatorname{diag}(\mathbf{f}/\lvert \mathbf{m} \rvert) + \mathcal{O}(\rho + \eta/\rho) \tag{4.25}$$

for some constant $c \sim 1$ to see

$$\langle V_{\mathrm{l}}, MD \rangle = c \langle \operatorname{diag}(\mathbf{pf})D \rangle + \mathcal{O}(\rho + \eta/\rho) \langle \operatorname{diag}(\mathbf{w}_1)D \rangle,$$

where $\mathbf{w}_1 \in \mathbb{C}^N$ is a deterministic vector with uniformly bounded entries. Since

$$\lvert \langle \operatorname{diag}(\mathbf{w}_1)D \rangle \rvert \prec \frac{\rho + \Xi}{N\eta}$$

by (4.22), the bound on the first term in (4.23) follows together with (4.19c) via Lemma 4.3.6. Now we consider the second term in (4.23). We split $D = D_{\mathrm{d}} + D_{\mathrm{o}}$ into its diagonal and off-diagonal components. Since $\mathcal{B}$ and $\mathcal{S}$ preserve the space of diagonal and the space of off-diagonal matrices we find

$$\begin{aligned} &\langle V_{\mathrm{l}}, M(\mathcal{S}\mathcal{B}^{-1}\mathcal{Q}[MD])(\mathcal{B}^{-1}\mathcal{Q}[MD]) \rangle \\ &\quad = \frac{1}{N^2} \sum_{i,j} u_{ij} d_{ii} d_{jj} + \langle V_{\mathrm{l}}, M(\mathcal{S}\mathcal{B}^{-1}\mathcal{Q}[MD_{\mathrm{o}}])(\mathcal{B}^{-1}\mathcal{Q}[MD_{\mathrm{o}}]) \rangle, \end{aligned} \tag{4.26}$$

with an appropriate deterministic matrix u_{ij} having bounded entries. In particular, the cross terms vanish and the first term is bounded by

$$\Big\lvert \frac{1}{N^2} \sum_{i,j} u_{ij} d_{ii} d_{jj} \Big\rvert \le \max_i \lvert d_{ii} \rvert \Big\lvert \frac{1}{N} \sum_j u_{ij} d_{jj} \Big\rvert \prec \Big(\frac{\rho + \Xi}{N\eta} \Big)^{3/2} \tag{4.27}$$

according to (4.22). By taking the off-diagonal part of (4.16a) and using the fact that M and V_{r} and therefore also $\mathcal{B}^{-1}\mathcal{Q}[M\mathcal{S}[V_{\mathrm{r}}]V_{\mathrm{r}}]$ are diagonal (cf. (b) above) we have

$$\Big\lvert \mathcal{B}^{-1}\mathcal{Q}[MD_{\mathrm{o}}] + G_{\mathrm{o}} \Big\rvert \prec \theta^3 + \theta \Big(\frac{\rho + \Xi}{N\eta} \Big)^{1/2} + \frac{\rho + \Xi}{N\eta} \lesssim N^{-\epsilon}\theta^2 + N^{\epsilon} \frac{\rho + \Xi}{N\eta}$$

for any ϵ such that $\theta \lesssim N^{-\epsilon}$ by Young's inequality in the last step. Together with (4.25), (4.22) and the assumption that $|G_{\mathrm{o}}| = |(G-M)_{\mathrm{o}}| \prec \Lambda$ we then compute

$$\begin{aligned}
&\langle V_{\mathrm{l}}, M(\mathcal{S}\mathcal{B}^{-1}\mathcal{Q}[MD_{\mathrm{o}}])(\mathcal{B}^{-1}\mathcal{Q}[MD_{\mathrm{o}}])\rangle \\
&= c\,\langle \operatorname{diag}(\mathbf{pf})(\mathcal{S}\mathcal{B}^{-1}\mathcal{Q}[MD_{\mathrm{o}}])(\mathcal{B}^{-1}\mathcal{Q}[MD_{\mathrm{o}}])\rangle + \mathcal{O}\Big(\Big(\rho + \frac{\eta}{\rho}\Big)\frac{\rho+\Xi}{N\eta}\Big) \\
&= c\,\langle \operatorname{diag}(\mathbf{pf})\mathcal{S}[G_{\mathrm{o}}]G_{\mathrm{o}}\rangle \\
&\qquad + \mathcal{O}\Big(\Big(\rho + \frac{\eta}{\rho}\Big)\frac{\rho+\Xi}{N\eta} + \Big(\Big(\frac{\rho+\Xi}{N\eta}\Big)^{1/2} + \Lambda\Big)\Big[N^{-\epsilon}\theta^2 + N^{\epsilon}\frac{\rho+\Xi}{N\eta}\Big]\Big).
\end{aligned}$$

Thus the bound on the second term on the lhs. in (4.23) follows together with (4.26)–(4.27) by $\mathcal{S}[G_{\mathrm{o}}] = T \odot G^t$ and (4.19d) via Lemma 4.3.6. This completes the proof of (4.23). □

With (4.22) and (4.23) the upper bound (4.16c) on the error ϵ_* of the cubic equation (4.16b) takes the same form as the rhs. of (4.23) if K is sufficiently large depending on ϵ. By the first estimate in (4.21b) we can redefine the control parameter Λ on $|G-M|$ as $\Lambda := \theta + ((\rho+\Xi)/N\eta)^{1/2}$ and the claim (4.21a) follows directly with (4.23), thus completing the proof of Lemma 4.3.8. □

4.3.3 Bootstrapping

Now we will show that the difference $G-M$ converges to zero uniformly for all spectral parameters $z \in \mathbb{D}_\zeta$ as defined in (4.20). For convenience we refer to existing bounds on $G-M$ far away from the real line to establish a rough bound on $G-M$ in, say, $\mathbb{D}_1$. We then iteratively lower the threshold on η by appealing to Proposition 4.3.4 and Lemma 4.3.8 until we establish the rough bound in all of $\mathbb{D}_\zeta$. As a second step we then improve the rough bound iteratively until we obtain Theorem 4.2.5.

Lemma 4.3.9 (Rough bound)**.** *For any $\zeta > 0$ there exists a constant $c > 0$ such that on the domain $\mathbb{D}_\zeta$ we have the rough bound*

$$|G-M| \prec N^{-c}. \tag{4.28}$$

Proof. The rough bound (4.28) in a neighbourhood of a cusp has first been established for Wigner-type random matrices in [9]. For the convenience of the reader we present a streamlined proof that is adapted to the current setting. The lemma is an immediate consequence of the following statement. Let $\zeta_{\mathrm{s}} > 0$ be a sufficiently small *step size*, depending on ζ. Then for any $\mathbb{N}_0 \ni k \le 1/\zeta_{\mathrm{s}}$ on the domain $\mathbb{D}_{\max\{1-k\zeta_{\mathrm{s}},\zeta\}}$ we have

$$|G-M| \prec N^{-4^{-k}\zeta}. \tag{4.29}$$

We prove (4.29) by induction over k. For sufficiently small ζ the induction start $k=0$ holds due to the local law away from the self-consistent spectrum, e.g. Theorem 2.2.1.

Now as induction hypothesis suppose that (4.29) holds on $\widetilde{\mathbb{D}}_k := \mathbb{D}_{\max\{1-k\zeta_{\mathrm{s}},\zeta\}}$, and in particular, $|G| \prec 1$, $\|G\|_p \le_{\epsilon,p} N^{\epsilon}$ for any ϵ, p according to Lemma 4.3.6. The monotonicity of the function $\eta \mapsto \eta\,\|G(\tau+\mathrm{i}\eta)\|_p$ (see e.g. the proof of Proposition 2.5.5) implies $\|G\|_p \le_{\epsilon,p} N^{\epsilon+\zeta_{\mathrm{s}}} \le N^{2\zeta_{\mathrm{s}}}$ and therefore, according to Lemma 4.3.6, that $|G| \prec N^{2\zeta_{\mathrm{s}}}$ on $\widetilde{\mathbb{D}}_{k+1}$. This, in turn, implies $|D| \prec N^{-\zeta/3}$ on $\widetilde{\mathbb{D}}_{k+1}$ by (4.19a) and Lemma 4.3.6, provided

ζ_{s} is chosen small enough. We now fix $\mathbf{x}, \mathbf{y}$ and a large integer K as the parameters of $\|\cdot\|_* = \|\cdot\|_*^{\mathbf{x},\mathbf{y},K}$ for the rest of the proof and omit them from the notation but we stress that all estimates will be uniform in $\mathbf{x}, \mathbf{y}$. We find $\sup_{z \in \mathbb{D}_{k+1,c}} \|D(z)\|_* \prec N^{-\zeta/3}$, by using a simple union bound and $\|\partial_z D\| \leq N^C$ for some $C > 0$. Thus, for K large enough, we can use (4.16a), (4.16b), (4.16c) and (4.17) to infer

$$\begin{aligned} \left|\Theta^3 + \xi_2\Theta^2 + \xi_1\Theta\right| &\lesssim N^{1/2K} \|D\|_* \prec N^{1/2K-\zeta/3}, \\ \|G - M\|_* &\lesssim |\Theta| + N^{1/K} \|D\|_* \prec |\Theta| + N^{1/K-\zeta/3}, \end{aligned} \tag{4.30}$$

on the event $\|G - M\|_* + \|D\|_* \lesssim N^{-10/K}$, and on $\widetilde{\mathbb{D}}_{k+1}$. Now we use the following lemma [12, Lemma 10.3] to translate the first estimate in (4.30) into a bound on $|\Theta|$. For the rest of the proof we keep $\tau = \Re z$ fixed and consider the coefficients ξ_1, ξ_2 and Θ as functions of η.

Lemma 4.3.10 (Bootstrapping cubic inequality). *For $0 < \eta_* < \eta^* < \infty$ let $\xi_1, \xi_2\colon [\eta_*, \eta^*] \to \mathbb{C}$ be complex valued functions and $\widetilde{\xi}_1, \widetilde{\xi}_2, d\colon [\eta_*, \eta^*] \to \mathbb{R}^+$ be continuous functions such that at least one of the following holds true:*

(i) $|\xi_1| \sim \widetilde{\xi}_1$, $|\xi_2| \sim \widetilde{\xi}_2$, and $\widetilde{\xi}_2^3/d, \widetilde{\xi}_1^3/d^2, \widetilde{\xi}_1^2/d\widetilde{\xi}_2$ are monotonically increasing, and $d^2/\widetilde{\xi}_1^3 + d\widetilde{\xi}_2/\widetilde{\xi}_1^2 \ll 1$ at η^*,

(ii) $|\xi_1| \sim \widetilde{\xi}_1$, $|\xi_2| \lesssim \widetilde{\xi}_1^{1/2}$, and $\widetilde{\xi}_1^3/d^2$ is monotonically increasing.

Then any continuous function $\Theta\colon [\eta_*, \eta^*] \to \mathbb{C}$ that satisfies the cubic inequality $|\Theta^3 + \xi_2\Theta^2 + \xi_1\Theta| \lesssim d$ on $[\eta_*, \eta^*]$, has the property

$$\textit{If } |\Theta| \lesssim \min\left\{d^{1/3}, \frac{d^{1/2}}{\widetilde{\xi}_2^{1/2}}, \frac{d}{\widetilde{\xi}_1}\right\} \textit{ at } \eta^*, \;\; \textit{then } |\Theta| \lesssim \min\left\{d^{1/3}, \frac{d^{1/2}}{\widetilde{\xi}_2^{1/2}}, \frac{d}{\widetilde{\xi}_1}\right\} \textit{ on } [\eta_*, \eta^*]. \tag{4.31}$$

With direct arithmetics we can now verify that the coefficients ξ_1, ξ_2 in (4.16b) and the auxiliary coefficients $\widetilde{\xi}_1, \widetilde{\xi}_2$ defined in (4.15e) satisfy the assumptions in Lemma 4.3.10 with the choice of the constant function $d = N^{-4^{-k}\zeta+\delta}$ for any $\delta > 0$, by using only the information on ξ_1, ξ_2 given by the comparison relations (4.16d). As an example, in the regime where τ_0 is a right edge and $\omega \sim \Delta$, we have $\widetilde{\xi}_1 \sim (\eta + \Delta)^{2/3}$ and $\widetilde{\xi}_2 \sim (\eta + \Delta)^{1/3}$ and both functions are monotonically increasing in η. Then Assumption (ii) of Lemma 4.3.10 is satisfied. All other regimes are handled similarly.

We now set $\eta^* := N^{-k\zeta_{\mathrm{s}}}$ and

$$\eta_* := \inf\left\{ \eta \in [N^{-(k+1)\zeta_{\mathrm{s}}}, \eta^*] \;\middle|\; \sup_{\eta' \geq \eta} \|G(\tau + \mathrm{i}\eta') - M(\tau + \mathrm{i}\eta')\|_* \leq N^{-10/K}/2 \right\}.$$

By the induction hypothesis we have $|\Theta(\eta^*)| \lesssim d \lesssim \min\{d^{1/3}, d^{1/2}\widetilde{\xi}_2^{-1/2}, d\widetilde{\xi}_1^{-1}\}$ with overwhelming probability, so that the condition in (4.31) holds, and conclude $|\Theta(\eta)| \prec d^{1/3} = N^{-(4^{-k}\zeta-\delta)/3}$ for $\eta \in [\eta_*, \eta^*]$. For small enough $\delta > 0$ the second bound in (4.30) implies $\|G - M\|_* \prec N^{-4^{k+1}\zeta}$. By continuity and the definition of η_* we conclude $\eta_* = N^{-(k+1)\zeta_{\mathrm{s}}}$, finishing the proof of (4.29). □

Proof of Theorem 4.2.5. The bounds within the proof hold true uniformly for $z \in \mathbb{D}_\zeta$, unless explicitly specified otherwise. We therefore suppress this qualifier in the following statements. First we apply Lemma 4.3.8 with the choice $\Xi = \Lambda$, i.e. we do not treat the imaginary part of the resolvent separately. With this choice the first inequality in (4.21b) becomes self-improving and after iteration shows that

$$|G - M| \prec \theta + \sqrt{\frac{\rho}{N\eta}} + \frac{1}{N\eta}, \tag{4.32}$$

and, in other words, (4.21a) holds with $\Xi = \theta + (\rho/N\eta)^{1/2} + 1/N\eta$. This implies that if $|\Theta| \prec \theta \lesssim N^{-c}$ for some arbitrarily small $c > 0$, then

$$\left|\Theta^3 + \xi_2\Theta^2 + \xi_1\Theta\right| \lesssim N^{5\widetilde{\epsilon}} d_* + N^{-\widetilde{\epsilon}}(\theta^3 + \widetilde{\xi}_2\theta^2) \tag{4.33}$$

holds for all sufficiently small $\widetilde{\epsilon}$ with overwhelming probability, where we defined

$$d_* := \widetilde{\xi}_2\Big(\frac{\widetilde{\rho}}{N\eta} + \frac{1}{(N\eta)^2}\Big) + \frac{1}{(N\eta)^3} + \Big(\frac{\widetilde{\rho}}{N\eta}\Big)^{3/2}. \tag{4.34}$$

For this conclusion we used the comparison relations (4.16d), Proposition 4.3.2(iv) as well as (4.15b), and the bound $\sqrt{\eta/\rho} \sim \sqrt{\eta/\widetilde{\rho}} \lesssim \widetilde{\xi}_2$.

The bound (4.33) is a self-improving estimate on $|\Theta|$ in the following sense. For $k \in \mathbb{N}$ and $l \in \mathbb{N} \cup \{*\}$ let

$$d_k := \max\{N^{-k\widetilde{\epsilon}}, N^{6\widetilde{\epsilon}} d_*\}, \qquad \theta_l := \min\Big\{d_l^{1/3}, \frac{d_l^{1/2}}{\widetilde{\xi}_2^{1/2}}, \frac{d_l}{\widetilde{\xi}_1}\Big\}.$$

Then (4.33) with $|\Theta| \prec \theta_k$ implies that

$$\left|\Theta^3 + \xi_2\Theta^2 + \xi_1\Theta\right| \lesssim N^{-\widetilde{\epsilon}} d_k.$$

Applying Lemma 4.3.10 with $d = N^{-\widetilde{\epsilon}} d_k$, $\eta^* \sim 1$, $\eta_* = N^{\zeta - 1}$ yields the improvement $|\Theta| \prec \theta_{k+1}$. Here we needed to check the condition in (4.31) but at $\eta^* \sim 1$ we have $\widetilde{\xi}_1 \sim 1$, so $|\Theta| \lesssim N^{-\widetilde{\epsilon}} d_k \le d_{k+1} \sim \theta_{k+1}$. After a k-step iteration until $N^{-k\widetilde{\epsilon}}$ becomes smaller than $N^{6\widetilde{\epsilon}} d_*$, we find $|\Theta| \prec \theta_*$, where we used that $\widetilde{\epsilon}$ can be chosen arbitrarily small. We are now ready to prove the following bound which we, for convenience, record as a proposition.

Proposition 4.3.11. *For any $\zeta > 0$ we have the bounds*

$$|G - M| \prec \theta_* + \sqrt{\frac{\rho}{N\eta}} + \frac{1}{N\eta}, \qquad |G - M|_{\mathrm{av}} \prec \theta_* + \frac{\rho}{N\eta} + \frac{1}{(N\eta)^2} \quad \text{in} \quad \mathbb{D}_\zeta, \tag{4.35}$$

where $\theta_* := \min\{d_*^{1/3}, d_*^{1/2}/\widetilde{\xi}_2^{1/2}, d_*/\widetilde{\xi}_1\}$, and $d_*, \widetilde{\rho}, \widetilde{\xi}_1, \widetilde{\xi}_2$ are given in (4.34), (4.15b) *and* (4.15e), *respectively.*

Proof. Using $|\Theta| \prec \theta_*$ proven above, we apply (4.32) with $\theta = \theta_*$ to conclude the first inequality in (4.35). For the second inequality in (4.35) we use the estimate on $|G - M|_{\mathrm{av}}$ from (4.21b) with $\theta = \theta_*$ and $\Xi = (\rho/N\eta)^{1/2} + 1/N\eta$. □

The bound on $|G - M|$ from (4.35) implies a complete delocalisation of eigenvectors uniformly at singularities of the scDOS. The following corollary was established already in [9, Corollary 1.14] and, given (4.35), the proof follows the same line of reasoning.

Corollary 4.3.12 (Eigenvector delocalisation). *Let* $\mathbf{u} \in \mathbb{C}^N$ *be an eigenvector of* H *corresponding to an eigenvalue* $\lambda \in \tau_0 + (-c, c)$ *for some sufficiently small positive constant* $c \sim 1$. *Then for any deterministic* $\mathbf{x} \in \mathbb{C}^N$ *we have*

$$|\langle \mathbf{u}, \mathbf{x} \rangle| \prec \frac{1}{\sqrt{N}} \, \|\mathbf{u}\| \, \|\mathbf{x}\| \, .$$

The bounds (4.35) simplify in the regime $\eta \geq N^{\zeta} \eta_{\mathrm{f}}$ above the typical eigenvalue spacing to

$$|G - M| \prec \sqrt{\frac{\rho}{N\eta}} + \frac{1}{N\eta}, \qquad |G - M|_{\mathrm{av}} \prec \frac{1}{N\eta}, \qquad \text{for} \quad \eta \geq N^{\zeta} \eta_{\mathrm{f}} \tag{4.36}$$

using Lemma 4.3.3 which implies $\theta_* \leq d_* / \widetilde{\xi}_1 \leq 1/N\eta$. The bound on $|G - M|_{\mathrm{av}}$ is further improved in the case when $\tau_0 = \mathfrak{e}_-$ is an edge and, in addition to $\eta \geq N^{\zeta} \eta_{\mathrm{f}}$, we assume $N^{\delta} \eta \leq \omega \leq \Delta/2$ for some $\delta > 0$, i.e. if ω is well inside a gap of size $\Delta \geq N^{\delta + \zeta} \eta_{\mathrm{f}}$. Then we find $\Delta > N^{-3/4}$ by the definition of $\eta_{\mathrm{f}} = \Delta^{1/9} / N^{2/3}$ in (4.7) and use Lemma 4.3.3 and (4.15b), (4.15e) to conclude

$$\begin{aligned} \theta_* + \frac{\widetilde{\rho}}{N\eta} + \frac{1}{(N\eta)^2} &\lesssim \frac{\widetilde{\xi}_2}{\widetilde{\xi}_1} \Big(\frac{\widetilde{\rho}}{N\eta} + \frac{1}{(N\eta)^2} \Big) \\ &\sim \frac{\Delta^{1/6}}{\omega^{1/2}} \Big(\frac{\eta}{\Delta^{1/6} \omega^{1/2}} + \frac{1}{N\eta} \Big) \frac{1}{N\eta} \lesssim \frac{N^{-\delta/2}}{N\eta}. \end{aligned} \tag{4.37}$$

In the last bound we used $1/N\omega \leq N^{-\delta}/N\eta$ and $\Delta^{1/6}/(N\eta\omega^{1/2}) \leq N^{-\delta/2}$. Using (4.37) in (4.35) yields the improvement

$$|G - M|_{\mathrm{av}} \prec \frac{N^{-\delta/2}}{N\eta}, \qquad \text{for} \quad \tau = \mathfrak{e}_- + \omega, \quad \Delta/2 \geq \omega \geq N^{\delta} \eta \geq N^{\zeta + \delta} \eta_{\mathrm{f}}. \tag{4.38}$$

The bounds on $|G - M|_{\mathrm{av}}$ from (4.36) and (4.38), inside and outside the self-consistent spectrum, allow us to show the uniform rigidity, Corollary 4.2.6. We postpone these arguments until after we finish the proof of Theorem 4.2.5. The uniform rigidity implies that for $\mathrm{dist}(z, \operatorname{supp} \rho) \geq N^{\zeta} \eta_{\mathrm{f}}$ we can estimate the imaginary part of the resolvent via

$$\Im \langle \mathbf{x}, G\mathbf{x} \rangle = \sum_{\lambda} \frac{\eta \, |\langle \mathbf{u}_\lambda, \mathbf{x} \rangle|^2}{\eta^2 + (\tau_0 + \omega - \lambda)^2} \prec \eta + \frac{1}{N} \sum_{|\lambda - \tau_0| \leq c} \frac{\eta}{\eta^2 + (\tau_0 + \omega - \lambda)^2} \prec \rho(z), \tag{4.39}$$

for any normalised $\mathbf{x} \in \mathbb{C}^N$, where $\mathbf{u}_\lambda$ denotes the normalised eigenvector corresponding to λ. For the first inequality in (4.39) we used Corollary 4.3.12 and for the second we applied Corollary 4.2.6 that allows us to replace the Riemann sum with an integral as $[\eta^2 + (\tau_0 + \omega - \lambda)^2]^{1/2} = |z - \lambda| \geq N^{\zeta} \eta_{\mathrm{f}}$.

Using with (4.39), we apply Lemma 4.3.8, repeating the strategy from the beginning of the proof. But this time we can choose the control parameter $\Xi = \rho$. In this way we find

$$|G - M| \prec \theta_{\#} + \sqrt{\frac{\rho}{N\eta}}, \quad |G - M|_{\mathrm{av}} \prec \theta_{\#} + \frac{\rho}{N\eta}, \quad \text{for } \mathrm{dist}(z, \operatorname{supp} \rho) \geq N^{\zeta} \eta_{\mathrm{f}}, \tag{4.40}$$

where we defined

$$\theta_\# := \min\left\{ \frac{d_\#}{\widetilde{\xi}_1}, \frac{d_\#^{1/2}}{\widetilde{\xi}_2^{1/2}}, d_\#^{1/3} \right\}, \qquad d_\# := \widetilde{\xi}_2 \frac{\widetilde{\rho}}{N\eta} + \left(\frac{\widetilde{\rho}}{N\eta}\right)^{3/2}.$$

Note that the estimates in (4.40) are simpler than those in (4.35). The reason is that the additional terms $1/N\eta, 1/(N\eta)^2$ and $1/(N\eta)^3$ in (4.35) are a consequence of the presence of Ξ in (4.21a), (4.21b). With $\Xi = \rho$ these are immediately absorbed into ρ and not present any more. The second term in the definition of $d_\#$ can be dropped since we still have $\widetilde{\xi}_2 \gtrsim (\rho/N\eta)^{1/2}$ (this follows from Lemma 4.3.3 if $\eta \geq N^\zeta \eta_{\mathrm{f}}$, and directly from (4.15b), (4.15e) if $\omega \geq N^\zeta \eta_{\mathrm{f}}$). This implies $\theta_\# \lesssim d_\#^{1/2}/\widetilde{\xi}_2^{1/2} \lesssim (\rho/N\eta)^{1/2}$, so the first bound in (4.40) proves (4.8a).

Now we turn to the proof of (4.8b). Given the second bound in (4.36), it is sufficient to consider the case when $\tau = \mathfrak{e}_- + \omega$ and $\eta \leq \omega \leq \Delta/2$ with $\omega \geq N^\zeta \eta_{\mathrm{f}}$. In this case Proposition 4.3.2 yields $\widetilde{\xi}_2\widetilde{\rho}/\widetilde{\xi}_1 + \widetilde{\rho} \lesssim \eta/\omega \sim \eta/\operatorname{dist}(z, \operatorname{supp}\rho)$. Thus we have

$$\theta_\# + \frac{\rho}{N\eta} \lesssim \frac{d_\#}{\widetilde{\xi}_1} + \frac{\widetilde{\rho}}{N\eta} \lesssim \frac{1}{N \operatorname{dist}(z, \operatorname{supp}\rho)}$$

and therefore the second bound in (4.40) implies (4.8b). This completes the proof of Theorem 4.2.5. □

4.3.4 Rigidity and absence of eigenvalues

The proofs of Corollaries 4.2.6 and 4.2.8 rely on the bounds on $|G - M|_{\mathrm{av}}$ from (4.36) and (4.38). As before, we may restrict ourselves to the neighbourhood of a local minimum $\tau_0 \in \operatorname{supp}\rho$ of the scDOS which is either an internal minimum with a small value of $\rho(\tau_0) > 0$, a cusp location or a right edge adjacent to a small gap of length $\Delta > 0$. All other cases, namely the bulk regime and regular edges adjacent to large gaps, have been treated prior to this work [9, DS4].

Proof of Corollary 4.2.8. Let us denote the empirical eigenvalue distribution of H by $\rho_H = \frac{1}{N}\sum_{i=1}^N \delta_{\lambda_i}$ and consider the case when $\tau_0 = \mathfrak{e}_-$ is a right edge, $\Delta \geq N^\delta \eta_{\mathrm{f}}$ for any $\delta > 0$ and $\eta_{\mathrm{f}} = \eta_{\mathrm{f}}(\mathfrak{e}_-) \sim \Delta^{1/9} N^{-2/3}$. Then we show that there are no eigenvalues in $\mathfrak{e}_- + [N^\delta \eta_{\mathrm{f}}, \Delta/2]$ with overwhelming probability. We apply [9, Lemma 5.1] with the choices

$$\nu_1 := \rho, \quad \nu_2 := \rho_H, \quad \eta_1 := \eta_2 := \epsilon := N^\zeta \eta_{\mathrm{f}}, \quad \tau_1 := \mathfrak{e}_- + \omega, \quad \tau_2 := \mathfrak{e}_- + \omega + N^\zeta \eta_{\mathrm{f}},$$

for any $\omega \in [N^\delta \eta_{\mathrm{f}}, \Delta/2]$ and some $\zeta \in (0, \delta/4)$. We use (4.38) to estimate the error terms J_1, J_2 and J_3 from [9, Eq. (5.2)] by $N^{2\zeta - \delta/2 - 1}$ and see that $(\rho_H - \rho)([\tau_1, \tau_2]) = \rho_H([\tau_1, \tau_2]) \prec N^{2\zeta - \delta/2 - 1}$, showing that with overwhelming probability the interval $[\tau_1, \tau_2]$ does not contain any eigenvalues. A simple union bound finishes the proof of Corollary 4.2.8. □

Proof of Corollary 4.2.6. Now we establish Corollary 4.2.6 around a local minimum $\tau_0 \in \operatorname{supp}\rho$ of the scDOS. Its proof has two ingredients. First we follow the strategy of the proof of [9, Corollary 1.10] to see that

$$|(\rho - \rho_H)((-\infty, \tau_0 + \omega])| \prec \frac{1}{N}, \tag{4.41}$$

for any $|\omega| \leq c$, i.e. we have a very precise control on ρ_H. In contrast to the statement in that corollary we have a local law (4.36) with uniform $1/N\eta$ error and thus the bound (4.41) does not deteriorate close to τ_0. We warn the reader that the standard argument inside the proof of [9, Corollary 1.10] has to be adjusted slightly to arrive at (4.41). In fact, when inside that proof the auxiliary result [9, Lemma 5.1] is used with the choice $\tau_1 = -10$, $\tau_2 = \tau$, $\eta_1 = \eta_2 = N^{\zeta-1}$ for some $\zeta > 0$, this choice should be changed to $\tau_1 = -C$, $\tau_2 = \tau$, $\eta_1 = N^{\zeta-1}$ and $\eta_2 = N^{\zeta}\eta_{\mathrm{f}}(\tau)$, where $C > 0$ is chosen sufficiently large such that τ_1 lies far to the left of the self-consistent spectrum.

The control (4.41) suffices to prove Corollary 4.2.6 for all $\tau = \tau_0 + \omega$ except for the case when $\tau_0 = \mathfrak{e}_-$ is an edge at a gap of length $\Delta \geq N^{\zeta}\eta_{\mathrm{f}}$ and $\omega \in [-N^{\zeta}\eta_{\mathrm{f}}, 0]$ for some fixed $\zeta > 0$ and $\eta_{\mathrm{f}} = \eta_{\mathrm{f}}(\mathfrak{e}_-) \sim \Delta^{1/9}/N^{2/3}$, i.e. except for some N^{ζ} eigenvalues close to the edge with arbitrarily small $\zeta > 0$. In all other cases, the proof follows the same argument as the proof of [9, Corollary 1.11] using the uniform $1/N$-bound from (4.41) and we omit the details here.

The reason for having to treat the eigenvalues very close to the edge $\mathfrak{e}_-$ separately is that (4.41) does not give information on which side of the gap these N^{ζ} eigenvalues are found. To get this information requires the second ingredient, the *band rigidity*,

$$\mathbf{P}\big[\rho((-\infty, \mathfrak{e}_- + \omega]) = \rho_H((-\infty, \mathfrak{e}_- + \omega])\big] \geq 1 - N^{-\nu}, \qquad (4.42)$$

for any $\nu \in \mathbb{N}$, $\Delta \geq \omega \geq N^{\zeta}\eta_{\mathrm{f}}$ and large enough N. The combination of (4.42) and (4.41) finishes the proof of Corollary 4.2.6.

Band rigidity has been shown in case Δ is bounded from below in Corollary 3.2.9. We will now adapt this proof to the case of small gap sizes $\Delta \geq N^{\zeta-3/4}$. Since by Corollary 4.2.8 with overwhelming probability there are no eigenvalues in $\mathfrak{e}_- + [N^{\zeta}\eta_{\mathrm{f}}, \Delta/2]$, it suffices to show (4.42) for $\omega = \Delta/2$. As in the proof of Corollary 3.2.9 we consider the interpolation

$$H_t := \sqrt{1-t}W + A - t\mathcal{S}M(\tau), \qquad t \in [0,1],$$

between the original random matrix $H = H_0$ and the deterministic matrix $H_1 = A - \mathcal{S}M(\tau)$, for $\tau = \mathfrak{e}_- + \Delta/2$. The interpolation is designed such that the solution M_t of the MDE corresponding to H_t is constant at spectral parameter τ, i.e. $M_t(\tau) = M(\tau)$. Let ρ_t denote the scDOS of H_t. Exactly as in the proof of Corollary 3.2.9 it suffices to show that no eigenvalue crosses the gap along the interpolation with overwhelming probability, i.e. that for any $\nu \in \mathbb{N}$ we have

$$\mathbf{P}\big[\mathfrak{a}_t \in \operatorname{Spec}(H_t) \text{ for some } t \in [0,1]\big] \leq \frac{C(\nu)}{N^{\nu}}. \qquad (4.43)$$

Here $t \to \mathfrak{a}_t \in \mathbb{R} \setminus \operatorname{supp} \rho_t$ is some spectral parameter inside the gap, continuously depending on t, such that $\mathfrak{a}_0 = \tau$. In the proof of Corollary 3.2.9, $\mathfrak{a}_t$ was chosen independent of t, but the argument remains valid with any other choice of $\mathfrak{a}_t$. We call I_t the connected component of $\mathbb{R} \setminus \operatorname{supp} \rho_t$ that contains $\mathfrak{a}_t$ and denote $\Delta_t = |I_t|$ the gap length. In particular, $\Delta_0 = \Delta$ and $\tau \in I_t$ for all $t \in [0,1]$ by [12, Lemma 8.1(ii)]. For concreteness we choose $\mathfrak{a}_t$ to be the spectral parameter lying exactly in the middle of I_t. The $1/3$-Hölder continuity of ρ_t, hence I_t and $\mathfrak{a}_t$ in t follows from [12, Proposition 10.1(a)]. Via a simple union bound it suffices to show that for any fixed $t \in [0,1]$ we have no eigenvalue in $\mathfrak{a}_t + [-N^{-100}, N^{-100}]$.

Since $\|W\| \lesssim 1$ with overwhelming probability, in the regime $t \geq 1 - \epsilon$ for some small constant $\epsilon > 0$, the matrix H_t is a small perturbation of the deterministic matrix H_1

whose resolvent $(H_1 - \tau)^{-1} = M(\tau)$ at spectral parameter τ is bounded by Assumption (4.C), in particular $\Delta_1 \gtrsim 1$. By $1/3$-Hölder continuity hence $\Delta_t \gtrsim 1$, and $\operatorname{Spec}(H_t) \subset \operatorname{Spec}(H_1) + [-C\epsilon^{1/3}, C\epsilon^{1/3}]$ for some $C \sim 1$ in this regime with very high probability. Since $\operatorname{Spec}(H_1) \subset \operatorname{supp} \rho_t + [-C\epsilon^{1/3}, C\epsilon^{1/3}]$ by [12, Proposition 10.1(a)] there are no eigenvalues of H_t in a neighbourhood of $\mathfrak{a}_t$, proving (4.43) for $t \geq 1 - \epsilon$.

For $t \in [\epsilon, 1-\epsilon]$ we will now show that $\Delta_t \sim_\epsilon 1$ for any $\epsilon > 0$. In fact, we have $\operatorname{dist}(\tau, \operatorname{supp} \rho_t) \gtrsim_\epsilon 1$. This is a consequence of [12, Lemma D.1]. More precisely, we use the equivalence of (iii) and (v) of that lemma. We check (iii) and conclude the uniform distance to the self-consistent spectrum by (v). Since $M_t(\tau) = M(\tau)$ and $\|M(\tau)\| \lesssim 1$ we only need to check that the stability operator $\mathcal{B}_t = t + (1-t)\mathcal{B}$ of H_t has a bounded inverse. We write $\mathcal{B}_t = \mathcal{C}(1 - (1-t)\widetilde{\mathcal{C}}\mathcal{F})\mathcal{C}^{-1}$ in terms of the saturated self-energy operator $\mathcal{F} = \mathcal{CSC}$, where $\mathcal{C}[R] := |M(\tau)|^{1/2} R \, |M(\tau)|^{1/2}$ and $\widetilde{\mathcal{C}}[R] := (\operatorname{sgn} M(\tau))R(\operatorname{sgn} M(\tau))$. Afterwards we use that $\|\mathcal{F}\|_{\mathrm{hs}\to\mathrm{hs}} \leq 1$ (cf. [8, Eq. (4.24)]) and $\|\widetilde{\mathcal{C}}\|_{\mathrm{hs}\to\mathrm{hs}} = 1$ to first show the uniform bound $\|\mathcal{B}_t\|_{\mathrm{hs}\to\mathrm{hs}} \lesssim 1/t$ and then improve the bound to $\|\mathcal{B}_t\| \lesssim 1/t$ using the trick of expanding in a geometric series from [8, Eqs. (4.60)–(4.63)]. This completes the argument that $\Delta_t \sim_\epsilon 1$. Now we apply Corollary 2.2.3 to see that there are no eigenvalues of H_t around $\mathfrak{a}_t$ as long as t is bounded away from zero and one, proving (4.43) for this regime.

Finally, we are left with the regime $t \in [0, \epsilon]$ for some sufficiently small $\epsilon > 0$. By [12, Proposition 10.1(a)] the self-consistent Green's function M_t corresponding to H_t is bounded even in a neighbourhood of τ, whose size only depends on model parameters. In particular, Assumptions (4.A)–(4.C) are satisfied for H_t and Corollary 4.2.8, which was already proved above, is applicable. Thus it suffices to show that the size Δ_t of the gap in $\operatorname{supp} \rho_t$ containing τ is bounded from below by $\Delta_t \geq N^{\zeta - 3/4}$ for some $\zeta > 0$. The size of the gap can be read off from the following relationship between the norm of the saturated self-energy operator and the size of the gap: Let H be a random matrix satisfying (4.A)–(4.C) and τ be well inside the interior of the gap of length $\Delta \in [0, c]$ in the self-consistent spectrum for a sufficiently small $c \sim 1$. Then

$$1 - \|\mathcal{F}(\tau)\|_{\mathrm{hs}\to\mathrm{hs}} \sim \lim_{\eta \searrow 0} \frac{\eta}{\rho(\tau + \mathrm{i}\eta)} \sim (\Delta + \operatorname{dist}(\tau, \operatorname{supp} \rho))^{1/6} \operatorname{dist}(\tau, \operatorname{supp} \rho)^{1/2} \sim \Delta^{2/3}, \tag{4.44}$$

where in the first step we used [8, Eqs. (4.23)–(4.25)], in the second step (4.15b), and in the last step that $\operatorname{dist}(\tau, \operatorname{supp} \rho) \sim \Delta$. Applying the analogue of (4.44) for H_t with $\mathcal{F}_t(\tau)$ and using that $\operatorname{dist}(\tau, \rho_t) \lesssim \Delta_t$, we obtain $1 - \|\mathcal{F}_t(\tau)\|_{\mathrm{hs}\to\mathrm{hs}} \lesssim \Delta_t^{2/3}$. Combining this inequality with (4.44) and using that $\mathcal{F}_t(\tau) = (1-t)F(\tau)$ for $t \in [0, c]$, we have $\Delta_t^{3/2} \gtrsim t + (1-t)\Delta^{2/3}$, i.e. $\Delta_t \gtrsim t^{3/2} + \Delta$. In particular, the gap size Δ_t never drops below $c\Delta \gtrsim N^{\zeta - 3/4}$. This completes the proof of the last regime in (4.43). □

4.4 Cusp fluctuation averaging and proof of Theorem 4.3.7

First we review some of the basic nomenclature from [DS3]. We consider random matrices $H = A + W$ with diagonal expectation A and complex Hermitian or real symmetric zero mean random component W indexed by some abstract set J of size $|J| = N$. We recall that Greek letters $\alpha, \beta, \ldots$ stand for labels, i.e. double-indices from $I = J \times J$, whereas Roman letters $a, b, \ldots$ stand for single indices. If $\alpha = (a, b)$, then we set $\alpha^t := (b, a)$ for its transpose. Underlined Greek letters stand for multisets of labels, whereas bold-faced

Greek letters stand for tuples of labels with the counting combinatorics being their – for our purposes – only relevant difference.

According to Proposition 2.3.2 with $\mathcal{N}(\alpha) = \{\alpha, \alpha^t\}$ it follows from the assumed independence that for general (conjugate) linear functionals $\Lambda^{(k)}$, of bounded norm $\|\Lambda^{(k)}\| = \mathcal{O}(1)$

$$\mathbf{E} \prod_{k\in[p]} \Lambda^{(k)}(D) = \mathbf{E} \prod_{l\in[p]} \Big(1 + \sum_{\alpha_l,\boldsymbol{\beta}_l}^{\sim(l)}\Big) \prod_{k\in[p]} \begin{cases} \Lambda^{(k)}_{\alpha_k,\underline{\beta}^k} & \text{if} \sum_{\alpha_k} \\ \Lambda^{(k)}_{\underline{\beta}^k_{<k},\underline{\beta}^k_{>k}} & \text{else} \end{cases} + \mathcal{O}\left(N^{-p}\right), \quad (4.45a)$$

where we recall that

$$\sum_{\alpha_l,\boldsymbol{\beta}_l}^{\sim(l)} := \sum_{\alpha_l\in I} \sum_{1\leq m<6p} \sum_{\boldsymbol{\beta}_l\in\{\alpha_l,\alpha_l^t\}^m} \frac{\kappa(\alpha_l,\underline{\beta}_l)}{m!} \sum_{\underline{\beta}_l^1\sqcup\dots\sqcup\underline{\beta}_l^p=\underline{\beta}_l} \mathbb{1}(|\underline{\beta}_l^l| = 0 \text{ if } |\underline{\beta}_l| = 1) \quad (4.45b)$$

and that

$$\begin{aligned} \Lambda_{\alpha_1,\dots,\alpha_k} &:= -(-1)^k \Lambda(\Delta^{\alpha_1} G \dots \Delta^{\alpha_k} G), \\ \Lambda_{\{\alpha_1,\dots,\alpha_m\}} &:= \sum_{\sigma\in S_m} \Lambda_{\alpha_{\sigma(1)},\dots,\alpha_{\sigma(m)}}, \quad \Lambda_{\alpha,\{\alpha_1,\dots,\alpha_m\}} := \sum_{\sigma\in S_m} \Lambda_{\alpha,\alpha_{\sigma(1)},\dots,\alpha_{\sigma(m)}}, \\ \Lambda_{\underline{\alpha},\underline{\beta}} &:= \sum_{\alpha\in\underline{\alpha}} \Lambda_{\alpha,\underline{\alpha}\cup\underline{\beta}\setminus\{\alpha\}}, \quad \underline{\beta}^k_{<k} := \bigsqcup_{j<k} \underline{\beta}^k_j, \quad \underline{\beta}^k_{>k} := \bigsqcup_{j>k} \underline{\beta}^k_j. \end{aligned} \quad (4.45c)$$

Some notations in (4.45) require further explanation. First, $\Delta^{(a,b)}$ denotes the matrix of all zeros except for an 1 in the (a, b)-th entry. The qualifier "if $\sum_{\alpha_k}$" is satisfied for those terms in which α_k is a summation variable when the brackets in the product $\prod_j(1 + \sum)$ are opened. The notation $\bigsqcup$ indicates the union of multisets.

For even p we apply (4.45) with $\Lambda^{(k)}(D) := \langle \mathrm{diag}(\mathbf{fp})D\rangle$ for $k \leq p/2$ and $\Lambda^{(k)}(D) := \overline{\langle \mathrm{diag}(\mathbf{fp})D\rangle}$ for $k > p/2$. This is obviously a special case of $\Lambda^{(k)}(D) = \langle BD\rangle$ which was considered in the so-called averaged case of [DS3] with arbitrary B of bounded operator norm since $\|\mathrm{diag}(\mathbf{fp})\| = \|\mathbf{fp}\|_\infty \leq C$. It was proved in [DS3] that

$$|\langle \mathrm{diag}(\mathbf{fp})D\rangle| \lesssim \frac{\rho}{N\eta},$$

which is not good enough at the cusp. We can nevertheless use the graphical language developed in [DS3] to estimate the complicated right hand side of (4.45).

4.4.1 Graphical representation via double index graphs

The graphs (or Feynman diagrams) introduced in [DS3] encode the structure of all terms in (4.45). Their (directed) edges correspond to resolvents G, while vertices correspond to Δ's. Loop edges are allowed while parallel edges are not. Resolvents G and their Hermitian conjugates G^* are distinguished by different types of edges. Each vertex v carries a label α_v and we need to sum up for all labels. Some labels are independently summed up, these are the α-labels in (4.45), while the β-labels are strongly restricted; in the independent case they can only be of the type α or α^t. These graphs will be called "double indexed" graphs since the vertices are naturally equipped with labels (double indices). Here we introduced the terminology "double indexed" for the graphs in [DS3] to distinguish them from the "single indexed" graphs to be introduced later in this paper.

To be more precise, the graphs in [DS3] were vertex-coloured graphs. The colours encoded a resummation of the terms in (4.45): vertices whose labels (or their transpose) appeared in one of the cumulants in (4.45) received the same colour. We then first summed up the colours and only afterwards we summed up all labels compatible with the given colouring. According to (2.49) for every even p it holds that

$$\mathbf{E}\,|\langle \operatorname{diag}(\mathbf{fp})D\rangle|^p = \sum_{\Gamma\in\mathcal{G}^{\mathrm{av}(p,6p)}} \operatorname{Val}(\Gamma) + \mathcal{O}\left(N^{-p}\right), \tag{4.46a}$$

where $\mathcal{G}^{\mathrm{av}(p,6p)}$ is a certain finite collection of vertex coloured directed graphs with p connected components, and $\operatorname{Val}(\Gamma)$, the value of the graph Γ, will be recalled below. According to [DS3] each graph $\Gamma\in\mathcal{G}^{\mathrm{av}(p,6p)}$ fulfils the following properties:

Proposition 4.4.1 (Properties of double index graphs). *There exists a finite set $\mathcal{G}^{av(p,6p)}$ of double index graphs Γ such that* (4.46) *hold. Each Γ fulfils the following properties.*

(a) There exist exactly p connected components, all of which are oriented cycles. Each vertex has one incoming and one outgoing edge.

(b) Each connected component contains at least one vertex and one edge. Single vertices with a looped edge are in particular legal connected components.

(c) Each colour colours at least two and at most $6p$ *vertices.*

(d) If a colour colours exactly two vertices, then these vertices are in different connected components.

(e) The edges represent the resolvent matrix G or its adjoint G^*. Within each component either all edges represent G or all edges represent G^*. Accordingly we call the components either G or G^*-cycles.

(f) Within each cycle there is one designated edge which is represented as a wiggled line in the graph. The designated edge represents the matrix G $\operatorname{diag}(\mathbf{pf})$ *in a G-cycle and the matrix* $\operatorname{diag}(\mathbf{pf})G^*$ *in a G^*-cycle.*

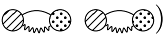
(g) For each colour there exists at least one component in which a vertex of that colour is connected to the matrix $\operatorname{diag}(\mathbf{fp})$. *According to (f) this means that if the relevant vertex is in a G-cycle, then the designated (wiggled) edge is its incoming edge. If the relevant vertex is in a G-cycle, then the designated edge is its outgoing edge.*

If V is the vertex set of Γ and for each colour $c\in C$, V_c denotes the c-coloured vertices then we recall that

$$\begin{aligned}\operatorname{Val}(\Gamma) = (-1)^{|V|}\,\mathbf{E}\Big(\prod_{c\in C}\prod_{v\in V_c}\sum_{\alpha_v}\frac{\kappa(\{\alpha_v\}_{v\in V_c})}{(|V_c|-1)!}\Big)&\\ \times\prod_{\operatorname{Cyc}(v_1,\dots,v_k)\in\Gamma}\begin{cases}\langle G\operatorname{diag}(\mathbf{fp})\Delta^{\alpha_{v_1}}G\dots G\Delta^{\alpha_{v_k}}\rangle\\ \langle\Delta^{\alpha_{v_k}}G^*\dots G^*\Delta^{\alpha_{v_1}}\operatorname{diag}(\mathbf{fp})G^*\rangle\end{cases}&\end{aligned}\tag{4.46b}$$

where the ultimate product is the product over all p of the cycles in the graph. By the notation $\operatorname{Cyc}(v_1,\dots,v_k)$ we indicate a directed cycle with vertices $v_1,\dots,v_k$. Depending upon whether a given cycle is a G-cycle or G^*-cycle, it then contributes with one of the factors indicated after the last curly bracket in (4.46b) with the vertex order chosen in such a way that the designated edge represents the $G\,\mathrm{diag}(\mathbf{fp})$ or $\mathrm{diag}(\mathbf{fp})G^*$ matrix. As an example illustrating (4.46b) we have

$$N^{-2} \sum_{\substack{\alpha_1,\beta_1\\ \alpha_2,\beta_2}} \kappa(\alpha_1,\beta_1)\kappa(\alpha_2,\beta_2) \left\langle G\,\mathrm{diag}(\mathbf{fp})\Delta^{\alpha_1} G \Delta^{\beta_2}\right\rangle \left\langle \Delta^{\beta_1} G^* \Delta^{\alpha_2}\,\mathrm{diag}(\mathbf{fp}) G^*\right\rangle \tag{4.47}$$

$= \mathrm{Val}\,($  $)$.

Actually in [DS3] the graphical representation of the graph Γ is simplified, it does not contain all information encoded in the graph. First, the direction of the edges are not indicated. In the picture both cycles should be oriented in a clockwise orientation. Secondly, the type of edges are not indicated, apart from the wiggled line. In fact, the edges in the second graph stand for G^*, while those in the first graph stand for G. To translate the pictorial representation directly let the striped vertices in the first and second cycle be associated with α_1, β_1 and the dotted vertices with α_2, β_2. Accordingly, the wiggled edge in the first cycle stands for $G\,\mathrm{diag}(\mathbf{fp})$, while the wiggled edge in the second cycle stands for $\mathrm{diag}(\mathbf{fp})G^*$. The reason why these details were omitted in the graphical representation of a double index graph is that they do not influence the basic power counting estimate of its value used in [DS3].

4.4.2 Single index graphs

In [DS3] we operated with double index graphs that are structurally simple and appropriate for bookkeeping complicated correlation structures, but they are not suitable for detecting the additional smallness we need at the cusp. The contribution of the graphs in [DS3] were estimated by a relatively simple power counting argument where only the number of (typically off-diagonal) resolvent elements were recorded. In fact, for many subleading graphs this procedure already gave a very good bound that is sufficient at the cusps as well. The graphs carrying the leading contribution, however, have now to be computed to a higher accuracy and this leads to the concept of "single index graphs". These are obtained by a certain refinement and reorganization of the double index graphs via a procedure we will call *graph resolution* to be defined later. The main idea is to restructure the double index graph in such a way that instead of labels (double indices) $\alpha = (a, b)$ its vertices naturally represent single indices a and b. Every double indexed graph will give rise to a finite number of resolved single index graphs. The double index graphs that require a more precise analysis compared with [DS3] will be resolved to single index graphs. After we explain the structure of the single index graphs and the graph resolution procedure, double index graphs will not be used in this paper any more. Thus, unless explicitly stated otherwise, by graph we will mean single index graph in the rest of this paper.

We now define the set $\mathcal{G}$ of single index graphs we will use in this paper. They are directed graphs, where parallel edges and loops are allowed. Let the graph be denoted by Γ with vertex set $V(\Gamma)$ and edge set $E(\Gamma)$. We will assign a value to each Γ which comprises weights assigned to the vertices and specific values assigned to the edges. Since an edge may represent different objects, we will introduce different types of edges that will be graphically distinguished by different line style. We now describe these ingredients precisely.

Vertices.

Each vertex $v \in V(\Gamma)$ is equipped with an associated index $a_v \in J$. Graphically the vertices are represented by small unlabelled bullets $\bullet$, i.e. in the graphical representation the actual index is not indicated. It is understood that all indices will be independently summed up over the entire index set J when we compute the value of the graph.

Vertex weights.

Each vertex $v \in V(\Gamma)$ carries some weight vector $\mathbf{w}^{(v)} \in \mathbb{C}^J$ which is evaluated $\mathbf{w}^{(v)}_{a_v}$ at the index a_v associated with the vertex. We generally assume these weights to be uniformly bounded in N, i.e. $\sup_N \|\mathbf{w}^{(v)}\|_\infty < \infty$. Visually we indicate vertex weights by incoming arrows as in $\mathbf{w} \rightarrow\!\bullet$. Vertices without explicitly indicated weight may carry an arbitrary bounded weight vector. We also use the notation $\mathbf{1} \rightarrow\!\bullet$ to indicate the constant $\mathbf{1}$ vector as the weight, this corresponds to summing up the corresponding index unweighted

G-edges.

The set of G-edges is denoted by $\mathrm{GE}(\Gamma) \subset E(\Gamma)$. These edges describe resolvents and there are four types of G-edges. First of all, there are directed edges corresponding to G and G^* in the sense that a directed G or G^*-edge $e = (v, u) \in E$ initiating from the vertex $v = i(e)$ and terminating in the vertex $u = t(e)$ represents the matrix elements $G_{a_v a_u}$ or respectively $G^*_{a_v a_u}$ evaluated in the indices a_v, a_u associated with the vertices v and u. Besides these two there are also edges representing $G - M$ and $(G - M)^*$. Distinguishing between G and $G - M$, for practical purposes, is only important if it occurs in a loop. Indeed, $(G - M)_{aa}$ is typically much smaller than G_{aa}, while $(G - M)_{ab}$ basically acts just like G_{ab} when a, b are summed independently. Graphically we will denote the four types of G-edges by

$$G = \bullet\!\longrightarrow\!\bullet, \quad G^* = \bullet\!\dashrightarrow\!\bullet, \quad G - M = \bullet\!\diamond\!\!\longrightarrow\!\bullet, \quad G^* - M^* = \bullet\!\diamond\!\!\dashrightarrow\!\bullet$$

where all these edges can also be loops. The convention is that continuous lines represent G, dashed lines correspond to G^*, while the diamond on both types of edges indicates the subtraction of M or M^*. An edge $e \in \mathrm{GE}(\Gamma)$ carries its type as its attribute, so as a short hand notation we can simply write G_e for $G_{a_{i(e)}, a_{t(e)}}$, $G^*_{a_{i(e)}, a_{t(e)}}$, $(G - M)_{a_{i(e)}, a_{t(e)}}$ and $(G - M)^*_{a_{i(e)}, a_{t(e)}}$ depending on which type of G-edge e represents. Due to their special role in the later estimates, we will separately bookkeep those $G - M$ or $G^* - M^*$ edges that appear looped. We thus define the subset $\mathrm{GE}_{g-m} \subset \mathrm{GE}$ as the set of G-edges $e \in \mathrm{GE}(\Gamma)$ of type $G - M$ or $G^* - M^*$ such that $i(e) = t(e)$. We write $g - m$ to refer to the fact that looped edges are evaluated on the diagonal $(g - m)_{a_v}$ of $(G - M)_{a_v a_v}$.

(G-)edge degree.

For any vertex v we define its in-degree $\deg^-(v)$ and out-degree $\deg^+(v)$ as the number of incoming and outgoing G-edges. Looped edges (v, v) are counted for both in- and out-degree. We denote the total degree by $\deg(v) = \deg^-(v) + \deg^+(v)$.

Interaction edges.

Besides the G-edges we also have interaction edges, $\mathrm{IE}(\Gamma)$, representing the cumulants κ. A directed interaction edge $e = (u, v)$ represents the matrix $R^{(e)} = (r^{(e)}_{ab})_{a,b\in J}$ given by the cumulant

$$\begin{aligned} r^{(u,v)}_{ab} &= \frac{1}{(\deg(u)-1)!}\kappa(\underbrace{ab,\dots,ab}_{\deg^-(u)\text{ times}}, \underbrace{ba,\dots,ba}_{\deg^+(u)\text{ times}}) \\ &= \frac{1}{(\deg(v)-1)!}\kappa(\underbrace{ab,\dots,ab}_{\deg^+(v)\text{ times}}, \underbrace{ba,\dots,ba}_{\deg^-(v)\text{ times}}). \end{aligned} \tag{4.48}$$

This relation is indeed compatible with exchanging the roles of u and v since $\deg^-(u) = \deg^+(v)$ and vice versa. For the important case when $\deg(u) = \deg(v) = 2$ it follows that the interaction from u to v is given by S if u has one incoming and one outgoing G-edge and T if u has two incoming G-edges, i.e.

$$s_{ab} = \kappa(ab, ba) \qquad t_{ab} = \kappa(ab, ab).$$

Visually we will represent interaction edges as

$$R = \bullet \xrightarrow{R} \bullet \qquad \text{and more specifically by} \qquad S = \bullet \xrightarrow{S} \bullet, \quad T = \bullet \xrightarrow{T} \bullet.$$

Although the interaction matrix $R^{(e)}$ is completely determined by the in- and out-degrees of the adjacent vertices $i(e), t(e)$ we still write out the specific S and T names because these will play a special role in the latter part of the proof. As a short hand notation we shall frequently use $R_e := R^{(e)}_{a_{i(e)},a_{t(e)}}$ to denote the matrix element selected by the indices $a_{i(e)}, a_{t(e)}$ associated with the initial and terminal vertex of e. We also note that we do not indicate the direction of edges associated with S as the matrix S is symmetric.

Generic weighted edges.

Besides the specific G-edges and interaction edges, additionally we also allow for generic edges reminiscent of the generic vertex weights introduced above. They will be called *generic weighted edges*, or *weighted edges* for short. To every weighted edge e we assign a weight matrix $K^{(e)} = (k^{(e)}_{ab})_{a,b\in J}$ which is evaluated as $k^{(e)}_{a_{i(e)},a_{t(e)}}$ when we compute the value of the graph by summing up all indices. To simplify the presentation we will not indicate the precise form of the weight matrix $K^{(e)}$ but only its entry-wise scaling as a function of N. A weighted edge presented as $\bullet \overset{N^{-l}}{\cdots\cdots} \bullet$ represents an arbitrary weight matrix $K^{(e)}$ whose entries scale like $|k^{(e)}_{ab}| \le cN^{-l}$. We denote the set of weighted edges by $\mathrm{WE}(\Gamma)$. For a given weighted edge $e \in \mathrm{WE}$ we record the entry-wise scaling of $K^{(e)}$ in an exponent $l(e) \ge 0$ in such a way that we always have $|k^{(e)}_{ab}| \le cN^{-l(e)}$.

Graph value.

For graphs $\Gamma \in \mathcal{G}$ we define their value

$$\begin{aligned}\mathrm{Val}(\Gamma) := (-1)^{|\mathrm{GE}(\Gamma)|}\Big(\prod_{v\in V(\Gamma)}\sum_{a_v\in J}\mathbf{w}^{(v)}_{a_v}\Big)\Big(\prod_{e\in \mathrm{IE}(\Gamma)} r^{(e)}_{a_{i(e)},a_{t(e)}}\Big)\\ \times\Big(\prod_{e\in\mathrm{WE}(\Gamma)} k^{(e)}_{a_{i(e)},a_{t(e)}}\Big)\mathbf{E}\Big(\prod_{e\in\mathrm{GE}(\Gamma)} G_e\Big),\end{aligned} \tag{4.49}$$

which differs slightly from that in (4.46b) because it applies to a different class of graphs.

4.4.3 Single index resolution

There is a natural mapping from double indexed graphs to a collection of single indexed graphs that encodes the rearranging of the terms in (4.46b) when the summation over labels α_v is reorganized into summation over single indices. Now we describe this procedure.

Definition 4.4.2 (Single index resolution). *By the* single index resolution *of a double vertex graph we mean the collection of single index graphs obtained through the following procedure.*

(i) For each colour, the identically coloured vertices of the double index graph are mapped into a pair of vertices of the single index graph.

(ii) The pair of vertices in the single index graph stemming from a fixed colour is connected by an interaction edge in the single index graph.

(iii) Every (directed) edge of the double index graph is naturally mapped to a G-edge of the single index graph. While mapping equally coloured vertices $x_1, \dots, x_k$ in the double index graph to vertices u, v connected by an interaction edge $e = (u, v)$ there are $k-1$ binary choices of whether we map the incoming edge of x_j to an incoming edge of u and the outgoing edge of x_j to an outgoing edge of v or vice versa. In this process we are free to consider the mapping of x_1 (or any other vertex, for that matter) as fixed by symmetry of $u \leftrightarrow v$.

(iv) If a wiggled G-edge is mapped to an edge from u to v, then v is equipped with a weight of $\mathbf{pf}$. If a wiggled G^*-edge is mapped to an edge from u to v, then u is equipped with a weight of $\mathbf{pf}$. All vertices with no weight specified in this process are equipped with the constant weight $\mathbf{1}$.

We define the set $\mathcal{G}(p) \subset \mathcal{G}$ as the set of all graphs obtained from the double index graphs $\mathcal{G}^{av(p,6p)}$ via the single index resolution procedure.

Remark 4.4.3.

(i) We note some ingredients described in Section 4.4.2 for a typical graph in $\mathcal{G}$ will be absent for graphs $\Gamma \in \mathcal{G}(p) \subset \mathcal{G}$. For example, $\mathrm{WE}(\Gamma) = \mathrm{GE}_{g-m}(\Gamma) = \emptyset$ for all $\Gamma \in \mathcal{G}(p)$.

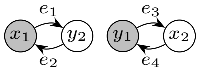
(ii) We also remark that loops in double index graphs are never mapped into loops in single index graphs along the single index resolution. Indeed, double index loops are always mapped to edges parallel to the interaction edge of the corresponding vertex.

A few simple facts immediately follow from the the single index construction in Definition 4.4.2. From (i) it is clear that the number of vertices in the single index graph is twice the number of colours of the double index graph. From (ii) it follows that the number of interaction edges in the single index graph equals the number of colours of the double index graph. Finally, from (iii) it is obvious that if for some colour c there are $k = k(c)$ vertices in the double index graph with colour c, then the resolution of this colour gives rise to $2^{k(c)-1}$ single indexed graph. Since these resolutions are done independently for each colour, we obtain that the number of single index graphs originating from one double index graph is

$$\prod_c 2^{k(c)-1}$$

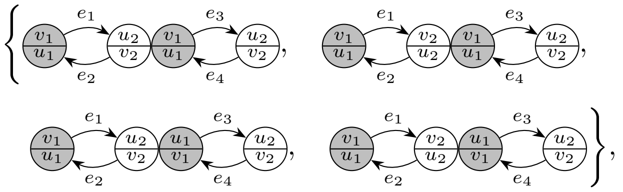
Since the number of double index graph in $\mathcal{G}^{\mathrm{av}(p,6p)}$ is finite, so is the number of graphs in $\mathcal{G}(p)$.

Let us present an example of single index resolution applied to the graph from (4.47) where we, for the sake of transparency, label all vertices and edges. Γ is a graph consisting of one 2-cycle on the vertices x_1, y_2 and one 2-cycle on the vertices x_2, y_1 as in

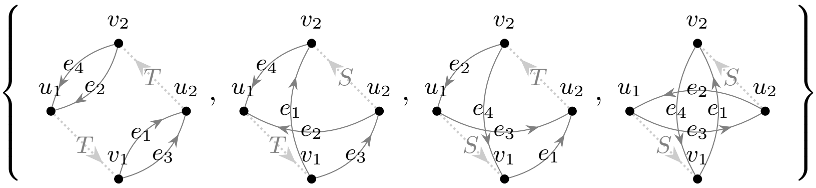


(4.50)

with x_1, y_1 and x_2, y_2 being of equal colour (i.e. being associated to labels connected through cumulants). In order to explain steps (i)-(iii) of the construction we first neglect that some edges may be wiggled, but we restore the orientation of the edges in the picture. We then fix the mapping of x_i to pairs of vertices (u_i, v_i) for $i = 1, 2$ in such a way that the incoming edges of x_i are incoming at u_i and the outgoing edges from x_i are outgoing from v_i. It remains to map y_i to (u_i, v_i) and for each i there are two choices of doing so that we obtain the four possibilities

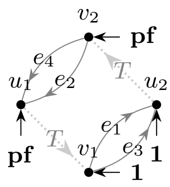


which translates to



(4.51)

in the language of single index graphs where the S, T assignment agrees with (4.48). Finally we want to visualize step (iv) in the single index resolution in our example. Suppose that in (4.50) the edges e_1 and e_2 are G-edges while e_3 and e_4 are G^* edges with e_2 and e_4 being wiggled (in agreement with (4.47)). According to (iv) it follows that the terminal vertex of e_2 and the initial vertex of e_4 are equipped with a weight of $\mathbf{pf}$ while the remaining vertices are equipped with a weight of $\mathbf{1}$. The first graph in (4.51) would thus be equipped with the weights



.

Single index graph expansion.

With the value definition in (4.49) it follows from Definition 4.4.2 that

$$\mathbf{E}\left|\langle \operatorname{diag}(\mathbf{fp})D\rangle\right|^p = N^{-p}\sum_{\Gamma\in\mathcal{G}(p)}\operatorname{Val}(\Gamma) + \mathcal{O}\left(N^{-p}\right). \tag{4.52}$$

We note that in contrast to the value definition for double index graphs (4.46), where each average in (4.46b) contains an $1/N$ prefactor, the single index graph value (4.49) does not include the N^{-p} prefactor. We chose this convention in this paper mainly because the exponent p in the prefactor N^{-p} cannot be easily read off from the single index graph itself, whereas in the double index graph p is simply the number of connected components.

We now collect some simple facts about the structure of these graphs in $\mathcal{G}(p)$ which directly follow from the corresponding properties of the double index graphs listed in Proposition 4.4.1.

Fact 4.1. *The interaction edges* $\operatorname{IE}(\Gamma)$ *form a perfect matching of* Γ, *in particular* $|V| = 2\,|\mathrm{IE}|$. *Moreover,* $1 \le |\mathrm{IE}(\Gamma)| \le p$ *and therefore the number of vertices in the graph is even and satisfies* $2 \le |V(\Gamma)| \le 2p$. *Finally, for* $(u,v) = e \in \operatorname{IE}(\Gamma)$ *we have* $\deg^-(u) = \deg^+(v)$, $\deg^+(u) = \deg^-(v)$ *and consequently also* $\deg(e) := \deg(u) = \deg(v)$. *The degree furthermore satisfies the bounds* $2 \le \deg(e) \le 6p$ *for each* $e \in \operatorname{IE}(\Gamma)$.

Fact 4.2. *The weights associated with the vertices are some non-negative powers of* $\mathbf{fp}$ *in such a way that the total power of all* $\mathbf{fp}$*'s is exactly p. The trivial zeroth power, i.e. the constant weight* $\mathbf{1}$ *is allowed. Furthermore, the* $\mathbf{fp}$ *weights are distributed in such a way that at least one non-trivial* $\mathbf{fp}$ *weight is associated with each interacting edge* $(u,v) = e \in \operatorname{IE}(\Gamma)$.

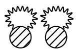
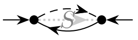
4.4.4 Examples of graphs

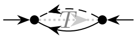
We now turn to some examples explaining the relation the of double index graphs from [DS3] and single index graphs. We note that the single index graphs actually contain more information because they specify edge direction, specify weights explicitly and differentiate between G and G^* edges. These information were not necessary for the power counting arguments used in [DS3], but for the improved estimates they will be crucial.

We start with the graphs representing the following simple equality following from $\kappa(\alpha,\beta) = \mathbf{E}\,w_\alpha w_\beta$

$$\begin{aligned} &N^2\,\mathbf{E}\sum_{\alpha,\beta}\kappa(\alpha,\beta)\,\langle\operatorname{diag}(\mathbf{fp})\Delta^\alpha G\rangle\,\langle G^*\Delta^\beta\operatorname{diag}(\mathbf{fp})^*\rangle \\ &\quad= \sum_{a,b}s_{ab}(pf)_a^2\,\mathbf{E}\,G_{ba}G^*_{ab} + \sum_{a,b}t_{ab}(pf)_a(pf)_b\,\mathbf{E}\,G_{ba}G^*_{ba} \end{aligned}$$

which can be represented as

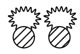
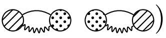
$N^2\,\mathrm{Val}\Big($  $\Big) = \mathrm{Val}\Big((\mathbf{pf})^2$ 

$\mathbf{1} \Big) + \mathrm{Val}\Big(\mathbf{pf}$ 

$\mathbf{pf} \Big)\,.$

We now turn to the complete graphical representation for the second moment in the case of Gaussian entries,

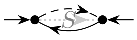
$$\begin{aligned}\mathbf{E}\,|\langle \mathrm{diag}(\mathbf{fp})D\rangle|^2 &= \mathbf{E}\,\langle \mathrm{diag}(\mathbf{fp})D\rangle\,\langle D^*\,\mathrm{diag}(\mathbf{fp})\rangle \qquad (4.53)\end{aligned}$$

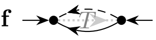
$= \mathrm{Val}\Big($  $\Big) + \mathrm{Val}\,($  $)$

$$\begin{aligned}&= \sum_{\alpha,\beta} \kappa(\alpha,\beta)\,\langle \mathrm{diag}(\mathbf{fp})\Delta^{\alpha} G\rangle\,\langle G^*\Delta^{\beta}\,\mathrm{diag}(\mathbf{fp})^*\rangle\\ &+ \sum_{\alpha_1,\beta_1}\sum_{\alpha_2,\beta_2} \kappa(\alpha_1,\beta_1)\kappa(\alpha_2,\beta_2) y\,\langle \mathrm{diag}(\mathbf{fp})\Delta^{\alpha_1}G\Delta^{\beta_2}G\rangle\,\langle G^*\Delta^{\beta_1}G^*\Delta^{\alpha_2}\,\mathrm{diag}(\mathbf{fp})^*\rangle\,,\end{aligned}$$

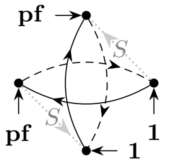
where we again stress that the double index graphs hide the specific weights and the fact that one of the connected components actually contains G^* edges. In terms of single index graphs, the rhs. of (4.53) can be represented as the sum over the values of the six graphs

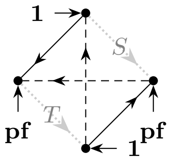
(4.54)

$N^2\,\mathbf{E}\,|\langle \mathrm{diag}(\mathbf{fp})D\rangle|^2 = \mathrm{Val}\Big((\mathbf{pf})^2$ 

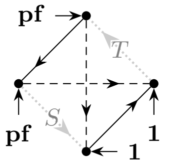
$\mathbf{1}\Big) + \mathrm{Val}\Big(\mathbf{pf}$ 

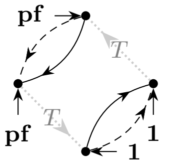
$\mathbf{pf}\Big)$

$+ \mathrm{Val}\Bigg($ 

$\Bigg) + \mathrm{Val}\Bigg($ 

$\Bigg)$

$+ \mathrm{Val}\Bigg($ 

$\Bigg) + \mathrm{Val}\Bigg($ 

$\Bigg)$

The first two graphs were already explained above. The additional four graphs come from the second term in the rhs. of (4.53). Since $\kappa(\alpha_1,\beta_1)$ is non-zero only if $\alpha_1=\beta_1$ or $\alpha_1=\beta_1^t$, there are four possible choices of relations among the α and β labels in the two kappa factors. For example, the first graph in the second line of (4.54) corresponds to the choice $\alpha_1^t=\beta_1$, $\alpha_2^t=\beta_2$. Written out explicitly with summation over single indices, this value is given by

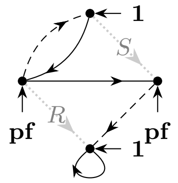
$$\sum_{a_1,b_1}\sum_{a_2,b_2} (pf)_{a_1}(pf)_{b_2} s_{a_1b_1}s_{a_2b_2}\,\mathbf{E}\,G_{a_2a_1}G_{b_1b_2}G^*_{a_1a_2}G^*_{b_2b_1}$$

where in the picture the left index corresponds to a_1, the top index to b_2, the right one to a_2 and the bottom one to b_1.

We conclude this section by providing an example of a graph with some degree higher than two which only occur in the non-Gaussian situation and might contain looped edges. For example, in the expansion of $N^2\,\mathbf{E}\,|\langle \mathrm{diag}(\mathbf{fp})D\rangle|^2$ in the non-Gaussian setup there is the term

$$\sum_{\substack{a_1,b_1\\a_2,b_2}} r_{a_1b_1}s_{a_2b_2}\,\mathbf{E}\left\langle \mathrm{diag}(\mathbf{fp})\Delta^{a_1b_1}G\Delta^{b_1a_1}G\Delta^{b_2a_2}G\right\rangle \left\langle G^*\Delta^{b_1a_1}G^*\Delta^{a_2b_2}\,\mathrm{diag}(\mathbf{fp})^*\right\rangle$$

$$= \mathrm{Val}\Bigg($$



$$\Bigg),$$

where $r_{ab} = \kappa(ab, ba, ba)/2$ and $s_{ab} = \kappa(ab, ba)$, in accordance with (4.48).

4.4.5 Simple Estimates on $\mathrm{Val}(\Gamma)$

In most cases we aim only at estimating the value of a graph instead of precisely computing it. The simplest power counting estimate on (4.49) uses that the matrix elements of G and those of the generic weight matrix K are bounded by an $\mathcal{O}(1)$ constant, while the matrix elements of $R^{(e)}$ are bounded by $N^{-\deg(e)/2}$. Thus the naive estimate on (4.49) is

$$\begin{aligned} |\mathrm{Val}(\Gamma)| &\lesssim \Big(\prod_{v\in V(\Gamma)} N\Big)\Big(\prod_{e\in \mathrm{IE}(\Gamma)} N^{-\deg(e)/2}\Big) \\ &= \prod_{e\in\mathrm{IE}(\Gamma)} N^{2-\deg(e)/2} \le \prod_{e\in\mathrm{IE}(\Gamma)} N \le N^p \end{aligned} \tag{4.55}$$

where we used that the interaction edges form a perfect matching and that $\deg(e) \ge 2$, $|\mathrm{IE}(\Gamma)| \le p$. The somewhat informal notation $\lesssim$ in (4.55) hides a technical subtlety. The resolvent entries G_{ab} are indeed bounded by an $\mathcal{O}(1)$ constant in the sense of very high moments but not almost surely. We will make bounds like the one in (4.55) rigorous in a high moments sense in Lemma 4.4.8.

The estimate (4.55) ignores the fact that typically only the diagonal resolvent matrix elements of G are of $\mathcal{O}(1)$, the off-diagonal matrix elements are much smaller. This is manifested in the *Ward-identity*

$$\sum_{a\in J} |G_{ab}|^2 = (G^*G)_{bb} = \frac{(G-G^*)_{bb}}{2i\eta} = \frac{\Im G_{bb}}{\eta}. \tag{4.56a}$$

Thus the sum of off-diagonal resolvent elements G_{ab} is usually smaller than its naive size of order N, at least in the regime $\eta \gg N^{-1}$. This is quantified by the so called Ward estimates

$$\sum_{a\in J}|G_{ab}|^2 = N\frac{\Im G_{bb}}{N\eta} \lesssim N\psi^2, \qquad \sum_{a\in J}|G_{ab}| \lesssim N\psi, \qquad \psi := \left(\frac{\rho}{N\eta}\right)^{1/2}. \tag{4.56b}$$

Similarly to (4.55) the inequalities $\lesssim$ in (4.56b) are meant in a power counting sense ignoring that the entries of $\Im G$ might not be bounded by ρ almost surely but only in some high moment sense.

As a consequence of (4.56b) we can gain a factor of ψ for each off-diagonal (that is, connecting two separate vertices) G-factor, but clearly only for at most two G-edges per adjacent vertex. Moreover, this gain can obviously only be used once for each edge and not twice, separately when summing up the indices at both adjacent vertices. As a consequence a careful counting of the total number of ψ-gains is necessary, see Section 2.4.3 for details.

Ward bounds for the example graphs from Section 4.4.4. From the single index graphs drawn in (4.54) we can easily obtain the known bound $\mathbf{E}\,|\langle \operatorname{diag}(\mathbf{fp})D\rangle|^2 \lesssim \psi^4$. Indeed, the last four graphs contribute a combinatorial factor of N^4 from the summations over four single indices and a scaling factor of N^{-2} from the size of S, T. Furthermore, we can gain a factor of ψ for each G-edge through Ward estimates and the bound follows. Similarly, the first two graphs contribute a factor of $N = N^{2-1}$ from summation and S/T and a factor of ψ^2 from the Ward estimates, which overall gives $N^{-1}\psi^2 \lesssim \psi^4$. As this example shows, the bookkeeping of available Ward-estimates is important and we will do so systematically in the following sections.

4.4.6 Improved estimates on $\operatorname{Val}(\Gamma)$: Wardable edges

For the sake of transparency we briefly recall the combinatorial argument used in [DS3], which also provides the starting point for the refined estimate in the present paper. Compared to [DS3], however, we phrase the counting argument directly in the language of the single index graphs. We only aim to gain from the G-edges adjacent to vertices of degree two or three; for vertices of higher degree the most naive estimate $|G_{ab}| \lesssim 1$ is already sufficient as demonstrated in [DS3]. We collect the vertices of degree two and three in the set $V_{2,3}$ and collect the G-edges adjacent to $V_{2,3}$ in the set $E_{2,3}$. In Section 2.4.3 a specific *marking procedure* on the G-edges of the graph is introduced that has the following properties. For each $v \in V_{2,3}$ we put a *mark* on at most two adjacent G-edges in such a way that those edges can be estimated via (4.56b) while performing the a_v summation. In this case we say that the mark comes from the v-perspective. An edge may have two marks coming from the perspective of each of its adjacent vertices. Later, marked edges will be estimated via (4.56b) while summing up a_v. After doing this for all of $V_{2,3}$ we call an edge in $E_{2,3}$ *marked effectively* if it either *(i)* has two marks, or *(ii)* has one mark and is adjacent to only one vertex from $V_{2,3}$. While subsequently using (4.56b) in the summation of a_v for $v \in V_{2,3}$ (in no particular order) on the marked edges (and estimating the remaining edges adjacent to v trivially) we can gain at least as many factors of ψ as there are *effectively marked* edges. Indeed, this follows simply from the fact that *effectively marked* edges are never estimated trivially during the procedure just described, no matter the order of vertex summation.

Fact 4.3. *For each $\Gamma \in \mathcal{G}(p)$ there is a marking of edges adjacent to vertices of degree at most 3 such that there are at least $\sum_{e\in \mathrm{IE}(\Gamma)}(4-\deg(e))_+$ effectively marked edges.*

Proof. On the one hand we find from Fact 4.1 (more specifically, from the equality $\deg(e) = \deg(u) = \deg(v)$ for $(u,v) = e \in \mathrm{IE}(\Gamma)$) that

$$|E_{2,3}| \ge \sum_{v\in V_{2,3}} \frac{1}{2}\deg(v) = \sum_{e\in \mathrm{IE}(\Gamma), \deg(e)\in\{2,3\}} \deg(e). \tag{4.57}$$

On the other it can be checked that for every pair $(u,v) = e \in \mathrm{IE}(\Gamma)$ with $\deg(e) = 2$ all G-edges adjacent to u or v can be marked from the u, v-perspective. Indeed, this is a direct consequence of Proposition 4.4.1(d): Because the two vertices in the double index graph being resolved to (u, v) cannot be part of the same cycle it follows that all of the (two, three or four) G-edges adjacent to the vertices with index u or v are not loops (i.e. do not represent diagonal resolvent elements). They are cyclically marked and can thereby bounded by using (4.56b). Similarly, it can be checked that for every edge $(u, v) = e \in \mathrm{IE}(\Gamma)$ with $\deg(e) = 3$ at most two G-edges adjacent to u or v can remain unmarked from the u, v-perspective. By combining these two observations it follows that at most

$$\sum_{e\in \mathrm{IE}(\Gamma),\deg(e)\in\{2,3\}} (2\deg(e) - 4) \tag{4.58}$$

edges in $E_{2,3}$ are *ineffectively marked* since those are counted as unmarked from the perspective of one of its vertices. Subtracting (4.58) from (4.57) it follows that in total at least

$$\sum_{e\in \mathrm{IE}(\Gamma)} (4 - \deg(e))_+ = \sum_{e\in \mathrm{IE}(\Gamma),\deg(e)\in\{2,3\}} (4 - \deg(e))$$

edges are marked effectively, just as claimed. □

In [DS3] it was sufficient to estimate the value of each graph in $\mathcal{G}(p)$ by subsequently estimating all effectively marked edges using (4.56b). For the purpose of improving the local law at the cusp, however, we need to introduce certain operations on the graphs of $\mathcal{G}(p)$ which allow to estimate the graph value to a higher accuracy. It is essential that during those operations we keep track of the number of edges we estimate using (4.56b). Therefore we now introduce a more flexible way of recording these edges. We first recall a basic definition [129] from graph theory.

Definition 4.4.4. *For $k \geq 1$ a graph $\Gamma = (V, E)$ is called* k-degenerate *if any induced subgraph has minimal degree at most k.*

It is well known that being k-degenerate is equivalent to the following sequential property[1]. We provide a short proof for convenience.

Lemma 4.4.5. *A graph $\Gamma = (V, E)$ is k-degenerate if and only if there exists an ordering of vertices $\{v_1, \dots, v_n\} = V$ such that for each $m \in [n] := \{1, \dots, n\}$ it holds that*

$$\deg_{\Gamma[\{v_1,\dots,v_m\}]}(v_m) \leq k \tag{4.59}$$

where for $V' \subset V$, $\Gamma[V']$ denotes the induced subgraph on the vertex set V'.

Proof. Suppose the graph is k-degenerate and let $n := |V|$. Then there exists some vertex $v_n \in V$ such that $\deg(v_n) \leq k$ by definition. We now consider the subgraph induced by $V' := V \setminus \{v_n\}$ and, by definition, again find some vertex $v_{n-1} \in V'$ of degree $\deg_{\Gamma[V']}(v_{n-1}) \leq k$. Continuing inductively we find a vertex ordering with the desired property.

Conversely, assume there exists a vertex ordering such that (4.59) holds for each m. Let $V' \subset V$ be an arbitrary subset and let $m := \max\{\, l \in [n] \mid V_l \in V' \,\}$. Then it holds that

$$\deg_{\Gamma[V']}(v_m) \leq \deg_{\Gamma[\{v_1,\dots,v_m\}]}(v_m) \leq k$$

and the proof is complete. □

[1] This equivalent property is commonly known as having a *colouring number* of at most $k + 1$, see e.g. [84].

The reason for introducing this graph theoretical notion is that it is equivalent to the possibility of estimating edges effectively using (4.56b). A subset GE' of G-edges in $\Gamma \in \mathcal{G}$ can be fully estimated using (4.56b) if and only if there exists a vertex ordering such that we can subsequently remove vertices in such a way that in each step at most two edges from GE' are removed. Due to Lemma 4.4.5 this is the case if and only if $\Gamma' = (V, \mathrm{GE}')$ is 2-degenerate. For example, the graph $\Gamma_{\mathrm{eff}} = (V, \mathrm{GE}_{\mathrm{eff}})$ induced by the effectively marked G-edges $\mathrm{GE}_{\mathrm{eff}}$ is a 2-degenerate graph. Indeed, each effectively marked edge is adjacent to at least one vertex which has degree at most 2 in Γ_{eff}: Vertices of degree 2 in (V, GE) are trivially at most of degree 2 in Γ_{eff}, and vertices of degree 3 in (V, GE) are also at most of degree 2 in Γ_{eff} as they can only be adjacent to 2 effectively marked edges. Consequently any induced subgraph of Γ_{eff} has to contain some vertex of degree at most 2 and thereby Γ_{eff} is 2-degenerate.

Definition 4.4.6. *For a graph $\Gamma = (V, \mathrm{GE} \cup \mathrm{IE} \cup \mathrm{WE}) \in \mathcal{G}$ we call a subset of G-edges $\mathrm{GE}_W \subset \mathrm{GE}$ Wardable if the subgraph (V, GE_W) is 2-degenerate.*

Lemma 4.4.7. *For each $\Gamma \in \mathcal{G}(p)$ there exists a Wardable subset $\mathrm{GE}_W \subset \mathrm{GE}$ of size*

$$|\mathrm{GE}_W| = \sum_{e \in \mathrm{IE}} (4 - \deg(e))_+. \tag{4.60}$$

Proof. This follows immediately from Fact 4.3, the observation that $(V, \mathrm{GE}_{\mathrm{eff}})$ is 2-degenerate and the fact that sub-graphs of 2-degenerate graphs remain 2-degenerate. □

For each $\Gamma \in \mathcal{G}(p)$ we choose a Wardable subset $\mathrm{GE}_{\mathrm{W}}(\Gamma) \subset \mathrm{GE}(\Gamma)$ satisfying (4.60). At least one such set is guaranteed to exist by the lemma. For graphs with several possible such sets, we arbitrarily choose one, and consider it permanently assigned to Γ. Later we will introduce certain operations on graphs $\Gamma \in \mathcal{G}(p)$ which produce families of derived graphs $\Gamma' \in \mathcal{G} \supset \mathcal{G}(p)$. During those operations the chosen Wardable subset $\mathrm{GE}_{\mathrm{W}}(\Gamma)$ will be modified in order to produce eligible sets of Wardable edges $\mathrm{GE}_{\mathrm{W}}(\Gamma')$ and we will select one among those to define the Wardable subset of Γ'. We stress that the relation (4.60) on the Wardable set is required only for $\Gamma \in \mathcal{G}(p)$ but not for the derived graphs Γ'.

We now give a precise meaning to the vague bounds of (4.55), (4.56b). We define the N-exponent, $n(\Gamma)$, of a graph $\Gamma = (V, \mathrm{GE} \cup \mathrm{IE} \cup \mathrm{WE})$ as the effective N-exponent in its value-definition, i.e. as

$$n(\Gamma) := |V| - \sum_{e \in \mathrm{IE}} \frac{\deg(e)}{2} - \sum_{e \in \mathrm{WE}} l(e).$$

We defer the proof of the following technical lemma to the appendix.

Lemma 4.4.8. *Let $\Gamma = (V, \mathrm{GE} \cup \mathrm{IE} \cup \mathrm{WE}) \in \mathcal{G}$ be a graph with Wardable edge set $\mathrm{GE}_W \subset \mathrm{GE}$ and at most $|V| \le cp$ vertices and at most $|\mathrm{GE}| \le cp^2$ G-edges. Then there exists a constant $0 < C < \infty$ such that for each $0 < \epsilon < 1$ it holds that*

$$|\mathrm{Val}(\Gamma)| \le_\epsilon N^{\epsilon p} (1 + \|G\|_q)^{Cp^2}\, \text{W-Est}(\Gamma), \tag{4.61a}$$

where

$$\text{W-Est}(\Gamma) := N^{n(\Gamma)} (\psi + \psi_q')^{|\mathrm{GE}_W|} (\psi + \psi_q' + \psi_q'')^{|\mathrm{GE}_{g-m}|}, \qquad q := Cp^3/\epsilon. \tag{4.61b}$$

Remark 4.4.9.

(i) We consider ϵ and p as fixed within the proof of Theorem 4.3.7 and therefore do not explicitly carry the dependence of them in quantities like W-Est.

(ii) We recall that the factors involving GE_{g-m} *and* WE *do not play any role for graphs* $\Gamma \in \mathcal{G}(p)$ *as those sets are empty in this restricted class of graphs (see Remark 4.4.3).*

(iii) Ignoring the difference between ψ *and* ψ_q', ψ_q'' *and the irrelevant order* $\mathcal{O}\left(N^{p\epsilon}\right)$ *factor in* (4.61)*, the reader should think of* (4.61) *as the heuristic inequality*

$$|\mathrm{Val}(\Gamma)| \lesssim N^{n(\Gamma)} \psi^{|\mathrm{GE}_W| + |\mathrm{GE}_{g-m}|}.$$

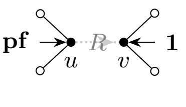
Using Lemma 4.4.7, $N^{-1/2} \lesssim \psi \lesssim 1$, $|V| = 2\,|\mathrm{IE}| \leq 2p$ *and* $\deg(e) \geq 2$ *(from Fact 4.1) we thus find*

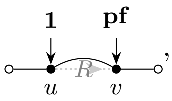
$$\begin{aligned} N^{-p}\,|\mathrm{Val}(\Gamma)| &\lesssim N^{|\mathrm{IE}|-p} \prod_{e \in \mathrm{IE}} N^{1-\deg(e)/2} \psi^{(4-\deg(e))_+} \\ &\lesssim \psi^{2|\mathrm{IE}|-2p} \prod_{e \in \mathrm{IE}} \psi^{\deg(e)-2+(4-\deg(e))_+} \leq \psi^{2p} \end{aligned} \tag{4.62}$$

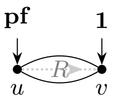
for any $\Gamma = (V, \mathrm{GE} \cup \mathrm{IE}) \in \mathcal{G}(p)$.

4.4.7 Improved estimates on $\mathrm{Val}(\Gamma)$ at the cusp: σ-cells

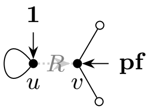
Definition 4.4.10. *For* $\Gamma \in \mathcal{G}$ *we call an interaction edge* $(u, v) = e \in \mathrm{IE}(\Gamma)$ *a* σ*-cell if the following four properties hold: (i)* $\deg(e) = 2$*, (ii) there are no* G*-loops adjacent to* u *or* v*, (iii) precisely one of* u, v *carries a weight of* **pf** *while the other carries a weight of* **1***, and (iv),* e *is not adjacent to any other non* GE*-edges. Pictorially, possible* σ*-cells are given by*



,


,


but not by



.

For $\Gamma \in \mathcal{G}$ *we denote the number of* σ*-cells in* Γ *by* $\sigma(\Gamma)$.

Next, we state a simple lemma, estimating $\text{W-Est}(\Gamma)$ of the graphs in the restricted class $\Gamma \in \mathcal{G}(p)$.

Lemma 4.4.11. *For each* $\Gamma = (V, \mathrm{IE} \cup \mathrm{GE}) \in \mathcal{G}(p)$ *it holds that*

$$N^{-p}\,|\text{W-Est}(\Gamma)| \leq_p \left(\sqrt{\eta/\rho}\right)^{p-\sigma(\Gamma)} (\psi + \psi_q')^{2p} \prod_{\substack{e \in \mathrm{IE} \\ \deg(e) \geq 4}} N^{2-\deg(e)/2}.$$

Proof. We introduce the short-hand notations $\mathrm{IE}_k := \{\, e \in \mathrm{IE} \mid \deg(e) = k \,\}$ and $\mathrm{IE}_{\geq k} := \bigcup_{l \geq k} \mathrm{IE}_l$. Starting from (4.61b) and Lemma 4.4.7 we find

$$\begin{aligned} N^{-p}\,|\text{W-Est}(\Gamma)| \leq N^{-(p-|\mathrm{IE}|)} &\left(\prod_{e \in \mathrm{IE}_2} (\psi + \psi_q')^2 \right) \left(\prod_{e \in \mathrm{IE}_3} \frac{\psi + \psi_q'}{\sqrt{N}} \right) \\ &\times \left(\prod_{e \in \mathrm{IE}_{\geq 4}} \frac{1}{N} \right) \left(\prod_{e \in \mathrm{IE}_{\geq 4}} N^{2-\deg(e)/2} \right). \end{aligned}$$

Using $N^{-1/2} = \psi\sqrt{\eta/\rho} \le C\psi$ it then follows that

$$\begin{aligned} N^{-p}\left|\text{W-Est}(\Gamma)\right| \le_p \left[\frac{\eta}{\rho}\psi^2\right]^{p-|\mathrm{IE}|} &\left(\prod_{e\in \mathrm{IE}_2}(\psi+\psi_q')^2\right) \\ \times &\left(\prod_{e\in\mathrm{IE}_{\ge 3}}\sqrt{\frac{\eta}{\rho}}(\psi+\psi_q')^2\right)\left(\prod_{e\in\mathrm{IE}_{\ge 4}} N^{2-\deg(e)/2}\right). \end{aligned} \tag{4.63}$$

It remains to relate (4.63) to the number $\sigma(\Gamma)$ of σ-cells in Γ. Since each interaction edge which is not a σ-cell has an additional weight $\mathbf{pf}$ attached to it, it follows from Fact 4.2 that $|\mathrm{IE}| - \sigma(\Gamma) \le p - |\mathrm{IE}|$. Therefore, from (4.63), $|\mathrm{IE}_2| \le |\mathrm{IE}|$ and $\eta/\rho \le C$ we have that

$$\begin{aligned} N^{-p}\left|\text{W-Est}(\Gamma)\right| \le_p &\left[\sqrt{\eta/\rho}(\psi+\psi_q')^2\right]^{p-|\mathrm{IE}|+\left|\mathrm{IE}_{\ge 3}\right|+|\mathrm{IE}_2|-\sigma(\Gamma)} \\ &\times\left[(\psi+\psi_q')^2\right]^{\sigma(\Gamma)}\left(\prod_{e\in\mathrm{IE}_{\ge 4}} N^{2-\deg(e)/2}\right), \end{aligned}$$

proving the claim. □

Using Lemma 4.4.8 and $\sqrt{\eta/\rho} \le \sigma_q$, the estimate in Lemma 4.4.11 has improved the previous bound (4.62) by a factor $\sigma_q^{p-\sigma(\Gamma)}$ (ignoring the irrelevant factors). In order to prove (4.19c), we thus need to remove the $-\sigma(\Gamma)$ from this exponent, in other words, we need to show that from each σ-cell we can multiplicatively gain a factor of σ_q. This is the content of the following proposition.

Proposition 4.4.12. *Let $\Gamma \in \mathcal{G}$ be a single index graph with at most cp vertices and cp^2 edges with a σ-cell $(u,v) = e \in \mathrm{IE}(\Gamma)$. Then there exists a finite collection of graphs $\{\Gamma_\sigma\} \sqcup \mathcal{G}_\Gamma$ with at most one additional vertex and at most $6p$ additional G-edges such that*

$$\begin{aligned} &\mathrm{Val}(\Gamma) = \sigma\,\mathrm{Val}(\Gamma_\sigma) + \sum_{\Gamma'\in\mathcal{G}_\Gamma}\mathrm{Val}(\Gamma') + \mathcal{O}\left(N^{-p}\right), \\ &\text{W-Est}(\Gamma_\sigma) = \text{W-Est}(\Gamma), \qquad \text{W-Est}(\Gamma') \le_p \sigma_q\,\text{W-Est}(\Gamma), \quad \Gamma'\in\mathcal{G}_\Gamma \end{aligned} \tag{4.64}$$

and all graphs Γ_σ and $\Gamma' \in \mathcal{G}_\Gamma$ have exactly one σ-cell less than Γ.

Using Lemma 4.4.8 and Lemma 4.4.11 together with the repeated application of Proposition 4.4.12 we are ready to present the proof of Theorem 4.3.7.

Proof of Theorem 4.3.7. We remark that the isotropic local law (4.19a) and the averaged local law (4.19b) are verbatim as in Theorem 2.4.1. We therefore only prove the improved bound (4.19c)–(4.19d) in the remainder of the section. We recall (4.52) and partition the set of graphs $\mathcal{G}(p) = \mathcal{G}_0(p) \cup \mathcal{G}_{\ge 1}(p)$ into those graphs $\mathcal{G}_0(p)$ with no σ-cells and those graphs $\mathcal{G}_{\ge 1}(p)$ with at least one σ-cell. For the latter group we then use Proposition 4.4.12 for some σ-cell to find

$$\begin{aligned} \mathbf{E}\left|\langle \mathrm{diag}(\mathbf{pf})D\rangle\right|^p = N^{-p}&\sum_{\Gamma\in\mathcal{G}_0(p)}\mathrm{Val}(\Gamma) \\ &+ N^{-p}\sum_{\Gamma\in\mathcal{G}_{\ge 1}(p)}\left(\sigma\,\mathrm{Val}(\Gamma_\sigma) + \sum_{\Gamma'\in\mathcal{G}_\Gamma}\mathrm{Val}(\Gamma')\right) + \mathcal{O}\left(N^{-2p}\right), \end{aligned} \tag{4.65}$$

where the number of σ-cells is reduced by 1 for Γ_σ and each $\Gamma' \in \mathcal{G}_\Gamma$ as compared to Γ. We note that the Ward-estimate $\text{W-Est}(\Gamma)$ from Lemma 4.4.11 together with Lemma 4.4.8 is already sufficient for the graphs in $\mathcal{G}_0(p)$. For those graphs $\mathcal{G}_1(p)$ with exactly one σ-cell the expansion in (4.65) is sufficient because $\sigma \le \sigma_q$ and, according to (4.64), each $\Gamma' \in \mathcal{G}_\Gamma$ has a Ward estimate which is already improved by σ_q. For the other graphs we iterate the expansion from Proposition 4.4.12 until no sigma cells are left.

It only remains to count the number of G-edges and vertices in the successively derived graphs to make sure that Lemma 4.4.8 and Proposition 4.4.12 are applicable and that the last two factors in (4.19c) come out as claimed. Since every of the $\sigma(\Gamma) \le p$ applications of Proposition 4.4.12 creates at most $6p$ additional G-edges and one additional vertex, it follows that $|\mathrm{GE}(\Gamma)| \le C'p^2$, $|V| \le C'p$ also in any successively derived graph. Finally, it follows from the last factor in Lemma 4.4.11 that for each $e \in \mathrm{IE}$ with $\deg(e) \ge 5$ we gain additional factors of $N^{-1/2}$. Since $|\mathrm{IE}| \le p$, we easily conclude that if there are more than $4p$ G-edges, then each of them comes with an additional gain of $N^{-1/2}$. Now (4.19c) follows immediately after taking the p-th root.

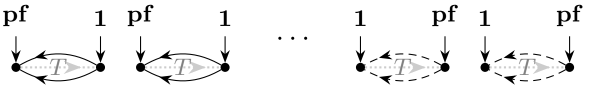
We turn to the proof of (4.19d). We first write out

$$\langle \operatorname{diag}(\mathbf{pf})[T \odot G^t]G\rangle = \frac{1}{N}\sum_{a,b}(pf)_a t_{ab} G_{ba} G_{ba}$$

and therefore can, for even p, write the p-th moment as the value

$$\mathbf{E}\left|\langle \operatorname{diag}(\mathbf{pf})[T \odot G^t]G\rangle\right|^p = N^{-p}\,\mathrm{Val}(\Gamma_0)$$

of the graph $\Gamma_0 = (V, \mathrm{GE}\cup\mathrm{IE}) \in \mathcal{G}$ which is given by p disjoint 2-cycles as

$\Gamma_0 =$ 

,

where there are $p/2$ cycles of G-edges and $p/2$ cycles of G^* edges. It is clear that (V, GE) is 2-degenerate and since $|\mathrm{GE}| = 2p$ it follows that

$$\text{W-Est}(\Gamma_0) \le N^p(\psi + \psi_q')^{2p}.$$

On the other hand each of the p interaction edges in Γ_0 is a σ-cell and we can use Proposition 4.4.12 p times to obtain (4.19d) just as in the proof of (4.19c). □

4.4.8 Proof of Proposition 4.4.12

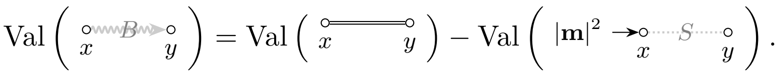
It follows from the MDE that

$$G = M - M\mathcal{S}[M]G - MWG = M - G\mathcal{S}[M]M - GWM,$$

which we use to locally expand a term of the form $G_{xa}G^*_{ay}$ for fixed a, x, y further. To make the computation local we allow for an arbitrary random function $f = f(W)$, which in practice encodes the remaining G-edges in the graph. A simple cumulant expansion shows

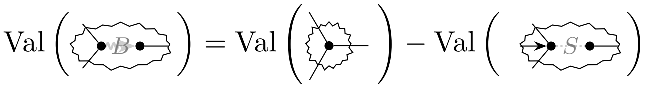
$$\begin{aligned}\sum_b B_{ab}\,\mathbf{E}\,G_{xb}G^*_{by}f &= \mathbf{E}\,M_{xa}G^*_{ay}f - \sum_{k=2}^{6p}\sum_b\sum_{\beta\in I^k}\kappa(ba,\underline{\beta})m_a\,\mathbf{E}\,\partial_\beta\Big[G_{xb}G^*_{ay}f\Big]\\ &\quad+\sum_b s_{ba}m_a\,\mathbf{E}\Big[G_{xa}(g-m)_bG^*_{ay}+G_{xb}\overline{(g-m)}_aG^*_{by}-G_{xb}G^*_{ay}\partial_{ab}\Big]f+\mathcal{O}\left(N^{-p}\right)\\ &\quad+\sum_b t_{ba}m_a\,\mathbf{E}\Big[G_{xb}(G-M)_{ab}G^*_{ay}+G_{xb}G^*_{ab}G^*_{ay}-G_{xb}G^*_{ay}\partial_{ba}\Big]f \qquad (4.66)\end{aligned}$$

where $\partial_\alpha := \partial_{w_\alpha}$ and introduced the stability operator $B := 1-\operatorname{diag}(|\mathbf{m}|^2)S$. The stability operator B appears from rearranging the equation obtained from the cumulant expansion to express the quantity $\mathbf{E}\,G_{xb}G^*_{by}f$. In our graphical representation, the stability operator is a special edge that we can also express as



(4.67)

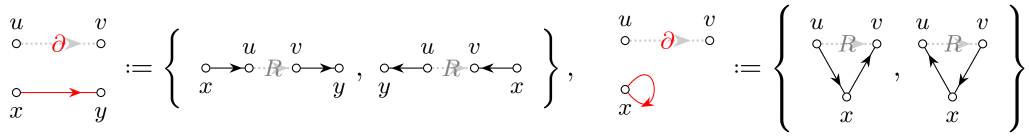
An equality like (4.67) is meant locally in the sense that the pictures only represent subgraphs of the whole graph with the empty, labelled vertices symbolizing those vertices which connect the subgraph to its complement. Thus (4.67) holds true for every fixed graph extending x, y consistently in all three graphs. The doubly drawn edge in (4.67) means that the external vertices x, y are identified with each other and the associated indices are set equal via a δ_{a_x,a_y} function. Thus (4.67) should be understood as the equality



(4.68)

where the outside edges incident at the merged vertices x, y are reconnected to one common vertex in the middle graph. For example, in the picture (4.68) the vertex x is connected to the rest of the graph by two edges, and the vertex y by one.

In order to represent (4.66) in terms of graphs we have to define a notion of *differential edge*. First, we define a *targeted differential edge* represented by an interaction edge with a red ∂-sign written on top and a red-coloured *target G-edge* to denote the collection of graphs

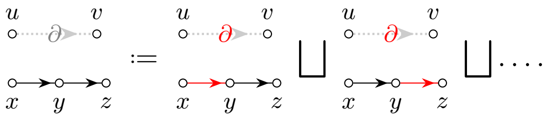


(4.69)

The second picture in (4.69) shows that the target G-edge may be a loop; the definition remains the same. This definition extends naturally to G^* edges and is exactly the same for $G-M$ edges (note that this is compatible with the usual notion of derivative as M does not depend on W). Graphs with the differential signs should be viewed only as an intermediate simplifying picture but they really mean the collection of graphs indicated in the right hand side of (4.69). They represent the identities

$$\begin{aligned}\sum_\alpha \kappa(uv,\alpha)\partial_{uv}G_{xy} &= -s_{uv}G_{xv}G_{uy}-t_{uv}G_{xu}G_{vy},\\ \sum_\alpha \kappa(uv,\alpha)\partial_{uv}G_{xx} &= -s_{uv}G_{xv}G_{ux}-t_{uv}G_{xu}G_{vx}\end{aligned}$$

In other words we introduced these graphs only to temporary encode expressions with derivatives (e.g. second term in the rhs. of (4.66)) *before* the differentiation is actually performed. We can then further define the action of an *untargeted differential edge* according the Leibniz rule as the collection of graphs with the differential edge being targeted on all G-edges of the graph one by one (in particular not only those in the displayed subgraph), i.e. for example



(4.70)

Here the union is a union in the sense of multisets, i.e. allows for repetitions in the resulting set (note that also this is compatible with the usual action of derivative operations). The $\sqcup\dots$ symbol on the rhs. of (4.70) indicates that the targeted edge cycles through *all* G-edges in the graph, not only the ones in the subgraph. For example, if there are k G-edges in the graph, then the picture (4.70) represents a collection of $2k$ graphs arising from performing the differentiation

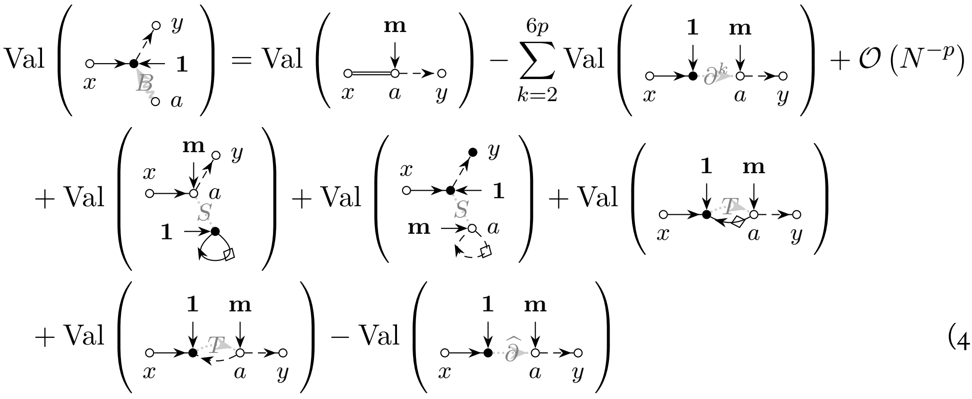
$$\begin{aligned}
&\sum_\alpha \kappa(uv,\alpha)\partial_{uv}\left[G_{xy}G_{yz}f\right]\\
&=\sum_\alpha \kappa(uv,\alpha)\left[\partial_{uv}G_{xy}\right]G_{yz}f+\sum_\alpha \kappa(uv,\alpha)G_{xy}\left[\partial_{uv}G_{yz}\right]f+\sum_\alpha \kappa(uv,\alpha)G_{xy}G_{yz}\left[\partial_{uv}f\right]\\
&=-s_{uv}\left[G_{xv}G_{uy}G_{yz}f+G_{xy}G_{yv}G_{uz}f+G_{xy}G_{yz}(\partial_{vu}f)\right]\\
&\qquad -t_{uv}\left[G_{xu}G_{vy}G_{yz}f+G_{xy}G_{yu}G_{vz}f+G_{xy}G_{yz}(\partial_{uv}f)\right],
\end{aligned}$$

where $f=f(W)$ represents the value of the G-edges outside the displayed subgraph.

Finally we introduce the notation that a differential edge which is targeted on all G-vertices except for those in the displayed subgraph. This differential edge targeted on the outside will be denoted by $\widehat{\partial}$.

Regarding the value of the graph, we define the value of a collection of graphs as the sum of their values. We note that this definition is for the collection of graphs encoded by the differential edges also consistent with the usual differentiation.

Written in a graphical form (4.66) reads



(4.71)

where the ultimate graph encodes the ultimate terms in the last two lines of (4.66).

We worked out the example for the resolution of the quantity $\mathbf{E}\, G_{xa}G^*_{ay}f$, but very similar formulas hold if the order of the fixed indices (x,y) and the summation index a

changes in the resolvents, as well as for other combinations of the complex conjugates. In graphical language this corresponds to changing the arrows of the two G-edges adjacent to a, as well as their types. In other words, equalities like the one in (4.71) hold true for other any degree two vertex but the stability operator changes slightly: In total there are 16 possibilities, four for whether the two edges are incoming or outgoing at a and another four for whether the edges are of type G or of type G^*. The general form for the stability operator is

$$B := 1 - \operatorname{diag}(\mathbf{m}^{\#_1}\mathbf{m}^{\#_2})R, \tag{4.72}$$

where $R = S$ if there is one incoming and one outgoing edge, $R = T$ if there are two outgoing edges and $R = T^t$ otherwise, and where $\#_1, \#_2$ represent complex conjugations if the corresponding edges are of G^* type. Thus for, for example, the stability operator in a for $G^*_{xa}G^*_{ya}$ is $1 - \operatorname{diag}(\overline{\mathbf{m}}^2)T^t$. Note that the stability operator at vertex with degree two is exclusively determined by the type and orientation of the two G-edges adjacent to a. In the sequel the letter B will refer to the appropriate stability operator, we will not distinguish their 9 possibilities ($R = S, T, T^t$ and $\mathbf{m}^{\#_1}\mathbf{m}^{\#_2} = |\mathbf{m}|^2, \mathbf{m}^2, \overline{\mathbf{m}}^2$) in the notation.

Lemma 4.4.13. *Let $\Gamma \in \mathcal{G}$ be a single index graph with at most cp vertices and cp^2 edges and let $a \in V(\Gamma)$ be a vertex of degree $\deg(a) = 2$. The insertion of the stability operator B (4.72) at a as in (4.71) produces a finite set of graphs with at most one additional vertex and $6p$ additional edges, denoted by $\mathcal{G}_\Gamma$, such that*

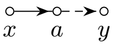
$$\operatorname{Val}(\Gamma) = \sum_{\Gamma' \in \mathcal{G}_\Gamma} \operatorname{Val}(\Gamma') + \mathcal{O}\left(N^{-p}\right),$$

and all of them have a Ward estimate

$$\text{W-Est}(\Gamma') \leq_p \left(\rho + \psi + \eta/\rho + \psi'_q + \psi''_q\right) \text{W-Est}(\Gamma) \leq_p \sigma_q \, \text{W-Est}(\Gamma), \qquad \Gamma' \in \mathcal{G}_\Gamma.$$

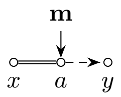
Moreover all σ-cells in Γ, except possibly a σ-cell adjacent to a, remain σ-cells also in each Γ'.

Proof. As the proofs for all of the 9 cases of B-operators are almost identical we prove the lemma for the case (4.71) for definiteness. Now we compare the value of the graph

$\Gamma :=$ 

with the graph in the lhs. of (4.71), i.e. when the stability operator B is attached to the vertex a. We remind the reader that the displayed graphs only show a certain subgraph of the whole graph. The goal is to show that $\text{W-Est}\,(\Gamma') \leq (\rho + \psi + \eta/\rho + \psi'_q + \psi''_q)\,\text{W-Est}(\Gamma)$ for each graph Γ' occurring on the rhs. of (4.71). The forthcoming reasoning is based on comparing the quantities $|V|$, $|\text{GE}_\text{W}|$, $|\text{GE}_{g-m}|$ and $\sum_{e \in \text{IE}} \deg(e)/2$ defining the Ward estimate W-Est from (4.61b) of the graph Γ and the various graphs Γ' occurring on the rhs. of (4.71).

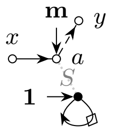
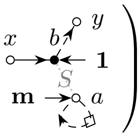
(a) We begin with the first graph and claim that

W-Est  $\leq \dfrac{1}{N\psi^2}\,\text{W-Est}(\Gamma) = \dfrac{\eta}{\rho}\,\text{W-Est}(\Gamma).$

Due to the double edge which identifies the x and a vertices it follows that $|V(\Gamma')| = |V(\Gamma)| - 1$. The degrees of all interaction edges remain unchanged when going from Γ to Γ'. As the 2-degenerate set of Wardable edges $\mathrm{GE}_\mathrm{W}(\Gamma')$ we choose $\mathrm{GE}_\mathrm{W}(\Gamma)\setminus N(a)$, i.e. the 2-degenerate edge set in the original graph except for the edge-neighbourhood $N(a)$ of a, i.e. those edges adjacent to a. As a subgraph of $(V, \mathrm{GE}_\mathrm{W}(\Gamma))$ it follows that $(V\setminus\{a\}, \mathrm{GE}_\mathrm{W}(\Gamma'))$ is again 2-degenerate. Thus $|\mathrm{GE}_\mathrm{W}(\Gamma)| \ge |\mathrm{GE}_\mathrm{W}(\Gamma')| \ge |\mathrm{GE}_\mathrm{W}(\Gamma)| - 2$ and the claimed bound follows since $|\mathrm{GE}_{g-m}(\Gamma')| = |\mathrm{GE}_{g-m}(\Gamma)|$ and

$$\frac{\text{W-Est}(\Gamma')}{\text{W-Est}(\Gamma)} = \frac{1}{N(\psi+\psi_q')^{|\mathrm{GE}_\mathrm{W}(\Gamma)|-|\mathrm{GE}_\mathrm{W}(\Gamma')|}} \le \frac{1}{N\psi^2}.$$

(b) Next, we consider the third and fourth graph and claim that

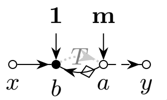
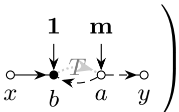
W-Est $\Bigg($  $\Bigg)$ + W-Est $\Bigg($  $\Bigg)$ $= 2(\psi + \psi_q' + \psi_q'')\,\text{W-Est}(\Gamma).$

Here there is one more vertex (corresponding to an additional summation index),

$$|V(\Gamma')| = |V(\Gamma)| + 1,$$

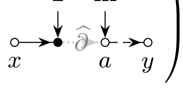
whose effect in (4.61b) is compensated by one additional interaction edge e of degree 2. Hence the N-exponent $n(\Gamma)$ remains unchanged. In the first graph we can simply choose $\mathrm{GE}_\mathrm{W}(\Gamma') = \mathrm{GE}_\mathrm{W}(\Gamma)$, whereas in the second graph we choose $\mathrm{GE}_\mathrm{W}(\Gamma') = \mathrm{GE}_\mathrm{W}(\Gamma)\setminus\{(x,a),(a,y)\}\cup\{(x,b),(b,y)\}$ which is 2-degenerate as a subgraph of a 2-degenerate graph together with an additional vertex of degree 2. Thus in both cases we can choose $\mathrm{GE}_\mathrm{W}(\Gamma')$ (if necessary, by removing excess edges from $\mathrm{GE}_\mathrm{W}(\Gamma')$ again) in such a way that $|\mathrm{GE}_\mathrm{W}(\Gamma')| = |\mathrm{GE}_\mathrm{W}(\Gamma)|$ but the number of $(g-m)$-loops is increased by 1, i.e. $|\mathrm{GE}_{g-m}(\Gamma')| = |\mathrm{GE}_{g-m}(\Gamma)| + 1$.

(c) Similarly, we claim for the fifth and sixth graph that

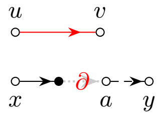
W-Est $\Bigg($  $\Bigg)$ + W-Est $\Bigg($  $\Bigg)$ $= 2(\psi + \psi_q')\,\text{W-Est}(\Gamma).$

There is one more vertex whose effect in (4.61b) is compensated by one more interaction edge of degree 2, whence the number N-exponent remains unchanged. The number of Wardable edges can be increased by one by setting $\mathrm{GE}_\mathrm{W}(\Gamma')$ to be a suitable subset of $\mathrm{GE}_\mathrm{W}(\Gamma)\setminus\{(x,a),(a,y)\}\cup\{(x,b),(a,b),(a,y)\}$ which is 2-degenerate as the subset of a 2-degenerate graph together with two vertices of degree 2. The number of $(g-m)$-loops remains unchanged.

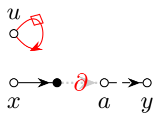
(d) For the last graph in (4.71), i.e. where the derivative targets an outside edge, we claim that

W-Est $\Bigg($  $\Bigg)$ $\le_p (\psi + \psi_q' + \psi_q'')\,\text{W-Est}(\Gamma).$

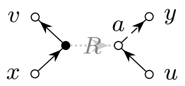
Here the argument on the lhs., Γ', stands for a whole collection of graphs but we essentially only have to consider two types: The derivative edge either hits a G-edge or a $(g-m)$-loop, i.e.



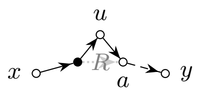
or



which encodes the graphs



and

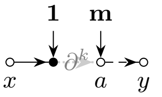


as well as the corresponding transpositions (as in (4.69)). In both cases the N-size of W-Est remains constant since the additional vertex is balanced by the additional degree two interaction edge. In both cases all four displayed edges can be included in $\mathrm{GE}_\mathrm{W}(\Gamma')$. So $|\mathrm{GE}_\mathrm{W}|$ can be increased by 1 in the first case and by 2 in the second case while the number of $(g-m)$-loops remains constant in the first case is decreased by 1 in the second case. The claim follows directly in the first case and from

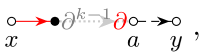
$$\frac{\text{W-Est}(\Gamma')}{\text{W-Est}(\Gamma)} = \frac{(\psi+\psi_q')^2}{\psi+\psi_q'+\psi_q''} \le \psi + \psi_q' + \psi_q''$$

in the second case.

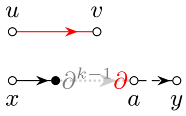
(e) It remains to consider the second graph in the rhs. of (4.71) with the higher derivative edge. We claim that for each $k\ge 2$ it holds that

W-Est () $\le_p (\psi+\psi_q')$ W-Est(Γ).

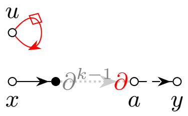
We prove the claim by induction on k starting from $k=2$. For any $k\ge 2$ we write $\partial^k = \partial^{k-1}\partial$. For the action of the last derivative we distinguish three cases: (i) action on an edge adjacent to the derivative edge, (ii) action on a non-adjacent G-edge and (iii) an action on a non-adjacent $(g-m)$-loop. Graphically this means



,

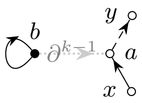


or

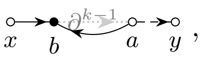


. (4.73)

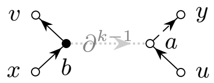
We ignored the case where the derivative acts on (a,y) since it is estimated identically to the first graph. We also neglected the possibility that the derivative acts on a g-loop, as this is estimated exactly as the last graph and the result is even better since no $(g-m)$-loop is destroyed. After performing the last derivative in (4.73) we obtain the following graphs Γ'



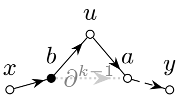
,



,



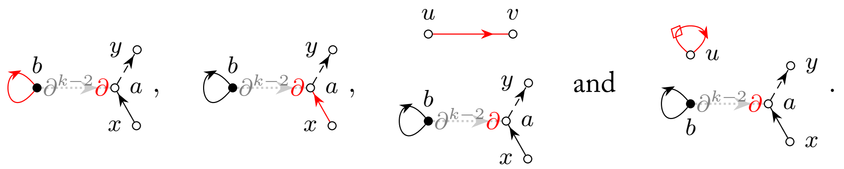
and



(4.74)

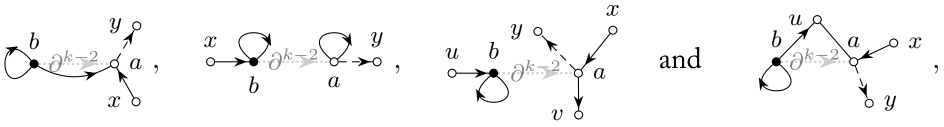
where we neglected the transposition of the third graph with u, v exchanged because this is equivalent with regard to the counting argument. First, we handle the second, third and fourth graphs in (4.74). In all these cases the set $\mathrm{GE}_{\mathrm{W}}(\Gamma')$ is defined simply by adding all edges drawn in (4.74) to the set $\mathrm{GE}_{\mathrm{W}}(\Gamma) \setminus \{(x, a), (a, y)\}$. The new set remains 2-degenerate since all these new edges are adjacent to vertices of degree 2. Compared to the original graph, Γ, we thus have increased $|\mathrm{GE}_{\mathrm{W}}| + |\mathrm{GE}_{g-m}|$ by at least 1.

We now continue with the first graph in (4.74), where we explicitly expand the action of another derivative (notice that this is the only graph where $k \geq 2$ is essentially used). We distinguish four cases, depending on whether the derivative acts on (i) the b-loop, (ii) an adjacent edge, (iii) a non-adjacent edge or (iv) a non-adjacent $(g - m)$-loop, i.e. graphically we have



(4.75)

After performing the indicated derivative, the encoded graphs Γ' are



(4.76)

where we again neglected the version of the third graph with u, v exchanged. We note that both the first and the second graph in (4.75) produce the first graph in (4.76). Now we define how to get the set $\mathrm{GE}_{\mathrm{W}}(\Gamma')$ from $\mathrm{GE}_{\mathrm{W}}(\Gamma) \setminus \{(x, a), (a, y)\}$ for each case. In the first graph of (4.76) we add all three non-loop edges to $\mathrm{GE}_{\mathrm{W}}(\Gamma')$, in the second graph we add both non-loop edges, in the third and fourth graph we add the non-looped edge adjacent to b as well as any two non-looped edges adjacent to a. Thus, compared to the original graph the number $|\mathrm{GE}_{\mathrm{W}}| + |\mathrm{GE}_{g-m}|$ is at least preserved. On the other hand the N-power counting is improved by $N^{-1/2}$. Indeed, there is one additional vertex b, yielding a factor N, which is compensated by the scaling factor $N^{-3/2}$ from the interaction edge of degree 3.

To conclude the inductive step we note that additional derivatives (i.e. the action of ∂^{k-2}) can only decrease the Ward-value of a graph. Indeed, any single derivative can at most decrease the number $|\mathrm{GE}_{\mathrm{W}}(\Gamma)| + |\mathrm{GE}_{g-m}|$ by 1 by either differentiating a $(g-m)$-loop or differentiating an edge from GE_{W}. Thus the number $|\mathrm{GE}_{\mathrm{W}}|+|\mathrm{GE}_{g-m}|$ is decreased by at most $k-2$ while the number $|\mathrm{GE}_{g-m}|$ is not increased. In particular, by choosing a suitable subset of Wardable edges, we can define $\mathrm{GE}_{\mathrm{W}}(\Gamma')$ in such a way that $|\mathrm{GE}_{\mathrm{W}}| + |\mathrm{GE}_{g-m}|$ is decreased by exactly $k - 2$. But at the same time each derivative provides a gain of $cN^{-1/2} \leq \psi \leq \psi + \psi_q'$ since the degree of the interaction edge is increased by one. Thus we have

$$\frac{\text{W-Est}(\Gamma')}{\text{W-Est}(\Gamma)} \leq_p (\psi + \psi_q')^{k-1+|\mathrm{GE}_{\mathrm{W}}(\Gamma')|+|\mathrm{GE}_{g-m}(\Gamma')|-|\mathrm{GE}_{\mathrm{W}}(\Gamma)|-|\mathrm{GE}_{g-m}(\Gamma)|} = \psi + \psi_q',$$

just as claimed. □

Lemma 4.4.13 shows that the insertion of the B-operator reduces the Ward-estimate by at least ρ. However, this insertion does not come for free since the inverse

$$B^{-1} = (1 - \operatorname{diag}(\mathbf{m}^{\#_1}\mathbf{m}^{\#_2})R)^{-1}$$

is generally not a uniformly bounded operator. For example, it follows from (4.2) that

$$\Im \mathbf{m} = \eta \left|\mathbf{m}\right|^2 + \left|\mathbf{m}\right|^2 S \Im \mathbf{m}$$

and therefore $(1 - \operatorname{diag}(|\mathbf{m}|^2)S)^{-1}$ is singular for small η with $\Im\mathbf{m}$ being the unstable direction. It turns out, however, that B is invertible on the subspace complementary to some bad direction $\mathbf{b}^{(B)}$. At this point we distinguish two cases. If B has a uniformly bounded inverse, i.e. if $\left\|B^{-1}\right\|_{\infty\to\infty} \le C$ for some constant $C > 0$, then we set $P_B := 0$. Otherwise we define P_B as the spectral projection operator onto the eigenvector $\mathbf{b}^{(B)}$ of B corresponding to the eigenvalue β with smallest modulus:

$$P_B := \frac{\langle \mathbf{l}^{(B)}, \cdot \rangle}{\langle \mathbf{l}^{(B)}, \mathbf{b}^{(B)} \rangle} \mathbf{b}^{(B)}, \qquad Q_B := 1 - P_B, \tag{4.77}$$

where $\langle \mathbf{v}, \mathbf{w} \rangle := N^{-1} \sum_a \overline{v_a} w_a$ denotes the normalized inner product and $\mathbf{l}^{(B)}$ is the corresponding left eigenvector, $(B^* - \beta)\mathbf{l}^{(B)} = 0$.

Lemma 4.4.14. *For all 9 possible B-operators in* (4.72) *it holds that*

$$\left\|B^{-1} Q_B\right\|_{\infty\to\infty} \le C < \infty \tag{4.78}$$

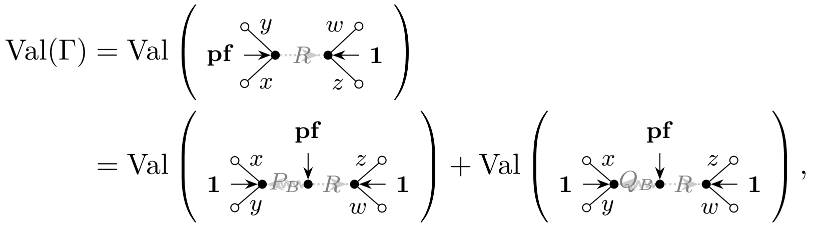
for some constant $C > 0$, depending only on model parameters.

Proof. First we remark that it is sufficient to prove the bound (4.78) on $B^{-1}Q_B$ as an operator on $\mathbb{C}^N$ with the Euclidean norm, i.e. $\left\|B^{-1}Q_B\right\| \le C$. For this insight we refer to [7, Proof of (5.28) and (5.40a)]. Recall that $R = S$, $R = T$ or $R = T^t$, depending on which stability operator we consider (cf. (4.72)). We begin by considering the complex hermitian symmetry class and the cases $R = T$ and $R = T^t$. We will now see that in this case B has a bounded inverse and thus $Q_B = 1$. Indeed, we have

$$\left\|B^{-1}\right\| \lesssim \frac{1}{1 - \|F^{(R)}\|},$$

where $F^{(R)}\mathbf{w} := |\mathbf{m}|\, R(|\mathbf{m}|\, \mathbf{w})$. The fullness Assumption (4.B) in (4.3) implies that $|t_{ij}| \le (1-c)s_{ij}$ for some constant $c > 0$ and thus $\|F^{(R)}\| \le (1-c)\left\|F^{(S)}\right\| \le 1-c$ for $R = T, T^t$. Here we used $\left\|F^{(S)}\right\| \le 1$, a general property of the saturated self-energy matrix $F^{(S)}$ that was first established in [10, Lemma 4.3] (see also [8, Eq. (4.24)] and [12, Eq. (4.5)]). Now we turn to the case $R = S$ for both the real symmetric and complex hermitian symmetry classes. In this case B is the restriction to diagonal matrices of an operator $\mathcal{T} : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$, where $\mathcal{T} \in \{\mathrm{Id} - M^*\mathcal{S}[\cdot]M, \mathrm{Id} - M\mathcal{S}[\cdot]M, \mathrm{Id} - M^*\mathcal{S}[\cdot]M^*\}$. All of these operators were covered in [12, Lemma 5.1] and thus (4.78) is a consequence of that lemma. Recall that the flatness (4.14) of $\mathcal{S}$ ensured the applicability of the lemma. □

We will insert the identity $1 = P_B + BB^{-1}Q_B$, and we will perform an explicit calculation for the P_B component, while using the boundedness of $B^{-1}Q_B$ in the other component. We are thus left with studying the effect of inserting B-operators and suitable projections into a σ-cell. To include all possible cases with regard to edge-direction and edge-type (i.e. G or G^*), in the pictures below we neither indicate directions of the G-edges nor their type but implicitly allow all possible assignments. We recall that both the R-interaction edge as well as the *relevant* B-operators (cf. (4.72)) are completely determined by the type of the four G-edges as well as their directions. To record the type of the inserted B, P_B, Q_B operators we call those inserted on the rhs. of the R-edge B', P_B' and Q_B' in the following graphical representations. Pictorially we start first decompose the σ-cell subgraph of some graph Γ as

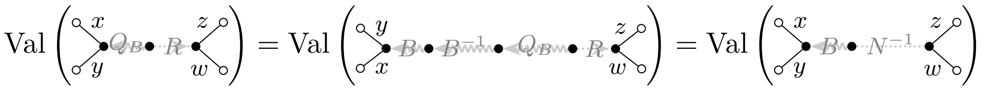


(4.79)

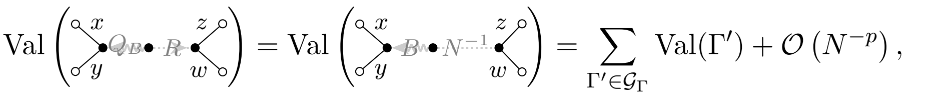
where we allow the vertices x, y to agree with z or w. With formulas, the insertion in (4.79) means the following identity

$$\sum_{ab} (pf)_a G_{ya} G_{xa} R_{ab} G_{bw} G_{bz} = \sum_{abc} (pf)_c G_{ya} G_{xa} (P_{ac} + Q_{ac}) R_{cb} G_{bw} G_{bz}$$

since $P_{ac} + Q_{ac} = \delta_{ac}$. We first consider with the second graph in (4.79), whose treatment is independent of the specific weights, so we already removed the weight information. We insert the B operator as



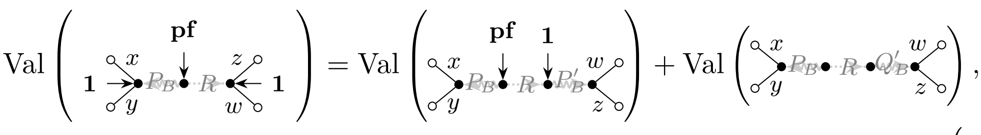
and notice that due to Lemma 4.4.14 the matrix $K = (B^{-1})^t Q_B^t R$, assigned to the weighted edge in the last graph, is entry-wise $|k_{ab}| \leq cN^{-1}$ bounded (the transpositions compensate for the opposite orientation of the participating edges). It follows from Lemma 4.4.13 that


Val
x
y
Q_B
R
z
w
= Val
x
y
B
N^{-1}
z
w
= $\sum_{\Gamma' \in \mathcal{G}_\Gamma} \mathrm{Val}(\Gamma') + \mathcal{O}(N^{-p})$,

(4.80)

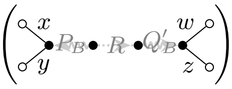
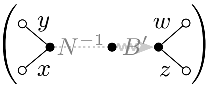
where all $\Gamma' \in \mathcal{G}_\Gamma$ satisfy W-Est$(\Gamma') \leq_p \sigma_q$ W-Est(Γ) and all σ-cells in Γ except for the currently expanded one remain σ-cells in Γ'. We note that it is legitimate to compare the Ward estimate of Γ' with that of Γ because with respect to the Ward-estimate there is no difference between Γ and the modification of Γ in which the R-edge is replaced by a generic N^{-1}-weighted edge.

We now consider the first graph in (4.79) and repeat the process of inserting projections $P_B' + Q_B'$ to the other side of the R-edge to find



(4.81)

where we already neglected those weights which are of no importance to the bound. The argument for the second graph in (4.81) is identical to the one we used in (4.80) and we find another finite collection of graphs $\mathcal{G}'_\Gamma$ such that

Val () = Val () $= \sum_{\Gamma' \in \mathcal{G}'_\Gamma} \mathrm{Val}\left(\Gamma'\right) + \mathcal{O}\left(N^{-p}\right),$

where the weighted edge carries the weight matrix $K = P^t_B R Q_{B'} B'^{-1}$, which is according to Lemma 4.4.14 indeed scales like $|k_{ab}| \le cN^{-1}$. The graphs $\Gamma' \in \mathcal{G}'_\Gamma$ also satisfy $\text{W-Est}(\Gamma') \le_p \sigma_q \,\text{W-Est}(\Gamma)$ and all σ-cells in Γ except for the currently expanded one remain σ-cells in Γ'.

It remains to consider the first graph in (4.81) in the situation where B does not have a bounded inverse. We compute the weight matrix of the $P^t_B R P'_B$ interaction edge as

$$\begin{aligned} P^t_B \operatorname{diag}(\mathbf{pf}) R P'_B &= \left(\frac{\langle \overline{\mathbf{b}^{(B)}}, \cdot \rangle}{\langle \overline{\mathbf{b}^{(B)}}, \overline{\mathbf{l}^{(B)}} \rangle} \overline{\mathbf{l}^{(B)}} \right) \left[\operatorname{diag}(\mathbf{pf}) R \frac{\langle \mathbf{l}^{(B')}, \cdot \rangle}{\langle \mathbf{l}^{(B')}, \mathbf{b}^{(B')} \rangle} \mathbf{b}^{(B')} \right] \\ &= \frac{\langle \mathbf{b}^{(B)} \mathbf{pf}(R\mathbf{b}^{(B')}) \rangle}{\langle \overline{\mathbf{b}^{(B)}}, \overline{\mathbf{l}^{(B)}} \rangle} \frac{\langle \mathbf{l}^{(B')}, \cdot \rangle \, \overline{\mathbf{l}^{(B)}}}{\langle \mathbf{l}^{(B')}, \mathbf{b}^{(B')} \rangle} \end{aligned}$$

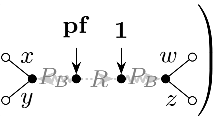
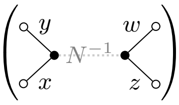
which we separate into the scalar factor

$$\frac{\langle \mathbf{b}^{(B)} \mathbf{pf}(R\mathbf{b}^{(B')}) \rangle \, \langle \mathbf{l}^{(B')}, \overline{\mathbf{l}^{(B)}} \rangle}{\langle \overline{\mathbf{b}^{(B)}}, \overline{\mathbf{l}^{(B)}} \rangle \, \langle \mathbf{l}^{(B')}, \mathbf{b}^{(B')} \rangle}$$

and the weighted edge

$$K = \frac{\langle \mathbf{l}^{(B')}, \cdot \rangle \, \overline{\mathbf{l}^{(B)}}}{\langle \mathbf{l}^{(B')}, \overline{\mathbf{l}^{(B)}} \rangle} \tag{4.82}$$

which scales like $|k_{ab}| \le cN^{-1}$. Thus we can write

Val () $= \dfrac{\langle \mathbf{b}^{(B)} \mathbf{pf}(R\mathbf{b}^{(B')}) \rangle \, \langle \mathbf{l}^{(B')}, \overline{\mathbf{l}^{(B)}} \rangle}{\langle \overline{\mathbf{b}^{(B)}}, \overline{\mathbf{l}^{(B)}} \rangle \, \langle \mathbf{l}^{(B')}, \mathbf{b}^{(B')} \rangle}$ Val () . (4.83)

Note that the B and B' operators are not completely independent: According to Fact 4.1 it follows that for an interaction edge $e = (u, v)$ associated with the matrix R the number of incoming G-edges in u is the same as the number of outgoing G-edges from v, and vice versa. Thus, according to (4.72), the B-operator at u comes with an S if and only if the B'-operator at v comes also with an S. Furthermore, if the B-operator comes with an T, then the B'-operator comes with an T^t, and vice versa. The distribution of the conjugation operators to B, B' in (4.72), however, can be arbitrary. We now use the fact that the scalar factor in (4.83) can be estimated by $|\sigma| + \rho + \eta/\rho$ (cf. Lemma 4.A.2). Summarising the above arguments, from (4.79)–(4.83), the proof of Proposition 4.4.12 is complete.

4.5 Cusp universality

The goal of this section is the proof of cusp universality in the sense of Theorem 4.2.3. Let H be the original Wigner-type random matrix with expectation $A := \mathbf{E}\, H$ and variance matrix $S = (s_{ij})$ with $s_{ij} := \mathbf{E}\,|h_{ij} - a_{ij}|^2$ and $T = (t_{ij})$ with $t_{ij} := \mathbf{E}(h_{ij} - a_{ij})^2$. We consider the Ornstein Uhlenbeck process $\{\, \widetilde{H}_t \mid t \geq 0 \,\}$ starting from $\widetilde{H}_0 = H$, i.e.

$$\mathrm{d}\widetilde{H}_t = -\frac{1}{2}(\widetilde{H}_t - A)\,\mathrm{d}t + \Sigma^{1/2}[\mathrm{d}B_t], \qquad \Sigma[R] := \mathbf{E}\,W\,\mathrm{Tr}(WR) \tag{4.84}$$

which preserves expectation and variance. In our setting of deformed Wigner-type matrices the covariance operator $\Sigma : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ is given by

$$\Sigma[R] := S \odot R + T \odot R^t.$$

The OU process effectively adds a small Gaussian component to $\widetilde{H}_t$ along the flow in the sense that $\widetilde{H}_t = A + e^{-t/2}(H - A) + \widetilde{U}_t$ in distribution with $\widetilde{U}_t$ being and independent centred Gaussian matrix with covariance $\mathbf{Cov}(\widetilde{U}) = (1 - e^{-t/2})\Sigma$. Due to the fullness Assumption (4.B) there exist small c, t_* such that $\widetilde{U}_t$ can be decomposed as $\widetilde{U}_t = \sqrt{ct}U + U_t'$ with $U \sim$ GUE and U_t' Gaussian and independent of U for $t \leq t_*$. Thus there exists a Wigner-type matrix H_t such that

$$\begin{aligned} &\widetilde{H}_t = H_t + \sqrt{ct}U, \qquad \mathcal{S}_t = \mathcal{S} - ct\mathcal{S}^{\mathrm{GUE}}, \\ &\mathbf{E}\,H_t = A, \qquad U \sim \mathrm{GUE}, \qquad \mathcal{S}^{\mathrm{GUE}}[R] := \langle R \rangle = \frac{1}{N}\,\mathrm{Tr}\,R \end{aligned} \tag{4.85}$$

with U independent of H_t. Note that we do not define H_t as a stochastic process and we will use the representation (4.85) only for one carefully chosen $t = N^{-1/2+\epsilon}$. We note that H_t satisfies the assumption of our local law from Theorem 4.2.5. It thus follows that $G_t := (H_t - z)^{-1}$ is well approximated by the solution $M_t = \mathrm{diag}(M_t)$ to the MDE

$$-M_t^{-1} = z - A + \mathcal{S}_t[M_t]. \qquad \rho_t(E) := \lim_{\eta \searrow 0} \frac{\Im\,\langle M_t(E + i\eta)\rangle}{\pi}.$$

In particular, by setting $t = 0$, M_0 well approximates the resolvent of the original matrix H and $\rho_0 = \rho$ is its self-consistent density. Note that the Dyson equation of $\widetilde{H}_t$ and hence its solution as well are independent of t, since they are entirely determined by the first and second moments of $\widetilde{H}_t$ that are the same A and S for any t. Thus the resolvent of $\widetilde{H}_t$ is well approximated by the same M_0 and the self-consistent density of $\widetilde{H}_t$ is given by $\rho_0 = \rho$ for any t. While H and $\widetilde{H}_t$ have identical self-consistent data, structurally they differ in a key point: $\widetilde{H}_t$ has a small Gaussian component. Thus the correlation kernel of the local eigenvalue statistics has a contour integral representation using a version of the Brézin-Hikami formulas, see Section 4.5.2.

The contour integration analysis requires a Gaussian component of size at least $ct \gg N^{-1/2}$ and a very precise description of the eigenvalues of H_t just above the scale of the eigenvalue spacing. This information will come from the optimal rigidity, Corollary 4.2.6, and the precise shape of the self-consistent density of states of H_t. The latter will be analysed in Section 4.5.1 where we describe the evolution of the density near the cusp under an additive GUE perturbation $\sqrt{s}U$. We need to construct H_t with a small gap carefully so that after a relatively long time $s = ct$ the matrix $H_t + \sqrt{ct}U$ develops a cusp exactly at the right location. In fact, we the process has two scales in the shifted variable $\nu = s - ct$ that indicates the time relative to the cusp formation. It turns out that the *locations* of the edges typically move linearly with ν, while the *length of the gap* itself scales like $(-\nu)_+^{3/2}$, i.e. it varies much slower and we need to fine-tune the evolution of both.

To understand this tuning process, we fix $t = N^{-1/2+\epsilon}$ and we consider the matrix flow $s \to H_t(s) := H_t + \sqrt{s}U$ for any $s \geq 0$ and not just for $s = ct$. It is well known that the corresponding self-consistent densities are given by the semicircular flow. Equivalently, these densities can be described by the free convolution of ρ_t with a scaled semicircular distribution ρ_{sc}. In short, the self-consistent density of $H_t(s)$ is given by $\rho_s^{\mathrm{fc}} := \rho_t \boxplus \sqrt{s}\rho_{\mathrm{sc}}$, where we omitted t from the notation ρ_s^{fc} since we consider t fixed. In particular we have $\rho_0^{\mathrm{fc}} = \rho_t$, the density of H_t and $\rho_{ct}^{\mathrm{fc}} = \rho$, the density of $\widetilde{H}_t = H_t + \sqrt{ct}U$ as well as that of H. Hence, as a preparation to the contour integration, in Section 4.5.1 we need to describe the cusp formation along the semicircular flow. Before going into details, we describe the strategy.

Since in the sequel the densities ρ_s^{fc} and their local minima and gaps will play an important role, we introduce the convention that properties of the original density ρ will always carry ρ as a superscript for the remainder of Section 4.5. In particular, the points $\mathfrak{c}, \mathfrak{e}_\pm, \mathfrak{m}$ and the gap size Δ from (4.4) and Theorem 4.2.3 will from now on be denoted by $\mathfrak{c}^\rho$, $\mathfrak{e}_\pm^\rho$, $\mathfrak{m}^\rho$ and Δ^ρ. In particular a superscript of ρ never denotes a power.

Proof strategy.

First we consider case (i) when ρ, the self-consistent density associated with H, has an exact cusp at the point $\mathfrak{c}^\rho \in \mathbb{R}$. Note that $\mathfrak{c}^\rho$ is also a cusp point of the self-consistent density of $\widetilde{H}_t$ for any t.

We set $t := N^{-1/2+\epsilon}$. Define the functions

$$\Delta(\nu) := (2\gamma)^2(\nu/3)^{3/2} \qquad \text{and} \qquad \rho^{\min}(\nu) := \gamma^2\sqrt{\nu}/\pi$$

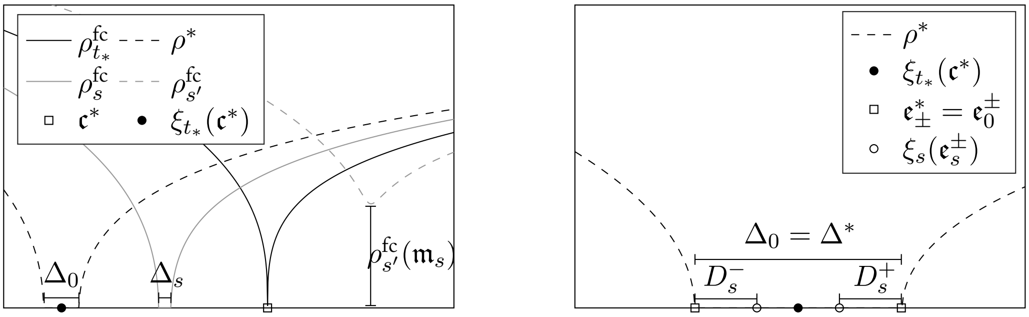
for any $\nu \geq 0$. For $s < ct$ denote the gap in the support of ρ_s^{fc} close to $\mathfrak{c}^\rho$ by $[\mathfrak{e}_s^-, \mathfrak{e}_s^+]$ and its length by $\Delta_s := \mathfrak{e}_s^+ - \mathfrak{e}_s^-$. In Section 4.5.1 we will prove that if ρ has an exact cusp in $\mathfrak{c}^\rho$ as in (4.4a), then ρ_s^{fc} has a gap of size $\Delta_s \approx \Delta(ct - s)$, and, in particular, $\rho_t = \rho_0^{\mathrm{fc}}$ has a gap of size $\Delta_0 \approx \Delta(ct) \sim t^{3/2}$, only depending on c, t and γ. The distance of $\mathfrak{c}^\rho$ from the gap is $\approx \mathrm{const} \cdot t$. This overall shift will be relatively easy to handle, but notice that it must be tracked very precisely since the gap changes much slower than its location. For $s > ct$ with $s - ct = \mathcal{O}(1)$ we will similarly prove that ρ_s^{fc} has no gap anymore close to $\mathfrak{c}^\rho$ but a unique local minimum in $\mathfrak{m}_s$ of size $\rho_s^{\mathrm{fc}}(\mathfrak{m}_s) \approx \rho^{\min}(s - ct)$.

Now we consider the case where ρ has no exact cusp but a small gap of size $\Delta^\rho > 0$. We parametrize this gap length via a parameter $t^\rho > 0$ defined by $\Delta^\rho = \Delta(t^\rho)$. It follows from the associativity (4.86b) of the free convolution that ρ_t has a gap of size $\Delta_0 \approx \Delta(ct + t^\rho)$.

Finally, the third case is where ρ has a local minimum of size $\rho(\mathfrak{m}^\rho)$. We parametrize it as $\rho(\mathfrak{m}^\rho) = \rho^{\min}(t^\rho)$ with $0 < t^\rho < ct$ then it follows that ρ_t has a gap of size $\Delta_0 \approx \Delta(ct - t^\rho)$.

Note that these conclusions follow purely from the considerations in Section 4.5.1 for exact cusps and the associativity of the free convolution. We note that in both almost cusp cases t^ρ should be interpreted as a time (or reverse time) to the cusp formation.

In the final part of the proof in Sections 4.5.2–4.5.3 we will write the correlation kernel of $H_t + \sqrt{ct}U$ as a contour integral purely in terms of the mesoscopic shape parameter γ and the gap size Δ_0 of the density ρ_t associated with H_t. If $\Delta_0 \approx \Delta(ct)$, then the gap closes after time $s \approx ct$ and we obtain a Pearcey kernel with parameter $\alpha = 0$. If $\Delta_0 \approx \Delta(ct + t^\rho)$ and $t^\rho \sim N^{-1/2}$, then the gap does not quite close at time $s = ct$ and we obtain a Pearcey kernel with $\alpha > 0$, while for $\Delta_0 \approx \Delta(ct - t^\rho)$ with $t^\rho \sim N^{-1/2}$ the gap after time $s = ct$ is transformed into a tiny local minimum and we obtain a Pearcey kernel with $\alpha < 0$. The



(a) Free semicircular flow around cusp.

(b) Location of ξ_s within the gap.

Figure 4.1: Figure 4.1(a) illustrates the evolution of ρ^{fc}_s along the semicircular flow at two times $0 < s < t_* < s'$ before and after the cusp. We recall that $\rho^* = \rho^{\text{fc}}_0$ and $\rho = \rho^{\text{fc}}_{t_*}$. Figure 4.1(b) shows the points $\xi_s(\mathfrak{e}^\pm_s)$ as well as their distances to the edges $\mathfrak{e}^\pm_0$.

precise value of α in terms of Δ^ρ and $\rho(\mathfrak{m}^\rho)$ are given in (4.6). Note that as an input to the contour integral analysis, in all three cases we use the local law only for H_t, i.e. in a situation when there is a small gap in the support of ρ_t, given by Δ_0 defined as above in each case.

4.5.1 Free convolution near the cusp

In this section we quantitatively investigate the free semi-circular flow before and after the formation of cusp. We first establish the exact rate at which a gap closes to form a cusp, and the rate at which the cusp is transformed into a non-zero local minimum. We now suppose that ρ^* is a general density with a small spectral gap $[\mathfrak{e}^*_-, \mathfrak{e}^*_+]$ whose Stieltjes transform m^* can be obtained from solving a Dyson equation. Let $\rho_{\text{sc}}(x) := \sqrt{(4-x^2)_+}/2\pi$ be the density of the semicircular distribution and let $s \geq 0$ be a time parameter. The free semicircular convolution ρ^{fc}_s of ρ^* with $\sqrt{s}\rho_{\text{sc}}$ is then defined implicitly via its Stieltjes transform

$$m^{\text{fc}}_s(z) = m^*(\xi_s(z)) = m^*(z + s m^{\text{fc}}_s(z)), \quad \xi_s(z) := z + s m^{\text{fc}}_s(z), \quad z, m^{\text{fc}}_s(z) \in \mathbb{H}. \tag{4.86a}$$

It follows directly from the definition that $s \mapsto m^{\text{fc}}_s$ is *associative* in the sense that

$$m^{\text{fc}}_{s+s'}(z) = m_s(z + s' m^{\text{fc}}_{s+s'}(z)), \qquad s, s' \geq 0. \tag{4.86b}$$

Figure 4.1(a) illustrates the quantities in the following lemma. We state the lemma for scDOSs from arbitrary data pairs $(A_*, \mathcal{S}_*)$ satisfying the conditions in [12], i.e.

$$\|A_*\| \leq C, \qquad c\langle R\rangle \leq \mathcal{S}_*[R] \leq C\langle R\rangle \tag{4.87}$$

for any self-adjoint $R = R^*$ and some constants $c, C > 0$.

Lemma 4.5.1. *Let ρ^* be the density of a Stieltjes transform $m^* = \langle M_*\rangle$ associated with some Dyson equation*

$$-1 = (z - A_* + \mathcal{S}_*[M_*])M_*,$$

with $(A_*, \mathcal{S}_*)$ satisfying (4.87). Then there exists a small constant c, depending only on the constants in Assumptions (4.87) such that the following statements hold true. Suppose that ρ^* has an *initial gap* $[\mathfrak{e}_-^*, \mathfrak{e}_+^*]$ *of size* $\Delta^* = \mathfrak{e}_+^* - \mathfrak{e}_-^* \leq c$. *Then there exists some critical time* $t_* \lesssim (\Delta^*)^{2/3}$ *such that* $m_{t_*}^{fc}$ *has exactly one exact cusp in some point* $\mathfrak{c}^*$ *with* $|\mathfrak{c}^* - \mathfrak{e}_\pm^*| \lesssim t_*$, *and that* $\rho_{t_*}^{fc}$ *is locally around* $\mathfrak{c}^*$ *given by* (4.4a) *for some* $\gamma > 0$. *Considering the time evolution* $[0, 2t_*] \ni s \mapsto m_s^{fc}$ *we then have the following asymptotics.*

(i) After the cusp. *For* $t_* < s \leq 2t_*$, ρ_s^{fc} *has a unique non-zero local minimum in some point* $\mathfrak{m}_s$ *such that*

$$\begin{aligned} \rho_s^{fc}(\mathfrak{m}_s) &= \frac{\sqrt{s - t_*}\gamma^2}{\pi}[1 + \mathcal{O}((s - t_*)^{1/2})], \\ \Big|\mathfrak{m}_s - \mathfrak{c}^* + (s - t_*)\Re m_s^{fc}(\mathfrak{m}_s)\Big| &\lesssim (s - t_*)^{3/2+1/4}. \end{aligned} \tag{4.88a}$$

Furthermore, $\mathfrak{m}_s$ *can approximately be found by solving a simple equation, namely there exists* $\widetilde{\mathfrak{m}}_s$ *such that*

$$\widetilde{\mathfrak{m}}_s - \mathfrak{c}^\rho + (s - t_*)\Re m_s^{fc}(\widetilde{\mathfrak{m}}_s) = 0, \quad |\mathfrak{m}_s - \widetilde{\mathfrak{m}}_s| \lesssim (s - t_*)^{3/2+1/4}, \quad \rho_s^{fc}(\widetilde{\mathfrak{m}}_s) \sim \sqrt{s - t_*}. \tag{4.88b}$$

(ii) Before the cusp. *For* $0 \leq s < t_*$, *the support of* ρ_s^{fc} *has a spectral gap* $[\mathfrak{e}_s^-, \mathfrak{e}_s^+]$ *of size* $\Delta_s := \mathfrak{e}_s^+ - \mathfrak{e}_s^-$ *near* $\mathfrak{c}^*$ *which satisfies*

$$\Delta_s = (2\gamma)^2 \Big(\frac{t_* - s}{3}\Big)^{3/2}[1 + \mathcal{O}((t_* - s)^{1/3})]. \tag{4.88c}$$

In particular we find that the initial gap $\Delta^* = \Delta_0$ *is related to* t_* *via*

$$\Delta^* = (2\gamma)^2(t_*/3)^{3/2}[1 + \mathcal{O}((t_* - s)^{1/3})].$$

Proof. Within the proof of the lemma we rely on the extensive shape analysis from [12]. We are doing so not only for the density $\rho^* = \rho_0^{\mathrm{fc}}$ and its Stieltjes transform, but also for ρ_s^{fc} and its Stieltjes transform m_s^{fc} for $0 \leq s \leq 2t_*$. The results from [12] also apply here since $m_s^{\mathrm{fc}}(z) = \langle M_*(\xi_s(z))\rangle$ can also be realized as the solution

$$\begin{aligned} -M_*(\xi_s(z))^{-1} &= z + s\,\langle M_*(\xi_s(z))\rangle - A_* + \mathcal{S}_*[M_*(\xi_s(z))] \\ &= z - A_* + (\mathcal{S}_* + s\mathcal{S}^{\mathrm{GUE}})[M_*(\xi_s(z))] \end{aligned}$$

to the Dyson equation with perturbed self-energy $\mathcal{S}_* + s\mathcal{S}^{\mathrm{GUE}}$. Since $t_* \lesssim 1$ it follows that the shape analysis from [12] also applies to ρ_s^{fc} for any $s \in [0, 2t_*]$.

We begin with part (i). Set $\nu := s - t_*$, then for $0 \leq \nu \leq t_*$ we want to find x_ν such that $\Im m_s^{\mathrm{fc}}$ has a local minimum in $\mathfrak{m}_s := \mathfrak{c}^* + x_\nu$ near $\mathfrak{c}^*$, i.e.

$$x_\nu := \arg\min_x \Im m_s^{\mathrm{fc}}(\mathfrak{c}^* + x), \qquad |x_\nu| \lesssim \nu.$$

First we show that x_ν with these properties exists and is unique by using the extensive shape analysis in [12]. Uniqueness directly follows from [12, Theorem 7.2(ii)]. For the existence, we set

$$a_\nu(x) := \Im m_{\mathrm{fc}}^s(\mathfrak{c}^* + x), \quad b_\nu(x) := \Re m_s^{\mathrm{fc}}(\mathfrak{c}^* + x), \quad a_\nu := a_\nu(x_\nu), \quad b_\nu := b_\nu(x_\nu).$$

Set $\delta := K\nu$ with a large constant K. Since $a_0(x) = \Im m_{t_*}(\mathfrak{c}^* + x) \sim |x|^{1/3}$, we have $a_0(\pm\delta) \sim \delta^{1/3}$ and $a_0(0) = 0$. Recall from [12, Proposition 10.1(a)] that the map $s \mapsto m_s^{\text{fc}}$ is $1/3$-Hölder continuous. It then follows that $a_\nu(\pm\delta) \sim \delta^{1/3} + \mathcal{O}\left(\nu^{1/3}\right)$, while $a_\nu(0) \lesssim \nu^{1/3}$. Thus a_ν necessarily has a local minimum in $(-\delta, \delta)$ if K is sufficiently large. This shows the existence of a local minimum with $|x_\nu| \lesssim K\nu \sim \nu$.

We now study the function $f_\nu(x) = x + \nu b_\nu(x)$ in a small neighbourhood around 0. From [12, Eqs. (7.62), (5.43)–(5.45)] it follows that

$$\begin{aligned} b_\nu'(x) &= \Re \frac{c_1(x) + \mathcal{O}\left(a_\nu(x)\right)}{-\mathrm{i}c_2(x)a_\nu(x) + a_\nu(x)^2 + \mathcal{O}\left(a_\nu(x)^3\right)} + \mathcal{O}\left(1\right) \\ &= \frac{c_1(x)}{c_2(x)^2 + a_\nu(x)^2} + \mathcal{O}\left(\frac{1}{c_2(x) + a_\nu(x)}\right) \end{aligned} \tag{4.89}$$

whenever $a_\nu(x) \ll 1$, with appropriate real functions[2] $c_1(x) \sim 1$ and $c_2(x) \geq 0$. Moreover, $|c_2(0)| \ll 1$ since $\mathfrak{c}^*$ is an almost cusp point for m_s^{fc} for any $s \in [0, 2t_*]$. Thus it follows that $b_\nu'(x) > 0$ whenever $a_\nu(x) + c_2(x) \ll 1$. Due to the $1/3$-Hölder continuity[3] of both $a_\nu(x)$ and $c_2(x)$ and $a_\nu(0) + |c_2(0)| \ll 1$, it follows that $b_\nu'(x) > 0$ whenever $|x| \ll 1$. We can thus conclude that f_ν satisfies $f_\nu' \geq 1$ in some $\mathcal{O}(1)$-neighbourhood of 0. As $|f_\nu(0)| \lesssim \nu$ we can conclude that there exists a root $\widetilde{x}_\nu$, $f_\nu(\widetilde{x}_\nu) = 0$ of size $|\widetilde{x}_\nu| \lesssim \nu$. With $\widetilde{\mathfrak{m}}_s := \mathfrak{c}^* + \widetilde{x}_\nu$ we have thus shown the first equality in (4.88b).

Using (4.4a), we now expand the defining equation

$$a_\nu(x) = \Im m_{t_*}^{\text{fc}}(\mathfrak{c}^* + x + \nu b_\nu(x) + \mathrm{i}\nu a_\nu(x))$$

for the free convolution in the regime for those x sufficiently close to $\widetilde{x}_\nu$ such that $|x + \nu b_\nu(x)| \lesssim \nu a_\nu(x)$ to find

$$\begin{aligned} a_\nu(x) &= \frac{\sqrt{3}\gamma^{4/3}}{2\pi}\nu a_\nu(x) \int_{\mathbb{R}} \frac{|\lambda|^{1/3} + \mathcal{O}\left(|\lambda|^{2/3}\right)}{(\lambda - x - \nu b_\nu(x))^2 + (\nu a_\nu(x))^2} \,\mathrm{d}\lambda \\ &= \frac{\sqrt{3}\gamma^{4/3}}{2\pi} \int_{\mathbb{R}} \frac{(\nu a_\nu(x))^{1/3}|\lambda|^{1/3}}{(\lambda - [x + \nu b_\nu(x)]/\nu a_\nu(x))^2 + 1} \,\mathrm{d}\lambda + \mathcal{O}\left((\nu a_\nu(x))^{2/3}\right) \\ &= (\nu a_\nu(x))^{1/3}\gamma^{4/3}\left[1 + \frac{1}{9}\left(\frac{x + \nu b_\nu(x)}{\nu a_\nu(x)}\right)^2 + \mathcal{O}\left(\left(\frac{x + \nu b_\nu(x)}{\nu a_\nu(x)}\right)^4 + (\nu a_\nu(x))^{1/3}\right)\right], \end{aligned}$$

i.e.

$$a_\nu(x) = \nu^{1/2}\gamma^2\left[1 + \frac{1}{9}\left(\frac{x + \nu b_\nu(x)}{\nu a_\nu(x)}\right)^2 + \mathcal{O}\left(\left(\frac{x + \nu b_\nu(x)}{\nu a_\nu(x)}\right)^4 + (\nu a_\nu(x))^{1/3}\right)\right]^{3/2}. \tag{4.90}$$

Note that (4.90) implies that $\nu a_\nu(\widetilde{x}_\nu) \sim \nu^{3/2}$, i.e. the last claim in (4.88b). We now pick some large K and note that from (4.90) it follows that $a_\nu(\widetilde{x}_\nu \pm K\nu^{7/4}) > a_\nu(\widetilde{x}_\nu)$. Thus the interval $[\widetilde{x}_\nu - K\nu^{7/4}, \widetilde{x}_\nu + K\nu^{7/4}]$ contains a local minimum of $a_\nu(x)$, but by the uniqueness this must then be x_ν. We thus have $|x_\nu - \widetilde{x}_\nu| \leq K\nu^{7/4}$, proving the second

[2] We have $c_1 = \pi/\psi$, $c_2 = 2\sigma/\psi$ with the notations ψ, σ in [12], where $\psi \sim 1$ and $|\sigma| \ll 1$ near the almost cusp, but we refrain from using these letters in the present context to avoid confusions.

[3] See [12, Lemma 5.5] for the $1/3$-Hölder continuity of quantities ψ, σ in the definition of c_2.

claim in (4.88b). By 1/3-Hölder continuity of $a_\nu(x)$ and by $a_\nu(\widetilde{x}_\nu) \sim \nu^{1/2}$ from (4.90), we conclude that $a_\nu = a_\nu(x_\nu) \sim \nu^{1/2}$ as well. Using that $\widetilde{x}_\nu + \nu b_\nu(\widetilde{x}_\nu) = 0$ and $b_\nu' \lesssim 1/\nu$ from (4.89) and $a_\nu(x) \gtrsim \sqrt{\nu}$, we conclude that $|x_\nu + \nu b_\nu(x_\nu)| \lesssim \nu^{7/4}$, i.e. the second claim in (4.88a). Plugging this information back into (4.90), we thus find $a_\nu = \gamma^2\sqrt{\nu}(1+\mathcal{O}\left(\nu^{1/2}\right))$ and have also proven the first claim in (4.88a).

We now turn to part (ii). It follows from the analysis in [12] that ρ_s^{fc} exhibits either a small gap, a cusp or a small local minimum close to $\mathfrak{c}^*$. It follows from (i) that a cusp is transformed into a local minimum, and a local minimum cannot be transformed into a cusp along the semicircular flow. Therefore it follows that the support of ρ_s^{fc} has a gap of size $\Delta_s = \mathfrak{e}_s^+ - \mathfrak{e}_s^-$ between the edges $\mathfrak{e}_s^\pm$. Evidently $\mathfrak{e}_{t_*}^- = \mathfrak{e}_{t_*}^+ = \mathfrak{c}^*$, $\mathfrak{e}_0^+ - \mathfrak{e}_0^- = \Delta_0$, $\mathfrak{e}_0^\pm = \mathfrak{e}_\pm^*$ and for $s > 0$ we differentiate (4.86a) to obtain

$$\frac{(m_s^{\mathrm{fc}})'(z)}{1+s(m_s^{\mathrm{fc}})'(z)} = m_*'(z+sm_s^{\mathrm{fc}}(z)) \quad \text{and conclude} \quad m_*'(\xi_s(\mathfrak{e}_s^\pm)) = 1/s \tag{4.91}$$

by considering the $z \to \mathfrak{e}_s^\pm$ limit and the fact that ρ_s^{fc} has a square root at edge (for $s < t_*$) hence $(m_s^{\mathrm{fc}})'$ blows up at this point. Denoting the $\mathrm{d}/\,\mathrm{d}s$ derivative by dot, from

$$\frac{\mathrm{d}}{\mathrm{d}s} m_s^{\mathrm{fc}}(\mathfrak{e}_s^\pm) = m_*'(\xi_s(\mathfrak{e}_s^\pm))\left(\dot{\mathfrak{e}}_s^\pm + m_s^{\mathrm{fc}}(\mathfrak{e}_s^\pm) + s\frac{\mathrm{d}}{\mathrm{d}s}m_s^{\mathrm{fc}}(\mathfrak{e}_s^\pm)\right) = \frac{\dot{\mathfrak{e}}_s^\pm + m_s^{\mathrm{fc}}(\mathfrak{e}_s^\pm)}{s} + \frac{\mathrm{d}}{\mathrm{d}s}m_s^{\mathrm{fc}}(\mathfrak{e}_s^\pm)$$

we can thus conclude that $\dot{\mathfrak{e}}_s^\pm = -m_s^{\mathrm{fc}}(\mathfrak{e}_s^\pm)$. This implies that the gap as a whole moves with linear speed (for non-zero $m_s^{\mathrm{fc}}(\mathfrak{e}_s^\pm)$), and, in particular, the distance of the gap of ρ^* to $\mathfrak{c}^*$ is an order of magnitude larger than the size of the gap. It follows that the size $\Delta_s := \mathfrak{e}_s^+ - \mathfrak{e}_s^-$ of the gap of ρ_s^{fc} satisfies

$$\begin{aligned}\dot{\Delta}_s = m_s^{\mathrm{fc}}(\mathfrak{e}_s^-) - m_s^{\mathrm{fc}}(\mathfrak{e}_s^+) &= \int_{\mathbb{R}} \Big[\frac{1}{x-\mathfrak{e}_s^-} - \frac{1}{x-\mathfrak{e}_s^+}\Big]\rho_s^{\mathrm{fc}}(x)\,\mathrm{d}x \\ &= -\Delta_s \int_{\mathbb{R}} \frac{\rho_s^{\mathrm{fc}}(x)}{(x-\mathfrak{e}_s^-)(x-\mathfrak{e}_s^+)}\,\mathrm{d}x.\end{aligned}$$

We now use the precise shape of ρ_s^{fc} close to $\mathfrak{e}_s^\pm$ according to (4.4b) which is given by

$$\begin{aligned}\rho_s^{\mathrm{fc}}(\mathfrak{e}_s^\pm \pm x) = {}& \frac{\sqrt{3}(2\gamma)^{4/3}\Delta_s^{1/3}}{2\pi} \\ & \times \Big((1+\mathcal{O}((t_*-t)^{1/3}))\Psi_{\mathrm{edge}}(x/\Delta_s) + \mathcal{O}\left(\Delta_s^{1/3}\Psi_{\mathrm{edge}}^2(x/\Delta_s)\right)\Big),\end{aligned} \tag{4.92}$$

where Ψ_{edge} defined in (4.4c) exhibits the limiting behaviour $\lim_{\Delta\to 0}\Delta^{1/3}\Psi_{\mathrm{edge}}(x/\Delta) = |x|^{1/3}/2^{4/3}$. Using (4.92), we compute

$$\begin{aligned}\dot{\Delta}_s &= -(1+\mathcal{O}((t_*-s)^{1/3}))\frac{\sqrt{3}(2\gamma)^{4/3}\Delta_s^{1/3}}{\pi}\int_0^\infty \frac{\Psi_{\mathrm{edge}}(x)}{x(1+x)}\,\mathrm{d}x \\ &= -\gamma^{4/3}(2\Delta_s)^{1/3}\Big[1+\mathcal{O}((t_*-s)^{1/3}+\Delta_s^{1/3})\Big],\end{aligned} \tag{4.93}$$

where the $(1+\mathcal{O}((t_*-s)^{1/3}))$ factor in (4.92) encapsulates two error terms; both are due to the fact that the shape factor γ_s of ρ_s^{fc} from (4.4b) is not exactly the same as γ, i.e. the one for $s = t_*$. To track this error in γ we go back to [12]. First, $|\sigma|$ in [12, Eq. (7.5a)] is of size

$(t_*-s)^{1/3}$ by the fact that σ vanishes at $s=t_*$ and is $1/3$-Hölder continuous according to [12, Lemma 10.5]. Secondly, according to [12, Lemma 10.5] the shape factor Γ (which is directly related to γ in the present context) is also $1/3$-Hölder continuous and therefore we know that the shape factors of ρ^* at $\mathfrak{e}_0^\pm$ are at most multiplicatively perturbed by a factor of $(1+\mathcal{O}((t_*-s)^{1/3}))$. By solving the differential equation (4.93) with the initial condition $\Delta_{t_*}=0$, the claim (4.88c) follows. □

Besides the asymptotic expansion for gap size and local minimum we also require some quantitative control on the location of $\xi_{t_*}(\mathfrak{c}^*)$, as defined in (4.86a), and some slight perturbations thereof within the spectral gap $[\mathfrak{e}_-^*,\mathfrak{e}_+^*]$ of ρ^*. We remark the the point $\xi^* := \xi_{t_*}(\mathfrak{c}^*)$ plays a critical role for the contour integration in Section 4.5.2 since it will be the critical point of the phase function. From (4.88c) we recall that the gap size scales as $t_*^{3/2}$ which makes it natural to compare distances on that scale. In the regime where $t' \ll t_*$ all of the following estimates thus identify points very close to the centre of the initial gap.

Lemma 4.5.2. *Suppose that we are in the setting of Lemma 4.5.1. We then find that $\xi_{t_*}(\mathfrak{c}^*)$ is very close to the centre of $[\mathfrak{e}_-^*,\mathfrak{e}_+^*]$ in the sense that*

$$\Big|\xi_{t_*}(\mathfrak{c}^*)-\frac{\mathfrak{e}_+^*+\mathfrak{e}_-^*}{2}\Big| \lesssim t_*^{3/2+1/3}. \tag{4.94a}$$

Furthermore, for $0\le t'\le t_*$ we have that

$$\begin{aligned}\Big|\xi_{t_*-t'}\Big(\frac{\mathfrak{e}_{t_*-t'}^+ + \mathfrak{e}_{t_*-t'}^-}{2}\Big)-\frac{\mathfrak{e}_+^*+\mathfrak{e}_-^*}{2}\Big| &\lesssim t_*^{3/2+1/9},\\ \Big|\xi_{t_*+t'}(\mathfrak{m}_{t_*+t'})-\frac{\mathfrak{e}_+^*+\mathfrak{e}_-^*}{2}\Big| &\lesssim t_*^{3/2}(t_*^{1/12}+(t'/t_*)^{1/2}).\end{aligned} \tag{4.94b}$$

Proof. We begin with proving (4.94a). For $s<t_*$ we denote the distance of $\xi_s(\mathfrak{e}_s^\pm)$ to the edges $\mathfrak{e}_0^\pm$ by $D_s^\pm := \pm(\mathfrak{e}_0^\pm - \xi_s(\mathfrak{e}_s^\pm))$, cf. Figure 4.1(b). We have, by differentiating $m_*'(\xi_s(\mathfrak{e}_s^\pm))=1/s$ from (4.91) that

$$\dot{D}_s^\pm = \mp\frac{\mathrm{d}}{\mathrm{d}s}\xi_s(\mathfrak{e}_s^\pm), \qquad -\frac{1}{s^2}=m_*''(\xi_s(\mathfrak{e}_s^\pm))\frac{\mathrm{d}}{\mathrm{d}s}\xi_s(\mathfrak{e}_s^\pm) \tag{4.95}$$

and by differentiating (4.86a),

$$(m_s^{\mathrm{fc}})'=m_*'(\xi_s)\xi_s', \quad \xi_s'(m_s^{\mathrm{fc}})''=m_*''(\xi_s)(\xi_s')^3+(m_s^{\mathrm{fc}})'\xi_s'', \quad m_*''(\xi_s)=\frac{(m_s^{\mathrm{fc}})''}{(1+s(m_s^{\mathrm{fc}})')^3}.$$

We now consider $z=\mathfrak{e}_s^\pm+\mathrm{i}\eta$ with $\eta\to 0$ and compute from (4.92), for any $s<t_*$,

$$\begin{aligned}\lim_{\eta\searrow 0}\sqrt{\eta}(m_s^{\mathrm{fc}})'(z) &= \lim_{\eta\searrow 0}\sqrt{\eta}\int_{\mathbb{R}}\frac{\rho_s^{\mathrm{fc}}(x)}{(x-z)^2}\,\mathrm{d}x = \lim_{\eta\searrow 0}\frac{\sqrt{3\eta}(2\gamma)^{4/3}\Delta_s^{1/3}}{2\pi}\int_0^\infty\frac{\Psi_{\mathrm{edge}}(x/\Delta_s)}{(x-\mathrm{i}\eta)^2}\,\mathrm{d}x\\ &= \frac{(2\gamma)^{4/3}}{2\sqrt{3}\Delta_s^{1/6}\pi}\int_0^\infty\frac{x^{1/2}}{(x-\mathrm{i})^2}\,\mathrm{d}x = \frac{(2\gamma)^{4/3}\sqrt{\mathrm{i}}}{4\sqrt{3}\Delta_s^{1/6}}\end{aligned}$$

and

$$\begin{aligned}\lim_{\eta\searrow 0}\eta^{3/2}(m_s^{\mathrm{fc}})''(z) &= \lim_{\eta\searrow 0}\eta^{3/2}2\int_{\mathbb{R}}\frac{\rho_s^{\mathrm{fc}}(x)}{(x-z)^3}\,\mathrm{d}x\\ s &= \lim_{\eta\searrow 0}\frac{\sqrt{3}\eta^{3/2}(2\gamma)^{4/3}\Delta_s^{1/3}}{\pi}\int_0^\infty\frac{\Psi_{\mathrm{edge}}(x/\Delta_s)}{(x-\mathrm{i}\eta)^3}\,\mathrm{d}x\\ &= \frac{(2\gamma)^{4/3}}{\sqrt{3}\Delta_s^{1/6}\pi}\int_0^\infty\frac{x^{1/2}}{(x-\mathrm{i})^3}\,\mathrm{d}x = \frac{(2\gamma)^{4/3}\mathrm{i}^{3/2}}{8\sqrt{3}\Delta_s^{1/6}}.\end{aligned}$$

Here we used that fact that the error terms in (4.92) become irrelevant in the $\eta\to 0$ limit. We conclude, together with (4.95), that

$$m_*''(\xi_s(\mathfrak{e}_s^\pm)) = \pm\frac{3(2\Delta_s)^{1/3}}{s^3\gamma^{8/3}},\quad \dot{D}_s^\pm = \pm(s^2m_*''(\xi_s(\mathfrak{e}_s^\pm)))^{-1} = \frac{s\gamma^2}{2\sqrt{3}\sqrt{t_*-s}}[1+\mathcal{O}(t_*^{1/3})].$$

Since $D_0^- = D_0^+ = 0$ and $\dot{D}_s^- \approx \dot{D}_s^+$ it follows that, to leading order, $D_s^+ \approx D_s^-$ and more precisely

$$D_s^\pm = \gamma^2\frac{2t_*^{3/2} - s\sqrt{t_*-s} - 2t_*\sqrt{t_*-s}}{3^{3/2}}[1+\mathcal{O}(t_*^{1/3})].$$

In particular it follows that $\left|\mathfrak{e}_0^\pm - \xi_{t_*}(\mathfrak{c}^*)\right| = [1+\mathcal{O}(t_*)^{1/3}]2\gamma^2t_*^{3/2}/3^{3/2}$. Together with the $s=0$ case from (4.88c) we thus find

$$\left|\xi_{t_*}(\mathfrak{c}^*) - \frac{\mathfrak{e}_+^* + \mathfrak{e}_-^*}{2}\right| \lesssim t_*^{3/2+1/3} = t_*^{11/6},$$

proving (4.94a).

We now turn to the proof of (4.94b) where we treat the small gap and small non-zero minimum separately. We start with the first inequality. We observe that (4.94a) in the setting where (ρ^*, t_*) are replaced by $(\rho_{t_*-t'}^{\mathrm{fc}}, t')$ implies

$$\left|\mathfrak{c}^* + t'm_{t_*}^{\mathrm{fc}}(\mathfrak{c}^*) - \frac{\mathfrak{e}_{t_*-t'}^+ + \mathfrak{e}_{t_*-t'}^-}{2}\right| \le (t')^{11/6}. \tag{4.96}$$

Furthermore, we infer from the definition of ξ and the associativity (4.86b) of the free convolution that

$$\xi_{t_*-t'}\Big(\mathfrak{c}^* + t'm_{t_*}^{\mathrm{fc}}(\mathfrak{c}^*)\Big) = \mathfrak{c}^* + t'm_{t_*}^{\mathrm{fc}}(\mathfrak{c}^*) + (t_*-t')m_{t_*-t'}^{\mathrm{fc}}\Big(\mathfrak{c}^* + t'm_{t_*}^{\mathrm{fc}}(\mathfrak{c}^*)\Big) = \xi_{t_*}(\mathfrak{c}^*)$$

and can therefore estimate

$$\begin{aligned}\left|\xi_{t_*-t'}\Big(\frac{\mathfrak{e}_{t_*-t'}^+ + \mathfrak{e}_{t_*-t'}^-}{2}\Big) - \xi_{t_*}(\mathfrak{c}^*)\right| &= \left|\xi_{t_*-t'}\Big(\frac{\mathfrak{e}_{t_*-t'}^+ + \mathfrak{e}_{t_*-t'}^-}{2}\Big) - \xi_{t_*-t'}\Big(\mathfrak{c}^* + t'm_{t_*}^{\mathrm{fc}}(\mathfrak{c}^*)\Big)\right|\\ &\lesssim (t')^{11/6} + t_*(t')^{11/18} \lesssim t_*^{29/18},\end{aligned}$$

just as claimed. In the last step we used (4.96) and the fact that

$$|\xi_s(a) - \xi_s(b)| \lesssim |a-b| + s\,|a-b|^{1/3}, \tag{4.97}$$

which directly follows from the definition of ξ and the $1/3$-Hölder continuity of m_s^{fc}.

Finally, we address the second inequality in (4.94b) and appeal to Lemma 4.5.1(i) to establish the existence of $\widetilde{\mathfrak{m}}_{t_*+t'}$ such that

$$\mathfrak{c}^* - \widetilde{\mathfrak{m}}_{t_*+t'} = t'\Re m^{\mathrm{fc}}_{t_*+t'}(\widetilde{\mathfrak{m}}_{t_*+t'}). \tag{4.98}$$

It thus follows from (4.88b) that $|\widetilde{\mathfrak{m}}_{t_*+t'} - \mathfrak{m}_{t_*+t'}| \lesssim (t')^{7/4}$ and therefore from (4.97) that

$$|\xi_{t_*+t'}(\widetilde{\mathfrak{m}}_{t_*+t'}) - \xi_{t_*+t'}(\mathfrak{m}_{t_*+t'})| \lesssim (t')^{7/4} + t_*(t')^{7/12} \lesssim t_*^{19/12}.$$

Using (4.98) twice, as well as the associativity (4.86b) of the free convolution and $\Im m^{\mathrm{fc}}_{t_*}(\mathfrak{c}^*) = 0$ we then further compute

$$\begin{aligned} \xi_{t_*+t'}(\widetilde{\mathfrak{m}}_{t_*+t'}) - \xi_{t_*}(\mathfrak{c}^*) &= \widetilde{\mathfrak{m}}_{t_*+t'} + (t_*+t')m^{\mathrm{fc}}_{t_*+t'}(\widetilde{\mathfrak{m}}_{t_*+t'}) - \mathfrak{c}^* - t_* m^{\mathrm{fc}}_{t_*}(\mathfrak{c}^*) \\ &= t_*\Re\Big[m^{\mathrm{fc}}_{t_*}(\mathfrak{c}^* + \mathrm{i}t'\Im m^{\mathrm{fc}}_{t_*+t'}(\widetilde{\mathfrak{m}}_{t_*+t'})) - m^{\mathrm{fc}}_{t_*}(\mathfrak{c}^*)\Big] + \mathrm{i}(t_*+t')\Im m^{\mathrm{fc}}_{t_*+t'}(\widetilde{\mathfrak{m}}_{t_*+t'}). \end{aligned} \tag{4.99}$$

By Hölder continuity we can, together with (4.94a) and $\Im m_{t_*+t'}(\widetilde{\mathfrak{m}}_{t_*+t'}) \sim (t')^{1/2}$ from (4.88b), conclude that

$$\begin{aligned} &\Big|\xi_{t_*+t'}(\mathfrak{m}_{t_*+t'}) - \frac{\mathfrak{e}^*_+ + \mathfrak{e}^*_-}{2}\Big| \\ &\lesssim |\xi_{t_*+t'}(\mathfrak{m}_{t_*+t'}) - \xi_{t_*+t'}(\widetilde{\mathfrak{m}}_{t_*+t'})| + |\xi_{t_*+t'}(\widetilde{\mathfrak{m}}_{t_*+t'}) - \xi_{t_*}(\mathfrak{c}^*)| + \Big|\xi_{t_*}(\mathfrak{c}^*) - \frac{\mathfrak{e}^*_+ + \mathfrak{e}^*_-}{2}\Big| \\ &\lesssim [t_*^{7/4} + t_*(t_*^{7/4})^{1/3}] + t_*(t')^{1/2} + t_*^{11/6} \lesssim t_*^{3/2}(t_*^{1/12} + (t'/t_*)^{1/2}). \end{aligned}$$

In the first term we used (4.97) and the second estimate of (4.88b). In the second term we used (4.99) together with $\Im m_{t_*+t'}(\widetilde{\mathfrak{m}}_{t_*+t'}) \sim (t')^{1/2}$ from (4.88b) and $1/3$-Hölder continuity of $m^{\mathrm{fc}}_{t_*}$. Finally, the last term was already estimated in the exact cusp case, i.e. in (4.94a). □

4.5.2 Correlation kernel as contour integral

We denote the eigenvalues of H_t by $\lambda_1, \dots, \lambda_N$. Following the work of Brézin and Hikami (see e.g. [50, Eq. (2.14)] or [74, Eq. (3.13)] for the precise version used in the present context) the correlation kernel of $\widetilde{H}_t = H_t + \sqrt{ct}U$ can be written as

$$\begin{aligned} \widehat{K}^t_N(u,v) &:= \frac{N}{(2\pi\mathrm{i})^2 ct} \\ &\times \int_\Upsilon \mathrm{d}z \int_\Gamma \mathrm{d}w \frac{\exp\left(N\left[w^2 - 2vw + v^2 - z^2 + 2zu - u^2\right]/2ct\right)}{w-z} \prod_i \frac{w-\lambda_i}{z-\lambda_i}, \end{aligned}$$

where Υ is any contour around all λ_i, and Γ is any vertical line not intersecting Υ. With this notation, the k-point correlation function of the eigenvalues of $\widetilde{H}_t$ is given by

$$p_k^{(N)}(x_1, \dots, x_k) = \det\Big(\frac{1}{N}\widehat{K}^t_N(x_i, x_j)\Big)_{i,j\in[k]}.$$

Due to the determinantal structure we can freely conjugate K_N with $v \mapsto e^{N(\xi v - v^2/2)/ct}$ for $\xi := \xi_{ct}(\mathfrak{b})$ to redefine the correlation kernel as

$$\begin{aligned} K^t_N(u,v) &:= \frac{N}{(2\pi\mathrm{i})^2 ct} \\ &\times \int_\Upsilon \mathrm{d}z \int_\Gamma \mathrm{d}w \frac{\exp\left(N\left[w^2 - 2v(w-\xi) - z^2 + 2u(z-\xi)\right]/2ct\right)}{w-z} \prod_i \frac{w-\lambda_i}{z-\lambda_i}. \end{aligned}$$

This redefinition K_N^t does not agree point-wise with the previous definition $\widehat{K_N^t}$, but gives rise to the same determinant, and in particular to the same k-point correlation function. Here $\mathfrak{b}$ is the base point chosen in Theorem 4.2.3. The central result concerning the correlation kernel is the following proposition.

Proposition 4.5.3. *Under the assumptions of Theorem 4.2.3, the rescaled correlation kernel*

$$\widetilde{K_N^t}(x,y) := \frac{1}{N^{3/4}\gamma} K_N^t\left(\mathfrak{b} + \frac{x}{N^{3/4}\gamma}, \mathfrak{b} + \frac{y}{N^{3/4}\gamma}\right) \tag{4.100}$$

around the base point $\mathfrak{b}$ *chosen in* (4.6) *converges uniformly to the Pearcey kernel from* (4.5) *in the sense that*

$$\left|\widetilde{K_N^t}(x,y) - K_\alpha(x,y)\right| \le CN^{-c}$$

for $x, y \in [-R, R]$. *Here* R *is an arbitrary large threshold,* $c > 0$ *is some universal constant,* $C > 0$ *is a constant depending only on the model parameters and* R*, and* α *is chosen according to* (4.6).

Proof. We now split the contour Υ into two parts, one encircling all eigenvalues λ_i to the left of $\xi = \mathfrak{b} + ct\,\langle M(\mathfrak{b})\rangle$, and the other one encircling all eigenvalues λ_i to the right of ξ, which does not change the value of K_N^t. We then move the vertical Γ contour so that it crosses the real axis in ξ. This does also not change the value K_N^t as the only pole is the one in z for which the residue reads

$$\frac{N}{(2\pi \mathrm{i})^2 ct}\int_\Upsilon \mathrm{d}z \exp\left(\frac{N}{ct\gamma}(u-v)(z-\xi)\right) = 0.$$

We now perform a linear change of variables $z \mapsto \xi + \Delta_0 z$, $w \mapsto \xi + \Delta_0 w$ in (4.100) to transform the contours Υ, Γ into contours

$$\widehat{\Gamma} := (\Gamma - \xi)/\Delta_0, \qquad \widehat{\Upsilon} := (\Upsilon - \xi)/\Delta_0 \tag{4.101}$$

to obtain

$$\widetilde{K_N^t}(x,y) = \frac{N^{1/4}\Delta_0}{(2\pi \mathrm{i})^2 ct\gamma}\int_{\widehat{\Upsilon}} \mathrm{d}z \int_{\widehat{\Gamma}} \mathrm{d}w \frac{\exp\left(\Delta_0 N^{1/4}\frac{xz-yw}{ct\gamma} + N\Delta_0^2 \frac{\widetilde{f}(w)-\widetilde{f}(z)}{ct}\right)}{w-z}, \tag{4.102}$$

where

$$\widetilde{f}(z) := \frac{z^2}{2} - \frac{ct}{\Delta_0^2}\int_\xi^{\xi+\Delta_0 z} \langle G_t(u) - M_t(\xi)\rangle \,\mathrm{d}u.$$

Here $\Delta_0 := \mathfrak{e}_0^+ - \mathfrak{e}_0^-$ indicates the length of the gap $[\mathfrak{e}_0^-, \mathfrak{e}_0^+]$ in the support of ρ_t. From Lemma 4.5.1 with $\rho^* = \rho_t$ and $t_* = ct$ we infer $\Delta_0 \sim t^{3/2} \sim N^{-3/4+3\epsilon/2}$. In order to obtain (4.102) we used the relation $\xi - \mathfrak{b} = ctm_{ct}^{\mathrm{fc}}(\mathfrak{b}) = ct\,\langle M_t(\mathfrak{b} + ctm_{ct}^{\mathrm{fc}}(\mathfrak{b}))\rangle = ct\,\langle M_t(\xi)\rangle$.

We begin by analysing the deterministic variant of $\widetilde{f}(z)$,

$$f(z) := \frac{z^2}{2} - \frac{ct}{\Delta_0^2}\int_\xi^{\xi+\Delta_0 z} \langle M_t(u) - M_t(\xi)\rangle \,\mathrm{d}u.$$

We separately analyse the large- and small-scale behaviour of $f(z)$. On the one hand, using the $1/3$-Hölder continuity of $u \mapsto \langle M_t(u)\rangle$, eq. (4.88c) and

$$\frac{ct}{\Delta_0^2}\int_\xi^{\xi+\Delta_0 z} |\langle M_t(u) - M_t(\xi)\rangle|\,\mathrm{d}u \lesssim \frac{t(\Delta_0\,|z|)^{4/3}}{\Delta_0^2} \lesssim |z|^{4/3}\,.$$

we conclude the large-scale asymptotics

$$f(z) = \frac{z^2}{2} + \mathcal{O}\left(|z|^{4/3}\right), \qquad |z| \gg 1. \tag{4.103}$$

We now turn to the small-scale $|z| \ll 1$ asymptotics. We first specialize Lemma 4.5.1 and Lemma 4.5.2 to $\rho^* = \rho_t$ and collect the necessary conclusions in the following Lemma.

Lemma 4.5.4. *Under the assumptions of Theorem 4.2.3 it follows that ρ_t has a spectral gap $[\mathfrak{e}_0^-, \mathfrak{e}_0^+]$ of size*

$$\Delta_0 = \mathfrak{e}_0^+ - \mathfrak{e}_0^- = \Delta(ct \pm t^\rho)\left[1 + \mathcal{O}\left(t^{1/3}\right)\right], \quad \pm t^\rho := \begin{cases} 0 & \textit{in case (i)} \\ 3(\Delta^\rho)^{2/3}/(2\gamma)^{4/3} & \textit{in case (ii)} \\ -\pi^2\rho(\mathfrak{m}^\rho)^2/\gamma^4 & \textit{in case (iii).} \end{cases} \tag{4.104a}$$

Furthermore, in all three cases we have that ξ is is very close to the centre of the gap in the support of ρ_t in the sense that

$$\left|\xi - \frac{\mathfrak{e}_0^+ + \mathfrak{e}_0^-}{2}\right| = \mathcal{O}\left(t^{3/2}N^{-\epsilon/2}\right). \tag{4.104b}$$

Proof. We prove (4.104a)–(4.104b) separately in cases (i), (ii) and (iii).

(i) Here (4.104a) follows directly from (4.88c) with $\rho^* = \rho_t, t_* = ct, s = 0$ and $\mathfrak{c}^* = \mathfrak{c}^\rho$. Furthermore (4.104b) follows from (4.94a) with $\rho^* = \rho_t, t_* = ct$ and $\mathfrak{c}^* = \mathfrak{c}^\rho$.

(ii) We apply (4.88c) with $\rho^* = \rho = \rho_{ct}^{\mathrm{fc}}$, $t_* = t^\rho$, $s = 0$ to conclude that $\Delta^\rho = (2\gamma)^2(t^\rho/3)^{3/2}[1 + \mathcal{O}((t^\rho)^{1/3})]$, and that $\rho_{ct+t^\rho}^{\mathrm{fc}}$ has an exact cusp in some point $\mathfrak{c}$. Thus (4.104a) follows from another application of (4.88c) with $\rho^* = \rho_t, t_* = ct + t^\rho$, $s = 0$ and $\mathfrak{c}^* = \mathfrak{c}$. Furthermore, (4.104b) follows again from (4.94b) but this time with $\rho^* = \rho_t, t_* = ct + t^\rho, t' = t^\rho$ and $\mathfrak{e}_{t_*-t'}^\pm = \mathfrak{e}_\pm^\rho$, and using that $t_*^{1/9} \le N^{-\epsilon/2}$ for sufficiently small ϵ.

(iii) From (4.88a) with $\rho^* = \rho_t$, $t_* = ct - t^\rho$, $s = ct$ to conclude $\rho(\mathfrak{m}^\rho) = [1 + \mathcal{O}((t^\rho)^{1/2})]\gamma^2\sqrt{t^\rho}/\pi$, and that ρ_{ct-t^ρ} has an exact cusp in some point $\mathfrak{c}$. Finally, (4.104b) follows again from (4.94b) but with $\rho^* = \rho_t$, $t_* = ct - t^\rho$, $t' = t^\rho$ and $\mathfrak{m}_{t_*+t'} = \mathfrak{m}^\rho$, and using $t'/t_* \lesssim t^\rho/ct \lesssim N^{-\epsilon}$ and $t_*^{1/12} \le N^{-\epsilon/2}$ for sufficiently small ϵ. □

Equipped with Lemma 4.5.4 we can now turn to the small scale analysis of $f(z)$ and write out the Stieltjes transform to find

$$\begin{aligned} f(z) &= \frac{z^2}{2} - \frac{ct}{\Delta_0^2}\int_{\mathbb{R}}\int_\xi^{\xi+\Delta_0 z} \frac{u-\xi}{(x-u)(x-\xi)}\rho_t(x)\,\mathrm{d}u\,\mathrm{d}x \\ &= \frac{z^2}{2} - \frac{ct}{\Delta_0}\int_{\mathbb{R}}\int_0^z \frac{u}{(x-u)x}\rho_t(\xi+\Delta_0 x)\,\mathrm{d}u\,\mathrm{d}x. \end{aligned}$$

Note that these integrals are not singular since $\rho_t(\xi+\Delta_0 x)$ vanishes for $|x|\le 1/2$. We now perform the u integration to find

$$f(z)=\frac{z^2}{2}-\frac{ct}{\Delta_0}\int_{\mathbb{R}}\left[\log x-\log(x-z)-\frac{z}{x}\right]\rho_t(\xi+\Delta_0 x)\,\mathrm{d}x. \tag{4.105}$$

By using the precise shape (4.92) (with $s=0$) of ρ_t close to the edges $\mathfrak{e}_0^{\pm}$, and recalling the gap size from (4.104a) and location of ξ from (4.104b) we can then write

$$f(z)=(1+\mathcal{O}(t^{1/3}))\widetilde{g}(z)+\mathcal{O}\left(|z|^2\,t^{1/3}\right) \tag{4.106}$$

with

$$\widetilde{g}(z):=\frac{z^2}{2}-\frac{3\sqrt{3}}{2\pi(1\pm t^\rho/ct)}\int_{\mathbb{R}}\left[\log x-\log(x-z)-\frac{z}{x}\right]\Psi_{\text{edge}}(|x|-1/2)\mathbb{1}_{|x|\ge 1/2}\,\mathrm{d}x$$

being the leading order contribution. Here ± indicates that the formula holds for all three cases (i), (ii) and (iii) simultaneously, where $t^\rho=0$ in case (i). The contribution of the error term in (4.92) to the integral in (4.105) is of order $\mathcal{O}(|z|^2\,t^{1/2})$ using that $\log x-\log(x-z)-z/x=\mathcal{O}(|z/x|^2)$ and that $|x|\ge 1/2$ on the support of $\rho_t(\xi+\Delta_0 x)$. By the explicit integrals

$$\frac{3\sqrt{3}}{2\pi}\int_0^\infty\frac{\Psi_{\text{edge}}(x)}{(x+1/2)^2}\,\mathrm{d}x=\frac{1}{2},\qquad \frac{3\sqrt{3}}{2\pi}\int_0^\infty\frac{\Psi_{\text{edge}}(x)}{(x+1/2)^4}\,\mathrm{d}x=\frac{8}{27}$$

and a Taylor expansion of the logarithm $\log(x-z)$ we find that the quadratic term $z^2/2$ almost cancels and we conclude the small-scale asymptotics

$$\widetilde{g}(z)=\left(\frac{\pm t^\rho}{ct}\frac{z^2}{2}-\frac{4z^4}{27}\right)\left(1+\mathcal{O}\left(t^\rho/t\right)\right)+\mathcal{O}\left(|z|^5\right),\qquad |z|\ll 1. \tag{4.107}$$

4.5.3 Contour deformations

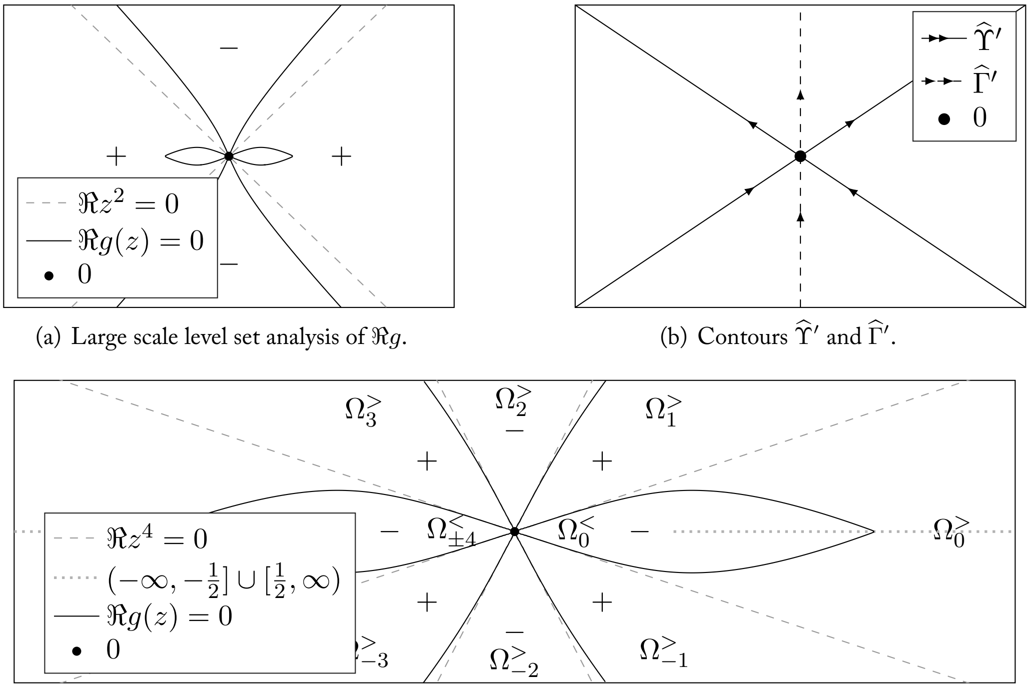
We now argue that we can deform the contours Υ,Γ and thereby via (4.101) the derived contours $\widehat{\Upsilon},\widehat{\Gamma}$, in a way which bounds the sign of $\Re g$ away from zero along the contours. Here $g(z)$ is the N-independent variant of $\widetilde{g}(z)$ given by

$$\begin{aligned} g(z)&:=\frac{z^2}{2}-\frac{3\sqrt{3}}{2\pi}\int_{\mathbb{R}}\left[\log x-\log(x-z)-\frac{z}{x}\right]\Psi_{\text{edge}}(|x|-1/2)\mathbb{1}_{|x|\ge 1/2}\,\mathrm{d}x\\ &=\widetilde{g}(z)+\mathcal{O}\left(N^{-\epsilon}\,|z|^2\right). \end{aligned} \tag{4.108}$$

The topological aspect of our argument is inspired by the approach in [95, 97, 96].

Lemma 4.5.5. *For all sufficiently small $\delta>0$ there exists $K=K(\delta)$ such that the following holds true. The contours Υ,Γ then can be deformed, without touching $(\operatorname{supp}\rho_t+[-1,1])\setminus\{\xi\}$ or each other, in such a way that the rescaled contours $\widehat{\Upsilon},\widehat{\Gamma}$ defined in (4.101) satisfy $\Re g\ge K$ on $\widehat{\Upsilon}\cap\{|z|>\delta\}$ and $\Re g\le -K$ on $\widehat{\Gamma}\cap\{|z|>\delta\}$. Furthermore, locally around 0 the contours can be chosen in such a way that*

$$\begin{aligned} &\widehat{\Gamma}\cap\{\,z\in\mathbb{C}\mid |z|\le\delta\,\}=(-\mathrm{i}\delta,\mathrm{i}\delta),\\ &\widehat{\Upsilon}\cap\{\,z\in\mathbb{C}\mid |z|\le\delta\,\}=(-\delta e^{\mathrm{i}\pi/4},\delta e^{\mathrm{i}\pi/4})\cup(-\delta e^{-\mathrm{i}\pi/4},\delta e^{-\mathrm{i}\pi/4}). \end{aligned} \tag{4.109}$$



(a) Large scale level set analysis of $\Re g$.

(b) Contours $\widehat{\Upsilon}'$ and $\widehat{\Gamma}'$.

(c) Small scale level set analysis of $\Re g$ where $\pm$ represents the sign of $\Re g(z)$.

Figure 4.2: Representative cusp analysis. Figures 4.2(c) and 4.2(a) show the level set $\Re g(z) = 0$. On a small scale $g(z) \sim z^4$, while on a large scale $g(z) \sim z^2$. Figure 4.2(b) shows the final deformed and rescaled contours $\widehat{\Upsilon}'$ and $\widehat{\Gamma}'$. Figure 4.2(c) furthermore shows the cone sections $\Omega_k^>$ and $\Omega_k^<$, where we for clarity do not indicate the precise area thresholds given by δ and R. We also do not specifically indicate $\Omega_k^<$ for $k = \pm 1, \pm 2, \pm 3$ as then $\mathrm{cc}(\Omega_k^<) = \mathrm{cc}(\Omega_k^>)$, cf. Claims 4–5 in the proof of Lemma 4.5.5.

Proof. Just as in (4.107) we have the expansion

$$g(z) = -\frac{4z^4}{27} + \mathcal{O}\left(|z|^5\right), \qquad |z| \ll 1. \tag{4.110}$$

It thus follows that for some small $\delta > 0$, and

$$\Omega_k^< := \left\{ z \in \mathbb{C} \;\middle|\; |z| < \delta, \left| \arg z - \frac{k\pi}{4} \right| < \delta \right\}$$

we have $\Omega_{\pm 1}^<, \Omega_{\pm 3}^< \subset \Omega_+ := \{ \Re g > 0 \}$ and $\Omega_0^<, \Omega_{\pm 2}^<, \Omega_4^< \subset \Omega_- := \{ \Re g < 0 \}$ in agreement with Figure 4.2(c). For large z, however, it also follows from (4.103) together with (4.108) and (4.106) that for some large R, and

$$\Omega_k^> := \left\{ z \in \mathbb{C} \;\middle|\; |z| > R, \frac{(k-1)\pi}{4} + \delta < \arg z < \frac{(k+1)\pi}{4} + \delta \right\}$$

we have $\Omega_0^>, \Omega_4^> \subset \Omega_+$ and $\Omega_{\pm 2}^> \subset \Omega_-$, in agreement with Figure 4.2(a). We denote the connected component of $\Omega_\pm$ containing some set A by $\mathrm{cc}(A)$.

Claim 1 – $\mathrm{cc}(\Omega_0^>)$, $\mathrm{cc}(\Omega_4^>)$ are the only two unbounded connected components of Ω_+: Suppose there was another unbounded connected component A of Ω_+. Since $\Omega_{\pm 2}^> \subset \Omega_-$ we would be able to find some $z_0 \in A$ with arbitrarily large $|\Re z_0|$. If $\Re z_0 > 0$, then we note that the map $x \mapsto \Re g(z_0 + x)$ is increasing, and otherwise we note that the map $x \mapsto \Re g(z_0 - x)$ is increasing. Thus it follows in both cases that the connected component A actually coincides with $\mathrm{cc}(\Omega_0^>)$ or with $\mathrm{cc}(\Omega_4^>)$, respectively.

Claim 2 – $\mathrm{cc}(\Omega_{\pm 2}^>)$ are the only two unbounded connected components of Ω_-: This follows very similarly to Claim 1.

Claim 3 – $\mathrm{cc}(\Omega_{\pm 1}^<)$, $\mathrm{cc}(\Omega_{\pm 2}^<)$, $\mathrm{cc}(\Omega_{\pm 3}^<)$ are unbounded: We note that the map $z \mapsto \Re g(z)$ is harmonic on $\mathbb{C} \setminus ([1/2, \infty) \cup (-\infty, -1/2])$ and subharmonic on $\mathbb{C}$. Therefore it follows that $\mathrm{cc}(\Omega_{\pm 1}^<), \mathrm{cc}(\Omega_{\pm 3}^<) \subset \Omega_+$ are unbounded. Since these sets are moreover symmetric with respect to the real axis it then also follows that $\mathrm{cc}(\Omega_{\pm 2}) \cap ((-\infty, -1/2] \cup [1/2, \infty)) = \emptyset$. This implies that $\Re g(z)$ is harmonic on $\mathrm{cc}(\Omega_{\pm 2}^<)$ and consequently also that $\mathrm{cc}(\Omega_{\pm 2}^<)$ are unbounded.

Claim 4 – $\mathrm{cc}(\Omega_1^<) = \mathrm{cc}(\Omega_{-1}^<) = \mathrm{cc}(\Omega_0^>)$ and $\mathrm{cc}(\Omega_3^<) = \mathrm{cc}(\Omega_{-3}^<) = \mathrm{cc}(\Omega_4^>)$: This follows from Claims 1–3.

Claim 5 – $\mathrm{cc}(\Omega_2^<) = \mathrm{cc}(\Omega_2^>)$ and $\mathrm{cc}(\Omega_{-2}^<) = \mathrm{cc}(\Omega_{-2}^>)$: This also follows from Claims 1–3.

The claimed bounds on $\Re g$ now follow from Claims 4–5 and compactness. The claimed small scale shape (4.109) follows by construction of the sets $\Omega_k^<$. □

From Lemma 4.5.5 and Lemma 4.2.8 it follows that K_N^t and thereby also $\widetilde{K_N^t}$ remain, with overwhelming probability, invariant under the chosen contour deformation. Indeed, K_N^t only has poles where $z = w$ or $z = \lambda_i$ for some i. Due to self-adjointness and Lemma 4.5.5, $z = \lambda_i$ can only occur if $\lambda_i = \xi$ or $\mathrm{dist}(\lambda_i, \operatorname{supp} \rho_t) > 1$. Both probabilities are exponentially small as a consequence of Lemma 4.2.8, since for the former we have $\eta_{\mathrm{f}}(\xi) \sim N^{-3/4+\epsilon/6}$ according to (4.7), while $\mathrm{dist}(\xi, \operatorname{supp} \rho_t) \sim N^{-3/4+3\epsilon/2}$.

For $z \in \widehat{\Gamma} \cup \widehat{\Upsilon}$ it follows from (4.109) that we can estimate

$$|f(z) - \widetilde{f}(z)| = \frac{ct}{\Delta_0^2} \left| \int_\xi^{\xi + \Delta_0 z} \langle \widetilde{G}_t(u) - M_t(u) \rangle \, \mathrm{d}u \right| \prec \frac{t\Delta_0 \, |z|}{N t^{3/2} \Delta_0^2} \sim \frac{|z|}{N t^2} = |z| \, N^{-2\epsilon}. \tag{4.111}$$

Indeed, for (4.111) we used (4.109) to obtain $\mathrm{dist}(\Re u, \operatorname{supp} \rho_t) \gtrsim t^{3/2}$, so that

$$|\langle \widetilde{G}_t(u) - M_t(u) \rangle| \prec 1/N t^{3/2}$$

follows from the local law from (4.8b).

We now distinguish three regimes: $|z| \lesssim N^{-\epsilon/2}$, $N^{-\epsilon/2} \lesssim |z| \ll 1$ and finally $|z| \gtrsim 1$ which we call microscopic, mesoscopic and macroscopic. We first consider the latter two regimes as they only contribute small error terms.

Macroscopic regime.

If either $|z| \geq \delta$ or $|w| \geq \delta$, it follows from Lemma 4.5.5 that $\Re g(w) \leq -K$ and/or $\Re g(z) \geq K$, and therefore together with (4.106),(4.108) and (4.111) that $\Re \widetilde{f}(w) \lesssim -K$

and/or $\Re\widetilde{f}(z) \gtrsim K$ with overwhelming probability. Using $\Delta_0 \sim N^{-3/4+3\epsilon/2}$ from (4.104a), we find that $N\Delta_0^2/ct \sim N^{2\epsilon}$ and $\Delta_0 N^{1/4}/ct\gamma \sim N^{\epsilon/2}$, so that the integrand in (4.102) in the considered regime is exponentially small.

Mesoscopic regime.

If either $\delta \geq |z| \gg N^{-\epsilon/2}$ or $\delta \geq |w| \gg N^{-\epsilon/2}$, then $\Re g(w) \sim -|w|^4 \ll -N^{-2\epsilon}$ and/or $\Re g(z) \sim |z|^4 \gg N^{-2\epsilon}$ from (4.110). Thus it follows from (4.106) and (4.108) that also $\Re f(w) \ll -N^{-2\epsilon}$ and/or $\Re f(z) \gg N^{-2\epsilon}$ and by (4.111) that with overwhelming probability $\Re\widetilde{f}(w) \ll -N^{-2\epsilon}$ and/or $\Re\widetilde{f}(z) \gg N^{-2\epsilon}$. Since $1/|w-z|$ is integrable over the contours it thus follows that the contribution to $\widetilde{K}_N^t(x,y)$, as in (4.102), from z, w with either $|z| \gg N^{-\epsilon/2}$ or $|w| \gg N^{-\epsilon/2}$ is negligible.

Microscopic regime.

We can now concentrate on the important regime where $|z|+|w| \lesssim N^{-\epsilon/2}$ and to do so perform another change of variables $z \mapsto ct\gamma z/\Delta_0 N^{1/4} \sim N^{-\epsilon/2} z$, $w \mapsto ct\gamma w/\Delta_0 N^{1/4} \sim N^{-\epsilon/2} w$ which gives rise to two new contours

$$\widehat{\Gamma}' := \frac{\Delta_0 N^{1/4}}{ct\gamma}\widehat{\Gamma}, \qquad \widehat{\Upsilon}' := \frac{\Delta_0 N^{1/4}}{ct\gamma}\widehat{\Upsilon},$$

as depicted in Figure 4.2(b), and the kernel

$$\widetilde{K}_N^t(x,y) = \frac{1}{(2\pi \mathrm{i})^2}\int_{\widehat{\Upsilon}'} \mathrm{d}z \int_{\widehat{\Gamma}'} \mathrm{d}w \frac{\exp\left(xz - yw + \frac{N\Delta_0^2}{ct}[\widetilde{f}(\frac{ct\gamma w}{\Delta_0 N^{1/4}}) - \widetilde{f}(\frac{ct\gamma z}{\Delta_0 N^{1/4}})]\right)}{w-z}. \tag{4.112}$$

We only have to consider w, z with $|w| + |z| \lesssim 1$ in (4.112) since $t/\Delta_0 N^{1/4} \sim N^{-\epsilon/2}$ and the other regime has already been covered in the previous paragraph before the change of variables.

We now separately estimate the errors stemming from replacing $\widetilde{f}(z)$ first by $f(z)$, then by $\widetilde{g}(z)$ and finally by $\pm t^\rho z^2/2ct - 4z^4/27$. We recall that $\Delta_0 \sim t^{3/2} = N^{-3/4+3\epsilon/2}$ from (4.104a), $t^\rho \lesssim N^{-1/2}$ from the definition of t^ρ in (4.104a), and that $t = N^{-1/2+\epsilon}$ which will be used repeatedly in the following estimates. According to (4.111), we have

$$\frac{N\Delta_0^2}{ct}\left|\widetilde{f}\Big(\frac{ct\gamma z}{\Delta_0 N^{1/4}}\Big) - f\Big(\frac{ct\gamma z}{\Delta_0 N^{1/4}}\Big)\right| \prec \frac{N\Delta_0^2}{t}\frac{t}{\Delta_0 N^{1/4}} N^{-2\epsilon}|z| \lesssim N^{-\epsilon/2}. \tag{4.113a}$$

Next, from (4.106) we have

$$\begin{aligned}\frac{N\Delta_0^2}{ct}\left|f\Big(\frac{ct\gamma z}{\Delta_0 N^{1/4}}\Big) - \widetilde{g}\Big(\frac{ct\gamma z}{\Delta_0 N^{1/4}}\Big)\right| &\lesssim t^{1/3}\left|\frac{ct\gamma z}{\Delta_0 N^{1/4}}\right|^2 \frac{N\Delta_0^2}{ct} + t^{1/3}\frac{N\Delta_0^2}{ct} \\ &\lesssim N^{-1/6+7\epsilon/3}.\end{aligned}$$

Finally, we have to estimate the error from replacing $\widetilde{g}(z)$ by its Taylor expansion with (4.107) and find

$$\frac{N\Delta_0^2}{ct}\left|\widetilde{g}\Big(\frac{ct\gamma z}{\Delta_0 N^{1/4}}\Big) - \frac{\pm t^\rho}{2ct}\Big(\frac{ct\gamma z}{\Delta_0 N^{1/4}}\Big)^2 + \frac{4}{27}\Big(\frac{ct\gamma z}{\Delta_0 N^{1/4}}\Big)^4\right| \lesssim N^{-\epsilon/2}. \tag{4.113b}$$

Finally, from (4.104a) and the definition of α from (4.6) we obtain that

$$\frac{N\Delta_0^2}{ct}\left[\frac{\pm t^\rho}{2ct}\left(\frac{ct\gamma z}{\Delta_0 N^{1/4}}\right)^2-\frac{4}{27}\left(\frac{ct\gamma z}{\Delta_0 N^{1/4}}\right)^4\right]=\left(\alpha\frac{z^2}{2}-\frac{z^4}{4}\right)[1+\mathcal{O}(t^{1/3})]. \quad (4.113c)$$

From (4.113) and the integrability of $1/|z-w|$ for small z,w along the contours we can thus conclude

$$\begin{aligned}\widetilde{K}_N^t(x,y)&=(1+\mathcal{O}(N^{-c}))\frac{1}{(2\pi\mathrm{i})^2}\\&\quad\times\int_{\widehat{\Upsilon}'}\mathrm{d}z\int_{\widehat{\Gamma}'}\mathrm{d}w\frac{\exp\left(xz-yw+z^4/4-\alpha z^2/2-w^4/4+\alpha w^2/2\right)}{w-z}.\end{aligned} \quad (4.114)$$

Furthermore, it follows from (4.109) that, as $N\to\infty$, the contours $\widehat{\Upsilon}', \widehat{\Gamma}'$ are those depicted in Figure 4.2(b), i.e.

$$\widehat{\Upsilon}'=(-e^{\mathrm{i}\pi/4}\infty,e^{\mathrm{i}\pi/4}\infty)\cup(-e^{-\mathrm{i}\pi/4}\infty,e^{-\mathrm{i}\pi/4}\infty),\qquad \widehat{\Gamma}':=(-\mathrm{i}\infty,\mathrm{i}\infty).$$

We recognize (4.114) as the extended Pearcey kernel from (4.5).

It is easy to see that all error terms along the contour integration are uniform in x,y running over any fixed compact set. This proves that $\widetilde{K}_N^t(x,y)$ converges to $K_\alpha(x,y)$ uniformly in x,y in a compact set. This completes the proof of Proposition 4.5.3. □

4.5.4 Green function comparison

We will now complete the proof of Theorem 4.2.3 by demonstrating that the local k-point correlation function at the common physical cusp location τ_0 of the matrices $\widetilde{H}_t$ does not change along the flow (4.84). Together with Proposition 4.5.3 this completes the proof of Theorem 4.2.3. A version of this *continuity of the matrix Ornstein-Uhlenbeck process* with respect to the local correlation functions that is valid in the bulk or at regular edges is the third step in the well known three step approach to universality [78]. We will present this argument in the more general setup of correlated random matrices, i.e. in the setting of [DS3]. In particular, we assume that the cumulants of the matrix elements w_{ab} satisfy the decay conditions (2.C)–(2.D), an assumption that is obviously fulfilled for deformed Wigner-type matrices.

We claim that the k-point correlation function $p_k^{(N)}$ of $H=\widetilde{H}_0$ and the corresponding k-point correlation function $\widetilde{p}_{k,t}^{(N)}$ of $\widetilde{H}_t$ stay close along the OU-flow in the sense that

$$\left|\int_{\mathbb{R}^k}F(\mathbf{x})\left[N^{k/4}p_k^{(N)}\left(\mathfrak{b}+\frac{\mathbf{x}}{\gamma N^{3/4}}\right)-\widetilde{p}_{k,t}^{(N)}\left(\mathfrak{b}+\frac{\mathbf{x}}{\gamma N^{3/4}}\right)\right]\mathrm{d}x_1\ldots\mathrm{d}x_k\right|=\mathcal{O}(N^{-c}), \quad (4.115)$$

for $\epsilon>0, t\le N^{-1/4-\epsilon}$, smooth functions F and some constant $c=c(k,\epsilon)$, where $\mathfrak{b}$ is the physical cusp point. The proof of (4.115) follows the standard arguments of computing t-derivatives of products of traces of resolvents $\widetilde{G}^{(t)}=(\widetilde{H}_t-z)$ at spectral parameters z just below the fluctuation scale of eigenvalues, i.e. for $\Im z\ge N^{-\zeta}\eta_f(\Re z)$. Since the procedure detailed e.g. in [78, Chapter 15] is well established and not specific to the cusp scaling, we keep our explanations brief.

The only cusp-specific part of the argument is estimating products of random variables

$$X_t=X_t(x):=N^{1/4}\langle\Im\widetilde{G}^{(t)}(\mathfrak{b}+\gamma^{-1}N^{-3/4}x+\mathrm{i}N^{-3/4-\zeta})\rangle$$

and we claim that

$$\mathbf{E}\Big[\prod_{j=1}^{k} X_t(x_j) - \prod_{j=1}^{k} X_0(x_j)\Big] \lesssim N^{-c} \tag{4.116}$$

as long as $t \le N^{-1/4-\epsilon}$ for some $c = c(k, \epsilon, \zeta)$. For simplicity we first consider $k = 1$ and find from Itô's Lemma that

$$\mathbf{E}\,\frac{\mathrm{d}X_t}{\mathrm{d}t} = \mathbf{E}\Big[-\frac{1}{2}\sum_{\alpha} w_\alpha \partial_\alpha X_t + \frac{1}{2}\sum_{\alpha,\beta} \kappa(\alpha,\beta)\partial_\alpha\partial_\beta X_t\Big], \tag{4.117}$$

which we further compute using a standard cumulant expansion, as already done in the bulk regime in the proof of Corollary 2.2.6 and in the edge regime in Section 3.6.2. We recall that $\kappa(\alpha,\beta)$, and more generally $\kappa(\alpha,\beta_1,\dots,\beta_k)$ denote the joint cumulants of the random variables w_α, w_β and $w_\alpha, w_{\beta_1},\dots,w_{\beta_k}$, respectively, which accordingly scale like N^{-1} and $N^{-(k+1)/2}$. Here greek letters $\alpha,\beta \in [N]^2$ are double indices. After cumulant expansion, the leading term in (4.117) cancels, and the next order contribution is

$$\sum_{\alpha,\beta_1,\beta_2} \kappa(\alpha,\beta_1,\beta_2)\,\mathbf{E}\big[\partial_\alpha\partial_{\beta_1}\partial_{\beta_2} X_t\big],$$

with $N^{-3/2}$ being the size of the cumulant $\kappa(\alpha,\beta_1,\beta_2)$. With $\alpha = (a,b)$ and $\beta_i = (a_i,b_i)$ we then estimate

$$\begin{aligned} N^{-3/4} \sum_{a,b,c} \sum_{a_1,b_1,a_2,b_2} &|\kappa(ab, a_1b_1, a_2b_2)|\,\mathbf{E}\Big|\widetilde{G}^{(t)}_{ca}\widetilde{G}^{(t)}_{ba_1}\widetilde{G}^{(t)}_{b_1a_2}\widetilde{G}^{(t)}_{b_2c}\Big| \\ &\le N^{-3/4-3/2+2+3/4+\zeta}\|\Im\widetilde{G}^{(t)}\|_3\|\widetilde{G}^{(t)}\|_3^2, \end{aligned}$$

where we used the Ward-identity and that $\max_\alpha \sum_{\beta_1,\beta_2}\kappa(\alpha,\beta_1,\beta_2) \lesssim N^{-3/2}$. We now use that according to the proof of Proposition 2.5.5, $\eta \mapsto \eta\|\widetilde{G}^{(t)}\|_p$ and similarly $\eta \mapsto \eta\|\Im\widetilde{G}^{(t)}\|_p$ are monotonically increasing with $\eta' = N^{-3/4+\zeta}$ to find $\|\Im\widetilde{G}^{(t)}\|_p \le_p N^{3\zeta-1/4}$ and $\|\widetilde{G}^{(t)}\|_p \le_p N^{3\zeta}$ from the local law from Theorem 4.2.5 and the scaling of ρ at η'. Since all other error terms can be handled similarly and give an even smaller contribution it follows that

$$\Big|\mathbf{E}\,\frac{\mathrm{d}X_t}{\mathrm{d}t}\Big| \lesssim N^{1/4+C\zeta} \quad \text{and more generally} \quad \Big|\mathbf{E}\,\frac{\mathrm{d}}{\mathrm{d}t}\prod_{j=1}^{k} X_t(x_j)\Big| \lesssim N^{1/4+Ck\zeta}, \tag{4.118}$$

for some constant $C > 0$. Now (4.116) and therefore (4.115) follow from (4.118) as in [78, Theorem 15.3] using the choice $t = N^{-1/2+\epsilon} \le N^{-1/4-\epsilon}$ and choosing ζ sufficiently small.

4.A Technical lemmata

Lemma 4.A.1. *Let $\mathbb{C}^{N\times N}$ be equipped with a norm $\|\cdot\|$. Let $\mathcal{A}\colon \mathbb{C}^{N\times N}\times\mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ be a bilinear form and let $\mathcal{B}\colon \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ a linear operator with a non-degenerate isolated eigenvalue β. Denote the spectral projection corresponding to β by $\mathcal{P}$ and by $\mathcal{Q}$ the one corresponding to the spectral complement of β, i.e.*

$$\mathcal{P} := -\lim_{\epsilon\searrow 0}\frac{1}{2\pi\mathrm{i}}\oint_{\partial B_\epsilon(\beta)}\frac{\mathrm{d}\omega}{\mathcal{B}-\omega} = \langle V_{\mathrm{l}},\cdot\rangle\, V_{\mathrm{r}}, \qquad \mathcal{Q} := 1-\mathcal{P},$$

where V_r *is the eigenmatrix corresponding to* β *and* $\langle V_\mathrm{l}, \cdot\rangle$ *a linear functional. Assume that for some positive constant* $\lambda > 1$ *the bounds*

$$\|\mathcal{A}\| + \left\|\mathcal{B}^{-1}\mathcal{Q}\right\| + \|\langle V_\mathrm{l}, \cdot\rangle\| + \|V_\mathrm{r}\| \le \lambda, \tag{4.119}$$

are satisfied, where we denote the induced norms on linear operators, linear functionals and bilinear forms on $\mathbb{C}^{N\times N}$ *by the same symbol* $\|\cdot\|$*. Then there exists a universal constant* $c > 0$ *such that for any* $\delta \in (0,1)$ *and any* $Y, X \in \mathbb{C}^{N\times N}$ *with* $\|Y\| + \|X\| \le c\lambda^{-4}$ *that satisfies the quadratic equation*

$$\mathcal{B}[Y] - \mathcal{A}[Y,Y] + X = 0, \tag{4.120}$$

the following holds: The scalar quantity

$$\Theta := \langle V_\mathrm{l}, Y\rangle ,$$

fulfils the cubic equation

$$\mu_3\Theta^3 + \mu_2\Theta^2 + \mu_1\Theta + \mu_0 = \lambda^{12}\mathcal{O}\left(\delta\,|\Theta|^3 + |\Theta|^4 + \delta^{-2}\,\|X\|^3\right), \tag{4.121}$$

with coefficients

$$\begin{aligned}
\mu_3 &= \langle V_\mathrm{l}, \mathcal{A}[V_\mathrm{r}, \mathcal{B}^{-1}\mathcal{Q}\mathcal{A}[V_\mathrm{r}, V_\mathrm{r}]] + \mathcal{A}[\mathcal{B}^{-1}\mathcal{Q}\mathcal{A}[V_\mathrm{r}, V_\mathrm{r}], V_\mathrm{r}]\rangle \\
\mu_2 &= \langle V_\mathrm{l}, \mathcal{A}[V_\mathrm{r}, V_\mathrm{r}]\rangle \\
\mu_1 &= -\langle V_\mathrm{l}, \mathcal{A}[\mathcal{B}^{-1}\mathcal{Q}[X], V_\mathrm{r}] + \mathcal{A}[V_\mathrm{r}, \mathcal{B}^{-1}\mathcal{Q}[X]]\rangle - \beta \\
\mu_0 &= \langle V_\mathrm{l}, \mathcal{A}[\mathcal{B}^{-1}\mathcal{Q}[X], \mathcal{B}^{-1}\mathcal{Q}[X]] - X\rangle .
\end{aligned} \tag{4.122}$$

Furthermore,

$$Y = \Theta V_\mathrm{r} - \mathcal{B}^{-1}\mathcal{Q}[X] + \Theta^2\mathcal{B}^{-1}\mathcal{Q}\mathcal{A}[V_\mathrm{r}, V_\mathrm{r}] + \lambda^7\mathcal{O}\left(|\Theta|^3 + |\Theta|\,\|X\| + \|X\|^2\right). \tag{4.123}$$

Here, the constants implicit in the $\mathcal{O}$*-notation depend on* c *only.*

Proof. We decompose Y as

$$Y = Y_1 + Y_2, \qquad Y_1 = \Theta V_\mathrm{r} - \mathcal{B}^{-1}\mathcal{Q}[X], \qquad Y_2 = \mathcal{Q}[Y] + \mathcal{B}^{-1}\mathcal{Q}[X].$$

Then (4.120) takes the form

$$\Theta\beta V_\mathrm{r} + \mathcal{P}[X] + \mathcal{B}\mathcal{Q}[Y_2] = \mathcal{A}[Y,Y]. \tag{4.124}$$

We project both sides with $\mathcal{Q}$, invert $\mathcal{B}$ and take the norm to conclude

$$\|Y_2\| = \lambda^2\mathcal{O}(\|Y_1\|^2 + \|Y_2\|^2),$$

Then we use the smallness of Y_2 by properly choosing δ and the definition of Y_1 to infer $Y_2 = \lambda^4\mathcal{O}_2$, where we introduced the notation

$$\mathcal{O}_k = \mathcal{O}(|\Theta|^k + \|X\|^k).$$

Inserting this information back into (4.124) and using $|\Theta| + \|X\| = \mathcal{O}(\lambda^{-3})$ reveals

$$Y_2 = \mathcal{B}^{-1}\mathcal{Q}\mathcal{A}[Y_1, Y_1] + \lambda^7\mathcal{O}_3. \tag{4.125}$$

In particular, (4.123) follows. Plugging (4.125) into (4.124) and applying the projection $\mathcal{P}$ yields

$$\begin{aligned}\Theta\beta V_{\mathrm{r}} + \mathcal{P}[X] &= \mathcal{P}\Big[\mathcal{A}[Y_1,Y_1] + \mathcal{A}[Y_1,Y_2] + \mathcal{A}[Y_2,Y_1]\Big] + \lambda^{11}\mathcal{O}_4 \\ &= \mathcal{P}\Big[\mathcal{A}[Y_1,Y_1] + \mathcal{A}[Y_1,\mathcal{B}^{-1}\mathcal{Q}\mathcal{A}[Y_1,Y_1]] + \mathcal{A}[\mathcal{B}^{-1}\mathcal{Q}\mathcal{A}[Y_1,Y_1],Y_1]\Big] + \lambda^{11}\mathcal{O}_4.\end{aligned}$$

For a linear operator $\mathcal{K}_1$ and a bilinear form $\mathcal{K}_2$ with $\|\mathcal{K}_1\| + \|\mathcal{K}_2\| \le 1$ we use the general bounds

$$\Theta\mathcal{K}_2[R,R] \le \delta\Theta^3 + \delta^{-1/2}\|R\|^3, \qquad \Theta^2\mathcal{K}_1[R] \le \delta\Theta^3 + \delta^{-2}\|R\|^3,$$

for any $R \in \mathbb{C}^{N\times N}$ and $\delta > 0$ to find

$$\begin{aligned}\Theta\beta V_{\mathrm{r}} + \mathcal{P}[X] = \mathcal{P}\Big[&\mathcal{A}[\Theta V_{\mathrm{r}} - \mathcal{B}^{-1}\mathcal{Q}[X], \Theta V_{\mathrm{r}} - \mathcal{B}^{-1}\mathcal{Q}[X]] + \Theta^3\mathcal{A}[V_{\mathrm{r}},\mathcal{B}^{-1}\mathcal{Q}\mathcal{A}[V_{\mathrm{r}},V_{\mathrm{r}}]] \\ &+ \Theta^3\mathcal{A}[\mathcal{B}^{-1}\mathcal{Q}\mathcal{A}[V_{\mathrm{r}},V_{\mathrm{r}}],V_{\mathrm{r}}]\Big] + \lambda^8\mathcal{O}(\delta|\Theta|^3 + \lambda^3|\Theta|^4 + \delta^{-2}\|X\|^3),\end{aligned}$$

which proves (4.121). □

Proof of Lemma 4.3.3. Due to the asymptotics

$$\Psi_{\mathrm{edge}} \sim \min\{\lambda^{1/2},\lambda^{1/3}\}, \quad \Psi_{\min} \sim \min\{\lambda^2,|\lambda|^{1/3}\}$$

and the classification of singularities in (4.4), we can infer the following behaviour of the self-consistent fluctuation scale from Definition 4.2.4. There exists a constant $c > 0$ only depending on the model parameters such that we have the following asymptotics. First of all, in the spectral bulk we trivially have that $\eta_{\mathrm{f}}(\tau) \sim N^{-1}$ as long as τ is at least a distance of $c > 0$ away from local minima of ρ. In the remaining cases we use the explicit shape formulae from (4.4) to compute η_{f} directly from Definition 4.2.4.

(a) *Non-zero local minimum or cusp.* Let τ be the location of a non-zero local minimum $\rho(\tau) = \rho_0 > 0$ or a cusp $\rho(\tau) = \rho_0 = 0$. Then

$$\eta_{\mathrm{f}}(\tau+\omega) \sim \begin{cases} 1/(N\max\{\rho_0,|\omega|^{1/3}\}), & \max\{\rho_0,|\omega|^{1/3}\} > N^{-1/4}, \\ N^{-3/4}, & \max\{\rho_0,|\omega|^{1/3}\} \le N^{-1/4}, \end{cases} \tag{4.126a}$$

for $\omega \in (-c,c)$.

(b) *Edge.* Let $\tau = \mathfrak{e}_\pm$ be the position of a left/right edge at a gap in $\operatorname{supp}\rho \cap (\mathfrak{e}_\pm - \kappa, \mathfrak{e}_\pm + \kappa)$ of size $\Delta \in (0,\kappa]$ (cf. (4.4b)). Then

$$\eta_{\mathrm{f}}(\mathfrak{e}_\pm \pm \omega) \sim \begin{cases} N^{-3/4}, & \omega \le \Delta \le N^{-3/4}, \\ \Delta^{1/6}/\omega^{1/2}N, & \Delta^{1/9}/N^{2/3} < \omega \le \Delta, \\ \Delta^{1/9}/N^{2/3}, & \omega \le \Delta^{1/9}/N^{2/3},\ \Delta > N^{-3/4}, \\ N^{-3/4}, & \Delta < \omega \le N^{-3/4}, \\ 1/\omega^{1/3}N, & \omega \ge N^{-3/4},\ \omega > \Delta, \end{cases} \tag{4.126b}$$

for $\omega \in [0,c)$.

The claimed bounds in Lemma 4.3.3 now follow directly from (4.15e) and (4.126) by distinguishing the respective regimes. □

Proof of Lemma 4.4.8. We start from (4.49) and estimate all vertex weights $\boldsymbol{w}^{(v)}$, interaction matrices $R^{(e)}$ and weight matrices $K^{(e)}$ trivially by

$$|w_a^{(v)}| \le C, \qquad |r_{ab}^{(e)}| \le CN^{-\deg(e)/2}, \qquad |k_{ab}^{(e)}| \le CN^{-l(e)}, \qquad \forall a, b$$

to obtain

$$|\mathrm{Val}(\Gamma)| \le C^{|V|+|\mathrm{IE}|+|\mathrm{WE}|} N^{n(\Gamma)-|V|} \Big\| \Big(\prod_{v \in V} \sum_{a_v \in J} \Big) \prod_{e \in \mathrm{GE}} G_e \Big\|_1 .$$

We now choose the vertex ordering $V = \{v_1, \dots, v_m\}$ as in Lemma 4.4.5. In the first step we partition the set of G-edges into three parts $\mathrm{GE} = E_1 \cup E_2 \cup E_3$: the edges not adjacent to v_m, $E_1 = \mathrm{GE} \setminus N(v_m)$, the non-Wardable edges adjacent to v_m, $E_2 = \mathrm{GE} \cap N(v_m) \setminus \mathrm{GE}_{\mathrm{W}}$ and the Wardable edges adjacent to v_m, $E_3 = \mathrm{GE}_{\mathrm{W}} \cap N(v_m)$. By the choice of ordering it holds that $|E_3| \le 2$. We introduce the shorthand notation $G_{E_i} = \prod_{e \in E_i} G_e$ and use the general Hölder inequality for any collection of random variables $\{X_A\}$ and $\{Y_A\}$ indexed by some arbitrary index set $\mathcal{A}$

$$\Big\| \sum_{A \in \mathcal{A}} |X_A Y_A| \Big\|_q \le \Big\| \sum_{A \in \mathcal{A}} |X_A| \Big\|_{q_1} |\mathcal{A}|^{1/q_2} \max_{A \in \mathcal{A}} \|Y_A\|_{q_2}, \qquad \frac{1}{q} = \frac{1}{q_1} + \frac{1}{q_2}$$

to compute

$$\begin{aligned} &\Big\| \sum_{a_{v_1}, \dots, a_{v_{m-1}}} |G_{E_1}| \sum_{a_{v_m}} |G_{E_2} G_{E_3}| \Big\|_q \\ &\le N^{(m-1)/q_2} \Big\| \sum_{a_{v_1}, \dots, a_{v_{m-1}}} |G_{E_1}| \Big\|_{q_1} \max_{a_1, \dots, a_{v_{m-1}}} \bigg(\Big\| \sum_{a_{v_m}} |G_{E_3}| \Big\|_{2q_2} N^{1/2q_2} \max_{a_{v_m}} \|G_{E_2}\|_{2q_2} \bigg), \end{aligned}$$

where we choose $1/q = 1/q_1 + 1/q_2$ in such a way that $q_2 \ge p/c\epsilon$. Since $|E_3| \le 2$ we can use (4.56a) to estimate

$$\Big\| \sum_{a_{v_m}} |G_{E_3}| \Big\|_{2q_2} \le N(\psi'_{2q_2})^{|E_3|} \le N(\psi + \psi'_{2q_2})^{|E_3|}$$

and it thus follows from

$$\|G_{E_2}\|_{2q_2} \le \prod_{e \in E_2} \|G_e\|_{2|E_2|q_2} = \|G - M\|_{2|E_2|q_2}^{|E_2 \cap \mathrm{GE}_{g-m}|} \|G\|_{2|E_2|q_2}^{|E_2 \setminus \mathrm{GE}_{g-m}|}$$

that

$$\begin{aligned} &\Big\| \sum_{a_{v_1}, \dots, a_{v_{m-1}}} |G_{E_1}| \sum_{a_{v_m}} |G_{E_2} G_{E_3}| \Big\|_q \qquad (4.127) \\ &\le N^{\epsilon/c} \Big\| \sum_{a_{v_1}, \dots, a_{v_{m-1}}} |G_{E_1}| \Big\|_{q_1} N(\psi + \psi'_{q'})^{|E_3|} (\psi + \psi'_{q'} + \psi''_{q'})^{|E_2 \cap \mathrm{GE}_{g-m}|} (1 + \|G\|_{q'})^{|E_2|} \end{aligned}$$

for $q' \geq 2q_2 \, |\mathrm{GE}|$. By using (4.127) inductively $m = |V| \leq cp$ times it thus follows that

$$\begin{aligned}&\Big\| \Big(\prod_{v \in V} \sum_{a_v \in J} \Big) \prod_{e \in \mathrm{GE}} G_e \Big\|_1 \\ &\qquad \leq N^{p\epsilon} N^{|V|} (\psi + \psi'_{q'})^{|\mathrm{GE}_{\mathrm{W}}|} (\psi + \psi'_{q'} + \psi''_{q'})^{|\mathrm{GE}_{g-m}|} (1 + \|G\|_{q'})^{|\mathrm{GE}|},\end{aligned}$$

proving the lemma. □

Lemma 4.A.2. *For the coefficient in* (4.83) *we have the expansion*

$$\frac{\langle \mathbf{b}^{(B)} \mathbf{pf}(R\mathbf{b}^{(B')}) \rangle \, \langle \mathbf{l}^{(B')}, \overline{\mathbf{l}^{(B)}} \rangle}{\langle \overline{\mathbf{b}^{(B)}}, \overline{\mathbf{l}^{(B)}} \rangle \, \langle \mathbf{l}^{(B')}, \mathbf{b}^{(B')} \rangle} = c\sigma \, \|F\| \, \langle |\mathbf{m}|^{-2} \, \mathbf{f}^2 \rangle + \mathcal{O}(\rho + \eta/\rho), \tag{4.128}$$

for some $|c| \sim 1$, *provided* $\|B^{-1}\|_{\infty \to \infty} \geq C$ *for some large enough constant* $C > 0$.

Proof. Recall from the explanation after (4.83) that $R' = S, T, T^t$ if $R = S, T^t, T$, respectively. As we saw in the proof of Lemma 4.4.14, in the case $R = T, T^t$ in the complex Hermitian symmetry class, the operator B as well as B' has a bounded inverse. Since we assume that $\|B^{-1}\|_{\infty \to \infty}$ is large, we have $R = R' = S$, which also includes the real symmetric symmetry class. In particular, we also have $\|(B')^{-1}\|_{\infty \to \infty} \geq C$ and all subsequent statements hold simultaneously for B and B'. We call $\mathbf{f}^{(S)}$ the normalised eigenvector corresponding to the eigenvalue with largest modulus of $F^{(S)} := |M| \, S \, |M|$, recalling $M = \mathrm{diag}(\mathbf{m})$. Since $B = |M| \, (1 - F^{(S)} + \mathcal{O}(\rho)) \, |M|^{-1}$ we can use perturbation theory of $F^{(S)}$ to analyse spectral properties of B. In particular, we find

$$\begin{aligned} \mathbf{b}^{(B)} &= |M| \, \mathbf{f}^{(S)} + \mathcal{O}(\rho), \qquad \mathbf{l}^{(B)} = |M|^{-1} \, \mathbf{f}^{(S)} + \mathcal{O}(\rho), \\ B^{-1} Q_B &= |M| \, (1 - F^{(S)})^{-1} (1 - P_{\mathbf{f}^{(S)}}) \, |M|^{-1} + \mathcal{O}(\rho), \end{aligned} \tag{4.129}$$

where $P_{\mathbf{f}^{(S)}}$ is the orthogonal projection onto the $\mathbf{f}^{(S)}$ direction. The error terms are measured in $\|\cdot\|_\infty$-norm. For the expansions (4.129) we used that F has a spectral gap in the sense that

$$\mathrm{Spec}(F^{(S)}/\|F^{(S)}\|) \subseteq [-1 + c, 1 - c] \cup \{1\},$$

for some constant $c > 0$, depending only on model parameters. By using (4.129) we see that the lhs. of (4.128) becomes $\pm \langle (\mathbf{f}^{(S)})^2 \mathbf{pf} \rangle \, \|F^{(S)}\| \, \langle |\mathbf{m}|^{-2} \, (\mathbf{f}^{(S)})^2 \rangle + \mathcal{O}\,(\rho)$. To complete the proof of the Lemma we note that $\mathbf{f}^{(S)} = \mathbf{f} / \, \|\mathbf{f}\| + \mathcal{O}\,(\eta/\rho)$ according to [12, Eq. (5.10)]. □

4.B Local law under uniform primitivity assumption

Here we explain the necessary changes to the proof of Theorem 4.2.5 and its corollaries when the fullness Assumption (4.B) is replaced by requiring only that the matrix of variances S is uniformly primitive, i.e. we verify Remark 4.2.9. We remark that this additional argument is also needed for the proof of the local law in the complex Hermitian Wigner-type case if we assume $s_{ij} \geq c/N$, but not necessarily fullness, Assumption (4.B), since in this case flatness in the sense of (4.14) may not hold.

For a uniformly primitive variance profile the flatness condition (4.14) may be violated. Thus we have to review all the instances in the proof of Theorem 4.2.5 where the lower bound in (4.14) was used (the upper bound follows from Assumption (4.A)). This happened at the following places:

(1) When we used [12, Proposition 3.5 and Lemma 4.8] to verify [12, Assumption 4.5] and thus applied [12, Lemma 5.1] to see that the stability operator $\mathcal{B}$ has a unique isolated eigenvalue β of smallest modulus in the paragraph proceeding Proposition 4.3.2. We also used [12, Assumption 4.5] (i) when we applied [12, Proposition 6.1] at the end of the proof of Proposition 4.3.4; (ii) when we referred to [12] inside the proof of Proposition 4.3.2 for various comparison relations (using [12, Eq. (5.15)], [12, Remark 7.3], [12, Proposition 6.1] and [12, Remark 10.4]) and finally (iii) when we imported the comparability of $\Im \mathbf{m}$ to its average $\Im \langle \mathbf{m} \rangle$ through the use of [12, Proposition 3.5] and asymptotic expansions for $V_{\mathrm{l}}, V_{\mathrm{r}}$ from [12, Corollary 5.2] for the proof of (4.23).

(2) When we imported the bounds (4.17) on the $\|\cdot\|_*$-norm from [DS4], where flatness was assumed.

(3) Inside the proof of Lemma 4.4.14, where [12, Lemma 5.1] was used again and where the fullness Assumption (4.B) was also used explicitly.

(4) Inside the proof of Lemma 4.A.2.

These are all instances where Assumption (4.B) was used either directly or indirectly through the flatness condition (4.14). We will now go through (1)-(4) one by one and show how the use of Assumption (4.B) can be avoided if the variance profile S is uniformly primitive. The proofs of Corollaries 4.2.6, 4.2.7 and 4.2.8 are not effected by this change in assumptions.

Modification of (1).

We will now show that [12, Assumption 4.5] still holds under the weaker uniform primitivity assumption and therefore all the mentioned results from [12] can still be used. For this purpose we consider the *saturated self-energy operator* $\mathcal{F} := Q\mathcal{S}[Q^* \cdot Q]Q^*$ from [12, Eq. (3.4)], where Q is defined as q in [12, Eq. (3.1)]. We see that $\mathcal{F}$ leaves the space of diagonal matrices and off-diagonal matrices invariant and splits as $\mathcal{F} = \mathcal{F}_{\mathrm{d}} + \mathcal{F}_{\mathrm{o}}$ with $\mathcal{F}_{\mathrm{d}}[\mathrm{diag}(\boldsymbol{r})] := \mathrm{diag}(F^{(S)}\boldsymbol{r})$, $F^{(S)} := |M|\, S\, |M|$ and $\mathcal{F}_{\mathrm{o}}[R] := |M|^{1/2} (T \odot (|M|^{1/2} R^t |M|^{1/2})) |M|^{1/2}$. Since $\|\mathcal{F}_{\mathrm{o}}\|_{\mathrm{hs}\to\mathrm{hs}} \lesssim 1/N$ by $\|M\| \lesssim 1$ and $|t_{ij}| \lesssim 1/N$ from Assumption (4.A) we obtain that [12, Assumption 4.5] reduces to a statement about the diagonal contribution $F^{(S)}$ of the saturated self-energy and follows from [7, Proposition 5.3].

Modification of (2).

The bounds (4.17) were proven in Lemma 3.3.4. Flatness was used in its proof only to establish the bound $\|\mathcal{B}^{-1}\mathcal{Q}\|_{\mathrm{hs}\to\mathrm{hs}} \lesssim 1$. However, since [12, Lemma 5.1] is still applicable according to the modification of (1) above, this bound remains valid.

Modification of (3).

Besides the use of flatness to justify the application of [12, Lemma 5.1], covered by the modification of (1), Assumption (4.B) was also used directly here to trivialise the case when H has complex valued entries and $R = T$ or $R = T^t$. In this case it was shown that $\|B^{-1}\| \lesssim 1$, cf. the proof of Lemma 4.4.14, and consequently it was possible to make the trivial choice

$P_B := 0$, $Q_B := 1$, see the paragraph before (4.77). This bound is no longer true under the uniform primitivity assumption since the assumption imposes no restriction on the self-adjoint matrices T, T^t, except the trivial one, $|t_{ij}| \le s_{ij}$.

In the general case Q_B is chosen as in (4.77). To show $\|B^{-1}Q_B\| \lesssim 1$ when $R = T$ or $R = T^t$ in (4.72), we write B in the form

$$B = 1 - \operatorname{diag}(\mathbf{m}^{\#_1}\mathbf{m}^{\#_2})R = |M|\,(1 - UF^{(R)})\,|M|^{-1}$$

with $F^{(R)} = |M|\,R\,|M|$, $M = \operatorname{diag}(\mathbf{m})$ and a diagonal unitary matrix U. Due to $\Im \mathbf{m} \sim \rho$ it is easy to check that $U = 1 + \mathcal{O}(\rho)$. Moreover, since $|M| \sim 1$ as quadratic forms and $1 - UF^{(R)} = 1 - F^{(R)} + \mathcal{O}(\rho)$ we can use perturbation theory of the self-adjoint matrix $F^{(R)}$ to invert $1 - UF^{(R)}$. We claim the following dichotomy for $F^{(R)}$: either $1 - F^{(R)}$ has a bounded inverse or a non-degenerate isolated eigenvalue close to zero with a spectral gap that is bounded from below. In the first case $1 - UF^{(R)}$ has a bounded inverse since ρ is small, hence $\|B^{-1}\| \lesssim 1$. In the second case B also has an isolated eigenvalue at 1 and Q_B projects to its spectral complement, so $\|B^{-1}Q_B\| \lesssim 1$ holds.

Now we formulate the dichotomy more precisely. We claim that for $Y := F^{(R)}/\|F^{(S)}\|$ with $\|F^{(S)}\|, \|Y\| \le 1$ we have either $\|Y\| \le 1 - \epsilon$ or

$$\operatorname{Spec}(Y/\|Y\|) \subseteq \{-1\} \cup [-1+\epsilon, 1-\epsilon] \cup \{1\}, \tag{4.130}$$

for some positive $\epsilon \sim 1$, where 1 is a non-degenerate eigenvalue in (4.130). Note that $F^{(S)}$ is the saturated self-energy and thus $\|F^{(S)}\| \le 1$ holds due to [10, Lemma 4.3], and $\|Y\| \le 1$ since $|t_{ij}| \le s_{ij}$. To verify (4.130) we apply the following lemma with $X := F^{(S)}/\|F^{(S)}\|$.

Lemma 4.B.1. *Let $\delta > 0$, and let $X = X^* \in \mathbb{R}^{N\times N}$ be a symmetric matrix with non-negative entries of norm $\|X\| = 1$. Assume that 1 is a non-degenerate eigenvalue with normalized $\|\mathbf{x}\|^2 = \langle |\mathbf{x}|^2\rangle = 1$ Perron-Frobenius eigenvector $X\mathbf{x} = \mathbf{x}$ with strictly positive entries $|x_i| \ge \delta$, and that X has a spectral gap $\|XQ_{\mathbf{x}}\| \le 1 - \delta$, where $Q_{\mathbf{x}} := 1 - \langle \mathbf{x}, \cdot\rangle\,\mathbf{x}$. Then there exists $\delta' = \delta'(\delta) > 0$ such that any self-adjoint matrix $Y \in \mathbb{C}^{N\times N}$ with $|y_{ij}| \le x_{ij}$ and $1 - \|Y\| \le \delta'$ has a normalized eigenvector $\mathbf{y}$ corresponding to the eigenvalue of largest modulus, $Y\mathbf{y} = \pm\|Y\|\,\mathbf{y}$, which satisfies*

$$\||\mathbf{y}| - \mathbf{x}\| \lesssim \sqrt{\epsilon}, \quad \langle |\mathbf{y}|, \mathbf{x}\rangle = 1 + \mathcal{O}(\epsilon), \quad \|YQ_{\mathbf{y}}\| \le 1 - \delta', \quad \||\mathbf{y}| - \mathbf{x}\|_1 \lesssim \epsilon\,|\log\epsilon| \tag{4.131}$$

with $\epsilon := 1 - \|Y\|$. Here $\|\cdot\|_1 := \langle|\cdot|\rangle$, and in the case $\epsilon = 0$ the rhs. of the ultimate inequality should be interpreted as 0.

The assumptions on $X = F^{(S)}/\|F^{(S)}\|$ in Lemma 4.B.1 are satisfied by [7, Proposition 5.3(iv, v)]. Note that the lemma shows that 1 and -1 cannot both be eigenvalues of $Y/\|Y\|$ in (4.130). This concludes the necessary modifications for (3) apart from the proof of Lemma 4.B.1, which we postpone until after the discussions of the modifications for (4).

Modification of (4).

Under the uniform primitivity assumption Lemma 4.A.2 does not hold in its current form. Instead an error term of the order $\|B^{-1}\|_{\infty\to\infty}^{-1}$ has to be added to the right hand side, i.e. it is replaced by the following lemma.

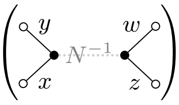
Lemma 4.B.2. *For the coefficient in* (4.83) *we have the expansion*

$$\frac{\langle \mathbf{b}^{(B)}\mathbf{pf}(R\mathbf{b}^{(B')})\rangle\,\langle \mathbf{l}^{(B')},\overline{\mathbf{l}^{(B)}}\rangle}{\langle \overline{\mathbf{b}^{(B)}},\overline{\mathbf{l}^{(B)}}\rangle\,\langle \mathbf{l}^{(B')},\mathbf{b}^{(B')}\rangle} = c\sigma\,\|F\|\,\langle |\mathbf{m}|^{-2}\,\mathbf{f}^2\rangle + \mathcal{O}(\rho+\eta/\rho+\left\|B^{-1}\right\|_{\infty\to\infty}^{-1}\log N), \tag{4.132}$$

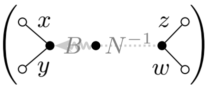
for some $|c|\sim 1$, *provided* $\|B^{-1}\|_{\infty\to\infty}\geq C$ *for some large enough constant* $C>0$.

Before proving Lemma 4.B.2 we will discuss how the proof of Proposition 4.4.12 continues after equation (4.83) when Lemma 4.B.2 is used instead of Lemma 4.A.2. This was the only instance where Lemma 4.A.2 was used in the proof of Theorem 4.2.5. By Lemma 4.B.2 the scalar factor in (4.83) is of the form $\sigma+\mathcal{O}(\rho+\eta/\rho+\|B^{-1}\|_{\infty\to\infty}^{-1}\log N)$, up to a bounded constant. Similarly to (4.80), we thus write (4.83) as the sum of three graph values

$$\frac{\langle \mathbf{b}^{(B)}\mathbf{pf}(R\mathbf{b}^{(B')})\rangle\,\langle \mathbf{l}^{(B')},\overline{\mathbf{l}^{(B)}}\rangle}{\langle \overline{\mathbf{b}^{(B)}},\overline{\mathbf{l}^{(B)}}\rangle\,\langle \mathbf{l}^{(B')},\mathbf{b}^{(B')}\rangle}\,\mathrm{Val}$$



$$= (\sigma+\mathcal{O}(\rho+\eta/\rho))\,\mathrm{Val}(\Gamma') + (\log N)\,\mathrm{Val} \tag{4.133}$$

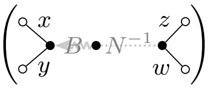


with W-Est(Γ') = W-Est(Γ), where we absorbed the $\|F\|\,\langle|\mathbf{m}|^{-2}\,\mathbf{f}^2\rangle\lesssim 1$ factor into the weight matrix of Γ'. Here we were able to insert the B-operator in (4.133) since

$$\left\|B^{-1}\right\|_{\infty\to\infty}^{-1}\left\|B^{-1}K\right\|_{\infty\to\infty}\leq\|K\|_{\infty\to\infty}\lesssim\frac{1}{N},$$

where K is as in (4.82). For the last graph in (4.133) we apply Lemma 4.4.13 to find

$$\mathrm{Val}$$



$$= \sum_{\Gamma''\in\mathcal{G}'_\Gamma}\mathrm{Val}(\Gamma'')+\mathcal{O}(N^{-p}),\qquad \text{W-Est}(\Gamma'')\leq_p\sigma_q\,\text{W-Est}(\Gamma).$$

Thus we gained a factor σ_q. This finishes the proof of a version of Proposition 4.4.12, where in (4.64) the factor σ_q is replaced by $\sigma_q\log N$, under the uniform primitivity assumption on S. The extra $\log N$-factor in Proposition 4.4.12 does not effect the proof of Theorem 4.3.7 because it is insignificant compared to the N^ϵ-factors in (4.19c) and (4.19d).

Altogether, this finishes the discussion of the modifications (1)-(4) and thus verifies the validity of Remark 4.2.9. We finish this section of the appendix by providing the remaining proofs of Lemmas 4.B.2 and 4.B.1.

Proof of Lemma 4.B.2. The case $R=S$ was already considered in Lemma 4.A.2, whence we can restrict ourselves to $R=T$ and $R=T^t$ here. For any of the possible choices of stability operators $B=1-\mathrm{diag}(\mathbf{m}^{\#_1}\mathbf{m}^{\#_2})R$ in (4.72) we call $\mathbf{f}^{(R)}$ the normalized eigenvector corresponding to the eigenvalue with largest modulus of $F^{(R)}:=|M|\,R\,|M|$. We may assume that $\|F^{(R)}\|\geq 1-\epsilon$ for some sufficiently small $\epsilon>0$ since otherwise B has a bounded inverse and this is not the situation in which Lemma 4.B.2 is used. Since $B=|M|\,(1-F^{(R)}+\mathcal{O}(\rho))\,|M|^{-1}$ we use perturbation theory of the self-adjoint matrix $F^{(R)}$ to analyse B. For $R=T,T^t$ we apply Lemma 4.B.1 with the choice $X:=F^{(S)}/\|F^{(S)}\|$, $Y:=F^{(R)}/\left\|F^{(S)}\right\|$. As we argued for the modification of (3) above, $F^{(R)}$ has a spectral gap in the sense of (4.130). Expanding around the isolated eigenvalue $\pm\|F^{(R)}\|$ of $F^{(R)}$ we still have (4.129) with $F^{(S)}$ replaced by $F^{(R)}$ and $\mathbf{f}^{(S)}$ replaced by $\mathbf{f}^{(R)}$. We have the following cases to consider:

$$B = (1 - \operatorname{diag}(\mathbf{m}^{\#_1}\mathbf{m}^{\#_2})R', \qquad B' = (1 - \operatorname{diag}(\mathbf{m}^{\#_1}\mathbf{m}^{\#_2})R,$$

where

$$R = T, \qquad R' = T^t, \qquad \text{or} \tag{4.134a}$$
$$R = T^t, \qquad R' = T. \tag{4.134b}$$

Recall that $\#_i, i = 1, 2, 3, 4$ stand for the identity or the complex conjugation operators, but the individual choices are not important as the $F^{(R)}$ and $F^{(R')}$ operators are not influenced by them. In cases (4.134a) and (4.134b) we have $\mathbf{f}^{(R')} = \overline{\mathbf{f}^{(R)}}$. Thus we find

$$\frac{\langle \mathbf{b}^{(B)}\mathbf{pf}(R\mathbf{b}^{(B')})\rangle \, \langle \mathbf{l}^{(B')}, \overline{\mathbf{l}^{(B)}}\rangle}{\langle \overline{\mathbf{b}^{(B)}}, \overline{\mathbf{l}^{(B)}}\rangle \, \langle \mathbf{l}^{(B')}, \mathbf{b}^{(B')}\rangle} = \pm \| F^{(R')}\| \, \langle |\mathbf{f}^{(R)}|^2 \mathbf{pf}\rangle \, \langle |\mathbf{m}|^{-2} \, |\mathbf{f}^{(R)}|^2\rangle + \mathcal{O}(\rho),$$

where we used (4.129), the eigenvalue equation $F^{(R')}\mathbf{f}^{(R')} = \pm \|F^{(R')}\| \mathbf{f}^{(R')}$ and that $\mathbf{f}^{(R)}$ and $\mathbf{f}^{(R')}$ are normalised.

Now we use (4.131) to approximate $|\mathbf{f}^{(R)}|$ with first $\mathbf{f}^{(S)}$, and then $\mathbf{f}$ as in the end of the proof of Lemma 4.A.2, to see that

$$\begin{aligned} \langle |\mathbf{f}^{(R)}|^2 \mathbf{pf}\rangle &= c\sigma + \mathcal{O}\left(\eta/\rho + (1 - \|F^{(R)}\|) \left|\log(1 - \|F^{(R)}\|)\right|\right), \\ \Big\langle \Big| \frac{\mathbf{f}^{(R)}}{|\mathbf{m}|}\Big|^2 \Big\rangle &= c' \Big\langle \Big[\frac{\mathbf{f}}{|\mathbf{m}|}\Big]^2 \Big\rangle + \mathcal{O}\left(\eta/\rho + (1 - \|F^{(R)}\|) \left|\log(1 - \|F^{(R)}\|)\right|\right) \end{aligned}$$

for some $|c|, |c'| \sim 1$. Since $B = |M| \, U(U^* - F^{(R)}) \, |M|^{-1}$ for some diagonal unitary $U = 1 + \mathcal{O}(\rho)$ implies $\|B^{-1}\|_{\infty\to\infty} \lesssim \|B^{-1}\| \lesssim \|\mathbf{m}\| \, \|\mathbf{m}^{-1}\| (1 - \|F^{(R)}\|)^{-1}$ and since $|\log(1 - \|F^{(R)}\|)| \lesssim \log N$ the bound in (4.132) follows. Here we used $1 - \|F^{(R)}\| \geq 1 - \|F\| \gtrsim \Im z \geq 1/N$ by [10, Lemma 4.3]. □

Proof of Lemma 4.B.1. Throughout the proof we consider δ as fixed and consider only the case where $\epsilon \ll 1$, as the statement is trivial otherwise. Within the proof we understand applications of functions (e.g. $|\cdot|$ or $\Re$) to vectors/matrices and also inequalities between vectors/matrices in an elementwise sense. Let $\mathbf{y}_1, \mathbf{y}_2$ be normalized eigenvectors corresponding to the, in modulus, largest and second largest eigenvalue of Y, i.e. $|\langle \mathbf{y}_1, Y\mathbf{y}_1\rangle| = \|Y\| =: s_1$, $\mathbf{y}_1 = \mathbf{y}$ and $\mathbf{y}_2 \in \arg\max_{\mathbf{y}\perp\mathbf{y}_1} |\langle \mathbf{y}, Y\mathbf{y}\rangle|$ and set $s_2 := |\langle \mathbf{y}_2, Y\mathbf{y}_2\rangle|$. Furthermore, let V_1, V_2 denote diagonal unitary matrices such that $\mathbf{y}_i = V_i \, |\mathbf{y}_i|$. We then compute

$$\begin{aligned} s_i = \sigma_i \, \langle \mathbf{y}_i, Y\mathbf{y}_i\rangle &= \langle |\mathbf{y}_i| \, , \sigma_i \Re V_i^* Y V_i \, |\mathbf{y}_i|\rangle \leq \langle |\mathbf{y}_i| \, , X \, |\mathbf{y}_i|\rangle \\ &\leq |\langle \mathbf{x}, |\mathbf{y}_i|\rangle|^2 + (1-\delta) \, \|Q_{\mathbf{x}} \, |\mathbf{y}_i|\|^2 = 1 - \delta \, \|Q_{\mathbf{x}} \, |\mathbf{y}_i|\|^2 \, , \end{aligned}$$

where $\sigma_i = \operatorname{sgn} \langle \mathbf{y}_i, Y\mathbf{y}_i\rangle$ and we used $|Y| \leq X$ in the first inequality. With $\epsilon_i := 1 - s_i \geq 0$ it then follows that $\|Q_{\mathbf{x}} \, |\mathbf{y}_i|\| \lesssim \sqrt{\epsilon_i}$, and by positivity of $\mathbf{x}$, that $\langle \mathbf{x}, |\mathbf{y}_i|\rangle = 1 + \mathcal{O}(\epsilon_i)$, $\||\mathbf{y}_i| - \mathbf{x}\| \lesssim \sqrt{\epsilon_i}$. Recalling $\epsilon_1 = \epsilon = 1 - \|Y\|$, this completes the proof of the first two inequalities in (4.131).

We now turn to the claimed bound on $\|YQ_{\mathbf{y}}\| = s_2$. We first note that

$$X - \sigma_i \Re V_i^* Y V_i = |X - \sigma_i \Re V_i^* Y V_i|$$

since $|V_i^* Y V_i| \leq X$ and therefore

$$\begin{aligned}\langle \mathbf{x}, |X - \sigma_i \Re V_i^* Y V_i| \, \mathbf{x} \rangle &= 1 - \sigma_i \langle \mathbf{x}, V_i^* Y V_i \mathbf{x} \rangle = 1 - \sigma_i \langle |\mathbf{y}_i|, V_i^* Y V_i |\mathbf{y}_i| \rangle + \mathcal{O}(\sqrt{\epsilon_i}) \\ &= 1 - s_i + \mathcal{O}(\sqrt{\epsilon_i}) = \mathcal{O}(\sqrt{\epsilon_i})\end{aligned}$$

and by taking the imaginary part and using the elementary inequality

$$(\Im z)^2 \leq 2 |z| \left| \Re z - |z| \right|$$

we also have $\langle \mathbf{x}, |\Im V_i^* Y V_i| \, \mathbf{x} \rangle \lesssim \sqrt[4]{\epsilon_i}$. It thus follows that also $\langle \mathbf{x}, |X - \sigma_i V_i^* Y V_i| \, \mathbf{x} \rangle \lesssim \sqrt[4]{\epsilon_i}$ and consequently

$$\begin{aligned}\mathcal{O}(\sqrt[4]{\epsilon_1}) &= \langle \mathbf{x}, V_2^* V_1 (X - \sigma_1 V_1^* Y V_1) V_1^* V_2 \mathbf{x} \rangle \\ &= \langle V_1^* V_2 \mathbf{x}, X V_1^* V_2 \mathbf{x} \rangle - \sigma_1 \sigma_2 + \mathcal{O}(\sqrt{\epsilon_2}),\end{aligned} \tag{4.135}$$

where the second equality used that $\|V_2 \mathbf{x} - \mathbf{y}_2\| \lesssim \sqrt{\epsilon_2}$ and $\langle \mathbf{y}_2, Y \mathbf{y}_2 \rangle = \sigma_2 s_2 = \sigma_2 + \mathcal{O}(\epsilon_2)$. But using the additional information that $\mathbf{y}_1 \perp \mathbf{y}_2$ it now follows that the projection of $V_1^* V_2 \mathbf{x}$ onto $\mathbf{x}$ almost vanishes in the sense $|\langle \mathbf{x}, V_1^* V_2 \mathbf{x} \rangle| \lesssim \sqrt{\epsilon_1} + \sqrt{\epsilon_2}$, while we recall that the matrix X is assumed to be bounded as $\|X Q_{\mathbf{x}}\| \leq 1 - \delta$ on the complement of $\mathbf{x}$. Therefore $|\langle V_1^* V_2 \mathbf{x}, X V_1^* V_2 \mathbf{x} \rangle| \leq 1 - \delta + \mathcal{O}(\sqrt{\epsilon_1} + \sqrt{\epsilon_2})$ and consequently $\delta \lesssim \sqrt[4]{\epsilon_1} + \sqrt{\epsilon_2}$ from (4.135). In the considered case $\epsilon_1 \ll 1$ it follows that $\epsilon_2 \gtrsim 1$ and $\|Y Q_{\mathbf{y}}\| \leq 1 - \epsilon_2$, confirming the third inequality in (4.131).

We now turn to the ultimate inequality in (4.131) and use $|\mathbf{y}| = \sigma_1 \|Y\|^{-1} V_1^* Y V_1 |\mathbf{y}| \leq \|Y\|^{-1} X |\mathbf{y}|$ and by iteration $|\mathbf{y}| \leq \|Y\|^{-k} X^k |\mathbf{y}|$ for integers k. Using $|x_i| \geq \delta$ and $\langle \mathbf{x}, |\mathbf{y}| - X^k |\mathbf{y}| \rangle = \langle \mathbf{x} - X^k \mathbf{x}, |\mathbf{y}| \rangle = \langle 0, |\mathbf{y}| \rangle = 0$ we find

$$\begin{aligned}\delta \left\| |\mathbf{y}| - X^k |\mathbf{y}| \right\|_1 &\leq \left\langle \mathbf{x}, \left| |\mathbf{y}| - X^k |\mathbf{y}| \right| \right\rangle = 2 \langle \mathbf{x}, (|\mathbf{y}| - X^k |\mathbf{y}|)_+ \rangle - \langle \mathbf{x}, |\mathbf{y}| - X^k |\mathbf{y}| \rangle \\ &\leq 2(\|Y\|^{-k} - 1) \langle \mathbf{x}, X^k |\mathbf{y}| \rangle,\end{aligned}$$

to infer $\left\| |\mathbf{y}| - X^k |\mathbf{y}| \right\|_1 \lesssim k \epsilon_1 = k(1 - \|Y\|)$ from which we conclude that

$$\| |\mathbf{y}| - \mathbf{x} \|_1 = \left\| \mathbf{x}(\langle |\mathbf{y}|, \mathbf{x} \rangle - 1) + |\mathbf{y}| - X^k |\mathbf{y}| + X^k Q_{\mathbf{x}} |\mathbf{y}| \right\|_1 \lesssim k\epsilon + (1 - \delta)^k,$$

where we used the second inequality in (4.131). Thus with the choice $k = \delta^{-1} |\log \epsilon|$ the ultimate inequality in (4.131) follows. Tracking the dependence on δ in the proof yields that we can choose $\delta' = c\delta^3$ for some universal constant c. □

Cusp Universality for Random Matrices II: The Real Symmetric Case 5

We prove that the local eigenvalue statistics of real symmetric Wigner-type matrices near the cusp points of the eigenvalue density are universal. Together with the companion paper [DS5], which proves the same result for the complex Hermitian symmetry class, this completes the last remaining case of the Wigner-Dyson-Mehta universality conjecture after bulk and edge universalities have been established in the last years. We extend the recent Dyson Brownian motion analysis at the edge [122] to the cusp regime using the optimal local law from [DS5] and the accurate local shape analysis of the density from [12]. We also present a novel method to improve the estimate on eigenvalue rigidity via the maximum principle of the heat flow related to the Dyson Brownian motion.

Published as G. Cipolloni, L. Erdős, T. Krüger, and D. Schröder, *Cusp universality for random matrices II: The real symmetric case.* preprint (2018), `arXiv:1811.04055`.

5.1 Introduction

We consider *Wigner-type* matrices, i.e. $N \times N$ Hermitian random matrices H with independent, not necessarily identically distributed entries above the diagonal; a natural generalization of the standard Wigner ensembles that have i.i.d. entries. The Wigner-Dyson-Mehta (WDM) conjecture asserts that the local eigenvalue statistics are universal, i.e. they are independent of the details of the ensemble and depend only on the *symmetry type*, i.e. on whether H is real symmetric or complex Hermitian. Moreover, different statistics emerge in the bulk of the spectrum and at the spectral edges with a square root vanishing behavior of the eigenvalue density. The WDM conjecture for both symmetry classes has been proven for Wigner matrices, see [78] for complete historical references. Recently it has been extended to more general ensembles including Wigner-type matrices in the bulk and edge regimes; we refer to the companion paper [DS5] for up to date references.

The key tool for the recent proofs of the WDM conjecture is the Dyson Brownian motion (DBM), a system of coupled stochastic differential equations. The DBM method has evolved during the last years. The original version, presented in the monograph [78], was in the spirit of a high dimensional analysis of a strongly correlated Gibbs measure and its dynamics. Starting in [79] with the analysis of the underlying parabolic equation and its short range approximation, the PDE component of the theory became prominent. With the coupling idea, introduced in [43, 40], the essential part of the proofs became fully deterministic, greatly simplifying the technical aspects. In the current paper we extend this trend and use PDE methods even for the proof of the rigidity bound, a key technical input, that earlier was obtained with direct random matrix methods.

The historical focus on the bulk and edge universalities has been motivated by the Wigner ensemble since, apart from the natural bulk regime, its semicircle density vanishes as a square root near the edges, giving rise to the Tracy-Widom statistics. Beyond the Wigner ensemble, however, the density profile shows a much richer structure. Already Wigner matrices with nonzero expectation on the diagonal, also called *deformed Wigner ensemble*, may have a density supported on several intervals and a cubic root cusp singularity in the density arises whenever two such intervals touch each other as some deformation parameter varies. Since local spectral universality is ultimately determined by the local behavior of the density near its vanishing points, the appearance of the cusp gives rise to a new type of universality. This was first observed in [50] and the local eigenvalue statistics at the cusp can be explicitly described by the Pearcey process in the complex Hermitian case [172]. The corresponding explicit formulas for the real symmetric case have not yet been established.

The key classification theorem [10] for the density of Wigner-type matrices showed that the density may vanish only as a square root (at regular edges) or as a cubic root (at cusps); no other singularity may occur. This result has recently been extended to a large class of matrices with correlated entries [12]. In other words, the cusp universality is the third and last universal spectral statistics for random matrix ensembles arising from natural generalizations of the Wigner matrices. We note that invariant β-ensembles may exhibit further universality classes, see [57].

In the companion paper [DS5] we established cusp universality for Wigner-type matrices in the complex Hermitian symmetry class. In the present work we extend this result to the real symmetric class and even to certain space-time correlation functions. In fact, we show the appearance of a natural one-parameter family of universal statistics associated to a family of singularities of the eigenvalue density that we call *physical cusps*. In both works we follow the *three step strategy*, a general method developed for proving local spectral universality for random matrices, see [78] for a pedagogical introduction. The first step is the *local law* or *rigidity*, establishing the location of the eigenvalues with a precision slightly above the typical local eigenvalue spacing. The second step is to establish universality for ensembles with a tiny Gaussian component. The third step is a perturbative argument to remove this tiny Gaussian component relying on the optimal local law. The first and third steps are insensitive to the symmetry type, in fact the optimal local law in the cusp regime has been established for both symmetry classes in [DS5] and it completes also the third step in both cases.

There are two different strategies for the second step. In the complex Hermitian symmetry class, the Brézin-Hikami formula [49] turns the problem into a saddle point analysis for a contour integral. This direct path was followed in [DS5] relying on the optimal local law. In the real symmetric case, lacking the Brézin-Hikami formula, only the second strategy via the analysis of Dyson Brownian motion (DBM) is feasible. This approach exploits the very fast decay to local equilibrium of DBM. It is the most robust and powerful method up to now to establish local spectral universality. In this paper we present a version of this method adjusted to the cusp situation. We will work in the real symmetric case for definiteness. The proof can easily be modified for the complex Hermitian case as well. The DBM method does not explicitly yield the local correlation kernel. Instead it establishes that the local statistics are universal and therefore can be identified from a reference ensemble that we will choose as the simplest Gaussian ensemble exhibiting a cusp singularity.

In this paper we partly follow the recent DBM analysis at the regular edges [122] and we extend it to the cusp regime, using the optimal local law from the companion paper [DS5] and the precise control of the density near the cusps [7, 12]. The main conceptual difference between [122] and the current work is that we obtain the necessary local law along the time evolution of DBM via novel DBM methods in Section 5.6. Some other steps, such as the Sobolev inequality, heat kernel estimates from [41] and the finite speed of propagation [79, 122, 40], require only moderate adjustments for the cusp regime, but for completeness we include them in the Appendix. The comparison of the short range approximation of the DBM with the full evolution, Lemma 5.7.2 and Lemma 5.C.1, will be presented in detail in Section 5.7 and in Appendix 5.C since it is more involved in the cusp setup, after the necessary estimates on the semicircular flow near the cusp are proven in Section 5.4.

We now outline the novelties and main difficulties at the cusp compared with the edge analysis in [122]. The basic idea is to interpolate between the time evolution of two DBM's, with initial conditions given by the original ensemble and the reference ensemble, respectively, after their local densities have been matched by shift and scaling. Beyond this common idea there are several differences.

The first difficulty lies in the rigidity analysis of the DBM starting from the interpolated initial conditions. The optimal rigidity from [DS5], that holds for very general Wigner–type matrices, applies for the flows of both the original and the reference matrices, but it does not directly apply to the interpolating process. The latter starts from a regular initial data but it runs for a very short time, violating the *flatness* (i.e. effective mean-field) assumption of [DS5]. While it is possible to extend the analysis of [DS5] to this case, here we chose a technically lighter and conceptually more interesting route. We use the maximum principle of the DBM to transfer rigidity information on the reference process to the interpolating one after an appropriate localization.

The second difficulty in the cusp regime is that the shape of the density is highly unstable under the semicircular flow that describes the evolution of the density under the DBM. The regular edge analysed in [122] remains of square root type along its dynamics and it can be simply described by its location and its multiplicative *slope parameter* – both vary regularly with time. In contrast, the evolution of the cusp is a relatively complicated process: it starts with a small gap that shrinks to zero as the cusp forms and then continues developing a small local minimum. The density is described by quite involved shape functions, see (5.2c), (5.2e), that have a two-scale structure, given in terms of a total of three parameters, each varying on different time scales. For example, the location of the gap moves linearly with time, the length of gap shrinks as the $3/2$-th power of the time, while the local minimum after the cusp increases as the $1/2$-th power of the time. The scaling behavior of the corresponding quantiles, that approximate the eigenvalues by rigidity, follows the same complicated pattern of the density. All these require a very precise description of the semicircular flow near the cusp as well as the optimal rigidity.

The third difficulty is that we need to run the DBM for a relatively long time in order to exploit the local decay; in fact this time scale, $N^{-1/2+\epsilon}$ is considerably longer than the characteristic time scale $N^{-3/4}$ on which the physical cusp varies under the semicircular flow. We need to tune the initial condition very precisely so that after a relatively long time it develops a cusp exactly at the right location with the right slope.

The fourth difficulty is that, unlike for the regular edge regime, the eigenvalues or quantiles on both sides of the (physical) cusp contribute to the short range approximation of the dynamics, their effect cannot be treated as mean-field. Moreover, there are two scaling regimes for quantiles corresponding to the two-scale structure of the density.

Finally, we note that the analysis of the semicircular flow around the cusp, partly completed already in the companion paper [DS5], is relatively short and transparent despite its considerably more complex pattern compared to the corresponding analysis around the regular edge. This is mostly due to strong results imported from the general shape analysis [7]. Not only the exact formulas for the density shapes are taken over, but we also heavily rely on the $1/3$-Hölder continuity in space and time of the density and its Stieltjes transform, established in the strongest form in [12].

Notations and conventions. We now introduce some custom notations we use throughout the paper. For integers n we define $[n] := \{1, \dots, n\}$. For positive quantities f, g, we write $f \lesssim g$ and $f \sim g$ if $f \le Cg$ or, respectively, $cg \le f \le Cg$ for some constants c, C that depend only on the *model parameters*, i.e. on the constants appearing in the basic Assumptions (5.A)–(5.C) listed in Section 5.2 below. Similarly, we write $f \ll g$ if $f \le cg$ for some tiny constant $c > 0$ depending on the model parameters. We denote vectors by bold-faced lower case Roman letters $\mathbf{x}, \mathbf{y} \in \mathbb{C}^N$, and matrices by upper case Roman letters $A, B \in \mathbb{C}^{N\times N}$. We write $\langle A\rangle := N^{-1} \operatorname{Tr} A$ and $\langle \mathbf{x}\rangle := N^{-1} \sum_{a\in[N]} x_a$ for the averaged trace and the average of a vector. We often identify diagonal matrices with the vector of its diagonal elements. Accordingly, for any matrix R, we denote by $\operatorname{diag}(R)$ the vector of its diagonal elements, and for any vector $\mathbf{r}$ we denote by $\operatorname{diag}(\mathbf{r})$ the corresponding diagonal matrix.

We will frequently use the concept of "with very high probability" meaning that for any fixed $D > 0$ the probability of the event is bigger than $1 - N^{-D}$ if $N \ge N_0(D)$.

Acknowledgement. The authors are very grateful to Johannes Alt for his invaluable contribution in helping improve several results of [12] tailored to the needs of this paper.

5.2 Main results

For definiteness we consider the real symmetric case $H \in \mathbb{R}^{N\times N}$. With small modifications the proof presented in this paper works for complex Hermitian case as well, but this case was already considered in [DS5] with a contour integral analysis. Let $W = W^* \in \mathbb{R}^{N\times N}$ be a symmetric random matrix and $A = \operatorname{diag}(\boldsymbol{a})$ be a deterministic diagonal matrix with entries $\boldsymbol{a} = (a_i)_{i=1}^N \in \mathbb{R}^N$. We say that W is of *Wigner-type* [9] if its entries w_{ij} for $i \le j$ are centred, $\mathbf{E}\, w_{ij} = 0$, independent random variables. We define the *variance matrix* or *self-energy matrix* $S = (s_{ij})_{i,j=1}^N$, $s_{ij} := \mathbf{E}\, w_{ij}^2$. In [9] it was shown that as N tends to infinity, the resolvent $G(z) := (H - z)^{-1}$ of the *deformed Wigner-type matrix* $H = A + W$ entrywise approaches a diagonal matrix $M(z) := \operatorname{diag}(\mathbf{m}(z))$ for $z \in \mathbb{H} := \{\, z \in \mathbb{C} \mid \Im z > 0 \,\}$. The entries $\mathbf{m} = (m_1 \ldots, m_N) \colon \mathbb{H} \to \mathbb{H}^N$ of M have positive imaginary parts and solve the *Dyson equation*

$$-\frac{1}{m_i(z)} = z - a_i + \sum_{j=1}^{N} s_{ij} m_j(z), \qquad z \in \mathbb{H} := \{\, z \in \mathbb{C} \mid \Im z > 0 \,\}, \quad i \in [N]. \quad (5.1)$$

We call M or $\mathbf{m}$ the *self-consistent Green's function*. The normalised trace $\langle M \rangle$ of M is the Stieltjes transform $\langle M(z) \rangle = \int_{\mathbb{R}} (\tau - z)^{-1} \rho(\mathrm{d}\tau)$ of a unique probability measure ρ on $\mathbb{R}$ that approximates the empirical eigenvalue distribution of $A + W$ increasingly well as $N \to \infty$. We call ρ the *self-consistent density of states* (scDOS). Accordingly, its support $\operatorname{supp} \rho$ is called the *self-consistent spectrum*. It was proven in [7] that under very general conditions, $\rho(\mathrm{d}\tau)$ is an absolutely continuous measure with a $1/3$-Hölder continuous density, $\rho(\tau)$. Furthermore, the self-consistent spectrum consists of finitely many intervals with square root growth of ρ at the *edges*, i.e. at the points in $\partial \operatorname{supp} \rho$.

We call a point $\mathfrak{c} \in \mathbb{R}$ a cusp of ρ if $\mathfrak{c} \in \operatorname{int} \operatorname{supp} \rho$ and $\rho(\mathfrak{c}) = 0$. Cusps naturally emerge when we consider a one-parameter family of ensembles and two support intervals of ρ merge as the parameter value changes. The cusp universality phenomenon is not restricted to the exact cusp; it also occurs for situations shortly before and after the merging of two such support intervals, giving rise to a one parameter family of universal statistics. More precisely, universality emerges if ρ has a *physical cusp*. The terminology indicates that all these singularities become indistinguishable from the exact cusp if the density is resolved with a local precision above the typical eigenvalue spacing. We say that ρ exhibits a physical cusp if it has a small gap $(\mathfrak{e}_-, \mathfrak{e}_+) \subset \mathbb{R} \setminus \operatorname{supp} \rho$ with $\mathfrak{e}_+, \mathfrak{e}_- \in \operatorname{supp} \rho$ in its support of size $\mathfrak{e}_+ - \mathfrak{e}_- \lesssim N^{-3/4}$ or a local minimum $\mathfrak{m} \in \operatorname{int} \operatorname{supp} \rho$ of size $\rho(\mathfrak{m}) \lesssim N^{-1/4}$. Correspondingly, we call the points $\mathfrak{b} := \frac{1}{2}(\mathfrak{e}_+ + \mathfrak{e}_-)$ and $\mathfrak{b} := \mathfrak{m}$ *physical cusp points*, respectively.

Our main result is cusp universality under the real symmetric analogues of the assumptions of [DS5]. Throughout this paper we make the following three assumptions:

Assumption (5.A) (Bounded moments). *The entries of the matrix $\sqrt{N} W$ have bounded moments and the expectation A is bounded, i.e. there are positive C_k such that*

$$|a_i| \le C_0, \qquad \mathbf{E} \, |w_{ij}|^k \le C_k N^{-k/2}, \qquad k \in \mathbb{N}.$$

Assumption (5.B) (Flatness). *We assume that the matrix S is flat in the sense $s_{ij} = \mathbf{E} \, w_{ij}^2 \ge c/N$ for some constant $c > 0$.*

Assumption (5.C) (Bounded self-consistent Green's function). *The scDOS ρ has a physical cusp point $\mathfrak{b}$, and in a neighbourhood of the physical cusp point $\mathfrak{b} \in \mathbb{R}$ the self-consistent Green's function is bounded, i.e. for positive C, κ we have*

$$|m_i(z)| \le C, \qquad z \in [\mathfrak{b} - \kappa, \mathfrak{b} + \kappa] + \mathrm{i}\mathbb{R}^+.$$

We call the constants appearing in Assumptions (5.A)–(5.C) *model parameters*. All generic constants in this paper may implicitly depend on these model parameters. Dependence on further parameters, however, will be indicated.

Remark 5.2.1. *The boundedness of $\mathbf{m}$ in Assumption (5.C) can be, for example, ensured by assuming some regularity of the variance matrix S. For more details we refer to [7, Chapter 6].*

According to the extensive analysis in [7, 12] it follows[1] that there exists some small $\delta_* \sim 1$ such that the self-consistent density ρ around the points where it is small exhibits one of the following three types of behaviours.

(i) *Exact cusp.* There is a cusp point $\mathfrak{c} \in \mathbb{R}$ in the sense that $\rho(\mathfrak{c}) = 0$ and $\rho(\mathfrak{c} \pm \delta) > 0$ for $0 \neq \delta \ll 1$. In this case the self-consistent density is locally around $\mathfrak{c}$ given by

$$\rho(\mathfrak{c} + \omega) = \frac{\sqrt{3}\gamma^{4/3}\,|\omega|^{1/3}}{2\pi}\Big[1 + \mathcal{O}\Big(|\omega|^{1/3}\Big)\Big] \tag{5.2a}$$

for $\omega \in [-\delta_*, \delta_*]$ and some $\gamma > 0$.

(ii) *Small gap.* There is a maximal interval $[\mathfrak{e}_-, \mathfrak{e}_+]$ of size $0 < \Delta := \mathfrak{e}_+ - \mathfrak{e}_- \ll 1$ such that $\rho|_{[\mathfrak{e}_-,\mathfrak{e}_+]} \equiv 0$. In this case the density around $\mathfrak{e}_\pm$ is, for some $\gamma > 0$, locally given by

$$\rho(\mathfrak{e}_\pm \pm \omega) = \frac{\sqrt{3}(2\gamma)^{4/3}\Delta^{1/3}}{2\pi}\Psi_{\mathrm{edge}}(\omega/\Delta)\left[1 + \mathcal{O}\left(\min\Big\{\omega^{1/3}, \frac{\omega^{1/2}}{\Delta^{1/6}}\Big\}\right)\right] \tag{5.2b}$$

for $\omega \in [0, \delta_*]$, where

$$\Psi_{\mathrm{edge}}(\lambda) := \frac{\sqrt{\lambda(1+\lambda)}}{(1 + 2\lambda + 2\sqrt{\lambda(1+\lambda)})^{2/3} + (1 + 2\lambda - 2\sqrt{\lambda(1+\lambda)})^{2/3} + 1} \tag{5.2c}$$

for $\lambda \geq 0$.

(iii) *Non-zero local minimum.* There is a local minimum at $\mathfrak{m} \in \mathbb{R}$ of ρ such that $0 < \rho(\mathfrak{m}) \ll 1$. In this case there exists some $\gamma > 0$ such that

$$\begin{aligned}\rho(\mathfrak{m} + \omega) &= \rho(\mathfrak{m}) + \rho(\mathfrak{m})\Psi_{\min}\left(\frac{3\sqrt{3}\gamma^4\omega}{2(\pi\rho(\mathfrak{m}))^3}\right)\\ &\quad\times\left[1 + \mathcal{O}\left(\min\Big\{\rho(\mathfrak{m})^{1/2}, \frac{\rho(\mathfrak{m})^4}{|\omega|}\Big\} + \min\Big\{\frac{\omega^2}{\rho(\mathfrak{m})^5}, |\omega|^{1/3}\Big\}\right)\right]\end{aligned} \tag{5.2d}$$

for $\omega \in [-\delta_*, \delta_*]$, where

$$\Psi_{\min}(\lambda) := \frac{\sqrt{1+\lambda^2}}{(\sqrt{1+\lambda^2}+\lambda)^{2/3} + (\sqrt{1+\lambda^2}-\lambda)^{2/3} - 1} - 1, \qquad \lambda \in \mathbb{R}. \tag{5.2e}$$

We note that the choices for the *slope* parameter γ in (5.2b)–(5.2d) are consistent with (5.2a) in the sense that in the regimes $\Delta \ll \omega \ll 1$ and $\rho(\mathfrak{m})^3 \ll |\omega| \ll 1$ the respective formulae asymptotically agree. The precise form of the pre-factors in (5.2) is also chosen such that in the universality statement γ is a linear rescaling parameter.

It is natural to express universality in terms of a rescaled k-point function $p_k^{(N)}$ which we define implicitly by

$$\mathbf{E}\binom{N}{k}^{-1}\sum_{\{i_1,\dots,i_k\}\subset[N]} f(\lambda_{i_1},\dots,\lambda_{i_k}) = \int_{\mathbb{R}^k} f(\boldsymbol{x})p_k^{(N)}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{5.3}$$

for test functions f, where the summation is over all subsets of k distinct integers from $[N]$.

[1] The claimed expansions (5.2a) and (5.2d) follow directly from [12, Theorem 7.2(c), (d)]. The error term in (5.2b) follows from [12, Theorem 7.1(a)], where we define γ according to h therein.

Theorem 5.2.2. *Let H be a real symmetric or complex Hermitian deformed Wigner-type matrix whose scDOS ρ has a physical cusp point $\mathfrak{b}$ such that Assumptions (5.A)–(5.C) are satisfied. Let $\gamma > 0$ be the slope parameter at $\mathfrak{b}$, i.e. such that ρ is locally around $\mathfrak{b}$ given by (5.2). Then the local k-point correlation function at $\mathfrak{b}$ is universal, i.e. for any $k \in \mathbb{N}$ there exists a k-point correlation function $p_{k,\alpha}^{\mathrm{GOE/GUE}}$ such that for any smooth compactly supported test function $F\colon \mathbb{R}^k \to \mathbb{R}$ it holds that*

$$\int_{\mathbb{R}^k} F(\boldsymbol{x}) \left[\frac{N^{k/4}}{\gamma^k} p_k^{(N)} \left(\mathfrak{b} + \frac{\boldsymbol{x}}{\gamma N^{3/4}} \right) - p_{k,\alpha}^{\mathrm{GOE/GUE}}(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} = \mathcal{O}\left(N^{-c(k)} \right),$$

where the parameter α is given by

$$\alpha := \begin{cases} 0 & \textit{in case (i)} \\ 3\,(\gamma\Delta/4)^{2/3}\, N^{1/2} & \textit{in case (ii)} \\ -\left(\pi\rho(\mathfrak{m})/\gamma\right)^2 N^{1/2} & \textit{in case (iii)} \end{cases} \tag{5.4}$$

and $c(k) > 0$ is a small constant only depending on k.

Remark 5.2.3.

(i) In the complex Hermitian symmetry class the k-point function is given by

$$p_{k,\alpha}^{\mathrm{GUE}}(\boldsymbol{x}) = \det\Big(K_{\alpha,\alpha}(x_i, x_j) \Big)_{i,j=1}^k.$$

Here the extended Pearcey kernel *$K_{\alpha,\beta}$ is given by*

$$\begin{aligned} K_{\alpha,\beta}(x,y) &= -\frac{\mathbb{1}_{\beta>\alpha}}{\sqrt{2\pi(\beta-\alpha)}} \exp\Big(-\frac{(y-x)^2}{2(\beta-\alpha)} \Big) \\ &+ \frac{1}{(2\pi\mathrm{i})^2} \int_\Xi \mathrm{d}z \int_\Phi \mathrm{d}w \frac{\exp(-w^4/4 + \beta w^2/2 - yw + z^4/4 - \alpha z^2/2 + xz)}{w-z}, \end{aligned} \tag{5.5}$$

where Ξ is a contour consisting of rays from $\pm\infty e^{\mathrm{i}\pi/4}$ to 0 and rays from 0 to $\pm\infty e^{-\mathrm{i}\pi/4}$, and Φ is the ray from $-\mathrm{i}\infty$ to $\mathrm{i}\infty$. For more details we refer to [50, 172, 4] and the references in [DS5].

(ii) The real symmetric k-point function $p_{k,\alpha}^{\mathrm{GOE}}$ is not known explicitly. In fact, it is not even known whether $p_{k,\alpha}^{\mathrm{GOE}}$ is determinantal. We will nevertheless establish the existence of $p_{k,\alpha}^{\mathrm{GOE}}$ in Section 5.3 as the limit of the correlation functions of a one parameter family of Gaussian comparison models.

Theorem 5.2.2 is a universality result about the spatial correlations of eigenvalues. Our method also allows us to prove the corresponding statement on space-time universality when we consider the time evolution of eigenvalues $(\lambda_i^t)_{i\in[N]}$ according to the Dyson Brownian motion $\mathrm{d}H^{(t)} = \mathrm{d}\mathfrak{B}_t$ with initial condition $H^{(0)} = H$, where, depending on the symmetry class, $\mathfrak{B}_t$ is a complex Hermitian or real symmetric matrix valued Brownian motion. For any ordered k-tuple $\boldsymbol{\tau} = (\tau_1, \dots, \tau_k)$ with $0 \le \tau_1 \le \cdots \le \tau_k \lesssim N^{-1/2}$ we then define the *time-dependent k-point function* as follows. Denote the unique values in the tuple $\boldsymbol{\tau}$ by

$\sigma_1 < \cdots < \sigma_l$ such that $\{\tau_1, \ldots, \tau_k\} = \{\sigma_1, \ldots, \sigma_l\}$ and denote the multiplicity of σ_j in $\boldsymbol{\tau}$ by k_j and note that $\sum k_j = k$. We then define $p_{k,\boldsymbol{\tau}}^{(N)}$ implicitly via

$$\mathbf{E} \prod_{j=1}^{l} \left[\binom{N}{k_j}^{-1} \sum_{\{i_1^j,\ldots,i_{k_j}^j\} \subset [N]} \right] f(\lambda_{i_1^1}^{\sigma_1}, \ldots, \lambda_{i_{k_1}^1}^{\sigma_1}, \ldots, \lambda_{i_1^l}^{\sigma_l}, \ldots, \lambda_{i_{k_l}^l}^{\sigma_l}) = \int_{\mathbb{R}^k} f(\boldsymbol{x}) p_{k,\boldsymbol{\tau}}^{(N)}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \tag{5.6}$$

for test functions f and note that (5.6) reduces to (5.3) in the case $\tau_1 = \cdots = \tau_k = 0$. We note that in (5.6) coinciding indices are allowed only for eigenvalues at different times. If the scDOS ρ of H has a physical cusp in $\mathfrak{b}$, then for $\tau \lesssim N^{-1/2}$ the scDOS ρ_τ of $H^{(\tau)}$ also has a physical cusp $\mathfrak{b}_\tau$ close to $\mathfrak{b}$ and we can prove space-time universality in the sense of the following theorem, whose proof we defer to Appendix 5.A.

Theorem 5.2.4. *Let H be a real symmetric or complex Hermitian deformed Wigner-type matrix whose scDOS ρ has a physical cusp point $\mathfrak{b}$ such that Assumptions (5.A)–(5.C) are satisfied. Let $\gamma > 0$ be the slope parameter at $\mathfrak{b}$, i.e. such that ρ is locally around $\mathfrak{b}$ given by (5.2). Then there exists a k-point correlation function $p_{k,\boldsymbol{\alpha}}^{\mathrm{GOE/GUE}}$ such that for any $0 \le \tau_1 \le \cdots \le \tau_k \lesssim N^{-1/2}$ and any test function F it holds that*

$$\int_{\mathbb{R}^k} F(\boldsymbol{x}) \left[\frac{N^{k/4}}{\gamma^k} p_{k,\boldsymbol{\tau}/\gamma^2}^{(N)} \left(\mathfrak{b}_{\boldsymbol{\tau}/\gamma^2} + \frac{\boldsymbol{x}}{\gamma N^{3/4}} \right) - p_{k,\boldsymbol{\alpha}}^{\mathrm{GOE/GUE}}(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} = \mathcal{O}\left(N^{-c(k)}\right),$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k)$, $\mathfrak{b}_{\boldsymbol{\tau}} = (\mathfrak{b}_{\tau_1}, \ldots, \mathfrak{b}_{\tau_k})$ and $\boldsymbol{\alpha} = \alpha - \boldsymbol{\tau} N^{1/2}$ with α from (5.4) and $c(k) > 0$ is a small constant only depending on k. In the case of the complex Hermitian symmetry class the k-point correlation function is known to be determinantal of the form

$$p_{\alpha_1,\ldots,\alpha_k}^{\mathrm{GUE}}(\boldsymbol{x}) = \det\Big(K_{\alpha_i,\alpha_j}(x_i, x_j)\Big)_{i,j=1}^{k}$$

with $K_{\alpha,\beta}$ as in (5.5).

Remark 5.2.5. *The extended Pearcey kernel $K_{\alpha,\beta}$ in Theorem 5.2.4 has already been observed for the double-scaling limit of non-intersecting Brownian bridges [172, 4]. However, in the random matrix setting our methods also allow us to prove that the space-time universality of Theorem 5.2.4 extends beyond the Gaussian DBM flow. If the times $0 \le \tau_1 \le \cdots \le \tau_k \lesssim N^{-1/2}$ are ordered, then the k-point correlation function of the DBM flow asymptotically agrees with the k-point correlation function of eigenvalues of the matrices*

$$H + \sqrt{\tau_1} W_1, H + \sqrt{\tau_1} W_1 + \sqrt{\tau_2 - \tau_1} W_2, \ldots, H + \sqrt{\tau_1} W_1 + \cdots + \sqrt{\tau_k - \tau_{k-1}} W_k$$

for independent standard Wigner matrices $W_1, \ldots, W_k$.

5.3 Ornstein-Uhlenbeck flow

Starting from this section we consider a more general framework that allows for random matrix ensembles with certain correlation among the entries. In this way we stress that our proofs regarding the semicircular flow and the Dyson Brownian motion are largely model independent, assuming the optimal local law holds. The independence assumption on the entries of W is made only because we rely on the local law from [DS5] that was proven for deformed Wigner-type matrices. We therefore present the flow directly in the more general framework of the *matrix Dyson equation* (MDE)

$$1 + (z - A + \mathcal{S}[M(z)])M(z) = 0, \qquad A := \mathbf{E}\, H, \qquad \mathcal{S}[R] := \mathbf{E}\, WRW, \tag{5.7}$$

with spectral parameter in the complex upper half plane, $\Im z > 0$, and positive definite imaginary part, $\frac{1}{2\mathrm{i}}(M(z) - M(z)^*) > 0$, of the solution M. The MDE generalizes (5.1). Note that in the Wigner-type case the *self-energy operator* $\mathcal{S}\colon \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ is related to the variance matrix S by $\mathcal{S}[\operatorname{diag} \mathbf{r}] = \operatorname{diag}(S\mathbf{r})$.

As in [DS5] we consider the Ornstein-Uhlenbeck flow

$$\mathrm{d}\widetilde{H}_s = -\frac{1}{2}(\widetilde{H}_s - A)\,\mathrm{d}s + \Sigma^{1/2}[\mathrm{d}\mathfrak{B}_s], \quad \Sigma[R] := \frac{\beta}{2}\,\mathbf{E}\, W \operatorname{Tr} WR, \quad \widetilde{H}_0 := H, \tag{5.8}$$

which preserves expectation and self-energy operator $\mathcal{S}$. Since we consider real symmetric H, the parameter β indicating the symmetry class is $\beta = 1$. In (5.8) with $\mathfrak{B}_s \in \mathbb{R}^{N\times N}$ we denote a real symmetric matrix valued standard (GOE) Brownian motion, i.e. $(\mathfrak{B}_s)_{ij}$ for $i < j$ and $(\mathfrak{B}_s)_{ii}/\sqrt{2}$ are independent standard Brownian motions and $(\mathfrak{B}_s)_{ji} = (\mathfrak{B}_s)_{ij}$. In case H were complex Hermitian, we would have $\beta = 2$ and $\mathrm{d}\mathfrak{B}_s$ would be an infinitesimal GUE matrix. This was the setting in [DS5]. The OU flow effectively adds a small Gaussian component of size $\sqrt{s}$ to $\widetilde{H}_s$. More precisely, we can construct a Wigner-type matrix H_s, satisfying Assumptions (5.A)–(5.C), such that, for any fixed s,

$$\widetilde{H}_s = H_s + \sqrt{cs}U, \qquad \mathcal{S}_s = \mathcal{S} - cs\mathcal{S}^{\mathrm{GOE}}, \qquad \mathbf{E}\, H_s = A, \qquad U \sim \mathrm{GOE}, \tag{5.9}$$

where U is independent of H_s. Here $c > 0$ is a small universal constant which depends on the constant in Assumption (5.B), $\mathcal{S}_s$ is the self-energy operator corresponding to H_s and $\mathcal{S}^{\mathrm{GOE}}[R] := \langle R \rangle + R^t/N$. Since $\mathcal{S}$ is flat in the sense $\mathcal{S}[R] \gtrsim \langle R \rangle$ and s is small it follows that also $\mathcal{S}_s$ is flat.

As a consequence of the well established Green function comparison technique the k-point function of $H = \widetilde{H}_0$ is comparable with the one of $\widetilde{H}_s$ as long as $s \le N^{-1/4-\epsilon}$ for some $\epsilon > 0$. Indeed, from (4.115) for any compactly supported test function $F\colon \mathbb{R}^k \to \mathbb{R}$, we find

$$\int_{\mathbb{R}^k} F(\boldsymbol{x}) N^{k/4} \Big[p_k^{(N)}\Big(\mathfrak{b} + \frac{\boldsymbol{x}}{\gamma N^{3/4}}\Big) - \widetilde{p}_{k,s}^{(N)}\Big(\mathfrak{b} + \frac{\boldsymbol{x}}{\gamma N^{3/4}}\Big)\Big] \mathrm{d}\boldsymbol{x} = \mathcal{O}\left(N^{-c}\right), \tag{5.10}$$

where $\widetilde{p}_{k,s}^{(N)}$ is the k-point correlation function of $\widetilde{H}_s$, and $c = c(k) > 0$ is some constant.

It follows from the flatness assumption that the matrix H_s satisfies the assumptions of the local law from Theorem 4.2.5 uniformly in $s \ll 1$. Therefore Corollary 4.2.6 implies that the eigenvalues of H_s are rigid down to the optimal scale. It remains to prove that for long enough times s the local eigenvalue statistics of $H_s + \sqrt{cs}U$ on a scale of $1/\gamma N^{3/4}$ around $\mathfrak{b}$ agree with the local eigenvalue statistics of the Gaussian reference ensemble around 0 at a scale of $1/N^{3/4}$. By a simple rescaling Theorem 5.2.2 then follows from (5.10) together with the following Proposition.

Proposition 5.3.1. *Let $t_1 := N^{-1/2+\omega_1}$ with some small $\omega_1 > 0$ and let t_* be such that $|t_* - t_1| \lesssim N^{-1/2}$. Assume that $H^{(\lambda)}$ and $H^{(\mu)}$* [2] *are Wigner-type matrices satisfying Assumptions (5.A)–(5.C) such that the scDOSs $\rho_{\lambda,t_*}, \rho_{\mu,t_*}$ of $H^{(\lambda)} + \sqrt{t_*}U^{(\lambda)}$ and $H^{(\mu)} + \sqrt{t_*}U^{(\mu)}$*

[2] We use the notation $H^{(\lambda)}$ and $H^{(\mu)}$ since we denote the eigenvalues of $H^{(\lambda)}$ and $H^{(\mu)}$ by λ_i and μ_i respectively, with $1 \le i \le N$ respectively.

with independent $U^{(\lambda)}, U^{(\mu)} \sim$ GOE *have cusps in some points* $\mathfrak{c}_\lambda$, $\mathfrak{c}_\mu$ *such that locally around* $\mathfrak{c}_r$, $r = \lambda, \mu$, *the densities* ρ_{r,t_*} *are given by* (5.2a) *with* $\gamma = 1$. *Then the local* k*-point correlation functions* $p_{k,t_1}^{(N,r)}$ *of* $H^{(r)} + \sqrt{t_1}U^{(r)}$ *around the respective physical cusps* $\mathfrak{b}_{r,t_1}$ *of* ρ_{r,t_1}, $j = 1, 2$, *asymptotically agree in the sense*

$$\int_{\mathbb{R}^k} F(\boldsymbol{x}) \left[N^{k/4} p_{k,t_1}^{(N,\lambda)}\left(\mathfrak{b}_{\lambda,t_1} + \frac{\boldsymbol{x}}{N^{3/4}}\right) - N^{k/4} p_{k,t_1}^{(N,\mu)}\left(\mathfrak{b}_{\mu,t_1} + \frac{\boldsymbol{x}}{N^{3/4}}\right)\right] \mathrm{d}\boldsymbol{x} = \mathcal{O}\left(N^{-c(k)}\right).$$

Proof of Theorem 5.2.2. Set $s := t_1/c\theta^2$ and $H^{(\lambda)} := \theta H_s$ where c is the constant from (5.9) and $\theta \sim 1$ is yet to be chosen. Note that $H^{(\lambda)} + \sqrt{t}U = \theta(H_s + \sqrt{t/\theta^2}U)$, and in particular $H^{(\lambda)} + \sqrt{t_1}U = \widetilde{H}_s$. Moreover, it follows from the semicircular flow analysis in Section 5.4 that for some t_* with $|t_* - t_1| \lesssim N^{-1/2}$, the scDOS $\theta\rho_{\lambda,t_*}(\lambda\cdot)$ of $H_s + \sqrt{t_*/\theta^2}U$ and thereby also ρ_{λ,t_*}, the one of $H^{(\lambda)} + \sqrt{t_*}U$, have exact cusps in $\mathfrak{c}_\lambda/\theta$ and $\mathfrak{c}_\lambda$, respectively. It follows from the $1/3$-Hölder continuity of the slope parameter, cf. [12, Lemma 10.5, Eq. (7.5a)], that locally around $\mathfrak{c}_\lambda/\theta$ the scDOS of $H_s + \sqrt{t_*/\theta^2}U$ is given by

$$\theta\rho_{\lambda,t_*}(\mathfrak{c}_\lambda + \theta\omega) = \theta\rho_{\lambda,t_*}\Big(\theta\Big(\frac{\mathfrak{c}_\lambda}{\theta} + \omega\Big)\Big) = \frac{\sqrt{3}\gamma^{4/3}\left|\omega\right|^{1/3}}{2\pi}\Big[1 + \mathcal{O}\left(\left|\omega\right|^{1/3} + \left|t_* - t_1\right|^{1/3}\right)\Big].$$

Whence we can choose $\theta = \gamma[1 + \mathcal{O}(\left|t_1 - t_*\right|^{1/3})]$ appropriately such that

$$\rho_{\lambda,t_*}(\mathfrak{c}_\lambda + \omega) = \frac{\sqrt{3}\left|\omega\right|^{1/3}}{2\pi}\Big[1 + \mathcal{O}\left(\left|\omega\right|^{1/3}\right)\Big]$$

and it follows that $H^{(\lambda)}$ satisfies the assumptions of Proposition 5.3.1, in particular the slope parameter of $H^{(\lambda)} + \sqrt{t_*}U$ is normalized to 1. Furthermore, the almost cusp $\mathfrak{b}_{\lambda,t_1}$ of $H^{(\lambda)} + \sqrt{t_1}U$ is given by $\mathfrak{b}_{\lambda,t_1} = \theta\mathfrak{b}$ with $\mathfrak{b}$ as in Theorem 5.2.2.

We now choose our Gaussian comparison model. For $\alpha \in \mathbb{R}$ we consider the *reference ensemble*

$$U_\alpha = U_\alpha^{(N)} := \operatorname{diag}(1, \dots, 1, -1, \dots, -1) + \sqrt{1 - \alpha N^{-1/2}}U \in \mathbb{R}^{N\times N}, \qquad (5.11)$$

where $U \sim$ GOE, with $\lfloor N/2 \rfloor$ and $\lceil N/2 \rceil$ times ± 1 in the deterministic diagonal. An elementary computation shows that for even N and $\alpha = 0$, the self-consistent density of U_α has an exact cusp of slope $\gamma = 1$ in $\mathfrak{c} = 0$, i.e. it is given by (5.2a). For odd N the exact cusp is at distance $\lesssim N^{-1}$ away from 0 which is well below the natural scale of order $N^{-3/4}$ of the eigenvalue fluctuation and therefore has no influence on the k-point correlation function. The reference ensemble U_α has for $0 \neq |\alpha| \sim 1$ a small gap of size $N^{-3/4}$ or small local minimum of size $N^{-1/4}$ at the physical cusp point $|\mathfrak{b}| \lesssim \frac{1}{N}$, depending on the sign of α. Using the definition in (5.11), let $H^{(\mu)} := U_{N^{1/2}t_*}$ from which it follows that $H^{(\mu)} + \sqrt{t_*}U \sim U_0$ has an exact cusp in 0 whose slope is 1 by an easy explicit computation in the case of even N. For odd N the cusp emerges at a distance of $\lesssim N^{-1}$ away from 0, which is well below the investigated scale. Thus also $H^{(2)}$ satisfies the assumptions of Proposition 5.3.1. The almost cusp $\mathfrak{b}_{\mu,t_1}$ is given by $\mathfrak{b}_{\mu,t_1} = 0$ by symmetry of the density ρ_{μ,t_1} in the case of even N and at a distance of $|\mathfrak{b}_{\mu,t_1}| \lesssim N^{-1}$ in the case of odd N. This fact follows, for example, from explicitly solving the 2d-quadratic equation. The perturbation of size $1/N$ is not visible on the scale of the k-point correlation functions.

Now Proposition 5.3.1 together with (5.10) and $s \sim N^{-1/2+\omega_1}$ implies that

$$\int F(\boldsymbol{x}) \left[\frac{N^{k/4}}{\theta^k} p_k^{(N)} \left(\mathfrak{b} + \frac{\boldsymbol{x}}{\theta N^{3/4}} \right) - N^{k/4} p_{k,\alpha,\mathrm{GOE}}^{(N)} \left(\frac{\boldsymbol{x}}{N^{3/4}} \right) \right] \mathrm{d}\boldsymbol{x} = \mathcal{O}\left(N^{-c}\right) \qquad (5.12)$$

with $\alpha = N^{1/2}(t_* - t_1)$, where $p_{k,\alpha,\mathrm{GOE}}^{(N)}$ denotes the k-point function of the comparison model U_α. This completes the proof of Theorem 5.2.2 modulo the comparison of $p_{k,\alpha,\mathrm{GOE}}^{(N)}$ with its limit by relating $t_* - t_1$ to the size of the gap and the local minimum of ρ via Lemma 4.5.1 (or (5.17a)–(5.17c) later) and recalling that $\theta = \gamma[1 + \mathcal{O}(\,|t_1 - t_*|^{1/3}\,)]$.

To complete the proof we claim that

$$\int F(\boldsymbol{x}) \left[N^{k/4} p_{k,\alpha,\mathrm{GOE}}^{(N)} \left(\frac{\boldsymbol{x}}{N^{3/4}} \right) - p_{k,\alpha}^{\mathrm{GOE}}\left(\boldsymbol{x}\right) \right] \mathrm{d}\boldsymbol{x} = \mathcal{O}\left(N^{-c}\right). \qquad (5.13)$$

The proof of (5.13) is a straightforward consequence of the proof of (5.12). Although for notational simplicity we gave the proof for the case when H and U_α are of the same dimension, it works without any modification when their dimensions are only comparable, see Remark 5.5.2. This allows us to use (5.12) to compare, in a weak sense, the k-point correlation functions of $U_\alpha^{(N)}$ and $U_\alpha^{(\frac{4}{3}N)}$, say, with an effective error of order $N^{-c(k)}$. Applying this to a sequence of ensembles $U_\alpha^{(N_n)}$ with $N_n = (4/3)^n$, we find that $p_{k,\alpha,\mathrm{GOE}}^{(N_n)}$, tested against a fixed smooth function F, forms a Cauchy sequence in n. This proves the existence of the limit in (5.13). □

5.4 Semicircular flow analysis

In this section we analyse various properties of the semicircular flow in order to prepare the Dyson Brownian motion argument in Section 5.6 and Section 5.7. If ρ is a probability density on $\mathbb{R}$ with Stieltjes transform m, then the free semicircular evolution $\rho_t^{\mathrm{fc}} = \rho \boxplus \sqrt{t}\rho_{\mathrm{sc}}$ of ρ is defined as the unique probability measure whose Stieltjes transform m_t^{fc} solves the implicit equation

$$m_t^{\mathrm{fc}}(\zeta) = m(\zeta + t m_t^{\mathrm{fc}}(\zeta)), \qquad \zeta \in \mathbb{H}, \quad t \geq 0. \qquad (5.14)$$

Here $\sqrt{t}\rho_{\mathrm{sc}}$ is the semicircular distribution of variance t.

We now prepare the Dyson Brownian motion argument in Section 5.7 by providing a detailed analysis of the scDOS along the semicircular flow. As in Proposition 5.3.1 we consider the setting of two densities ρ_λ, ρ_μ whose semicircular evolutions reach a cusp of the same slope at the same time. Within the whole section we shall assume the following setup: Let ρ_λ, ρ_μ be densities associated with solutions M_λ, M_μ to some Dyson equations satisfying Assumptions (5.A)–(5.C) (or their matrix counterparts). We consider the free convolutions $\rho_{\lambda,t} := \rho_\lambda \boxplus \sqrt{t}\rho_{\mathrm{sc}}, \rho_{\mu,t} := \rho_\mu \boxplus \sqrt{t}\rho_{\mathrm{sc}}$ of ρ_λ, ρ_μ with semicircular distributions of variance t and assume that after a time $t_* \sim N^{-1/2+\omega_1}$ both densities $\rho_{\lambda,t_*}, \rho_{\mu,t_*}$ have cusps in points $\mathfrak{c}_\lambda, \mathfrak{c}_\mu$ around which they can be approximated by (5.2a) with the same $\gamma = \gamma_\lambda(t_*) = \gamma_\mu(t_*)$. It follows from the semicircular flow analysis in Lemma 4.5.1 that for $0 \leq t \leq t_*$ both densities have small gaps $[\mathfrak{e}_{r,t}^-, \mathfrak{e}_{r,t}^+]$, $r = \lambda, \mu$ in their supports, while for $t_* \leq t \leq 2t_*$ they have non-zero local minima in some points $\mathfrak{m}_{r,t}$, $r = \lambda, \mu$. We now define the concept of *interpolating densities* following [122, Section 3.1.1].

Definition 5.4.1. *For $\alpha \in [0,1]$ define the α-*interpolating density *$\rho_{\alpha,t}$ as follows. For any $0 \le E \le \delta_*$ and $r = \lambda, \mu$ let*

$$n_{r,t}(E) := \begin{cases} \int_{\mathfrak{e}_{r,t}^+}^{\mathfrak{e}_{r,t}^+ + E} \rho_{r,t}(\omega)\,\mathrm{d}\omega, & 0 \le t \le t_* \\ \int_{\mathfrak{m}_{r,t}}^{\mathfrak{m}_{r,t}+E} \rho_{r,t}(\omega)\,\mathrm{d}\omega, & t_* \le t \le 2t_* \end{cases}$$

be the counting functions and $\varphi_{\lambda,t}$, $\varphi_{\mu,t}$ their inverses, i.e. $n_{r,t}(\varphi_{r,t}(s)) = s$. Define now

$$\varphi_{\alpha,t}(s) := \alpha\varphi_{\lambda,t}(s) + (1-\alpha)\varphi_{\mu,t}(s) \tag{5.15}$$

for $s \in [0, \delta_{**}]$ where $\delta_{**} \sim 1$ depends on δ_* and is chosen in such a way that $\varphi_{\alpha,t}$ is invertible[3]*. We thus define $n_{\alpha,t}(E)$ to be the inverse of $\varphi_{\alpha,t}(s)$ near zero. Furthermore, for $0 \le t \le t_*$ set*

$$\begin{aligned} \mathfrak{e}_{\alpha,t}^{\pm} &:= \alpha\mathfrak{e}_{\lambda,t}^{\pm} + (1-\alpha)\mathfrak{e}_{\mu,t}^{\pm}, \\ \rho_{\alpha,t}(\mathfrak{e}_{\alpha,t}^+ + E) &:= \frac{\mathrm{d}}{\mathrm{d}E} n_{\alpha,t}(E), \quad E \in [0,\delta_*] \end{aligned} \tag{5.16}$$

and for $t \ge t_*$ set

$$\begin{aligned} \mathfrak{m}_{\alpha,t} &:= \alpha\mathfrak{m}_{\lambda,t} + (1-\alpha)\mathfrak{m}_{\mu,t}, \\ \rho_{\alpha,t}(\mathfrak{m}_{\alpha,t} + E) &:= \alpha\rho_{\lambda,t}(\mathfrak{m}_{\lambda,t}) + (1-\alpha)\rho_{\mu,t}(\mathfrak{m}_{\mu,t}) + \frac{\mathrm{d}}{\mathrm{d}E} n_{\alpha,t}(E), \quad E \in [-\delta_*, \delta_*]. \end{aligned}$$

We define $\rho_{\alpha,t}(E)$ for $0 \le t \le t_*$ and $E \in [\mathfrak{e}_{\alpha,t}^- - \delta_*, \mathfrak{e}_{\alpha,t}^-]$ analogously.

The motivation for the interpolation mode in Definition 5.4.1 is that (5.15) ensures that the quantiles of $\rho_{\alpha,t}$ are the convex combination of the quantiles of $\rho_{\lambda,t}$ and $\rho_{\mu,t}$, see (5.26c) later. The following two lemmas collect various properties of the interpolating density. Recall that $\rho_{\lambda,t}$ and $\rho_{\mu,t}$ are asymptotically close near the cusp regime, up to a trivial shift, since they develop a cusp with the same slope at the same time. In Lemma 5.4.2 we show that $\rho_{\alpha,t}$ shares this property. Lemma 5.4.3 shows that $\rho_{\alpha,t}$ inherits the regularity properties of $\rho_{\lambda,t}$ and $\rho_{\mu,t}$ from [12].

Lemma 5.4.2 (Size of gaps and minima along the flow). *For $t \le t_*$ and $r = \alpha, \lambda, \mu$ the supports of $\rho_{r,t}$ have small gaps $[\mathfrak{e}_{r,t}^-, \mathfrak{e}_{r,t}^+]$ near $\mathfrak{c}_*$ of size*

$$\begin{aligned} \Delta_{r,t} &:= \mathfrak{e}_{r,t}^+ - \mathfrak{e}_{r,t}^- = (2\gamma)^2\Big(\frac{t_*-t}{3}\Big)^{3/2}\big[1 + \mathcal{O}((t_*-t)^{1/3})\big], \\ \Delta_{r,t} &= \Delta_{\mu,t}\big[1 + \mathcal{O}((t_*-t)^{1/3})\big] \end{aligned} \tag{5.17a}$$

and the densities are close in the sense

$$\rho_{r,t}(\mathfrak{e}_{r,t}^{\pm} \pm \omega) = \rho_{\mu,t}(\mathfrak{e}_{\mu,t}^{\pm} \pm \omega)\Big[1 + \mathcal{O}\Big((t_*-t)^{1/3} + \min\Big\{\omega^{1/3}, \frac{\omega^{1/2}}{(t_*-t)^{1/4}}\Big\}\Big)\Big] \tag{5.17b}$$

[3] Invertibility in a small neighbourhood follows from the form of the explicit shape functions in (5.2b) and (5.2d).

for $0 \leq \omega \leq \delta_*$. *For* $t_* < t \leq 2t_*$ *the densities* $\rho_{r,t}$ *have small local minima* $\mathfrak{m}_{r,t}$ *of size*

$$\begin{aligned} \rho_{r,t}(\mathfrak{m}_{r,t}) &= \frac{\gamma^2\sqrt{t-t_*}}{\pi}\big[1+\mathcal{O}((t-t_*)^{1/2})\big], \\ \rho_{r,t}(\mathfrak{m}_{r,t}) &= \rho_{\mu,t}(\mathfrak{m}_{\mu,t})\big[1+\mathcal{O}((t-t_*)^{1/2})\big] \end{aligned} \tag{5.17c}$$

and the densities are close in the sense

$$\begin{aligned} &\frac{\rho_{r,t}(\mathfrak{m}_{r,t}+\omega)}{\rho_{\mu,t}(\mathfrak{m}_{\mu,t}+\omega)} - 1 \\ &= \mathcal{O}\Big((t-t_*)^{1/2} + \min\Big\{(t-t_*)^{1/4}, \frac{(t-t_*)^2}{|\omega|}\Big\} + \min\Big\{\frac{\omega^2}{(t-t_*)^{5/2}}, |\omega|^{1/3}\Big\}\Big) \end{aligned} \tag{5.17d}$$

for $\omega \in [-\delta_*, \delta_*]$. *Here* $\delta_*, \delta_{**} \sim 1$ *are small constants depending on the model parameters in Assumptions (5.A)–(5.C).*

Lemma 5.4.3. *The density* $\rho_{\alpha,t}$ *from Definition 5.4.1 is well defined and is a* $1/3$*-Hölder continuous density. More precisely, in the pre-cusp regime, i.e. for* $t \leq t_*$, *we have*

$$\Big|\rho'_{\alpha,t}(\mathfrak{e}^{\pm}_{\alpha,t} \pm x)\Big| \lesssim \frac{1}{\rho_{\alpha,t}(\mathfrak{e}^{\pm}_{\alpha,t} \pm x)\Big(\rho_{\alpha,t}(\mathfrak{e}^{\pm}_{\alpha,t} \pm x) + \Delta^{1/3}_{\alpha,t}\Big)} \tag{5.18a}$$

for $0 \leq x \leq \delta_*$. *Moreover, the Stieltjes transform* $m_{\alpha,t}$ *satisfies the bounds*

$$\begin{aligned} \Big|m_{\alpha,t}(\mathfrak{e}^{\pm}_{\alpha,t} \pm x)\Big| &\lesssim 1, \\ \Big|m_{\alpha,t}(\mathfrak{e}^{\pm}_{\alpha,t} \pm (x+y)) - m_{\alpha,t}(\mathfrak{e}^{\pm}_{\alpha,t} \pm x)\Big| &\lesssim \frac{|y|\,|\log|y||}{\rho_{\alpha,t}(\mathfrak{e}^{\pm}_{\alpha,t} \pm x)(\rho_{\alpha,t}(\mathfrak{e}^{\pm}_{\alpha,t} \pm x) + \Delta^{1/3}_{\alpha,t})} \end{aligned} \tag{5.18b}$$

for $|x| \leq \delta_*/2$, $|y| \ll x$. *In the small minimum case, i.e. for* $t \geq t_*$, *we similarly have*

$$\Big|\rho'_{\alpha,t}(\mathfrak{m}_{\alpha,t}+x)\Big| \lesssim \frac{1}{\rho^2_{\alpha,t}(\mathfrak{m}_{\alpha,t}+x)} \tag{5.19a}$$

for $|x| \leq \delta_*$ *and*

$$\begin{aligned} |m_{\alpha,t}(\mathfrak{m}_{\alpha,t}+x)| &\lesssim 1, \\ |m_{\alpha,t}(\mathfrak{m}_{\alpha,t}+(x+y)) - m_{\alpha,t}(\mathfrak{m}_{\alpha,t}+x)| &\lesssim \frac{|y|\,|\log|y||}{\rho^2_{\alpha,t}(\mathfrak{m}_{\alpha,t}+x)} \end{aligned} \tag{5.19b}$$

for $|x| \leq \delta_*$ *and* $|y| \ll |x|$.

Proof of Lemma 5.4.2. We first consider the two densities $r = \lambda, \mu$ only. The first claims in (5.17a) and (5.17c) follow directly from Lemma 4.5.1, while the second claims follow immediately from the first ones. For the proof of (5.17b) and (5.17d) we first note that by elementary calculus

$$\Psi_{\text{edge}}((1+\epsilon)\lambda) = \Psi_{\text{edge}}(\lambda)\big[1+\mathcal{O}(\epsilon)\big], \qquad \Psi_{\min}((1+\epsilon)\lambda) = \Psi_{\min}(\lambda)\big[1+\mathcal{O}(\epsilon)\big]$$

so that

$$\Delta^{1/3}_{\lambda,t}\Psi_{\text{edge}}(\omega/\Delta_{\lambda,t}) = \Delta^{1/3}_{\mu,t}\Psi_{\text{edge}}(\omega/\Delta_{\mu,t})\Big[1+\mathcal{O}\Big((t_*-t)^{1/3}\Big)\Big]$$

and the claimed approximations follow together with (5.2b) and (5.2d). Here the exact cusp case $t = t_*$ is also covered by interpreting $0^{1/3}\Psi_{\mathrm{edge}}(\omega/0) = \omega^{1/3}/2^{4/3}$.

In order to prove the corresponding statements for the interpolating densities $\rho_{\alpha,t}$, we first have to establish a quantitative understanding of the counting function $n_{r,t}$ and its inverse. We claim that for $r = \alpha, \lambda, \mu$ they satisfy for $0 \le E \le \delta_*$, $0 \le s \le \delta_{**}$ that

$$\begin{aligned} n_{r,t}(E) &\sim \min\left\{\frac{E^{3/2}}{\Delta_{r,t}^{1/6}}, E^{4/3}\right\}, \qquad \varphi_{r,t}(s) \sim \max\left\{s^{3/4}, s^{2/3}\Delta_{r,t}^{1/9}\right\}, \\ \frac{\varphi_{r,t}(s)}{\varphi_{\lambda,t}(s)} &\sim \min\left\{\varphi_{\lambda,t}^{1/3}(s), \frac{\varphi_{\lambda,t}^{1/2}(s)}{\Delta_{\lambda,t}^{1/6}}\right\} \end{aligned} \tag{5.20a}$$

for $t \le t_*$ and

$$\begin{aligned} n_{r,t}(E) &\sim \max\{E^{4/3}, E\rho_{r,t}(\mathfrak{m}_{r,t})\}, \qquad \varphi_{r,t}(s) \sim \min\left\{s^{3/4}, \frac{s}{\rho_{r,t}(\mathfrak{m}_{r,t})}\right\} \\ \frac{\varphi_{r,t}(s)}{\varphi_{\lambda,t}(s)} &\sim \min\left\{\varphi_{\lambda,t}^{1/3}(s), \frac{\varphi_{\lambda,t}(s)}{\rho_{r,t}^{2}(\mathfrak{m}_{r,t})}, \frac{\varphi_{\lambda,t}^{2}(s)}{\rho_{r,t}^{11/2}(\mathfrak{m}_{r,t})}\right\} \end{aligned} \tag{5.20b}$$

for $t \ge t_*$.

Proof of (5.20). We begin with the proof of (5.20a) for $r = \lambda, \mu$. Recall that the shape function Ψ_{edge} satisfies the scaling $\Delta^{1/3}\Psi_{\mathrm{edge}}(\omega/\Delta) \sim \min\{\omega^{1/3}, \omega^{1/2}/\Delta^{1/6}\}$. We first find by elementary integration that

$$\int_0^q \min\Big\{\omega^{1/3}, \frac{\omega^{1/2}}{\Delta^{1/6}}\Big\}\,\mathrm{d}\omega = \frac{9q^{4/3}\min\{q,\Delta\}^{1/6} - \min\{q,\Delta\}^{3/2}}{12\Delta^{1/6}} \sim \min\Big\{\frac{q^{3/2}}{\Delta^{1/6}}, q^{4/3}\Big\}$$

from which we conclude the first relation in (5.20a), and by inversion also the second relation. Together with the estimate for the error integral for $\rho_{\lambda,t}(\mathfrak{e}_{\lambda,t}^{+} + \omega) - \rho_{\mu,t}(\mathfrak{e}_{\mu,t}^{+} + \omega) \lesssim \min\{\omega^{2/3}, \omega/\Delta_{\lambda,t}^{1/3}\}$,

$$\int_0^q \min\Big\{\omega^{2/3}, \frac{\omega}{\Delta^{1/3}}\Big\}\,\mathrm{d}\omega = \frac{6q^{5/3}\min\{q,\Delta\}^{1/3} - \min\{q,\Delta\}^{2}}{10\Delta^{1/3}} \sim \min\Big\{\frac{q^{2}}{\Delta^{1/3}}, q^{5/3}\Big\}$$

we can thus conclude also the third relation in (5.20a).

We now turn to the case $t > t_*$ where both densities $\rho_{\lambda,t}, \rho_{\mu,t}$ exhibit a small local minimum. We first record the elementary integral

$$\begin{aligned} \int_0^q \Big(\rho + \min\Big\{\omega^{1/3}, \frac{\omega^2}{\rho^5}\Big\}\Big)\,\mathrm{d}\omega &= \frac{q^{4/3}\min\{\rho^3, q\}^{5/3} + 12q\rho^6 - 5\min\{q,\rho^3\}^3}{12\rho^5} \\ &\sim \max\{q^{4/3}, q\rho\} \end{aligned}$$

for $q, \rho \ge 0$ and easily conclude the first two relation in (5.20b). For the error integral we obtain

$$\int_0^q \min\Big\{\omega^{1/3}, \frac{\omega^2}{\rho^5}\Big\}\Big[\min\Big\{\rho^{1/2}, \frac{\rho^4}{\omega}\Big\} + \min\Big\{\omega^{1/3}, \frac{\omega^2}{\rho^5}\Big\}\Big]\,\mathrm{d}\omega \sim \min\Big\{q^{5/3}, \frac{q^2}{\rho}, \frac{q^3}{\rho^{9/2}}\Big\}$$

from which the third relation in (5.20b) follows. Finally, the claims (5.20a) and (5.20b) for $r = \alpha$ follow immediately from Definition 5.4.1 and the corresponding statements for $r = \lambda, \mu$. This completes the proof of (5.20). □

We now turn to the density $\rho_{\alpha,t}$ for which the claims (5.17a), (5.17c) follow immediately from Definition 5.4.1 and the corresponding statements for $\rho_{\lambda,t}$ and $\rho_{\mu,t}$. For $t \leq t_*$ we now continue by differentiating $E = \varphi_{r,t}(n_{r,t}(E))$ to obtain

$$\begin{aligned}\rho_{\alpha,t}(\mathfrak{e}_{\alpha,t}^+ + \varphi_{\alpha,t}(s)) &= \frac{1}{\varphi'_{\alpha,t}(s)} = \frac{1}{\alpha\varphi'_{\lambda,t}(s) + (1-\alpha)\varphi'_{\mu,t}(s)} \\ &= \left(\frac{\alpha}{\rho_{\lambda,t}(\mathfrak{e}_{\lambda,t}^+ + \varphi_{\lambda,t}(s))} + \frac{1-\alpha}{\rho_{\mu,t}(\mathfrak{e}_{\mu,t}^+ + \varphi_{\mu,t}(s))}\right)^{-1} \\ &= \rho_{\lambda,t}(\mathfrak{e}_{\lambda,t}^+ + \varphi_{\lambda,t}(s))\left(\alpha + (1-\alpha)\frac{\rho_{\lambda,t}(\mathfrak{e}_{\lambda,t}^+ + \varphi_{\lambda,t}(s))}{\rho_{\mu,t}(\mathfrak{e}_{\mu,t}^+ + \varphi_{\mu,t}(s))}\right)^{-1},\end{aligned} \tag{5.21}$$

from which we can easily conclude (5.17b) for $r = \alpha$ together with (5.17b) for $r = \lambda$ and (5.20a). The proof of (5.17d) for $r = \alpha$ follows by the same argument and replacing $\mathfrak{e}_{r,t}^+$ by $\mathfrak{m}_{r,t}$. This finishes the proof of Lemma 5.4.2 □

Proof of Lemma 5.4.3. By differentiating we find

$$\begin{aligned}\frac{\rho'_{\alpha,t}(\mathfrak{e}_{\alpha,t}^+ + \varphi_{\alpha,t}(s))}{\rho_{\alpha,t}(\mathfrak{e}_{\alpha,t}^+ + \varphi_{\alpha,t}(s))} &= -\frac{\alpha\varphi''_{\lambda,t}(s) + (1-\alpha)\varphi''_{\mu,t}(s)}{\left(\alpha\varphi'_{\lambda,t}(s) + (1-\alpha)\varphi'_{\mu,t}(s)\right)^2} \\ &= \left[\alpha\frac{\rho'_{\lambda,t}(\mathfrak{e}_{\lambda,t}^+ + \varphi_{\lambda,t}(s))}{\rho^3_{\lambda,t}(\mathfrak{e}_{\lambda,t}^+ + \varphi_{\lambda,t}(s))} + (1-\alpha)\frac{\rho'_{\mu,t}(\mathfrak{e}_{\mu,t}^+ + \varphi_{\mu,t}(s))}{\rho^3_{\mu,t}(\mathfrak{e}_{\mu,t}^+ + \varphi_{\mu,t}(s))}\right] \\ &\quad\times\left(\frac{\alpha}{\rho_{\lambda,t}(\mathfrak{e}_{\lambda,t}^+ + \varphi_{\lambda,t}(s))} + \frac{1-\alpha}{\rho_{\mu,t}(\mathfrak{e}_{\mu,t}^+ + \varphi_{\mu,t}(s))}\right)^{-2},\end{aligned}$$

from which we conclude the claimed bound (5.18a) together with the fact that the densities ρ_λ and ρ_μ fulfil the same bound according to [12, Remark 10.7], and the estimates from Lemma 5.4.2. Similarly, the bound in (5.19a) follows by the same argument by replacing $\mathfrak{e}_{\alpha,t}^\pm$ by $\mathfrak{m}_{\alpha,t}$. The bound $|\rho'| \leq \rho^{-2}$ on the derivative implies $\frac{1}{3}$-Hölder continuity.

We now turn to the claimed bound on the Stieltjes transform and compute

$$m_{\alpha,t}(\mathfrak{e}_{\alpha,t}^+ + x) = \int_0^{\delta_*}\frac{\rho_{\alpha,t}(\mathfrak{e}_{\alpha,t}^+ + \omega)}{\omega - x}\,\mathrm{d}\omega + \int_{-\delta_*}^0\frac{\rho_{\alpha,t}(\mathfrak{e}_{\alpha,t}^- + \omega)}{\omega - \Delta_{\alpha,t} - x}\,\mathrm{d}\omega,$$

out of which for $x > 0$ the first term can be bounded by

$$\begin{aligned}\int_0^{\delta_*}\frac{\rho_{\alpha,t}(\mathfrak{e}_{\alpha,t}^+ + \omega)}{\omega - x}\,\mathrm{d}\omega &\lesssim \int_0^{\delta_*}\frac{|\omega - x|^{1/3}}{\omega - x}\,\mathrm{d}\omega + \int_{2x}^{\delta_*}\frac{\rho_{\alpha,t}(\mathfrak{e}_{\alpha,t}^+ + x)}{\omega - x}\,\mathrm{d}\omega \\ &\lesssim |x|^{1/3}\,|\log x| + |\delta_* - x|^{1/3},\end{aligned}$$

while the second term can be bounded by

$$\left|\int_{-\delta_*}^0\frac{\rho_{\alpha,t}(\mathfrak{e}_{\alpha,t}^- + \omega)}{\omega - \Delta_{\alpha,t} - x}\,\mathrm{d}\omega\right| \lesssim |\delta_* - \Delta_{\alpha,t} - x|^{1/3} + |\Delta_{\alpha,t} + x|^{1/3}\,|\log(\Delta_{\alpha,t} + x)|,$$

both using the $1/3$-Hölder continuity of $\rho_{\alpha,t}$. The corresponding bounds for $x < 0$ are similar, completing the proof of the first bound in (5.18b).

The proof of the first bound in (5.19b) is very similar and follows from

$$\begin{aligned}|m_{\alpha,t}(\mathfrak{m}_{\alpha,t}+x)| &\lesssim \left|\int_{-\delta_*}^{\delta_*}\frac{|\omega-x|^{1/3}}{\omega-x}\,\mathrm{d}\omega\right| + \left|\int_{[-\delta_*,\delta_*]\setminus[x-\delta_*/2,x+\delta_*/2]}\frac{\rho_{\alpha,t}(\mathfrak{m}_{\alpha,t}+x)}{\omega-x}\,\mathrm{d}\omega\right|\\ &\lesssim 1.\end{aligned}$$

We now turn to the second bound in (5.18b) which is only non-trivial in the case $x > 0$. To simplify the following integrals we temporarily use the short-hand notations $m = m_{\alpha,t}$, $\mathfrak{e}^+ = \mathfrak{e}^+_{\alpha,t}$, $\rho = \rho_{\alpha,t}$, $\Delta = \Delta_{\alpha,t}$ and compute

$$m(\mathfrak{e}^+ + x + y) - m(\mathfrak{e}^+ + x) = \int_{-\Delta-\delta_*}^{\delta_*}\frac{\rho(\mathfrak{e}^+ + \omega)}{\omega - x - y}\,\mathrm{d}\omega - \int_{-\Delta-\delta_*}^{\delta_*}\frac{\rho(\mathfrak{e}^+ + \omega)}{\omega - x}\,\mathrm{d}\omega$$

where we now focus on the integration regime $\omega \geq 0$ as this is the regime containing the two critical singularities. We first observe that

$$\begin{aligned}&\int_{-y}^{\delta_*-y}\frac{\rho(\mathfrak{e}^+ + \omega + y)}{\omega - x}\,\mathrm{d}\omega - \int_0^{\delta_*}\frac{\rho(\mathfrak{e}^+ + \omega)}{\omega - x}\,\mathrm{d}\omega\\ &\quad= \int_0^{\delta_*}\frac{\rho(\mathfrak{e}^+ + \omega + y) - \rho(\mathfrak{e}^+ + \omega)}{\omega - x}\,\mathrm{d}\omega + \int_{-y}^{0}\frac{\rho(\mathfrak{e}^+ + \omega + y)}{\omega - x}\,\mathrm{d}\omega + \mathcal{O}(y),\end{aligned}$$

where the second integral is easily bounded by

$$\int_{-y}^{0}\frac{\rho(\mathfrak{e} + \omega + y)}{\omega - x}\,\mathrm{d}\omega \lesssim \frac{1}{x}\min\left\{y^{4/3}, y^{3/2}\Delta^{-1/6}\right\} \lesssim \frac{y}{\rho(\mathfrak{e}^+ + x)(\rho(\mathfrak{e}^+ + x) + \Delta^{1/3})}.$$

We split the remaining integral into three regimes $[0, x/2]$, $[x/2, 3x/2]$ and $[3x/2, \delta_*]$. In the first one we use (5.18a) as well as the scaling relation $\rho(\mathfrak{e}^+ + \omega) \sim \min\{\omega^{1/3}, \omega^{1/2}\Delta^{-1/6}\}$ to obtain

$$\begin{aligned}&\int_0^{x/2}\frac{\rho(\mathfrak{e}^+ + \omega + y) - \rho(\mathfrak{e}^+ + \omega)}{\omega - x}\,\mathrm{d}\omega \lesssim \frac{y}{x}\int_0^{x/2}\frac{1}{\rho(\mathfrak{e}^+ + \omega)\Big(\rho(\mathfrak{e}^+ + \omega) + \Delta^{1/3}\Big)}\,\mathrm{d}\omega\\ &\lesssim \frac{y}{x}\min\left\{\frac{x^{1/2}}{\Delta^{1/6}}, x^{1/3}\right\} \sim \frac{y}{\max\{x^{2/3}, x^{1/2}\Delta^{1/6}\}} \lesssim \frac{y}{\rho(\mathfrak{e}^+ + x)(\rho(\mathfrak{e}^+ + x) + \Delta^{1/3})}.\end{aligned}$$

The integral in the regime $[3x/2, \delta_*]$ is completely analogous and contributes the same bound. Finally, we are left with the regime $[x/2, 3x/2]$ which we again subdivide into $[x - y, x + y]$ and $[x/2, 3x/2] \setminus [x - y, x + y]$. In the first of those we have

$$\begin{aligned}&\int_{x-y}^{x+y}\frac{\rho(\mathfrak{e}^+ + \omega + y) - \rho(\mathfrak{e}^+ + \omega)}{\omega - x}\,\mathrm{d}\omega\\ &\quad= \int_{x-y}^{x+y}\frac{\rho(\mathfrak{e}^+ + \omega + y) - \rho(\mathfrak{e}^+ + x + y) - \rho(\mathfrak{e}^+ + \omega) + \rho(\mathfrak{e}^+ + x)}{\omega - x}\,\mathrm{d}\omega\\ &\quad\lesssim \frac{y}{\rho(\mathfrak{e}^+ + x)(\rho(\mathfrak{e}^+ + x) + \Delta^{1/3})},\end{aligned}$$

while in the second one we obtain

$$\begin{aligned}\int_{[x/2,3x/2]\setminus[x-y,x+y]}&\frac{\rho(\mathfrak{e}^+ +\omega+y)-\rho(\mathfrak{e}^+ +x+y)-\rho(\mathfrak{e}^+ +\omega)+\rho(\mathfrak{e}^+ +x)}{\omega-x}\,\mathrm{d}\omega\\ &\lesssim\frac{y}{\rho(\mathfrak{e}^+ +x)(\rho(\mathfrak{e}^+ +x)+\Delta^{1/3})}\int_{[x/2,3x/2]\setminus[x-y,x+y]}|\omega-x|^{-1}\,\mathrm{d}\omega\\ &\lesssim\frac{y\,|\log y|}{\rho(\mathfrak{e}^+ +x)(\rho(\mathfrak{e}^+ +x)+\Delta^{1/3})}.\end{aligned}$$

Collecting the various estimates completes the proof of (5.18b).

The second bound in (5.19b) follows by a similar argument and we focus on the most critical term

$$\begin{aligned}&\int_{-\delta_*/2}^{\delta_*/2}\frac{\rho(\mathfrak{m}+\omega+y)-\rho(\mathfrak{m}+\omega)}{\omega-x}\,\mathrm{d}\omega\\ &\quad=\Big(\int_{-\delta_*/2}^{x-y}+\int_{x-y}^{x+y}+\int_{x+y}^{\delta_*/2}\Big)\frac{\rho(\mathfrak{m}+\omega+y)-\rho(\mathfrak{m}+\omega)}{\omega-x}\,\mathrm{d}\omega.\end{aligned}$$

Here we can bound the middle integral by

$$\begin{aligned}&\Big|\int_{x-y}^{x+y}\frac{\rho(\mathfrak{m}+\omega+y)-\rho(\mathfrak{m}+\omega)}{\omega-x}\,\mathrm{d}\omega\Big|\\ &\quad=\Big|\int_{x-y}^{x+y}\frac{\rho(\mathfrak{m}+\omega+y)-\rho(\mathfrak{m}+x+y)-\rho(\mathfrak{m}+\omega)+\rho(\mathfrak{m}+x)}{\omega-x}\,\mathrm{d}\omega\Big|\lesssim\frac{|y|}{\rho^2(\mathfrak{m}+x)},\end{aligned}$$

while for the first integral we have

$$\begin{aligned}&\Big|\int_{-\delta_*/2}^{x-y}\frac{\rho(\mathfrak{m}+\omega+y)-\rho(\mathfrak{m}+x+y)-\rho(\mathfrak{m}+\omega)+\rho(\mathfrak{m}+x)}{\omega-x}\,\mathrm{d}\omega\Big|\\ &\quad\lesssim\frac{|y|}{\rho^2(\mathfrak{m}+x)}\int_{-\delta_*/2}^{x-y}\frac{1}{|\omega-x|}\,\mathrm{d}\omega\lesssim\frac{|y|\,|\log|y||}{\rho^2(\mathfrak{m}+x)}.\end{aligned}$$

The third integral is completely analogous, completing the proof of (5.19b). □

5.4.1 Movement of edges and minima

For the analysis of the Dyson Brownian motion it is necessary to have a precise understanding of the movement of the reference points $\mathfrak{e}_{r,t}^{\pm}$ and $\mathfrak{m}_{r,t}$, $r=\lambda,\mu$. For technical reasons it is slightly easier to work with an auxiliary quantity $\widetilde{\mathfrak{m}}_{r,t}$ which is very close to $\mathfrak{m}_{r,t}$. According to Lemma 4.5.1 the minimum $\mathfrak{m}_{r,t}$ can approximately be found by solving the implicit equation

$$\widetilde{\mathfrak{m}}_{r,t}=\mathfrak{c}_r-(t-t_*)\Re m_{r,t}(\widetilde{\mathfrak{m}}_{r,t}),\qquad\widetilde{\mathfrak{m}}_{r,t}\in\mathbb{R},\quad r=\lambda,\mu.\tag{5.22a}$$

The explicit relation (5.22a) is the main reason why it is more convenient to study the movement of $\widetilde{\mathfrak{m}}_t$ rather than the one of $\mathfrak{m}_t$. We claim that $\widetilde{\mathfrak{m}}_{r,t}$ is indeed a very good approximation for $\mathfrak{m}_{r,t}$ in the sense that

$$|\mathfrak{m}_{r,t}-\widetilde{\mathfrak{m}}_{r,t}|\lesssim(t-t_*)^{3/2+1/4},\quad\Im m_{r,t}(\widetilde{\mathfrak{m}}_{r,t})=\gamma^2(t-t_*)^{1/2}+\mathcal{O}\left(t-t_*\right)\tag{5.22b}$$

for $r=\lambda,\mu$.

Proof of (5.22b). The first claim in (5.22b) is a direct consequence of Lemma 4.5.1. For the second claim we refer to (4.90) which implies

$$\begin{aligned}\Im m_{r,t}(\widetilde{\mathfrak{m}}_{r,t}) &= (t-t_*)^{1/2}\gamma^2\Big[1+\mathcal{O}\left((t-t_*)^{1/3}[\Im m_{r,t}(\widetilde{\mathfrak{m}}_{r,t})]^{1/3}\right)\Big]\\ &= \gamma^2(t-t_*)^{1/2}+\mathcal{O}\left(t-t_*\right). \qquad \square\end{aligned}$$

For the t-derivative of $\mathfrak{e}^+_{r,t}$ and $\widetilde{\mathfrak{m}}_{r,t}$ we have the explicit relations

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathfrak{e}^+_{r,t} = -m_{r,t}(\mathfrak{e}^+_{r,t}), \quad 0\le t\le t_* \tag{5.22c}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mathfrak{m}}_{r,t} = -\Re m_{r,t}(\widetilde{\mathfrak{m}}_{r,t}) + \mathcal{O}\left(t-t_*\right), \quad t_*\le t\le 2t_*. \tag{5.22d}$$

Proof of (5.22c) *and* (5.22d). The claim in (5.22c) was already observed in the proof of Lemma 4.5.1. For (5.22d) we begin by computing the integral

$$\begin{aligned}m'_{r,t_*}(\mathfrak{c}_r+\mathrm{i}\eta) &= \int_{\mathbb{R}}\frac{\rho_{t_*}(\mathfrak{c}_r+x)}{(x-\mathrm{i}\eta)^2}\,\mathrm{d}x = \int_{\mathbb{R}}\frac{\sqrt{3}\gamma^{4/3}\,|x|^{1/3}+\mathcal{O}(\,|x|^{2/3}\,)}{2\pi(x-\mathrm{i}\eta)^2}\,\mathrm{d}x\\ &= \frac{\gamma^{4/3}}{3\eta^{2/3}}+\mathcal{O}\left(\eta^{-1/3}\right),\end{aligned} \tag{5.23}$$

so that by definition $m_{r,t}(z)=m_{r,t_*}(z+(t-t_*)m_{r,t}(z))$ of the free semicircular flow,

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}m_{r,t}(\widetilde{\mathfrak{m}}_{r,t})\\ &= m'_{r,t_*}(\widetilde{\mathfrak{m}}_{r,t}+(t-t_*)m_{r,t}(\widetilde{\mathfrak{m}}_{r,t}))\Big[\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mathfrak{m}}_{r,t}+m_{r,t}(\widetilde{\mathfrak{m}}_{r,t})+(t-t_*)\frac{\mathrm{d}}{\mathrm{d}t}m_{r,t}(\widetilde{\mathfrak{m}}_{z,t})\Big]\\ &= \Big(\frac{1}{3(t-t_*)}+\mathcal{O}\left((t-t_*)^{-1/2}\right)\Big)\Big[\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mathfrak{m}}_{r,t}+m_{r,t}(\widetilde{\mathfrak{m}}_{r,t})+(t-t_*)\frac{\mathrm{d}}{\mathrm{d}t}m_{r,t}(\widetilde{\mathfrak{m}}_{r,t})\Big]\\ &= \mathrm{i}\Big(\frac{1}{3(t-t_*)}+\mathcal{O}\left((t-t_*)^{-1/2}\right)\Big)\Big[\Im m_{r,t}(\widetilde{\mathfrak{m}}_{r,t})+(t-t_*)\frac{\mathrm{d}}{\mathrm{d}t}\Im m_{r,t}(\widetilde{\mathfrak{m}}_{r,t})\Big]\\ &= \Big(\mathrm{i}\frac{\gamma^2}{3(t-t_*)^{1/2}}+\frac{\mathrm{i}}{3}\frac{\mathrm{d}}{\mathrm{d}t}\Im m_{r,t}(\widetilde{\mathfrak{m}}_{r,t})\Big)\Big[1+\mathcal{O}\left((t-t_*)^{1/2}\right)\Big].\end{aligned}$$

Here we used (5.22a), (5.22b) together with (5.23) in the second step. The third step follows from taking the t-derivative of (5.22a). The ultimate inequality is again a consequence of (5.22b). By considering real and imaginary part separately it thus follows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Im m_{r,t}(\widetilde{\mathfrak{m}}_{r,t}) = \frac{\gamma^2}{2(t-t_*)^{1/2}}\Big[1+\mathcal{O}\left((t-t_*)^{1/2}\right)\Big], \qquad \frac{\mathrm{d}}{\mathrm{d}t}\Re m_{r,t}(\widetilde{\mathfrak{m}}_{r,t}) = \mathcal{O}\left(1\right)$$

and therefore (5.22d) follows by differentiating (5.22a). □

5.4.2 Quantiles

Finally we consider the locations of quantiles of $\rho_{r,t}$ for $r = \alpha, \lambda, \mu$ and their fluctuation scales. For $0 \leq t \leq t_*$ we define the shifted quantiles $\widehat{\gamma}_{r,i}(t)$, and for $t_* \leq t \leq 2t_*$ the shifted quantiles[4] $\check{\gamma}_{r,i}(t)$ in such a way that

$$\int_0^{\widehat{\gamma}_{r,i}(t)} \rho_{r,t}(\mathfrak{e}^+_{r,t} + \omega)\, \mathrm{d}\omega = \frac{i}{N}, \quad \int_0^{\check{\gamma}_{r,i}(t)} \rho_{r,t}(\mathfrak{m}_{r,t} + \omega)\, \mathrm{d}\omega = \frac{i}{N}, \quad |i| \ll N. \tag{5.24}$$

Notice that for $i = 0$ we always have $\widehat{\gamma}_{r,0}(t) = \check{\gamma}_{r,0}(t) = 0$. We will also need to define the semiquantiles, distinguished by star from the quantiles, defined as follows:

$$\int_0^{\widehat{\gamma}^*_{r,i}(t)} \rho_{r,t}(\mathfrak{e}^+_{r,t} + \omega)\, \mathrm{d}\omega = \frac{i - \frac{1}{2}}{N}, \quad \int_0^{\check{\gamma}^*_{r,i}(t)} \rho_{r,t}(\mathfrak{m}_{r,t} + \omega)\, \mathrm{d}\omega = \frac{i - \frac{1}{2}}{N}, \quad 1 \leq i \ll N$$

and

$$\int_0^{\widehat{\gamma}^*_{r,i}(t)} \rho_{r,t}(\mathfrak{e}^+_{r,t} + \omega)\, \mathrm{d}\omega = \frac{i + \frac{1}{2}}{N}, \quad \int_0^{\check{\gamma}^*_{r,i}(t)} \rho_{r,t}(\mathfrak{m}_{r,t} + \omega)\, \mathrm{d}\omega = \frac{i + \frac{1}{2}}{N}, \quad -N \ll i \leq -1 \tag{5.25}$$

Note that the definition is slightly different for positive and negative i's, in particular $\widehat{\gamma}^*_i \in [\widehat{\gamma}_{i-1}, \widehat{\gamma}_i]$ for $i \geq 1$ and $\widehat{\gamma}^*_i \in [\widehat{\gamma}_i, \widehat{\gamma}_{i+1}]$ for $i < 0$. The semiquantiles are not defined for $i = 0$.

Lemma 5.4.4. *For $1 \leq |i| \ll N$, $r = \alpha, \lambda, \mu$ and $0 \leq t \leq t_*$ we have*

$$\begin{aligned} \widehat{\gamma}_{r,i}(t) &\sim \operatorname{sgn}(i) \max\Big\{ \Big(\frac{|i|}{N}\Big)^{3/4}, \Big(\frac{|i|}{N}\Big)^{2/3} (t_* - t)^{1/6} \Big\} - \begin{cases} 0, & i > 0 \\ \Delta_{r,t}, & i < 0 \end{cases} \\ \widehat{\gamma}_{r,i}(t) &= \widehat{\gamma}_{\mu,i}(t) \Big[1 + \mathcal{O}\Big((t_* - t)^{1/3} + \min\Big\{ \frac{\widehat{\gamma}_{\mu,i}(t)^{1/2}}{(t_* - t)^{1/4}}, \widehat{\gamma}_{\mu,i}(t)^{1/3} \Big\} \Big) \Big], \end{aligned} \tag{5.26a}$$

while for $t_* \leq t \leq 2t_*$ we have

$$\check{\gamma}_{r,i}(t) \sim \operatorname{sgn}(i) \min\Big\{ \Big(\frac{|i|}{N}\Big)^{3/4}, \frac{|i|}{N} (t_* - t)^{-1/2} \Big\}, \tag{5.26b}$$

$$\check{\gamma}_{r,i}(t) = \check{\gamma}_{\mu,i}(t) \Big[1 + \mathcal{O}\Big((t_* - t)^{1/2} + \min\Big\{ \frac{\check{\gamma}_{\mu,i}(t)^2}{(t_* - t)^{11/4}}, \frac{\check{\gamma}_{\mu,i}(t)}{t_* - t}, \check{\gamma}_{\mu,i}(t)^{1/3} \Big\} \Big) \Big].$$

Moreover, the quantiles of $\rho_{\alpha,t}$ are the convex combination

$$\widehat{\gamma}_{\alpha,i}(t) = \alpha \widehat{\gamma}_{\lambda,i}(t) + (1 - \alpha) \widehat{\gamma}_{\mu,i}(t), \quad \check{\gamma}_{\alpha,i}(t) = \alpha \check{\gamma}_{\lambda,i}(t) + (1 - \alpha) \check{\gamma}_{\mu,i}(t). \tag{5.26c}$$

Proof. The proof follows directly from the estimates in (5.20a) and (5.20b). The relation (5.26c) follows directly from (5.15) in the definition of the α-interpolating density. □

[4] We use a separate variable name $\check{\gamma}$ because in Section 5.8 the name $\widehat{\gamma}$ is used for the quantiles with respect to the base point $\widetilde{\mathfrak{m}}$ instead of $\mathfrak{m}$.

5.4.3 Rigidity scales

The fluctuation scale $\eta_{\mathrm{f}}^{\rho}(\tau)$ of any density function $\rho(\omega)$ around τ is defined via

$$\int_{\tau-\eta_{\mathrm{f}}^{\rho}(\tau)}^{\tau+\eta_{\mathrm{f}}^{\rho}(\tau)} \rho(\omega)\,\mathrm{d}\omega = \frac{1}{N}$$

for $\tau \in \operatorname{supp}\rho$ and by the value $\eta_{\mathrm{f}}(\tau) := \eta_{\mathrm{f}}(\tau')$ where $\tau' \in \operatorname{supp}\rho$ is the edge closest to τ for $\tau \notin \operatorname{supp}\rho$. If this edge is not unique, an arbitrary choice can be made between the two possibilities. From (5.26a) we immediately obtain for $0 \le t \le t_*$ and $1 \le i \le N$, that

$$\begin{aligned}\eta_{\mathrm{f}}^{\rho_{r,t}}(\mathfrak{e}_{r,t}^{+} + \widehat{\gamma}_{r,\pm i}(t)) &\sim \max\Big\{\frac{\Delta_{r,t}^{1/9}}{N^{2/3}i^{1/3}}, \frac{1}{N^{3/4}i^{1/4}}\Big\}\\ &\sim \max\Big\{\frac{(t_*-t)^{1/6}}{N^{2/3}i^{1/3}}, \frac{1}{N^{3/4}i^{1/4}}\Big\}, \qquad r=\alpha,\lambda,\mu,\end{aligned} \tag{5.27a}$$

while for $t_* \le t \le 2t_*$, $1 \le |i| \ll N$ we obtain from (5.26b) that

$$\begin{aligned}\eta_{\mathrm{f}}^{\rho_{r,t}}(\mathfrak{m}_{r,t} + \check{\gamma}_{r,i}(t)) &\sim \min\Big\{\frac{1}{N\rho_{r,t}(\mathfrak{m}_{r,t})}, \frac{1}{N^{3/4}\,|i|^{1/4}}\Big\}\\ &\sim \min\Big\{\frac{1}{N(t-t_*)^{1/2}}, \frac{1}{N^{3/4}\,|i|^{1/4}}\Big\}, \qquad r=\alpha,\lambda,\mu.\end{aligned} \tag{5.27b}$$

In the second relations we used (5.17a) and (5.17c). For reference purposes we also list for $0 < i, j \ll N$ the bounds

$$|\widehat{\gamma}_{r,i}(t) - \widehat{\gamma}_{r,j}(t)| \sim \max\Big\{\frac{\Delta_{r,t}^{1/9}\,|i-j|}{N^{2/3}(i+j)^{1/3}}, \frac{|i-j|}{N^{3/4}(i+j)^{1/4}}\Big\}, \tag{5.28}$$

in case $t \le t_*$ and

$$|\check{\gamma}_{r,i}(t) - \check{\gamma}_{r,j}(t)| \sim \min\Big\{\frac{|i-j|}{\rho_{r,t}(\mathfrak{m}_{r,t})N}, \frac{|i-j|}{N^{3/4}(i+j)^{1/4}}\Big\} \tag{5.29}$$

in case $t > t_*$. Furthermore we have

$$\rho_{r,t}(\mathfrak{e}_{r,t}^{+} + \widehat{\gamma}_{r,i}(t)) \sim \min\Big\{\frac{i^{1/3}}{N^{1/3}(t_*-t)^{1/6}}, \frac{i^{1/4}}{N^{1/4}}\Big\} \tag{5.30}$$

and

$$\rho_{r,t}(\mathfrak{m}_{r,t} + \check{\gamma}_{r,i}(t)) \sim \max\Big\{\rho_{r,t}(\mathfrak{m}_{r,t}), \frac{i^{1/4}}{N^{1/4}}\Big\}. \tag{5.31}$$

5.4.4 Stieltjes transform bounds

It follows from (5.17b) and (5.17d) that also the real parts of the Stieltjes transforms $m_{\alpha,t}$, $m_{\lambda,t}$, $m_{\mu,t}$ are close. We claim that for $r=\lambda,\alpha$, $\lambda \in [-\delta_*, \delta_*]$ and $0 \le t \le t_*$ we have

$$\begin{aligned}&\Big|\Re\Big[\Big(m_{r,t}(\mathfrak{e}_{r,t}^{+}+\lambda) - m_{r,t}(\mathfrak{e}_{r,t}^{+})\Big) - \Big(m_{\mu,t}(\mathfrak{e}_{\mu,t}^{+}+\lambda) - m_{\mu,t}(\mathfrak{e}_{\mu,t}^{+})\Big)\Big]\Big|\\ &\quad\lesssim |\lambda|^{1/3}\Big[|\lambda|^{1/3} + (t_*-t)^{1/3}\Big]\,|\log|\lambda|| + (t_*-t)^{11/18}\mathbf{1}(\lambda \le -\Delta_{\mu,t}/2),\end{aligned} \tag{5.32a}$$

while for $t_* \le t \le 2t_*$ we have

$$\begin{aligned}&\Big|\Re\Big[\Big(m_{r,t}(\mathfrak{m}_{r,t}+\lambda)-m_{r,t}(\mathfrak{m}_{r,t})\Big)-\Big(m_{\mu,t}(\mathfrak{m}_{\mu,t}+\lambda)-m_{\mu,t}(\mathfrak{m}_{\mu,t})\Big)\Big]\Big|\\ &\quad\lesssim \Big[|\lambda|^{1/3}(t-t_*)^{1/4}+(t_*-t)^{3/4}+|\lambda|^{2/3}\Big]\,|\log|\lambda||\,.\end{aligned}\tag{5.32b}$$

Proof of (5.32). We first recall from Lemma 5.4.3 that also the density $\rho_{\alpha,t}$ is $1/3$-Hölder continuous which we will use repeatedly in the following proof. We begin with the proof of (5.32a) and compute for $r=\alpha,\lambda,\mu$

$$\begin{aligned}&\Re\Big[m_{r,t}(\mathfrak{e}^+_{r,t}+\lambda)-m_{r,t}(\mathfrak{e}^+_{r,t})\Big]\\ &\quad=\int_0^\infty \frac{\lambda\rho_{r,t}(\mathfrak{e}^+_{r,t}+\omega)}{(\omega-\lambda)\omega}\,\mathrm{d}\omega+\int_0^\infty\frac{\lambda\rho_{r,t}(\mathfrak{e}^-_{r,t}-\omega)}{(\omega+\Delta_{r,t}+\lambda)(\omega+\Delta_{r,t})}\,\mathrm{d}\omega.\end{aligned}\tag{5.33}$$

For $\lambda>0$ the first of the two terms is the more critical one. Our goal is to obtain a bound on

$$\int_0^\infty\frac{\lambda}{(\omega-\lambda)\omega}\Big[\rho_{\lambda,t}(\mathfrak{e}^+_{\lambda,t}+\omega)-\rho_{\mu,t}(\mathfrak{e}^+_{\mu,t}+\omega)\Big]\,\mathrm{d}\omega$$

by using (5.17b). Let $0<\epsilon<\lambda/2$ be a small parameter for which we separately consider the two critical regimes $0\le\omega\le\epsilon$ and $|\lambda-\omega|\le\epsilon$. We use

$$\rho_{r,t}(\mathfrak{e}^+_{r,t}+\omega)\lesssim\omega^{1/3}\quad\text{and}\quad\rho_{r,t}(\mathfrak{e}^+_{r,t}+\omega)=\rho_{r,t}(\mathfrak{e}^+_{r,t}+\lambda)+\mathcal{O}\left(|\omega-\lambda|^{1/3}\right)\tag{5.34}$$

for $r=\lambda,\mu$, from the $1/3$-Hölder continuity of $\rho_{r,t}$ and the fact that the integral over $1/(\omega-\lambda)$ from $\lambda-\epsilon$ to $\lambda+\epsilon$ vanishes by symmetry to estimate, for $r=\lambda,\mu$,

$$\left|\int_0^\epsilon\frac{\lambda}{(\omega-\lambda)\omega}\rho_{r,t}(\mathfrak{e}^+_{r,t}+\omega)\,\mathrm{d}\omega\right|\lesssim\int_0^\epsilon|\omega|^{-2/3}\,\mathrm{d}\omega\lesssim\epsilon^{1/3}$$

and

$$\begin{aligned}\left|\int_{\lambda-\epsilon}^{\lambda+\epsilon}\left[\frac{\rho_{r,t}(\mathfrak{e}^+_{r,t}+\omega)}{\omega-\lambda}-\frac{\rho_{r,t}(\mathfrak{e}^+_{r,t}+\omega)}{\omega}\right]\mathrm{d}\omega\right|&\lesssim\int_{\lambda-\epsilon}^{\lambda+\epsilon}|\omega-\lambda|^{-2/3}\,\mathrm{d}\omega+\epsilon\lambda^{-2/3}\\ &\lesssim\epsilon^{1/3}+\epsilon\lambda^{-2/3}.\end{aligned}$$

Next, we consider the remaining integration regimes where we use (5.17b) and (5.34) to estimate

$$\begin{aligned}&\left|\int_\epsilon^{\lambda-\epsilon}\frac{\lambda}{(\omega-\lambda)\omega}\Big[\rho_{r,t}(\mathfrak{e}^+_{r,t}+\omega)-\rho_{\mu,t}(\mathfrak{e}^+_{\mu,t}+\omega)\Big]\,\mathrm{d}\omega\right|\\ &\quad\lesssim\int_\epsilon^{\lambda/2}\frac{\omega^{1/3}(t_*-t)^{1/3}+\omega^{2/3}}{\omega}\,\mathrm{d}\omega+\int_{\lambda/2}^{\lambda-\epsilon}\Big(\frac{\lambda^{1/3}(t_*-t)^{1/3}}{\omega-\lambda}+\frac{\lambda^{2/3}}{\omega-\lambda}\Big)\,\mathrm{d}\omega\\ &\quad\lesssim\lambda^{1/3}\Big((t_*-t)^{1/3}+\lambda^{1/3}\Big)\,|\log\epsilon|\end{aligned}$$

and similarly

$$\left|\int_{\lambda+\epsilon}^\infty\frac{\lambda}{(\omega-\lambda)\omega}\Big[\rho_{r,t}(\mathfrak{e}^+_{r,t}+\omega)-\rho_{\mu,t}(\mathfrak{e}^+_{\mu,t}+\omega)\Big]\,\mathrm{d}\omega\right|\lesssim\lambda^{1/3}\Big((t_*-t)^{1/3}+\lambda^{1/3}\Big)\,|\log\epsilon|\,.$$

We now consider the difference of the first terms in (5.33) for $r = \lambda, \mu$ and for $\lambda < 0$ where the bound is simpler because the integration regime close to λ does not have to be singled out. Using (5.17b) we find

$$\left| \int_0^\infty \frac{\lambda}{(\omega-\lambda)\omega} \Big[\rho_{r,t}(\mathfrak{e}_{r,t}^+ + \omega) - \rho_{\mu,t}(\mathfrak{e}_{\mu,t}^+ + \omega) \Big] \mathrm{d}\omega \right| \lesssim |\lambda|^{2/3} + (t_* - t)^{1/3} |\lambda|^{1/3} .$$

Finally, it remains to consider the difference of the second terms in (5.33). We first treat the regime where $\lambda \geq -\frac{3}{4}\Delta_{r,t}$ and split the difference into the sum of two terms

$$\begin{aligned}
&\left| \int_0^\infty \left(\frac{\lambda \rho_{r,t}(\mathfrak{e}_{r,t}^- - \omega)}{(\omega + \Delta_{r,t} + \lambda)(\omega + \Delta_{r,t})} - \frac{\lambda \rho_{r,t}(\mathfrak{e}_{r,t}^- - \omega)}{(\omega + \Delta_{\mu,t} + \lambda)(\omega + \Delta_{\mu,t})} \right) \mathrm{d}\omega \right| \\
&\quad \leq |\lambda| \, |\Delta_{r,t} - \Delta_{\mu,t}| \int_0^\infty \frac{\rho_{r,t}(\mathfrak{e}_{r,t}^- - \omega) \big[2\Delta_{r,t} + 2\omega + |\lambda| \big]}{(\omega + \Delta_{r,t} + \lambda)^2 (\omega + \Delta_{r,t})^2} \mathrm{d}\omega \\
&\quad \lesssim \frac{|\Delta_{r,t} - \Delta_{\mu,t}|}{\Delta_{r,t}^{2/3}} - \frac{|\Delta_{r,t} - \Delta_{\mu,t}|}{(\Delta_{r,t} + |\lambda|)^{2/3}} \lesssim (t_* - t)^{1/3} |\lambda|^{1/3}
\end{aligned}$$

and

$$\begin{aligned}
&\left| \int_0^\infty \left(\frac{\lambda \rho_{r,t}(\mathfrak{e}_{r,t}^- - \omega)}{(\omega + \Delta_{\mu,t} + \lambda)(\omega + \Delta_{\mu,t})} - \frac{\lambda \rho_{\mu,t}(\mathfrak{e}_{\mu,t}^- - \omega)}{(\omega + \Delta_{\mu,t} + \lambda)(\omega + \Delta_{\mu,t})} \right) \mathrm{d}\omega \right| \\
&\quad \lesssim |\lambda|^{2/3} + (t_* - t)^{1/3} |\lambda|^{1/3} .
\end{aligned}$$

Here we used $\rho_{r,t}(\mathfrak{e}_{r,t}^- - \omega) \lesssim \omega^{1/3}$ as well as (5.17a) for the first and (5.17a),(5.17b) for the second computation. By collecting the various error terms and choosing $\epsilon = \lambda^2$ we conclude (5.32a).

We define $\kappa := -\lambda - \Delta_{r,t}$. Then we are left with the regime $\lambda < -\frac{3}{4}\Delta_{r,t}$ or equivalently $\kappa > -\frac{1}{4}\Delta_{r,t}$ and use

$$m_{r,t}(\mathfrak{e}_{r,t}^+ + \lambda) - m_{r,t}(\mathfrak{e}_{r,t}^+) = (m_{r,t}(\mathfrak{e}_{r,t}^- - \kappa) - m_{r,t}(\mathfrak{e}_{r,t}^-)) + (m_{r,t}(\mathfrak{e}_{r,t}^-) - m_{r,t}(\mathfrak{e}_{r,t}^+)) ,$$

as well as

$$\begin{aligned}
m_{\mu,t}(\mathfrak{e}_{\mu,t}^+ + \lambda) - m_{\mu,t}(\mathfrak{e}_{\mu,t}^+) &= (m_{\mu,t}(\mathfrak{e}_{\mu,t}^- - \kappa + \Delta_{\mu,t} - \Delta_{r,t}) - m_{\mu,t}(\mathfrak{e}_{\mu,t}^- - \kappa)) \\
&\quad + (m_{\mu,t}(\mathfrak{e}_{\mu,t}^- - \kappa) - m_{\mu,t}(\mathfrak{e}_{\mu,t}^-)) + (m_{\mu,t}(\mathfrak{e}_{\mu,t}^-) - m_{\mu,t}(\mathfrak{e}_{\mu,t}^+))
\end{aligned}$$

in the left hand side of (5.32a). Thus we have to estimate the three expressions,

$$\left| \Re \Big[\Big(m_{r,t}(\mathfrak{e}_{r,t}^- - \kappa) - m_{r,t}(\mathfrak{e}_{r,t}^-) \Big) - \Big(m_{\mu,t}(\mathfrak{e}_{\mu,t}^- - \kappa) - m_{\mu,t}(\mathfrak{e}_{\mu,t}^-) \Big) \Big] \right| , \tag{5.35a}$$

$$\left| \Re \Big[\Big(m_{r,t}(\mathfrak{e}_{r,t}^-) - m_{r,t}(\mathfrak{e}_{r,t}^+) \Big) - \Big(m_{\mu,t}(\mathfrak{e}_{\mu,t}^-) - m_{\mu,t}(\mathfrak{e}_{\mu,t}^+) \Big) \Big] \right| , \tag{5.35b}$$

$$\left| \Re \Big[m_{\mu,t}(\mathfrak{e}_{\mu,t}^- - \kappa + \Delta_{\mu,t} - \Delta_{r,t}) - m_{\mu,t}(\mathfrak{e}_{\mu,t}^- - \kappa) \Big] \right| . \tag{5.35c}$$

In order to bound the first term we use that estimating (5.35a) for $\kappa \geq -\frac{3}{4}\Delta_{r,t}$ is equivalent to estimating the left hand side of (5.32a) for $\lambda \geq -\frac{3}{4}\Delta_{r,t}$, i.e. the regime we already considered above. This equivalence follows by using the reflection $A \to -A$ of the expectation (cf. (5.7)) that turns every left edge $\mathfrak{e}_{z,t}^+$ into a right edge $\mathfrak{e}_{z,t}^-$. In particular, by the analysis that we already performed (5.35a) is bounded by $|\kappa|^{1/3}\,[|\kappa|^{1/3}+(t_*-t)^{1/3}]\,|\log|\kappa||$. Since $|\kappa|\le|\lambda|$ this is the desired bound.

For the second term (5.35b) we see from (5.33) that we have to estimate the difference between the expressions

$$\int_0^\infty \frac{\Delta_{r,t}\rho_{r,t}(\mathfrak{e}_{r,t}^+ +\omega)}{\omega(\omega+\Delta_{r,t})}\,\mathrm{d}\omega+\int_0^\infty \frac{\Delta_{r,t}\rho_{r,t}(\mathfrak{e}_{r,t}^- -\omega)}{\omega(\omega+\Delta_{r,t})}\,\mathrm{d}\omega, \tag{5.36}$$

for $r=\alpha,\lambda,\mu$. The summands in (5.36) are treated analogously, so we focus on the first summand. We split the integrand of the difference between the first summands and estimate

$$\begin{aligned}&\frac{(\Delta_{r,t}-\Delta_{\mu,t})\rho_{r,t}(\mathfrak{e}_{r,t}^+ +\omega)}{(\omega+\Delta_{r,t})(\omega+\Delta_{\mu,t})}+\frac{\Delta_{\mu,t}}{\omega(\omega+\Delta_{\mu,t})}(\rho_{r,t}(\mathfrak{e}_{r,t}^+ +\omega)-\rho_{\mu,t}(\mathfrak{e}_{\mu,t}^+ +\omega))\\&\qquad\lesssim\frac{\Delta(\omega^{1/3}+(t_*-t)^{1/3})}{\omega^{2/3}(\omega+\Delta)}\end{aligned}$$

where $\Delta:=\Delta_{r,t}\sim\Delta_{\mu,t}$ and we used (5.17a), (5.17b) and the first inequality of (5.34). Thus

$$\left|\int_0^\infty \frac{\Delta_{r,t}\rho_{r,t}(\mathfrak{e}_{r,t}^+ +\omega)}{\omega(\omega+\Delta_{r,t})}\,\mathrm{d}\omega-\int_0^\infty \frac{\Delta_{\mu,t}\rho_{\mu,t}(\mathfrak{e}_{\mu,t}^+ +\omega)}{\omega(\omega+\Delta_{\mu,t})}\,\mathrm{d}\omega\right|\lesssim\Delta^{2/3}+\Delta^{1/3}(t_*-t)^{1/3}.$$

Since $|\lambda|\gtrsim\Delta$ this finishes the estimate on (5.35b).

For (5.35c) we use the $1/3$-Hölder regularity of $m_{\mu,t}$ and (5.17a) to get an upper bound $\Delta^{1/3}(t_*-t)^{1/9}\lesssim(t_*-t)^{11/18}$. This finishes the proof of (5.32a).

We now turn to the case of a small local minimum in (5.32b) and compute for $r=\alpha,\lambda,\mu$ and $\lambda\neq 0$ that

$$\Re\Big[m_{r,t}(\mathfrak{m}_{r,t}+\lambda)-m_{r,t}(\mathfrak{m}_{r,t})\Big]=\int_{\mathbb{R}}\frac{\lambda\rho_{r,t}(\mathfrak{m}_{r,t}+\omega)}{(\omega-\lambda)\omega}\,\mathrm{d}\omega.$$

Without loss of generality, we consider the case $\lambda>0$ as $\lambda<0$ is completely analogous. As before, we first pick a threshold $\epsilon\le\lambda/2$ and single out the integration over $[-\epsilon,\epsilon]$ and $[\lambda-\epsilon,\lambda+\epsilon]$. From the $1/3$-Hölder continuity of $\rho_{r,t}$ we have, for $r=\lambda,\mu$,

$$\rho_{r,t}(\mathfrak{m}_{r,t}+\omega)=\rho_{r,t}(\mathfrak{m}_{r,t}+\lambda)+\mathcal{O}\left(|\lambda-\omega|^{1/3}\right)$$

and therefore

$$\left|\int_{-\epsilon}^{\epsilon}\frac{\rho_{r,t}(\mathfrak{m}_{r,t}+\omega)}{\omega-\lambda}\,\mathrm{d}\omega\right|\lesssim\frac{\epsilon}{\lambda},\qquad\left|\int_{-\epsilon}^{\epsilon}\frac{\rho_{r,t}(\mathfrak{m}_{r,t}+\omega)}{\omega}\,\mathrm{d}\omega\right|\lesssim\int_{-\epsilon}^{\epsilon}|\omega|^{-2/3}\,\mathrm{d}\omega\lesssim\epsilon^{1/3}$$

and

$$\begin{aligned}\left|\int_{\lambda-\epsilon}^{\lambda+\epsilon}\frac{\rho_{r,t}(\mathfrak{m}_{r,t}+\omega)}{\omega-\lambda}\,\mathrm{d}\omega\right|&\lesssim\int_{\lambda-\epsilon}^{\lambda+\epsilon}|\omega-\lambda|^{-2/3}\,\mathrm{d}\omega\lesssim\epsilon^{1/3},\\\left|\int_{\lambda-\epsilon}^{\lambda+\epsilon}\frac{\rho_{r,t}(\mathfrak{m}_{r,t}+\omega)}{\omega}\,\mathrm{d}\omega\right|&\lesssim\frac{\epsilon}{\lambda}.\end{aligned}$$

We now consider the difference between $\rho_{r,t}$ and $\rho_{\mu,t}$ for which we have

$$|\rho_{r,t}(\mathfrak{m}_{r,t}+\omega)-\rho_{\mu,t}(\mathfrak{m}_{\mu,t}+\omega)|\lesssim(t-t_*)\,|\omega|^{1/3}\,(t-t_*)^{1/4}+(t-t_*)^{3/4}+|\omega|^{2/3}$$

from (5.17d), (5.17c) and the $1/3$-Hölder continuity of $\rho_{r,t}$. Thus we can estimate

$$\begin{aligned}&\left|\left[\int_{-\infty}^{-\epsilon} + \int_{\epsilon}^{\lambda-\epsilon} + \int_{\lambda+\epsilon}^{\infty}\right] \frac{\lambda(\rho_{\lambda,t}(\mathfrak{m}_{r,t}+\omega) - \rho_{r,t}(\mathfrak{m}_{r,t}+\omega))}{(\omega-\lambda)\omega} \,\mathrm{d}\omega\right| \\ &\quad\lesssim \left[\int_{-\infty}^{-\epsilon} + \int_{\epsilon}^{\lambda-\epsilon} + \int_{\lambda+\epsilon}^{\infty}\right] \frac{\lambda(|\omega|^{1/3}\,(t-t_*)^{1/4} + (t-t_*)^{3/4} + |\omega|^{2/3})}{|\omega-\lambda|\,\omega} \,\mathrm{d}\omega \\ &\quad\lesssim |\log\epsilon| \left[\lambda^{1/3}(t-t_*)^{1/4} + (t-t_*)^{3/4} + \lambda^{2/3}\right].\end{aligned}$$

We again choose $\epsilon = \lambda^2$ and by collecting the various error estimates can conclude (5.32b). □

5.5 Index matching for two DBM

For two real symmetric matrix valued standard (GOE) Brownian motions $\mathfrak{B}_t^{(\lambda)}, \mathfrak{B}_t^{(\mu)} \in \mathbb{R}^{N\times N}$ we define the matrix flows

$$H_t^{(\lambda)} := H^{(\lambda)} + \mathfrak{B}_t^{(\lambda)}, \quad H_t^{(\mu)} := H^{(\mu)} + \mathfrak{B}_t^{(\mu)}. \tag{5.37}$$

In particular, by (5.37) it follows that

$$H_t^{(\lambda)} \overset{d}{=} H^{(\lambda)} + \sqrt{t}U^{(\lambda)}, \quad H_t^{(\mu)} \overset{d}{=} H^{(\mu)} + \sqrt{t}U^{(\mu)}, \tag{5.38}$$

for any fixed $0 \le t \le t_1$, where $U^{(\lambda)}$ and $U^{(\mu)}$ are GOE matrices. In (5.38) with $X \overset{d}{=} Y$ we denote that the two random variables X and Y are equal in distribution.

We will prove Proposition 5.3.1 by comparing the two Dyson Brownian motions for the eigenvalues of the matrices $H_t^{(\lambda)}$ and $H_t^{(\mu)}$ for $0 \le t \le t_1$, see (5.39)–(5.40) below. To do this, we will use the coupling idea of [40] and [43], where the DBMs for the eigenvalues of $H_t^{(\lambda)}$ and $H_t^{(\mu)}$ are coupled in such a way that the difference of the two DBMs obeys a discrete parabolic equation with good decay properties. In order to analyse this equation we consider a *short range approximation* for the DBM, first introduced in [79]. Coupling only the short range approximation of the DBMs leads to a parabolic equation whose heat kernel has a rapid off diagonal decay by *finite speed of propagation* estimates. In this way the kernels of both DBMs are locally determined and thus can be directly compared by optimal rigidity since locally the two densities, hence their quantiles, are close. Technically it is much easier to work with a one parameter interpolation between the two DBM's and consider its derivative with respect to the parameter, as introduced in [40]; the proof of the finite speed propagation for this dynamics does not require to establish level repulsion unlike in several previous works [79, 77, 121]. However, it requires to establish (almost) optimal rigidity for the interpolating dynamics as well. Note that optimal rigidity is known for $H_t^{(\lambda)}$ and $H_t^{(\mu)}$ from [DS5], see Lemma 5.6.1 later, but not for the interpolation.

In Section 5.6 we will establish rigidity for the interpolating process by DBM methods. Armed with this rigidity, in Section 5.7 we prove Proposition 5.3.1 for the small gap and the exact cusp case, i.e. $t_1 \le t_*$. Some estimates are slightly different for the small minimum case, i.e. $t_* \le t_1 \le 2t_*$, the modifications are given in Section 5.8. We recall that t_* is the time at which both $H_{t_*}^{(\lambda)}$ and $H_{t_*}^{(\mu)}$ have an exact cusp. Some technical details on the corresponding Sobolev inequality and heat kernel estimates as well as finite speed of propagation and short range approximation are deferred to the Appendix: these are similar to the corresponding estimates for the edge case, see [41] and [122], respectively.

In the rest of this section we prepare the proof of Proposition 5.3.1 by setting up the appropriate framework. While we are interested only in the eigenvalues near the physical cusp, the DBM is highly non-local, so we need to define the dynamics for all eigenvalues. In the setup of Proposition 5.3.1 we could easily assume that the cusps for the two matrix flows are formed at the same time and their slope parameters coincide – these could be achieved by a rescaling and a trivial time shift. However, the number of eigenvalues to the left of the cusp may macroscopically differ for the two ensembles which would mean that the labels of the ordered eigenvalues near the cusp would not be constant along the interpolation. To resolve this discrepancy, we will pad the system with N fictitious particles in addition to the original flow of N eigenvalues similarly as in [120], giving sufficient freedom to match the labels of the eigenvalues near the cusp. These artificial particles will be placed very far from the cusp regime and from each other so that their effect on the dynamics of the relevant particles is negligible.

With the notation of Section 5.4, we let $\rho_{\lambda,t}$, $\rho_{\mu,t}$ denote the (self-consistent) densities at time $0 \le t \le t_1$ of $H_t^{(\lambda)}$ and $H_t^{(\mu)}$, respectively. In particular, $\rho_{\lambda,0} = \rho_\lambda$ and $\rho_{\mu,0} = \rho_\mu$, where ρ_λ, ρ_μ are the self consistent densities of $H^{(\lambda)}$ and $H^{(\mu)}$ and $\rho_{\lambda,t}$, $\rho_{\mu,t}$ are their semicircular evolutions. For each $0 \le t \le t_*$ both densities $\rho_{\lambda,t}$, $\rho_{\mu,t}$ have a small gap, denoted by $[\mathfrak{e}_{\lambda,t}^-, \mathfrak{e}_{\lambda,t}^+]$ and $[\mathfrak{e}_{\mu,t}^-, \mathfrak{e}_{\mu,t}^+]$ and we let

$$\Delta_{\lambda,t} := \mathfrak{e}_{\lambda,t}^+ - \mathfrak{e}_{\lambda,t}^-, \qquad \Delta_{\mu,t} := \mathfrak{e}_{\mu,t}^+ - \mathfrak{e}_{\mu,t}^-$$

denote the length of these gaps. In case of $t_* \le t \le 2t_*$ the densities $\rho_{\lambda,t}$, $\rho_{\mu,t}$ have a small minimum denoted by $\mathfrak{m}_{\lambda,t}$ and $\mathfrak{m}_{\mu,t}$ respectively. Since we always assume $0 \le t \le t_1 \ll 1$, both $H_t^{(\lambda)}$ and $H_t^{(\mu)}$ will always have exactly one physical cusp near $\mathfrak{c}_\lambda$ and $\mathfrak{c}_\mu$, respectively, using that the Stieltjes transform of the density is a Hölder continuous function of t, see [12, Proposition 10.1].

Let i_λ and i_μ be the indices defined by

$$\int_{-\infty}^{\mathfrak{e}_{\lambda,0}^-} \rho_\lambda = \frac{i_\lambda - 1}{N}, \quad \int_{-\infty}^{\mathfrak{e}_{\mu,0}^-} \rho_\mu = \frac{i_\mu - 1}{N}.$$

By band rigidity (see Remark 3.2.10) i_λ and i_μ are integers. Note that by the explicit expression of the density in (5.2a)-(5.2b) it follows that $cN \le i_\lambda, i_\mu \le (1-c)N$ with some small $c > 0$, because the density on both sides of a physical cusp is macroscopic.

We let $\lambda_i(t)$ and $\mu_i(t)$ denote the eigenvalues of $H_t^{(\lambda)}$ and $H_t^{(\mu)}$, respectively. Let $\{B_i\}_{i\in[-N,N]\setminus\{0\}}$ be a family of independent standard (scalar) Brownian motions. It is well known [67] that the eigenvalues of $H_t^{(\lambda)}$ satisfy the equation for *Dyson Brownian motion*, i.e. the following system of coupled SDE's

$$\mathrm{d}\lambda_i = \sqrt{\frac{2}{N}}\,\mathrm{d}B_{i-i_\lambda+1} + \frac{1}{N}\sum_{j\neq i} \frac{1}{\lambda_i - \lambda_j}\,\mathrm{d}t \tag{5.39}$$

with initial conditions $\lambda_i(0) = \lambda_i(H^{(\lambda)})$. Similarly, for the eigenvalues of $H_t^{(\mu)}$ we have

$$\mathrm{d}\mu_i = \sqrt{\frac{2}{N}}\,\mathrm{d}B_{i-i_\mu+1} + \frac{1}{N}\sum_{j\neq i} \frac{1}{\mu_i - \mu_j}\,\mathrm{d}t \tag{5.40}$$

with initial conditions $\mu_i(0) = \mu_i(H^{(\mu)})$. Note that we chose the Brownian motions for λ_i and $\mu_{i+i_\mu-i_\lambda}$ to be identical. This is the key ingredient for the coupling argument, since in this way the stochastic differentials will cancel when we take the difference of the two DBMs or we differentiate it with respect to an additional parameter.

For convenience of notation, we will shift the indices so that the same index labels the last quantile before the gap in ρ_λ and ρ_μ. This shift was already prepared by choosing the Brownian motions for μ_{i_μ} and λ_{i_λ} to be identical. We achieve this shift by adding N "ghost" particles very far away and relabelling, as in [120]. We thus embed λ_i and μ_i into the enlarged processes $\{x_i\}_{i\in[-N,N]\setminus\{0\}}$ and $\{y_i\}_{i\in[-N,N]\setminus\{0\}}$. Note that the index 0 is always omitted.

More precisely, the processes x_i are defined by the following SDE *(extended Dyson Brownian motion)*

$$\mathrm{d}x_i = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \frac{1}{N}\sum_{j\neq i}\frac{1}{x_i - x_j}\,\mathrm{d}t, \qquad 1 \le |i| \le N, \tag{5.41}$$

with initial data

$$x_i(0) = \begin{cases} -N^{200} + iN & \text{if } -N \le i \le -i_\lambda \\ \lambda_{i+i_\lambda}(0) & \text{if } 1 - i_\lambda \le i \le -1 \\ \lambda_{i+i_\lambda-1}(0) & \text{if } 1 \le i \le N+1-i_\lambda \\ N^{200} + iN & \text{if } N+2-i_\lambda \le i \le N, \end{cases}$$

and the y_i are defined by

$$\mathrm{d}y_i = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \frac{1}{N}\sum_{j\neq i}\frac{1}{y_i - y_j}\,\mathrm{d}t, \qquad 1 \le |i| \le N, \tag{5.42}$$

with initial data

$$y_i(0) = \begin{cases} -N^{200} + iN & \text{if } -N \le i \le -i_\mu \\ \mu_{i+i_\mu}(0) & \text{if } 1 - i_\mu \le i \le -1 \\ \mu_{i+i_\mu-1}(0) & \text{if } 1 \le i \le N+1-i_\mu \\ N^{200} + iN & \text{if } N+2-i_\mu \le i \le N. \end{cases}$$

The summations in (5.41) and (5.42) extend to all j with $1 \le |j| \le N$ except $j = i$.

The following lemma shows that the additional particles at distance N^{200} have negligible effect on the dynamics of the re-indexed eigenvalues, thus we can study the processes x_i and y_i instead of the eigenvalues λ_i, μ_i. The proof of this lemma follows by Appendix C of [120].

Lemma 5.5.1. *With very high probability the following estimates hold:*

$$\sup_{0\le t\le 1}\ \sup_{1\le i\le N+1-i_\lambda} |x_i(t) - \lambda_{i+i_\lambda-1}(t)| \le N^{-100},$$

$$\sup_{0\le t\le 1}\ \sup_{1-i_\lambda\le i\le N+1-i_\lambda} |x_i(t) - \lambda_{i+i_\lambda}(t)| \le N^{-100},$$

$$\sup_{0\le t\le 1}\ \sup_{1\le i\le N+1-i_\mu} |y_i(t) - \mu_{i+i_\mu-1}(t)| \le N^{-100},$$

$$\begin{aligned}&\sup_{0\le t\le 1}\sup_{1-i_\mu\le i\le N+1-i_\mu}|y_i(t)-\mu_{i+i_\mu}(t)|\le N^{-100},\\ &\sup_{0\le t\le 1} x_{-i_\lambda}(t)\lesssim -N^{200},\quad \sup_{0\le t\le 1} x_{N+2-i_\lambda}(t)\gtrsim N^{200},\\ &\sup_{0\le t\le 1} y_{-i_\mu}(t)\lesssim -N^{200},\quad \sup_{0\le t\le 1} y_{N+2-i_\mu}(t)\gtrsim N^{200}.\end{aligned}$$

Remark 5.5.2. *For notational simplicity we assumed that $H^{(\lambda)}$ and $H^{(\mu)}$ have the same dimensions, but our proof works as long as the corresponding dimensions N_λ and N_μ are merely comparable, say $\frac{2}{3}N_\lambda\le N_\mu\le\frac{3}{2}N_\lambda$. The only modification is that the times in* (5.37) *need to be scaled differently in order to keep the strength of the stochastic differential terms in* (5.39)–(5.40) *identical. Furthermore, the number of additional "ghost" particles in the* extended Dyson Brownian motion *(see* (5.41) *and* (5.42)*) will be different to ensure that we have the same total number of particles, say $2N := 2N_\mu$, after the extension. Hence, there will be $N = N_\mu$ particles added to the DBM of the eigenvalues of $H^{(\mu)}$ and $2N_\mu - N_\lambda$ particles added to the DBM of $H^{(\lambda)}$.*

We now construct the analogues of the self-consistent densities $\rho_{\lambda,t}$, $\rho_{\mu,t}$ for the $x(t)$ and $y(t)$ processes as well as for their α-interpolations. We start with $\rho_{x,t}$. Recall $\rho_{\lambda,t}$ from Section 5.4, and set

$$\rho_{x,t}(E) := \rho_{\lambda,t}(E)+\frac{1}{N}\sum_{i=-N}^{-i_\lambda}\psi(E-x_i(t))+\frac{1}{N}\sum_{i=N+2-i_\lambda}^{N}\psi(E-x_i(t)) \tag{5.43}$$

for $E\in\mathbb{R}$, where ψ is a non-negative symmetric approximate delta-function on scale N^{-1}, i.e. it is supported in an N^{-1} neighbourhood of zero, $\int\psi=1$, $\|\psi\|_\infty\lesssim N$ and $\|\psi'\|_\infty\lesssim N^2$. Note that the total mass is $\int_{\mathbb{R}}\rho_{x,t}=2$. For the Stieltjes transform $m_{x,t}$ of $\rho_{x,t}$, we have $\sup_{z\in\mathbb{C}^+}|m_{x,t}(z)|\le C$ since the same bound holds for $\rho_{\lambda,t}$ by the shape analysis. Note that $\rho_{\lambda,t}$ is the semicircular flow with initial condition $\rho_{\lambda,t=0}=\rho_\lambda$ by definition, but $\rho_{x,t}$ is not exactly the semicircular evolution of $\rho_{x,0}$. We will not need this information, but in fact, the effect of the far away padding particles on the density near the cusp is very tiny.

Since $\rho_{x,t}$ coincides with $\rho_{\lambda,t}$ in a big finite interval, their edges and local minima near the cusp regime coincide, i.e we can identify

$$\mathfrak{e}^{\pm}_{x,t}=\mathfrak{e}^{\pm}_{\lambda,t},\qquad \mathfrak{m}_{x,t}=\mathfrak{m}_{\lambda,t}.$$

The shifted quantiles and semiquantiles $\widehat{\gamma}_{x,i}(t)$, $\widecheck{\gamma}_{x,i}(t)$ and $\widehat{\gamma}^*_{x,i}(t)$, $\widecheck{\gamma}^*_{x,i}(t)$ of $\rho_{x,t}$ are defined by the obvious analogues of the formulas (5.24)–(5.25) except that r subscript is replaced with x and the indices run over the entire range $1\le|i|\le N$. As before, $\gamma_{x,0}(t)=\mathfrak{e}^+_{x,t}$. The unshifted quantiles are defined by

$$\gamma_{x,i}(t)=\widehat{\gamma}_{x,i}(t)+\mathfrak{e}^+_{x,t},\quad 0\le t\le t_*,\qquad \gamma_{x,i}(t)=\widecheck{\gamma}_{x,i}(t)+\mathfrak{m}_{x,t},\quad t_*\le t\le 2t_*$$

and similarly for the semiquantiles.

So far we explained how to construct $\rho_{x,t}$ and its quantiles from $\rho_{\lambda,t}$, exactly in the same way we obtain $\rho_{y,t}$ from $\rho_{\mu,t}$ with straightforward notations.

Now for any $\alpha\in[0,1]$ we construct the α-interpolation of $\rho_{x,t}$ and $\rho_{y,t}$ that we will denote by $\overline{\rho}_t$. The bar will indicate quantities related to α-interpolation that implicitly depend on α; a dependence that we often omit from the notation. The interpolating measure will be constructed via its quantiles, i.e. we define

$$\overline{\gamma}_i(t):=\alpha\widehat{\gamma}_{x,i}(t)+(1-\alpha)\widehat{\gamma}_{y,i}(t),\qquad \overline{\gamma}^*_i(t):=\alpha\widehat{\gamma}^*_{x,i}(t)+(1-\alpha)\widehat{\gamma}^*_{y,i}(t) \tag{5.44}$$

for $1 \leq |i| \leq N, 0 \leq t \leq t_*$, and similarly for $t_* \leq t \leq 2t_*$ involving $\breve{\gamma}$'s. We also set the interpolating edges

$$\overline{\mathfrak{e}}_t^{\pm} = \alpha \mathfrak{e}_{x,t}^{\pm} + (1-\alpha)\mathfrak{e}_{y,t}^{\pm}. \tag{5.45}$$

Recall the parameter δ_* describing the size of a neighbourhood around the physical cusp where the shape analysis for ρ_λ and ρ_μ in Section 5.2 holds. Choose $i(\delta_*) \sim N$ such that $\left|\overline{\gamma}_{x,-i(\delta_*)}(t)\right| \leq \delta_*$ as well as $\left|\overline{\gamma}_{x,i(\delta_*)}(t)\right| \leq \delta_*$ hold for all $0 \leq t \leq 2t_*$. Then define, for any $E \in \mathbb{R}$, the function

$$\begin{aligned}\overline{\rho}_t(E) &:= \rho_{\alpha,t}(E) \cdot \mathbf{1}\big(\overline{\gamma}_{-i(\delta_*)}(t) + \overline{\mathfrak{e}}_t^+ \leq E \\ &\leq \overline{\gamma}_{i(\delta_*)}(t) + \overline{\mathfrak{e}}_t^+\big) + \frac{1}{N} \sum_{i(\delta_*)<|i|\leq N} \psi(E - \overline{\mathfrak{e}}_t^+ - \overline{\gamma}_i^*(t)),\end{aligned} \tag{5.46}$$

where $\rho_{\alpha,t}$ is the α-interpolation, constructed in Definition 5.4.1, between $\rho_{\lambda,t}(E) = \rho_{x,t}(E)$ and $\rho_{\mu,t}(E) = \rho_{y,t}(E)$ for $|E| \leq \delta_*$. By this construction (using also the symmetry of ψ) we know that all shifted semiquantiles of $\overline{\rho}_t$ are exactly $\overline{\gamma}_i^*(t)$. The same holds for all shifted quantiles $\overline{\gamma}_i(t)$ at least in the interval $[-\delta_*, \delta_*]$ since here $\overline{\rho}_t \equiv \rho_{\alpha,t}$ and the latter was constructed exactly by the requirement of linearity of the quantiles (5.44), see (5.26c).

We also record $\int \overline{\rho}_t = 2$ and that for the Stieltjes transform $\overline{m}_t(z)$ of $\overline{\rho}_t$ we have

$$\max_{|\Re z - \overline{\mathfrak{e}}_t^+| \leq \frac{1}{2}\delta_*} |\overline{m}_t(z)| \leq C \tag{5.47}$$

for all $0 \leq t \leq 2t_*$. The first bound follows easily from the same boundedness of the Stieltjes transform of $\rho_{\alpha,t}$. Moreover, $\overline{m}_t(z)$ is $\frac{1}{3}$-Hölder continuous in the regime $\left|\Re z - \overline{\mathfrak{e}}_t^+\right| \leq \frac{1}{2}\delta_*$ since in this regime $\overline{\rho}_t = \rho_{\alpha,t}$ and $\rho_{\alpha,t}$ is $\frac{1}{3}$-Hölder continuous by Lemma 5.4.3.

5.6 Rigidity for the short range approximation

In this section we consider Dyson Brownian Motion (DBM), i.e. a system of $2N$ coupled stochastic differential equations for $z(t) = \{z_i(t)\}_{[-N,N]\setminus\{0\}}$ of the form

$$\mathrm{d}z_i = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \frac{1}{N}\sum_j \frac{1}{z_i - z_j}\,\mathrm{d}t, \qquad 1 \leq |i| \leq N, \tag{5.48}$$

with some initial condition $z_i(t=0) = z_i(0)$, where

$$B(s) = (B_{-N}(s), \dots, B_{-1}(s), B_1(s) \dots, B_N(s))$$

is the vector of $2N$ independent standard Brownian motions. We use the indexing convention that all indices i, j, etc., run from $-N$ to N but zero index is excluded.

We will assume that $z_i(0)$ is an α-linear interpolation of $x_i(0), y_i(0)$ for some $\alpha \in [0,1]$:

$$z_i(0) = z_i(0,\alpha) := \alpha x_i(0) + (1-\alpha)y_i(0).$$

In the following of this section we will refer to the process defined by (5.48) using $z(t,\alpha)$ in order to underline the α dependence of the process. Clearly for $\alpha = 0, 1$ we recover the original $y(t)$ and $x(t)$ processes, $z(t, \alpha = 0) = y(t)$, $z(t, \alpha = 1) = x(t)$. For these processes we have the following optimal rigidity estimate that immediately follows from Corollary 4.2.6 and Lemma 5.5.1:

Lemma 5.6.1. *Let $r_i(t) = x_i(t)$ or $r_i(t) = y_i(t)$ and $r = x, y$. Then, there exists a fixed small $\epsilon > 0$, depending only on the model parameters, such that for each $1 \le |i| \le \epsilon N$, we have*

$$\sup_{0 \le t \le 2t_*} |r_i(t) - \gamma_{r,i}(t)| \le N^{\xi} \eta_{\mathrm{f}}^{\rho_{r,t}}(\gamma_{r,i}(t)), \tag{5.49}$$

for any $\xi > 0$ with very high probability, where we recall that the behavior of $\eta_{\mathrm{f}}^{\rho_{r,t}}(\gamma_{r,i}(t))$, with $r = x, y$, is given by (5.27a).

Note that, by (5.17a), (5.17c) and (5.27), for all $1 \le |i| \le \epsilon N$ and for all $0 \le t \le t_*$ we have that

$$\eta_{\mathrm{f}}^{\rho_{r,t}}(\gamma_{r,i}(t)) \lesssim \frac{N^{\frac{\omega_1}{6}}}{|i|^{\frac{1}{4}} N^{\frac{3}{4}}},$$

with $r = x, y$.

In particular, we know that $z(0, \alpha)$ lie close to the quantiles (5.44) of an *α-interpolating density* $\rho_z = \overline{\rho}_0$, see the definition in (5.46). This means that ρ_z has a small gap $[\mathfrak{e}_z^-, \mathfrak{e}_z^+]$ of size $\Delta_z \sim t_*^{3/2}$ (i.e. it will develop a physical cusp in a time of order t_*) and it is an α-interpolation between $\rho_{x,0}$ and $\rho_{y,0}$. Here interpolation refers to the process introduced in Section 5.5 that guarantees that the corresponding quantiles are convex linear combinations of the two initial densities with weights α and $1 - \alpha$, i.e.

$$\gamma_{z,i} = \alpha \gamma_{x,i} + (1 - \alpha) \gamma_{y,i}.$$

In this section we will prove rigidity results for $z(t, \alpha)$ and for its appropriate short range approximation. Since the group velocity of the entire cusp regime is different for $\rho_{x,t}$ and $\rho_{y,t}$, the interpolated process will have an intermediate group velocity. Since we have to follow the process for time scales $t \sim N^{-\frac{1}{2}+\omega_1}$, much bigger than the relevant rigidity scale $N^{-\frac{3}{4}}$ we have to determine the group velocity quite precisely. Technically, we will encode this information by defining an appropriately shifted process $\widetilde{z}(t, \alpha) = z(t, \alpha) - \mathrm{Shift}(t, \alpha)$. It is essential that the shift function is independent of the indices i to preserve the local statistics of the process. In the next section we explain how to choose the shift.

5.6.1 Choice of the shifted process $\widetilde{z}$

The remainder of Section 5.6 is formulated for the small gap regime, i.e. for $0 \le t \le t_*$. We will comment on the modifications in the small minimum regime in Section 5.8. To match the location of the gap, the natural guess would be to study the shifted process $z_i(t, \alpha) - \mathfrak{e}_{z,t}^+$ where $[\mathfrak{e}_{z,t}^-, \mathfrak{e}_{z,t}^+]$ is the gap of the semicircular evolution $\rho_{z,t}$ of ρ_z near the physical cusp, and approximate $z_i(t, \alpha) - \mathfrak{e}_{z,t}^+$ by the shifted semiquantiles $\widehat{\gamma}_{z,i}^*(t)$ of $\rho_{z,t}$. However, the evolution of the semicircular flow $t \to \rho_{z,t}$ near the cusp is not sufficiently well understood. We circumvent this technical problem by considering the quantiles of another approximating density $\overline{\rho}_t$ defined by the requirement that its quantiles are exactly the α-linear combinations of the quantiles of $\rho_{x,t}$ and $\rho_{y,t}$ as described in Section 5.5. The necessary regularity properties of $\overline{\rho}_t$ follow directly from its construction. The precise description below assumes that $0 \le t \le 2t_*$, i.e. we are in the small gap situation. For $t_* \le t \le t_*$ an identical construction works but the reference point $\mathfrak{e}_{r,t}^+$ is replaced with the approximate minimum $\widetilde{\mathfrak{m}}_{r,t}$, for $r = x, y$. For simplicity we present all formulas for $0 \le t \le t^*$ and we will comment on the other case in Section 5.8.

More concretely, for any fixed $\alpha \in [0,1]$ recall the (semi)quantiles from (5.44). These are the (semi)quantiles of the interpolating density $\overline{\rho} = \overline{\rho}_t$ defined in (5.46) and let its Stieltjes transform be denoted by $\overline{m} = \overline{m}_t$. Bar will refer to quantities related to this interpolation; implicitly all quantities marked by bar depend on the interpolation parameter α, which dependence will be omitted from the notation. Notice that $\overline{\rho}_t$ has a gap $[\overline{\mathfrak{e}}_t^-, \overline{\mathfrak{e}}_t^+]$ near the cusp satisfying (5.45). Initially at $t = 0$ we have $\overline{\rho}_{t=0} = \rho_z$, in particular $\overline{\gamma}_i(t = 0) = \widehat{\gamma}_{z,i}(t = 0)$ and $\overline{\mathfrak{e}}_0^\pm = \mathfrak{e}_z^\pm$. We will choose the shift in the definition of the $\widetilde{z}_i(t,\alpha)$ process so that we could use $\overline{\gamma}_i^*(t)$ to trail it.

The semicircular flow and the α-interpolation do not commute hence $\overline{\gamma}_i(t)$ are not the same as the quantiles $\widehat{\gamma}_{z,i}(t)$ of the semicircular evolution $\rho_{z,t}$ of the initial density ρ_z. We will, however, show that they are sufficiently close near the cusp and up to times relevant for us, modulo an irrelevant time dependent shift. Notice that the evolution of $\widehat{\gamma}_{z,i}(t)$ is hard to control since analysing $\frac{\mathrm{d}}{\mathrm{d}t}\widehat{\gamma}_{z,i}(t) = -\Re m_{z,t}(\gamma_{z,i}(t)) + \Re m_{z,t}(\mathfrak{e}_{z,t}^+)$ would involve knowing the evolved density $\rho_{z,t}$ quite precisely in the critical cusp regime. While this necessary information is in principle accessible from the explicit expression for the semicircular flow and the precise shape analysis of ρ_z obtained from that of ρ_x and ρ_y, here we chose a different, technically lighter path by using $\overline{\gamma}_i(t)$. Note that unlike $\widehat{\gamma}_{z,i}(t)$, the derivative of $\overline{\gamma}_i(t)$ involves only the Stieltjes transform of the densities $\rho_{x,t}$ and $\rho_{y,t}$ for which shape analysis is available.

However, the global group velocities of $\overline{\gamma}(t)$ and $\widehat{\gamma}_z(t)$ are not the same near the cusp. We thus need to define $\widetilde{z}(t,\alpha)$ not as $z(t,\alpha) - \overline{\mathfrak{e}}_t^+$ but with a modified time dependent shift to make up for this velocity difference so that $\overline{\gamma}(t)$ indeed correctly follows $\widetilde{z}(t,\alpha)$. To determine this shift, we first define the function

$$h^*(t,\alpha) := \Re\Big[-\overline{m}_t(\overline{\mathfrak{e}}_t^+) + (1-\alpha)m_{y,t}(\mathfrak{e}_{y,t}^+) + \alpha m_{x,t}(\mathfrak{e}_{x,t}^+)\Big], \tag{5.50}$$

where recall that $\overline{m}_t$ is the Stieltjes transform of the measure $\overline{\rho}_t$. Note that $h^*(t) = \mathcal{O}(1)$ following from the boundedness of the Stieltjes transforms $m_{x,t}$, $m_{y,t}$ and $\overline{m}_t(\overline{\mathfrak{e}}_t^+)$. The boundedness of $m_{x,t}$ and $m_{y,t}$ follows by (5.14) and $\left|\overline{m}_t(\overline{e}_t^+)\right| \le C$ by (5.47).

We note that

$$h^*(t,\alpha=0) = m_{y,t}(\mathfrak{e}_{y,t}^+) - \overline{m}_t(\overline{\mathfrak{e}}_t^+) = m_{y,t}(\mathfrak{e}_{y,t}^+) - \overline{m}_t(\mathfrak{e}_{y,t}^+)$$

since for $\alpha = 0$ we have $\mathfrak{e}_{y,t}^+ = \overline{\mathfrak{e}}_t^+$ by construction. At $\alpha = 0$ the measure $\overline{\rho}_t$ is given exactly by the density $\rho_{y,t}$ in an $\mathcal{O}(1)$ neighbourhood of the cusp. Away from the cusp, depending on the precise construction in the analogue of (5.46), the continuous $\rho_{y,t}$ is replaced by locally smoothed out Dirac measures at the quantiles. Similar statement holds at $\alpha = 1$. It is easy to see that the difference of the corresponding Stieltjes transforms evaluated at the cusp regime is of order N^{-1}, i.e.

$$h^*(t,\alpha=0) + h^*(t,\alpha=1) = \mathcal{O}\left(N^{-1}\right). \tag{5.51}$$

Since later in (5.132) we will need to give some very crude estimate on the α-derivative of $h^*(t,\alpha)$, but it actually blows up since $\overline{m}_t'$ is singular at the edge, we introduce a tiny regularization of h^*, i.e. we define the function

$$h^{**}(t,\alpha) := \Re\Big[-\overline{m}_t(\overline{\mathfrak{e}}_t^+ + \mathrm{i}N^{-100}) + (1-\alpha)m_{y,t}(\mathfrak{e}_{y,t}^+) + \alpha m_{x,t}(\mathfrak{e}_{x,t}^+)\Big]. \tag{5.52}$$

Note that by the $\frac{1}{3}$-Hölder continuity of $\overline{m}_t$ in the cusp regime, i.e. for $z \in \mathbb{H}$ such that $\left|\Re z - \overline{\mathfrak{e}}_t^+\right| \le \frac{\delta_*}{2}$, it follows that

$$h^{**}(t,\alpha) = h^*(t,\alpha) + \mathcal{O}\left(N^{-30}\right).$$

Then, we define

$$h(t) = h(t,\alpha) := h^{**}(t,\alpha) - \alpha h^{**}(t,1) - (1-\alpha)h^{**}(t,0) = \mathcal{O}\left(1\right)$$

to ensure that

$$h(t,\alpha=0) = h(t,\alpha=1) = 0.$$

In particular, we have

$$h(t,\alpha) = \Re\Big[-\overline{m}_t(\overline{\mathfrak{e}}_t^+) + (1-\alpha)m_{y,t}(\mathfrak{e}_{y,t}^+) + \alpha m_{x,t}(\mathfrak{e}_{x,t}^+)\Big] + \mathcal{O}\left(N^{-1}\right). \tag{5.53}$$

Define its antiderivative

$$H(t,\alpha) := \int_0^t h(s,\alpha)\,\mathrm{d}s, \qquad H(0,\alpha) = 0, \qquad \max_{0\le t\le t_*} |H(t,\alpha)| \lesssim N^{-1/2+\omega_1}.$$

Now we are ready to define the correctly shifted process

$$\widetilde{z}_i(t) = \widetilde{z}_i(t,\alpha) := z_i(t) - \left[\alpha\mathfrak{e}_{x,t}^+ + (1-\alpha)\mathfrak{e}_{y,t}^+\right] - H(t,\alpha), \tag{5.54}$$

that will be trailed by $\overline{\gamma}_i(t)$. It satisfies the shifted DBM

$$\mathrm{d}\widetilde{z}_i = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \left[\frac{1}{N}\sum_{j\ne i}\frac{1}{\widetilde{z}_i - \widetilde{z}_j} + \Phi_\alpha(t)\right]\mathrm{d}t \tag{5.55}$$

with

$$\Phi(t) := \Phi_\alpha(t) = \alpha\Re m_{x,t}(\mathfrak{e}_{x,t}^+) + (1-\alpha)\Re m_{y,t}(\mathfrak{e}_{y,t}^+) - h(t,\alpha), \tag{5.56}$$

and with initial conditions $\widetilde{z}(0) := z(0) - \mathfrak{e}_z^+$ by (5.45) and $H(0,\alpha) = 0$. The shift function satisfies

$$\Phi_\alpha(t) = \Re[\overline{m}_t(\overline{\mathfrak{e}}_t^+)] + \mathcal{O}\left(N^{-1}\right). \tag{5.57}$$

Notice that for $\alpha = 0, 1$ this definition gives back the naturally shifted $x(t)$ and $y(t)$ processes since we clearly have

$$\widetilde{z}(t,\alpha=1) = \widetilde{x}(t) := x(t) - \mathfrak{e}_{x,t}^+, \qquad \widetilde{z}(t,\alpha=0) = \widetilde{y}(t) := y(t) - \mathfrak{e}_{y,t}^+,$$

that are trailed by the shifted semiquantiles

$$\overline{\gamma}_i^*(t,\alpha=1) = \widehat{\gamma}_{x,i}^*(t) := \gamma_{x,i}^*(t) - \mathfrak{e}_{x,t}^+, \qquad \overline{\gamma}_i^*(t,\alpha=0) = \widehat{\gamma}_{y,i}^*(t) := \gamma_{y,i}^*(t) - \mathfrak{e}_{y,t}^+.$$

As we explained, the time dependent shift $H(t,\alpha)$ in (5.54) makes up for the difference between the true edge velocity of the semicircular flow (which we do not compute directly) and the naive guess which is $\frac{\mathrm{d}}{\mathrm{d}t}\left[\alpha\mathfrak{e}_{x,t}^+ + (1-\alpha)\mathfrak{e}_{y,t}^+\right]$ hinted by the linear combination procedure. The precise expression (5.50) will come out of the proof. The key point is that this adjustment is global, i.e. it is only time dependent but independent of i since this expresses a group velocity of the entire cusp regime.

In the following three subsections we prove an almost optimal rigidity not directly for $\widetilde{z}_i(t)$ but for its appropriate short range approximation $\widehat{z}_i(t)$. This will be sufficient for the proof of the universality. The proof of the rigidity will be divided into three phases, which we first explain informally, as follows.

Phase 1. (Subsection 5.6.2) The main result is a rigidity for $\widetilde{z}_i(t) - \overline{\gamma}_i(t)$ for $1 \le |i| \lesssim \sqrt{N}$ on scale $N^{-\frac{3}{4}+C\omega_1}$ without i-dependence in the error term. First we prove a crude rigidity on scale $N^{-1/2+C\omega_1}$ for all indices i. Using this rigidity, we can define a short range approximation $\mathring{z}$ of the original dynamics $\widetilde{z}$ and show that $\widetilde{z}_i$ and $\mathring{z}_i$ are close by $N^{-\frac{3}{4}+C\omega_1}$ for $1 \le |i| \lesssim \sqrt{N}$. Then we analyse the short range process $\mathring{z}$ that has a finite speed of propagation, so we can localize the dynamics. Finally, we can directly compare $\mathring{z}$ with a deterministic particle dynamics because the effect of the stochastic term $\sqrt{2/N}\,\mathrm{d}B_i$, i.e. $\sqrt{t_*/N} = N^{-3/4+\omega_1/2} \ll N^{-3/4+C\omega_1}$, remains below the rigidity scale of interest in this Phase 1.

However, to understand this deterministic particle dynamics we need to compare it with the corresponding continuum evolution; this boils down to estimating the difference of a Stieltjes transform and its Riemann sum approximation at the semiquantiles. Since the Stieltjes transform is given by a singular integral, this approximation relies on quite delicate cancellations which require some strong regularity properties of the density. We can easily guarantee this regularity by considering the density $\overline{\rho}_t$ of the linear interpolation between the quantiles of $\rho_{x,t}$ and $\rho_{y,t}$.

Phase 2. (Subsection 5.6.3) In this section we improve the rigidity from scale $N^{-\frac{3}{4}+C\omega_1}$ to scale $N^{-\frac{3}{4}+\frac{1}{6}\omega_1}$, for a smaller range of indices, but we can achieve this not for $\widetilde{z}$ directly, but for its short range approximation $\widehat{z}$. Unlike $\mathring{z}$ in Phase 1, this time we choose a very short scale approximation $\widehat{z}$ on scale $N^{4\omega_\ell}$ with $\omega_1 \ll \omega_\ell \ll 1$. As an input, we need the rigidity of $\widetilde{z}_i$ on scale $N^{-\frac{3}{4}+C\omega_1}$ for $1 \le |i| \lesssim \sqrt{N}$ obtained in Phase 1. We use heat kernel contraction for a direct comparison with the $y_i(t)$ dynamics for which we know optimal rigidity by Corollary 4.2.6, with the precise matching of the indices (*band rigidity*). In particular, when the gap is large, this guarantees that band rigidity is transferred to the $\widehat{z}$ process from the $\widehat{y}$ process.

Phase 3. (Subsection 5.6.4) Finally, we establish the optimal i-dependence in the rigidity estimate for $\widehat{z}_i$ from Phase 2, i.e. we get a precision $N^{-\frac{3}{4}+\frac{1}{6}\omega_1}\,|i|^{-1/4}$. The main method we use in Phase 3 is maximum principle. We compare $\widehat{z}_i$ with $\widehat{y}_{i-K}$, a slightly shifted element of the $\widehat{y}$ process, where $K = N^{\xi}$ with some tiny ξ. This method allows us to prove the optimal i-dependent rigidity (with a factor $N^{\frac{1}{6}\omega_1}$) but only for indices $|i| \gg K$ because otherwise $\widehat{z}_i$ and $\widehat{y}_{i-K}$ may be on different sides of the gap for small i. For very small indices, therefore, we need to rely on band rigidity for $\widehat{z}$ from Phase 2.

The optimal i-dependence allows us to replace the random particles $\widehat{z}$ by appropriate quantiles with a precision so that

$$|\widehat{z}_i - \widehat{z}_j| \lesssim N^{\frac{\omega_1}{6}} \left|\overline{\gamma}_i - \overline{\gamma}_j\right| \sim N^{-\frac{3}{4}+\frac{\omega_1}{6}} \left||i|^{\frac{3}{4}} - |j|^{\frac{3}{4}}\right|.$$

Such upper bound on $|\widehat{z}_i - \widehat{z}_j|$, hence a lower bound on the interaction kernel $\mathcal{B}_{ij} = |\widehat{z}_i - \widehat{z}_j|^{-2}$ of the differentiated DBM (see (5.128) later) with the correct dependence on the indices i, j, is essential since this gives the heat kernel contraction which eventually drives the precision below the rigidity scale in order to prove universality. On a time scale $t_* = N^{-\frac{1}{2}+\omega_1}$ the $\ell^p \to \ell^\infty$ contraction of the heat kernel gains a factor $N^{-\frac{4}{15}\omega_1}$ with the convenient choice of $p = 5$. Notice that $\frac{4}{15} > \frac{1}{6}$, so the contraction wins over the imprecision in the rigidity $N^{\frac{1}{6}\omega_1}$ from Phase 3, but not over $N^{C\omega_1}$ from Phase 1, showing that both Phase 2 and Phase 3 are indeed necessary.

5.6.2 Phase 1: Rigidity for $\widetilde{z}$n scale $N^{-3/4+C\omega_1}$.

The main result of this section is the following proposition:

Proposition 5.6.2. *Fix $\alpha \in [0,1]$. Let $\widetilde{z}(t,\alpha)$ solve (5.55) with initial condition $\widetilde{z}_i(0,\alpha)$ satisfying the crude rigidity bound for all indices*

$$\max_{1\le|i|\le N} |\widetilde{z}_i(0,\alpha) - \overline{\gamma}_i^*(0)| \lesssim N^{-1/2+2\omega_1}. \tag{5.58}$$

We also assume that

$$\|m_{x,0}\|_\infty + \|m_{y,0}\|_\infty + \left|\overline{m}_t(\mathfrak{e}_t^\pm)\right| \le C. \tag{5.59}$$

Then we have a weak but uniform rigidity

$$\sup_{0\le t\le t_*} \max_{1\le|i|\le N} |\widetilde{z}_i(t,\alpha) - \overline{\gamma}_i^*(t)| \lesssim N^{-1/2+2\omega_1}, \tag{5.60}$$

with very high probability. Moreover, for small $|i|$, i.e. $1 \le |i| \le i_*$, with $i_* := N^{1/2+C_*\omega_1}$ for some large $C_* > 100$, we have a stronger rigidity:

$$\sup_{0\le t\le t_*} \max_{1\le|i|\le i_*} |\widetilde{z}_i(t,\alpha) - \overline{\gamma}_i^*(t)| \lesssim \max_{1\le|i|\le 2i_*} |\widetilde{z}_i(0,\alpha) - \overline{\gamma}_i^*(0)| + \frac{N^{C\omega_1}}{N^{3/4}} \tag{5.61}$$

with very high probability.

In our application, (5.58) is satisfied and the right hand side of (5.61) is simply $N^{-\frac{3}{4}+C\omega_1}$ since

$$\widetilde{z}_i(0,\alpha)-\overline{\gamma}_i^*(0) = \alpha(x_i(0)-\gamma_{x,i}(0))+(1-\alpha)(y_i(0)-\gamma_{y,i}(0)) = \mathcal{O}\left(\frac{N^\xi N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}|i|^{\frac{1}{4}}}\right), \tag{5.62}$$

for any $\xi > 0$ with very high probability, by optimal rigidity for $x_i(0)$ and $y_i(0)$ from Corollary 4.2.6. Similarly, the assumption (5.59) is trivially satisfied by (5.47). However, we stated Proposition 5.6.2 under the slightly weaker conditions (5.58), (5.59) to highlight what is really needed for its proof.

Before starting the proof, we recall a formula on the derivative of the (shifted) semi-quantiles of a density which evolves by the semicircular flow:

$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{\gamma}_{i,r}^*(t) = -\Re m_{r,t}(\gamma_{r,i}^*(t)) + \Re m_{r,t}(\mathfrak{e}_{r,t}^+), \qquad r = x, y. \tag{5.63}$$

This formula is well known, it follows from differentiating the quantile equation

$$\int_{\mathfrak{e}^+_{r,t}}^{\gamma^*_{i,r}(t)} \rho_{r,t}(E)\,\mathrm{d}E = \frac{i-\frac{1}{2}}{N}, \qquad i>0$$

that, using the notation $\gamma := \gamma^*_{i,r}(t)$, gives

$$\dot\gamma = -\frac{1}{\rho(\gamma)}\int_{\mathfrak{e}^+}^{\gamma} \dot\rho(E)\,\mathrm{d}E$$

using that $\rho(\mathfrak{e}^+) = 0$. In the previous equality we denoted the time derivative $\frac{\mathrm{d}}{\mathrm{d}t}$ by dot. We also recall that from the defining equation (5.14) of the semicircular flow it follows that the Stieltjes transform $m = m_t(\zeta)$ of ρ_t satisfies the Burgers equation:

$$\dot m = mm' = \frac{1}{2}(m^2)',$$

where prime denotes the $\frac{\mathrm{d}}{\mathrm{d}\zeta}$ derivative. Thus

$$\begin{aligned}\dot\gamma &= -\frac{1}{\Im m(\gamma)}\Im\int_{\mathfrak{e}^+}^{\gamma} \dot m(E)\,\mathrm{d}E = -\frac{1}{\Im m(\gamma)}\frac{1}{2}\Im\int_{\mathfrak{e}^+}^{\gamma}(m^2)'(E)\,\mathrm{d}E\\ &= -\frac{1}{\Im m(\gamma)}\frac{1}{2}\Im m^2(\gamma) = -\Re m(\gamma).\end{aligned}$$

A similar formula holds for the time derivative of $\mathfrak{e}^+$, giving (5.63).

Proof of Proposition 5.6.2. We start with the proof of the crude rigidity (5.60), then we introduce a short range approximation and finally, with its help, we prove the refined rigidity (5.61). The main technical input of the last step is a refined estimate on the forcing term. These four steps will be presented in the next four subsections.

5.6.2.1 Proof of the crude rigidity:

For the proof of (5.60), using (5.63) twice in (5.44), we notice that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\overline{\gamma}^*_i(t) &= \alpha\big[-\Re m_{x,t}(\gamma^*_{x,i}(t)) + \Re m_{x,t}(\mathfrak{e}^+_{x,t})\big]\\ &\qquad + (1-\alpha)\big[-\Re m_{y,t}(\gamma^*_{y,i}(t)) + \Re m_{y,t}(\mathfrak{e}^+_{y,t})\big] = \mathcal{O}(1)\end{aligned}$$

since $m_{x,t}$ and $m_{y,t}$ are bounded recalling that the semicircular flow preserves (or reduces) the ℓ^∞ norm of the Stieltjes transform by (5.14), so $\|m_{x,t}\|_\infty \le \|m_{x,0}\|_\infty \le C$, similarly for $m_{y,t}$. This gives

$$|\overline{\gamma}^*_i(t) - \overline{\gamma}^*_i(0)| \lesssim N^{-1/2+\omega_1}. \tag{5.64}$$

Thus in order to prove (5.60) it is sufficient to prove

$$\|\widetilde{z}(t,\alpha) - \widetilde{z}(0,\alpha)\|_\infty \le N^{-1/2+2\omega_1}, \tag{5.65}$$

for any fixed $\alpha \in [0,1]$. To do that, we compare the dynamics of (5.55) with the dynamics of the y-semiquantiles, i.e. set

$$u_i := u_i(t,\alpha) = \widetilde{z}_i(t) - \widehat{\gamma}^*_{y,i}(t),$$

for all $0 \leq t \leq t_*$.

Compute

$$\mathrm{d}u_i = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + (\widetilde{\mathcal{B}}u)_i\,\mathrm{d}t + \widetilde{F}_i(t)\,\mathrm{d}t \tag{5.66}$$

with

$$(\widetilde{\mathcal{B}}f)_i := \frac{1}{N}\sum_{j\neq i}\frac{f_j - f_i}{(\widetilde{z}_i - \widetilde{z}_j)(\widehat{\gamma}^*_{y,i} - \widehat{\gamma}^*_{y,j})} \tag{5.67}$$

and

$$\widetilde{F}_i(t) := \frac{1}{N}\sum_{j\neq i}\frac{1}{\widehat{\gamma}^*_{y,i} - \widehat{\gamma}^*_{y,j}} + \Re m_{y,t}(\gamma^*_{y,i}(t)) + \alpha\big[\Re m_{x,t}(\mathfrak{e}^+_{x,t}) - \Re m_{y,t}(\mathfrak{e}^+_{y,t})\big] - h(t).$$

The operator $\widetilde{\mathcal{B}}$ is defined on $\mathbb{C}^{2N}$ and we label the vectors $f \in \mathbb{C}^{2N}$ as

$$f = (f_{-N}, f_{-N+1}, \dots, f_{-1}, f_1, \dots, f_N),$$

, i.e. we omit the $i = 0$ index. Accordingly, in the summations the $j = 0$ term is always omitted since $\widetilde{z}_j$, $\widehat{z}_j$ and $\widehat{\gamma}^*_{y,j}$ are defined for $1 \leq |j| \leq N$. Furthermore in the summation of the interaction terms, the $j = i$ term is always omitted.

We now show that

$$\left\|\widetilde{F}(t)\right\|_\infty \lesssim \log N, \qquad 0 \leq t \leq t^*. \tag{5.68}$$

By the boundedness of $m_{x,t}, m_{y,t}$ and the $1/3$-Hölder continuity of $\overline{m}_t$ in the cusp regime, it remains to control

$$\frac{1}{N}\sum_{j\neq i}\frac{1}{\widehat{\gamma}^*_{y,i}(t) - \widehat{\gamma}^*_{y,j}(t)} \lesssim \sum_{1\leq|j-i|\leq N}\frac{1}{|i-j|} \lesssim \log N$$

since $|\widehat{\gamma}^*_{y,j} - \widehat{\gamma}^*_{y,i}| \geq c\,|i-j|\,/N$ as the density $\rho_{y,t}$ is bounded.

Let $\widetilde{\mathcal{U}}(s,t)$ be the fundamental solution of the heat evolution with kernel $\widetilde{\mathcal{B}}$ from (5.67), i.e, for any $0 \leq s \leq t$

$$\partial_t\widetilde{\mathcal{U}}(s,t) = \widetilde{\mathcal{B}}(t)\widetilde{\mathcal{U}}(s,t), \qquad \widetilde{\mathcal{U}}(s,s) = I.$$

Note that $\widetilde{\mathcal{U}}$ is a contraction on every ℓ^p space and the same is true for its adjoint $\widetilde{\mathcal{U}}^*(s,t)$. In particular, for any indices a, b and times s, t we have

$$\widetilde{\mathcal{U}}_{ab}(s,t) \leq 1, \quad \widetilde{\mathcal{U}}^*_{ab}(s,t) \leq 1. \tag{5.69}$$

By Duhamel principle, the solution to the SDE (5.66) is given by

$$u(t) = \widetilde{\mathcal{U}}(0,t)u(0) + \sqrt{\frac{2}{N}}\int_0^t \widetilde{\mathcal{U}}(s,t)\,\mathrm{d}B(s) + \int_0^t \widetilde{\mathcal{U}}(s,t)\widetilde{F}(s)\,\mathrm{d}s, \tag{5.70}$$

where $B(s) = (B_{-N}(s), \dots, B_{-1}(s), B_1(s)\dots, B_N(s))$ are the $2N$ independent Brownian motions from (5.48).

For the second term in (5.70) we fix an index i and consider the martingale

$$M_t := \sqrt{\frac{2}{N}}\int_0^t \sum_j \widetilde{\mathcal{U}}_{ij}(s,t)\,\mathrm{d}B_j(s)$$

with its quadratic variation process

$$[M]_t := \frac{2}{N}\int_0^t \sum_j \big(\widetilde{\mathcal{U}}_{ij}(s,t)\big)^2 \,\mathrm{d}s = \frac{2}{N}\int_0^t \Big\|\widetilde{\mathcal{U}}^*(s,t)\delta_i\Big\|_2^2 \,\mathrm{d}s \le \frac{2t}{N}.$$

By the Burkholder maximal inequality for martingales, for any $p > 1$ we have that

$$\mathbf{E}\sup_{0\le t\le T} |M_t|^{2p} \le C_p\,\mathbf{E}[M]_T^p \le C_p\frac{T^p}{N^p}.$$

By Markov inequality we obtain that

$$\sup_{0\le t\le T}|M_t| \le N^{\xi}\sqrt{\frac{T}{N}} \tag{5.71}$$

with probability more than $1 - N^{-D}$, for any (large) $D > 0$ and (small) $\xi > 0$.

The last term in (5.70) is estimated, using (5.68), by

$$\left|\int_0^t \widetilde{\mathcal{U}}(s,t)\widetilde{F}(s)\,\mathrm{d}s\right| \le t\max_{s\le t}\Big\|\widetilde{F}(s)\Big\|_\infty \lesssim t\log N.$$

This, together with (5.71) and the contraction property of $\widetilde{\mathcal{B}}$ implies from (5.70) that

$$\|u(t)-u(0)\|_\infty \lesssim N^{-3/4+\omega_1} + t\log N \lesssim N^{-1/2+2\omega_1}$$

with very high probability. Recalling the definition of u and (5.64), we get (5.65) since

$$\|\widetilde{z}(t)-\widetilde{z}(0)\|_\infty \le \|u(t)-u(0)\|_\infty + \Big\|\widehat{\gamma}_y^*(t)-\widehat{\gamma}_y^*(0)\Big\|_\infty \lesssim N^{-1/2+2\omega_1}.$$

This completes the proof of the crude rigidity bound (5.60).

5.6.2.2 Crude short range approximation.

Now we turn to the proof of (5.61) by introducing a short range approximation of the dynamics (5.55). Fix an integer L. Let $\mathring{z}_i = \mathring{z}_i(t)$ solve the $\boldsymbol{L}$-localized short scale DBM

$$\mathrm{d}\mathring{z}_i = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \frac{1}{N}\sum_{j:|j-i|\le L}\frac{1}{\mathring{z}_i-\mathring{z}_j}\,\mathrm{d}t + \left[\frac{1}{N}\sum_{j:|j-i|>L}\frac{1}{\overline{\gamma}_i^*-\overline{\gamma}_j^*} + \Phi(t)\right]\mathrm{d}t \tag{5.72}$$

for each $1\le|i|\le N$ and with initial condition $\mathring{z}_i(0) := \widetilde{z}_i(0)$. Then, we have the following comparison:

Lemma 5.6.3. *Fix $\alpha\in[0,1]$. Assume that*

$$\max_{1\le|i|\le N}|\widetilde{z}_i(0,\alpha)-\overline{\gamma}_i^*(0)| \lesssim N^{-1/2+2\omega_1}. \tag{5.73}$$

Consider the short scale DBM (5.72) with a range $L = N^{1/2+C_1\omega_1}$ with a constant $10\le C_1\ll C_*$, in particular L is much smaller than i_*. Then we have a weak uniform comparison

$$\sup_{0\le t\le t_*}\max_{1\le|i|\le N}|\mathring{z}_i(t,\alpha)-\widetilde{z}_i(t,\alpha)| \lesssim N^{-1/2+2\omega_1}, \tag{5.74}$$

and a stronger comparison for small i:

$$\sup_{0\le t\le t_*}\max_{1\le|i|\le i_*}|\mathring{z}_i(t,\alpha)-\widetilde{z}_i(t,\alpha)|\lesssim N^{-3/4+C\omega_1}, \tag{5.75}$$

both with very high probability.

Proof. For any fixed $\alpha\in[0,1]$ and for all $0\le t\le t_*$, set $w:=w(t,\alpha)=\mathring{z}(t,\alpha)-\widetilde{z}(t,\alpha)$ and subtract (5.72) and (5.55) to get

$$\partial_t w=\mathring{\mathcal{B}}_1 w+\mathring{F},$$

where

$$(\mathring{\mathcal{B}}_1 f)_i:=\frac{1}{N}\sum_{j:|j-i|\le L}\frac{f_j-f_i}{(\mathring{z}_i-\mathring{z}_j)(\widetilde{z}_i-\widetilde{z}_j)},\qquad \mathring{F}_i:=\frac{1}{N}\sum_{j:|j-i|>L}\Big[\frac{1}{\overline{\gamma}^*_i-\overline{\gamma}^*_j}-\frac{1}{\widetilde{z}_i-\widetilde{z}_j}\Big].$$

We estimate

$$\big|\mathring{F}_i\big|\le\frac{1}{N}\sum_{j:|j-i|>L}\frac{|\widetilde{z}_i-\overline{\gamma}^*_i|+\big|\widetilde{z}_j-\overline{\gamma}^*_j\big|}{(\overline{\gamma}^*_i-\overline{\gamma}^*_j)(\widetilde{z}_i-\widetilde{z}_j)}\lesssim\frac{N^{-1/2+2\omega_1}}{N}\sum_{j:|j-i|>L}\frac{1}{(\overline{\gamma}^*_i-\overline{\gamma}^*_j)(\widetilde{z}_i-\widetilde{z}_j)},$$

where we used the crude rigidity (5.60) (applicable by (5.73)), and we chose C_1 in $L=N^{1/2+C_1\omega_1}$ large enough so that $\big|\overline{\gamma}^*_i-\overline{\gamma}^*_j\big|$ for any $|i-j|\ge L$ be much bigger than the rigidity scale $N^{-1/2+2\omega_1}$ in (5.60). This is guaranteed since

$$\big|\overline{\gamma}^*_i-\overline{\gamma}^*_j\big|=\alpha\,\big|\widehat{\gamma}^*_{x,i}-\widehat{\gamma}^*_{x,j}\big|+(1-\alpha)\,\big|\widehat{\gamma}^*_{y,i}-\widehat{\gamma}^*_{y,j}\big|\gtrsim\frac{|i-j|}{N}\gtrsim N^{-1/2+C_1\omega}$$

with very high probability. By this choice of L we have $|\widetilde{z}_i-\widetilde{z}_j|\sim\big|\overline{\gamma}^*_i-\overline{\gamma}^*_j\big|$ and therefore

$$\begin{aligned}\big|\mathring{F}_i\big|&\lesssim\frac{N^{-\frac12+2\omega_1}}{N}\sum_{j:|j-i|>L}\frac{1}{(\overline{\gamma}^*_i-\overline{\gamma}^*_j)^2}\\&\lesssim N^{1/2+2\omega_1}\sum_{j:|j-i|>L}\frac{1}{|i-j|^2}\lesssim N^{-(\frac12 C_1-2)\omega_1}\le 1,\quad\forall\,|i|\le N.\end{aligned}\tag{5.76}$$

Since $\mathcal{B}_1$ is positivity preserving, its evolution is a contraction, so by Duhamel formula, similarly to (5.70), we get

$$\|\mathring{z}(t)-\widetilde{z}(t)\|_\infty=\|w(t)\|_\infty\le\|w(0)\|_\infty+t\max_{s\le t}\big\|\mathring{F}(s)\big\|_\infty\lesssim N^{-1/2+\omega_1}$$

with very high probability.

Next, we proceed with the proof of (5.75).

In fact, for $1\le|i|\le 2i_*$, with i_* much bigger than L, we have a better bound:

$$\begin{aligned}\big|\mathring{F}_i\big|&\lesssim\frac{N^{-\frac12+2\omega_1}}{N}\sum_{j:|j-i|>L}\frac{1}{(\overline{\gamma}^*_i-\overline{\gamma}^*_j)^2}\\&\lesssim\sum_{j:|j-i|>L}\frac{N^{2\omega_1}}{\big|\,|i|^{3/4}-|j|^{3/4}\,\big|^2}\lesssim N^{-\frac14-(\frac12 C_1-2)\omega_1}\le N^{-\frac14},\quad|i|\le 2i_*,\end{aligned}\tag{5.77}$$

which we can use to get the better bound (5.75). To do so, we define a continuous interpolation $v(t,\beta)$ between $\widetilde{z}$ and $\mathring{z}$. More precisely, for any fixed $\beta \in [0,1]$ we set $v(t,\beta) = \{v(t,\beta)_i\}_{i=-N}^{N}$ as the solution to the SDE

$$\begin{aligned} \mathrm{d}v_i =& \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \frac{1}{N}\sum_{j:|j-i|\le L}\frac{1}{v_i - v_j}\,\mathrm{d}t + \Phi_\alpha(t)\,\mathrm{d}t \\ &+ \frac{1-\beta}{N}\sum_{j:|j-i|>L}\frac{1}{\widetilde{z}_i - \widetilde{z}_j}\,\mathrm{d}t + \frac{\beta}{N}\sum_{j:|j-i|>L}\frac{1}{\overline{\gamma}_i^* - \overline{\gamma}_j^*}\,\mathrm{d}t \end{aligned} \tag{5.78}$$

with initial condition $v(t=0,\beta) = (1-\beta)\widetilde{z}_i(0) + \beta\mathring{z}_i(0)$. Clearly $v(t,\beta=0) = \widetilde{z}(t)$ and $v(t,\beta=1) = \mathring{z}(t)$.

Differentiating in β, for $u := u(t,\beta) = \partial_\beta v(t,\beta)$ we obtain the SDE

$$\mathrm{d}u_i = (\mathcal{B}^v u)_i\,\mathrm{d}t + \mathring{F}_i\,\mathrm{d}t, \quad \text{with} \quad (\mathcal{B}^v f)_i := \frac{1}{N}\sum_{j:|j-i|\le L}\frac{f_j - f_i}{(v_i - v_j)^2},$$

with initial condition $u(t=0,\beta) = \mathring{z}(0) - \widetilde{z}(0) = 0$. By the contraction property of the heat evolution kernel $\mathcal{U}^v$ of $\mathcal{B}^v$, with a simple Duhamel formula, we have for any fixed β

$$\sup_{0\le t\le t_*}\|u(t,\beta)\|_\infty \le t_*\|\mathring{F}\|_\infty \le N^{-1/2+\frac{3}{2}\omega_1}, \tag{5.79}$$

with very high probability, where we used (5.76). After integration in β we get

$$\|v(t,\beta) - \overline{\gamma}^*(t)\|_\infty \le \|v(t,0) - \overline{\gamma}^*(t)\|_\infty + \left\|\int_0^\beta u(t,\beta')\,\mathrm{d}\beta'\right\|_\infty, \quad 0 \le t \le t_*. \tag{5.80}$$

From (5.79) we have

$$\mathbf{E}\left\|\int_0^\beta u(t,\beta')\,\mathrm{d}\beta'\right\|_\infty^p \le \int_0^\beta \mathbf{E}\,\|u(t,\beta')\|^p\,\mathrm{d}\beta' \lesssim (N^{-1/2+\frac{3}{2}\omega_1})^p \tag{5.81}$$

for any exponent p. Hence, using a high moment Markov inequality, we have

$$\mathbf{P}\left(\left\|\int_0^\beta u(t,\beta')\,\mathrm{d}\beta'\right\|_\infty \ge N^{-1/2+\frac{3}{2}\omega_1+\xi}\right) \le N^{-D} \tag{5.82}$$

for any (large) $D > 0$ and (small) $\xi > 0$ by choosing p large enough. Since $v(t,0) = \widetilde{z}(t)$, for which we have rigidity in (5.60), by (5.80) and (5.82) we conclude that

$$\sup_{0\le t\le t_*}\|v(t,\beta) - \overline{\gamma}^*(t)\|_\infty \lesssim N^{-\frac{1}{2}+2\omega_1}$$

with very high probability for any $\beta \in [0,1]$.

In particular L is much larger than the rigidity scale of $v = v(t,\beta)$. This means that

$$\Big||v_i - v_j| - |\overline{\gamma}_i^* - \overline{\gamma}_j^*|\Big| \lesssim N^{-\frac{1}{2}+2\omega_1}$$

and $\left|\overline{\gamma}^*_i - \overline{\gamma}^*_j\right| \gtrsim \frac{|i-j|}{N} \geq N^{-\frac{1}{2}+C_1\omega_1} \gg N^{-\frac{1}{2}+2\omega_1}$ whenever $|i-j| \geq L$, so we have

$$|v_i - v_j| \sim |\overline{\gamma}^*_i - \overline{\gamma}^*_j|, \qquad |i-j| \geq L. \tag{5.83}$$

Since i_* is much bigger than L and L is much larger than the rigidity scale of $v_i(t,\beta)$ in the sense of (5.83), the heat evolution kernel $\mathcal{U}^v$ satisfies the following finite speed of propagation estimate (the proof is given in Appendix 5.B):

Lemma 5.6.4. *With the notations above we have*

$$\sup_{0\leq s\leq t\leq t_*} [\mathcal{U}^v_{pi} + \mathcal{U}^v_{ip}] \leq N^{-D}, \qquad 1 \leq |i| \leq i_*, \quad |p| \geq 2i_* \tag{5.84}$$

for any D if N is sufficiently large.

Using a Duhamel formula again, for any fixed β, we have

$$u_i(t) = \sum_p \mathcal{U}^v_{ip} u_p(0) + \int_0^t \sum_p \mathcal{U}^v_{ip}(s,t) \mathring{F}_p(s)\,\mathrm{d}s.$$

We can split the summation and estimate

$$|u_i(t)| \leq \Big[\sum_{|p|\leq 2i_*} + \sum_{|p|>2i_*}\Big] \mathcal{U}^v_{ip}\,|u_p(0)| + \int_0^t \Big[\sum_{|p|\leq 2i_*} + \sum_{|p|>2i_*}\Big] \mathcal{U}^v_{ip}(s,t)|\mathring{F}_p(s)|\,\mathrm{d}s.$$

For $|i| \leq i_*$, the terms with $|p| > 2i_*$ are negligible by (5.84) and the trivial bounds (5.76) and (5.79). For $1 \leq |p| \leq 2i_*$ we use the improved bound (5.77). This gives

$$|u_i(t,\beta)| \leq \max_{1\leq|j|\leq 2i_*} |u_j(0,\beta)| + N^{-3/4+\omega_1} = N^{-3/4+\omega_1}, \qquad |i| \leq i_*,$$

since $u(t=0,\beta) = 0$. Integrating from $\beta = 0$ to $\beta = 1$, and recalling that $v(\beta = 0) = \widetilde{z}$ and $v(\beta = 1) = \mathring{z}$, by high moment Markov inequality, we conclude

$$|\widetilde{z}_i(t) - \mathring{z}_i(t)| \lesssim N^{-\frac{3}{4}+\omega_1}, \qquad 1 \leq |i| \leq i_*,$$

with very high probability. This yields (5.75) and completes the proof of Lemma 5.6.3.

We remark that it would have been sufficient to require that $|\widetilde{z}_j(0) - \mathring{z}_j(0)| \leq N^{-\frac{3}{4}+\omega_1}$ for all $1 \leq |j| \leq 2i_*$ instead of setting $\mathring{z}(0) := \widetilde{z}(0)$ initially. Later in Section 5.6.3 we will use a similar finite speed of propagation mechanism to show that changing the initial condition for large indices has negligible effect. □

5.6.2.3 Refined rigidity for small $|i|$.

Finally, in the last but main step of the proof of (5.61) in Proposition 5.6.2 we compare $\mathring{z}_i$ with $\overline{\gamma}^*_i$ for small $|i|$ with a much higher precision than the crude bound $N^{-\frac{1}{2}+C\omega_1}$ which directly follows from (5.74) and (5.60). Notice that we use the semiquantiles for comparison since $\overline{\gamma}^*_i \in [\overline{\gamma}_{i-1}, \overline{\gamma}_i]$ and $\overline{\gamma}^*_i$ is typically close to the midpoint of this interval. In particular, $\overline{\rho}_t(\overline{\gamma}^*_i(t))$ is never zero, in fact we have $\overline{\rho}_t(\overline{\gamma}^*_i(t)) \geq cN^{-1/3}$, because by band rigidity quantiles may fall exactly at spectral edges, but semiquantiles cannot. This lower bound makes the semiquantiles much more convenient reference points than the quantiles.

Proposition 5.6.5. *Fix $\alpha \in [0,1]$, then with the notations above for the localized DBM $\mathring{z}(t,\alpha)$ on short scale $L = N^{1/2+C_1\omega_1}$ with $10 \le C_1 \le \frac{1}{10}C_*$, defined in (5.72), we have*

$$\big|(\mathring{z}_i(t,\alpha) - \overline{\gamma}_i^*(t)) - (\mathring{z}_i(0,\alpha) - \overline{\gamma}_i^*(0))\big| \le N^{-3/4+C\omega_1}, \quad 1 \le |i| \le i_*, \tag{5.85}$$

where $i_* = N^{\frac{1}{2}+C_*\omega_1}$, with very high probability.

Combining (5.85) with (5.75) and noticing that

$$\mathring{z}_i(0,\alpha) - \overline{\gamma}_i^*(0) = \widetilde{z}_i(0,\alpha) - \overline{\gamma}_i^*(0) = \mathcal{O}\left(\frac{N^{\xi}N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}\,|i|^{\frac{1}{4}}}\right)$$

for any $\xi > 0$ with very high probability by (5.62), we obtain (5.61) and complete the proof of Proposition 5.6.2. □

Proof of Proposition 5.6.5. We recall from (5.63) that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\overline{\gamma}_i^*(t) = \alpha\big[-\Re m_{x,t}(\gamma^*_{x,i}(t)) + \Re m_{x,t}(\mathfrak{e}^+_{x,t})\big] \\ + (1-\alpha)\big[-\Re m_{y,t}(\gamma^*_{y,i}(t)) + \Re m_{y,t}(\mathfrak{e}^+_{y,t})\big].\end{aligned} \tag{5.86}$$

Next, we define a dynamics that interpolates between $\mathring{z}_i(t,\alpha)$ and $\overline{\gamma}_i^*(t)$, i.e. between (5.72) and (5.86). Let $\beta \in [0,1]$ and for any fixed β define the process $v = v(t,\beta) = \{v_i(t,\beta)\}_{i=-N}^N$ as the solution of the following interpolating DBM

$$\begin{aligned}\mathrm{d}v_i =& \beta\sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \frac{1}{N}\sum_{j:|j-i|\le L}\frac{1}{v_i - v_j}\,\mathrm{d}t + \beta\left[\frac{1}{N}\sum_{j:|j-i|>L}\frac{1}{\overline{\gamma}_i^* - \overline{\gamma}_j^*}\,\mathrm{d}t + \Phi(t)\right]\mathrm{d}t \\ &+ (1-\beta)\left[\frac{\mathrm{d}}{\mathrm{d}t}\overline{\gamma}_i^*(t) - \frac{1}{N}\sum_{j:|j-i|\le L}\frac{1}{\overline{\gamma}_i^* - \overline{\gamma}_j^*}\right]\mathrm{d}t, \qquad 1 \le |i| \le N,\end{aligned} \tag{5.87}$$

with initial condition $v_i(0,\beta) := \beta\mathring{z}_i(0) + (1-\beta)\overline{\gamma}_i^*(0)$. Notice that

$$v_i(t,\beta=0) = \overline{\gamma}_i^*(t), \qquad v_i(t,\beta=1) = \mathring{z}_i(t). \tag{5.88}$$

Here we use the same letter v as in (5.78) within the proof of Lemma 5.6.3, but this is now a new interpolation. Since both appearances of the letter v are used only within the proofs of separate lemmas, this should not cause any confusion. The same remark applies to the letter u below.

Let $u := u(t,\beta) = \partial_\beta v(t,\beta)$, then it satisfies the equation

$$\mathrm{d}u_i = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \sum_{j\ne i}\mathcal{B}_{ij}(u_i - u_j)\,\mathrm{d}t + F_i\,\mathrm{d}t, \qquad 1 \le |i| \le N, \tag{5.89}$$

with a time dependent short range kernel (omitting the time argument and the β parameter)

$$\mathcal{B}_{ij}(t) = \mathcal{B}_{ij} := -\frac{1}{N}\frac{\mathbf{1}(|i-j| \le L)}{(v_i - v_j)^2} \tag{5.90}$$

and external force

$$
\begin{aligned}
F_i = F_i(t) :={}& -\frac{1}{N}\sum_j \frac{1}{\overline{\gamma}_j^*(t) - \overline{\gamma}_i^*(t)} + \alpha \Re m_{x,t}(\gamma_{x,i}^*(t)) \\
&+ (1-\alpha)\Re m_{y,t}(\gamma_{y,i}^*(t)) - h(t,\alpha), \quad 1 \le |i| \le N.
\end{aligned}
\tag{5.91}
$$

Since the density $\overline{\rho}$ is regular, at least near the cusp regime, we can replace the sum over j with an integral with very high precision for small i; this integral is $\Re\overline{m}(\overline{\mathfrak{e}}^+ + \overline{\gamma}_i^*)$. A simple rearrangement of various terms yields

$$
F_i = \left[\Re\overline{m}(\overline{\mathfrak{e}}^+ + \overline{\gamma}_i^*) - \frac{1}{N}\sum_j \frac{1}{\overline{\gamma}_{j\neq i}^* - \overline{\gamma}_i^*}\right] - (1-\alpha)D_{y,i} - \alpha D_{x,i} + \mathcal{O}\left(N^{-1}\right), \tag{5.92}
$$

with

$$
D_{r,i} := \Re\Big[\big(\overline{m}(\overline{\mathfrak{e}}^+ + \overline{\gamma}_i^*) - \overline{m}(\overline{\mathfrak{e}}^+)\big) - \big(m_r(\gamma_{r,i}^*) - m_r(\mathfrak{e}_r^+)\big)\Big], \qquad r = x, y,
$$

where we used the formula for h from (5.53) and the definition of Φ from (5.56). The choice of the shift h was governed by the idea to replace the last three terms in (5.91) by $\Re\overline{m}(\overline{\mathfrak{e}}^+ + \overline{\gamma}_i^*)$. However, the shift cannot be i dependent as it would result in an i-dependent shift in the definition of $\widetilde{z}_i$, see (5.54), which would mean that the differences (gaps) of the processes z_i and $\widetilde{z}_i$ are not the same. Therefore, we defined the shift $h(t)$ by the similar formula evaluated at the edge, justifying the choice (5.53). The discrepancy is expressed by $D_{x,i}$ and $D_{y,i}$ which are small. Indeed we have, for $r = x, y$ and $1 \le |i| \le 2i_*$ that

$$
\begin{aligned}
|D_{r,i}| \le{}& \left|\Re\Big[\big(\overline{m}(\overline{\mathfrak{e}}^+ + \widehat{\gamma}_{r,i}^*) - \overline{m}(\overline{\mathfrak{e}}^+)\big) - \big(m_r(\mathfrak{e}_r^+ + \widehat{\gamma}_{r,i}^*) - m_r(\mathfrak{e}_r^+)\big)\Big]\right| \\
&+ \left|\overline{m}(\overline{\mathfrak{e}}^+ + \widehat{\gamma}_{r,i}^*) - \overline{m}(\overline{\mathfrak{e}}^+ + \overline{\gamma}_i^*)\right| \\
\lesssim{}& |\widehat{\gamma}_{r,i}^*|^{1/3}\left[|\widehat{\gamma}_{r,i}^*|^{1/3} + N^{-\frac{1}{6}+\frac{\omega_1}{3}}\right]\left|\log|\widehat{\gamma}_{r,i}^*|\right| + N^{-\frac{11}{36}+\omega_1} + \frac{|\widehat{\gamma}_{r,i}^* - \overline{\gamma}_i^*|}{\overline{\rho}(\overline{\gamma}_i^*)^2} \\
\lesssim{}& \left[\left(\frac{|i|}{N}\right)^{1/2} + \left(\frac{|i|}{N}\right)^{1/4} N^{-\frac{1}{6}+\frac{\omega_1}{3}}\right](\log N) + N^{-\frac{11}{36}+\omega_1} \\
&+ \frac{(|i|/N) + (|i|/N)^{3/4}N^{-\frac{1}{6}+\omega_1}}{(|i|/N)^{1/2}} \lesssim N^{-\frac{1}{4}+C\omega_1},
\end{aligned}
\tag{5.93}
$$

where from the first to the second line we used (5.32a) and the bound on the derivative of $\overline{m}$, see (5.18b). In the last inequality we used (5.26a) to estimate $|\widehat{\gamma}_{r,i}^*| \lesssim (|i|/N)^{3/4}N^{C\omega_1}$ and similarly $|\widehat{\gamma}_{r,i}^* - \overline{\gamma}_i^*|$ in the regime $|i| \le i_* = N^{\frac{1}{2}+C_*\omega_1}$, furthermore we used that $\overline{\rho}(\overline{\gamma}_i^*) \ge (|i|/N)^{1/4}$ and also $|\overline{\gamma}_i^*| \ge c/N$, since a semiquantile is always away from the edge.

Let $\mathcal{U}(s,t)$ be the fundamental solution of the heat evolution with kernel $\mathcal{B}$ from (5.90). Similarly to (5.70), the solution to the SDE (5.89) is given by

$$
u(t) = \mathcal{U}(0,t)u + \sqrt{\frac{2}{N}}\int_0^t \mathcal{U}(s,t)\,\mathrm{d}B(s) + \int_0^t \mathcal{U}(s,t)F(s)\,\mathrm{d}s. \tag{5.94}
$$

The middle martingale term can be estimated as in (5.71). The last term in (5.94) is estimated by

$$\left|\int_0^t \mathcal{U}(s,t)F(s)\,\mathrm{d}s\right| \le t \max_{0\le s\le t} \|F(s)\|_\infty . \tag{5.95}$$

First we use these simple Duhamel bounds to obtain a crude rigidity bound on $v_i(t,\beta)$ by integrating the bound on u

$$\begin{aligned}|v_i(t,\beta)-v_i(t,\beta=0)| &\le \beta \max_{\beta'\in[0,\beta]} |u_i(t,\beta')| \\ &\le \max_{\beta'\in[0,1]} \left\|u(0,\beta')\right\|_\infty + N^{-1/2+\omega_1+\xi}, \qquad 1\le |i|\le N,\end{aligned} \tag{5.96}$$

for any $\xi>0$ with very high probability, using (5.71), (5.94), (5.95) and that $\mathcal{U}$ is a contraction. Note that in the first inequality of (5.96) we used that it holds with very high probability by Markov inequality as in (5.81)-(5.82). We also used the trivial bound

$$\max_{0\le s\le t_*} \|F(s)\|_\infty \lesssim \log L \sim \log N, \tag{5.97}$$

which easily follows from (5.91), (5.93) and the fact that $\left|\overline{\gamma}_j^*(t)-\overline{\gamma}_i^*(t)\right| \gtrsim |i-j|/N$.

Recalling that $v_i(t,\beta=0)=\overline{\gamma}_i^*(t)$ and $u_i(0,\beta')=\mathring{z}_i(0)-\overline{\gamma}_i^*(0)$, together with (5.74) and (5.60), by (5.96), we obtain the crude rigidity

$$|v_i(t,\beta)-\overline{\gamma}_i^*(t)| \le N^{-\frac{1}{2}+2\omega_1}, \qquad 1\le |i|\le N, \tag{5.98}$$

with very high probability.

The main technical result is a considerable improvement of the bound (5.98) at least for i near the cusp regime. This is the content of the following proposition whose proof is postponed:

Proposition 5.6.6. *The vector F defined in* (5.91) *satisfies the bound*

$$\max_{s\le t_*} |F_i(s)| \le N^{-\frac{1}{4}+C\omega_1}, \qquad 1\le |i|\le 2i_*. \tag{5.99}$$

Since i_* is much bigger than $L=N^{\frac{1}{2}+C_1\omega_1}$ with a large C_1, and we have the rigidity (5.98) on scale much smaller than L, similarly to Lemma 5.6.4, we have the following finite speed of propagation result. The proof is identical to that of Lemma 5.6.4.

Proposition 5.6.7. *For the short range dynamics $\mathcal{U}=\mathcal{U}^{\mathcal{B}}$ defined by the operator* (5.90)*:*

$$\sup_{0\le s\le t\le t_*} \left[\mathcal{U}_{pi}(s,t)+\mathcal{U}_{ip}(s,t)\right] \le N^{-D}, \qquad 1\le |i|\le i_*, \quad |p|\ge 2i_*. \tag{5.100}$$

for any D if N is sufficiently large. □

Armed with these two propositions, we can easily complete the proof of Proposition 5.6.5. For any $1\le |i|\le i_*$ we have from (5.70), using (5.69), (5.71), (5.100) and that $\mathcal{U}$ is a contraction on ℓ^∞ that

$$\begin{aligned}|u_i(t)| \le& N^{-3/4+\omega_1+\xi} + \sum_p \mathcal{U}_{ip}\,|u_p(0)| + \int_0^t \sum_p \mathcal{U}_{ip}(s,t)\,|F_p(s)|\,\mathrm{d}s \\ \le& N^{-3/4+\omega_1+\xi} + \max_{|p|\le 2i_*} |u_p(0)| + t \max_{0\le s\le t_*}\max_{|p|\le 2i_*} |F_p(s)| + N^{-D}\max_{0\le s\le t} \|F(s)\|_\infty .\end{aligned} \tag{5.101}$$

The trivial bound (5.97) together with (5.99) completes the proof of (5.85) by integrating back the bound (5.101) for $u = \partial_\beta v$ in β, using a high moment Markov inequality similar to (5.81)-(5.82), and recalling (5.88). This completes the proof of Proposition 5.6.5. □

5.6.2.4 Estimate of the forcing term.

Proof of Proposition 5.6.6. Within this proof we will use $\gamma_i := \overline{\gamma}_i(t)$, $\gamma_i^* := \overline{\gamma}_i^*(t)$, $\rho = \overline{\rho}_t$, $m = \overline{m}_t$ and $\mathfrak{e}^+ = \overline{\mathfrak{e}}_t^+$ for brevity. For notational simplicity we may assume within this proof that $\mathfrak{e}^+ = 0$ by a simple shift. The key input is the following bound on the derivative of the density, proven in [12] for self-consistent densities of Wigner type matrices

$$|\rho'(x)| \leq \frac{C}{\rho(x)[\rho(x) + \Delta^{1/3}]}, \qquad |x| \leq \delta_* \tag{5.102}$$

where $\Delta = \overline{\Delta}_t$ is the length of the unique gap in the support of $\rho = \overline{\rho}_t$ in a small neighbourhood of size $\delta_* \sim 1$ around $\mathfrak{e}^+ = 0$. If there is no such gap, then we set $\Delta = 0$ in (5.102). By the definition of the interpolated density $\overline{\rho}_t$ in (5.46) clearly follows that it satisfies (5.102) by (5.4.3). Notice that (5.102) implies local Hölder continuity, i.e.

$$|\rho(x) - \rho(y)| \leq \min\left\{ |x-y|^{1/3}, |x-y|^{1/2} \Delta^{-1/6} \right\} \tag{5.103}$$

for any x, y in a small neighbourhood of the gap or the local minimum.

Throughout the entire proof we fix an i with $1 \leq |i| \leq 2i_*$. For simplicity, we assume $i > 0$, the case $i < 0$ is analogous. We rewrite F_i from (5.92) as follows

$$F_i = G_1 + G_2 + G_3 + G_4 \tag{5.104}$$

with

$$G_1 := \sum_{1 \leq |j-i| \leq L} \int_{\gamma_{j-1}}^{\gamma_j} \left[\frac{1}{x - \gamma_i^*} - \frac{1}{\gamma_j^* - \gamma_i^*} \right] \rho(x)\,\mathrm{d}x, \quad G_2 := \int_{\gamma_{i-1}}^{\gamma_i} \frac{\rho(x)\,\mathrm{d}x}{x - \gamma_i^*},$$
$$G_3 := \sum_{|j-i| > L} \int_{\gamma_{j-1}}^{\gamma_j} \left[\frac{1}{x - \gamma_i^*} - \frac{1}{\gamma_j^* - \gamma_i^*} \right] \rho(x)\,\mathrm{d}x,$$
$$G_4 := -(1-\alpha) D_{y,i} - \alpha D_{x,i} + \mathcal{O}\left(\frac{1}{N}\right).$$

The term G_4 was already estimated in (5.93). In the following we will show separately that $|G_a| \lesssim N^{-1/4}, a = 1, 2, 3$.

Estimate of G_3. By elementary computations, using the crude rigidity (5.60), it follows that

$$|G_3| \lesssim \frac{N^{-\frac{1}{2}+2\omega_1}}{N} \sum_{j : |j-i| > L} \frac{1}{(\gamma_i^* - \gamma_j^*)^2}.$$

Then, the estimate $|G_3| \lesssim N^{-\frac{1}{4}}$ follows using the same computations as in (5.77).

Estimate of G_2. We write

$$G_2 = \int_{\gamma_{i-1}}^{\gamma_i} \frac{\rho(x)\,\mathrm{d}x}{x - \gamma_i^*} = \int_{\gamma_{i-1}}^{\gamma_i} \frac{\rho(x) - \rho(\gamma_i^*)}{x - \gamma_i^*}\,\mathrm{d}x + \rho(\gamma_i^*) \int_{\gamma_{i-1}}^{\gamma_i} \frac{\mathrm{d}x}{x - \gamma_i^*} \tag{5.105}$$

and we will show that both summands are bounded by $CN^{-1/4}$. We make the convention that if γ_{i-1} is exactly at the left edge of a gap, then for the purpose of this proof we redefine it to be the right edge of the same gap and similarly, if γ_i is exactly at the right edge of the gap, then we set it to be left edge. This is just to make sure that $[\gamma_{i-1}, \gamma_i]$ is always included in the support of ρ.

In the first integral we use (5.103) to get

$$\left| \int_{\gamma_{i-1}}^{\gamma_i} \frac{\rho(x) - \rho(\gamma_i^*)}{x - \gamma_i^*} \,\mathrm{d}x \right| \lesssim \min\left\{ (\gamma_i - \gamma_{i-1})^{1/3}, (\gamma_i - \gamma_{i-1})^{1/2} \Delta^{-1/6} \right\} = \mathcal{O}\left(N^{-1/4} \right). \tag{5.106}$$

Here we used that the local eigenvalue spacing (with the convention above) is bounded by

$$\gamma_i - \gamma_{i-1} \lesssim \max\left\{ \frac{\Delta^{1/9}}{N^{2/3}}, \frac{1}{N^{3/4}} \right\}. \tag{5.107}$$

For the second integral in (5.105) is an explicit calculation

$$\rho(\gamma_i^*) \int_{\gamma_{i-1}}^{\gamma_i} \frac{\mathrm{d}x}{x - \gamma_i^*} = \rho(\gamma_i^*) \log \frac{\gamma_i - \gamma_i^*}{\gamma_i^* - \gamma_{i-1}}. \tag{5.108}$$

Using the definition of the quantiles and (5.103), we have

$$\begin{aligned} \frac{1}{2N} &= \int_{\gamma_{i-1}}^{\gamma_i^*} \rho(x) \,\mathrm{d}x \\ &= \rho(\gamma_i^*)(\gamma_i^* - \gamma_{i-1}) + \mathcal{O}\left(\min\left\{ |\gamma_i^* - \gamma_{i-1}|^{4/3}, |\gamma_i^* - \gamma_{i-1}|^{3/2} \Delta^{-1/6} \right\} \right), \end{aligned}$$

and similarly

$$\frac{1}{2N} = \int_{\gamma_i^*}^{\gamma_i} \rho(x) \,\mathrm{d}x = \rho(\gamma_i^*)(\gamma_i - \gamma_i^*) + \mathcal{O}\left(\min\left\{ |\gamma_i^* - \gamma_i|^{4/3}, |\gamma_i^* - \gamma_i|^{3/2} \Delta^{-1/6} \right\} \right).$$

The error terms are comparable and they are $\mathcal{O}\left(N^{-1}\right)$ using (5.107), thus, subtracting these two equations, we have

$$|(\gamma_i - \gamma_i^*) - (\gamma_i^* - \gamma_{i-1})| \lesssim \frac{\min\left\{ |\gamma_i^* - \gamma_i|^{4/3}, |\gamma_i^* - \gamma_i|^{3/2} \Delta^{-1/6} \right\}}{\rho(\gamma_i^*)}.$$

Expanding the logarithm in (5.108), we have

$$\begin{aligned} \left| \rho(\gamma_i^*) \int_{\gamma_{i-1}}^{\gamma_i} \frac{\mathrm{d}x}{x - \gamma_i^*} \right| &\lesssim \rho(\gamma_i^*) \frac{|(\gamma_i - \gamma_i^*) - (\gamma_i^* - \gamma_{i-1})|}{\gamma_i^* - \gamma_{i-1}} \\ &\lesssim \min\left\{ |\gamma_i^* - \gamma_i|^{1/3}, |\gamma_i^* - \gamma_i|^{1/2} \Delta^{-1/6} \right\} \lesssim N^{-1/4} \end{aligned}$$

as in (5.106). This completes the estimate

$$|G_2| \lesssim N^{-1/4}. \tag{5.109}$$

Estimate of G_1. Fix $i > 0$ and set $n = n(i)$ as follows

$$n(i) := \min\left\{ n \in \mathbb{N} \mid \min\left\{ |\gamma_{i-n-1} - \gamma_i^*|, |\gamma_{i+n} - \gamma_i^*| \right\} \geq cN^{-3/4} \right\} \tag{5.110}$$

with some small constant $c > 0$.

Next, we estimate $n(i)$. Notice that for $i = 1$ we have $n(i) = 0$. If $i \geq 2$, then we notice that one can choose c sufficiently small depending only on the model parameters, such that

$$\frac{1}{2} \leq \frac{\rho(x)}{\rho(\gamma_i^*)} \leq 2 \ : \ \forall x \in [\gamma_{i-n(i)-1}, \gamma_{i+n(i)}], \quad i \geq 2. \tag{5.111}$$

Let

$$m(i) := \max\Big\{ m \in \mathbb{N} \ : \ \frac{1}{2} \leq \frac{\rho(x)}{\rho(\gamma_i^*)} \leq 2 \ : \ \forall x \in [\gamma_{i-m-1}, \gamma_{i+m}] \Big\},$$

then, in order to verify (5.111), we need to prove that $m(i) \geq n(i)$.

Then by a case by case calculation it follows that $m(i) \geq c_1 |i|$ and thus

$$\min\Big\{ \left|\gamma_{i-m(i)-1} - \gamma_i^*\right|, \left|\gamma_{i+m(i)} - \gamma_i^*\right| \Big\} \gtrsim \max\Big\{ \Big(\frac{i}{N}\Big)^{2/3} \Delta^{1/9}, \Big(\frac{i}{N}\Big)^{3/4} \Big\} \geq c_2 N^{-3/4}.$$

with some c_1, c_2. Hence (5.111) will hold if $c \leq c_2$ is chosen in the definition (5.110). Notice that in these estimates it is important that the semiquantiles are always at a certain distance away from the quantiles.

Now we give an upper bound on $n(i)$ when γ_i^* is near a (possible small) gap as in the proof above. The local eigenvalue spacing is

$$\gamma_i - \gamma_i^* \sim \max\Big\{ \frac{\Delta^{1/9}}{N^{2/3}(i)^{1/3}}, \frac{1}{N^{3/4}(i)^{1/4}} \Big\},$$

which is bigger than $cN^{-3/4}$ if $i \leq \Delta^{1/3} N^{1/4}$. So in this case $n(i) = 0$ and we may now assume that $i \geq \Delta^{1/3} N^{1/4}$ and still $i \geq 2$.

Consider first the so-called *cusp case* when $i \geq N\Delta^{4/3}$, in this case, as long as $n \leq \frac{1}{2} i$, we have

$$\gamma_{i+n} - \gamma_i^* \sim \frac{n}{N^{3/4}(i+1)^{1/4}}.$$

This is bigger than $cN^{-3/4}$ if $n \geq i^{1/4}$, thus we have $n(i) \leq i^{1/4}$ in this case.

In the opposite case, the so-called *edge case*, $i \leq N\Delta^{4/3}$, which together with the above assumption $i \geq \Delta^{1/3} N^{1/4}$ also implies that $\Delta \geq N^{-3/4}$. In this case, as long as $n \leq \frac{1}{2} i$, we have

$$\gamma_{i+n} - \gamma_i^* \sim \frac{n\Delta^{1/9}}{N^{2/3} i^{1/3}}.$$

This is bigger than $cN^{-3/4}$ if $n \geq \Delta^{-1/9} N^{-1/12} i^{1/3}$. So we have $n(i) \leq \Delta^{-1/9} N^{-1/12} i^{1/3} \leq i^{1/3}$ in this case.

We split the sum in the definition of G_1, see (5.104), as follows:

$$\begin{aligned} G_1 &= \sum_{1 \leq |j-i| \leq L} \int_{\gamma_{j-1}}^{\gamma_j} \frac{x - \gamma_j^*}{(\gamma_i^* - \gamma_j^*)(x - \gamma_i^*)} \rho(x) \, \mathrm{d}x \\ &= \Big(\sum_{n(i) < |j-i| \leq L} + \sum_{1 \leq |j-i| \leq n(i)} \Big) =: S_1 + S_2. \end{aligned} \tag{5.112}$$

For the first sum we use $|x-\gamma_j^*|\le \gamma_{j+1}^*-\gamma_j^*$, $|\gamma_i^*-x|\sim|\gamma_i^*-\gamma_j^*|$. Moreover, we have

$$\rho(\gamma_i^*)(\gamma_i-\gamma_{i-1})\sim\frac{1}{N}$$

from the definition of the semiquantiles. Thus we restore the integration in the first sum S_1 and estimate

$$\begin{aligned}|S_1| &\lesssim \frac{1}{N}\Big[\int_{-\infty}^{\gamma_{i-n(i)-1}}+\int_{\gamma_{i+n(i)}}^{\infty}\Big]\frac{\mathrm{d}x}{|x-\gamma_i^*|^2}\\ &\lesssim\frac{1}{N}\Big[\frac{1}{|\gamma_{i-n(i)-1}-\gamma_i^*|}+\frac{1}{|\gamma_{i+n(i)}-\gamma_i^*|}\Big]\le CN^{-1/4}.\end{aligned}$$

In the last step we used the definition of $n(i)$.

Now we consider S_2. Notice that this sum is non-empty only if $n(i)\neq 0$ In this case to estimate S_2 we have to symmetrize. Fix $1\le n\le n(i)$, assume $i>n$ and consider together

$$\begin{aligned}&\int_{\gamma_{i-n-1}}^{\gamma_{i-n}}\frac{x-\gamma_{i-n}^*}{(\gamma_i^*-\gamma_{i-n}^*)(x-\gamma_i^*)}\rho(x)\,\mathrm{d}x+\int_{\gamma_{i+n-1}}^{\gamma_{i+n}}\frac{x-\gamma_{i+n}^*}{(\gamma_i^*-\gamma_{i+n}^*)(x-\gamma_i^*)}\rho(x)\,\mathrm{d}x\\ &=\frac{1}{\gamma_i^*-\gamma_{i-n}^*}\int_{\gamma_{i-n-1}}^{\gamma_{i-n}}\frac{x-\gamma_{i-n}^*}{x-\gamma_i^*}\rho(x)\,\mathrm{d}x+\frac{1}{\gamma_i^*-\gamma_{i+n}^*}\int_{\gamma_{i+n-1}}^{\gamma_{i+n}}\frac{x-\gamma_{i+n}^*}{x-\gamma_i^*}\rho(x)\,\mathrm{d}x\\ &=\frac{1}{N}\Big[\frac{1}{\gamma_i^*-\gamma_{i-n}^*}+\frac{1}{\gamma_i^*-\gamma_{i+n}^*}\Big]+\left[\int_{\gamma_{i-n-1}}^{\gamma_{i-n}}\frac{\rho(x)\,\mathrm{d}y}{x-\gamma_i^*}+\int_{\gamma_{i+n-1}}^{\gamma_{i+n}}\frac{\rho(x)\,\mathrm{d}x}{x-\gamma_i^*}\right]\\ &=:B_1(n)+B_2(n).\end{aligned}\tag{5.113}$$

We now use $\frac{1}{3}$-Hölder regularity

$$\rho(x)=\rho(\gamma_i^*)+\mathcal{O}\left(|x-\gamma_i^*|^{1/3}\right).$$

We thus have

$$\sum_{n\le n(i)}\int_{\gamma_{i-n-1}}^{\gamma_{i-n}}\frac{\rho(x)\,\mathrm{d}y}{x-\gamma_i^*}=\sum_{n\le n(i)}\rho(\gamma_i^*)\log\frac{\gamma_{i-n-1}-\gamma_i^*}{\gamma_{i-n}-\gamma_i^*}+\mathcal{O}\left(\int_{\gamma_{i-n(i)-1}}^{\gamma_{i+n(i)}}\frac{\mathrm{d}x}{|x-\gamma_i^*|^{2/3}}\right)$$

and similarly

$$\sum_{n\le n(i)}\int_{\gamma_{i+n-1}}^{\gamma_{i+n}}\frac{\rho(x)\,\mathrm{d}y}{x-\gamma_i^*}=\sum_{n\le n(i)}\rho(\gamma_i^*)\log\frac{\gamma_{i+n-1}-\gamma_i^*}{\gamma_{i+n}-\gamma_i^*}+\mathcal{O}\left(\int_{\gamma_{i-n(i)-1}}^{\gamma_{i+n(i)}}\frac{\mathrm{d}x}{|x-\gamma_i^*|^{2/3}}\right).$$

The error terms are bounded by $CN^{-1/4}$ using (5.110) and therefore we have

$$\sum_{n\le n(i)}B_2(n)=\sum_{n\le n(i)}\rho(\gamma_i^*)\Big[\log\frac{\gamma_i^*-\gamma_{i-n-1}}{\gamma_i^*-\gamma_{i-n}}-\log\frac{\gamma_{i+n}-\gamma_i^*}{\gamma_{i+n-1}-\gamma_i^*}\Big]+\mathcal{O}\left(N^{-1/4}\right)$$

$$=\sum_{n\le n(i)}\rho(\gamma_i^*)\Big[\log\frac{\gamma_i^*-\gamma_{i-n-1}}{\gamma_{i+n}-\gamma_i^*}+\log\frac{\gamma_{i+n-1}-\gamma_i^*}{\gamma_i^*-\gamma_{i-n}}\Big]+\mathcal{O}\left(N^{-1/4}\right).$$

We now use the bound

$$|\rho(x)-\rho(\gamma_i^*)| \lesssim \frac{|x-\gamma_i^*|}{\rho(\gamma_i^*)^2+\rho(\gamma_i^*)\Delta^{1/3}}, \qquad x\in[\gamma_{i-n(i)-1},\gamma_{i+n(i)}],$$

which follows from the derivative bound (5.102) if ϵ in the definition of $i_* = \epsilon N$ is chosen sufficiently small, depending on δ since throughout the proof $1 \le |i| \le 2i_*$ and $n(i) \ll i_*$.

Note that

$$\frac{n}{N} = \int_{\gamma_{i-n}}^{\gamma_i} \rho(x)\,\mathrm{d}x = \rho(\gamma_i^*)[\gamma_i-\gamma_{i-n}] + \mathcal{O}\left(\frac{|\gamma_{i-n}-\gamma_i^*|^2}{\rho(\gamma_i^*)^2+\rho(\gamma_i^*)\Delta^{1/3}}\right) \tag{5.114}$$

Thus, using (5.114) also for $\gamma_{i+n}-\gamma_i$, equating the two equations and dividing by $\rho(\gamma_i^*)$, we have

$$\gamma_i-\gamma_{i-n} = \gamma_{i+n}-\gamma_i + \mathcal{O}\left(\frac{|\gamma_{i-n}-\gamma_i^*|^2}{\rho(\gamma_i^*)^3+\rho(\gamma_i^*)^2\Delta^{1/3}}\right).$$

Similar relation holds for the semiquantiles:

$$\gamma_i^*-\gamma_{i-n}^* = \gamma_{i+n}^*-\gamma_i^* + \mathcal{O}\left(\frac{|\gamma_{i-n}^*-\gamma_i^*|^2}{\rho(\gamma_i^*)^3+\rho(\gamma_i^*)^2\Delta^{1/3}}\right) \tag{5.115}$$

and for the mixed relations among quantiles and semiquantiles:

$$\gamma_i^*-\gamma_{i-n} = \gamma_{i+n-1}-\gamma_i^* + \mathcal{O}\left(\frac{|\gamma_{i-n}-\gamma_i^*|^2}{\rho(\gamma_i^*)^3+\rho(\gamma_i^*)^2\Delta^{1/3}}\right)$$

$$\gamma_i^*-\gamma_{i-n-1} = \gamma_{i+n}-\gamma_i^* + \mathcal{O}\left(\frac{|\gamma_{i-n}-\gamma_i^*|^2}{\rho(\gamma_i^*)^3+\rho(\gamma_i^*)^2\Delta^{1/3}}\right).$$

Thus, using $\gamma_i^*-\gamma_{i-n-1}\sim\gamma_{i+n}-\gamma_i^*$, we have

$$\begin{aligned}\rho(\gamma_i^*)\left|\log\frac{\gamma_i^*-\gamma_{i-n-1}}{\gamma_{i+n}-\gamma_i^*}\right| &\lesssim \frac{\rho(\gamma_i^*)}{\gamma_{i+n}-\gamma_i^*}\mathcal{O}\left(\frac{|\gamma_{i-n-1}-\gamma_i^*|^2}{\rho(\gamma_i^*)^3+\rho(\gamma_i^*)^2\Delta^{1/3}}\right)\\ &\lesssim \frac{|\gamma_{i-n-1}-\gamma_i^*|}{\rho(\gamma_i^*)^2+\rho(\gamma_i^*)\Delta^{1/3}}.\end{aligned}$$

Using $n \le n(i)$ and (5.110), we have $|\gamma_{i-n-1}-\gamma_i^*| \lesssim N^{-3/4}$. The contribution of this term to $\sum_n B_2(n)$ is thus

$$N^{-3/4}\sum_{n\le n(i)}\frac{1}{\rho(\gamma_i^*)^2+\rho(\gamma_i^*)\Delta^{1/3}} \le \frac{n(i)N^{-3/4}}{\rho(\gamma_i^*)^2+\rho(\gamma_i^*)\Delta^{1/3}}. \tag{5.116}$$

In the bulk regime we have $\rho(\gamma_i^*)\sim 1$ and $n(i)\sim N^{1/4}$, so this contribution is much smaller than $N^{-1/4}$.

In the cusp regime, i.e. when $\Delta \le (i/N)^{3/4}$, then we have $\gamma_i^*\sim(i/N)^{3/4}$ and $\rho(\gamma_i^*)\sim(i/N)^{1/4}$, thus we get

$$(5.116) \le \frac{n(i)N^{-3/4}}{\rho(\gamma_i^*)^2+\rho(\gamma_i^*)\Delta^{1/3}} \le \frac{n(i)N^{-3/4}}{\rho(\gamma_i^*)^2} \lesssim N^{-1/4}n(i)i^{-1/2} \lesssim N^{-1/4}$$

since $n(i) \leq i^{1/4}$.

In the edge regime, i.e. when $\Delta \geq (i/N)^{3/4}$, then we have $\gamma_i^* \sim \Delta^{1/9}(i/N)^{2/3}$ and $\rho(\gamma_i^*) \sim \Delta^{-1/9}(i/N)^{1/3}$, thus we get

$$(5.116) \leq \frac{n(i)N^{-3/4}}{\rho(\gamma_i^*)^2 + \rho(\gamma_i^*)\Delta^{1/3}} \leq \frac{n(i)N^{-3/4}}{\rho(\gamma_i^*)\Delta^{1/3}} \lesssim \frac{n(i)N^{-5/12}}{\Delta^{2/9}i^{1/3}} \leq \frac{N^{-5/12}}{\Delta^{2/9}} \leq N^{-1/4}$$

since $n(i) \leq i^{1/3}$ and $\Delta \geq N^{-3/4}$. This completes the proof of $\sum_n B_2(n) \lesssim N^{-1/4}$.

Finally the $\sum_n B_1(n)$ term from (5.113) is estimated as follows by using (5.115):

$$\begin{aligned}
&\sum_n \frac{1}{N}\Big[\frac{1}{\gamma_i^* - \gamma_{i-n-1}^*} + \frac{1}{\gamma_i^* - \gamma_{i+n-1}^*}\Big] \\
&= \sum_n \frac{1}{N}\frac{1}{(\gamma_i^* - \gamma_{i-n}^*)^2}\mathcal{O}\left(\frac{(\gamma_i - \gamma_{i-n-1})^2}{\rho(\gamma_i^*)^2[\rho(\gamma_i^*) + \Delta^{1/3}]}\right) \lesssim \frac{n(i)}{N\rho(\gamma_i^*)^2[\rho(\gamma_i^*) + \Delta^{1/3}]}.
\end{aligned}$$

In the bulk regime this is trivially bounded by $CN^{-3/4}$. In the cusp regime, $\Delta \leq (i/N)^{3/4}$, we have

$$\frac{n(i)}{N\rho(\gamma_i^*)^2[\rho(\gamma_i^*) + \Delta^{1/3}]} \leq \frac{n(i)}{N\rho(\gamma_i^*)^3} \lesssim \frac{n(i)}{N^{1/4}i^{3/4}} \lesssim N^{-1/4}$$

since $n(i) \leq i^{1/4}$.

Finally, in the edge regime, $\Delta \geq (i/N)^{3/4}$, we just use

$$\frac{n(i)}{N\rho(\gamma_i^*)^2[\rho(\gamma_i^*) + \Delta^{1/3}]} \leq \frac{n(i)}{N\rho(\gamma_i^*)^2\Delta^{1/3}} \lesssim \frac{n(i)}{N^{1/4}i^{3/4}} \lesssim N^{-1/4}$$

since $n(i) \leq i^{1/3}$. This gives $\sum_n B_1(n) \lesssim N^{-1/4}$. Together with the estimate on $\sum_n B_2(n)$ we get $|S_2| \lesssim N^{-1/4}$, see (5.112) and (5.113). This completes the estimate of G_1 in (5.104), which, together with (5.109) and (5.93) finishes the proof of Proposition 5.6.6. □

5.6.3 Phase 2: Rigidity of $\widehat{z}$ on scale $N^{-3/4+\omega_1/6}$, without i dependence

For any fixed $\alpha \in [0,1]$ recall the definition of the shifted process $\widetilde{z}(t,\alpha)$ (5.55) and the shifted α-interpolating semiquantiles $\overline{\gamma}_i^*(t)$ from (5.44) that trail $\widetilde{z}$. Furthermore, for all $0 \leq t \leq t_*$ we consider the interpolated density $\overline{\rho}_t$ with a small gap $[\overline{\mathfrak{e}}_t^-, \overline{\mathfrak{e}}_t^+]$, and its Stieltjes transform $\overline{m}_t$. In particular,

$$\overline{\mathfrak{e}}_t^{\pm} = \alpha\mathfrak{e}_{x,t}^{\pm} + (1-\alpha)\mathfrak{e}_{y,t}^{\pm}.$$

We recall that by Proposition 5.6.2 and (5.62) we have that

$$\sup_{0\leq t\leq t_*}\max_{1\leq|i|\leq i_*} |\widetilde{z}_i(t,\alpha) - \overline{\gamma}_i^*(t)| \leq N^{-\frac{3}{4}+C\omega_1}, \tag{5.117}$$

holds with very high probability for some $i_* = N^{\frac{1}{2}+C_*\omega_1}$.

In this section we improve the rigidity (5.117) from scale $N^{-\frac{3}{4}+C\omega_1}$ to the almost optimal, but still i-independent rigidity of order $N^{-\frac{3}{4}+\frac{\omega_1}{6}+\xi}$ but only for a new short range approximation $\widehat{z}_i(t,\alpha)$ of $\widetilde{z}_i(t,\alpha)$. The range of this new approximation $\ell^4 = N^{4\omega_\ell}$ with some $\omega_\ell \ll 1$ is much shorter than that of $\mathring{z}_i(t,\alpha)$ in Section 5.6.2. Furthermore, the result will hold only for $1 \leq |i| \leq N^{4\omega_\ell+\delta_1}$, for some small $\delta_1 > 0$. The rigorous statement is in Proposition 5.6.9 below, after we give the definition of the short range approximation.

5.6.3.1 Short range approximation on fine scale.

Adapting the idea of [122] to the cusp regime, we now introduce a new short range approximation process $\widehat{z}(t,\alpha)$ for the solution to (5.55). The short range approximation in this section will always be denoted by hat, $\widehat{z}$, in distinction to the other short range approximation $\mathring{z}$ used in Section 5.6.2, see (5.72). Not only the length scale is shorter for $\widehat{z}$, but the definition of $\widehat{z}$ is more subtle than in (5.72)

The new short scale approximation is characterized by two exponents ω_ℓ and ω_A. In particular, we will always assume that $\omega_1 \ll \omega_\ell \ll \omega_A \ll 1$, where recall that $t_* \sim N^{-\frac{1}{2}+\omega_1}$ is defined in such a way $\overline{\rho}_{t_*}$ has an exact cusp. The key quantity is $\ell := N^{\omega_\ell}$ that determines the scale on which the interaction term in (5.55) will be cut off and replaced by its mean-field value. This scale is not constant, it increases away from the cusp at a certain rate. The cutoff will be effective only near the cusp, for indices beyond $\frac{i_*}{2}$, with $i_* = N^{\frac{1}{2}+C_*\omega_1}$, no cutoff is made. Finally, the intermediate scale N^{ω_A} is used for a technical reason: closer to the cusp, for indices less than N^{ω_A}, we always use the density $\rho_{y,t}$ of the reference process $y(t)$ to define the mean field approximation of the cutoff long range terms. Beyond this scale we use the actual density $\overline{\rho}_t$. In this way we can exploit the closeness of the density $\overline{\rho}_t$ to the reference density $\rho_{y,t}$ near the cusp and simplify the estimate. This choice will guarantee that the error term ζ_0 in (5.127) below is non zero only for $|i| > N^{\omega_A}$.

Now we define the $\widehat{z}$ process precisely. Let

$$\mathcal{A} := \left\{(i,j) : |i-j| \le \ell(10\ell^3 + |i|^{\frac{3}{4}} + |j|^{\frac{3}{4}})\right\} \cup \left\{(i,j) : |i|, |j| > \frac{i_*}{2}\right\}.$$

One can easily check that for each i with $1 \le |i| \le \frac{i_*}{2}$ the set $\{j : (i,j) \in \mathcal{A}\}$ is an interval of the nonzero integers and that $(i,j) \in \mathcal{A}$ if and only if $(j,i) \in \mathcal{A}$. For each such fixed i we denote the smallest and the biggest j such that $(i,j) \in \mathcal{A}$ by $j_-(i)$ and $j_+(i)$, respectively. We will use the notation

$$\sum_j^{\mathcal{A},(i)} := \sum_{\substack{j:(i,j)\in\mathcal{A}\\ i\neq j}}, \qquad \sum_j^{\mathcal{A}^c,(i)} := \sum_{j:(i,j)\notin\mathcal{A}}.$$

Assuming for simplicity that i_* is divisible by four, we introduce the intervals

$$\mathcal{J}_z(t) := \left[\overline{\gamma}_{-\frac{3i_*}{4}}(t), \overline{\gamma}_{\frac{3i_*}{4}}(t)\right], \tag{5.118}$$

and for each $0 < |i| \le \frac{i_*}{2}$ we define

$$\mathcal{I}_{z,i}(t) := [\overline{\gamma}_{j_-(i)}(t), \overline{\gamma}_{j_+(i)}(t)]. \tag{5.119}$$

For a fixed $\alpha \in [0,1]$ and $N \ge |i| > \frac{i_*}{2}$ we let

$$\begin{aligned} \mathrm{d}\widehat{z}_i(t,\alpha) = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \Bigg[\frac{1}{N}\sum_j^{\mathcal{A},(i)} \frac{1}{\widehat{z}_i(t,\alpha)-\widehat{z}_j(t,\alpha)} \\ + \frac{1}{N}\sum_j^{\mathcal{A}^c,(i)} \frac{1}{\widetilde{z}_i(t,\alpha)-\widetilde{z}_j(t,\alpha)} + \Phi_\alpha(t)\Bigg]\mathrm{d}t \end{aligned} \tag{5.120}$$

for $0 < |i| \leq N^{\omega_A}$

$$\mathrm{d}\widehat{z}_i(t,\alpha) = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \Bigg[\frac{1}{N}\sum_j^{\mathcal{A},(i)} \frac{1}{\widehat{z}_i(t,\alpha)-\widehat{z}_j(t,\alpha)} \\ + \int_{\mathcal{I}_{y,i}(t)^c} \frac{\rho_{y,t}(E+\mathfrak{e}^+_{y,t})}{\widehat{z}_i(t,\alpha)-E}\,\mathrm{d}E + \Re[m_{y,t}(\mathfrak{e}^+_{y,t})]\Bigg]\,\mathrm{d}t, \tag{5.121}$$

and for $N^{\omega_A} < |i| \leq \frac{i_*}{2}$

$$\mathrm{d}\widehat{z}_i(t,\alpha) = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \Bigg[\frac{1}{N}\sum_j^{\mathcal{A},(i)} \frac{1}{\widehat{z}_i(t,\alpha)-\widehat{z}_j(t)} + \int_{\mathcal{I}_{z,i}(t)^c\cap\mathcal{J}_z(t)} \frac{\overline{\rho}_t(E+\mathfrak{e}^+_t)}{\widehat{z}_i(t,\alpha)-E}\,\mathrm{d}E \\ + \frac{1}{N}\sum_{|j|\geq\frac{3}{4}i_*} \frac{1}{\widetilde{z}_i(t,\alpha)-\widetilde{z}_j(t,\alpha)} + \Phi_\alpha(t)\Bigg]\,\mathrm{d}t, \tag{5.122}$$

with initial data

$$\widehat{z}_i(0,\alpha) := \widetilde{z}_i(0,\alpha), \tag{5.123}$$

where we recall that $\widetilde{z}_i(0,\alpha) = \alpha\widetilde{x}_i(0) + (1-\alpha)\widetilde{y}_i(0)$ for any $\alpha \in [0,1]$. In particular, $\widehat{z}(t,1) = \widehat{x}(t)$ and $\widehat{z}(t,0) = \widehat{y}(t)$, that are the short range approximations of the $\widetilde{x}(t) := x(t) - \mathfrak{e}^+_{x,t}$ and $\widetilde{y}(t) := x(t) - \mathfrak{e}^+_{y,t}$ processes.

Using the rigidity estimates in (5.60) and (5.117) we will prove the following lemma in Appendix 5.C.

Lemma 5.6.8. *Assuming that the rigidity estimates* (5.60) *and* (5.117) *hold. Then, for any fixed* $\alpha \in [0,1]$ *we have*

$$\sup_{1\leq|i|\leq N}\ \sup_{0\leq t\leq t_*} |\widehat{z}_i(t,\alpha) - \widetilde{z}_i(t,\alpha)| \leq \frac{N^{C\omega_1}}{N^{\frac{3}{4}}}, \tag{5.124}$$

with very high probability.

In particular, since (5.60) and (5.117) have already been proven, we conclude from (5.117) and (5.124) that

$$\sup_{0\leq t\leq t_*} |\widehat{z}_i(t,\alpha) - \overline{\gamma}_i(t)| \leq \frac{N^{C\omega_1}}{N^{\frac{3}{4}}}, \qquad 1 \leq |i| \leq i_*, \tag{5.125}$$

for any fixed $\alpha \in [0,1]$.

Now we state the improved rigidity for $\widehat{z}$, the main result of Section 5.6.3:

Proposition 5.6.9. *Fix any* $\alpha \in [0,1]$. *There exists a constant* $C > 0$ *such that if* $0 < \delta_1 < C\omega_\ell$ *then*

$$\sup_{0\leq t\leq t_*} |\widehat{z}_i(t,\alpha) - \overline{\gamma}_i(t)| \lesssim \frac{N^\xi N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}}, \qquad 1 \leq |i| \leq N^{4\omega_\ell+\delta_1} \tag{5.126}$$

for any $\xi > 0$ *with very high probability.*

Proof. Recall that initially $\widetilde{z}_i(0,\alpha)$ is a linear interpolation between $\widetilde{x}_i(0)$ and $\widetilde{y}_i(0)$ and thus for $\widetilde{z}_i(0,\alpha)$ optimal rigidity (5.62) holds. We define the derivative process

$$w_i(t,\alpha) := \partial_\alpha \widehat{z}_i(t,\alpha).$$

In particular, we find that $w = w(t,\alpha)$ is the solution of

$$\partial_t w = \mathcal{L} w + \zeta^{(0)}, \qquad \mathcal{L} := \mathcal{B} + \mathcal{V}, \tag{5.127}$$

with initial data

$$w_i(0,\alpha) = \widehat{x}_i(0) - \widehat{y}_i(0).$$

Here, for any $1 \leq |i| \leq N$, the (short range) operator $\mathcal{B}$ is defined on any vector $f \in \mathbb{C}^{2N}$ as

$$(\mathcal{B}f)_i := \sum_j^{\mathcal{A},(i)} \mathcal{B}_{ij}(f_i - f_j), \qquad \mathcal{B}_{ij} := -\frac{1}{N}\frac{1}{(\widehat{z}_i(t,\alpha) - \widehat{z}_j(t,\alpha))^2}. \tag{5.128}$$

Moreover, $\mathcal{V}$ is a multiplication operator, i.e. $(\mathcal{V}f)_i = \mathcal{V}_i f_i$, where $\mathcal{V}_i$ is defined in different regimes of i as follows:

$$\begin{aligned} \mathcal{V}_i &:= -\int_{\mathcal{I}_{y,i}(t)^c} \frac{\rho_{y,t}(E + \mathfrak{e}_{y,t}^+)}{(\widehat{z}_i(t,\alpha) - E)^2}\,\mathrm{d}E, \qquad 1 \leq |i| \leq N^{\omega_A} \\ \mathcal{V}_i &:= -\int_{\mathcal{I}_{z,i}(t)^c \cap \mathcal{J}_z(t)} \frac{\overline{\rho}_t(E + \overline{\mathfrak{e}}_t^+)}{(\widehat{z}_i(t,\alpha) - E)^2}\,\mathrm{d}E, \qquad N^{\omega_A} < |i| \leq \frac{i_*}{2} \end{aligned} \tag{5.129}$$

and $\mathcal{V}_i = 0$ for $|i| > \frac{i_*}{2}$. The error term $\zeta_i^{(0)} = \zeta_i^{(0)}(t)$ in (5.127) is defined as follows: for $|i| > \frac{i_*}{2}$ we have

$$\zeta_i^{(0)} := \frac{1}{N}\sum_j^{\mathcal{A}^c,(i)} \frac{\partial_\alpha \widetilde{z}_j(t,\alpha) - \partial_\alpha \widetilde{z}_i(t,\alpha)}{(\widetilde{z}_i(t,\alpha) - \widetilde{z}_j(t,\alpha))^2} + \partial_\alpha \Phi_\alpha(t) =: Z_1 + \partial_\alpha \Phi_\alpha(t) \tag{5.130}$$

for $N^{\omega_A} < |i| \leq \frac{i_*}{2}$ we have

$$\begin{aligned} \zeta_i^{(0)} := \frac{1}{N}\sum_{|j| \geq \frac{3i_*}{4}} \frac{\partial_\alpha \widetilde{z}_j(t,\alpha) - \partial_\alpha \widetilde{z}_i(t,\alpha)}{(\widetilde{z}_i(t,\alpha) - \widetilde{z}_j(t,\alpha))^2} + \int_{\mathcal{I}_{z,i}(t)^c \cap \mathcal{J}_z(t)} \frac{\partial_\alpha[\overline{\rho}_t(E + \overline{\mathfrak{e}}_t^+)]}{\widehat{z}_i(t,\alpha) - E}\,\mathrm{d}E \quad (5.131) \\ + \Big(\partial_\alpha \int_{\mathcal{I}_{z,i}(t)^c \cap \mathcal{J}_z(t)}\Big) \frac{\overline{\rho}_t(E + \overline{\mathfrak{e}}_t^+)}{\widehat{z}_i(t,\alpha) - E}\,\mathrm{d}E + \partial_\alpha \Phi_\alpha(t) =: Z_2 + Z_3 + Z_4 + \partial_\alpha \Phi_\alpha(t), \end{aligned}$$

and finally for $1 \leq |i| \leq N^{\omega_A}$ we have $\zeta_i^{(0)} = 0$. We recall that $\mathcal{I}_{z,i}(t)$ and $\mathcal{J}_z(t)$ in (5.131) are defined by (5.119) and (5.118) respectively. Next, we prove that the error term $\zeta^{(0)}$ in (5.127) is bounded by some large power of N.

Lemma 5.6.10. *There exists a large constant $C > 0$ such that*

$$\sup_{0 \leq t \leq t_*} \max_{1 \leq |i| \leq N} \left|\zeta_i^{(0)}(t)\right| \leq N^C. \tag{5.132}$$

Proof. By (5.56), it follows that

$$\partial_\alpha \Phi_\alpha(t) = \partial_\alpha \Re[\overline{m}_t(\overline{\mathfrak{e}}_t^+ + \mathrm{i}N^{-100})] + h^{**}(t,1) - h^{**}(t,0),$$

with $h^{**}(t,\alpha)$ defined by (5.52). Since the two h^{**} terms are small by (5.51), for each fixed t, we have that

$$|\partial_\alpha \Phi_\alpha(t)| \lesssim \left| \partial_\alpha \int_{\mathbb{R}} \frac{\overline{\rho}_t(\overline{\mathfrak{e}}_t^+ + E)}{E - \mathrm{i}N^{-100}} \, \mathrm{d}E \right| + N^{-1} = U_1 + U_2 + N^{-1}, \tag{5.133}$$

where

$$\begin{aligned} U_1 &:= \left| \partial_\alpha \int_{\overline{\gamma}_{-i(\delta_*)}}^{\overline{\gamma}_{i(\delta_*)}} \frac{\overline{\rho}_t(\overline{\mathfrak{e}}_t^+ + E)}{E - \mathrm{i}N^{-100}} \, \mathrm{d}E \right| = \left| \partial_\alpha \int_{I_*} \frac{\overline{\rho}_t(\overline{\mathfrak{e}}_t^+ + \varphi_{\alpha,t}(s))}{\varphi_{\alpha,t}(s) - \mathrm{i}N^{-100}} \varphi_{\alpha,t}'(s) \, \mathrm{d}s \right| \\ U_2 &:= \left| \frac{1}{N} \sum_{i_*(\delta) < |i| \le N} \partial_\alpha \int_{\mathbb{R}} \frac{\psi(E - \overline{\gamma}_i^*(t))}{E - \mathrm{i}N^{-100}} \, \mathrm{d}E \right|, \end{aligned}$$

using the notation $\overline{\gamma}_{i(\delta_*)} = \overline{\gamma}_{i(\delta_*)}(t)$ and the definition of $\overline{\rho}_t$ from (5.46). In U_1 we changed variables, i.e. $E = \varphi_{\alpha,t}(s)$, using that $s \to \varphi_{\alpha,t}(s)$ is strictly increasing. In particular, in order to compute the limits of integration we used that $\varphi_{\alpha,t}(i/N) = \overline{\gamma}_i(t)$ by (5.15) and defined the α-independent interval $I_* := [-i(\delta_*)/N, i(\delta_*)/N]$. Furthermore, in U_1 we denoted by prime the s-derivative.

For U_1 we have that (omitting the t dependence, $\overline{\rho} = \overline{\rho}_t$, etc.)

$$\begin{aligned} U_1 \lesssim &\left| \int_{I_*} \frac{\partial_\alpha[\overline{\rho}(\overline{\mathfrak{e}}^+ + \varphi_\alpha(s))]}{\varphi_\alpha(s) - \mathrm{i}N^{-100}} \varphi_\alpha'(s) \, \mathrm{d}s \right| + \left| \int_{I_*} \frac{\overline{\rho}(\overline{\mathfrak{e}}^+ + \varphi_\alpha(s))}{(\varphi_\alpha(s) - \mathrm{i}N^{-100})^2} (\varphi_\alpha'(s))^2 \, \mathrm{d}s \right| \\ &+ \left| \int_{I_*} \frac{\overline{\rho}(\overline{\mathfrak{e}}^+ + \varphi_\alpha(s))}{\varphi_\alpha(s) - \mathrm{i}N^{-100}} \partial_\alpha \varphi_\alpha'(s) \, \mathrm{d}s \right|. \end{aligned} \tag{5.134}$$

For $s \in I_*$, by the definition of $\varphi_\alpha(s)$ and (5.16) it follows that

$$1 = n_\alpha'(\varphi_\alpha(s))\varphi_\alpha'(s) = \rho_\alpha(\varphi_\alpha(s))\varphi_\alpha(s),$$

and so that

$$\varphi_\alpha'(s) = \frac{1}{\rho_\alpha(\varphi_\alpha(s))} \lesssim s^{-\frac{1}{4}}, \tag{5.135}$$

where in the last inequality we used that $\rho_\alpha(\omega) \sim \min\{\omega^{1/3}, \omega^{1/2}\Delta^{-1/6}\}$ and $\varphi_\alpha(s) \sim \max\{s^{\frac{3}{4}}, s^{\frac{2}{3}}\Delta^{1/9}\}$ by (5.20a).

In the first integral in (5.134) we use that

$$\overline{\rho}(\overline{\mathfrak{e}}^+ + \varphi_\alpha(s)) = \rho_\alpha(\overline{\mathfrak{e}}^+ + \varphi_\alpha(s)), \qquad s \in I_*$$

by (5.46) and that $\partial_\alpha[\rho_\alpha(\overline{\mathfrak{e}}^+ + \varphi_\alpha(s))]$ is bounded by the explicit relation in (5.21). For the other two integrals in (5.134) we use that $\overline{\rho}$ is bounded on the integration domain and that $(\varphi_\alpha'(s))^2 \lesssim s^{-1/2}$ from (5.135), hence it is integrable. In the third integral we also observe that

$$\partial_\alpha \varphi_\alpha(s) = \varphi_\lambda(s) - \varphi_\mu(s)$$

by (5.15), thus $|\partial_\alpha \varphi'_\alpha(s)| \lesssim s^{-1/4}$ similarly to (5.135). Using $|\varphi_\alpha(s) - \mathrm{i}N^{-100}| \gtrsim N^{-100}$, we thus conclude that

$$U_1 \lesssim N^{200}.$$

Next, we proceed with the estimate for U_2.

Notice that $|\partial_\alpha \psi(E - \overline{\gamma}_i^*(t))| \le \|\psi'\|_\infty |\widehat{\gamma}_{x,i}(t) - \widehat{\gamma}_{y,i}(t)|$ by (5.44). Furthermore, since $\left|E - \mathrm{i}N^{-100}\right| \gtrsim \delta_*$ on the domain of integration of U_2, we conclude that

$$U_2 \lesssim N^{200} \left\|\psi'\right\|_\infty,$$

and therefore from (5.133) we have

$$|\partial_\alpha \Phi_\alpha(t)| \lesssim N^{202}. \tag{5.136}$$

since $\left\|\psi'\right\|_\infty \lesssim N^2$ by the choice of ψ, see below (5.43).

Similarly, we conclude that

$$|Z_3| \lesssim N^{200} \|\psi'\|_\infty.$$

To estimate Z_2, by (5.55), it follows that

$$\mathrm{d}(\partial_\alpha \widetilde{z}_i) = \left[\frac{1}{N}\sum_j \frac{\partial_\alpha \widetilde{z}_j - \partial_\alpha \widetilde{z}_i}{(\widetilde{z}_i - \widetilde{z}_j)^2} + \partial_\alpha \Phi_\alpha(t)\right] \mathrm{d}t,$$

with initial data

$$\partial_\alpha \widetilde{z}_i(0,\alpha) = \widetilde{x}_i(0) - \widetilde{y}_i(0),$$

for all $1 \le |i| \le N$. Since $|\partial_\alpha \widetilde{z}_i(0,\alpha)| \lesssim N^{200}$ for all $1 \le |i| \le N$, by Duhamel principle and contraction, it follows that

$$|\partial_\alpha \widetilde{z}_i(t,\alpha)| \lesssim N^{200} + t_* \max_{0 \le \tau \le t_*} |\partial_\alpha \Phi_\alpha(\tau)| \lesssim N^{202} \tag{5.137}$$

for all $0 \le t \le t_*$. In particular, by (5.137) it follows that

$$|Z_2| \lesssim N^{202}\sqrt{N}$$

since for all j in the summation in Z_2 we have that $|i - j| \gtrsim i_* \sim N^{\frac{1}{2}}$ and thus $|\widetilde{z}_i - \widetilde{z}_j| \gtrsim \left|i - j\right|/N \gtrsim N^{-1/2}$.

Finally, we estimate Z_4 using the fact that the endpoints of $\mathcal{I}_{z,i}(t)^c \cap \mathcal{J}_z(t)$ are quantiles $\overline{\gamma}_i(t)$ whose α-derivatives are bounded by (5.44). Hence

$$|Z_4| \lesssim \left|\frac{\overline{\rho}_t(\overline{\gamma}_{j_+} + \overline{\mathfrak{e}}_t^+)}{\widehat{z}_i - \overline{\gamma}_{j_+}}\right| + \left|\frac{\overline{\rho}_t(\overline{\gamma}_{j_-} + \overline{\mathfrak{e}}_t^+)}{\widehat{z}_i - \overline{\gamma}_{j_-}}\right| + \left|\frac{\overline{\rho}_t(\overline{\gamma}_{\frac{3i_*}{4}} + \overline{\mathfrak{e}}_t^+)}{\widehat{z}_i - \overline{\gamma}_{\frac{3i_*}{4}}}\right| \lesssim N \tag{5.138}$$

by rigidity. Combining (5.136)-(5.138) we conclude (5.132), completing also the proof of Lemma 5.6.10. □

Continuing the analysis of the equation (5.127), for any fixed α let us define $w^{\#} = w^{\#}(t,\alpha)$ as the solution of

$$\partial_t w^{\#} = \mathcal{L}w^{\#}, \tag{5.139}$$

with cutoff initial data

$$w_i^{\#}(0,\alpha) = \mathbb{1}_{\{|i|\leq N^{4\omega_\ell+\delta}\}}w_i(0,\alpha),$$

with some $0<\delta<C\omega_\ell$ where $C>10$ a constant such that $(4+C)\omega_\ell<\omega_A$.

By the rigidity (5.125) the finite speed estimate (5.207), with $\delta' := \delta$, for the propagator $\mathcal{U}^{\mathcal{L}}$ of $\mathcal{L}$ holds. Let $0<\delta_1<\frac{\delta}{2}$, then, using Duhamel principle and (5.132), it easily follows that

$$\sup_{0\leq t\leq t_*}\max_{|i|\leq N^{4\omega_\ell+\delta_1}}\left|w_i^{\#}(t,\alpha)-w_i(t,\alpha)\right|\leq N^{-100}, \tag{5.140}$$

for any $\alpha\in[0,1]$. In other words, the initial conditions far away do not influence the w-dynamics, hence they can be set zero.

Next, we use the heat kernel contraction for the equation in (5.139). By the optimal rigidity of $\widehat{x}_i(0)$ and $\widehat{y}_i(0)$, since $w_i^{\#}(0,\alpha)$ is non zero only for $1\leq|i|\leq N^{4\omega_\ell+\delta}$, it follows that

$$\max_{1\leq|i|\leq N}\left|w_i^{\#}(0,\alpha)\right|\leq\frac{N^{\xi}N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}},$$

and so, by heat kernel contraction and Duhamel principle

$$\sup_{0\leq t\leq t_*}\max_{1\leq|i|\leq N}\left|w_i^{\#}(t,\alpha)\right|\leq\frac{N^{\xi}N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}}. \tag{5.141}$$

Next, we recall that $\widehat{z}(t,\alpha=0)=\widehat{y}(t)$.

Combining (5.140) and (5.141), integrating $w_i(t,\alpha')$ over $\alpha'\in[0,\alpha]$, by high moment Markov inequality as in (5.81)-(5.82), we conclude that

$$\sup_{0\leq t\leq t_*}|\widehat{z}_i(t,\alpha)-\widehat{y}_i(t)|\leq\frac{N^{\xi}N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}},\qquad 1\leq|i|\leq N^{4\omega_\ell+\delta_1},$$

for any fixed $\alpha\in[0,1]$ with very high probability for any $\xi>0$. Since

$$|\widehat{z}_i(t,\alpha)-\overline{\gamma}_i(t)|\leq|\widehat{y}_i(t)-\widehat{\gamma}_{y,i}(t)|+|\overline{\gamma}_i(t)-\widehat{\gamma}_{y,i}(t)|+\frac{N^{\xi}N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}},$$

for all $1\leq|i|\leq N^{4\omega_\ell+\delta_1}$ and $\alpha\in[0,1]$, by (5.28) and the optimal rigidity of $\widehat{y}_i(t)$, see (5.49), we conclude that

$$\sup_{0\leq t\leq t_*}|\widehat{z}_i(t,\alpha)-\overline{\gamma}_i(t)|\leq\frac{N^{\xi}N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}},\qquad 1\leq|i|\leq N^{4\omega_\ell+\delta_1}$$

for any fixed $\alpha\in[0,1]$, for any $\xi>0$ with very high probability. This concludes the proof of (5.126). □

5.6.4 Phase 3: Rigidity for $\widehat{z}$ with the correct i-dependence.

In this subsection we will prove almost optimal i-dependent rigidity for the short range approximation $\widehat{z}_i(t,\alpha)$ (see (5.120)–(5.123)) for $1 \le |i| \le N^{4\omega_\ell+\delta_1}$.

Proposition 5.6.11. *Let δ_1 be defined in Proposition 5.6.9, then, for any fixed $\alpha \in [0,1]$, we have that*

$$\sup_{0\le t\le t_*} |\widehat{z}_i(t,\alpha) - \overline{\gamma}_i(t)| \lesssim \frac{N^\xi N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}|i|^{\frac{1}{4}}}, \qquad 1 \le |i| \le N^{4\omega_\ell+\delta_1}, \tag{5.142}$$

for any $\xi > 0$ with very high probability.

Proof. Define

$$K := \lceil N^\xi \rceil,$$

then (5.126) (with $\xi \to \xi/2$) implies (5.142) for all $1 \le |i| \le 2K$. Next, we prove (5.142) for all $2K \le |i| \le N^{4\omega_\ell+\delta_1}$ by coupling $\widetilde{x}_i(t)$ with $\widetilde{y}_{\langle i-K\rangle}(t)$, where we make the following notational convention:

$$\langle i-K\rangle := \begin{cases} i-K & \text{if } i \in [K+1,N]\cup[-N,-1], \\ i-K-1 & \text{if } i\in[1,K]. \end{cases} \tag{5.143}$$

This slight complication is due to our indexing convention that excludes $i=0$.

In order to couple the Brownian motion of $\widetilde{x}_i(t)$ with the one of $\widetilde{y}_{\langle i-K\rangle}(t)$ we construct a new process $\widetilde{z}^*(t,\alpha)$ satisfying

$$\mathrm{d}\widetilde{z}_i^*(t,\alpha) = \sqrt{\frac{2}{N}}\,\mathrm{d}B_{\langle i-K\rangle} + \left[\frac{1}{N}\sum_{j\neq i}\frac{1}{\widetilde{z}_i^*(t,\alpha)-\widetilde{z}_j^*(t,\alpha)} + \Phi_\alpha(t)\right]\mathrm{d}t, \quad 1\le |i|\le N \tag{5.144}$$

with initial data

$$\widetilde{z}_i^*(0,\alpha) = \alpha\widetilde{x}_i(0) + (1-\alpha)\widetilde{y}_{\langle i-K\rangle}(0),$$

for any $\alpha\in[0,1]$. Notice that the only difference with respect to $\widetilde{z}_i(t,\alpha)$ from (5.55) is a shift in the index of the Brownian motion, i.e. $\widetilde{z}$ and $\widetilde{z}^*$ (almost) coincide in distribution, but their coupling to the y-process is different. The slight discrepancy comes from the effect of the few extreme indices. Indeed, to make the definition (5.144) unambiguous even for extreme indices, $i\in[-N,-N+K-1]$, additionally we need to define independent Brownian motions B_j and initial padding particles $\widetilde{y}_j(0) = -jN^{300}$ for $j=-N-1,\dots-N-K$. Similarly to Lemma 5.5.1, the effect of these very distant additional particles is negligible on the dynamics of the particles for $1\le|i|\le\epsilon N$ for some small ϵ.

Next, we define the process $\widehat{z}^*(t,\alpha)$ as the short range approximation of $\widetilde{z}^*(t,\alpha)$, given by (5.120)–(5.122) but B_i replaced with $B_{\langle i-K\rangle}$ and we use initial data $\widehat{z}^*(0,\alpha) = \widetilde{z}^*(0,\alpha)$. In particular,

$$\widehat{z}_i^*(t,1) = \widehat{x}_i(t) + \mathcal{O}\left(N^{-100}\right), \quad \widehat{z}_i^*(t,0) = \widehat{y}_{\langle i-K\rangle}(t) + \mathcal{O}\left(N^{-100}\right), \quad 1\le|i|\le\epsilon N,$$

the discrepancy again coming from the negligible effect of the additional K distant particles on the particles near the cusp regime.

Let $w_i^*(t,\alpha) := \partial_\alpha\widehat{z}_i^*(t,\alpha)$, i.e. $w^* = w^*(t,\alpha)$ is a solution of

$$\partial_t w^* = \mathcal{B}w^* + \mathcal{V}w^* + \zeta^{(0)}$$

with initial data

$$w_i^*(0,\alpha) = \widehat{x}_i^*(0) - \widehat{y}_{\langle i-K\rangle}(0).$$

The operators $\mathcal{B}$, $\mathcal{L}$ and the error term $\zeta^{(0)}$ are defined as in (5.128)-(5.131) with all $\widetilde{z}$ and $\widehat{z}$ replaced by $\widetilde{z}^*$ and $\widehat{z}^*$, respectively.

We now define $(w^*)^\#$ as the solution of

$$\partial_t (w^*)^\# = \mathcal{L}(w^*)^\#,$$

with cutoff initial data

$$(w_i^*)^\#(0,\alpha) = \mathbb{1}_{\{|i|\le N^{4\omega_\ell+\delta}\}} w_i^*(0,\alpha),$$

with $0 < \delta < C\omega_\ell$ with $C > 10$ such that $(4+C)\omega_\ell < \omega_A$.

We claim that

$$(w_i^*)^\#(0,\alpha) \ge 0, \qquad 1 \le |i| \le N. \tag{5.145}$$

We need to check it for $1 \le |i| \le N^{4\omega_\ell+\delta}$, otherwise $(w_i^*)^\#(0,\alpha) = 0$ by the cutoff. In the regime $1 \le |i| \le N^{4\omega_\ell+\delta}$ we use the optimal rigidity (Lemma 5.6.1 with $\xi \to \xi/10$) for $\widehat{x}_i^*(0)$ and $\widehat{y}_{\langle i-K\rangle}(0)$ that yields

$$\begin{aligned}(w_i^*)^\#(0,\alpha) = \widehat{x}_i^*(0) - \widehat{y}_{\langle i-K\rangle}(0) \ge -N^{\frac{\xi}{10}}\eta_f(\gamma_{x,i}^*(0)) + \widehat{\gamma}_{x,i}(0)\\ - \widehat{\gamma}_{y,\langle i-K\rangle}(0) - N^{\frac{\xi}{10}}\eta_f(\gamma_{y,\langle i-K\rangle}^*(0)).\end{aligned} \tag{5.146}$$

We now check that $\widehat{\gamma}_{x,i}(0) - \widehat{\gamma}_{y,\langle i-K\rangle}(0)$ is sufficiently positive to compensate for the $N^{\frac{\xi}{10}}\eta_f$ error terms. Indeed, by (5.26a) and (5.28), for all $|i| \ge 2K$ we have

$$\widehat{\gamma}_{x,i}(t) - \widehat{\gamma}_{y,\langle i-K\rangle}(t) \gtrsim K\eta_f(\gamma_{x,i}^*(t)) \gg N^{\frac{\xi}{10}}\eta_f(\gamma_{x,i}^*(t))$$

and that

$$\eta_f(\gamma_{y,\langle i-K\rangle}^*(t)) \sim \eta_f(\gamma_{x,i}^*(t)).$$

This shows (5.145) in the $2K \le |i| \le N^{4\omega_\ell+\delta}$ regime. If $K \le |i| \le 2K$ or $-K \le i \le -1$ we have that $(w_i^*)^\#(0,\alpha) \ge 0$ since

$$\begin{aligned}\widehat{\gamma}_{x,i}(0) - \widehat{\gamma}_{y,\langle i-K\rangle}(0) &\gtrsim \max\left\{\frac{K^{3/4}}{N^{3/4}}, (t_* - t)^{1/6}\frac{K^{2/3}}{N^{2/3}}\right\}\\ &\gtrsim K \max\left\{\eta_f(\gamma_{x,i}^*(0)), \eta_f(\gamma_{y,\langle i-K\rangle}^*(0))\right\},\end{aligned}$$

so $\widehat{\gamma}_{x,i}(0) - \widehat{\gamma}_{y,\langle i-K\rangle}(0)$ beats the error terms $N^{\frac{\xi}{10}}\eta_f$ as well. Finally, if $1 \le i \le K-1$ the bound in (5.146) is easy since $\widehat{\gamma}_{x,i}(0)$ and $\widehat{\gamma}_{y,\langle i-K\rangle}(0)$ have opposite sign, i.e. they are in two different sides of the small gap and one of them is at least of order $(K/N)^{3/4}$, beating $N^{\frac{\xi}{10}}\eta_f$. This proves (5.145). Hence, by the maximum principle we conclude that

$$(w_i^*)^\#(t,\alpha) \ge 0, \qquad 0 \le t \le t_*, \qquad \alpha \in [0,1].$$

Let $\delta_1 < \frac{\delta}{2}$ be defined in Proposition 5.6.9. The rigidity estimate in (5.125) holds for $\widehat{z}^*$ as well, since $\widehat{z}$ and $\widehat{z}^*$ have the same distribution. Furthermore, by (5.125) the propagator

$\mathcal{U}$ of $\mathcal{L} := \mathcal{B} + \mathcal{V}$ satisfies the finite speed estimate in Lemma 5.B.3. Then, using Duhamel principle and (5.132), we obtain

$$\sup_{0 \le t \le t_*} \max_{1 \le |i| \le N^{4\omega_\ell + \delta_1}} \left| (w_i^*)^{\#}(t,\alpha) - w_i^*(t,\alpha) \right| \le N^{-100}, \tag{5.147}$$

for any $\alpha \in [0,1]$ with very high probability.

By (5.147), integrating $w_i^*(t,\alpha')$ over $\alpha' \in [0,\alpha]$, we conclude that

$$\widehat{z}_i^*(t,\alpha) - \widehat{y}_{\langle i-K \rangle}(t) \ge -N^{-100}, \qquad 1 \le |i| \le N,^{4\omega_\ell + \delta_1} \tag{5.148}$$

for all $\alpha \in [0,1]$ and $0 \le t \le t_*$ with very high probability. Note that in order to prove (5.148) with very high probability we used a Markov inequality as in (5.81)-(5.82). Hence,

$$\begin{aligned} \widehat{z}_i^*(t,\alpha) - \overline{\gamma}_i(t) &\ge [\widehat{y}_{\langle i-K \rangle}(t) - \widehat{\gamma}_{y,\langle i-K \rangle}(t)] + [\widehat{\gamma}_{y,\langle i-K \rangle}(t) - \widehat{\gamma}_{y,i}(t)] \\ &\quad + [\widehat{\gamma}_{y,i}(t) - \overline{\gamma}_i(t)] - N^{-100} \\ &\gtrsim -K(\eta_f(\gamma^*_{y,\langle i-K \rangle}(t)) + \eta_f(\gamma^*_{y,i}(t))) - \gamma_i^*(t) t_*^{1/3} \\ &\ge -2K(\eta_f(\gamma^*_{y,\langle i-K \rangle}(t)) + \eta_f(\gamma^*_{y,i}(t))) \end{aligned}$$

for all $1 \le |i| \le N^{4\omega_\ell + \delta_1}$, where we used the optimal rigidity (5.49) and (5.28) in going to the second line. In particular, since for $|i| \ge 2K$ we have that $\eta_f(\gamma^*_{y,i}(t)) \sim \eta_f(\gamma^*_{y,i-K}(t))$, we conclude that

$$\widehat{z}_i^*(t,\alpha) - \overline{\gamma}_i(t) \ge -\frac{CKN^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}} |i|^{\frac{1}{4}}}, \qquad 2K \le |i| \le N^{4\omega_\ell + \delta_1}, \tag{5.149}$$

for all $0 \le t \le t_*$ and for any $\alpha \in [0,1]$. This implies the lower bound in (5.142).

In order to prove the upper bound in (5.142) we consider a very similar process $\widetilde{z}_i^*(t,\alpha)$ (we continue to denote it by star) where the index shift in y is in the other direction. More precisely, it is defined as a solution of

$$\mathrm{d}\widetilde{z}_i^*(t,\alpha) = \sqrt{\frac{2}{N}}\, \mathrm{d}B_{\langle i+K \rangle} + \left[\frac{1}{N} \sum_{j \ne i} \frac{1}{\widetilde{z}_i^*(t,\alpha) - \widetilde{z}_j^*(t,\alpha)} + \Phi_\alpha(t) \right] \mathrm{d}t$$

with initial data

$$\widetilde{z}_i(0,\alpha) = \alpha \widetilde{y}_{\langle i+K \rangle}(0) + (1-\alpha)\widetilde{x}_i(0),$$

for any $\alpha \in [0,1]$. Here $\langle i+K \rangle$ is defined analogously to (5.143). Then, by similar computations, we conclude that

$$\widehat{z}_i^*(t,\alpha) - \overline{\gamma}_i(t) \le \frac{KN^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}} |i|^{\frac{1}{4}}}, \qquad 2K \le |i| \le N^{4\omega_\ell + \delta_1}, \tag{5.150}$$

for all $0 \le t \le t_*$ and for any $\alpha \in [0,1]$. Combining (5.149) and (5.150) we conclude (5.142) and complete the proof of Proposition 5.6.11. □

5.7 Proof of Proposition 5.3.1: Dyson Brownian motion near the cusp

In this section $t_1 \le t_*$, indicating that we are before the cusp formation. The main result of this section is the following proposition from which we can quickly prove Proposition 5.3.1 for $t_1 \le t_*$. If $t_1 > t_*$ we conclude Proposition 5.3.1 using the analogous Proposition 5.8.1 instead of Proposition 5.7.1 exactly in the same way.

Proposition 5.7.1. *For $t_1 \le t_*$, with very high probability we have that*

$$\left|(\lambda_j(t_1) - \mathfrak{e}^+_{\lambda,t_1}) - (\mu_{j+i_\mu - i_\lambda}(t_1) - \mathfrak{e}^+_{\mu,t_1})\right| \le N^{-\frac{3}{4}-c\omega_1} \tag{5.151}$$

for some small constant $c > 0$ and for any j such that $|j - i_\lambda| \le N^{\omega_1}$.

The proof of Proposition 5.7.1 will be given at the end of the section after several auxiliary lemmas.

Proof of Proposition 5.3.1. Using the change of variables $\boldsymbol{x} = N^{\frac{3}{4}}(\boldsymbol{x}' - \mathfrak{b}_{r,t_1})$, for $r = \lambda, \mu$, and the definition of correlation function, for each Lipschitz continuous and compactly supported test function F, we have that

$$\begin{aligned}
&\int_{\mathbb{R}^k} F(\boldsymbol{x}) \left[N^{k/4} p^{(N,\lambda)}_{k,t_1}\left(\mathfrak{b}_{\lambda,t_1} + \frac{\boldsymbol{x}}{N^{3/4}}\right) - N^{k/4} p^{(N,\mu)}_{k,t_1}\left(\mathfrak{b}_{\mu,t_1} + \frac{\boldsymbol{x}}{N^{3/4}}\right)\right] \mathrm{d}\boldsymbol{x} \\
&= N^k \binom{N}{k}^{-1} \sum_{\{i_1,\dots,i_k\}\subset[N]} \Bigg[\mathbf{E}_{H^{(\lambda)}_{t_1}} F\left(N^{\frac{3}{4}}(\lambda_{i_1} - \mathfrak{b}_{\lambda,t_1}),\dots,N^{\frac{3}{4}}(\lambda_{i_k} - \mathfrak{b}_{\lambda,t_1})\right) \\
&\qquad\qquad - \mathbf{E}_{H^{(\mu)}_{t_1}} F(\lambda \to \mu)\Bigg],
\end{aligned} \tag{5.152}$$

where $\lambda_1,\dots,\lambda_N$ and $\mu_1,\dots,\mu_N$ are the eigenvalues, labelled in increasing order, of $H^{(\lambda)}_{t_1}$ and $H^{(\mu)}_{t_1}$ respectively. In $\mathbf{E}_{H^{(\mu)}_{t_1}} F(\lambda \to \mu)$ we also replace $\mathfrak{b}_{\lambda,t_1}$ by $\mathfrak{b}_{\mu,t_1}$.

In order to apply Proposition 5.7.1 we split the sum in the rhs. of (5.152) into two sums:

$$\sum_{\substack{\{i_1,\dots,i_k\}\subset[N] \\ |i_1 - i_\lambda|,\dots,|i_k - i_\lambda| < N^\epsilon}} \qquad \text{and its complement} \qquad \sum{}', \tag{5.153}$$

where ϵ is a positive exponent with $\epsilon \ll \omega_1$.

We start with the estimate for the second sum of (5.153). In particular, we will estimate only the term $\mathbf{E}_{H^{(\lambda)}_{t_1}}(\cdot)$, the estimate for $\mathbf{E}_{H^{(\mu)}_{t_1}}(\cdot)$ will follow in an analogous way.

Since the test function F is compactly supported and in $\sum'$ there is at least one index i_l such that $|i_l - i_\lambda| \ge N^\epsilon$, we have that

$$\begin{aligned}
&\sum{}' \mathbf{E}_{H^{(\lambda)}_{t_1}} F\left(N^{\frac{3}{4}}(\lambda_{i_1} - \mathfrak{b}_{\lambda,t_1}),\dots,N^{\frac{3}{4}}(\lambda_{i_k} - \mathfrak{b}_{\lambda,t_1})\right) \\
&\quad \lesssim N^{k-1} \sum_{i_l \colon |i_l - i_\lambda| \ge N^\epsilon} \mathbf{P}_{H^{(\lambda)}_{t_1}}\left(|\lambda_{i_l} - \mathfrak{b}_{\lambda,t_1}| \lesssim N^{-\frac{3}{4}}\right).
\end{aligned}$$

Let $\gamma_{\lambda,i} = \widehat{\gamma}_{\lambda,i} + \mathfrak{e}^+_{\lambda,t_1}$ be the classical eigenvalue locations of $\rho_\lambda(t_1)$ defined by (5.24) for all $1 - i_\lambda \leq i \leq N + 1 - i_\lambda$. Then, by the rigidity estimate from Corollary 4.2.6, we have that

$$\mathbf{P}_{H_{t_1}^{(\lambda)}}\Big(|\lambda_{i_l} - \mathfrak{b}_{\lambda,t_1}| \lesssim N^{-\frac{3}{4}}, |i_l - i_\lambda| \geq N^\epsilon\Big) \leq N^{-D}, \tag{5.154}$$

for each $D > 0$. Indeed, by rigidity it follows that

$$\begin{aligned} |\lambda_{i_l} - \mathfrak{b}_{\lambda,t_1}| &\geq |\gamma_{\lambda,i_l} - \gamma_{\lambda,i_\lambda}| - |\lambda_{i_l} - \gamma_{\lambda,i_l}| - |\mathfrak{b}_{\lambda,t_1} - \gamma_{\lambda,i_\lambda}| \\ &\gtrsim \frac{N^{c\epsilon}}{N^{\frac{3}{4}}} - \frac{N^{c\xi}}{N^{\frac{3}{4}}} \gtrsim \frac{N^{c\epsilon}}{N^{\frac{3}{4}}} \end{aligned} \tag{5.155}$$

with very high probability, if $N^\epsilon \leq |i_l - i_\lambda| \leq \tilde{c}N$, for some $0 < \tilde{c} < 1$. In (5.155) we used the rigidity from Corollary 4.2.6 in the form

$$|\lambda_i - \gamma_{\lambda,i}| \leq \frac{N^\xi}{N^{\frac{3}{4}}},$$

for any $\xi > 0$, with very high probability. Note that (5.154) and (5.155) hold for any $\epsilon \gtrsim \xi$. If $|i_l - i_\lambda| \geq \tilde{c}N$, then $|\gamma_{i_l} - \gamma_{i_\lambda}| \sim 1$ and the bound in (5.155) clearly holds. A similar estimate holds for $H_{t_1}^{(\mu)}$, hence, choosing $D > k + 1$ we conclude that the second sum in (5.153) is negligible.

Next, we consider the first sum in (5.153). For $t_1 \leq t_*$ we have, by (5.17a) that

$$\Big|(\mathfrak{e}^+_{\lambda,t_1} - \mathfrak{b}_{\lambda,t_1}) - (\mathfrak{e}^+_{\mu,t_1} - \mathfrak{b}_{\mu,t_1})\Big| = \frac{1}{2}|\Delta_{\lambda,t_1} - \Delta_{\mu,t_1}| \lesssim \Delta_{\mu,t_1}(t_* {-} t_1)^{1/3} \leq N^{-\frac{3}{4}-\frac{1}{6}+C\omega_1}.$$

Hence, by (5.151), using that $|F(\boldsymbol{x}) - F(\boldsymbol{x}')| \lesssim \|\boldsymbol{x} - \boldsymbol{x}'\|$, we conclude that

$$\begin{aligned} \sum_{\substack{\{i_1,\dots,i_k\}\subset[N] \\ |i_1-i_\lambda|,\dots,|i_k-i_\lambda|\leq N^\epsilon}} \Big[\mathbf{E}_{H_{t_1}^{(\lambda)}} F\Big(N^{\frac{3}{4}}(\lambda_{i_1} - \mathfrak{b}_{\lambda,t_1}), \dots, N^{\frac{3}{4}}(\lambda_{i_k} - \mathfrak{b}_{\lambda,t_1})\Big) \\ - \mathbf{E}_{H_{t_1}^{(\mu)}} F(\lambda \to \mu)\Big] \leq C_k \frac{N^{k\epsilon}}{N^{c\omega_1}}, \end{aligned}$$

for some $c > 0$. Then, using that

$$\frac{N^k(N-k)!}{N!} = 1 + \mathcal{O}_k(N^{-1}),$$

we easily conclude the proof of Proposition 5.3.1. □

5.7.1 Interpolation.

In order to prove Proposition 5.7.1 we recall a few concepts introduced previously. In Section 5.5 we introduced the padding particles $x_i(t)$, $y_i(t)$, for $1 \leq |i| \leq N$, that are good approximations of the eigenvalues $\lambda_j(t)$, $\mu_j(t)$ respectively, for $1 \leq j \leq N$, in the sense of Lemma 5.5.1. They satisfy a Dyson Brownian Motion equation (5.41), (5.42) mimicking the DBM of genuine eigenvalue processes (5.39), (5.40). It is more convenient to consider shifted processes where the edge motion is subtracted.

More precisely, for $r = x, y$ and $r(t) = x(t), y(t)$, we defined

$$\widetilde{r}_i(t) := r_i(t) - \mathfrak{e}^+_{r,t}, \qquad 1 \leq |i| \leq N,$$

for all $0 \leq t \leq t_*$. In particular, $\widetilde{r}(t)$ is a solution of

$$\mathrm{d}\widetilde{r}_i(t) = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \left(\frac{1}{N}\sum_{j\neq i}\frac{1}{\widetilde{r}_i(t) - \widetilde{r}_j(t)} + \Re[m_{r,t}(\mathfrak{e}^+_{r,t})]\right)\mathrm{d}t,$$

with initial data

$$\widetilde{r}_i(0) = r_i(0) - \mathfrak{e}^+_{r,0},$$

for all $1 \leq |i| \leq N$.

Next, following a similar idea of [122], we also introduced in (5.55) an interpolation process between $\widetilde{x}(t)$ and $\widetilde{y}(t)$. For any $\alpha \in [0,1]$ we defined the process $\widetilde{z}(t,\alpha)$ as the solution of

$$\mathrm{d}\widetilde{z}_i(t,\alpha) = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \left(\frac{1}{N}\sum_{j\neq i}\frac{1}{\widetilde{z}_i(t,\alpha) - \widetilde{z}_j(t,\alpha)} + \Phi_\alpha(t)\right)\mathrm{d}t,$$

with initial data

$$\widetilde{z}_i(0,\alpha) = \alpha\widetilde{x}_i(0) + (1-\alpha)\widetilde{y}_i(0),$$

for each $1 \leq |i| \leq N$. Recall that $\Phi_\alpha(t)$ was defined in (5.56) and it is such that $\Phi_0(t) = \Re[m_{y,t}(\mathfrak{e}^+_{y,t})]$ and $\Phi_1(t) = \Re[m_{x,t}(\mathfrak{e}^+_{x,t})]$. Note that $\widetilde{z}_i(t,1) = \widetilde{x}_i(t)$ and $\widetilde{z}_i(t,0) = \widetilde{y}_i(t)$ for all $1 \leq |i| \leq N$ and $0 \leq t \leq t_*$.

We recall the definition of the interpolated quantiles from (5.44) of Section 5.5;

$$\overline{\gamma}_i(t) := \alpha\widehat{\gamma}_{x,i}(t) + (1-\alpha)\widehat{\gamma}_{y,i}(t), \qquad \alpha \in [0,1],$$

where $\widehat{\gamma}_{x,i}$ and $\widehat{\gamma}_{y,i}$ are the shifted quantiles of $\rho_{x,t}$ and $\rho_{y,t}$ respectively, defined in Section 5.5. In particular,

$$\overline{\mathfrak{e}}^\pm_t = \alpha\mathfrak{e}^\pm_{x,t} + (1-\alpha)\mathfrak{e}^\pm_{y,t}, \qquad \alpha \in [0,1].$$

We denoted the interpolated density, whose quantiles are the $\overline{\gamma}_i(t)$, by $\overline{\rho}_t$ (5.46), and its Stieltjes transform by $\overline{m}_t$.

Let $\widehat{z}(t,\alpha)$ be the short range approximation of $\widetilde{z}(t,\alpha)$ defined by (5.120)-(5.122), with exponents $\omega_1 \ll \omega_\ell \ll \omega_A \ll 1$ and with initial data $\widehat{z}(0,\alpha) = \widetilde{z}(0,\alpha)$ and $i_* = N^{\frac{1}{2}+C_*\omega_1}$, for some large constant $C_* > 0$. In particular, $\widehat{x}(t) = \widehat{z}(t,1)$ and $\widehat{y}(t) = \widehat{z}(t,0)$. Assuming optimal rigidity in (5.49) for $\widetilde{r}_i(t) = \widetilde{x}_i(t), \widetilde{y}_i(t)$, the following lemma shows that the process $\widetilde{r}$ and its short range approximation $\widehat{r} = \widehat{x}, \widehat{y}$ stay very close to each other, i.e. $|\widehat{r}_i - \widetilde{r}_i| \leq N^{-\frac{3}{4}-c}$, for some small $c > 0$. This is the analogue of Lemma 3.7 in [122] and its proof, given in Appendix 5.C, follows similar lines. It assumes the optimal rigidity, see (5.156) below, which is ensured by Corollary 4.2.6, see Lemma 5.6.1.

Lemma 5.7.2. *Let $i_* = N^{\frac{1}{2}+C_*\omega_1}$. Assume that $\widetilde{z}(t,0)$ and $\widetilde{z}(t,1)$ satisfy the optimal rigidity*

$$\sup_{0\leq t\leq t_1}|\widetilde{z}_i(t,\alpha) - \widehat{\gamma}_{r,i}(t)| \leq N^\xi \eta_{\mathrm{f}}^{\rho_{r,t}}(e^+_{r,t} + \widehat{\gamma}_{r,\pm i}(t)), \quad 1 \leq |i| \leq i_*, \tag{5.156}$$

with $r = x, y$, $\alpha = 0, 1$, *for any* $\xi > 0$, *with very high probability. Then, for* $\alpha = 0$ *or* $\alpha = 1$ *we have that*

$$\begin{aligned} &\sup_{1 \le |i| \le N} \sup_{0 \le t \le t_1} |\widetilde{z}_i(t, \alpha) - \widehat{z}_i(t, \alpha)| \\ &\quad \lesssim \frac{N^{\frac{\omega_1}{6}} N^{\xi}}{N^{\frac{3}{4}}} \left(\frac{N^{\omega_1}}{N^{3\omega_\ell}} + \frac{N^{\omega_1}}{N^{\frac{1}{8}}} + \frac{N^{C\omega_1} N^{\frac{\omega_A}{2}}}{N^{\frac{1}{6}}} + \frac{N^{\frac{\omega_A}{2}} N^{C\omega_1}}{N^{\frac{1}{4}}} + \frac{N^{C\omega_1}}{N^{\frac{1}{18}}} \right), \end{aligned} \tag{5.157}$$

for any $\xi > 0$, *with very high probability.*

In particular, (5.157) implies that there exists a small fixed universal constant $c > 0$ such that

$$\sup_{1 \le |i| \le N} \sup_{0 \le t \le t_1} |\widetilde{z}_i(t, \alpha) - \widehat{z}_i(t, \alpha)| \lesssim N^{-\frac{3}{4} - c}, \qquad \alpha = 0, 1 \tag{5.158}$$

with very high probability.

Remark 5.7.3. *Note the denominator in the first error term in* (5.157)*: the factor* $N^{3\omega_\ell}$ *is better than* $N^{2\omega_\ell}$ *in Lemma 3.7 in [122], this is because of the natural cusp scaling. The fact that this power is at least* $N^{(1+\epsilon)\omega_\ell}$ *was essential in [122] since this allowed to transfer the optimal rigidity from* $\widetilde{z}$ *to the* $\widehat{z}$ *process for all* $\alpha \in [0, 1]$. *Optimal rigidity for* $\widehat{z}$ *is essential (i) for the heat kernel bound for the propagator of* $\mathcal{L}$, *see* (5.127)*–*(5.128), *and (ii) for a good* ℓ^p*-norm for the initial condition in* (5.168). *With our approach, however, this power in* (5.157) *is not critical since we have already obtained an even better,* i*-dependent rigidity for the* $\widehat{z}$ *process for any* α *by using maximum principle, see Proposition 5.6.11. We still need* (5.157) *for the* x *and* y *processes (i.e. only for* $\alpha = 0, 1$*), but only with a precision below the rigidity scale, therefore the denominator in the first term has only to beat* $N^{\frac{7}{6}\omega_1 + \xi}$.

5.7.2 Differentiation.

Next, we consider the α-derivative of the process $\widehat{z}(t, \alpha)$. Let

$$u_i(t) = u_i(t, \alpha) := \partial_\alpha \widehat{z}_i(t, \alpha), \qquad 1 \le |i| \le N,$$

then u is a solution of the equation

$$\partial_t u = \mathcal{L} u + \zeta^{(0)}, \tag{5.159}$$

where $\zeta^{(0)}$, defined by (5.130)-(5.131), is an error term that is non zero only for $|i| > N^{\omega_A}$ and such that $\left|\zeta_i^{(0)}\right| \lesssim N^C$, for some large constant $C > 0$ with very high probability, by (5.132), and the operator $\mathcal{L} = \mathcal{B} + \mathcal{V}$ acting on $\mathbb{R}^{2N}$ is defined by (5.128)-(5.129).

In the following with $\mathcal{U}^{\mathcal{L}}$ we denote the semigroup associated to (5.159), i.e. by Duhamel principle

$$u(t) = \mathcal{U}^{\mathcal{L}}(0, t) u(0) + \int_0^t \mathcal{U}^{\mathcal{L}}(s, t) \zeta^{(0)}(s) \, \mathrm{d}s$$

and $\mathcal{U}^{\mathcal{L}}(s, s) = \mathrm{Id}$ for all $0 \le s \le t$. Furthermore, for each a, b such that $|a|, |b| \le N$, with $\mathcal{U}^{\mathcal{L}}_{ab}$ we denote the entries of $\mathcal{U}^{\mathcal{L}}$, which can be either seen as the solution of the equation (5.159) with initial condition $u_a(0) = \delta_{ab}$.

By Proposition 5.6.2 and Lemma 5.C.1, for any fixed $\alpha \in [0, 1]$, it follows that

$$\sup_{0\le t\le t_*} |\widehat{z}_i(t,\alpha)-\overline{\gamma}_i(t)| \lesssim \frac{N^{C\omega_1}}{N^{\frac{1}{2}}}, \qquad 1\le |i|\le N, \tag{5.160}$$

and

$$\sup_{0\le t\le t_*} |\widehat{z}_i(t,\alpha)-\overline{\gamma}_i(t)| \lesssim \frac{N^{C\omega_1}}{N^{\frac{3}{4}}}, \qquad 1\le |i|\le i_*, \tag{5.161}$$

with very high probability. Then, using (5.161), as a consequence of Lemma 5.B.3 we have the following:

Lemma 5.7.4. *There exists a constant $C>0$ such that for any $0<\delta<C\omega_\ell$, if $1\le |a|\le \frac{1}{2}N^{4\omega_\ell+\delta}$ and $|b|\ge N^{4\omega_\ell+\delta}$, then*

$$\sup_{0\le s\le t\le t_*} \mathcal{U}^{\mathcal{L}}_{ab}(s,t)+\mathcal{U}^{\mathcal{L}}_{ba}(s,t)\le N^{-D}$$

for any $D>0$ with very high probability.

Furthermore, by Proposition 5.6.11, for any fixed $\alpha\in[0,1]$, we have that

$$\sup_{0\le t\le t_*} |\widehat{z}_i(t,\alpha)-\overline{\gamma}_i(t)| \lesssim \frac{N^{\xi}N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}|i|^{\frac{1}{4}}}, \qquad 1\le |i|\le N^{4\omega_\ell+\delta_1}, \tag{5.162}$$

for some small fixed $\delta_1>0$ and for any $\xi>0$ with very high probability.

Next, we introduce the ℓ^p norms

$$\|u\|_p := \left(\sum_i |u_i|^p\right)^{\frac{1}{p}}, \quad \|u\|_\infty := \max_i |u_i|\,.$$

Following a similar scheme to [41], [79] with some minor modifications we will prove the following Sobolev-type inequalities in Appendix 5.D.

Lemma 5.7.5. *For any small $\eta>0$ there exists $c_\eta>0$ such that*

$$\begin{aligned}\sum_{i\ne j\in\mathbb{Z}_+} \frac{(u_i-u_j)^2}{|i^{\frac{3}{4}}-j^{\frac{3}{4}}|^{2-\eta}} &\ge c_\eta\Big(\sum_{i\ge 1}|u_i|^p\Big)^{\frac{2}{p}},\\ \sum_{i\ne j\in\mathbb{Z}_-} \frac{(u_i-u_j)^2}{|\,|i|^{\frac{3}{4}}-|j|^{\frac{3}{4}}\,|^{2-\eta}} &\ge c_\eta\Big(\sum_{i\le -1}|u_i|^p\Big)^{\frac{2}{p}}\end{aligned} \tag{5.163}$$

hold, with $p=\frac{8}{2+3\eta}$, for any function $\|u\|_p<\infty$.

Using the Sobolev inequality in (5.163) and the finite speed estimate of Lemma 5.7.4, we prove the following lemma on the heat kernel decay in Appendix 5.E.

Lemma 5.7.6. *Assume (5.160), (5.161) and (5.162). Let $0<\delta_4<\frac{\delta_1}{10}$ and $w_0\in\mathbb{R}^{2N}$ such that $|(w_0)_i|\le N^{-100}\|w_0\|_1$, for $|i|\ge \ell^4N^{\delta_4}$. Then, for any small $\eta>0$ there exists a constant $C>0$ independent of η and a constant c_η such that for all $0\le s\le t\le t_*$*

$$\|\mathcal{U}^{\mathcal{L}}(s,t)w_0\|_2 \le \left(\frac{N^{C\eta+\frac{\omega_1}{3}}}{c_\eta N^{\frac{1}{2}}(t-s)}\right)^{1-3\eta}\|w_0\|_1, \tag{5.164}$$

and

$$\|\mathcal{U}^{\mathcal{L}}(0,t)w_0\|_\infty \leq \left(\frac{N^{C\eta+\frac{\omega_1}{3}}}{c_\eta N^{\frac{1}{2}} t} \right)^{\frac{2(1-3\eta)}{p}} \|w_0\|_p, \tag{5.165}$$

for each $p \geq 1$.

Let $0 < \delta_v < \frac{\delta_4}{2}$. Define $v_i = v_i(t,\alpha)$ to be the solution of

$$\partial_t v = \mathcal{L} v, \qquad v_i(0,\alpha) = u_i(0,\alpha)\mathbf{1}_{\{|i| \leq N^{4\omega_\ell+\delta_v}\}}. \tag{5.166}$$

Then, by Lemma 5.7.4 the next lemma follows.

Lemma 5.7.7. *Let* u *be the solution of the equation in* (5.159) *and* v *defined by* (5.166)*, then we have that*

$$\sup_{0 \leq t \leq t_1} \sup_{|i| \leq \ell^4} |u_i(t) - v_i(t)| \leq N^{-100}, \tag{5.167}$$

with very high probability.

Proof. By (5.159) and (5.166) have that

$$\begin{aligned} u_i(t) - v_i(t) = \sum_{j=-N}^{N} \mathcal{U}_{ij}^{\mathcal{L}}(0,t)u_j(0) - \sum_{j=-N^{4\omega_\ell+\delta_v}}^{N^{4\omega_\ell+\delta_v}} \mathcal{U}_{ij}^{\mathcal{L}}(0,t)u_j(0) \\ + \int_0^t \sum_{|j| \geq N^{\omega_A}} \mathcal{U}_{ij}^{\mathcal{L}}(s,t)\zeta_j^{(0)}(s) \, \mathrm{d}s. \end{aligned}$$

Then, using that $\zeta_i^{(0)} = 0$ for $1 \leq |i| \leq N^{\omega_A}$ and (5.132), the bound in (5.167) follows by Lemma 5.7.4. □

Proof of Proposition 5.7.1. We consider only the $j = i_\lambda$ case. By Lemma 5.5.1 and (5.158) we have that

$$\begin{aligned} &\left| (\lambda_{i_\lambda}(t_1) - \mathfrak{e}_{\lambda,t_1}^{+}) - (\mu_{i_\mu}(t_1) - \mathfrak{e}_{\mu,t_1}^{+}) \right| \\ &\leq |\widetilde{x}_1(t_1) - \widehat{x}_1(t_1)| + |\widehat{x}_1(t_1) - \widehat{y}_1(t_1)| + |\widehat{y}_1(t_1) - \widetilde{y}_1(t_1)| \leq |\widehat{x}_1(t_1) - \widehat{y}_1(t_1)| + N^{-\frac{3}{4}-c} \end{aligned}$$

with very high probability.

Since $\widehat{z}_i(t_1,1) = \widehat{x}_i(t_1)$ and $\widehat{z}_i(t_1,0) = \widehat{y}_i(t_1)$ for all $1 \leq |i| \leq N$, by the definition of $u_i(t,\alpha)$, it follows that

$$\widehat{x}_1(t_1) - \widehat{y}_1(t_1) = \int_0^1 u_1(t_1,\alpha) \, \mathrm{d}\alpha.$$

Furthermore, by a high moment Markov inequality as in (5.81)-(5.82) and Lemma 5.7.7, we get

$$\int_0^1 |u_1(t_1,\alpha)| \, \mathrm{d}\alpha \lesssim N^{-100} + \int_0^1 |v_1(t_1,\alpha)| \, \mathrm{d}\alpha.$$

Since $v_i(0) = u_i(0)\mathbf{1}_{\{|i|\le N^{4\omega_\ell+\delta_v}\}}$ and, by (5.28) and (5.49), for $1 \le |i| \le N^{4\omega_\ell+\delta_v}$ we have that

$$\begin{aligned}|u_i(0)| &\lesssim |\widehat{x}_i(0) - \widehat{\gamma}_{x,i}(0)| + |\widehat{y}_i(0) - \widehat{\gamma}_{y,i}(0)| + |\widehat{\gamma}_{x,i}(0) - \widehat{\gamma}_{y,i}(0)| \\ &\lesssim \frac{N^{\frac{\omega_1}{6}}}{|i|^{\frac{1}{4}} N^{\frac{3}{4}}} + \frac{|i|^{\frac{3}{4}} N^{\frac{\omega_1}{2}}}{N^{\frac{11}{12}}} \lesssim \frac{N^{\frac{\omega_1}{6}}}{|i|^{\frac{1}{4}} N^{\frac{3}{4}}},\end{aligned}$$

we conclude that

$$\|v(0)\|_5 \lesssim \frac{N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}} \tag{5.168}$$

with very high probability. Hence, by Lemma 5.7.6 and Markov's inequality again, we get

$$\int_0^1 |v_1(t_1,\alpha)|\,\mathrm{d}\alpha \le \sup_{\alpha\in[0,1]} \|v(t_1,\alpha)\|_\infty \lesssim \frac{N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}} N^{\frac{4\omega_1}{15}}} = \frac{1}{N^{\frac{3}{4}} N^{\frac{\omega_1}{10}}},$$

with very high probability. This completes the proof of Proposition 5.7.1. □

5.8 Case of $t \ge t_*$: small minimum

In this section we consider the case when the densities $\rho_{x,t}$, $\rho_{y,t}$, hence their interpolation $\overline{\rho}_t$ as well, have a small minimum, i.e. $t_* \le t \le 2t_*$. We deal with the small minimum case in this separate section mainly for notational reasons: for $t_* \le t \le 2t_*$ the processes $x(t)$ and $y(t)$, and consequently the associated quantiles and densities, are shifted by $\widetilde{\mathfrak{m}}_{r,t}$, for $r = x, y$, instead of $\mathfrak{e}_{r,t}^+$. We recall that $\widetilde{\mathfrak{m}}_{r,t}$, defined in (5.22a), denotes a close approximation of the actual local minimum $\mathfrak{m}_{r,t}$ near the physical cusp. We chose to shift $x(t)$ and $y(t)$ by the tilde approximation of the minimum instead of the minimum itself for technical reasons, namely because the t-derivative of $\widetilde{\mathfrak{m}}_{r,t}$, $r = x, y$, satisfies the convenient relation in (5.22d).

As we explained at the beginning of Section 5.7, in order to prove universality, i.e. Proposition 5.3.1 at time $t_1 \ge t_*$, it is enough to prove the following:

Proposition 5.8.1. *For $t_1 \ge t_*$, we have, with very high probability, that*

$$\big|(\lambda_j(t_1) - \mathfrak{m}_{\lambda,t_1}) - (\mu_{j+i_\mu-i_\lambda}(t_1) - \mathfrak{m}_{\mu,t_1})\big| \le N^{-\frac{3}{4}-c}$$

for some small constant $c > 0$ and for any j such that $|j - i_\lambda| \le N^{\omega_1}$. Here $\mathfrak{m}_{\lambda,t_1}$ and $\mathfrak{m}_{\mu,t_1}$ are the local minimum of ρ_{λ,t_1} and ρ_{μ,t_1}, respectively.

We introduce the shifted process $\widetilde{r}_i(t) = \widetilde{x}_i(t), \widetilde{y}_i(t)$ for $t \ge t_*$. Let us define

$$\widetilde{r}_i(t) := r_i(t) - \widetilde{\mathfrak{m}}_{r,t}, \qquad 1 \le |i| \le N,$$

for $r = x, y$, hence, by (5.22d), the shifted points satisfy the following DBM

$$\mathrm{d}\widetilde{r}_i(t) = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \frac{1}{N}\sum_{j\ne i}\frac{1}{\widetilde{r}_i(t) - \widetilde{r}_j(t)}\,\mathrm{d}t - \Big(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mathfrak{m}}_{r,t}\Big)\,\mathrm{d}t.$$

Furthermore we recall that by $\widehat{\gamma}_{r,i}(t)$ we the denote the quantiles of $\rho_{r,t}$, with $r = x, y$, for all $t_* \le t \le 2t_*$, i.e.

$$\widehat{\gamma}_{r,i} = \gamma_{r,i} - \widetilde{\mathfrak{m}}_{r,t}, \qquad 1 \le |i| \le N.$$

By the rigidity estimate of Corollary 4.2.6, using Lemma 5.5.1 and the fluctuation scale estimate in (5.27a) the proof of the following lemma is immediate.

Lemma 5.8.2. *Let $\widetilde{r}(t) = \widetilde{x}(t), \widetilde{y}(t)$. There exists a fixed small $\epsilon > 0$, such that for each $1 \leq |i| \leq \epsilon N$, we have*

$$\sup_{t_* \leq t \leq t_1} |\widetilde{r}_i(t) - \widehat{\gamma}_{r,i}(t)| \leq N^{\xi} \eta_{\mathrm{f}}^{\rho_{r,t}}(\gamma_{r,i}(t)), \tag{5.169}$$

for any $\xi > 0$ with very high probability, where we recall that the behavior of $\eta_{\mathrm{f}}^{\rho_{r,t}}(\mathfrak{e}_{r,t}^{+} + \widehat{\gamma}_{r,\pm i}(t))$, with $r = x, y$, is given by (5.27b).

In order to prove Proposition 5.8.1, by Lemma 5.5.1 and (5.22b), it is enough to prove the following proposition.

Proposition 5.8.3. *For $t_1 \geq t_*$ we have, with very high probability, that*

$$|(x_i(t_1) - \widetilde{\mathfrak{m}}_{x,t_1}) - (y_i(t_1) - \widetilde{\mathfrak{m}}_{y,t_1})| \leq N^{-\frac{3}{4}-c}$$

for some small constant $c > 0$ and for any $1 \leq |i| \leq N^{\omega_1}$.

The remaining part of this section is devoted to the proof of Proposition 5.8.3. We start with some preparatory lemmas. We recall the definition of the interpolated quantiles given in Section 5.5,

$$\overline{\gamma}_i(t) := \alpha\widehat{\gamma}_{x,i}(t) + (1-\alpha)\widehat{\gamma}_{y,i}(t), \tag{5.170}$$

for all $\alpha \in [0,1]$ and $t_* \leq t \leq 2t_*$, as well as

$$\overline{\mathfrak{m}}_t := \alpha\widetilde{\mathfrak{m}}_{x,t} + (1-\alpha)\widetilde{\mathfrak{m}}_{y,t},$$

for all $\alpha \in [0,1]$ and $t_* \leq t \leq 2t_*$. Furthermore by $\overline{\rho}_t$ from (5.46) we denote the interpolated density between $\rho_{x,t}$ and $\rho_{y,t}$ and by $\overline{m}_t$ its Stieltjes transform.

We now define the process $\widetilde{z}_i(t,\alpha)$ whose initial data are given by the linear interpolation of $\widetilde{x}(0)$ and $\widetilde{y}(0)$. Analogously to the small gap case, we define the function $\Psi_\alpha(t)$, for $t_* \leq t \leq 2t_*$, that represents the correct shift of the process $\widetilde{z}(t,\alpha)$, in order to compensate the discrepancy of our choice of the interpolation for $\overline{\rho}_t$ with respect to the semicircular flow evolution of the density $\overline{\rho}_0$.

Analogously to the edge case, see (5.50)-(5.53), we define $h(t,\alpha)$ with the following properties

$$\begin{aligned} h(t,\alpha) = \alpha\Re[m_{x,t}(\widetilde{\mathfrak{m}}_{x,t})] + (1-\alpha)\Re[m_{y,t}(\widetilde{\mathfrak{m}}_{y,t})] \\ - \Re[\overline{m}_t(\overline{\mathfrak{m}}_t + \mathrm{i}N^{-100})] + \mathcal{O}\left(N^{-1}\right) \end{aligned} \tag{5.171}$$

and $h(t,0) = h(t,1) = 0$. Then, similarly to the edge case, we define

$$\Psi_\alpha(t) := -\alpha\frac{\mathrm{d}}{\mathrm{d}t}[m_{x,t}(\widetilde{\mathfrak{m}}_{x,t})] - (1-\alpha)\frac{\mathrm{d}}{\mathrm{d}t}[m_{y,t}(\widetilde{\mathfrak{m}}_{y,t})] - h(t,\alpha). \tag{5.172}$$

In particular, by our definition of $h(t,\alpha)$ in (5.171) it follows that $\Psi_0(t) = \frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mathfrak{m}}_{y,t}$, $\Psi_1(t) = \frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mathfrak{m}}_{x,t}$ and that

$$\Psi_\alpha(t) = \Re[\overline{m}_t(\overline{\mathfrak{m}}_t)] + \mathcal{O}\left(N^{-\frac{1}{2}+\omega_1}\right). \tag{5.173}$$

Note that the error in (5.173) is somewhat weaker than in the analogous equation (5.57) due to the additional error in (5.22d) compared with (5.22c).

More precisely, the process $\widetilde{z}(t,\alpha)$ is defined by

$$\mathrm{d}\widetilde{z}_i(t,\alpha) = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \left[\frac{1}{N}\sum_{j\neq i}\frac{1}{\widetilde{z}_i(t,\alpha)-\widetilde{z}_j(t,\alpha)} + \Psi_\alpha(t)\right]\mathrm{d}t, \tag{5.174}$$

with initial data

$$\widetilde{z}_i(t_*,\alpha) := \alpha\widetilde{x}_i(t_*) + (1-\alpha)\widetilde{y}_i(t_*), \tag{5.175}$$

for all $1 \leq |i| \leq N$ and for all $\alpha \in [0,1]$.

We recall that $\omega_1 \ll \omega_\ell \ll \omega_A \ll 1$ and that $i_* = N^{\frac{1}{2}+C_*\omega_1}$ with some large constant C_*.

Next, we define the analogue of $\mathcal{J}_z(t)$ and $\mathcal{I}_{z,i}(t)$ for the small minimum by (5.118) and (5.119) using the definition in (5.170) for the quantiles. Then, for each $t_* \leq t \leq t_1$, we define the short range approximation $\widehat{z}_i(t,\alpha)$ of $\widetilde{z}(t,\alpha)$ by the following SDE.

For $|i| > \frac{i_*}{2}$ we let

$$\begin{aligned}\mathrm{d}\widehat{z}_i(t,\alpha) = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \Bigg[&\frac{1}{N}\sum_j^{\mathcal{A},(i)}\frac{1}{\widehat{z}_i(t,\alpha)-\widehat{z}_j(t,\alpha)} \\ &+ \frac{1}{N}\sum_j^{\mathcal{A}^c,(i)}\frac{1}{\widetilde{z}_i(t,\alpha)-\widetilde{z}_j(t,\alpha)} + \Psi_\alpha(t)\Bigg]\mathrm{d}t,\end{aligned} \tag{5.176}$$

for $|i| \leq N^{\omega_A}$

$$\begin{aligned}\mathrm{d}\widehat{z}_i(t,\alpha) = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \Bigg[&\frac{1}{N}\sum_j^{\mathcal{A},(i)}\frac{1}{\widehat{z}_i(t,\alpha)-\widehat{z}_j(t,\alpha)} \\ &+ \int_{\mathcal{I}_{y,i}(t)^c}\frac{\rho_{y,t}(E+\widetilde{\mathfrak{m}}_{y,t}^+)}{\widehat{z}_i(t,\alpha)-E}\,\mathrm{d}E\Bigg]\mathrm{d}t - \Big(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mathfrak{m}}_{r,t}\Big)\,\mathrm{d}t,\end{aligned}$$

and for $N^{\omega_A} < |i| \leq \frac{i_*}{2}$

$$\begin{aligned}\mathrm{d}\widehat{z}_i(t,\alpha) = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \Bigg[&\frac{1}{N}\sum_j^{\mathcal{A},(i)}\frac{1}{\widehat{z}_i(t,\alpha)-\widehat{z}_j(t,\alpha)} + \int_{\mathcal{I}_{z,i}(t)^c\cap\mathcal{J}_z(t)}\frac{\overline{\rho}_t(E+\overline{\mathfrak{m}}_t^+)}{\widehat{z}_i(t,\alpha)-E}\,\mathrm{d}E \\ &+ \sum_{|j|\geq\frac{3}{4}i_*}\frac{1}{\widetilde{z}_i(t,\alpha)-\widetilde{z}_j(t,\alpha)} + \Psi_\alpha(t)\Bigg]\mathrm{d}t,\end{aligned}$$

with initial data

$$\widehat{z}_i(t_*,\alpha) := \widetilde{z}_i(t_*,\alpha). \tag{5.177}$$

Next, by Lemma 5.C.2, by the optimal rigidity in (5.169) for $\widetilde{x}(t)$ and $\widetilde{y}(t)$, the next lemma follows immediately.

Lemma 5.8.4. *For $\alpha = 0$ and $\alpha = 1$, with very high probability, we have*

$$\sup_{1\leq|i|\leq N}\ \sup_{t_*\leq t\leq t_1}|\widetilde{z}_i(t,\alpha)-\widehat{z}_i(t,\alpha)| \lesssim \frac{N^\xi}{N^{\frac{3}{4}}}\left(\frac{N^{\omega_1}}{N^{3\omega_\ell}} + \frac{N^{C\omega_1}}{N^{\frac{1}{24}}}\right),$$

for any $\xi > 0$ and $C > 0$ a large universal constant.

In order to proceed with the heat-kernel estimates we need an optimal i-dependent rigidity for $\widehat{z}_i(t,\alpha)$ for $1 \leq |i| \leq N^{4\omega_\ell+\delta}$, for some $0 < \delta < C\omega_\ell$. In particular, analogously to Proposition 5.6.11, we have:

Proposition 5.8.5. *Fix any $\alpha \in [0,1]$. There exists a small fixed $0 < \delta_1 < C\omega_\ell$, for some constant $C > 0$, such that*

$$\sup_{t_* \leq t \leq 2t_*} |\widehat{z}_i(t,\alpha) - \overline{\gamma}_i(t)| \lesssim \frac{N^\xi N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}} |i|^{\frac{1}{4}}}, \qquad 1 \leq |i| \leq N^{4\omega_\ell+\delta_1}$$

for any $\xi > 0$ with very high probability.

Proof. We can adapt the arguments in Section 5.6 to the case of the small minimum, $t \geq t_*$, in a straightforward way. In Section 5.6, as the main input, we used the precise estimates on the density $\rho_{r,t}$ (5.17b), (5.30), on the quantiles $\widehat{\gamma}_{r,i}(t)$ (5.26a), on the quantile gaps (5.28), on the fluctuation scale (5.27a) and on the Stieltjes transform (5.32a); all formulated for the small gap case, $0 \leq t \leq t_*$. In the small minimum case, $t \geq t_*$, the corresponding estimates are all available in Section 5.4, see (5.17d), (5.31), (5.26b), (5.29), (5.27b) and (5.32b), respectively. In fact, the semicircular flow is more regular after the cusp formation, see e.g. the better (larger) exponent in the $(t - t_*)$ error terms when comparing (5.17b) with (5.17d). This makes handling the small minimum case easier. The most critical part in Section 5.6 is the estimate of the forcing term (Proposition 5.6.6), where the derivative of the density (5.18a) was heavily used. The main mechanism of this proof is the delicate cancellation between the contributions to S_2 from the intervals $[\gamma_{i-n-1}, \gamma_{i-n}]$ and $[\gamma_{i+n-1}, \gamma_{i+n}]$, see (5.113). This cancellation takes place away from the edge. The proof is divided into two cases; the so-called "edge regime", where the gap length Δ is relatively large and the "cusp regime", where Δ is small or zero. The adaptation of this argument to the small minimum case, $t \geq t_*$, will be identical to the proof for the small gap case in the "cusp regime". In this regime the derivative bound (5.18a) is used only in the form $|\rho'| \lesssim \rho^{-2}$ which is available in the small minimum case, $t \geq t_*$, as well, see (5.19a). This proves Proposition 5.6.6 for $t \geq t_*$. The rest of the argument is identical to the proof in the small minimum case up to obvious notational changes; the details are left to the reader. □

Let us define $u_i(t,\alpha) := \partial_\alpha \widehat{z}_i(t,\alpha)$, for $t_* \leq t \leq 2t_*$. In particular, u is a solution of the equation

$$\partial_t u = \mathcal{L}u + \zeta^{(0)} \tag{5.178}$$

with initial condition $u(t_*,\alpha) = \widetilde{x}(t_*) - \widetilde{y}(t_*)$ from (5.175). The error term $\zeta^{(0)}$ is defined analogously to (5.130)-(5.131) but replacing Φ_α and $\overline{\mathfrak{e}}_t^+$ with Ψ_α and $\widetilde{\mathfrak{m}}_t$, respectively. Note that this error term is non zero only for $|i| \geq N^{\omega_A}$ and for any i we have $\left|\zeta_i^{(0)}\right| \leq N^C$ with very high probability, for some large $C > 0$. Furthermore, $\mathcal{L} = \mathcal{B} + \mathcal{V}$ is defined as in (5.128)-(5.129) replacing $\mathfrak{e}_{y,t}^+$ and $\overline{\mathfrak{e}}_t^+$ by $\widetilde{\mathfrak{m}}_{y,t}$ and $\overline{\mathfrak{m}}_t$, respectively. In the following by $\mathcal{U}^\mathcal{L}$ we denote the propagator of the operator $\mathcal{L}$.

Let $0 < \delta_v < \frac{\delta_4}{2}$, with δ_4 defined in Lemma 5.7.6. Define $v_i = v_i(t,\alpha)$ to be the solution of

$$\partial_t v = \mathcal{L}v, \qquad v_i(t_*,\alpha) = u_i(t_*,\alpha)\mathbf{1}_{\{|i| \leq N^{4\omega_\ell+\delta_v}\}}. \tag{5.179}$$

By the finite speed of propagation estimate in Lemma 5.B.3, similarly to the proof of Lemma 5.7.7, we immediately obtain the following:

Lemma 5.8.6. *Let u be the solution of the equation in* (5.178) *and v defined by* (5.179)*, then we have that*

$$\sup_{t_* \le t \le 2t_*} \sup_{1 \le |i| \le \ell^4} |u_i(t) - v_i(t)| \le N^{-100}$$

with very high probability.

Collecting all the previous lemmas we conclude this section with the proof of Proposition 5.8.3.

Proof of Proposition 5.8.3. We consider only the $i = 1$ case. By Lemma 5.5.1 and Lemma 5.8.4 we have that

$$\begin{aligned}&|(x_1(t_1) - \widetilde{\mathfrak{m}}_{x,t_1}) - (y_1(t_1) - \widetilde{\mathfrak{m}}_{y,t_1})|\\ &\le |\widetilde{x}_1(t_1) - \widehat{x}_1(t_1)| + |\widehat{x}_1(t_1) - \widehat{y}_1(t_1)| + |\widehat{y}_1(t_1) - \widetilde{y}_1(t_1)| \le |\widehat{x}_1(t_1) - \widehat{y}_1(t_1)| + \frac{1}{N^{\frac{3}{4}+c}}\end{aligned}$$

with very high probability. Since $u(t,\alpha) = \partial_\alpha \widehat{z}(t,\alpha)$, using $\widehat{x}_1(t_1) - \widehat{y}_1(t_1) = \int_0^1 u(t_1,\alpha)\, \mathrm{d}\alpha$ and Lemma 5.8.6 it will be sufficient to estimate $\int_0^1 |v_1(t_1,\alpha)|\, \mathrm{d}\alpha$. By rigidity from (5.169), we have

$$|v_i(t_*,\alpha)| = |u_i(t_*,\alpha)| = |\widetilde{y}_i(t_*) - \widetilde{x}_i(t_*)| \lesssim \frac{N^\xi}{N^{\frac{3}{4}} |i|^{\frac{1}{4}}},$$

for any $1 \le |i| \le N^{4\omega_\ell + \delta_v}$ hence

$$\|v(t_*,\alpha)\|_5 \lesssim \frac{N^\xi}{N^{\frac{3}{4}}},$$

for any $\xi > 0$ with very high probability.

Finally, using the heat kernel estimate in (5.165) for $\mathcal{U}^{\mathcal{L}}(0,t)$ for $t_* \le t \le 2t_*$, we conclude, after a Markov inequality as in (5.81)-(5.82),

$$\int_0^1 |v_1(t_1,\alpha)|\, \mathrm{d}\alpha \lesssim \frac{N^\xi}{N^{\frac{3}{4}} N^{\frac{4\omega_1}{15}}},$$

with very high probability. □

5.A Proof of Theorem 5.2.4

We now briefly outline the changes required for the proof of Theorem 5.2.4 compared to the proof of Theorem 5.2.2. We first note that for $0 \le \tau_1 \le \cdots \le \tau_k \lesssim N^{-1/2}$ in distribution $(H^{(\tau_1)}, \dots, H^{(\tau_k)})$ agrees with

$$\Big(H + \sqrt{\tau_1} U_1, H + \sqrt{\tau_1} U_1 + \sqrt{\tau_2 - \tau_1} U_2, \dots, H + \sqrt{\tau_1} U_1 + \cdots + \sqrt{\tau_k - \tau_{k-1}} U_k\Big), \tag{5.180}$$

where $U_1, \dots, U_k$ are independent GOE matrices. Next, we claim and prove later by Green function comparison that the time-dependent k-point correlation function of (5.180) asymptotically agrees with the one of

$$\Big(\widetilde{H}_t + \sqrt{\tau_1} U_1, \widetilde{H}_t + \sqrt{\tau_1} U_1 + \sqrt{\tau_2 - \tau_1} U_2, \dots, \widetilde{H}_t + \sqrt{\tau_1} U_1 + \cdots + \sqrt{\tau_k - \tau_{k-1}} U_k\Big), \tag{5.181}$$

and thereby also with the one of

$$\begin{aligned}\Big(H_t + \sqrt{ct}U + \sqrt{\tau_1}U_1, H_t + \sqrt{ct}U + \sqrt{\tau_1}U_1 + \sqrt{\tau_2 - \tau_1}U_2,\\ \dots, H_t + \sqrt{ct}U + \sqrt{\tau_1}U_1 + \cdots + \sqrt{\tau_k - \tau_{k-1}}U_k\Big),\end{aligned} \tag{5.182}$$

for any fixed $t \leq N^{-1/4-\epsilon}$, where we recall that $\widetilde{H}_t$ and H_t constructed as in Section 5.3 (see (5.9)). Finally, we notice that the joint eigenvalue distribution of the matrices in (5.182) is precisely given by the joint distribution of

$$\Big(\lambda_i(ct+\tau_1), \dots, \lambda_i(ct+\tau_k),\ i \in [N]\Big)$$

of eigenvalues λ_i^s evolved according to the DBM

$$\mathrm{d}\lambda_i(s) = \sqrt{\frac{2}{N}}\,\mathrm{d}B_i + \sum_{j\neq i} \frac{1}{\lambda_i(s) - \lambda_j(s)}\,\mathrm{d}s, \quad \lambda_i(0) = \lambda_i(H_t). \tag{5.183}$$

The high probability control on the eigenvalues evolved according to (5.183) in Propositions 5.7.1 and 5.8.1 allows to simultaneously compare eigenvalues at different times with those of the Gaussian reference ensemble automatically.

In order to establish Theorem 5.2.4 it thus only remains to argue that the k-point functions of (5.180) and (5.181) are asymptotically equal. For the sake of this argument we consider only the randomness in H and condition on the randomness in $U_1, \dots, U_k$. Then the OU-flow

$$\mathrm{d}\widetilde{H}_s' = -\frac{1}{2}(\widetilde{H}_s' - A - \sqrt{\tau_1}U_1 - \cdots - \sqrt{\tau_l - \tau_{l-1}}U_l)\,\mathrm{d}s + \Sigma^{1/2}[\mathrm{d}\mathfrak{B}_s]$$

with initial conditions

$$\widetilde{H}_0' = H + \sqrt{\tau_1}U_1 + \cdots + \sqrt{\tau_l - \tau_{l-1}}U_l$$

for fixed $U_1, \dots, U_l$ is given by

$$\widetilde{H}_s' = \widetilde{H}_s + \sqrt{\tau_1}U_1 + \cdots + \sqrt{\tau_l - \tau_{l-1}}U_l,$$

i.e. we view $\sqrt{\tau_1}U_1 + \cdots + \sqrt{\tau_l - \tau_{l-1}}U_l$ as an additional expectation matrix. Thus we can appeal to the standard Green function comparison technique already used in Section 5.3 to compare the k-point functions of (5.180) and (5.181). Here we can follow the standard resolvent expansion argument from (4.115) and note that the proof therein verbatim also allows us to compare products of traces of resolvents with differing expectations. Finally we then take the $\mathbf{E}_{U_1} \dots \mathbf{E}_{U_k}$ expectation to conclude that not only the conditioned k-point functions of (5.180) and (5.181) asymptotically agree, but also the k-point functions themselves.

5.B Finite speed of propagation estimate

In this section we prove a finite speed of propagation estimate for the time evolution of the α-derivative of the short range dynamics defined in (5.120)–(5.122). It is an adjustment to the analogous proof of Lemma 4.1 in [122]. For concreteness, we present the proof for the propagator $\mathcal{U}^{\mathcal{L}}$ where $\mathcal{L} = \mathcal{B} + \mathcal{V}$ is defined in (5.127)–(5.129). The point is that once the dynamics is localized, i.e. the range of the interaction term $\mathcal{B}$ is restricted to a local scale $|i-j| \leq |j_+(i) - j_-(i)|$, with $|j_+(i) - j_-(i)| \gtrsim N^{4\omega_\ell} =: L$, and the time is also restricted, $0 \leq t \leq 2t_* \lesssim N^{-\frac{1}{2}+\omega_1}$, then the propagation cannot go beyond a scale that is much bigger than the interaction scale. This mechanism is very general and will also be used in a slightly different (simpler) setup of Lemma 5.6.4 and Proposition 5.6.7 where the interaction scale is much bigger $L \sim \sqrt{N}$. We will give the necessary changes for the proof of Lemma 5.6.4 and Proposition 5.6.7 at the end of this section.

Lemma 5.B.1. *Let $\widehat{z}(t) = \widehat{z}(t,\alpha)$ be the solution to the short range dynamics* (5.120)–(5.122) *with $i_* = N^{\frac{1}{2}+C_*\omega_1}$, exponents $\omega_1 \ll \omega_\ell \ll \omega_A \ll 1$ and propagator $\mathcal{L} = \mathcal{B} + \mathcal{V}$ from* (5.127)–(5.129)*. Let us assume that*

$$\sup_{0 \leq t \leq t_*} |\widehat{z}_i(t) - \overline{\gamma}_i(t)| \leq \frac{N^{C\omega_1}}{N^{\frac{3}{4}}}, \qquad 1 \leq |i| \leq i_*, \tag{5.184}$$

where $\overline{\gamma}_i(t)$ are the quantiles from (5.44)*. Then, there exists a constant $C' > 0$ such that for any $0 < \delta < C'\omega_\ell$, $|a| \geq LN^\delta$ and $|b| \leq \frac{3}{4}LN^\delta$, for any fixed $0 \leq s \leq t_*$, we have that*

$$\sup_{s \leq t \leq t_*} \mathcal{U}^{\mathcal{L}}_{ab}(s,t) + \mathcal{U}^{\mathcal{L}}_{ba}(s,t) \leq N^{-D} \tag{5.185}$$

for any $D > 0$, with very high probability. The same result holds for the short range dynamics after the cusp defined in (5.178) *for $t_* \leq s \leq 2t_*$.*

Proof of Lemma 5.B.1. For concreteness we assume that $0 \leq s \leq t \leq t_*$, i.e. we are in the small gap regime. For $t_* \leq s \leq t \leq 2t_*$ the proof is analogous using the definition (5.170) for the $\overline{\gamma}_i(t)$, the definition of the short range approximation in (5.176)-(5.177) for the $\widehat{z}_i(t,\alpha)$ and replacing $\overline{\mathfrak{e}}_t^+$ by $\overline{\mathfrak{m}}_t$. With these adjustments the proof follows in the same way except for (5.200) below, where we have to use the estimates in (5.32b) instead of (5.32a).

First we consider the $s = 0$ case, then in Lemma 5.B.3 below we extend the proof for all $0 \leq s \leq t$. Let $\psi(x)$ be an even 1-Lipschitz real function, i.e. $\psi(x) = \psi(-x)$, $\|\psi'\|_\infty \leq 1$ such that

$$\psi(x) = |x| \quad \text{for} \quad |x| \leq \frac{L^{\frac{3}{4}} N^{\frac{3}{4}\delta}}{N^{\frac{3}{4}}}, \quad \psi'(x) = 0 \quad \text{for} \quad |x| \geq 2\frac{L^{\frac{3}{4}} N^{\frac{3}{4}\delta}}{N^{\frac{3}{4}}}.$$

and

$$\|\psi''\|_\infty \lesssim \frac{N^{\frac{3}{4}}}{L^{\frac{3}{4}} N^{\frac{3\delta}{4}}}. \tag{5.186}$$

We consider a solution of the equation

$$\partial_t f = \mathcal{L} f, \qquad 0 \leq t \leq t_*$$

with some discrete Dirac delta initial condition $f_i(0) = \delta_{ip_*}$ at p_* for any $|p_*| \geq N^{4\omega_\ell} N^\delta$. For concreteness, assume $p_* > 0$ and set $p := N^{4\omega_\ell} N^\delta$. Define

$$\phi_i(t,\alpha) := e^{\frac{1}{2}\nu\psi(\widehat{z}_i(t,\alpha) - \overline{\gamma}_p(t))}, \quad m_i(t,\alpha) := f_i(t,\alpha)\phi_i(t,\alpha), \quad \nu = \frac{N^{\frac{3}{4}}}{L^{\frac{3}{4}} N^{\delta'}} \tag{5.187}$$

with some $\delta' \geq \frac{\delta}{2}$ to be chosen later. Let $\widehat{z}_i = \widehat{z}_i(t,\alpha)$ and set

$$F(t) := \sum_i f_i^2 e^{\nu\psi(\widehat{z}_i - \overline{\gamma}_p(t))} = \sum_i m_i^2.$$

Since

$$2\sum_i f_i(\mathcal{B}f)_i \phi_i^2 = \sum_{(i,j)\in\mathcal{A}} \mathcal{B}_{ij}(m_i - m_j)^2 - \sum_{(i,j)\in\mathcal{A}} \mathcal{B}_{ij} m_i m_j \left(\frac{\phi_i}{\phi_j} + \frac{\phi_j}{\phi_i} - 2\right),$$

using Ito's formula, we conclude that

$$\mathrm{d}F = \sum_{(i,j)\in\mathcal{A}} \mathcal{B}_{ij}(m_i - m_j)^2 \,\mathrm{d}t + 2\sum_i \mathcal{V}_i m_i^2 \,\mathrm{d}t \tag{5.188}$$

$$- \sum_{(i,j)\in\mathcal{A}} \mathcal{B}_{ij} m_i m_j \left(\frac{\phi_i}{\phi_j} + \frac{\phi_j}{\phi_i} - 2\right) \mathrm{d}t \tag{5.189}$$

$$+ \sum_i \nu m_i^2 \psi'(\widehat{z}_i - \overline{\gamma}_p)\,\mathrm{d}(\widehat{z}_i - \overline{\gamma}_p) \tag{5.190}$$

$$+ \sum_i m_i^2 \left(\frac{\nu^2}{N}\psi'(\widehat{z}_i - \overline{\gamma}_p)^2 + \frac{\nu}{N}\psi''(\widehat{z}_i - \overline{\gamma}_p)\right) \mathrm{d}t. \tag{5.191}$$

Let $\tau_1 \leq t_*$ be the first time such that $F \geq 5$ and let τ_2 be stopping time so that the estimate (5.184) holds with $t \leq \tau_2$ instead of $t \leq t_*$; the condition (5.184) then says that $\tau_2 = t_*$ with very high probability. Define $\tau := \tau_1 \wedge \tau_2 \wedge t_*$, our goal is to show that $\tau = t_*$. In the following we assume $t \leq \tau$.

Now we estimate the terms in (5.188)–(5.191) one by one. We start with (5.189). Note that the rigidity scale $N^{-\frac{3}{4}+C\omega_1}$ in (5.184) is much smaller than $N^{-\frac{3}{4}(1-\delta)+3\omega_\ell}$, the range of the support of ψ', which, in turn, is comparable with $\left|\overline{\gamma}_i - \overline{\gamma}_p\right| \gtrsim (p/N)^{3/4}$ for any $i \geq 2p = 2LN^{\delta}$. Therefore $\psi'(\widehat{z}_i - \overline{\gamma}_p) = 0$ unless $|i| \lesssim LN^{\delta}$. Moreover, if $|i| \lesssim LN^{\delta}$ and $(i,j) \in \mathcal{A}$, then $|j| \lesssim LN^{\delta}$. Hence, the nonzero terms in the sum in (5.189) have both indices $|i|, |j| \lesssim N^{4\omega_\ell+\delta}$. By (5.184) and $C\omega_1 \ll \omega_\ell$, for such terms we have

$$|\widehat{z}_i - \widehat{z}_j| \lesssim \frac{|i-j|}{N^{\frac{3}{4}}\min\{|i|,|j|\}^{\frac{1}{4}}} + \frac{N^{C\omega_1}}{N^{\frac{3}{4}}} \lesssim \frac{L^{\frac{3}{4}}N^{\frac{\delta}{2}}}{N^{\frac{3}{4}}}.$$

Note that $\nu\,|\widehat{z}_i - \widehat{z}_j| \lesssim 1$. Therefore, by Taylor expanding in the exponent, we have

$$\begin{aligned}\left|\frac{\phi_i}{\phi_j} + \frac{\phi_j}{\phi_i} - 2\right| &= \left(e^{\frac{\nu}{2}(\psi(\widehat{z}_j - \overline{\gamma}_p) - \psi(\widehat{z}_i - \overline{\gamma}_p))} - e^{\frac{\nu}{2}(\psi(\widehat{z}_i - \overline{\gamma}_p) - \psi(\widehat{z}_j - \overline{\gamma}_p))}\right)^2 \\ &\lesssim \nu^2 \left|\psi(\widehat{z}_i - \overline{\gamma}_p) - \psi(\widehat{z}_j - \overline{\gamma}_p)\right|^2,\end{aligned}$$

and thus

$$\left|\mathcal{B}_{ij}\left(\frac{\phi_i}{\phi_j} + \frac{\phi_j}{\phi_i} - 2\right)\right| \lesssim \nu^2 \frac{\left|\psi(\widehat{z}_i - \overline{\gamma}_p) - \psi(\widehat{z}_j - \overline{\gamma}_p)\right|^2}{N(\widehat{z}_i - \widehat{z}_j)^2} \lesssim \frac{\nu^2}{N},$$

where in the last inequality we used that ψ is Lipschitz continuous. Hence we conclude the estimate of (5.189) as

$$\left|\sum_{(i,j)\in\mathcal{A}}\mathcal{B}_{ij}m_im_j\left(\frac{\phi_i}{\phi_j}+\frac{\phi_j}{\phi_i}-2\right)\right|\lesssim\frac{\nu^2}{N}\sum_i m_i^2\sum_j^{\mathcal{A},(i)}\mathbf{1}_{\{\phi_j\neq\phi_i\}}\lesssim\frac{\nu^2LN^{\frac{3}{4}\delta}}{N}F(t),$$

since the number of j's in the summation is at most

$$|j_+(i)-j_-(i)|\le\ell^4+\ell\,|i|^{3/4}\le LN^{3\delta/4}. \tag{5.192}$$

By (5.186) and since $|\psi'(x)|\le 1$, (5.191) is bounded as follows

$$\left|\sum_i m_i^2\left(\frac{\nu^2}{N}\psi'(\widehat{z}_i-\overline{\gamma}_p)^2+\frac{\nu}{N}\psi''(\widehat{z}_i-\overline{\gamma}_p)\right)\right|\lesssim\left(\frac{\nu^2}{N}+\frac{\nu}{N^{\frac{1}{4}}L^{\frac{3}{4}}N^{\frac{3}{4}\delta}}\right)F(t).$$

The next step is to get a bound for (5.190). Since $\psi'(\widehat{z}_i-\overline{\gamma}_p)=0$ unless $|i|\lesssim N^{4\omega_\ell+\delta}\ll N^{\omega_A}$, choosing $C>0$ such that $(4+C)\omega_\ell<\omega_A$ and using (5.121) we get

$$\mathrm{d}(\widehat{z}_i(t)-\overline{\gamma}_p(t))=\sqrt{\frac{2}{N}}\,\mathrm{d}B_i+\frac{1}{N}\sum_j^{\mathcal{A},(i)}\frac{1}{\widehat{z}_i(t)-\widehat{z}_j(t)}+Q_i(t) \tag{5.193}$$

with

$$\begin{aligned}Q_i(t):=&\int_{\mathcal{I}_{y,i}(t)^c}\frac{\rho_{y,t}(E+\mathfrak{e}_{y,t}^+)}{\widehat{z}_i(t)-E}\,\mathrm{d}E+\alpha\Big(\Re[m_{x,t}(\widehat{\gamma}_{x,p}(t)+\mathfrak{e}_{x,t}^+)-m_{x,t}(\mathfrak{e}_{x,t}^+)]\Big)\\&+(1-\alpha)\Big(\Re[m_{y,t}(\widehat{\gamma}_{y,p}(t)+\mathfrak{e}_{y,t}^+)-m_{y,t}(\mathfrak{e}_{y,t}^+)]\Big)+\Re[m_{y,t}(\mathfrak{e}_{y,t}^+)].\end{aligned} \tag{5.194}$$

We insert (5.193) into (5.190) and estimate all three terms separately in the regime $|i|\lesssim LN^\delta$. For the stochastic differential, by the definition of $\tau\le t_*$ and the Burkholder-Davis-Gundy inequality we have that

$$\sup_{0\le t\le\tau}\int_0^t\sqrt{\frac{2}{N}}\nu\sum_i m_i^2\psi'(\widehat{z}_i-\overline{\gamma}_p)\,\mathrm{d}B_i\le N^{\epsilon'}\frac{\nu}{\sqrt{N}}\sqrt{t_*}\sup_{0\le t\le\tau}F(t)\lesssim\nu N^{\epsilon'}N^{-\frac{3}{4}+\frac{1}{2}\omega_1}, \tag{5.195}$$

for any $\epsilon'>0$, with very high probability. In (5.195) we used that $\tau\le t_*\sim N^{-\frac{1}{2}+\omega_1}$, and that, by the definition of τ, $F(t)$ is bounded for all $0\le t\le\tau$.

The contribution of the second term in (5.193) to (5.190) is written, after symmetrisation, as follows

$$\begin{aligned}\frac{\nu}{N}\sum_{(i,j)\in\mathcal{A}}\frac{\psi'(\widehat{z}_i-\overline{\gamma}_p)m_i^2}{\widehat{z}_j-\widehat{z}_i}=&\frac{\nu}{2N}\sum_{(i,j)\in\mathcal{A}}\frac{\psi'(\widehat{z}_i-\overline{\gamma}_p)(m_i^2-m_j^2)}{\widehat{z}_j-\widehat{z}_i}\\&+\frac{\nu}{2N}\sum_{(i,j)\in\mathcal{A}}m_i^2\frac{\psi'(\widehat{z}_i-\overline{\gamma}_p)-\psi'(\widehat{z}_j-\overline{\gamma}_p)}{\widehat{z}_j-\widehat{z}_i}.\end{aligned} \tag{5.196}$$

Using (5.186) and (5.192), the second sum in (5.196) is bounded by

$$\begin{aligned}&\left|\frac{\nu}{2N}\sum_{(i,j)\in\mathcal{A}}m_i^2\frac{\psi'(\widehat{z}_i-\overline{\gamma}_p)-\psi'(\widehat{z}_j-\overline{\gamma}_p)}{\widehat{z}_j-\widehat{z}_i}\right|\\&\qquad\lesssim\frac{\nu}{N^{\frac{1}{4}}L^{\frac{3}{4}}N^{\frac{3}{4}\delta}}\sum_i m_i^2\sum_j^{\mathcal{A},(i)}\mathbf{1}_{\{\psi'(\widehat{z}_i-\overline{\gamma}_p)\neq\psi'(\widehat{z}_j-\overline{\gamma}_p)\}}\lesssim\frac{\nu L^{\frac{1}{4}}}{N^{\frac{1}{4}}}F.\end{aligned}$$

Using $m_i^2 - m_j^2 = (m_i - m_j)(m_i + m_j)$ and Schwarz inequality, the first sum in (5.196) is bounded as follows

$$\begin{aligned} \frac{\nu}{2N} \sum_{(i,j)\in\mathcal{A}} \frac{\psi'(\widehat{z}_i - \overline{\gamma}_p)(m_i^2 - m_j^2)}{\widehat{z}_j - \widehat{z}_i} &\le -\frac{1}{100} \sum_{(i,j)\in\mathcal{A}} \mathcal{B}_{ij}(m_i - m_j)^2 \\ &\quad + \frac{C\nu^2}{2N} \sum_{(i,j)\in\mathcal{A}} \psi'(\widehat{z}_i - \overline{\gamma}_p)^2 (m_i^2 + m_j^2). \end{aligned} \tag{5.197}$$

The second sum in (5.197), using (5.192), is bounded by

$$\frac{C\nu^2}{2N} \sum_{(i,j)\in\mathcal{A}} \psi'(\widehat{z}_i - \overline{\gamma}_p)(m_i^2 + m_j^2) \le \frac{C\nu^2 L N^{\frac{3\delta}{4}}}{2N} F,$$

hence we conclude that

$$\begin{aligned} \frac{\nu}{N} \sum_{(i,j)\in\mathcal{A}} \frac{\psi'(\widehat{z}_i - \overline{\gamma}_p) m_i^2}{\widehat{z}_j - \widehat{z}_i} &\le -\frac{1}{100} \sum_{(i,j)\in\mathcal{A}} \mathcal{B}_{ij}(m_i - m_j)^2 \\ &\quad + C\left(\frac{\nu L^{\frac{1}{4}}}{N^{\frac{1}{4}}} + \frac{\nu^2 L N^{\frac{3\delta}{4}}}{N} \right) F. \end{aligned} \tag{5.198}$$

Note that the first term on the right-hand side of (5.198) can be incorporated in the first, dissipative term in (5.188).

To conclude the estimate of (5.190) we write the third term in (5.193)

$$\begin{aligned} Q_i(t) &= \left(\int_{\mathcal{I}_{y,i}(t)^c} \frac{\rho_{y,t}(E + \mathfrak{e}_{y,t}^+)}{\widehat{z}_i(t) - E} \,\mathrm{d}E + \Re[m_{y,t}(\overline{\gamma}_p(t) + \mathfrak{e}_{y,t}^+)] \right) \\ &\quad + \alpha\Big(\Re[m_{x,t}(\widehat{\gamma}_{x,p}(t) + \mathfrak{e}_{x,t}^+) - m_{x,t}(\mathfrak{e}_{x,t}^+)] - \Re[m_{y,t}(\widehat{\gamma}_{x,p}(t) + \mathfrak{e}_{y,t}^+) - m_{y,t}(\mathfrak{e}_{y,t}^+)] \Big) \\ &\quad + \alpha\Big(\Re[m_{y,t}(\widehat{\gamma}_{x,p}(t) + \mathfrak{e}_{y,t}^+)] - \Re[m_{y,t}(\overline{\gamma}_p(t) + \mathfrak{e}_{y,t}^+)] \Big) \\ &\quad + (1-\alpha)\Big(\Re[m_{y,t}(\widehat{\gamma}_{y,p}(t) + \mathfrak{e}_{y,t}^+)] - \Re[m_{y,t}(\overline{\gamma}_p(t) + \mathfrak{e}_{y,t}^+) \Big) =: A_1 + A_2 + A_3 + A_4. \end{aligned} \tag{5.199}$$

Similarly to the estimates in (5.93), for A_2 we use (5.32a) while for A_3, A_4 we use (5.18b), then we use the asymptotic behavior of $\widehat{\gamma}_p, \overline{\gamma}_p$ by (5.26a) and $p = LN^\delta$ to conclude that

$$|A_2| + |A_3| + |A_4| \lesssim \frac{L^{\frac{1}{4}} N^{\frac{\delta}{4}} N^{C\omega_1} \log N}{N^{\frac{1}{4}} N^{\frac{1}{6}}}. \tag{5.200}$$

For the A_1 term we write it as

$$\begin{aligned} A_1 &= \int_{\mathcal{I}_{y,i}(t)^c} \frac{\overline{\gamma}_p(t) - \widehat{z}_i(t)}{(\widehat{z}_i(t) - E)(\overline{\gamma}_p(t) - E)} \rho_{y,t}(E + \mathfrak{e}_{y,t}^+) \,\mathrm{d}E \\ &\quad + \int_{\mathcal{I}_{y,i}(t)} \frac{\rho_{y,t}(E + \mathfrak{e}_{y,t}^+)}{\overline{\gamma}_p(t) - E} \,\mathrm{d}E. \end{aligned} \tag{5.201}$$

Since $i \leq Cp$, we have $\rho_{y,t}(E + \mathfrak{e}_{y,t}^{+}) \leq \rho_{y,t}(\overline{\gamma}_{Cp}(t) + \mathfrak{e}_{y,t}^{+}) \lesssim L^{\frac{1}{4}} N^{-\frac{1}{4}+\frac{\delta}{4}}$ for any $E \in \mathcal{I}_{y,i}(t)$, the second term in (5.201) is bounded by $L^{\frac{1}{4}} N^{-\frac{1}{4}+\frac{\delta}{4}} \log N$. In the first term in (5.201) we use that

$$|\widehat{z}_i(t) - E| \geq |\overline{\gamma}_i(t) - E| - |\widehat{z}_i(t) - \overline{\gamma}_i(t)| \gtrsim \overline{\gamma}_p(t)$$

for $E \notin \mathcal{I}_{y,i}(t)$, by rigidity (5.184) and by the fact that in the $i \leq Cp$ regime

$$\left| \overline{\gamma}_i(t) - \overline{\gamma}_{i \pm j_{\pm}(i)}(t) \right| \gtrsim \overline{\gamma}_p(t) \gg N^{-\frac{3}{4}+C\omega_1}$$

since $\omega_1 \ll \omega_\ell$ and $= LN^{\delta} = N^{4\omega_\ell + \omega_1}$.

We thus conclude that the first term in (5.201) is bounded by

$$\left| \widehat{z}_i(t) - \overline{\gamma}_p(t) \right| \frac{\Im[m_{y,t}(\mathfrak{e}_{y,t}^{+} + \mathrm{i}\overline{\gamma}_p(t))]}{\overline{\gamma}_p(t)} \lesssim \overline{\gamma}_p^{\frac{1}{3}} \lesssim L^{\frac{1}{4}} N^{-\frac{1}{4}+\frac{\delta}{4}},$$

where we used again the rigidity (5.184). In summary, we have

$$|A_1| \lesssim L^{\frac{1}{4}} N^{-\frac{1}{4}+\frac{\delta}{4}} \log N. \tag{5.202}$$

In particular (5.199)–(5.202) imply that

$$Q := \sup_{0 \leq t \leq t_*} \sup_{|i| \lesssim LN^{\delta}} |Q_i(t)| \lesssim L^{\frac{1}{4}} N^{-\frac{1}{4}+\frac{\delta}{4}} \log N. \tag{5.203}$$

Collecting all the previous estimates using the choice of ν from (5.187) with $\delta' \geq \frac{\delta}{2}$ and that F is bounded up to $t \leq \tau$, we integrate (5.188)–(5.191) from 0 up to time $0 \leq t \leq t_*$ and conclude that

$$\begin{aligned} \sup_{0 \leq t \leq \tau} F(t) - F(0) &\lesssim \left(\frac{\nu^2 L N^{\frac{3\delta}{4}+\omega_1}}{N^{\frac{3}{2}}} + \frac{\nu L^{\frac{1}{4}} N^{\omega_1}}{N^{\frac{3}{4}}} + \frac{\nu Q N^{\omega_1}}{N^{\frac{1}{2}}} \right) \\ &\lesssim \frac{N^{\frac{3\delta}{4}+\omega_1}}{L^{\frac{1}{2}} N^{2\delta'}} + \frac{N^{\omega_1}}{L^{\frac{1}{2}} N^{\delta'}} + \frac{N^{\omega_1+\frac{\delta}{4}}}{L^{\frac{1}{2}} N^{\delta'}} \log N \leq 1 \end{aligned} \tag{5.204}$$

for large N and with very high probability, where we used the choice of ν (5.187) and that $\omega_1 \ll \omega_\ell$ in the last line. Since $F(0) = 1$, we get that $\tau = t_*$ with very high probability, and so

$$\sup_{0 \leq t \leq t_*} F(t) \leq 5, \tag{5.205}$$

with very high probability.

Furthermore, since $p = LN^{\delta}$, if $i \leq \frac{3}{4} LN^{\delta}$, choosing $\delta' = \frac{3\delta}{4} - \epsilon_1$, with $\epsilon_1 \leq \frac{\delta}{4}$, then by Proposition 5.6.2 we have that

$$\nu \psi(\widehat{z}_i(t) - \overline{\gamma}_p) = \nu \left| \widehat{z}_i(t) - \overline{\gamma}_p \right| \gtrsim \nu \frac{|i-p|}{N^{\frac{3}{4}} |p|^{\frac{1}{4}}} \gtrsim \frac{N^{\frac{3\delta}{4}}}{N^{\delta'}} = N^{\epsilon_1}$$

with very high probability.

Note that (5.205) implies

$$f_i(t) \leq 5e^{-\frac{1}{2}\nu\psi(\widehat{z}_i(t)-\overline{\gamma}_p)}.$$

Therefore, if $i \leq \frac{3LN^\delta}{4}$ and $p_* \geq p$, then for each fixed $0 \leq t \leq t_*$ we have that

$$\mathcal{U}^{\mathcal{L}}_{ip_*}(0,t) \leq N^{-D},$$

for any $D > 0$ with very high probability. Similar estimate holds if i and p_* are negative or have opposite sign. This proves the estimate on the first term in (5.185) for any fixed s. The estimate for $\mathcal{U}^{\mathcal{L}}_{p_*i}(s,t)$ is analogous with initial condition $f = \delta_i$. This proves Lemma 5.B.1.

□

Next, we enhance this result to a bound uniform in $0 \leq s \leq t_*$. We first have:

Lemma 5.B.2. *Let u be a solution of*

$$\partial_t u = \mathcal{L}u,$$

with non-negative initial condition $u_i(0) \geq 0$. Then, for each $0 \leq t \leq t_*$ we have

$$\frac{1}{2}\sum_i u_i(0) \leq \sum_i u_i(t) \leq \sum_i u_i(0) \tag{5.206}$$

with very high probability.

Proof. Since $\mathcal{U}^{\mathcal{L}}$ is a contraction semigroup the upper bound in (5.206) is trivial. Notice that $\partial_t \sum_i u_i = \sum_i \mathcal{V}_i u_i$. Thus the lower bound will follow once we prove $-\mathcal{V}_i \lesssim N^{\frac{1}{2}}L^{-\frac{1}{2}}$ with very high probability since $t_* N^{\frac{1}{2}}L^{-\frac{1}{2}}$ is much smaller than 1 by $\omega_1 \ll \omega_\ell$.

The estimate $-\mathcal{V}_i \lesssim N^{\frac{1}{2}}L^{-\frac{1}{2}}$ proceeds similarly to (5.201). Indeed, for $1 \leq |i| < N^{\omega_A}$, we use $\rho_{y,t}(E + \mathfrak{e}^+_{y,t}) \lesssim |E|^{\frac{1}{3}}$ and that $|\widehat{z}_i(t) - E| \sim |\overline{\gamma}_i(t) - E|$ by rigidity (5.184) and by the fact that

$$|j_+(i) - i|\,,\ |j_-(i) - i| \gtrsim N^{4\omega_\ell} + N^{\omega_\ell}\,|i|^{\frac{3}{4}}$$

is much bigger than the rigidity scale. Therefore, we have

$$\begin{aligned} -\mathcal{V}_i &= \int_{\mathcal{I}_{y,i}(t)^c} \frac{\rho_{r,t}(E + \mathfrak{e}^+_{r,t})}{(\widehat{z}_i(t) - E)^2}\,\mathrm{d}E \\ &\lesssim \int_{\mathcal{I}_{y,i}(t)^c} \frac{1}{|E - \overline{\gamma}_i(t)|^{\frac{5}{3}}}\,\mathrm{d}E + \int_{\mathcal{I}_{y,i}(t)^c} \frac{|\overline{\gamma}_i|^{\frac{1}{3}}}{(E - \overline{\gamma}_i(t))^2}\,\mathrm{d}E \lesssim \frac{N^{\frac{1}{2}}}{N^{2\omega_\ell}} = \frac{N^{\frac{1}{2}}}{L^{\frac{1}{2}}}. \end{aligned}$$

The estimate of $-\mathcal{V}_i$ for $N^{\omega_A} < |i| \leq \frac{i_*}{2}$ is similar. This concludes the proof of Lemma 5.B.2.

□

Finally we prove the following version of Lemma 5.B.1 that is uniform in s:

Lemma 5.B.3. *Under the same hypotheses of Lemma 5.B.1, for any $\delta' > 0$, such that $\delta' < C'\omega_\ell$, with $C' > 0$ the constant defined in Lemma 5.B.1, $|a| \leq \frac{LN^{\delta'}}{2}$ and $|b| \geq LN^{\delta'}$ we have that*

$$\sup_{0 \leq s \leq t \leq t_*} \mathcal{U}^{\mathcal{L}}_{ab}(s,t) + \mathcal{U}^{\mathcal{L}}_{ba}(s,t) \leq N^{-D} \tag{5.207}$$

with very high probability. The same result holds for $t_* \leq s \leq t \leq 2t_*$ as well.

Proof. By the semigroup property for any $0 \leq s \leq t \leq t_*$ and any j we have that

$$\mathcal{U}^{\mathcal{L}}_{aj}(0,t) \geq \mathcal{U}^{\mathcal{L}}_{ab}(s,t)\mathcal{U}^{\mathcal{L}}_{bj}(0,s). \tag{5.208}$$

Furthermore, by Lemma 5.B.2 for the dual dynamics we have that

$$\frac{1}{2}\sum_j u_j(0) \leq \sum_j u_j(s) = \sum_i \sum_j \left(\mathcal{U}^{\mathcal{L}}_{ji}(0,s)\right)^T u_i(0),$$

and so, by choosing $u(0) = \delta_b$ we conclude that

$$\sum_j \mathcal{U}^{\mathcal{L}}_{bj}(0,s) \geq \frac{1}{2}, \qquad \forall\, 0 \leq s \leq t_*.$$

From the last inequality and since $\sup_{s\leq t_*} \mathcal{U}^{\mathcal{L}}_{bj}(0,s) \leq N^{-100}$ with very high probability for any $|j| \leq \frac{3}{4}LN^{\delta'}$ by Lemma 5.B.1, it follows that there exists an $j_* = j_*(s)$, maybe depending on s, with $|j_*(s)| \geq \frac{3}{4}LN^{\delta'}$, such that $\mathcal{U}^{\mathcal{L}}_{bj_*(s)}(0,s) \geq \frac{1}{4N}$. Furthermore, by the finite speed propagation estimate in Lemma 5.B.1 (this time with $|a| \geq \frac{3}{4}LN^{\delta}$ and $|b| \leq \frac{1}{2}LN^{\delta}$; note that its proof only used that $|a-b| \gtrsim LN^{\delta}$), we have that

$$\sup_{t\leq t_*} \mathcal{U}^{\mathcal{L}}_{aj_*}(0,t) \leq N^{-D}, \qquad \forall\, |j_*| \geq \frac{3}{4}LN^{\delta'}$$

with very high probability. Hence we get from (5.208) with $j = j_*(s)$ that $\sup_{s\leq t} \mathcal{U}^{\mathcal{L}}_{ab}(s,t) \lesssim N^{-D+1}$ with very high probability. The estimate for $\mathcal{U}^{\mathcal{L}}_{ba}(s,t)$ follows in a similar way. This concludes the proof of Lemma 5.B.3. □

Finally, we prove Lemma 5.6.4 and Proposition 5.6.7 which are versions of Lemma 5.B.3 but for the short range approximation on scale $L = N^{1/2+C_1\omega_1}$ needed in Section 5.6.2.2.

Proof of Lemma 5.6.4. Choosing $L = N^{\frac{1}{2}+C_1\omega_1}$, the proof of Lemma 5.B.1 is exactly the same except for the estimate of Q in (5.203), since, for any $\alpha \in [0,1]$, $Q_i(t)$ from (5.78) is now defined as

$$Q_i(t) := \frac{\beta}{N}\sum_{j:|j-i|>L} \frac{1}{\overline{\gamma}^*_i - \overline{\gamma}^*_j} + \frac{1-\beta}{N}\sum_{j:|j-i|>L} \frac{1}{\widetilde{z}_i - \widetilde{z}_j}\,\mathrm{d}t + \Phi_\alpha(t), \tag{5.209}$$

with $\Phi_\alpha(t)$ given in (5.56) instead of (5.194). Then Lemma 5.B.2 and Lemma 5.B.3 follow exactly in the same way.

By (5.209) it easily follows that

$$Q := \sup_{0\leq t\leq t_*}\ \sup_{|i|\leq LN^{\delta'}} |Q_i(t)| \lesssim \log N. \tag{5.210}$$

Hence, by an estimate similar to (5.204), we conclude that

$$\begin{aligned}\sup_{0\leq t\leq \tau} F(t) - F(0) &\lesssim \left(\frac{\nu^2 LN^{\frac{3\delta}{4}+\omega_1}}{N^{\frac{3}{2}}} + \frac{\nu L^{\frac{1}{4}}N^{\omega_1}}{N^{\frac{3}{4}}} + \frac{\nu Q N^{\omega_1}}{N^{\frac{1}{2}}}\right) \\ &\lesssim \frac{N^{\frac{3\delta}{4}+\omega_1}}{L^{\frac{1}{2}}N^{\delta'}} + \frac{N^{\omega_1}}{L^{\frac{1}{2}}N^{\delta'}} + \frac{N^{\frac{3}{4}+\omega_1}}{L^{\frac{3}{4}}N^{\frac{1}{2}}N^{\delta'}}\log N \leq 1,\end{aligned}$$

with very high probability. Note that in the last inequality we used that $L = N^{\frac{1}{2}+C_1\omega_1}$. □

Proof of Proposition 5.6.7. This proof is almost identical to the previous one, except that $Q_i(t)$ is now defined from (5.87) as

$$Q_i(t) := \beta\left[\frac{1}{N}\sum_{j:|j-i|>L}\frac{1}{\overline{\gamma}_i^* - \overline{\gamma}_j^*} + \Phi(t)\right] + (1-\beta)\left[\frac{\mathrm{d}}{\mathrm{d}t}\overline{\gamma}_i^*(t) - \frac{1}{N}\sum_{j:|j-i|\le L}\frac{1}{\overline{\gamma}_i^* - \overline{\gamma}_j^*}\right],$$

which satisfies the same bound (5.210). The rest of the proof is unchanged. □

5.C Short-long approximation

In this section we estimate the difference of the solution of the DBM $\widetilde{z}(t,\alpha)$ and its short range approximation $\widehat{z}(t,\alpha)$, closely following the proof of Lemma 3.7 in [122] and adapting it to the more complicated cusp situation. In particular, in Section 5.C.1 we estimate $|\widetilde{z}(t,\alpha) - \widehat{z}(t,\alpha)|$ for $0 \le t \le t_*$, i.e. until the formation of an exact cusp; in Section 5.C.2, instead, we estimate $|\widetilde{z}(t,\alpha) - \widehat{z}(t,\alpha)|$ for $t_* < t \le 2t_*$, i.e. after the formation of a small minimum. The precision of this approximation depends on the rigidity bounds we put as a condition. We consider a two-scale rigidity assumption, a weaker rigidity valid for all indices and a stronger rigidity valid for $1 \le |i| \lesssim i_* = N^{\frac{1}{2}+C_*\omega_1}$; both described by an exponent.

5.C.1 Short-long approximation: Small gap and exact cusp.

In this subsection we estimate the difference of the solution of the DBM $\widetilde{z}(t,\alpha)$ defined in (5.55) and its short range approximation $\widehat{z}(t,\alpha)$ defined by (5.120)-(5.123) for $0 \le t \le t_*$. We formulate Lemma 5.C.1 (for $0 \le t \le t_*$) below a bit more generally than we need in order to indicate the dependence of the approximation precision on these two exponents. For our actual application in Lemma 5.6.8 and Lemma 5.7.2 we use specific exponents.

Lemma 5.C.1. *Let* $\omega_1 \ll \omega_\ell \ll \omega_A \ll 1$. *Let* $0 < a_0 \le \frac{1}{4} + C\omega_1$, $C > 0$ *a universal constant and* $0 < a \le C\omega_1$. *Let* $i_* := N^{\frac{1}{2}+C_*\omega_1}$ *with* C_* *defined in Proposition 5.6.2. We assume that*

$$|\widetilde{z}_i(t,\alpha) - \overline{\gamma}_i(t)| \le \frac{N^{a_0}}{N^{\frac{3}{4}}}, \qquad 1 \le |i| \le N, \quad 0 \le t \le t_* \tag{5.211}$$

and that

$$|\widetilde{z}_i(t,\alpha) - \overline{\gamma}_i(t)| \le \frac{N^{a}}{N^{\frac{3}{4}}}, \qquad 1 \le |i| \le i_*, \quad 0 \le t \le t_*. \tag{5.212}$$

Then, for any $\alpha \in [0,1]$, *we have that*

$$\begin{aligned} &\sup_{1\le|i|\le N}\sup_{0\le t\le t_*}|\widehat{z}_i(t,\alpha) - \widetilde{z}_i(t,\alpha)| \qquad (5.213)\\ &\le \frac{N^a N^{C\omega_1}}{N^{\frac{3}{4}}}\left(\frac{1}{N^{2\omega_\ell}} + \frac{N^{\frac{\omega_A}{2}}\log N}{N^{\frac{1}{6}}N^a} + \frac{N^{\frac{\omega_A}{2}}\log N}{N^{\frac{1}{4}}N^a} + \frac{N^{\frac{\omega_A}{8}}}{N^{\frac{a}{2}}i_*^{\frac{1}{4}}} + \frac{N^{a_0}}{N^a i_*^{\frac{1}{2}}} + \frac{1}{N^{\frac{1}{18}}N^a}\right), \end{aligned}$$

with very high probability.

Proof of Lemma 5.6.8. Use Lemma 5.C.1 with the choice $a_0 = \frac{1}{4} + C\omega_1$ and $a = C\omega_1$, for some universal constant $C > 0$. The conditions (5.211) and (5.212) are guaranteed by (5.60) and (5.61). □

Proof of Lemma 5.C.1. Let $w_i := \widehat{z}_i - \widetilde{z}_i$, hence w is a solution of

$$\partial_t w = \mathcal{B}_1 w + \mathcal{V}_1 w + \zeta, \tag{5.214}$$

where the operator $\mathcal{B}_1$ is defined for any $f \in \mathbb{C}^{2N}$ by

$$(B_1 f)_i = \frac{1}{N} \sum_j^{\mathcal{A},(i)} \frac{f_j - f_i}{(\widetilde{z}_i(t,\alpha) - \widetilde{z}_j(t,\alpha))(\widehat{z}_i(t,\alpha) - \widehat{z}_j(t,\alpha))}. \tag{5.215}$$

The diagonal operator $\mathcal{V}_1$ is defined by $(\mathcal{V}_1 f)_i = \mathcal{V}_1(i) f_i$, where

$$\mathcal{V}_1(i) := -\int_{\mathcal{I}_{y,i}(t)^c} \frac{\rho_{y,t}(E + \mathfrak{e}_{y,t}^+)}{(\widetilde{z}_i(t,\alpha) - E)(\widehat{z}_i(t,\alpha) - E)} \mathrm{d}E, \quad \text{for} \quad 0 < |i| \le N^{\omega_A},$$

and

$$\mathcal{V}_1(i) := -\int_{\mathcal{I}_{z,i}(t)^c \cap \mathcal{J}_z(t)} \frac{\overline{\rho}_t(E + \overline{\mathfrak{e}}_t^+)}{(\widetilde{z}_i(t,\alpha) - E)(\widehat{z}_i(t,\alpha) - E)} \mathrm{d}E, \quad \text{for} \quad N^{\omega_A} < |i| \le \frac{i_*}{2}. \tag{5.216}$$

Finally, $\mathcal{V}_1(i) = 0$ for $|i| \ge \frac{i_*}{2}$. The vector ζ in (5.214) collects various error terms.

We define the stopping time

$$T := \max\left\{ t \in [0, t_*] \;\middle|\; \sup_{s,\alpha} |\widetilde{z}_i(s,\alpha) - \widehat{z}_i(s,\alpha)| \le \frac{\min\{|\mathcal{I}_{z,i}(t)|, |\mathcal{I}_{y,i}(t)|\}}{2} \right\}, \tag{5.217}$$

where $\sup_{s,\alpha} = \sup_{s\in[0,t]} \sup \alpha \in [0,1]$, where we recall that $|\mathcal{I}_{z,i}(t)| \sim |\mathcal{I}_{y,i}(t)| \sim N^{-\frac{3}{4}+3\omega_\ell}$.

For $0 \le t \le T$ we have that $\mathcal{V}_1 \le 0$. Therefore, since $\sum_i (\mathcal{B}f)_i = 0$, by the symmetry of $\mathcal{A}$, the semigroup of $\mathcal{B}_1 + \mathcal{V}_1$, denoted by $\mathcal{U}^{\mathcal{B}_1+\mathcal{V}_1}$, is a contraction on every ℓ^p space. Hence, since $w(0) = 0$ by (5.123), we have that

$$w(t) = \int_0^t \mathcal{U}^{\mathcal{B}_1+\mathcal{V}_1}(s,t)\zeta(s)\, \mathrm{d}s$$

and so

$$\|w(t)\|_\infty \le t \sup_{0\le s\le t} \|\zeta(s)\|_\infty \le N^{-\frac{1}{2}+\omega_1} \sup_{0\le s\le t} \|\zeta(s)\|_\infty. \tag{5.218}$$

Thus, to prove (5.213) it is enough to estimate $\|\zeta(s)\|_\infty$, for all $0 \le s \le t_*$.

The error term ζ is given by $\zeta_i = 0$ for $|i| > \frac{i_*}{2}$, then for $1 \le |i| \le N^{\omega_A}$, ζ_i is defined as

$$\zeta_i = \int_{\mathcal{I}_{y,i}(t)^c} \frac{\rho_{y,t}(E + \mathfrak{e}_{y,t}^+)}{\widetilde{z}_i(t,\alpha) - E} \mathrm{d}E - \frac{1}{N} \sum_j^{\mathcal{A}^c,(i)} \frac{1}{\widetilde{z}_i(t,\alpha) - \widetilde{z}_j(t,\alpha)} + \Phi_\alpha(t) - \Re[m_{y,t}(\mathfrak{e}_{y,t}^+)], \tag{5.219}$$

with $\Phi_\alpha(t)$ defined in (5.56), and for $N^{\omega_A} < |i| \le \frac{i_*}{2}$ as

$$\zeta_i = \int_{\mathcal{I}_{z,i}(t)^c \cap \mathcal{J}_z(t)} \frac{\overline{\rho}_t(E + \overline{\mathfrak{e}}_t^+)}{\widetilde{z}_i(t,\alpha) - E} \mathrm{d}E - \frac{1}{N} \sum_{1 \le |j| < \frac{3i_*}{4}}^{\mathcal{A}^c,(i)} \frac{1}{\widetilde{z}_i(t,\alpha) - \widetilde{z}_j(t,\alpha)}. \tag{5.220}$$

Note that in the sum in (5.220) we do not have the summation over $|j| \geq \frac{3i_*}{4}$ since if $1 \leq |i| \leq \frac{i_*}{2}$ and $|j| \geq \frac{3i_*}{4}$ then $(i, j) \in \mathcal{A}^c$.

In the following we will often omit the t and the α arguments from $\widetilde{z}_i$ and $\overline{\gamma}_i$ for notational simplicity.

First, we consider the error term (5.220) for $N^{\omega_A} < |i| \leq \frac{i_*}{2}$. We start with the estimate

$$\begin{aligned} |\zeta_i| &= \left| \int_{\mathcal{I}^c_{z,i}(t)\cap\mathcal{J}_z(t)} \frac{\overline{\rho}_t(E+\overline{\mathfrak{e}}_t^+)}{\widetilde{z}_i - E}\, \mathrm{d}E - \frac{1}{N} \sum_{1\leq|j|<\frac{3i_*}{4}}^{\mathcal{A}^c,(i)} \frac{1}{\widetilde{z}_i - \widetilde{z}_j} \right| \qquad (5.221) \\ &\lesssim \left| \sum_{1\leq|j|<\frac{3i_*}{4}}^{\mathcal{A}^c,(i)} \int_{\overline{\gamma}_j}^{\overline{\gamma}_{j+1}} \frac{\overline{\rho}_t(E+\overline{\mathfrak{e}}_t^+)(E-\overline{\gamma}_j)}{(\widetilde{z}_i - E)(\widetilde{z}_i - \overline{\gamma}_j)}\, \mathrm{d}E \right| + \left| \frac{1}{N} \sum_{1\leq|j|<\frac{3i_*}{4}}^{\mathcal{A}^c,(i)} \frac{\widetilde{z}_j - \overline{\gamma}_j}{(\widetilde{z}_i - \widetilde{z}_j)(\widetilde{z}_i - \overline{\gamma}_j)} \right| \\ &\quad + \left| \int_{\overline{\gamma}_{j_+}}^{\overline{\gamma}_{j_+ +1}} \frac{\overline{\rho}_t(E+\overline{\mathfrak{e}}_t^+)}{\widetilde{z}_i - E}\, \mathrm{d}E \right| + \left| \int_{\overline{\gamma}_{-\frac{3i_*}{4}}}^{\overline{\gamma}_{-\frac{3i_*}{4}+1}} \frac{\overline{\rho}_t(E+\overline{\mathfrak{e}}_t^+)}{\widetilde{z}_i - E}\, \mathrm{d}E \right| + \left| \int_0^{\overline{\gamma}_1} \frac{\overline{\rho}_t(E+\overline{\mathfrak{e}}_t^+)}{\widetilde{z}_i - E}\, \mathrm{d}E \right|. \end{aligned}$$

Since $|j_+ - i| \geq N^{4\omega_\ell} + N^{\omega_\ell} |i|^{\frac{3}{4}}$ and N^{ω_A}, i.e.

$$\left| \overline{\gamma}_{j_+} - \overline{\gamma}_i \right| \geq \frac{N^{\omega_\ell} |i|^{\frac{1}{2}}}{N^{\frac{3}{4}}}$$

is bigger than the rigidity scale (5.212), all terms in the last line of (5.221) are bounded by $N^{-\frac{1}{4}-3\omega_\ell}$.

Then, using the rigidity estimate in (5.212) for the first and the second term of the rhs. of (5.221), we conclude that

$$|\zeta_i| \lesssim \frac{N^a}{N^{\frac{7}{4}}} \sum_{1\leq|j|<\frac{3i_*}{4}}^{\mathcal{A}^c,(i)} \frac{1}{(\overline{\gamma}_i - \overline{\gamma}_j)^2} + N^{-\frac{1}{4}-3\omega_\ell}. \qquad (5.222)$$

The sum on the rhs. of (5.222) is over all the j, negative and positive, but the main contribution comes from i and j with the same sign, because if i and j have opposite sign then

$$\frac{1}{(\overline{\gamma}_i - \overline{\gamma}_j)^2} \leq \frac{1}{(\overline{\gamma}_{-i} - \overline{\gamma}_j)^2}.$$

Hence, assuming that i is positive (for negative i's we proceed exactly in the same way), we conclude that

$$|\zeta_i| \lesssim \frac{N^a}{N^{\frac{7}{4}}} \sum_{1\leq j<\frac{3i_*}{4}}^{\mathcal{A}^c,(i)} \frac{1}{(\overline{\gamma}_i - \overline{\gamma}_j)^2} + N^{-\frac{1}{4}-3\omega_\ell}. \qquad (5.223)$$

From now we assume that both i and j are positive. In order to estimate (5.223) we use the explicit expression of the quantiles from (5.26a), i.e.

$$\overline{\gamma}_j \sim \max\left\{ \left(\frac{j}{N}\right)^{2/3} \overline{\Delta}_t^{\frac{1}{9}}, \left(\frac{j}{N}\right)^{3/4} \right\},$$

where $\overline{\Delta}_t \lesssim t_*^{3/2}$ denotes the length of the small gap of $\overline{\rho}_t$, for all $|j| \leq i_* \sim N^{\frac{1}{2}}$. A simple calculation from (5.26a) shows that in the regime $i \geq N^{\omega_A}$ and $j \in \mathcal{A}^c$ we may replace

$\left|\overline{\gamma}_i - \overline{\gamma}_j\right| \sim |\gamma_{y,i}(t) - \gamma_{y,j}(t)| \sim \left|i^{3/4} - j^{3/4}\right| / N^{3/4}$, hence

$$|\zeta_i| \lesssim \frac{N^a}{N^{\frac{1}{4}}} \sum_{1\le j<\frac{3i_*}{4}}^{\mathcal{A}^c,(i)} \frac{i^{\frac{1}{2}} + j^{\frac{1}{2}}}{(i-j)^2} + N^{-\frac{1}{4}-3\omega_\ell}. \tag{5.224}$$

In fact, the same replacement works if either $i \geq N^{4\omega_\ell}$ or $j \geq N^{4\omega_\ell}$ and at least one of these two inequalities always hold as $(i,j) \in \mathcal{A}^c$. Using $i \leq \frac{i_*}{2}$ and that by the restriction $(i,j) \in \mathcal{A}^c$ we have $|j-i| \geq \ell(\ell^3 + i^{\frac{3}{4}})$, elementary calculation gives

$$|\zeta_i| \lesssim \frac{N^a}{N^{\frac{1}{4}} N^{2\omega_\ell}}. \tag{5.225}$$

Since analogous computations hold for i and j both negative, we have

$$|\zeta_i| \lesssim \frac{N^a}{N^{\frac{1}{4}} N^{2\omega_\ell}}, \qquad \text{for any} \quad N^{\omega_A} < |i| \leq \frac{i_*}{2}. \tag{5.226}$$

with very high probability.

Next, we proceed with the bound for ζ_i for $|i| \leq N^{\omega_A}$. From (5.219) we have

$$\begin{aligned}
\zeta_i = &\left(\int_{\mathcal{I}_{z,i}(t)^c \cap \mathcal{J}_z(t)} \frac{\overline{\rho}_t(E + \overline{\mathfrak{e}}_t^+)}{\widetilde{z}_i - E} \,\mathrm{d}E - \frac{1}{N} \sum_{|j|<\frac{3i_*}{4}}^{\mathcal{A}^c,(i)} \frac{1}{\widetilde{z}_i - \widetilde{z}_j} \right) \\
&+ \left(\int_{\mathcal{J}_z(t)^c} \frac{\overline{\rho}_t(E + \overline{\mathfrak{e}}_t^+)}{\widetilde{z}_i - E} \,\mathrm{d}E - \frac{1}{N} \sum_{|j|\geq\frac{3i_*}{4}}^{\mathcal{A}^c,(i)} \frac{1}{\widetilde{z}_i - \widetilde{z}_j} \right) \\
&+ \Phi_\alpha(t) - \Re[\overline{m}_t(\widetilde{z}_i + \overline{\mathfrak{e}}_t^+)] + \Re[m_{y,t}(\widetilde{z}_i + \mathfrak{e}_{y,t}^+)] - \Re[m_{y,t}(\mathfrak{e}_{y,t}^+)] \\
&+ \left(\int_{\mathcal{I}_{z,i}(t)} \frac{\overline{\rho}_t(E + \overline{\mathfrak{e}}_t^+)}{\widetilde{z}_i - E} \,\mathrm{d}E - \int_{\mathcal{I}_{y,i}(t)} \frac{\rho_{y,t}(E + \mathfrak{e}_{y,t}^+)}{\widetilde{z}_i - E} \,\mathrm{d}E \right) =: A_1 + A_2 + A_3 + A_4.
\end{aligned} \tag{5.227}$$

By the remark after (5.224), the estimate of A_1 proceeds as in (5.224) and so we conclude that

$$|A_1| \lesssim \frac{N^a}{N^{\frac{1}{4}} N^{2\omega_\ell}}. \tag{5.228}$$

To estimate A_2, we first notice that the restriction $(i,j) \in \mathcal{A}^c$ in the summation is superfluous for $|i| \leq N^{\omega_A}$ and $|j| \geq \frac{3}{4} i_*$. Let $\eta_1 \in [N^{-\frac{3}{4}+\frac{1}{8}\omega_A}, N^{-\frac{3}{4}+\frac{3}{4}\omega_A}]$ be an auxiliary scale we will determine later in the proof, then we write A_2 as follows:

$$\begin{aligned}
A_2 = &\left(\int_{\mathcal{J}_z(t)^c} \frac{\overline{\rho}_t(E + \overline{\mathfrak{e}}_t^+)}{\widetilde{z}_i - E} \,\mathrm{d}E - \int_{\mathcal{J}_z(t)^c} \frac{\overline{\rho}_t(E + \overline{\mathfrak{e}}_t^+)}{\widetilde{z}_i - E + \mathrm{i}\eta_1} \,\mathrm{d}E \right) \\
&+ \left(\frac{1}{N} \sum_{|j|\geq\frac{3i_*}{4}} \frac{1}{\widetilde{z}_i - \widetilde{z}_j + \mathrm{i}\eta_1} - \frac{1}{N} \sum_{|j|\geq\frac{3i_*}{4}} \frac{1}{\widetilde{z}_i - \widetilde{z}_j} \right) \\
&+ \left(\frac{1}{N} \sum_{|j|<\frac{3i_*}{4}} \frac{1}{\widetilde{z}_i - \widetilde{z}_j + \mathrm{i}\eta_1} - \int_{\mathcal{J}_z(t)} \frac{\overline{\rho}_t(E + \overline{\mathfrak{e}}_t^+)}{\widetilde{z}_i - E + \mathrm{i}\eta_1} \,\mathrm{d}E \right) \\
&+ (\overline{m}_t(\widetilde{z}_i + \mathrm{i}\eta_1) - m_{2N}(\widetilde{z}_i + \mathrm{i}\eta_1, t, \alpha)) =: A_{2,1} + A_{2,2} + A_{2,3} + A_{2,4},
\end{aligned} \tag{5.229}$$

where we introduced

$$m_{2N}(z,t,\alpha) := \frac{1}{N}\sum_{|j|\le N}\frac{1}{z_j(t,\alpha)-z}, \qquad z\in\mathbb{H}.$$

For $1\le|i|\le N^{\omega_A}$ and $|j|>\frac{3i_*}{4}$, the term $A_{2,2}$ is bounded by the crude rigidity (5.211) as

$$|A_{2,2}|\le\frac{1}{N}\sum_{|j|>\frac{3i_*}{4}}\frac{\eta_1}{(\widetilde{z}_i-\widetilde{z}_j)^2}\lesssim\frac{N^{\frac{1}{2}}\eta_1}{i_*^{\frac{1}{2}}}. \tag{5.230}$$

Exactly the same estimate holds for $A_{2,1}$.

Next, using the rigidity estimates in (5.211) and (5.212) we conclude that

$$\begin{aligned}|A_{2,4}|&\lesssim\frac{1}{N}\sum_{1\le|j|\le i_*}\frac{\left|\widetilde{z}_j-\overline{\gamma}_j\right|}{|\widetilde{z}_i-\widetilde{z}_j+\mathrm{i}\eta_1|^2}+\frac{1}{N}\sum_{i_*\le|j|\le N}\frac{\left|\widetilde{z}_j-\overline{\gamma}_j\right|}{|\widetilde{z}_i-\widetilde{z}_j+\mathrm{i}\eta_1|^2}\\&\lesssim\frac{N^a}{N^{\frac{3}{4}}\eta_1}\Im m_N(\overline{\gamma}_i+\mathrm{i}\eta_1)+\frac{N^{a_0}}{N^{\frac{7}{4}}}\sum_{i_*\le|j|\le N}\frac{1}{(\overline{\gamma}_i-\overline{\gamma}_j)^2}\\&\lesssim\frac{N^a}{N^{\frac{3}{4}}\eta_1}\left(\frac{N^{\frac{3\omega_A}{4}}}{N^{\frac{3}{4}}}+\eta_1\right)^{\frac{1}{3}}+\frac{N^{a_0}}{N^{\frac{1}{4}}i_*^{\frac{1}{2}}}\lesssim\frac{N^aN^{\frac{\omega_A}{4}}}{N\eta_1}+\frac{N^{a_0}}{i_*^{\frac{1}{2}}N^{\frac{1}{4}}}.\end{aligned}\tag{5.231}$$

Here we used that the rigidity scale near i for $1\,|i|\le N^{\omega_A}$ is much smaller than $\eta_1\ge N^{-\frac{3}{4}+\frac{1}{8}\omega_A}$. In particular, we know that $\Im m_N(\overline{\gamma}_i+\mathrm{i}\eta_1)$ can be bounded by the density $\overline{\rho}_t(\overline{\gamma}_i+\eta_1)$ which in turn is bounded by $(\overline{\gamma}_i+\eta_1)^{1/3}$. Similarly we conclude that

$$|A_{2,3}|\le\frac{N^aN^{\frac{\omega_A}{4}}}{N\eta_1}.$$

Optimizing (5.230) and (5.231) for η_1, we choose

$$\eta_1=\left(\frac{i_*^{\frac{1}{2}}N^aN^{\frac{\omega_A}{4}}}{N^{\frac{3}{2}}}\right)^{\frac{1}{2}}$$

which falls into the required interval for η_1. Collecting all estimates for the parts of A_2 in (5.229), we therefore conclude that

$$|A_2|\le\frac{N^{\frac{a}{2}}N^{\frac{\omega_A}{8}}}{i_*^{\frac{1}{4}}N^{\frac{1}{4}}}+\frac{N^{a_0}}{i_*^{\frac{1}{2}}N^{\frac{1}{4}}}. \tag{5.232}$$

Next, we treat A_3 from (5.227). $\Phi_\alpha(t)=\Re[\overline{m}_t(\overline{\mathfrak{e}}_t^+)]+\mathcal{O}\left(N^{-1}\right)$ by (5.57), then by (5.32a) we conclude that

$$\begin{aligned}|A_3|&=\left|\Re[\overline{m}_t(\overline{\mathfrak{e}}_t^+)]-\Re[\overline{m}_t(\widetilde{z}_i+\overline{\mathfrak{e}}_t^+)]+\Re[m_{y,t}(\widetilde{z}_i+\mathfrak{e}_{y,t}^+)]-\Re[m_{y,t}(\mathfrak{e}_{y,t}^+)]\right|\\&\lesssim\left(\frac{|i|^{\frac{1}{4}}N^{\frac{7\omega_1}{18}}}{N^{\frac{1}{4}}N^{\frac{1}{6}}}+\frac{|i|^{\frac{1}{2}}}{N^{\frac{1}{2}}}\right)|\log|\overline{\gamma}_i||\lesssim\frac{N^{\frac{\omega_A}{4}}N^{\frac{7\omega_1}{18}}\log N}{N^{\frac{1}{4}}N^{\frac{1}{6}}}+\frac{N^{\frac{\omega_A}{2}}\log N}{N^{\frac{1}{2}}}.\end{aligned}\tag{5.233}$$

We proceed writing A_4 as

$$\begin{aligned}A_4 = &\left(\int_{\mathcal{I}_{z,i}(t)} \frac{\overline{\rho}_t(E+\overline{\mathfrak{e}}_t^+)}{\widetilde{z}_i - E}\,\mathrm{d}E - \int_{\mathcal{I}_{z,i}(t)} \frac{\rho_{y,t}(E+\mathfrak{e}_{y,t}^+)}{\widetilde{z}_i - E}\,\mathrm{d}E\right)\\ &+\left(\int_{\mathcal{I}_{z,i}(t)} \frac{\rho_{y,t}(E+\mathfrak{e}_{y,t}^+)}{\widetilde{z}_i - E}\,\mathrm{d}E - \int_{\mathcal{I}_{y,i}(t)} \frac{\rho_{y,t}(E+\mathfrak{e}_{y,t}^+)}{\widetilde{z}_i - E}\,\mathrm{d}E\right) =: A_{4,1} + A_{4,2}.\end{aligned}$$

We start with the estimate for $A_{4,2}$. By (5.119) and the comparison estimates between $\overline{\gamma}_{z,i}$ and $\widehat{\gamma}_{y,i}$ by (5.28) we have that

$$\begin{aligned}|\mathcal{I}_{z,i}(t)\Delta\mathcal{I}_{y,i}(t)| &\lesssim \left|\overline{\gamma}_{z,i-j_-(i)} - \widehat{\gamma}_{y,i-j_-(i)}\right| + \left|\overline{\gamma}_{z,i+j_+(i)} - \widehat{\gamma}_{y,i+j_+(i)}\right|\\ &\lesssim \frac{N^{\frac{\omega_1}{2}}(\ell^3 + |i|^{\frac{3}{4}})}{N^{\frac{11}{12}}},\end{aligned} \tag{5.234}$$

where Δ is the symmetric difference. In the second inequality of (5.234) we used that $|i \pm j_\pm(i)| \lesssim N^{\omega_A}$ and $\omega_A \ll 1$. For $E \in \mathcal{I}_{z,i}\Delta\mathcal{I}_{y,i}$ we have that

$$\left|\frac{\rho_{y,t}(E+\mathfrak{e}_{y,t}^+)}{\widetilde{z}_i - E}\right| \lesssim \frac{N^{\frac{1}{2}}(\ell^2 + |i|^{\frac{1}{2}})}{\ell^3 + |i|^{\frac{3}{4}}},$$

and so, using $|i| \le N^{\omega_A}$,

$$|A_{4,2}| \lesssim \frac{N^{\frac{\omega_1}{2}} N^{\frac{\omega_A}{2}}}{N^{\frac{5}{12}}} = \frac{N^{\frac{\omega_1}{2}} N^{\frac{\omega_A}{2}}}{N^{\frac{1}{4}} N^{\frac{1}{6}}} \tag{5.235}$$

with very high probability.

To estimate the integral in $A_{4,1}$ we have to deal with the logarithmic singularity due to the values of E close to $\widetilde{z}_i(t)$. For $\max\{\overline{\mathfrak{e}}_t^-, \mathfrak{e}_{y,t}^-\} < E \le 0$ we have that

$$\rho_{y,t}(E+\mathfrak{e}_{y,t}^+) = \overline{\rho}_t(E+\overline{\mathfrak{e}}_t^+) = 0. \tag{5.236}$$

For $\min\{\overline{\mathfrak{e}}_t^-, \mathfrak{e}_{y,t}^-\} \le E \le \max\{\overline{\mathfrak{e}}_t^-, \mathfrak{e}_{y,t}^-\}$, using the $\frac{1}{3}$-Hölder continuity of $\overline{\rho}_t$ and $\rho_{y,t}$ and (5.17a) we have that

$$\left|\rho_{y,t}(E+\mathfrak{e}_{y,t}^+) - \overline{\rho}_t(E+\overline{\mathfrak{e}}_t^+)\right| \lesssim \Delta_{y,t}^{\frac{1}{3}}(t_* - t)^{\frac{1}{9}} \lesssim \frac{N^{\frac{11\omega_1}{18}}}{N^{\frac{11}{36}}},$$

for all $0 \le t \le t_*$. In the last inequality we used that $\Delta_{y,t} \le \Delta_{y,0} \lesssim N^{-\frac{3}{4}+\frac{3\omega_1}{2}}$ for all $t \le t_*$. Similarly, for $E \le \min\{\overline{\mathfrak{e}}_t^-, \mathfrak{e}_{y,t}^-\}$ we have that

$$\left|\rho_{y,t}(E+\mathfrak{e}_{y,t}^+) - \overline{\rho}_t(E+\overline{\mathfrak{e}}_t^+)\right| \lesssim \left|\rho_{y,t}(E'+\mathfrak{e}_{y,t}^-) - \overline{\rho}_t(E'+\overline{\mathfrak{e}}_t^-)\right| + \Delta_{y,t}^{\frac{1}{3}}(t_*-t)^{\frac{1}{9}}, \tag{5.237}$$

with $E' \le 0$.

Using (5.17b) for $E \ge 0$ and combining (5.17b) with (5.236)-(5.237) for $E < 0$, we have that

$$\begin{aligned}|A_{4,1}| \lesssim &\left(\frac{(\ell + |i|^{\frac{1}{4}})N^{\frac{\omega_1}{3}}}{N^{\frac{1}{4}}N^{\frac{1}{6}}} + \frac{(\ell^2 + |i|^{\frac{1}{2}})}{N^{\frac{1}{2}}} + \frac{N^{\frac{11\omega_1}{18}}}{N^{\frac{11}{36}}}\right)\int_{\mathcal{I}_{,i}(t)\cap\{|E-\widetilde{z}_i|>N^{-60}\}} \frac{1}{|\widetilde{z}_i - E|}\,\mathrm{d}E\\ &+\left|\int_{|E-\widetilde{z}_i|\le N^{-60}} \frac{\overline{\rho}_t(E+\overline{\mathfrak{e}}_t^+) - \rho_{y,t}(E+\mathfrak{e}_{y,t}^+)}{\widetilde{z}_i - E}\,\mathrm{d}E\right|.\end{aligned} \tag{5.238}$$

The two singular integrals in the second line are estimated separately. By the $\frac{1}{3}$-Hölder continuity $\rho_{y,t}$ we conclude that

$$\left|\int_{\left|E-\widetilde{z}_i\right|\leq N^{-60}} \frac{\rho_{y,t}(E+\mathfrak{e}_{y,t}^{+})}{\widetilde{z}_i-E}\,\mathrm{d}E\right| = \left|\int_{\left|E-\widetilde{z}_i\right|\leq N^{-60}} \frac{\rho_{y,t}(E+\mathfrak{e}_{y,t}^{+})-\rho_{y,t}(\widetilde{z}_i+\mathfrak{e}_{y,t}^{+})}{\widetilde{z}_i-E}\,\mathrm{d}E\right|$$
$$\lesssim \int_{\left|E-\widetilde{z}_i\right|\leq N^{-60}} \frac{1}{|\widetilde{z}_i-E|^{\frac{2}{3}}}\,\mathrm{d}E \lesssim N^{-20}.$$

The same bound holds for the other singular integral in (5.238) by using the $\frac{1}{3}$-Hölder continuity of $\overline{\rho}_t$. Hence, for $1\leq |i|\leq N^{\omega_A}$, by (5.238) we have that

$$|A_{4,1}|\leq \frac{N^{\frac{\omega_A}{4}}N^{\frac{\omega_1}{3}}\log N}{N^{\frac{1}{4}}N^{\frac{1}{6}}}+\frac{N^{\frac{\omega_A}{2}}\log N}{N^{\frac{1}{2}}}+\frac{N^{\frac{11\omega_1}{18}}\log N}{N^{\frac{11}{36}}}, \tag{5.239}$$

with very high probability.

Collecting all the estimates (5.226), (5.228), (5.232), (5.233), (5.235) and (5.239), and recalling $\omega_1\ll\omega_\ell\ll\omega_A\ll 1$, we see that (5.228) is the largest term and thus $|\zeta|\lesssim N^{-\frac{1}{4}-2\omega_\ell}N^{C\omega_1}$ as $a\leq C\omega_1$. Thus, using (5.218), we conclude that the estimate in (5.213) is satisfied for all $0\leq t\leq T$. In particular, this means that

$$|\widehat{z}_i(t,\alpha)-\widetilde{z}_i(t,\alpha)|\leq N^{-\frac{3}{4}+C\omega_1},\qquad 0\leq t\leq T,$$

for some small constant $C>0$. We conclude the proof of this lemma showing that $T\geq t_*$.

Suppose by contradiction that $T<t_*$, then, since the solution of the DBM have continuous paths (see Theorem 12.2 of [78]), we have that

$$|\widehat{z}_i(T+\tilde{t},\alpha)-\widetilde{z}_i(T+\tilde{t},\alpha)|\leq \frac{N^aN^{c\omega_1}}{N^{\frac{3}{4}}N^{2\omega_\ell}},$$

for some tiny $\tilde{t}>0$ and for any $\alpha\in[0,1]$. This bound is much smaller than the threshold $|\mathcal{I}_{y,i}(t)|,|\mathcal{I}_{z,i}(t)|\sim N^{-\frac{3}{4}+3\omega_\ell}$ in the definition of T. But this is a contradiction by the maximality in the definition of T, hence $T=t_*$, proving (5.213) for all $0\leq t\leq t_*$. This completes the proof of Lemma 5.C.1. □

Proof of Lemma 5.7.2. The proof of this lemma is very similar to that of Lemma 5.C.1, hence we will only sketch the proof by indicating the differences. The main difference is that in this lemma we have optimal i-dependent rigidity for all $1\leq|i|\leq i_*$. Hence, we can give a better estimate on the first two terms in (5.221) as follows (recall that $N^{\omega_A}\leq i\leq\frac{i_*}{2}$)

$$|\zeta_i|\lesssim \frac{N^\xi N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}}\sum_{|j|<\frac{3i_*}{4}}\frac{1}{(\overline{\gamma}_i-\overline{\gamma}_j)^2|j|^{\frac{1}{4}}}\lesssim\frac{N^\xi N^{\frac{\omega_1}{6}}}{N^{\frac{3}{4}}}\sum_{|j|<\frac{3i_*}{4}}\frac{|i|^{\frac{1}{2}}+|j|^{\frac{1}{2}}}{(|i|-|j|)^2|j|^{\frac{1}{4}}}\lesssim\frac{N^\xi N^{\frac{\omega_1}{6}}}{N^{\frac{1}{4}}N^{3\omega_\ell}}.$$

Compared with (5.225), the additional N^{ω_ℓ} factor in the denominator comes from the $|j|^{1/4}$ factor beforehand that is due to the optimal dependence of the rigidity on the index. Consequently, using the optimal rigidity in (5.49), we improve the denominator in the first term on the rhs. of (5.213) from $N^{2\omega_\ell}$ to $N^{3\omega_\ell}$ with respect Lemma 5.C.1.

Furthermore, by (5.49),

$$|A_{2,3}|\,,|A_{2,4}| \le \frac{N^{\xi}}{N\eta_1}, \qquad \text{and} \qquad |A_{2,1}|\,,|A_{2,2}| \lesssim \frac{N^{\frac{1}{2}}\eta_1}{i_*^{\frac{1}{2}}} \lesssim N^{\frac{1}{4}-\frac{C_*\omega_1}{2}}\eta_1,$$

since $i_* = N^{\frac{1}{2}+C_*\omega_1}$, hence, choosing $\eta_1 = N^{-\frac{5}{8}}$, we conclude that

$$|A_1| + |A_2| \lesssim \frac{N^{\xi}N^{\frac{\omega_1}{6}}}{N^{\frac{1}{4}}N^{3\omega_\ell}} + \frac{N^{\xi}}{N^{\frac{3}{8}}}.$$

All other estimates follow exactly in the same way of the proof of Lemma 5.C.1. This concludes the proof of Lemma 5.7.2. □

5.C.2 Short-long approximation: Small minimum.

In this subsection we estimate the difference of the solution of the DBM $\widetilde{z}(t,\alpha)$ defined by (5.174) and its short range approximation $\widehat{z}(t,\alpha)$ defined by (5.176)-(5.177) for $t_* \le t \le 2t_*$.

Lemma 5.C.2. *Under the same assumption of Lemma 5.C.1 and assuming that the rigidity bounds* (5.211) *and* (5.212) *hold for the $\widetilde{z}(t,\alpha)$ dynamics* (5.174) *for all $t_* \le t \le 2t_*$, we conclude that*

$$\sup_{1\le|i|\le N}\sup_{t_*\le t\le 2t_*}|\widetilde{z}_i(t,\alpha)-\widehat{z}_i(t,\alpha)| \lesssim \frac{N^aN^{C\omega_1}}{N^{\frac{3}{4}}}\left(\frac{1}{N^{2\omega_\ell}}+\frac{N^{\frac{\omega_A}{8}}}{i_*^{\frac{1}{4}}N^{\frac{a}{2}}}+\frac{N^{a_0}}{N^a i_*^{\frac{1}{2}}}+\frac{1}{N^aN^{\frac{1}{24}}}\right), \tag{5.240}$$

with very high probability, for any $\alpha \in [0,1]$.

Proof. The proof of this lemma is similar to the proof of Lemma 5.C.1, but some estimates for the semicircular flow are slightly different mainly because in this lemma the $\widetilde{z}_i(t,\alpha)$ are shifted by $\overline{\mathfrak{m}}_t$ instead of $\overline{\mathfrak{e}}_t^+$. Hence, we will skip some details in this proof, describing carefully only the estimates that are different respect to Lemma 5.C.1.

Let $w_i := \widehat{z}_i - \widetilde{z}_i$, hence w is a solution of

$$\partial_t = \mathcal{B}_1 w + \mathcal{V}_1 w + \zeta,$$

where $\mathcal{B}_1$ and $\mathcal{V}_1$ are defined as in (5.215)-(5.216) substituting $\overline{\mathfrak{e}}_t^+$ with $\overline{\mathfrak{m}}_t$.

Without loss of generality we assume that $\mathcal{V}_1 \le 0$ for all $t_* \le t \le T$ (see (5.217) in the proof of Lemma 5.C.1 but now we have $t_* \le t \le 2t_*$ in the definition of the stopping time). This implies that $\mathcal{U}^{\mathcal{B}_1+\mathcal{V}_1}$ is a contraction semigroup and so in order to prove (5.240) it is enough to estimate

$$\sup_{t_*\le t\le T}\|\zeta(s)\|_\infty.$$

At the end, exactly as at the end of the proof of Lemma 5.C.1, by continuity of the paths, we can easily establish $T = 2t_*$ for the stopping time.

The error term ζ is given by $\zeta_i = 0$ for $|i| > \frac{i_*}{2}$, then ζ_i for $1 \le |i| \le N^{\omega_A}$ is defined as

$$\zeta_i = \int_{\mathcal{I}_{y,i}(t)^c}\frac{\rho_{y,t}(E+\widetilde{\mathfrak{m}}_{y,t})}{\widetilde{z}_i - E}\,\mathrm{d}E - \frac{1}{N}\sum_j^{\mathcal{A}^c,(i)}\frac{1}{\widetilde{z}_i-\widetilde{z}_j} + \Psi_\alpha(t) + \frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mathfrak{m}}_{y,t},$$

with $\Psi_\alpha(t)$ defined in (5.172), and for $N^{\omega_A} < |i| \leq \frac{i_*}{2}$ as

$$\zeta_i = \int_{\mathcal{I}_{z,i}(t)^c \cap \mathcal{J}_z(t)} \frac{\overline{\rho}_t(E + \overline{\mathfrak{m}}_t)}{\widetilde{z}_i - E} \, \mathrm{d}E - \frac{1}{N} \sum_{|j| < \frac{3i_*}{4}}^{\mathcal{A}^c,(i)} \frac{1}{\widetilde{z}_i - \widetilde{z}_j}.$$

We start to estimate the error term for $N^{\omega_A} < |i| \leq \frac{i_*}{2}$. A similar computation as the one leading to (5.226) in Lemma 5.C.1, using (5.212), we conclude that

$$|\zeta_i| = \left| \int_{\mathcal{I}_{i,z}^c(t) \cap \mathcal{J}_z(t)} \frac{\overline{\rho}_t(E + \overline{\mathfrak{m}}_t)}{\widetilde{z}_i - E} \, \mathrm{d}E - \frac{1}{N} \sum_{|j| < \frac{3i_*}{4}}^{\mathcal{A}^c,(i)} \frac{1}{\widetilde{z}_i - \widetilde{z}_j} \right| \lesssim \frac{N^a}{N^{\frac{1}{4}} N^{2\omega_\ell}}, \quad N^{\omega_A} < |i| \leq \frac{i_*}{2}.$$

Next, we proceed with the bound for ζ_i for $1 \leq |i| \leq N^{\omega_A}$. We rewrite ζ_i as

$$\begin{aligned} \zeta_i = &\left(\int_{\mathcal{I}_{i,z}^c(t)} \frac{\overline{\rho}_t(E + \overline{\mathfrak{m}}_t)}{\widetilde{z}_i - E} \, \mathrm{d}E - \frac{1}{N} \sum_j^{\mathcal{A}^c,(i)} \frac{1}{\widetilde{z}_i - \widetilde{z}_j} \right) \\ &+ \Re[m_{y,t}(\widetilde{z}_i + \widetilde{\mathfrak{m}}_{y,t})] + \frac{\mathrm{d}}{\mathrm{d}t} \widetilde{\mathfrak{m}}_{y,t} + \Psi_\alpha(t) - \Re[\overline{m}_t(\widetilde{z}_i + \overline{\mathfrak{m}}_t)] \\ &+ \left(\int_{\mathcal{I}_{z,i}(t)} \frac{\overline{\rho}_t(E + \overline{\mathfrak{m}}_t)}{\widetilde{z}_i - E} \, \mathrm{d}E - \int_{\mathcal{I}_{y,i}(t)} \frac{\rho_{y,t}(E + \widetilde{\mathfrak{m}}_{y,t})}{\widetilde{z}_i - E} \, \mathrm{d}E \right) =: (A_1 + A_2) + A_3 + A_4. \end{aligned} \tag{5.241}$$

where $(A_1 + A_2)$ indicates that for the actual estimates we split the first line in (5.241) into two terms as in (5.227). By similar computations as in Lemma 5.C.1, see (5.228) and (5.232), we conclude that

$$|A_1| + |A_2| \lesssim \frac{N^a}{N^{\frac{1}{4}} N^{2\omega_\ell}} + \frac{N^{\frac{a}{2}} N^{\frac{\omega_A}{8}}}{i_*^{\frac{1}{4}} N^{\frac{1}{4}}} + \frac{N^{a_0}}{i_*^{\frac{1}{2}} N^{\frac{1}{4}}}. \tag{5.242}$$

By (5.22b), (5.22d), (5.32b) and the definition of $\Psi_\alpha(t)$ in (5.172) it follows that

$$\begin{aligned} |A_3| &\lesssim |\Re[m_{y,t}(\widetilde{z}_i + \widetilde{\mathfrak{m}}_{y,t}) - m_{y,t}(\widetilde{\mathfrak{m}}_{y,t})] - \Re[\overline{m}_t(\overline{\mathfrak{m}}_t) - \overline{m}_t(\widetilde{z}_i + \overline{\mathfrak{m}}_t)]| + \frac{N^{\omega_1}}{N} \\ &\lesssim \left(\frac{N^{\frac{\omega_A}{4}} N^{\frac{\omega_1}{4}}}{N^{\frac{1}{4}} N^{\frac{1}{8}}} + \frac{N^{\frac{3\omega_1}{4}}}{N^{\frac{3}{8}}} + \frac{N^{\frac{\omega_A}{2}}}{N^{\frac{1}{2}}} \right) |\log |\widehat{\gamma}_i(t)|| + \frac{N^{\frac{7\omega_1}{12}}}{N^{\frac{7}{24}}} \lesssim \frac{N^{\frac{7\omega_1}{12}}}{N^{\frac{7}{24}}}. \end{aligned} \tag{5.243}$$

We proceed writing A_4 as

$$\begin{aligned} A_4 = &\left(\int_{\mathcal{I}_{z,i}(t)} \frac{\overline{\rho}_t(E + \overline{\mathfrak{m}}_t)}{\widetilde{z}_i - E} \, \mathrm{d}E - \int_{\mathcal{I}_{z,i}(t)} \frac{\rho_{y,t}(E + \widetilde{\mathfrak{m}}_{y,t})}{\widetilde{z}_i - E} \, \mathrm{d}E \right) \\ &+ \left(\int_{\mathcal{I}_{z,i}(t)} \frac{\rho_{y,t}(E + \widetilde{\mathfrak{m}}_{y,t})}{\widetilde{z}_i - E} \, \mathrm{d}E - \int_{\mathcal{I}_{y,i}(t)} \frac{\rho_{y,t}(E + \widetilde{\mathfrak{m}}_{y,t})}{\widetilde{z}_i - E} \, \mathrm{d}E \right) =: A_{4,1} + A_{4,2}. \end{aligned}$$

We start with the estimate for $A_{4,2}$. By (5.29) we have that

$$|\mathcal{I}_{z,i}(t) \Delta \mathcal{I}_{y,i}(t)| \lesssim \frac{N^\xi (\ell + |i|)}{N},$$

where Δ is the symmetric difference. Note that this bound is somewhat better than the analogous (5.234) due to the better bound in (5.29) compared with (5.28). For $E \in \mathcal{I}_{z,i}(t)\Delta\mathcal{I}_{y,i}(t)$ we have that

$$\left|\frac{\rho_{y,t}(E+\overline{\mathfrak{m}}_t)}{\widetilde{z}_i - E}\right| \lesssim \frac{N^{\frac{1}{2}}(\ell^2 + |i|^{\frac{1}{2}})}{\ell^3 + |i|^{\frac{3}{4}}},$$

and so

$$|A_{4,2}| \lesssim \frac{N^{\frac{3\omega_A}{4}}}{N^{\frac{1}{2}}} \tag{5.244}$$

with very high probability.

To estimate the integral in $A_{4,1}$, we combine (5.17d) and (5.22b) to obtain that

$$\begin{aligned}|\overline{\rho}_t(\overline{\mathfrak{m}}_t + E) - \rho_{y,t}(\widetilde{\mathfrak{m}}_{y,t} + E)| \le |\rho_{x,t}(\alpha\mathfrak{m}_{x,t} + (1-\alpha)\mathfrak{m}_{y,t} + E) - \rho_{y,t}(\mathfrak{m}_{y,t} + E)|\\ + (t - t_*)^{\frac{7}{12}}\end{aligned}$$

Proceeding similarly to the estimate of $|A_{4,1}|$ at the end of the proof of Lemma 5.C.1, we conclude that

$$\begin{aligned}|A_{4,1}| \lesssim \left(\frac{N^{\xi}(\ell^2 + |i|^{\frac{1}{2}})}{N^{\frac{1}{2}}} + \frac{N^{\frac{7\omega_1}{12}}}{N^{\frac{7}{24}}}\right)\int_{\mathcal{I}_{z,i}(t)\cap\{|E-\widetilde{z}_i| > N^{-60}\}} \frac{1}{|\widetilde{z}_i - E|}\,\mathrm{d}E\\ + \left|\int_{|E-\widetilde{z}_i|\le N^{-60}} \frac{\overline{\rho}_t(E+\overline{\mathfrak{m}}_t) - \rho_{y,t}(E+\widetilde{\mathfrak{m}}_{y,t})}{\widetilde{z}_i - E}\,\mathrm{d}E\right|.\end{aligned} \tag{5.245}$$

Furthermore, similarly to the estimate in the singular integral in (5.238), but substituting $\overline{\mathfrak{e}}_t^+$ and $\mathfrak{e}_{y,t}^+$ by $\overline{\mathfrak{m}}_t$ and $\widetilde{\mathfrak{m}}_{y,t}$ respectively, we conclude that that the last term in (5.245) is bounded by N^{-20}. Therefore,

$$|A_{41}| \lesssim \frac{N^{\xi}(\ell^2 + |i|^{\frac{1}{2}})}{N^{\frac{1}{2}}} + \frac{N^{\frac{7\omega_1}{12}}}{N^{\frac{7}{24}}} \lesssim \frac{N^{\frac{7\omega_1}{12}}}{N^{\frac{7}{24}}}, \tag{5.246}$$

for any $|i| \le N^{\omega_A}$. Collecting (5.242), (5.243), (5.244) and (5.246) completes the proof of Lemma 5.C.2. □

5.D Sobolev-type inequality

The proof of the Sobolev-type inequality in the cusp case is essentially identical to that in the edge case presented in Appendix B of [41]; only the exponents need adjustment to the cusp scaling. We give some details for completeness.

Proof of Lemma 5.7.5. We will prove only the first inequality in (5.163). The proof for the second one is exactly the same. We start by proving a continuous version of (5.163) and then we will conclude the proof by linear interpolation. We claim that for any small η there exists a constant $c_\eta > 0$ such that for any real function $f \in L^p(\mathbb{R}_+)$ we have that

$$\int_0^{+\infty}\int_0^{+\infty} \frac{(f(x) - f(y))^2}{\left|x^{\frac{3}{4}} - y^{\frac{3}{4}}\right|^{2-\eta}}\,\mathrm{d}x\,\mathrm{d}y \ge c_\eta \left(\int_0^{+\infty} |f(x)|^p\,\mathrm{d}x\right)^{\frac{2}{p}}. \tag{5.247}$$

We recall the representation formula for fractional powers of the Laplacian: for any $0 < \alpha < 2$ and for any function $f \in L^p(\mathbb{R})$ for some $p \in [1, \infty)$ we have

$$\langle f, |p|^\alpha f\rangle = C(\alpha) \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{(f(x) - f(y))^2}{|x-y|^{1+\alpha}} \,\mathrm{d}x \,\mathrm{d}y,$$

with some explicit constant $C(\alpha)$, where $|p| := \sqrt{-\Delta}$.

Since for $0 < x < y$ we have that

$$y^{\frac{3}{4}} - x^{\frac{3}{4}} = \frac{4}{3} \int_x^y s^{-\frac{1}{4}} \,\mathrm{d}s \leq C(y-x)(xy)^{-\frac{1}{8}},$$

in order to prove (5.247) it is enough to show that

$$\int_0^{+\infty} \int_0^{+\infty} \frac{(f(x) - f(y))^2}{|x-y|^{2-\eta}} (xy)^q \,\mathrm{d}x \,\mathrm{d}y \geq c_\eta \left(\int_0^{+\infty} |f(x)|^p \,\mathrm{d}x \right)^{\frac{2}{p}}, \tag{5.248}$$

where $q := \frac{1}{4} - \frac{\eta}{8}$ and $p := \frac{8}{2+3\eta}$. Let $\tilde{f}(x)$ be the symmetric extension of f to the whole real line, i.e. $\tilde{f}(x) := f(x)$ for $x > 0$ and $\tilde{f}(x) := f(-x)$ for $x < 0$. Then, by a simple calculation we have

$$4 \int_0^{+\infty} \int_0^{+\infty} \frac{(f(x) - f(y))^2}{|x-y|^{2-\eta}} (xy)^q \,\mathrm{d}x \,\mathrm{d}y \geq \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{(\tilde{f}(x) - \tilde{f}(y))^2}{|x-y|^{2-\eta}} |xy|^q \,\mathrm{d}x \,\mathrm{d}y.$$

Introducing $\phi(x) := |x|^q$ and dropping the tilde for f the estimate in (5.248) would follow from

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \frac{(f(x) - f(y))^2}{|x-y|^{2-\eta}} \phi(x)\phi(y) \,\mathrm{d}x \,\mathrm{d}y \geq c'_\eta \left(\int_{\mathbb{R}} |f(x)|^p \,\mathrm{d}x \right)^{\frac{2}{p}}. \tag{5.249}$$

By the same computation as in the proof of Proposition 10.5 in [41] we conclude that

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \frac{(f(x) - f(y))^2}{|x-y|^{2-\eta}} \phi(x)\phi(y) \,\mathrm{d}x \,\mathrm{d}y = \langle \phi f, |p|^{1-\eta} \phi f\rangle + C_0(\eta) \int_{\mathbb{R}} \frac{|\phi(x) f(x)|^2}{|x|^{1-\eta}} \,\mathrm{d}x$$

with some $C_0(\eta) > 0$, hence for the proof of (5.249) it is enough to show that

$$\langle \phi f, |p|^{1-\eta} \phi f\rangle \geq c_\eta \left(\int_{\mathbb{R}} |f|^p \right)^{\frac{2}{p}}.$$

Let $g := |p|^{\frac{1}{2}(1-\eta)} |x|^q f$, we need to prove that

$$\|g\|_2 \geq c_\eta \| |x|^{-q} |p|^{-\frac{1}{2}(1-\eta)} g\|_p.$$

By the n-dimensional Hardy-Littlewood-Sobolev inequality in [165] we have that

$$\left\| |x|^{-q} \int |x-y|^{-a} g(y) \,\mathrm{d}y \right\|_p \leq C \|g\|_r,$$

where $\frac{1}{r} + \frac{a+q}{n} = 1 + \frac{1}{p}, 0 \leq q < \frac{n}{p}$ and $0 < a < n$. In our case $a = \frac{1+\eta}{2}, r = 2, n = 1$ and all the conditions are satisfied if we take $0 < \eta < 1$. This completes the proof of (5.247).

Next, in order to prove (5.163), we proceed by linear interpolation as in Proposition B.2 in [79]. Given $u : \mathbb{Z} \to \mathbb{R}$, let $\psi : \mathbb{R} \to \mathbb{R}$ be its linear interpolation, i.e. $\psi(i) := u_i$ for $i \in \mathbb{Z}$ and

$$\psi(x) := u_i + (u_{i+1} - u_i)(x - i) = u_{i+1} - (u_{i+1} - u_i)(i + 1 - x),$$

for $x \in [i, i+1]$. It is easy to see that for each $p \in [2, +\infty]$ (i.e. $\eta \leq 2/3$), there exists a constant C_p such that

$$C_p^{-1} \|\psi\|_{L^p(\mathbb{R})} \leq \|u\|_{L^p(\mathbb{Z})} \leq C_p \|\psi\|_{L^p(\mathbb{R})}. \tag{5.250}$$

In order to prove (5.163) we claim that

$$\int_0^{+\infty} \int_0^{+\infty} \frac{|\psi(x) - \psi(y)|^2}{\left|x^{\frac{3}{4}} - y^{\frac{3}{4}}\right|^{2-\eta}} \, \mathrm{d}x \, \mathrm{d}y \leq c_\eta \sum_{i \neq j \in \mathbb{Z}_+} \frac{(u_i - u_j)^2}{\left|i^{\frac{3}{4}} - j^{\frac{3}{4}}\right|^{2-\eta}}, \tag{5.251}$$

for some constant $c_\eta > 0$. Indeed, combining (5.250) and (5.251) with (5.247) we conclude (5.163). Finally, the proof of (5.251) is a simple exercise along the lines of the proof of Proposition B.2 in [79]. □

5.E Heat-kernel estimates

The proof of the heat kernel estimates relies on the Nash method. In the edge scaling regime a similar bound was proven in [41] for a compact interval, extended to non-compact interval but with compactly supported initial data w_0 in [122]. Here we closely follow the latter proof, adjusted to the cusp regime, where interactions on both sides of the cusp play a role unlike in the edge regime.

Proof of Lemma 5.7.6. We start proving (5.164), then (5.165) follows by (5.164) by duality. Without loss of generality we assume $\|w_0\|_1 = 1$ and that

$$\|w(\tilde{s})\|_p \geq N^{-100} \tag{5.252}$$

for each $s \leq \tilde{s} \leq t$, where $w(\tilde{s}) = \mathcal{U}^{\mathcal{L}}(s, \tilde{s}) w_0$. Otherwise, by ℓ^p-contraction we had $\|w(\tilde{s})\|_p \leq N^{-100}$ implying (5.164) directly.

In the following we use the convention $w := w(\tilde{s})$ if there is no confusion. By (5.163), we have that

$$\|w\|_p^2 \lesssim \sum_{\substack{i,j \geq 1 \\ i \neq j}} \frac{(w_i - w_j)^2}{\left|i^{\frac{3}{4}} - j^{\frac{3}{4}}\right|^{2-\eta}} + \sum_{\substack{i,j \leq -1 \\ i \neq j}} \frac{(w_i - w_j)^2}{\left|\,|i|^{\frac{3}{4}} - |j|^{\frac{3}{4}}\,\right|^{2-\eta}}.$$

First we assume that both i and j are positive. Let $\delta_4 < \delta_2 < \delta_3 < \frac{\delta_1}{2}$. We start with the following estimate

$$\sum_{\substack{i,j \geq 1 \\ i \neq j}} \frac{(w_i - w_j)^2}{\left|i^{\frac{3}{4}} - j^{\frac{3}{4}}\right|^{2-\eta}} \lesssim \sum_{\substack{(i,j) \in \mathcal{A} \\ i,j \geq 1}} \frac{(w_i - w_j)^2}{\left|i^{\frac{3}{4}} - j^{\frac{3}{4}}\right|^{2-\eta}} + \sum_{i \geq 1} \sum_{j \geq 1}^{\mathcal{A}^c,(i)} \frac{w_i^2}{\left|i^{\frac{3}{4}} - j^{\frac{3}{4}}\right|^{2-\eta}}. \tag{5.253}$$

We proceed by writing

$$\sum_{\substack{(i,j)\in\mathcal{A}\\ i,j\geq 1}} \frac{(w_i-w_j)^2}{\left|i^{\frac{3}{4}}-j^{\frac{3}{4}}\right|^{2-\eta}} \lesssim \sum_{\substack{(i,j)\in\mathcal{A}:\, i,j\geq 1\\ i \text{ or } j\leq \ell^4 N^{\delta_2}}} \frac{(w_i-w_j)^2}{\left|i^{\frac{3}{4}}-j^{\frac{3}{4}}\right|^{2-\eta}} + \sum_{\substack{(i,j)\in\mathcal{A}\\ i,j\geq \ell^4 N^{\delta_2}}} \frac{(w_i-w_j)^2}{\left|i^{\frac{3}{4}}-j^{\frac{3}{4}}\right|^{2-\eta}}.$$

By Lemma 5.B.3 we have that

$$\sum_{\substack{(i,j)\in\mathcal{A}\\ i,j\geq \ell^4 N^{\delta_2}}} \frac{(w_i-w_j)^2}{\left|i^{\frac{3}{4}}-j^{\frac{3}{4}}\right|^{2-\eta}} \lesssim N^{-200}, \tag{5.254}$$

since $i\geq \ell^4N^{\delta_2}$ and $|(w_0)_j|\leq N^{-100}$ for $j\geq \ell^4 N^{\delta_4}$ by our hypotheses. Indeed, for $i\geq \ell^4N^{\delta_2}$, we have that

$$w_i=\left(\mathcal{U}^{\mathcal{L}}(s,\tilde{s})w_0\right)_i=\sum_{j=-N}^{N}\mathcal{U}^{\mathcal{L}}_{ij}(w_0)_j=\sum_{j=-\ell^4N^{\delta_4}}^{\ell^4N^{\delta_4}}\mathcal{U}^{\mathcal{L}}_{ij}(w_0)_j+N^{-100}\lesssim N^{-100}, \tag{5.255}$$

with very high probability. If $(i,j)\in\mathcal{A}$, $i,j\geq 1$ and i or j are smaller than $\ell^4N^{\delta_2}$ then both i and j are smaller than $\ell^4N^{\delta_3}$. Hence, for such i and j, by (5.162), we have that

$$|\widehat{z}_i(t,\alpha)-\widehat{z}_j(t,\alpha)|\lesssim \frac{N^{\frac{\omega_1}{6}}\left|i^{\frac{3}{4}}-j^{\frac{3}{4}}\right|}{N^{\frac{3}{4}}}, \tag{5.256}$$

for any fixed $\alpha\in[0,1]$ and for all $0\leq t\leq t_*$, where $\widehat{z}_i(t,\alpha)$ is defined by (5.128)-(5.129). If i and j are both negative the estimates in (5.253)-(5.256) follow in the same way.

In the following of the proof $\mathcal{B}$, $\mathcal{B}_{ij}$ and $\mathcal{V}_i$ are defined in (5.128)-(5.129). By (5.256) it follows that

$$\begin{aligned}\sum_{\substack{(i,j)\in\mathcal{A}:\, i,j\geq 1\\ i \text{ or } j\leq \ell^4N^{\delta_2}}} \frac{(w_i-w_j)^2}{\left|i^{\frac{3}{4}}-j^{\frac{3}{4}}\right|^{2-\eta}} &+ \sum_{\substack{(i,j)\in\mathcal{A}:\, i,j\leq -1\\ i \text{ or } j\geq -\ell^4N^{\delta_2}}} \frac{(w_i-w_j)^2}{\left|i^{\frac{3}{4}}-j^{\frac{3}{4}}\right|^{2-\eta}}\\ &\lesssim -N^{-\frac{1}{2}}N^{\frac{\omega_1}{3}+C\eta}\sum_{(i,j)\in\mathcal{A}}\mathcal{B}_{ij}(w_i-w_j)^2=-2N^{-\frac{1}{2}}N^{\frac{\omega_1}{3}+C\eta}\langle w,\mathcal{B}w\rangle.\end{aligned} \tag{5.257}$$

Furthermore, since $1\leq |i|\leq \ell^4N^{\delta_3}$, we have that

$$\sum_j^{\mathcal{A}^c,(i)}\frac{1}{\left||i|^{\frac{3}{4}}-|j|^{\frac{3}{4}}\right|^{2-\eta}}\lesssim\frac{N^{\frac{\omega_1}{3}+C\eta}}{N^{\frac{3}{2}}}\sum_j^{\mathcal{A}^c,(i)}\frac{1}{(\widehat{z}_i-\widehat{z}_j)^2}. \tag{5.258}$$

By the rigidity (5.160), (5.161) and (5.162), we can replace $\widehat{z}_j$ by $\overline{\gamma}_j$ in the sum on the rhs. of (5.258) and so approximate it by an integral, then using that $\overline{\rho}_t(E)\lesssim\rho_{y,t}(E)$ in the cusp regime, i.e. $|E|\leq\delta_*$, with δ_* defined in Definition 5.4.1, we conclude that

$$\frac{1}{N}\sum_j^{\mathcal{A}^c,(i)}\frac{1}{(\widehat{z}_i(t)-\widehat{z}_j(t))^2}\lesssim\int_{I_{i,y}(t)^c}\frac{\rho_{y,t}(E+\mathfrak{e}^+_{y,t})}{(\widehat{z}_i(t)-E)^2}\,\mathrm{d}E=-\mathcal{V}_i. \tag{5.259}$$

Hence, by (5.259), we conclude that

$$\begin{aligned}\sum_i \sum_j^{\mathcal{A}^c,(i)} \frac{w_i^2}{\left||i|^{\frac{3}{4}}-|j|^{\frac{3}{4}}\right|^{2-\eta}} &\lesssim \sum_{1\le|i|\le\ell^4N^{\delta_3}} \sum_j^{\mathcal{A}^c,(i)} \frac{w_i^2}{\left||i|^{\frac{3}{4}}-|j|^{\frac{3}{4}}\right|^{2-\eta}}+N^{-200} \\ &\lesssim -N^{-\frac{1}{2}}N^{\frac{\omega_1}{3}+C\eta}\sum_{|i|\le\ell^4N^{\delta_3}} w_i^2\mathcal{V}_i+N^{-200} \\ &\lesssim -N^{-\frac{1}{2}}N^{\frac{\omega_1}{3}+C\eta}\langle w,\mathcal{V}w\rangle+N^{-200}.\end{aligned} \tag{5.260}$$

Note that in the first inequality of (5.260) we used (5.255).

Summarizing (5.254), (5.257) and (5.260) and rewriting N^{-200} into an ℓ^p-norm using (5.252), we obtain

$$\|w\|_p^2\le -N^{-\frac{1}{2}}N^{\frac{\omega_1}{3}+C\eta}\langle w,\mathcal{L}w\rangle+\frac{1}{10}\|w\|_p^2.$$

Hence, using Hölder inequality, we have that

$$\begin{aligned}\partial_t\|w\|_2^2=\langle w,\mathcal{L}w\rangle &\le -c_\eta N^{\frac{1}{2}}N^{-\frac{\omega_1}{3}-C\eta}\|w\|_p^2 \\ &\le -c_\eta N^{\frac{1}{2}}N^{-\frac{\omega_1}{3}-C\eta}\|w\|_2^{\frac{6-3\eta}{2}}\|w\|_1^{-\frac{2-3\eta}{2}} \\ &\le -c_\eta N^{\frac{1}{2}}N^{-\frac{\omega_1}{3}-C\eta}\|w\|_2^{\frac{6-3\eta}{2}}\|w_0\|_1^{-\frac{2-3\eta}{2}}.\end{aligned} \tag{5.261}$$

In the last inequality of (5.261) we used the ℓ^1-contraction of $\mathcal{U}^{\mathcal{L}}$. Integrating (5.261) back in time, it easily follows that

$$\|\mathcal{U}^{\mathcal{L}}(s,t)w_0\|_2\le\left(\frac{N^{C\eta+\frac{\omega_1}{3}}}{c_\eta N^{\frac{1}{2}}(t-s)}\right)^{1-3\eta}\|w_0\|_1, \tag{5.262}$$

proving (5.164). The same bound also holds for the transpose operator $(\mathcal{U}^{\mathcal{L}})^T$.

In order to prove (5.165) we follow Lemma 3.11 of [122]. Let $\chi(i):=\mathbf{1}_{\{|i|\le\ell^4N^{\delta_5}\}}$, with $\delta_4<\delta_5<\frac{\delta_1}{2}$, and $v\in\mathbb{R}^{2N}$. Then, we have that

$$\langle\mathcal{U}^{\mathcal{L}}(0,t)w_0,v\rangle=\langle w_0,(\mathcal{U}^{\mathcal{L}})^T\chi v\rangle+\langle w_0,(\mathcal{U}^{\mathcal{L}})^T(1-\chi)v\rangle.$$

By Lemma 5.B.3 we have that

$$\left|\langle w_0,(\mathcal{U}^{\mathcal{L}})^T(1-\chi)v\rangle\right|\le N^{-100}\|w_0\|_2\|v\|_1. \tag{5.263}$$

By (5.164) and Cauchy-Schwarz inequality we have that

$$\left|\langle w_0,(\mathcal{U}^{\mathcal{L}})^T\chi v\rangle\right|\le\|w_0\|_2\|(\mathcal{U}^{\mathcal{L}})^T\chi v\|_2\le\|w_0\|_2\left(\frac{N^{C\eta+\frac{\omega_1}{3}}}{c_\eta N^{\frac{1}{2}}t}\right)^{1-3\eta}\|v\|_1. \tag{5.264}$$

Hence, combining (5.263) and (5.264), we conclude that

$$\|\mathcal{U}^{\mathcal{L}}(0,t)w_0\|_\infty\le\left(\frac{N^{C\eta+\frac{\omega_1}{3}}}{c_\eta N^{\frac{1}{2}}t}\right)^{1-3\eta}\|w_0\|_2, \tag{5.265}$$

and so, by (5.262), that

$$\begin{aligned}\|\mathcal{U}^{\mathcal{L}}(0,t)w_0\|_\infty &= \|\mathcal{U}^{\mathcal{L}}(t/2,t)\mathcal{U}^{\mathcal{L}}(0,t/2)w_0\|_\infty \\ &\lesssim \left(\frac{N^{C\eta+\frac{\omega_1}{3}}}{c_\eta N^{\frac{1}{2}}t}\right)^{1-3\eta} \|\mathcal{U}^{\mathcal{L}}(0,t/2)w_0\|_2 \lesssim \left(\frac{N^{C\eta+\frac{\omega_1}{3}}}{c_\eta N^{\frac{1}{2}}t}\right)^{2(1-3\eta)} \|w_0\|_1,\end{aligned} \tag{5.266}$$

where in the first inequality we used that $\mathcal{U}^{\mathcal{L}}(0,t/2)w_0$ satisfies the hypothesis of Lemma 5.7.6, since $\left|(\mathcal{U}^{\mathcal{L}}(0,t/2)w_0)_i\right| \leq N^{-100}$ for $|i| \geq \ell^4 N^{2\delta_4}$ by the finite speed estimate of Lemma 5.B.3. Combining (5.265) and (5.266) then (5.165) follows by interpolation. □

Fluctuations of Rectangular Young Diagrams of Interlacing Wigner Eigenvalues 6

We prove a new CLT for the difference *of linear eigenvalue statistics of a Wigner random matrix H and its minor $\widehat{H}$ and find that the fluctuation is much smaller than the fluctuations of the individual linear statistics, as a consequence of the strong correlation between the eigenvalues of H and $\widehat{H}$. In particular our theorem identifies the fluctuation of Kerov's rectangular Young diagrams, defined by the interlacing eigenvalues of H and $\widehat{H}$, around their asymptotic shape, the Vershik-Kerov-Logan-Shepp curve. Young diagrams equipped with the Plancherel measure follow the same limiting shape. For this, algebraically motivated, ensemble a CLT has been obtained in [105] which is structurally similar to our result but the variance is different, indicating that the analogy between the two models has its limitations. Moreover, our theorem shows that Borodin's result [34] on the convergence of the spectral distribution of Wigner matrices to a Gaussian free field also holds in derivative sense.*

Published as L. Erdős and D. Schröder, *Fluctuations of rectangular Young diagrams of interlacing Wigner eigenvalues*, Int. Math. Res. Not. IMRN, 3255–3298 (2016), MR3805203.

6.1 Introduction

There is a rich history of probabilistic models of essentially algebraic nature with surprising connections to random matrix theory. Examples include the longest increasing subsequence in random permutations [22], queuing processes [25], random tilings of a hexagon [110], poly-nuclear growth processes [149] and 1+1 dimensional exclusion processes (see e.g. [37] for a good overview of the topic). Recent years have seen a spectacular progress towards the KPZ universality that is detected in the extreme regimes. The intuition for the KPZ universality often comes from relating these model to the extreme eigenvalues of random matrices. Many of these models are related to a classical algebraic problem, the statistics of Young tableaux from representation theory. In this paper we focus on the bulk regime and we investigate the analogy between large Young diagrams equipped with the classical Plancherel measure and Kerov's *rectangular Young diagrams*, originating from eigenvalue statistics of minors of large random Wigner random matrices. Their limiting shape curves coincide. Here we identify the fluctuation of the rectangular Young diagrams and establish the precise conditions when it is Gaussian and we compute its correlation. We find that the limiting behavior of the two diagram ensembles are not the same, even though in the extreme regime their statistics coincide.

Given an integer N and a partition $N = \lambda_1 + \lambda_2 + \dots$ of N into integers $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$, the planar figure obtained as the union of consecutive rows consisting of $\lambda_1, \lambda_2, \dots$ unit square cells, is called the *Young diagram* corresponding to λ of size $|\lambda| = N$. Young diagrams of size N are commonly considered as a probability space equipped with the *Plancherel measure* $P_N(\lambda) := d_\lambda^2/N!$, where d_λ is the number of Young tableaux with given shape λ (see, e.g. [86])

The first major connection between Young diagrams and random matrix theory was established by Baik, Deift and Johansson who showed in [22] that the distribution of $\lambda_1/\sqrt{N}$ with respect to $P_N(\lambda)$ asymptotically agrees with the distribution of the largest eigenvalue of an $N \times N$ GUE matrix, hence it follows the Tracy Widom law [170]. Similar result [20] holds for $\lambda_2/\sqrt{N}$ and the second largest eigenvalue, and Okounkov [139] established that the joint distribution of $\lambda_1/\sqrt{N}, \dots, \lambda_k/\sqrt{N}$ asymptotically follows that of the k largest eigenvalues of the GUE. Alternative proofs are given in [38, 108]. In fact, in [38] also sine kernel universality in the bulk regime (that is, correlation functions of rows λ_k with $k \sim \sqrt{N}$) has been proven.

To study the bulk behavior of Young diagrams, it is convenient to draw them in the Russian convention which is rotated by $45°$ from the horizontal convention (see Figure 6.1). In this way we can view the upper boundary the diagram as a continuous function $E \mapsto \lambda(E)$ such that $\lambda(E) \geq |E|$ and $\lambda'(E) = \pm 1$, whenever it is defined. We can continuously extend this function by $\lambda(E) = |E|$ outside the extent of the diagram. The limiting shape and the fluctuation of this curve under the Plancherel measure, after proper rescaling, has been determined:

$$\frac{1}{\sqrt{N}}\lambda(\sqrt{N}E) \approx \Omega(E) + \frac{2}{\sqrt{N}}\Delta(E), \qquad N \to \infty, \tag{6.1}$$

where

$$\Omega(E) := \begin{cases} |E| & \text{if } |E| \geq 2 \\ \frac{2}{\pi}\left[E \arcsin \frac{E}{2} + \sqrt{4 - E^2}\right] & \text{else} \end{cases}$$

is the *Vershik-Kerov-Logan-Shepp curve*. The fluctuation term $\Delta(E)$ is a generalized Gaussian process on the interval $[-2, 2]$ that can be defined by the trigonometric series

$$\Delta(2\cos\theta) = \frac{1}{\pi}\sum_{k\geq 2}\frac{\xi_k \sin k\theta}{\sqrt{k}}$$

of independent standard Gaussian random variables ξ_k. The limit shape has been independently identified in [130] and [175], the fluctuation was proved in [105] following Kerov's unpublished notes.

A direct connection between random matrices and Young diagrams in the bulk regime was found by Kerov in [114]. He showed that for a Wigner random matrix $H \in \mathbb{C}^{N\times N}$ and an independent random $N - 1$ dimensional hyperplane h with uniformly distributed

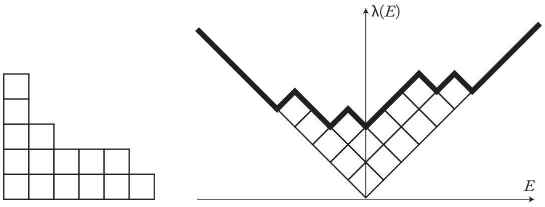


Figure 6.1: Young diagram in French and Russian convention corresponding to the partition $15 = 6 + 5 + 2 + 1 + 1$, together with the curve $\lambda(E)$

normal vector, the eigenvalues $\lambda_1, \ldots, \lambda_N$, and $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{N-1}$ of H and $P_h H P_h$, where P_h is the projection onto h, can be used to construct a curve very similar to Young diagrams. He defined a *rectangular Young diagram* (for a more general context, see [141]) as the function

$$w_N(E) := \sum_{k=1}^{N} |\lambda_k - E| - \sum_{k=1}^{N-1} \left|\widehat{\lambda}_k - E\right|, \qquad E \in \mathbb{R}.$$

It is easy to see that w_N is the unique piecewise-linear continuous function with local minima in $\lambda_1 \leq \ldots \leq \lambda_N$ and local maxima in $\widehat{\lambda}_1 \leq \ldots \leq \widehat{\lambda}_{N-1}$ such that the slope is ± 1 whenever it exists and $w_N(E) = \left|E - \sum \lambda_k + \sum \widehat{\lambda}_k\right|$ for large enough $|E|$. It was then shown that

$$\lim_{N\to\infty} \mathbf{E}\, w_N(E) = \Omega(E),$$

uniformly in E.

Bufetov in [51] has recently improved this result in two directions. First, he showed that the randomness in the choice of the projection is not needed; it is sufficient to consider the eigenvalues of H and its minor $\widehat{H} = (h_{ij})_{i,j\geq 2}$ (where the choice of removed row/column is, of course, arbitrary). Second, he improved the convergence in expectation to convergence in probability;

$$\lim_{N\to\infty} \sup_{E} |w_N(E) - \Omega(E)| = 0. \tag{6.2}$$

We note that $\sum_k \lambda_k = \sum_k \widehat{\lambda}_{k-1} + h_{11}$, so $w_N(E) = |E - h_{11}|$ for large E and thus it does not exactly match $\Omega(E)$ even outside of the limiting spectrum $[-2, 2]$. To remedy this, we will also consider the *shifted diagram*

$$\widetilde{w}_N(E) := w_N(E + h_{11})$$

which agrees identically with $\Omega(E)$ outside the spectrum. This modification is irrelevant for the limit shape but it becomes relevant when we consider fluctuations. Figure 6.2 shows realizations of $\widetilde{w}_N$ for different values of N together with the limiting curve Ω.

In the present work we upgrade the law of large numbers type results (6.2) to a central limit theorem (CLT) as in (6.1), and thus demonstrate that a certain analogy between random matrices and representation theory extends beyond the macroscopic behavior. Specifically, we prove that

$$w_N(E) \approx \Omega(E) + \frac{1}{\sqrt{N}}\left[\widehat{\Delta}(E) + \xi_{11}\frac{E\sqrt{(4-E^2)_+} + 4\arcsin E/2}{2\pi}\right], \tag{6.3}$$

$$\widetilde{w}_N(E) \approx \Omega(E) + \frac{1}{\sqrt{N}}\left[\widehat{\Delta}(E) + \xi_{11}\frac{E\sqrt{(4-E^2)_+}}{2\pi}\right] \tag{6.4}$$

where $\widehat{\Delta}(E)$ are collections of centered Gaussian random variables whose covariance structure we explicitly compute and $\xi_{11} = \sqrt{N}h_{11}$ is independent of them. Therefore the fluctuations of w_N and $\widetilde{w}_N$ are Gaussian if and only if h_{11} is Gaussian. We also conclude from our explicit formulas for the variances that although (6.3) resembles (6.1), the distribution of the Gaussian part of the fluctuation, $\widehat{\Delta}(E)$ and $\Delta(E)$ do not agree. For example – in contrast to $\Delta(E)$ – the fluctuation term $\widehat{\Delta}(E)$ has a finite variance.

Motivated by the preprint of the current paper, Sasha Sodin [159] considered another rectangular Young diagram w_N^* obtained from the interlacing roots and extrema of the characteristic polynomial of H. He found that

$$w_N^*(E) \approx \Omega(E) + \frac{1}{N}\widetilde{\Delta}(E), \tag{6.5}$$

albeit in a weaker sense than (6.3), where $\widetilde{\Delta}(E)$ is a generalized Gaussian process closely related to $\Delta(E)$ in (6.1). In particular the fluctuations of w_N^* are always Gaussian; the distribution of any specific matrix entry does not play a distinguished role. The difference between w_N and w_N^* can be understood via the Markov correspondence (see, e.g. [115]). Sodin pointed out that the rectangular Young diagram w_N created by a random matrix H and its minor $\widehat{H}$ is related to the entrywise spectral measure ρ_N, defined as $\int f \,\mathrm{d}\rho_N := f(H)_{11}$, while the empirical spectral density $\mu_N = \frac{1}{N}\sum_k \delta_{\lambda_k}$ corresponds to the rectangular Young diagram w_N^*. Thus the behavior of w_N^* is directly related to $\frac{1}{N}\operatorname{Tr} f(H)$ and not to $f(H)_{11}$ which also explains the difference in the size of the fluctuations between (6.3) and (6.5). For more details on the relation of w_N and w_N^* we refer to [159].

We prove our results (6.3)–(6.4) as corollaries to a new central limit theorem for the *difference* in linear eigenvalue statistics of a Wigner random matrix and its minor. For many classes of random matrices $H = H^{(N)} \in \mathbb{C}^{N\times N}$ the empirical spectral density, i.e., the normalized counting measure of eigenvalues, $\frac{1}{N}\sum_{k=1}^N \delta_{\lambda_k}$ converges weakly to a deterministic measure ρ as $N \to \infty$, which may be viewed as type of law of large numbers. Phrased in terms of an appropriate test function f,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^N f(\lambda_k(H^{(N)})) = \int f(x)\rho(\mathrm{d}x),$$

naturally raises the question whether the fluctuations in this convergence also follow an analogue of the central limit theorem. The object $\sum_{k=1}^N f(\lambda_k(H^{(N)})) = \operatorname{Tr} f(H^{(N)})$, called the *linear eigenvalue statistics* of $H^{(N)}$, has been studied for many types of random matrices [15, 132, 154, 24, 157, 162, 106] and large classes of test functions f. Contrary to the classical CLT, the fluctuations of the linear eigenvalue statistics do not grow with N, at least if f is sufficiently regular. The fluctuations are typically Gaussian, but there are also some pathological examples where this is not the case, e.g. for certain invariant ensembles with density

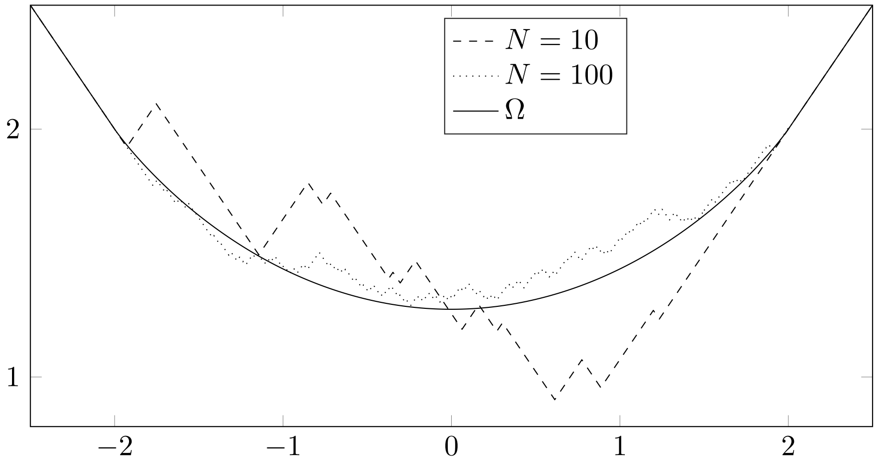


Figure 6.2: Sample rectangular Young diagrams $\widetilde{w}_N$ with limiting shape Ω

supported on several intervals [142]. For polynomial test functions f the Gaussian fluctuation can be proved by the elementary moment method, see e.g. [14, Theorem 2.1.13], but a simple approximation argument does not suffice to extend the result to less regular f. CLT still holds, for example it has been shown in [132] that for GOE random matrices and test functions f with bounded derivative $\operatorname{Tr} f(H^{(N)}) - \mathbf{E} \operatorname{Tr} f(H^{(N)})$ converges in distribution to a centered Gaussian random variable of variance

$$\frac{1}{2\pi^2} \int_{-2}^{2} \int_{-2}^{2} \left(\frac{f(x) - f(x')}{x - x'} \right)^2 \frac{4 - xx'}{\sqrt{4 - x^2}\sqrt{4 - x'^2}} \,\mathrm{d}x \,\mathrm{d}x'.$$

The currently weakest regularity conditions on f for CLT are found in [162]; $f \in H^{1+\epsilon}$ is necessary for general Wigner matrices and $f \in C^{1/2+\epsilon}$ suffices for GUE. We stress that linear statistics are very sensitive to regularity of the test function; while polynomial test functions do not require understanding of any local eigenvalue statistics (the global moment method works), the proof in [162] for the Wigner case heavily relied on techniques developed to prove local semicircle laws [82], while the GUE case even used the Brézin-Hikami formula and saddle point analysis of the determinantal kernel by Johansson [109].

All previous work concerned linear statistics of a single Wigner matrix except two papers by Borodin [34, 35] and a few recent works motivated by them. In these papers joint fluctuations of linear statistics of Wigner matrices and its minors were investigated (see also [111] where a similar question was discussed for d-regular graphs). Borodin considered general families of regularly nested minors and identified the limit of their joint spectral counting functions as a Gaussian free field (GFF), but the test function was polynomial and thus a relatively simple extension of the moment method [14] worked. The class of test functions was extended to include functions with a high Sobolev regularity ($H^{2.5+\epsilon}$ for Gaussian and $H^{5.5+\epsilon}$ for general Wigner matrices) using a Chebyshev basis decomposition [128] (see also [112] where not only nested but overlapping matrices were considered). However, all these results identify the joint fluctuations on order one scale, whose correlations are typically strictly between 0 and 1 for a collection of minors whose sizes asymptotically differ by cN. Our work detects the small fluctuation of order $N^{-1/2}$ resulting from the very strong correlation between minors of almost the same size. This fine effect is not visible on the scale of the analysis in [34, 35, 128]. Nevertheless, one may ask whether the fine scale covariance structure proven in our main Theorem 6.2.1 is consistent with the covariance formula in [34, 35, 128] if one formally applies it to H and its immediate minor $\widehat{H}$ disregarding the interchange of limits. Effectively this question is equivalent to asking whether the convergence of the spectral counting functions of the minors to the GFF also holds in derivative sense. In Appendix 6.A we show that the derivative of the GFF predicts the correct variance of the fluctuations but fails to identify their distribution, in general. This is essentially due to the fact that our fine scale result depends on the precise distribution of h_{11} while the macroscopic formula does not depend on any individual matrix entry.

Inspired by Kerov's rectangular Young diagrams, in the present work we study the difference of two linear statistics $f_N := \operatorname{Tr} f(H) - \operatorname{Tr} f(\widehat{H})$ of a Wigner matrix and its minor for a large class of test functions that includes $f(x) = |x - E|$. We find that the expectation of f_N converges to

$$\Omega_f := \frac{1}{\pi}\int_{-2}^{2} \frac{f(x)}{\sqrt{4-x^2}}\,\mathrm{d}x$$

and its fluctuations around Ω_f are of order $N^{-1/2}$. In particular, the fluctuations we identify are much smaller than those of the individual linear statistics, as a result of the strong correlation of the eigenvalues of H and $\widehat{H}$. Moreover, we prove that the fluctuations are Gaussian if and only if h_{11} follows a normal distribution. It is clear that h_{11} plays a special role, since for example with $f(x) = x$, we have $f_N = \operatorname{Tr} H - \operatorname{Tr} \widehat{H} = h_{11}$. Since our test function has a relatively low regularity, our proof requires to understand the spectral statistics on small mesoscopic scales. In practice, we jointly analyze the Green functions $G(z) = (H-z)^{-1}$ and $\widehat{G}(z) = (\widehat{H}-z)^{-1}$ on a spectral scale $\Im z \geq N^{-2/3}$.

After completing this manuscript, we learned[1] from Vadim Gorin that he and Lingfu Zhang have obtained [92] the exact analogue of our result for the multilevel extension of the β-Jacobi ensemble that was introduced in [36] as an analogue of the minor process for general β-ensemble.

Acknowlodgment. The authors are grateful for discussions with Zhigang Bao and for advice on references from Alexei Borodin. We thank Vadim Gorin for motivating the observation discussed in Appendix 6.A.

6.2 Main Results

We consider complex Hermitian and real symmetric random matrices and their minors of the form

$$H := \begin{pmatrix} h_{11} & \dots & h_{N1} \\ \vdots & \ddots & \vdots \\ h_{1N} & \dots & h_{NN} \end{pmatrix}, \qquad \widehat{H} := \begin{pmatrix} h_{22} & \dots & h_{N2} \\ \vdots & \ddots & \vdots \\ h_{2N} & \dots & h_{NN} \end{pmatrix}$$

with $(h_{ij})_{i,j=1}^{N}$ being independent (up to the symmetry constraint $h_{ij} = \overline{h_{ji}}$) random variables satisfying

$$\mathbf{E}\, h_{ij} = 0, \quad \mathbf{E}\,|h_{ij}|^2 = \frac{s_{ij}}{N} \quad \text{and} \quad \mathbf{E}\,|h_{ij}|^p \leq \frac{\mu_p}{N^{p/2}} \tag{6.6}$$

for all i, j, p and some absolute constants μ_p. Our main result about the difference of linear eigenvalue statistics of a Wigner random matrix and its minor is as follows.

[1] Private communication

Theorem 6.2.1. *Let the Wigner matrix H satisfy* (6.6), $s_{ij} = 1$ *for* $i \neq j$ *and* $s_{ii} \leq C$ *for all* i, $\mathbf{E}\,|h_{1j}|^4 = \sigma_4/N^2$ *for* $j = 2, \ldots, N$ *and* $\mathbf{E}\,h_{ij}^2 = \sigma_2/N$ *for* $i < j$. *Moreover, let* $f \in H^2([-10, 10])$ *be some real-valued function. Then the random variables*

$$f_N := \operatorname{Tr} f(H) - \operatorname{Tr} f(\widehat{H}) = \sum_{k=1}^{N} f(\lambda_k) - \sum_{k=1}^{N-1} f(\widehat{\lambda}_k)$$

and

$$\widetilde{f}_N := \sum_{k=1}^{N} f(\lambda_k - h_{11}) - \sum_{k=1}^{N-1} f(\widehat{\lambda}_k - h_{11})$$

are approximately given by

$$\begin{aligned} f_N &\approx \Omega_f + \frac{1}{\sqrt{N}} \left[\Delta_f + \xi_{11} \int_{-2}^{2} f'(x)\rho(x)\,\mathrm{d}x \right] \\ \widetilde{f}_N &\approx \Omega_f + \frac{1}{\sqrt{N}} \left[\Delta_f + \xi_{11} \int_{-2}^{2} \frac{x f''(x)}{2} \rho(x)\,\mathrm{d}x \right], \end{aligned} \tag{6.7}$$

where $\rho(x) := \frac{1}{2\pi}\sqrt{4 - x^2}$ *is the density of the semicircle law,*

$$\Omega_f := \frac{1}{\pi} \int_{-2}^{2} \frac{f(x)}{\sqrt{4 - x^2}}\,\mathrm{d}x$$

and Δ_f *is a centered Gaussian random variable, independent of* ξ_{11}. *Its variance is given by the explicit formulas*

$$\begin{aligned} &\mathbf{E}(\Delta_f)^2 = V_f := V_{f,1} + |\sigma_2|^2\, V_{\sigma_2} + (\sigma_4 - 1) V_{f,2} \\ &V_{f,1} = \int_{-2}^{2} f'(x)^2 \rho(x)\,\mathrm{d}x - \left(\int_{-2}^{2} x f'(x)\rho(x)\,\mathrm{d}x \right)^2 - \left(\int_{-2}^{2} f'(x)\rho(x)\,\mathrm{d}x \right)^2, \qquad (6.8) \\ &V_{f,2} = \left(\int_{-2}^{2} x f'(x)\rho(x)\,\mathrm{d}x \right)^2, \qquad (6.9) \end{aligned}$$

where V_{σ_2}, *as defined in eq.* (6.42), *is a correction term only needed when* $\sigma_2 \neq 0$. *For the special case of symmetric Wigner matrices* H *where* $\sigma_2 = 1$ *holds automatically, we have* $V_{\sigma_2} = V_{f,1}$.

More precisely, for any fixed $\epsilon > 0$,

$$\mathbf{E}\,f_N = \Omega_f + \mathcal{O}\left(N^{-2/3+\epsilon}\right), \qquad \mathbf{E}\,\widetilde{f}_N = \Omega_f + \mathcal{O}\left(N^{-2/3+\epsilon}\right),$$

and

$$\begin{aligned} \sqrt{N}\,[f_N - \Omega_f] - \xi_{11} \int_{-2}^{2} f'(x)\rho(x)\,\mathrm{d}x &\Rightarrow \Delta_f, \\ \sqrt{N}\left[\widetilde{f}_N - \Omega_f\right] - \xi_{11} \int_{-2}^{2} \frac{x f''(x)}{2}\rho(x)\,\mathrm{d}x &\Rightarrow \Delta_f \end{aligned}$$

converge in distribution to Δ_f. *Any fixed moment of these random variables converges at least at a rate of* $\mathcal{O}\left(N^{-1/6+\epsilon}\right)$ *to the corresponding Gaussian moments.*

Remark 6.2.2. *Theorem 6.2.1 shows that the fluctuations of f_N and $\widetilde{f}_N$ are always Gaussian if $\int f'(x)\rho(x)\,\mathrm{d}x = 0$ or $\int f''(x)x\rho(x)\,\mathrm{d}x = 0$, respectively. For generic f not fulfilling these conditions the fluctuations are Gaussian if and only if h_{11} follows a Gaussian distribution.*

By polarization identity, the limiting covariances of $\sqrt{N}[f_N - \Omega_f]$ and $\sqrt{N}[g_N - \Omega_g]$ may be obtained for any pair of functions f, g. In particular, Theorem 6.2.1 extends to complex test functions f by considering its real and imaginary parts separately. We also note that the condition $f \in H^2$ is not essential. The theorem holds for any $f \in H^1$, provided that $\int_{-2}^{2} |\rho'(x)xf'(x)|\,\mathrm{d}x < \infty$. Finally, we remark that the same statement holds for generalized Wigner matrices where we assume $s_{ij} = 1$ only for $i = 1$ and $j > 1$. For $i \geq 2$ we only need to assume

$$\sum_{j\geq 2} s_{ij} = N - 1, \qquad \max_{i,j} s_{ij} \leq C$$

for some constant C. We leave it to the reader to check that our proof carries over with minor modifications to this general case, as well.

Applied to rectangular Young diagrams, this result translates to:

Theorem 6.2.3. *Let the Wigner matrix H satisfy* (6.6), *$s_{ij} = 1$ for $(i,j) \neq (1,1)$, $\mathbf{E}\,|h_{1j}|^4 = \sigma_4/N^2$ for $j = 2, \ldots, N$ and $\mathbf{E}\,h_{ij}^2 = \sigma_2/N$ for $i < j$. Then – in the sense of Theorem 6.2.1 and with the same error bounds – we asymptotically have*

$$w_N(E) \approx \Omega(E) + \frac{1}{\sqrt{N}}\left[\widehat{\Delta}(E) + \xi_{11}\frac{E\sqrt{(4-E^2)_+} + 4\arcsin E/2}{2\pi}\right]$$

and

$$\widetilde{w}_N(E) \approx \Omega(E) + \frac{1}{\sqrt{N}}\left[\widehat{\Delta}(E) + \xi_{11}\frac{E\sqrt{(4-E^2)_+}}{2\pi}\right],$$

where $\widehat{\Delta}(E)$ is a centered Gaussian, independent of ξ_{11}. Its variance is given by the explicit formulas

$$\mathbf{E}[\widehat{\Delta}(E)]^2 = V(E) := V_1(E) + |\sigma_2|^2\,V_{\sigma_2}(E) + (\sigma_4 - 1)V_2(E),$$

$$V_1(E) = 1 - \frac{(4-E^2)_+^3}{9\pi^2} - \frac{\left(E\sqrt{(4-E^2)_+} + 4\arcsin E/2\right)^2}{4\pi^2}, \quad V_2(E) = \frac{(4-E^2)_+^3}{9\pi^2},$$

where it is understood that $\arcsin(\pm x) = \pm\pi/2$ for $x > 2$. The correction term $V_{\sigma_2}(E)$, that is only needed when $\sigma_2 \neq 0$, can be obtained via the general formula for V_{σ_2} from (6.42). *For the special case of real symmetric H, we have $V_{\sigma_2}(E) = V_1(E)$.*

A simple inspection also shows that $\widetilde{w}_N$ not only becomes deterministic for $|E| \geq 2$, but it has smaller fluctuation than w_N everywhere. Furthermore, both w_N and $\widetilde{w}_N$ have fluctuations precisely of order $N^{-1/2}$ in $E \in [-2, 2]$, while outside of this interval only w_N has fluctuations of precisely order $N^{-1/2}$ and $\widetilde{w}_N$ has strictly smaller fluctuations.

6.3 Variance Computation

In this section we prove Theorem 6.2.1 in the sense of mean and variance. The proof for higher moments and the convergence of distribution will be given in Section 6.4. We first introduce a commonly used (see, e.g., [73]) notion of high-probability bound which helps in keeping the notation compact.

Definition 6.3.1 (Stochastic Domination)**.** *If*

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

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

$$\sup_{u \in U^{(N)}} \mathbf{P}\left[X^{(N)}(u) > N^{\epsilon} Y^{(N)}(u)\right] \le N^{-D}$$

for large enough $N \ge N_0(\epsilon, D)$. In this case we use the notation $X \prec Y$. Moreover, if we have $|X| \prec Y$, we also write $X = \mathcal{O}_{\prec}(Y)$.

It can be checked (see [73, Lemma 4.4]) that $\prec$ satisfies the usual arithmetic properties, e.g. if $X_1 \prec Y_1$ and $X_2 \prec Y_2$, then also $X_1 + X_2 \prec Y_1 + Y_2$ and $X_1 X_2 \prec Y_1 Y_2$. We will say that a (sequence of) events $A = A^{(N)}$ holds with *overwhelming probability* if $\mathbf{P}(A^{(N)}) \ge 1 - N^{-D}$ for any $D > 0$ and $N \ge N_0(D)$. In particular, under the conditions (6.6), we have $h_{ij} \prec N^{-1/2}$ and $\max_k |\lambda_k| \le 3$ with overwhelming probability.

Let $\chi \colon \mathbb{R} \to \mathbb{R}$ be a smooth cut-off function which is constant 1 inside $[-5, 5]$ and constant 0 outside $[-10, 10]$. Now define

$$f_\chi(x) := f(x)\chi(x)$$

and its almost analytic extension

$$f_{\mathbb{C}}(x + i\eta) := \left[f_\chi(x) + i\eta f_\chi'(x)\right] \chi(\eta).$$

Clearly, $f_{\mathbb{C}}$ is bounded and compactly supported. Then,

$$\begin{aligned} \partial_{\bar{z}} f_{\mathbb{C}}(x + i\eta) &= \frac{1}{2}\frac{\partial}{\partial x} f_{\mathbb{C}}(x + i\eta) + \frac{i}{2}\frac{\partial}{\partial \eta} f_{\mathbb{C}}(x + i\eta) \\ &= \frac{i\eta}{2}\chi(\eta) f_\chi''(x) + \frac{i}{2}\chi'(\eta)\left[f_\chi(x) + i\eta f_\chi'(x)\right] \end{aligned}$$

and we note that for small η,

$$\partial_{\bar{z}} f_{\mathbb{C}}(x + i\eta) = \mathcal{O}(\eta) \quad \text{and} \quad \partial_\eta \partial_{\bar{z}} f_{\mathbb{C}}(x + i\eta) = \mathcal{O}(1). \tag{6.10}$$

For real λ we have

$$f_\chi(\lambda) = \frac{1}{2i\pi}\int_{\mathbb{C}} \frac{\partial_{\bar{z}} f_{\mathbb{C}}(z)}{\lambda - z}\, \mathrm{d}\bar{z} \wedge \mathrm{d}z = \frac{1}{\pi}\int_{\mathbb{R}^2} \frac{\partial_{\bar{z}} f_{\mathbb{C}}(x + i\eta)}{\lambda - x - i\eta}\, \mathrm{d}x\, \mathrm{d}\eta$$

whenever $f \in C^2(\mathbb{R})$, as follows from Cauchy's Theorem. Since the left hand side of this equality is real, it suffices to integrate the real part of the integrand on the right hand side which conveniently is symmetric with respect to the real axis. Consequently,

$$f_\chi(\lambda) = \frac{2}{\pi}\Re \int_{\mathbb{R}}\int_{\mathbb{R}_+} \frac{\partial_{\bar{z}} f_{\mathbb{C}}(x+i\eta)}{\lambda - x - i\eta}\,\mathrm{d}\eta\,\mathrm{d}x. \tag{6.11}$$

Eq. (6.11) is commonly known as the *Helffer-Sjöstrand formula* [101]. One can easily check that eq. (6.11) extends to $H^2(\mathbb{R})$ functions. The cut-off was chosen in such a way that with overwhelming probability $f(\lambda_k) = f_{\mathbb{C}}(\lambda_k)$ and $f(\lambda_k - h_{11}) = f_{\mathbb{C}}(\lambda_k - h_{11})$ and therefore eq. (6.11) yields

$$f_N = \frac{2}{\pi}\Re \int_{\mathbb{R}}\int_{\mathbb{R}_+} \partial_{\bar{z}} f_{\mathbb{C}}(x+i\eta)\left[\operatorname{Tr} G(x+i\eta) - \operatorname{Tr}\widehat{G}(x+i\eta)\right]\mathrm{d}\eta\,\mathrm{d}x \tag{6.12}$$

and

$$\widetilde{f}_N = \frac{2}{\pi}\Re \int_{\mathbb{R}}\int_{\mathbb{R}_+} \partial_{\bar{z}} f_{\mathbb{C}}(x+i\eta)\left[\operatorname{Tr} G(x+h_{11}+i\eta) - \operatorname{Tr}\widehat{G}(x+h_{11}+i\eta)\right]\mathrm{d}\eta\,\mathrm{d}x, \tag{6.13}$$

where for convenience we defined

$$H = \begin{pmatrix} h_{11} & h^* \\ h & \widehat{H}\end{pmatrix},\quad H^{(1)} = \begin{pmatrix} 0 & 0 \\ 0 & \widehat{H}\end{pmatrix},$$
$$G(z) := (H-z)^{-1},\quad \widehat{G}(z) := (\widehat{H}-z)^{-1},\quad G^{(1)}(z) := (H^{(1)}-z)^{-1}.$$

We also introduce the short hand notations

$$\Delta_N(z) := \operatorname{Tr} G(z) - \operatorname{Tr}\widehat{G}(z) \quad\text{and}\quad \widetilde{\Delta}_N(z) := \operatorname{Tr} G(z+h_{11}) - \operatorname{Tr}\widehat{G}(z+h_{11}).$$

From the Schur complement formula we find

$$\Delta_N(z) = \frac{1 + \langle h, \widehat{G}(z)^2 h\rangle}{h_{11} - z - \langle h, \widehat{G}(z) h\rangle} \quad\text{and}\quad \widetilde{\Delta}_N(z) = \frac{1 + \langle h, \widehat{G}(z+h_{11})^2 h\rangle}{-z - \langle h, \widehat{G}(z+h_{11}) h\rangle}. \tag{6.14}$$

The basic strategy now is that we identify the leading order behavior of these two expressions and then handle the fluctuations separately. To do so, we firstly exclude a critical area very close to the real line. Since

$$\left|\eta + \eta\,\langle h, \widehat{G}(x+i\eta)^2 h\rangle\right| \le \eta + \Im\,\langle h, \widehat{G}(x+i\eta) h\rangle \le \left|x_0 + x + i\eta + \langle h, \widehat{G}(x+i\eta) h\rangle\right|$$

for any $x_0 \in \mathbb{R}$ we find that

$$|\eta\,\Delta_N(x+i\eta)| \le 1 \quad\text{and}\quad \left|\eta\,\widetilde{\Delta}_N(x+i\eta)\right| \le 1$$

for all $\eta > 0$. Therefore we can restrict our integrations in (6.12)–(6.13) to the domain $\Im z > \eta_0 := N^{-2/3}$ and find that

$$f_N = \frac{2}{\pi}\Re \int_{\mathbb{R}}\int_{\eta_0}^{10} \partial_{\bar{z}} f_{\mathbb{C}}(x+i\eta)\Delta_N(x+i\eta)\,\mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_\prec\left(N^{-2/3}\right)$$

and

$$\widetilde{f}_N = \frac{2}{\pi}\Re\int_{\mathbb{R}}\int_{\eta_0}^{10} \partial_{\bar{z}} f_{\mathbb{C}}(x+i\eta)\widetilde{\Delta}_N(x+i\eta)\,\mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_{\prec}\left(N^{-2/3}\right).$$

For $\Im z = \eta \geq \eta_0$ we claim that the leading order of Δ_N and $\widetilde{\Delta}_N$ is given by

$$\widehat{\Delta}_N(z) := \frac{1+\frac{1}{N}\operatorname{Tr}\widehat{G}(z)^2}{-z-\frac{1}{N}\operatorname{Tr}\widehat{G}(z)}. \tag{6.15}$$

Accordingly, we split the proof effectively into two parts. We define

$$\widehat{\Omega}_f := \frac{2}{\pi}\Re\int_{\mathbb{R}}\int_{\eta_0}^{10} \partial_{\bar{z}} f_{\mathbb{C}}(x+i\eta)\widehat{\Delta}_N(x+i\eta)\,\mathrm{d}\eta\,\mathrm{d}x \tag{6.16}$$

and

$$\begin{aligned} F_N &:= \frac{2}{\pi}\Re\int_{\mathbb{R}}\int_{\eta_0}^{10} \partial_{\bar{z}} f_{\mathbb{C}}(x+i\eta)\left[\Delta_N(x+i\eta)-\widehat{\Delta}_N(x+i\eta)\right]\mathrm{d}\eta\,\mathrm{d}x \qquad (6.17)\\ \widetilde{F}_N &:= \frac{2}{\pi}\Re\int_{\mathbb{R}}\int_{\eta_0}^{10} \partial_{\bar{z}} f_{\mathbb{C}}(x+i\eta)\left[\widetilde{\Delta}_N(x+i\eta)-\widehat{\Delta}_N(x+i\eta)\right]\mathrm{d}\eta\,\mathrm{d}x, \end{aligned}$$

so that

$$f_N = \widehat{\Omega}_f + F_N + \mathcal{O}_{\prec}\left(N^{-2/3}\right) \quad \text{and} \quad \widetilde{f}_N = \widehat{\Omega}_f + \widetilde{F}_N + \mathcal{O}_{\prec}\left(N^{-2/3}\right).$$

Proposition 6.3.2 (Leading Order). *Under the assumptions of Theorem 6.2.1 we have that*

$$\widehat{\Omega}_f = \Omega_f + \mathcal{O}_{\prec}\left(N^{-2/3}\right).$$

Proposition 6.3.3 (Fluctuations). *Under the assumptions of Theorem 6.2.1 we have that*

$$\mathbf{E}\,F_N^2 = \frac{1}{N}V_f + \mathcal{O}_{\prec}\left(N^{-7/6}\right) \quad \textit{and} \quad \mathbf{E}\,\widetilde{F}_N^2 = \frac{1}{N}\widetilde{V}_f + \mathcal{O}_{\prec}\left(N^{-7/6}\right).$$

Note that the error terms in these propositions are deterministic and hence could also be written as $\mathcal{O}\left(N^{-2/3+\epsilon}\right)$ or $\mathcal{O}\left(N^{-7/6+\epsilon}\right)$ for any $\epsilon > 0$, respectively, but for simplicity we keep the $\mathcal{O}_{\prec}(\ldots)$ notation for deterministic quantities as well.

The positivity of V_f and $\widetilde{V}_f$ defined (6.8)–(6.9) follows from $1 = \sigma_2 \leq \sigma_4$ and from simple Schwarz inequalities

$$\begin{aligned} \left(\int_{-2}^{2} x\rho(x)f'(x)\,\mathrm{d}x\right)^2 &\leq \int_{-2}^{2}\rho(x)f'(x)^2\,\mathrm{d}x,\\ \left(\int_{-2}^{2} x\rho(x)\left(f'(x)-\int\rho f'\right)\mathrm{d}x\right)^2 &\leq \left(\int_{-2}^{2}\rho(x)\left(f(x)-\int\rho f'\right)^2\rho(x)\,\mathrm{d}x\right), \end{aligned}$$

using that the semicircle density ρ is symmetric and $\int x^2\rho(x)\,\mathrm{d}x = 1$.

6.3.1 Leading Order Integral

This section is devoted to the proof of Proposition 6.3.2. We rely on the local semicircle law in the averaged form (see [82] or [73, Theorem 2.3])

$$m_N(z) := \frac{1}{N} \operatorname{Tr} \widehat{G}(z) = m(z) + \mathcal{O}_\prec\left(\frac{1}{N\eta}\right) \tag{6.18}$$

and the entry-wise form

$$\widehat{G}_{ij}(z) - \delta_{ij} m(z) \prec \frac{1}{\sqrt{N\eta}} \tag{6.19}$$

which holds true for all $\eta = \Im z > \eta_0$. Here $m(z)$ is the Stieltjes transform of the semicircle distribution, i.e.,

$$m(z) := \int_{-2}^{2} \frac{1}{x-z} \rho(x) \, \mathrm{d}x = \frac{1}{2\pi} \int_{-2}^{2} \frac{\sqrt{4-x^2}}{x-z} \, \mathrm{d}x = \frac{-z + \sqrt{z^2-4}}{2},$$

where we chose the branch of the square root with positive imaginary part. Note that $\widehat{G}$ is an $(N-1) \times (N-1)$ matrix but we still normalize its trace by $1/N$ to define m_N; this unconventional notation will simplify some formulas later. Strictly speaking, the sum of the variances in each row of $\widehat{H}$ is not exactly one as required in [82, 73], partly due to the removal of one column and partly due to the relaxed bound $\mathbf{E} |h_{ii}|^2 \leq C/N$ on the diagonal elements. Nevertheless, we still have $\sum_{i=2}^N \mathbf{E} |h_{ij}|^2 = 1 + \mathcal{O}\left(N^{-1}\right)$ for each $j = 2, 3, \ldots, N$ and the proof of [73, Theorem 2.3] goes through. The only small change is that the $\mathcal{O}\left(N^{-1}\right)$ error term above gives rise to an additional term of size $\mathcal{O}\left(1/N\eta\right)$ in the definition of Υ_i in (5.7)-(5.8) of [73] using the trivial bound $|v_i| = |G_{ii} - m| \leq 2/\eta$ with the notation of that paper. Since the error bound on Υ_i used in that proof is bigger than $\mathcal{O}\left(1/N\eta\right)$, see [73, Lemma 5.2], the rest of the proof is unchanged.

Thus

$$\widehat{\Delta}_N(z) = \frac{1 + \frac{1}{N} \operatorname{Tr} \widehat{G}(z)^2}{-z - m(z)} + \mathcal{O}_\prec\left(\frac{1}{N\eta}\right) = m(z) \left[1 + \frac{1}{N} \operatorname{Tr} \widehat{G}(z)^2\right] + \mathcal{O}_\prec\left(\frac{1}{N\eta}\right),$$

where we used the relation $m(z) = 1/(-z - m(z))$. Since $\partial_{\bar{z}} f_{\mathbb{C}}(x + i\eta) = \mathcal{O}\left(\eta\right)$ for small η the error term, when inserted in (6.16) only gives a contribution of $1/N$. Thus eq. (6.16) becomes

$$\widehat{\Omega}_f = \frac{2}{\pi} \Re \int_{\mathbb{R}} \int_{\eta_0}^{10} \partial_{\bar{z}} f_{\mathbb{C}}(z) m(z) \left[1 + \frac{1}{N} \operatorname{Tr} \widehat{G}(z)^2\right] \mathrm{d}\eta \, \mathrm{d}x + \mathcal{O}_\prec\left(N^{-1}\right),$$

where from now on we shall always use the shorthand notation $z = x + i\eta$. Noting that

$$1 + \frac{1}{N} \operatorname{Tr} \widehat{G}(x+i\eta)^2 = \partial_\eta \left[\eta - i \frac{1}{N} \operatorname{Tr} \widehat{G}(x+i\eta)\right] = \partial_\eta \left[\eta - i m_N(x+i\eta)\right],$$

and $-i\partial_\eta h(z) = \partial_z h(z)$ for analytic h, we can now perform an integration by parts to find

$$\begin{aligned}
\widehat{\Omega}_f &= \frac{2}{\pi}\Re\int_{\mathbb{R}} \partial_{\bar z} f_{\mathbb{C}}(z_0) m(z_0)\,[\eta - i m_N(z_0)]\,\mathrm{d}x \\
&\quad - \frac{2}{\pi}\Re\int_{\mathbb{R}}\int_{\eta_0}^{10} \partial_\eta\left(\partial_{\bar z} f_{\mathbb{C}}(z) m(z)\right)[\eta - i m_N(z)]\,\mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_\prec\left(N^{-1}\right) \\
&= \frac{2}{\pi}\Re\int_{\mathbb{R}} \partial_{\bar z} f_{\mathbb{C}}(z_0) m(z_0)\,[\eta - i m(z_0)]\,\mathrm{d}x \\
&\quad - \frac{2}{\pi}\Re\int_{\mathbb{R}}\int_{\eta_0}^{10} \partial_\eta\left(\partial_{\bar z} f_{\mathbb{C}}(z) m(z)\right)[\eta - i m(z)]\,\mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_\prec\left(N^{-1}\right) \\
&\quad + \frac{2}{\pi}\Re\int_{\mathbb{R}} \partial_{\bar z} f_{\mathbb{C}}(z_0) m(z_0) i\,[m_N(z_0) - m(z_0)]\,\mathrm{d}x \\
&\quad - \frac{2}{\pi}\Re\int_{\mathbb{R}}\int_{\eta_0}^{10} \partial_\eta\left(\partial_{\bar z} f_{\mathbb{C}}(z) m(z)\right) i\,[m_N(z) - m(z)]\,\mathrm{d}\eta\,\mathrm{d}x \\
&= \frac{2}{\pi}\Re\int_{\mathbb{R}}\int_{\eta_0}^{10} \partial_{\bar z} f_{\mathbb{C}}(z) m(z)\,[1 + m'(z)]\,\mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_\prec\left(N^{-1}\right) + \mathcal{O}_\prec\left(N^{-1}\lvert\log\eta_0\rvert\right)
\end{aligned}$$

where we used that $\partial_{\bar z} f_{\mathbb{C}}(x + i\eta)$ scales like η near the real axis and the local semicircle law from eq. (6.18). For the main term we need the following simple lemma.

Lemma 6.3.4. *Let $\phi, \psi\colon [-10, 10] \times [0, 10i] \to \mathbb{C}$ be functions such that $\partial_{\bar z}\psi(z) \equiv 0$, $\phi, \psi \in H^1$ and ϕ vanishes at the left, right and top boundary of the integration region. Then for any $\eta_0 \in [0, 10]$, we have*

$$\int_{-10}^{10}\int_{\eta_0}^{10} [\partial_{\bar z}\phi(z)]\psi(z)\,\mathrm{d}\eta\,\mathrm{d}x = \frac{1}{2i}\int_{-10}^{10} \phi(x + i\eta_0)\psi(x + i\eta_0)\,\mathrm{d}x, \qquad z = x + i\eta.$$

Proof. This follows from the computation

$$\begin{aligned}
\int_{-10}^{10}\int_{\eta_0}^{10} [\partial_{\bar z}\phi(z)]\psi(z)\,\mathrm{d}\eta\,\mathrm{d}x &= \frac{1}{2i}\int_{-10}^{10}\int_{\eta_0}^{10} [\partial_{\bar z}\phi(z)]\psi(z)\,\mathrm{d}\bar z \wedge \mathrm{d}z \\
&= \frac{1}{2i}\int_{-10}^{10}\int_{\eta_0}^{10} \mathrm{d}(\phi(z)\psi(z)\,\mathrm{d}z) = \frac{1}{2i}\int_{-10}^{10} \phi(x + i\eta_0)\psi(x + i\eta_0)\,\mathrm{d}x,
\end{aligned}$$

where we used Stokes' Theorem in the ultimate step. □

We apply this together with $\Im m(x)[1 + m'(x)] = (4 - x^2)^{-1/2}$ and (6.10) to extend the integration to the real axis and conclude that

$$\begin{aligned}
\widehat{\Omega}_f &= \frac{2}{\pi}\Re\int_{\mathbb{R}}\int_{\eta_0}^{10} \partial_{\bar z} f_{\mathbb{C}}(z)\widehat{\Delta}_N(z)\,\mathrm{d}\eta\,\mathrm{d}x \\
&= \Omega_f + \mathcal{O}_\prec\left(N^{-2/3}\right) = \frac{1}{\pi}\int_{-2}^{2} \frac{f(x)}{\sqrt{4 - x^2}}\,\mathrm{d}x + \mathcal{O}_\prec\left(N^{-2/3}\right),
\end{aligned}$$

completing the proof of Proposition 6.3.2.

6.3.2 Fluctuation Integral

We now turn to the proof of Proposition 6.3.3. We formulate the main estimate as a lemma:

Lemma 6.3.5. *For any $\eta > \eta_0$ we have that*

$$\Delta_N(z) - \widehat{\Delta}_N(z) = \partial_z \frac{\langle h, \widehat{G}(z)h\rangle - m_N(z) - h_{11}}{-z - m_N(z)} + \mathcal{O}_\prec\left(\frac{1}{N\eta^2}\right) \tag{6.20}$$

and

$$\widetilde{\Delta}_N(z) - \widehat{\Delta}_N(z) = \partial_z \frac{\langle h, \widehat{G}(z+h_{11})h\rangle - m_N(z)}{-z - m_N(z)} + \mathcal{O}_\prec\left(\frac{1}{N\eta^2}\right). \tag{6.21}$$

Proof. This lemma relies on the following large deviation bound (see, e.g. [73, Theorem C.1])

$$\langle h, Ah\rangle = \frac{1}{N}\operatorname{Tr} A + \mathcal{O}_\prec\left(\frac{1}{N}\sqrt{\operatorname{Tr}|A|^2}\right). \tag{6.22}$$

To prove eq. (6.20) we write the difference $\Delta_N - \widehat{\Delta}_N$ from (6.14) and (6.15) as

$$\frac{(-z - m_N(z))\left(\langle h, \widehat{G}(z)^2 h\rangle - m_N'(z)\right) - (-1 - m_N'(z))\left(\langle h, \widehat{G}(z)h\rangle - m_N(z) - h_{11}\right)}{(-z - m_N(z))^2 - (-z - m_N(z))\left(\langle h, \widehat{G}(z)h\rangle - m_N(z) - h_{11}\right)}.$$

Now it follows from eq. (6.22) and (6.18) that

$$\langle h, \widehat{G}(z)h\rangle - m_N(z) \prec \frac{1}{N}\sqrt{\operatorname{Tr}\left|\widehat{G}(z)\right|^2} \le \frac{1}{N}\sqrt{\frac{1}{\eta}\Im \operatorname{Tr}\widehat{G}(z)} \prec \frac{1}{\sqrt{N\eta}} \tag{6.23}$$

and also

$$\langle h, \widehat{G}(z)^2 h\rangle - m_N'(z) \prec \frac{1}{N}\sqrt{\operatorname{Tr}\left|\widehat{G}(z)\right|^4} \le \frac{1}{N\eta}\sqrt{\operatorname{Tr}\left|\widehat{G}(z)\right|^2} \prec \frac{1}{\sqrt{N\eta^3}}.$$

We can therefore conclude that $\Delta_N(z) - \widehat{\Delta}_N(z)$ can be estimated as

$$\partial_z \frac{\langle h, \widehat{G}(z)h\rangle - m_N(z) - h_{11}}{-z - m_N(z)} + \mathcal{O}_\prec\left(\frac{1}{N\eta^2}\right).$$

The proof of eq. (6.21) is identical and shall be omitted. □

We now use eq. (6.20) to start estimating the fluctuations F_N of f_N as defined in eq. (6.17) via an integration by parts (with $z_0 = x + i\eta_0$)

$$\begin{aligned} F_N = &-\frac{2}{\pi}\Re\int_{\mathbb{R}} \partial_{\bar z} f_{\mathbb{C}}(z_0) i \frac{\langle h, \widehat{G}(z_0)h\rangle - m_N(z_0) - h_{11}}{-z_0 - m_N(z_0)} \,\mathrm{d}x \\ &+ \frac{2}{\pi}\Re\int_{\mathbb{R}}\int_{\eta_0}^{10} \partial_\eta \partial_{\bar z} f_{\mathbb{C}}(z) i \frac{\langle h, \widehat{G}(z)h\rangle - m_N(z) - h_{11}}{-z - m_N(z)} \,\mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_\prec\left(\frac{-\log\eta_0}{N}\right) \end{aligned}$$

and continue with the estimate

$$\frac{\langle h, \widehat{G}(z)h\rangle - m_N(z) - h_{11}}{-z - m_N(z)} = \frac{\langle h, \widehat{G}(z)h\rangle - m_N(z) - h_{11}}{-z - m(z)} + \mathcal{O}_\prec\left(\frac{1}{(N\eta)^{3/2}}\right)$$

from (6.23) and (6.18) to find that

$$\begin{aligned} F_N = &- \frac{2}{\pi}\Re \int_{\mathbb{R}} m(z_0)\partial_{\bar z} f_{\mathbb{C}}(z_0) i \left[\langle h, \widehat{G}(z_0)h\rangle - m_N(z_0) - h_{11}\right] \mathrm{d}x \qquad (6.24)\\ &+ \frac{2}{\pi}\Re \int_{\mathbb{R}}\int_{\eta_0}^{10} m(z)\partial_\eta\partial_{\bar z} f_{\mathbb{C}}(z) i \left[\langle h, \widehat{G}(z)h\rangle - m_N(z) - h_{11}\right] \mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_\prec\left(N^{-2/3}\right)\\ =&- \frac{2}{\pi}\Im \int_{\mathbb{R}}\int_{\eta_0}^{10} m(z)\partial_\eta\partial_{\bar z} f_{\mathbb{C}}(z) \left[\langle h, \widehat{G}(z)h\rangle - m_N(z) - h_{11}\right] \mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_\prec\left(N^{-2/3}\right), \end{aligned}$$

where we used in the last step that

$$\left|\partial_{\bar z} f_{\mathbb{C}}(z_0)\left[\langle h, \widehat{G}(z_0)h\rangle - m_N(z_0) - h_{11}\right]\right| \prec \sqrt{\frac{\eta_0}{N}} \le N^{-2/3}$$

from (6.23) and (6.10). Similarly one finds that

$$\begin{aligned} \widetilde{F}_N :=& \frac{2}{\pi}\Re \int_{\mathbb{R}}\int_{\eta_0}^{10} \partial_{\bar z} f_{\mathbb{C}}(z)[\widetilde{\Delta}_N(z) - \widehat{\Delta}_N(z)]\,\mathrm{d}\eta\,\mathrm{d}x \qquad (6.25)\\ =& \frac{2}{\pi}\Re \int_{\mathbb{R}}\int_{\eta_0}^{10} m(z)\partial_\eta\partial_{\bar z} f_{\mathbb{C}}(z) i \left[\langle h, \widehat{G}(z+h_{11})h\rangle - m_N(z)\right] \mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_\prec\left(N^{-2/3}\right)\\ =&\frac{2}{\pi}\Re \int_{\mathbb{R}}\int_{\eta_0}^{10} m(z)\partial_\eta\partial_{\bar z} f_{\mathbb{C}}(z) i \Big[\langle h, \widehat{G}(z+h_{11})h\rangle - m_N(z+h_{11})\\ &\qquad + m(z+h_{11}) - m(z)\Big] \mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_\prec\left(N^{-2/3}\right)\\ =& -\frac{2}{\pi}\Im \int_{\mathbb{R}}\int_{\eta_0}^{10} m(z)\partial_\eta\partial_{\bar z} f_{\mathbb{C}}(z) \Big[\langle h, \widehat{G}(z+h_{11})h\rangle - m_N(z+h_{11})\\ &\qquad + h_{11}m'(z)\Big] \mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_\prec\left(N^{-2/3}\right) \end{aligned}$$

where in the penultimate step we used the local semicircle law (6.18) and integrated the error term $(N\eta)^{-1}$ at an expense of $N^{-1}|\log\eta_0|$ and in the last step estimated

$$m(z+h_{11}) - m(z) = h_{11}m'(z) + \mathcal{O}_\prec\left(\frac{1}{\eta^{3/2}N}\right),$$

where the error term, after integration, contributes an error of at most $N^{-2/3}$.

Both fluctuation estimates from eqs. (6.24) and (6.25) have two convenient properties: Firstly, the leading order expressions for F_N and $\widetilde{F}_N$ have zero mean and secondly, the fluctuations in them stemming from h_{11} and the ones from h and $\widehat{G}(z)$ can be separated. Indeed,

$$\begin{aligned} &\mathbf{E}\left[\langle h, \widehat{G}(z+h_{11})h\rangle - m_N(z+h_{11}) + h_{11}m'(z)\right]^2\\ &\quad = \mathbf{E}\left[\langle h, \widehat{G}(z+h_{11})h\rangle - m_N(z+h_{11})\right]^2 + \mathbf{E}\left[h_{11}m'(z)\right]^2 \end{aligned}$$

since the expectation with respect to h, conditioned on h_{11} of the first term on the rhs. is zero and h and h_{11} are independent. Similarly,

$$\mathbf{E}\left[\langle h, \widehat{G}(z)h\rangle - m_N(z) - h_{11}\right]^2 = \mathbf{E}\left[\langle h, \widehat{G}(z)h\rangle - m_N(z)\right]^2 + \mathbf{E}\left[h_{11}\right]^2.$$

Therefore we can start computing the variances as

$$\begin{aligned}\mathbf{E}\, F_N^2 = \mathbf{E}\left(\frac{2}{\pi}\Im\int_{\mathbb{R}}\int_{\eta_0}^{10} m(z)\partial_\eta\partial_{\bar z} f_{\mathbb{C}}(z)\left[\langle h, \widehat{G}(z)h\rangle - m_N(z)\right]\mathrm{d}\eta\,\mathrm{d}x\right)^2 \\ + \frac{s_{11}}{N}\left(\frac{2}{\pi}\Im\int_{\mathbb{R}}\int_0^{10} m(z)\partial_\eta\partial_{\bar z} f_{\mathbb{C}}(z)\,\mathrm{d}\eta\,\mathrm{d}x\right)^2 + \mathcal{O}_\prec\left(N^{-7/6}\right)\end{aligned} \tag{6.26}$$

and

$$\begin{aligned}\mathbf{E}\, \widetilde{F}_N^2 = \mathbf{E}\left(\frac{2}{\pi}\Im\int_{\mathbb{R}}\int_{\eta_0}^{10} m(z)\partial_\eta\partial_{\bar z} f_{\mathbb{C}}(z)\left[\langle h, \widehat{G}(z+h_{11})h\rangle - m_N(z+h_{11})\right]\mathrm{d}\eta\,\mathrm{d}x\right)^2 \\ + \frac{s_{11}}{N}\left(\frac{2}{\pi}\Im\int_{\mathbb{R}}\int_0^{10} m(z)m'(z)\partial_\eta\partial_{\bar z} f_{\mathbb{C}}(z)\,\mathrm{d}\eta\,\mathrm{d}x\right)^2 + \mathcal{O}_\prec\left(N^{-7/6}\right).\end{aligned} \tag{6.27}$$

Note that in the second terms we extended the integration domain of η starting from 0 instead of η_0 at a negligible error. The second terms are already deterministic and explicitly computable using Lemma 6.3.4 and they give rise to the integral coefficients in (6.7). When taking expectations, we frequently use the property that if $X = \mathcal{O}_\prec(Y), Y \geq 0$ and $|X| \leq N^C$ for some constant C, then $\mathbf{E}\,|X| \prec \mathbf{E}\,Y$, or, equivalently, $\mathbf{E}\,|X| \leq N^\epsilon\,\mathbf{E}\,Y$ for any $\epsilon > 0$ and $N \geq N_0(\epsilon)$.

For the first term we introduce short-hand notations

$$g(z) := \frac{2}{\pi} m(z)\partial_\eta\partial_{\bar z} f_{\mathbb{C}}(z), \qquad X(z) := \sqrt{N}\left[\langle h, \widehat{G}(z)h\rangle - m_N(z)\right] \tag{6.28}$$

to write

$$\begin{aligned}F_N' &:= \frac{1}{\sqrt{N}}\mathbf{E}\left(\Im\int_{\mathbb{R}}\int_{\eta_0}^{10} g(z)X(z)\,\mathrm{d}\eta\,\mathrm{d}x\right)^2, \\ \widetilde{F}_N' &:= \frac{1}{\sqrt{N}}\mathbf{E}\left(\Im\int_{\mathbb{R}}\int_{\eta_0}^{10} g(z)X(z+h_{11})\,\mathrm{d}\eta\,\mathrm{d}x\right)^2.\end{aligned}$$

For complex numbers z, w we can expand

$$(\Im z)(\Im w) = \frac{1}{2}\Re\left[\bar{z}w - zw\right]$$

to write out

$$F_N' = \frac{1}{N}\frac{1}{2}\Re\iint\limits_{\mathbb{R}}\iint\limits_{\eta_0}^{10}\left[g(z)g(\bar{z}')\,\mathbf{E}\,X(z)X(\bar{z}') - g(z)g(z')\,\mathbf{E}\,X(z)X(z')\right]\mathrm{d}\eta\,\mathrm{d}\eta'\,\mathrm{d}x\,\mathrm{d}x' \tag{6.29}$$

where we used that $\overline{X(z)} = X(\bar{z})$ and $\overline{g(z)} = g(\bar{z})$. To work out the expectations, we expand

$$X(z)X(z') = N\left(\sum_{i\neq j}\overline{h_i}G_{ij}h_j + \sum_i\left[|h_i|^2 - \frac{1}{N}\right]G_{ii}\right)\left(\sum_{l\neq k}\overline{h_l}G_{lk}'h_k + \sum_l\left[|h_l|^2 - \frac{1}{N}\right]G_{ll}'\right)$$

where we introduced the shorthand notations

$$G = \widehat{G}(z), \qquad G' = \widehat{G}(z').$$

Note that we have redefined the notation G but it should not create any confusion since the full resolvent matrix $G(z)$ will not appear any more in the rest of the paper. To keep the notation simple we generally index the $(N-1)\times(N-1)$ matrices G, G' and the $(N-1)$ vector h by integers $\{2,\dots,N\}$. In particular, all sums involving G and G' run from 2 to N if not stated otherwise. We then compute the expectation $\mathbf{E}_1 = \mathbf{E}(\cdot|H^{(1)})$ conditioned on $H^{(1)}$ to find

$$\begin{aligned}
\mathbf{E}_1[X(z)X(z')] &= N\sum_{i\neq j}\Big(G_{ij}G'_{ji}\,\mathbf{E}\,|h_i|^2\,|h_j|^2 + G_{ij}G'_{ij}\,\mathbf{E}\,\overline{h_i^2}h_j^2\Big) \qquad (6.30)\\
&\quad + N\sum_i \mathbf{E}\left[|h_i|^2 - \frac{1}{N}\right]^2 G_{ii}G'_{ii}\\
&= \frac{1}{N}\sum_{i\neq j}\Big(G_{ij}G'_{ji} + |\sigma_2|^2\,G_{ij}G'_{ij}\Big) + \frac{\sigma_4-1}{N}\sum_i G_{ii}G'_{ii}\\
&= \frac{1}{N}\sum_{i\neq j}\Big(G_{ij}G'_{ji} + |\sigma_2|^2\,G_{ij}G'_{ij}\Big) + (\sigma_4-1)m(z)m(z')\\
&\quad + \mathcal{O}_\prec\left(\frac{1}{\sqrt{N\eta}} + \frac{1}{\sqrt{N\eta'}} + \frac{1}{N\sqrt{\eta\eta'}}\right),
\end{aligned}$$

where we recall that $\mathbf{E}\,h_{ij}^2 = \sigma_2/N$ for $i<j$ and $\mathbf{E}\,h_{ij} = \overline{\sigma_2}/N$ for $i>j$. For the computation of the first term we need a lemma:

Lemma 6.3.6. *Let $\eta,\eta'>0$. Then for z, z' with $|\Im z| = \eta$ and $|\Im z'| = \eta'$ it holds that*

$$\frac{1}{N}\sum_{i\neq j}G_{ij}G'_{ji} = \frac{m(z)^2m(z')^2}{1-m(z)m(z')} + \mathcal{O}_\prec\left(\frac{1}{(\eta+\eta')\sqrt{N\eta\eta'}}\left[\frac{1}{\sqrt{\eta}} + \frac{1}{\sqrt{\eta'}} + \frac{1}{\sqrt{N\eta\eta'}}\right]\right) \tag{6.31}$$

and

$$\begin{aligned}
\frac{1}{N}\sum_{i\neq j}G_{ij}G'_{ij} &= m(z)m(z')\frac{(1+m(z)m(z')\Re\sigma_2)\frac{\tan[m(z)m(z')\Im\sigma_2]}{m(z)m(z')\Im\sigma_2} - 1}{1-\Re\sigma_2\frac{\tan[m(z)m(z')\Im\sigma_2]}{\Im\sigma_2}} \qquad (6.32)\\
&\quad + \mathcal{O}_\prec\left(\frac{1}{(\eta+\eta')\sqrt{N\eta\eta'}}\left[\frac{1}{\sqrt{\eta}} + \frac{1}{\sqrt{\eta'}} + \frac{1}{\sqrt{N\eta\eta'}}\right]\right)
\end{aligned}$$

(if $\Im\sigma_2 = 0$, then we use the convention that $\tan x/x = 0$ *for $x=0$).*

We remark that the $(\eta+\eta')^{-1}$ factor in the error term can be substantially improved if $\Im z$ and $\Im z'$ has the same sign, see e.g. [70] for the special $z=z'$ case, but the same argument works in the general case.

Proof. The proof of this lemma follows the techniques used in [70]. We let $G^{(j)}$ denote the resolvent of the minor of $\widehat{H}$ after removing the j-th row and column. We have the resolvent identity

$$G_{ij} = -G_{ii}\sum_j^{(i)} G_{ik}^{(j)}h_{kj}, \qquad i\neq j,$$

where the summation runs over all $j = 2, 3, \dots, N$ except $j = i$; this exclusion is indicated with the upper index on the summation. Using the local semicircle law (6.18), we find that for any fixed i

$$\begin{aligned}
\frac{1}{N}\sum_j^{(i)} \mathbf{E}_j[G_{ij}G'_{ji}] &= \frac{1}{N}\sum_j^{(i)} \mathbf{E}_j\left[\frac{m(z)m(z')}{G_{jj}G'_{jj}}G_{ij}G'_{ji}\right] + \mathcal{O}_{\prec}(\Psi) \\
&= \frac{1}{N}m(z)m(z')\sum_j^{(i)} \mathbf{E}_j\left[\left(\sum_k^{(j)} G^{(j)}_{ik}h_{kj}\right)\left(\sum_l^{(j)} h_{jl}G'^{(j)}_{li}\right)\right] + \mathcal{O}_{\prec}(\Psi) \\
&= \frac{1}{N^2}m(z)m(z')\sum_j^{(i)}\sum_k^{(j)} G^{(j)}_{ik}G'^{(j)}_{ki} + \mathcal{O}_{\prec}(\Psi) \\
&= \frac{1}{N^2}m(z)m(z')\sum_j^{(i)}\left[\sum_k^{(ij)} G_{ik}G'_{ki} + G_{ii}G'_{ii}\right] + \mathcal{O}_{\prec}(\Psi) \\
&= \frac{1}{N}m(z)m(z')\left[\sum_k^{(i)} G_{ik}G'_{ki} + m(z)m(z')\right] + \mathcal{O}_{\prec}(\Psi)
\end{aligned}$$

where in the fourth equality we used

$$G^{(j)}_{ik} = G_{ik} - \frac{G_{ij}G_{jk}}{G_{jj}} = G_{ik} + \mathcal{O}_{\prec}\left(\frac{1}{N\eta}\right)$$

and the analogous identity for G' and we introduced the short hand notation

$$\Psi = \frac{1}{\sqrt{N^3\eta^2\eta'}} + \frac{1}{\sqrt{N^3\eta\eta'^2}} + \frac{1}{N^2\eta\eta'}$$

for the error term. We now follow the fluctuation averaging analysis from [70, Proof of Prop. 5.3 in Sections 6–7]. This proof was given for the case when the spectral parameters of the resolvents were identical, $z = z'$, but a simple inspection shows that the argument verbatim also applies to the $z \neq z'$ case. We conclude that

$$\frac{1}{N}\sum_j^{(i)} G_{ij}G'_{ji} = \frac{1}{N}\sum_j^{(i)} \mathbf{E}_j[G_{ij}G'_{ji}] + \mathcal{O}_{\prec}(\Psi)\,. \tag{6.33}$$

Therefore, after summing over i we have

$$[1 - m(z)m(z')]\,\frac{1}{N}\sum_{j\neq i} G_{ij}G'_{ji} = m(z)^2m(z')^2 + \mathcal{O}_{\prec}(N\Psi)\,.$$

To finish the proof, we note that by an elementary calculation

$$\frac{1}{|1 - m(z)m(z')|} \leq \frac{C}{\eta + \eta'}$$

since

$$|m(x + i\eta)| \leq 1 - c\,|\eta| \tag{6.34}$$

and therefore

$$\frac{1}{N}\sum_{i\neq j} G_{ij}G'_{ji} = \frac{m(z)^2 m(z')^2}{1-m(z)m(z')} + \mathcal{O}_\prec\left(\frac{N\Psi}{\eta+\eta'}\right).$$

This completes the proof of (6.31).

For the proof of eq. (6.32) we have to derive a vector self-consistent equation instead of the scalar one. We again start by noting that for $i \neq j$

$$\begin{aligned}\mathbf{E}_j\, G_{ij}G'_{ij} &= m(z)m(z')\sum_k^{(j)} G_{ik}G'_{ik}\,\mathbf{E}\,h_{kj}^2 + \mathcal{O}_\prec(\Psi)\\ &= m(z)m(z')\sum_k^{(i)} G_{ik}G'_{ik}\,\mathbf{E}\,h_{kj}^2 + m(z)^2m(z')^2\,\mathbf{E}\,h_{ij}^2 + \mathcal{O}_\prec(\Psi)\\ &= m(z)m(z')\sum_k^{(i)} F_{jk}\,\mathbf{E}_k\,G_{ik}G'_{ik} + m(z)^2m(z')^2F_{ji} + \mathcal{O}_\prec(\Psi),\end{aligned}$$

where we introduced the matrix F with matrix elements

$$F_{jk} := \mathbf{E}\,h_{kj}^2 = \frac{1}{N}\Big[\mathbb{1}(k<j)\sigma_2 + \mathbb{1}(k>j)\overline{\sigma_2} + \mathbb{1}(k=j)\Big].$$

For every fixed i, we have therefore derived a self-consistent equation for the (column) vector

$$v^{(i)} = \Big((1-\delta_{ij})\,\mathbf{E}_j\,G_{ij}G'_{ij}\Big)_{j=2}^N$$

which can be written as

$$\left[\mathbb{1} - m(z)m(z')F\right]v^{(i)} = m(z)^2m(z')^2\left[F - \frac{1}{N}\mathbb{1}\right]e_i + \mathcal{O}_\prec(\Psi),$$

where $e_i = (0,0,\dots 1,\dots 0)^T$ is the standard i-th basis vector of $\mathbb{C}^{N-1}$. To invert this equation while controlling the error term, we have estimate

$$\left\|\left[\mathbb{1} - m(z)m(z')F\right]^{-1}\right\|_{\ell^\infty\to\ell^\infty}.$$

To do so, we first note that

$$\left\|\left[\mathbb{1} - m(z)m(z')F\right]^{-1}\right\|_{\ell^2\to\ell^2} \le \left(1 - |m(z)|\,|m(z')|\,\|F\|_{\ell^2\to\ell^2}\right)^{-1} \le \frac{C}{\eta+\eta'},$$

where we used that F is Hermitian and of norm at most 1 and (6.34) (the norm here is induced by the usual ℓ^2 norm $\|u\|_2 := (\sum_i |u_i|^2)^{1/2}$ on $\mathbb{C}^{N-1}$). Next, if $(1-m(z)m(z')F)u = v$, then

$$\|u\|_\infty \le \|v\|_\infty + \|Fu\|_\infty \le \frac{C}{\eta+\eta'}\|v\|_\infty,$$

where we used

$$\begin{aligned}\|Fu\|_\infty &\le \frac{1}{N}\|u\|_1 \le \frac{1}{\sqrt{N}}\|u\|_2 = \frac{1}{\sqrt{N}}\left\|\left[\mathbb{1}-m(z)m(z')F\right]^{-1}v\right\|_2\\ &\le \frac{C}{\eta+\eta'}\frac{1}{\sqrt{N}}\|v\|_2 \le \frac{C}{\eta+\eta'}\|v\|_\infty,\end{aligned}$$

so that also

$$\left\| [\mathbb{1} - m(z)m(z')F]^{-1} \right\|_{\ell^\infty \to \ell^\infty} \leq \frac{C}{\eta + \eta'}$$

for $\eta, \eta' \leq C$. After inversion we find that

$$v^{(i)} = m(z)^2 m(z')^2 \left(\mathbb{1} - m(z)m(z')F\right)^{-1} \left(F - \frac{1}{N}\mathbb{1}\right) e_i + \mathcal{O}_\prec \left(\frac{\Psi}{\eta + \eta'}\right).$$

Using fluctuation averaging once more (see (6.33)) we can conclude that

$$\begin{aligned} \frac{1}{N}\sum_{i \neq j} G_{ij} G'_{ij} &= \frac{1}{N}\sum_{i \neq j} \mathbf{E}_j\, G_{ij} G'_{ij} + \mathcal{O}_\prec (N\Psi) \qquad (6.35) \\ &= m(z)^2 m(z')^2 e^T \left(\mathbb{1} - m(z)m(z')F\right)^{-1} \left(F - \frac{1}{N}\mathbb{1}\right) e + \mathcal{O}_\prec \left(\frac{N\Psi}{\eta + \eta'}\right), \end{aligned}$$

where $e = N^{-1/2}(1, \dots, 1)^T \in \mathbb{C}^{N-1}$. We now introduce the $(N-1) \times (N-1)$ matrix

$$S := \frac{1}{N} \begin{pmatrix} 0 & 1 & \dots & 1 \\ -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ -1 & \dots & -1 & 0 \end{pmatrix}.$$

Notice that $F = \frac{1}{N}\mathbb{1} + (\Re\sigma_2)(ee^T - \frac{1}{N}\mathbb{1}) + i(\Im\sigma_2)S$. We find, through an elementary computation, that

$$\begin{aligned} &m(z)m(z') \left\langle e, \left(\mathbb{1} - m(z)m(z')F\right)^{-1} \left(F - \frac{1}{N}\mathbb{1}\right) e\right\rangle \\ &\quad = \frac{(1 + m(z)m(z')\Re\sigma_2) \left\langle e, \left(\mathbb{1} - im(z)m(z')\Im\sigma_2 S\right)^{-1} e\right\rangle - 1}{1 - m(z)m(z')\Re\sigma_2 \left\langle e, \left(\mathbb{1} - im(z)m(z')\Im\sigma_2 S\right)^{-1} e\right\rangle} + \mathcal{O}_\prec \left(N^{-1}\right). \end{aligned}$$

It remains to compute

$$\langle e, (\mathbb{1} - \alpha S)^{-1} e\rangle = \sum_{k=0}^{\infty} \alpha^k \langle e, S^k e\rangle$$

for $\alpha \in \mathbb{C}$ with $|\alpha| < 1$. For any vector $f \in \mathbb{C}^{N-1}$,

$$(Sf)_n = -\frac{1}{N}\sum_{n' < n} f_{n'} + \frac{1}{N}\sum_{n' > n} f_{n'} = \frac{1}{N}\sum_{n'=2}^{N} h_{n-n'} f_{n'}$$

where $h_k := \mathbb{1}(k < 0) - \mathbb{1}(k > 0)$. Therefore

$$\begin{aligned} \langle e, S^k e\rangle &= N^{-1/2}\sum_{n=2}^{N} (S^k e)_n = N^{-3/2} \sum_{n,n'=2}^{N} h_{n-n'} (S^{k-1}e)_{n'} \\ &= \dots = N^{-k-1} \sum_{n_0, \dots, n_k = 2}^{N} h_{n_0 - n_1} \dots h_{n_{k-1} - n_k}. \end{aligned}$$

By symmetry, $\langle e, S^k e\rangle = 0$ for odd k. Otherwise one finds via a Riemann sum approximation that

$$\langle e, S^{2k} e\rangle = \int_0^1 \cdots \int_0^1 h(x_0 - x_1) \ldots h(x_{2k-1} - x_{2k}) \, \mathrm{d}x_0 \ldots \mathrm{d}x_{2k} + \mathcal{O}\left(N^{-1}\right),$$

where $h(x) = \mathbb{1}(x < 0) - \mathbb{1}(x > 0)$ is the Heaviside function and where we added the missing $n_i = 1$ terms at an expense of $\mathcal{O}\left(N^{-1}\right)$. Via an easy induction we see that

$$\langle e, S^{2k} e\rangle = (-1)^k \frac{2^{2k}(2^{2k}-1)}{(2k)!} B_{2k} + \mathcal{O}\left(N^{-1}\right),$$

where B_k is the k-th Bernoulli number. Consequently,

$$\langle e, (\mathbb{1} - \alpha S)^{-1} e\rangle = \frac{\tanh \alpha}{\alpha} + \mathcal{O}\left(N^{-1}\right).$$

We now use this with $\alpha = i m(z) m(z') \Im \sigma_2$ to conclude that

$$\begin{aligned} & m(z)m(z') \left\langle e, \left(\mathbb{1} - m(z)m(z')F\right)^{-1} \left(F - \frac{1}{N}\mathbb{1}\right) e\right\rangle \\ &= \frac{(1 + m(z)m(z')\Re\sigma_2)\frac{\tan[m(z)m(z')\Im\sigma_2]}{m(z)m(z')\Im\sigma_2} - 1}{1 - m(z)m(z')\Re\sigma_2\frac{\tan[m(z)m(z')\Im\sigma_2]}{m(z)m(z')\Im\sigma_2}} + \mathcal{O}_{\prec}\left(N^{-1}\right). \end{aligned}$$

Combining this with (6.35), we obtain

$$\frac{1}{N}\sum_{i \neq j} G_{ij} G'_{ij} = m(z)m(z') \frac{(1 + m(z)m(z')\Re\sigma_2)\frac{\tan[m(z)m(z')\Im\sigma_2]}{m(z)m(z')\Im\sigma_2} - 1}{1 - \Re\sigma_2 \frac{\tan[m(z)m(z')\Im\sigma_2]}{\Im\sigma_2}} + \mathcal{O}_{\prec}\left(\frac{N\Psi}{\eta + \eta'}\right).$$

We note that, in general, this is a finite expression since $|\Re\sigma_2| \le \sqrt{1 - (\Im\sigma_2)^2}$ and thus in the non-trivial case where $\Re\sigma_2 \neq 0$ and $\Im\sigma_2 \neq 0$,

$$\left|\Re\sigma_2 \frac{\tan[m(z)m(z')\Im\sigma_2]}{\Im\sigma_2}\right| < \left|\sqrt{1 - (\Im\sigma_2)^2}\frac{\tan[\Im\sigma_2]}{\Im\sigma_2}\right| \le 1. \qquad \square$$

We readily check that integrating the error terms in (6.29) from (6.30) and Lemma 6.3.6

only contributes an error of magnitude $N^{-7/6}$ and conclude that if $\sigma_2 = 0$, then

$$\begin{aligned}
F_N' &= \frac{1}{2N}\Re \iint\limits_{-10}^{10} \iint\limits_{\eta_0}^{10} g(z)g(\bar z')\left[\frac{m(z)^2 m(\bar z')^2}{1-m(z)m(\bar z')} + (\sigma_4 - 1)m(z)m(\bar z')\right] \\
&\qquad - g(z)g(z')\left[\frac{m(z)^2 m(z')^2}{1-m(z)m(z')} + (\sigma_4 - 1)m(z)m(z')\right] \mathrm{d}\eta\,\mathrm{d}\eta'\,\mathrm{d}x\,\mathrm{d}x' + \mathcal{O}_\prec\left(N^{-7/6}\right) \\
&= \frac{1}{2N}\Re \iint\limits_{-10}^{10} \iint\limits_{\eta_0}^{10} g(z)g(\bar z')\left[\sum_{k=2}^{\infty}[m(z)m(\bar z')]^k + (\sigma_4 - 1)m(z)m(\bar z')\right] \\
&\qquad - g(z)g(z')\left[\sum_{k=2}^{\infty}[m(z)m(z')]^k + (\sigma_4 - 1)m(z)m(z')\right] \mathrm{d}\eta\,\mathrm{d}\eta'\,\mathrm{d}x\,\mathrm{d}x' + \mathcal{O}_\prec\left(N^{-7/6}\right) \\
&= \frac{1}{N}\sum_{k=2}^{\infty}\left(\Im\int_{-10}^{10}\int_{\eta_0}^{10} g(z)m(z)^k\,\mathrm{d}\eta\,\mathrm{d}x\right)^2 \\
&\qquad + \frac{\sigma_4 - 1}{N}\left(\Im\int_{-10}^{10}\int_{\eta_0}^{10} g(z)m(z)\,\mathrm{d}\eta\,\mathrm{d}x\right)^2 + \mathcal{O}_\prec\left(N^{-7/6}\right) \\
&= \frac{1}{N}\sum_{k=2}^{\infty}\left(\frac{1}{\pi}\Im\int_{-10}^{10}\frac{\partial_\eta f_{\mathbb{C}}(z_0)}{i}m(z_0)^{k+1}\,\mathrm{d}x\right)^2 \\
&\qquad + \frac{\sigma_4 - 1}{N}\left(\frac{1}{\pi}\Im\int_{-10}^{10}\partial_x f_{\mathbb{C}}(z_0)m(z_0)^2\,\mathrm{d}x\right)^2 + \mathcal{O}_\prec\left(N^{-7/6}\right) \\
&= \frac{1}{N}\sum_{k=0}^{\infty}\left(\frac{1}{\pi}\Im\int_{-10}^{10}\frac{\partial_\eta f_{\mathbb{C}}(z_0)}{i}m(z_0)^{k+1}\,\mathrm{d}x\right)^2 + \frac{\sigma_4 - 2}{N}\left(\frac{1}{\pi}\Im\int_{-10}^{10} f'(x)m(x)^2\,\mathrm{d}x\right)^2 \\
&\qquad - \frac{1}{N}\left(\frac{1}{\pi}\Im\int_{-10}^{10} f'(x)m(x)\,\mathrm{d}x\right)^2 + \mathcal{O}_\prec\left(N^{-7/6}\right),
\end{aligned}$$

where $z_0 = x + i\eta_0$ and in the penultimate step we used Lemma 6.3.4 to write

$$\begin{aligned}
\Im\int_{-10}^{10}\int_{\eta_0}^{10} g(z)m(z)^k\,\mathrm{d}\eta\,\mathrm{d}x &= \frac{2}{\pi}\Im\int_{-10}^{10}\int_{\eta_0}^{10}[\partial_{\bar z}\partial_\eta f_{\mathbb{C}}(z)]m(z)^{k+1}\,\mathrm{d}\eta\,\mathrm{d}x \\
&= \Im\frac{1}{i\pi}\int_{-10}^{10}\partial_\eta f_{\mathbb{C}}(z_0)m(z_0)^{k+1}\,\mathrm{d}\eta\,\mathrm{d}x
\end{aligned}$$

and that

$$\frac{\partial_\eta f_{\mathbb{C}}(z_0)}{i} = \partial_x f_{\mathbb{C}}(z_0) + \mathcal{O}(\eta_0) = f'(x) + \mathcal{O}(\eta_0). \tag{6.36}$$

Now that we reduced the area integral to a line integral, we go the geometric series steps backwards to further simplify the first term as

$$\begin{aligned}
&\frac{1}{N}\sum_{k=0}^{\infty}\left(\frac{1}{\pi}\Im\int_{-10}^{10}\frac{\partial_\eta f_{\mathbb{C}}(z_0)}{i}m(z_0)^{k+1}\,\mathrm{d}x\right)^2 \qquad (6.37)\\
&= \frac{1}{2N\pi^2}\Re\iint\limits_{-10}^{10}\left[\left(\frac{\partial_\eta f_{\mathbb{C}}(z_0)}{i}\right)\overline{\left(\frac{\partial_\eta f_{\mathbb{C}}(z_0')}{i}\right)}\frac{m(z_0)\overline{m(z_0')}}{1-m(z_0)\overline{m(z_0')}}\right. \\
&\qquad\qquad \left. - \left(\frac{\partial_\eta f_{\mathbb{C}}(z_0)}{i}\right)\left(\frac{\partial_\eta f_{\mathbb{C}}(z_0')}{i}\right)\frac{m(z_0)m(z_0')}{1-m(z_0)m(z_0')}\right]\mathrm{d}x\,\mathrm{d}x'.
\end{aligned}$$

We would now like to approximate (6.37) using (6.36). For doing so, we have to control the error terms via the following lemma whose proof we postpone to the end of the section.

Lemma 6.3.7. *There exists an absolute constant C such that for $z_0 = x + i\eta_0$ and $z_0' = x' + i\eta_0$ with $0 < \eta_0 \leq 1/2$ it holds that*

$$\iint_{-10}^{10} \frac{1}{\left|1 - m(z_0)\overline{m(z_0')}\right|} \,\mathrm{d}x\,\mathrm{d}x' \leq C\left|\log \eta_0\right|, \qquad \iint_{-10}^{10} \frac{1}{\left|1 - m(z_0)m(z_0')\right|} \,\mathrm{d}x\,\mathrm{d}x' \leq C\left|\log \eta_0\right|. \tag{6.38}$$

Using Lemma 6.3.7 and (6.36) we can rewrite (6.37) as

$$\frac{1}{2N\pi^2}\Re \iint_{\mathbb{R}} f'(x)f'(x')\left[\frac{m(z_0)\overline{m(z_0')}}{1 - m(z_0)\overline{m(z_0')}} - \frac{m(z_0)m(z_0')}{1 - m(z_0)m(z_0')}\right]\mathrm{d}x\,\mathrm{d}x' + \mathcal{O}\left(\eta_0\left|\log \eta_0\right|\right).$$

Now, an explicit computation shows

$$\begin{aligned}&\Re\left[\frac{m(z_0)\overline{m(z_0')}}{1 - m(z_0)\overline{m(z_0')}} - \frac{m(z_0)m(z_0')}{1 - m(z_0)m(z_0')}\right] \\ &\quad = \Re\frac{-2i\Im m(z_0')}{-x - i\eta_0 - 2\Re m(z_0') + m(z_0)[|m(z_0')|^2 - 1]}.\end{aligned} \tag{6.39}$$

and therefore for small η_0 and (x, x') outside the square $[-2, 2]^2$ the integrand of (6.39) negligible. For $(x, x') \in [-2, 2]^2$ and small η_0 we have

$$\Re\frac{-2i\Im m(z_0')}{-x - i\widetilde{\eta} - 2\Re m(z_0') + m(z_0)[|m(z_0')|^2 - 1]} = \frac{\sqrt{4 - x'^2}\eta_0}{(x - x')^2 + \eta_0^2} + \mathcal{O}\left(\eta_0\right).$$

This expression acts like

$$\pi\sqrt{4 - x^2}\delta(x' - x)$$

for small η_0. More formally, it is well known that for any L^2-function h

$$\lim_{\eta\to 0}\int_{\mathbb{R}} \frac{\eta}{(x - x')^2 + \eta^2} h(x')\,\mathrm{d}x' = \pi h(x)$$

in L^2-sense. Working out an effective error term for $h \in H^1$, this allows us to conclude

$$\begin{aligned}F_N' = \frac{1}{N}\Bigg[&\int_{-2}^{2}\rho(x)f'(x)^2\,\mathrm{d}x - \left(\int_{-2}^{2}\rho(x)f'(x)\,\mathrm{d}x\right)^2 \\ &+ (\sigma_4 - 2)\left(\int_{-2}^{2}\rho(x)xf'(x)\,\mathrm{d}x\right)^2\Bigg] + \mathcal{O}_{\prec}\left(N^{-7/6}\right).\end{aligned}$$

The computation for $\widetilde{F}_N'$ from (6.27), still assuming $\sigma_2 = 0$, is completely analogous and there we also have

$$\begin{aligned}\widetilde{F}_N' = \frac{1}{N}\Bigg[&\int_{-2}^{2}\rho(x)f'(x)^2\,\mathrm{d}x - \left(\int_{-2}^{2}\rho(x)f'(x)\,\mathrm{d}x\right)^2 \\ &+ (\sigma_4 - 2)\left(\int_{-2}^{2}\rho(x)xf'(x)\,\mathrm{d}x\right)^2\Bigg] + \mathcal{O}_{\prec}\left(N^{-7/6}\right).\end{aligned}$$

We can now conclude from eqs. (6.26) and (6.27) that

$$\mathbf{E}\, F_N^2 = F_N' + \frac{s_{11}}{N}\left(\int_{-2}^{2}\rho(x)f'(x)\,\mathrm{d}x\right)^2 + \mathcal{O}_\prec\left(N^{-7/6}\right) = \frac{V_f}{N} + \mathcal{O}_\prec\left(N^{-7/6}\right) \tag{6.40}$$

and

$$\mathbf{E}\, \widetilde{F}_N^2 = \widetilde{F}_N' + \frac{s_{11}}{N}\left(\int_{-2}^{2}\rho(x)\frac{xf''(x)}{2}\,\mathrm{d}x\right)^2 + \mathcal{O}_\prec\left(N^{-7/6}\right) = \frac{\widetilde{V}_f}{N} + \mathcal{O}_\prec\left(N^{-7/6}\right). \tag{6.41}$$

So far we assumed $\sigma_2 = 0$ in (6.30). We now consider the general case for which we need (6.32) instead of (6.31). A similar analysis shows that we have to add an additional term $|\sigma_2|^2\, V_{\sigma_2}$ to V_f and $\widetilde{V}_f$ both in (6.40) and (6.41), given by

$$\begin{aligned} V_{\sigma_2} :=& \frac{1}{2\pi^2}\Re\iint_{\mathbb{R}} f'(x)f'(x')\Bigg[m(z_0)^2\overline{m(z_0')^2} \\ &\times\frac{(1+m(z_0)\overline{m(z_0')}\Re\sigma_2)\frac{\tan[m(z)\overline{m(z_0')}\Im\sigma_2]}{\Im\sigma_2} - m(z)\overline{m(z_0')}}{1-\Re\sigma_2\frac{\tan[m(z_0)\overline{m(z_0')}\Im\sigma_2]}{\Im\sigma_2}} \\ &- m(z_0)^2m(z_0')^2\frac{(1+m(z_0)m(z_0')\Re\sigma_2)\frac{\tan[m(z_0)m(z_0')\Im\sigma_2]}{\Im\sigma_2} - m(z_0)m(z_0')}{1-\Re\sigma_2\frac{\tan[m(z_0)m(z_0')\Im\sigma_2]}{\Im\sigma_2}}\Bigg]\,\mathrm{d}x\,\mathrm{d}x'. \end{aligned} \tag{6.42}$$

For the special case $\sigma_2 \in \mathbb{R}$ eq. (6.32) simplifies to

$$\begin{aligned} \frac{1}{N}\sum_{i\neq j} G_{ij}G_{ij}' =& \frac{m(z)^2m(z')^2\Re\sigma_2}{1-m(z)m(z')\Re\sigma_2} \\ &+ \mathcal{O}_\prec\left(\frac{1}{(\eta+\eta')\sqrt{N\eta^2\eta'}} + \frac{1}{(\eta+\eta')\sqrt{N\eta\eta'^2}} + \frac{1}{N(\eta+\eta')\eta\eta'}\right). \end{aligned}$$

In particular, for symmetric H, where $\sigma_2 = 1$ we find that eq. (6.42) simplifies to $V_{\sigma_2} = V_{f,1}$. This completes the proof of Proposition 6.3.3, modulo the proof of Lemma 6.3.7.

Proof of Lemma 6.3.7. The proof of the second inequality is similar to the first one and will be left to the reader. For the first inequality, we split the integration in two regimes. We shall make use of the fact (see, e.g., [73]) that on a compact domain, say $|z_0| \leq 10$, we have

$$\left|1-m(z_0)^2\right| \asymp \sqrt{\kappa_x+\eta} \quad\text{and}\quad \Im m(z_0) \asymp \begin{cases} \sqrt{\kappa_x+\eta_0} & \text{if } |x|\leq 2, \\ \frac{\eta_0}{\sqrt{\kappa_x+\eta_0}} & \text{else,} \end{cases} \tag{6.43}$$

where $\kappa_x = ||x|-2|$ is the distance to the edge.

Firstly in the region where $\max\{|x|, |x'|\} \geq 2$, we find

$$\left|1-m(z_0)\overline{m(z_0')}\right| \geq \frac{1}{2}\left[1-|m(z_0)|^2+1-\left|m(z_0')\right|^2\right] \geq c\sqrt{\kappa_{\max\{|x|,|x'|\}}+\eta_0},$$

where $c > 0$ is a universal constant, due to the fact that $1-|m(z_0)|^2 = \eta_0/\Im m(z_0)$ and (6.43).

Secondly, in the region where $|x|, |x'| < 2$, we write

$$1 - m(z_0)\overline{m(z_0')} = 1 - |m(z_0')|^2 + (m(z_0') - m(z_0))\overline{m(z_0')}$$

and estimate

$$\left|(m(z_0') - m(z_0))\overline{m(z_0')}\right| \geq c\,|x' - x|$$

for some positive constant c. This inequality follows from writing

$$\Re[m(z_0') - m(z_0)] = \int_x^{x'} \Re m'(u + i\eta_0)\,\mathrm{d}u$$

and from the estimate

$$\Re m'(u + i\eta_0) = -\frac{2(\Im m(u + i\eta_0))^2}{|1 - m(u + i\eta_0)^2|^2} \leq -c$$

for $|u| \leq 2$, where we used (6.43) in the last step. Consequently,

$$\left|1 - m(z_0)\overline{m(z_0')}\right| \geq c\,|x - x'| - \left|1 - |m(z_0')|^2\right| \geq c\,|x - x'| - C\frac{\eta_0}{\sqrt{\kappa_{x'} + \eta_0}}$$

and it follows that $\left|1 - m(z_0)\overline{m(z_0')}\right| \geq c\,|x - x'|/2$ whenever

$$|x - x'| \geq 2(C/c)\eta_0/\sqrt{\kappa_{x'} + \eta_0}.$$

Together with the trivial bound $\left|1 - m(z_0)\overline{m(z_0')}\right| \geq c\eta$ we find that the integral in (6.38) is bounded by $C\,|\log \eta_0|$. □

6.4 Computation of Higher Moments

We now turn to the computation of higher order moments and thereby to the completing the proof of Theorem 6.2.1. We recall from (6.24)–(6.25) that

$$F_N = -\frac{1}{\sqrt{N}}\Im \int_{\mathbb{R}} \int_{\eta_0}^{10} g(z)\left[X(z) - \sqrt{N}h_{11}\right]\mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_{\prec}\left(N^{-2/3}\right)$$

and

$$\widetilde{F}_N = -\frac{1}{\sqrt{N}}\Im \int_{\mathbb{R}} \int_{\eta_0}^{10} g(z)\left[X(z + h_{11}) + \sqrt{N}h_{11}m'(z)\right]\mathrm{d}\eta\,\mathrm{d}x + \mathcal{O}_{\prec}\left(N^{-2/3}\right),$$

where g and X were defined in (6.28). In order to compute moments of F_N and $\widetilde{F}_N$ we have to compute

$$\mathbf{E}[X(z_1)\ldots X(z_k)]$$

for any $k \in \mathbb{N}$ and $z_l \in \mathbb{C}\backslash\mathbb{R}, l = 1, \ldots, k$. We will first take the expectation with respect to the vector h in the X's which leads to a cyclic contraction of the indices of $\widehat{G}$. After taking the expectation with respect to $\widehat{H}$, we will show that the leading order terms come from cycles of length two. This will effectively show that the Wick theorem holds for the random variables X. The following lemma shows that cyclic products of at least three resolvents are in fact of lower order (the same phenomenon already was observed in [70]):

Lemma 6.4.1. *For closed cycles of length $k > 2$ we have that*

$$N^{-k/2} \sum_{i_1,\dots,i_k}^{\sim} \mathbf{E}\, G^{(1)}_{i_1 i_2} \dots G^{(k-1)}_{i_{k-1} i_k} G^{(k)}_{i_k i_1} \prec \frac{1}{(\max_a \eta_a)\sqrt{N\eta_1 \dots \eta_k}} \sum_{a=1}^{k} \frac{1}{\sqrt{\eta_a}}, \tag{6.44}$$

and for open cycles of any length $k > 1$ we have that

$$N^{-(k+1)/2} \sum_{i_1,\dots,i_k}^{\sim} \mathbf{E}\, G^{(1)}_{i_1 i_2} \dots G^{(k-1)}_{i_{k-1} i_k} \prec \frac{1}{\sqrt{N\eta_1 \dots \eta_{k-1}}} \sum_{a=1}^{k-1} \frac{1}{\sqrt{\eta_a}}, \tag{6.45}$$

where $G^{(l)} := \widehat{G}(z_l)$, $z_l \in \mathbb{C} \setminus \mathbb{R}$ with $\eta_l = |\Im z_l|$ for $l = 1, \dots, k$ and $\overset{\sim}{\sum}$ indicates that the sum is performed over pairwise distinct indices. Moreover, the same bound holds true when any of the $G^{(l)}$ are replaced by their transposes or Hermitian conjugates.

Proof. We first prove eq. (6.44) and assume a real symmetric H. To do so, we let $\epsilon > 0$ be arbitrary and will actually prove

$$N^{-k/2} \sum_{i_1,\dots,i_k}^{\sim} \mathbf{E}\, G^{(1)}_{i_1 i_2} \dots G^{(k-1)}_{i_{k-1} i_k} G^{(k)}_{i_k i_1} \prec \frac{N^{\epsilon}}{(\eta_1 + \eta_k)\sqrt{N\eta_1 \dots \eta_k}} \sum_{a=1}^{k} \frac{1}{\sqrt{\eta_a}},$$

from which (6.44) follows due to the definition of $\prec$ in Definition 6.3.1. We make use of the resolvent identity $G^{(1)} = \widehat{H} G^{(1)}/z_1 - 1/z_1$ to write

$$N^{-k/2} \sum_{i_1,\dots,i_k}^{\sim} \mathbf{E}\, G^{(1)}_{i_1 i_2} \dots G^{(k)}_{i_k i_1} = \frac{1}{N^{k/2} z_1} \sum_{i_1,\dots,i_k}^{\sim} \sum_{n} \mathbf{E}\, h_{i_1 n} G^{(1)}_{n i_2} G^{(2)}_{i_2 i_3} \dots G^{(k)}_{i_k i_1}. \tag{6.46}$$

We use the standard cumulant expansion (introduced in the context of random matrices in [116]) up to the third order term with a truncation

$$\begin{aligned} \mathbf{E}\, h f(h) = \mathbf{E}\, h\, \mathbf{E}\, f(h) + \mathbf{E}\, h^2\, \mathbf{E}\, f'(h) + \mathcal{O}\left(\mathbf{E}\left|h^3 \mathbb{1}(|h| > N^{\tau - 1/2})\right| \|f''\|_\infty\right) \\ + \mathcal{O}\left(\mathbf{E}\,|h|^3 \sup_{|x| \le N^{\tau-1/2}} |f''(x)|\right), \end{aligned} \tag{6.47}$$

where f is any smooth function of a real random variable h, such that the expectations exist and $\tau > 0$ is arbitrary (for a recent similar use of this formula with truncation see [98, Lemma 3.1]). This yields

$$\begin{aligned} \mathbf{E}\, h_{i_1 n} G^{(1)}_{n i_2} G^{(2)}_{i_2 i_3} \dots G^{(k)}_{i_k i_1} &= \frac{1}{N} \mathbf{E} \frac{\partial \left[G^{(1)}_{n i_2} G^{(2)}_{i_2 i_3} \dots G^{(k)}_{i_k i_1} \right]}{\partial h_{i_1 n}} + R \\ &= \frac{1}{N} \mathbf{E} \frac{\partial G^{(1)}_{n i_2}}{\partial h_{i_1 n}} G^{(2)}_{i_2 i_3} \dots G^{(k)}_{i_k i_1} + \frac{1}{N} \sum_{a=2}^{k} \mathbf{E} \frac{\partial G^{(a)}_{i_a i_{a+1}}}{\partial h_{i_1 n}} G^{(1)}_{n i_2} \prod_{a \ne b = 2}^{k} G^{(b)}_{i_b i_{b+1}} + R, \end{aligned} \tag{6.48}$$

where it is understood that $i_{k+1} = i_1$ and R is the error term resulting from the cumulant expansion. Using the identity

$$\frac{\partial G_{ij}}{\partial h_{kl}} = -(G_{ik} G_{lj} + G_{il} G_{kj})/(1 + \delta_{kl}),$$

and the local law (6.19), the first term on the rhs. of eq. (6.48) becomes

$$-(G_{ni_1}^{(1)}G_{ni_2}^{(1)}+G_{nn}^{(1)}G_{i_1i_2}^{(1)})G_{i_2i_3}^{(2)}\dots G_{i_ki_1}^{(k)}=-m(z_1)G_{i_1i_2}^{(1)}\dots G_{i_ki_1}^{(k)}+\mathcal{O}_{\prec}\left(\frac{1}{N^{k/2+1/2}\sqrt{\eta\eta_1}}\right),$$

whenever $n\neq i_1,i_2$ and where $\eta:=\eta_1\dots\eta_k$. If $n=i_1$ or $n=i_2$, we shall make use of the trivial estimate

$$-(G_{ni_1}^{(1)}G_{ni_2}^{(1)}+G_{nn}^{(1)}G_{i_1i_2}^{(1)})G_{i_2i_3}^{(2)}\dots G_{i_ki_1}^{(k)}\prec\frac{1}{N^{k/2}\sqrt{\eta}}.$$

The $a=k$ summand of the second term in eq. (6.48) becomes

$$\begin{aligned}&-(G_{i_ki_1}^{(k)}G_{ni_1}^{(k)}+G_{i_kn}^{(k)}G_{i_1i_1}^{(k)})G_{ni_2}^{(1)}\dots G_{i_{k-1}i_k}^{(k-1)}\\&\quad=-m(z_k)G_{ni_2}^{(1)}\dots G_{i_kn}^{(k)}+\mathcal{O}_{\prec}\left(\frac{1}{N^{k/2+1/2}\sqrt{\eta\eta_k}}\right)\end{aligned}\qquad(6.49)$$

whenever $n\neq i_1,i_k$. For these exceptional n we shall again use the trivial $N^{-k/2}\eta^{-1/2}$ estimate. For $a\neq k$ the summand in the second term of eq. (6.48) can always be estimated by

$$-(G_{i_ai_1}^{(a)}G_{ni_{a+1}}^{(a)}+G_{i_an}^{(a)}G_{i_1i_{a+1}}^{(a)})G_{ni_2}^{(1)}\prod_{a\neq b=2}^{k}G_{i_bi_{b+1}}^{(b)}\prec\frac{1}{N^{k/2}\sqrt{\eta}}$$

and this bound can be improved to

$$-(G_{i_ai_1}^{(a)}G_{ni_{a+1}}^{(a)}+G_{i_an}^{(a)}G_{i_1i_{a+1}}^{(a)})G_{ni_2}^{(1)}\prod_{a\neq b=2}^{k}G_{i_bi_{b+1}}^{(b)}\prec\frac{1}{N^{k/2+1/2}\sqrt{\eta\eta_a}},$$

whenever $n\notin\{i_1,\dots,i_k\}$. Thus for most of the $\mathcal{O}\left(N^{k+1}\right)$ terms in the sum in eq. (6.46) we have the improved bound, while for $\mathcal{O}\left(N^{k}\right)$ terms, where $n=i_l$ for some l, we use the weaker bound and we find that

$$\begin{aligned}&N^{-k/2}\sum_{i_1,\dots,i_k}^{\sim}\sum_n\frac{1}{N}\mathbf{E}\frac{\partial\left[G_{ni_2}^{(1)}G_{i_2i_3}^{(2)}\dots G_{i_ki_1}^{(k)}\right]}{\partial h_{i_1n}}\\&\quad=\frac{1}{N^{k/2+1}z_1}\sum_{n,i_1,\dots,i_k}^{\sim}\left[-m(z_1)\,\mathbf{E}\,G_{i_1i_2}^{(1)}\dots G_{i_ki_1}^{(k)}-m(z_k)\,\mathbf{E}\,G_{ni_2}^{(1)}\dots G_{i_kn}^{(k)}\right]\\&\qquad+\frac{1}{z_1}\mathcal{O}_{\prec}\left(\sum_{a=1}^{k}\frac{1}{\sqrt{N\eta\eta_a}}\right).\end{aligned}\qquad(6.50)$$

It remains to estimate the error R. To do so we have to compute the second derivatives

$$\frac{\partial^2\left[G_{ni_2}^{(1)}G_{i_2i_3}^{(2)}\dots G_{i_ki_1}^{(k)}\right]}{\partial h_{i_1n}^2}$$

which is a polynomial in $G_{ab}^{(l)}$ for $l\in\{1,\dots,k\},a,b\in\{i_1,\dots,i_k,n\}$ of total degree $k+2$ with at most 2 diagonal factors for $n\notin\{i_1,\dots,i_k\}$, and otherwise with at most 3 diagonal factors in every monomial. These factors each satisfy the entry-wise local law (6.19), but now we need these estimates even uniformly for all $|h_{i_1 n}| \le N^{\tau-1/2}$ which does not directly follow from the concept of stochastic domination. To circumvent this technical issue, we need to explicitly display the dependence of the resolvents $G^{(l)}$ on $h_{i_1 n}$. We therefore write $\widetilde{H}$ for the matrix $\widehat{H}$ with the (i_1, n) and (n, i_1) entries set to 0 and $\widetilde{G}^{(l)} = (\widetilde{H} - z_l)^{-1}$. Note that $\widetilde{G}^{(l)}$ is independent of $h_{i_1 n}$. Since $\widetilde{G}^{(l)}$ is the resolvent of a generalized Wigner matrix, from [82, 73] we have the usual resolvent estimates (6.18)–(6.19) for $\widetilde{G}^{(l)}$. Moreover, if $i_1 \neq n$, then by the resolvent identity

$$\begin{aligned} G^{(l)}_{ab} = \widetilde{G}^{(l)}_{ab} &- h_{i_1 n}\left[\widetilde{G}^{(l)}_{an}\widetilde{G}^{(l)}_{i_1 b} + \widetilde{G}^{(l)}_{a i_1}\widetilde{G}^{(l)}_{nb}\right] \\ &+ h^2_{i_1 n}\left[\widetilde{G}^{(l)}_{an}\widetilde{G}^{(l)}_{i_1 n}G^{(l)}_{i_1 b} + \widetilde{G}^{(l)}_{an}\widetilde{G}^{(l)}_{i_1 i_1}G^{(l)}_{nb} + \widetilde{G}^{(l)}_{a i_1}\widetilde{G}^{(l)}_{nn}G^{(l)}_{i_1 b} + \widetilde{G}^{(l)}_{a i_1}\widetilde{G}^{(l)}_{n i_1}G^{(l)}_{nb}\right] \end{aligned}$$

and we can estimate

$$\max_{a \neq b} \sup_{|h_{i_1 n}| \le N^{-1/2+\tau}} G^{(l)}_{ab} \prec \frac{N^{\tau}}{\sqrt{N\eta_l}}, \qquad \max_{a} \sup_{|h_{i_1 n}| \le N^{-1/2+\tau}} G^{(l)}_{aa} \prec 1$$

whenever $\tau < 1/12$ where we used the trivial bound $G^{(l)}_{ab} \le 1/\eta_l \le N^{2/3}$. On the other hand, if $i_1 = n$, then we have

$$G^{(l)}_{ab} = \widetilde{G}^{(l)}_{ab} - h_{nn}\widetilde{G}^{(l)}_{an}\widetilde{G}^{(l)}_{nb} + h^2_{nn}\widetilde{G}^{(l)}_{an}\widetilde{G}^{(l)}_{nn}G^{(l)}_{nb}$$

and therefore again

$$\max_{a \neq b} \sup_{|h_{nn}| \le N^{-1/2+\tau}} G^{(l)}_{ab} \prec \frac{N^{\tau}}{\sqrt{N\eta_l}}, \qquad \max_{a} \sup_{|h_{nn}| \le N^{-1/2+\tau}} G^{(l)}_{aa} \prec 1$$

whenever $\tau < 1/12$. Therefore

$$\sup_{|h_{i_1 n}| < N^{-1/2+\tau}} \left| \frac{\partial^2 \left[G^{(1)}_{n i_2} G^{(2)}_{i_2 i_3} \dots G^{(k)}_{i_k i_1}\right]}{\partial h^2_{i_1 n}} \right| \prec \sum_{a=1}^{k} \frac{N^{k\tau} N^{-k/2}}{\sqrt{\eta\eta_a}}$$

and we can conclude

$$\frac{1}{N^{k/2} z_1} \widetilde{\sum_{i_1,\dots,i_k}} \sum_n \mathbf{E}\left|h_{1_1 n}\right|^3 \sup_{|h_{i_1 n}| < N^{-1/2+\tau}} \left| \frac{\partial^2 \left[G^{(1)}_{n i_2} G^{(2)}_{i_2 i_3} \dots G^{(k)}_{i_k i_1}\right]}{\partial h^2_{i_1 n}} \right| \prec \sum_{a=1}^{k} \frac{N^{k\tau} N^{-1/2}}{\sqrt{\eta\eta_a}}. \tag{6.51}$$

We can now pick $\tau = \min\{\frac{1}{12}, \frac{\epsilon}{k}\}$ to have a final estimate of order

$$\sum_{a=1}^{k} \frac{N^{\epsilon}}{\sqrt{N\eta\eta_a}}$$

for the error originating from the last term in the truncated cumulant expansion (6.47). The remaining error

$$\mathbf{E}\left|h^3_{i_1 n}\mathbb{1}(|h_{i_1 n})| > N^{\tau-1/2}\right| \sup_{h_{i_1 n}} \left| \frac{\partial^2 \left[G^{(1)}_{n i_2} G^{(2)}_{i_2 i_3} \dots G^{(k)}_{i_k i_1}\right]}{\partial h^2_{i_1 n}} \right| \tag{6.52}$$

is negligible for any fixed k since the expectation is smaller than any power of $N^{-\tau}$ due to the arbitrary polynomial decay (6.6).

Putting together (6.50), the identity

$$z_1 + m(z_1) + m(z_k) = \frac{m(z_1)m(z_k) - 1}{m(z_1)}$$

and the estimates on R from (6.51)–(6.52) we have shown that

$$\begin{aligned} N^{-k/2} \sum_{i_1 \neq \cdots \neq i_k} \mathbf{E}\, G^{(1)}_{i_1 i_2} \ldots G^{(k)}_{i_k i_1} &= \frac{m(z_1)}{1 - m(z_1)m(z_k)} \mathcal{O}\left(\sum_{a=1}^{k} \frac{N^{\epsilon}}{\sqrt{N\eta\eta_a}} \right) \\ &= \mathcal{O}_{\prec}\left(\sum_{a=1}^{k} \frac{N^{\epsilon}}{(\eta_1 + \eta_k)\sqrt{N\eta\eta_a}} \right). \end{aligned}$$

Since the lhs. of this estimate is cyclic in $i_1, \ldots, i_k$, we can replace $\eta_1 + \eta_k$ in the error term by $\max_a \eta_a$.

For the proof of eq. (6.45) we follow essentially the same steps but for the last $a = k-1$ term we find

$$-(G^{(k-1)}_{i_{k-1} i_1} G^{(k-1)}_{n i_k} + G^{(k-1)}_{i_{k-1} n} G^{(k-1)}_{i_1 i_k}) G^{(1)}_{n i_2} \ldots G^{(k-2)}_{i_{k-2} i_{k-1}} \prec \frac{1}{N^{k/2}\sqrt{\eta\eta_{k-1}}}$$

instead of eq. (6.49). Consequently, eq. (6.50) becomes

$$\begin{aligned} &N^{-(k+1)/2} \sum_{i_1 \neq \cdots \neq i_k} \mathbf{E}\, G^{(1)}_{i_1 i_2} \ldots G^{(k-1)}_{i_{k-1} i_k} \\ &= \frac{1}{N^{(k+1)/2+1} z_1} \sum_{n \neq i_1 \neq \cdots \neq i_k} \left[-m(z_1)\, \mathbf{E}\, G^{(1)}_{i_1 i_2} \ldots G^{(k-1)}_{i_{k-1} i_k} \right] + \frac{1}{z_1} \mathcal{O}_{\prec}\left(\sum_{a=1}^{k-1} \frac{N^{\epsilon}}{\sqrt{N\eta\eta_a}} \right) \end{aligned}$$

from which eq. (6.45) follows immediately.

For the last claim, note that none of the estimates above relied on the order of the indices of any $G^{(l)}$ and the same bound holds true in the case of any transpositions.

The proof of the Hermitian case is similar, but the cumulant expansion has to be replaced by a complex variant (as in, e.g. [98, Lemma 7.1]). □

Next, we note that the bounds (6.44)–(6.45) also hold true without taking expectations:

Corollary 6.4.2. *In the setup of Lemma 6.4.1, for closed cycles of length* $k \geq 2$ *we have that*

$$N^{-k/2} \widetilde{\sum_{i_1, \ldots, i_k}} G^{(1)}_{i_1 i_2} \ldots G^{(k-1)}_{i_{k-1} i_k} G^{(k)}_{i_k i_1} \prec \frac{1}{(\max_a \eta_a)\sqrt{N\eta_1 \ldots \eta_k}} \sum_{a=1}^{k} \frac{1}{\sqrt{\eta_a}}, \tag{6.53}$$

and for open cycles of any length $k > 1$ *we have that*

$$N^{-(k+1)/2} \widetilde{\sum_{i_1, \ldots, i_k}} G^{(1)}_{i_1 i_2} \ldots G^{(k-1)}_{i_{k-1} i_k} \prec \frac{1}{\sqrt{N\eta_1 \ldots \eta_{k-1}}} \sum_{a=1}^{k-1} \frac{1}{\sqrt{\eta_a}}. \tag{6.54}$$

Proof. We note that the fluctuation averaging analysis from [70, Proof of Prop. 5.3 in Sections 6–7] does not rely on the fact $z_1 = \cdots = z_k$ and therefore also applies to the present case. □

The following lemma shows an asymptotic Wick theorem for X's, i.e. that higher moments of X to leading order only involve pairings:

Lemma 6.4.3. *For $k \geq 2$ and $z_1, \dots, z_k \in \mathbb{C}$ with $|\Im z_l| = \eta_l > 0$ we have that*

$$\mathbf{E}[X(z_1)\dots X(z_k)] = \sum_{\pi \in P_2([k])} \prod_{\{a,b\}\in\pi} \mathbf{E}[X(z_a)X(z_b)] + \mathcal{O}_{\prec}\left(\frac{1}{\sqrt{N\eta_1 \dots \eta_k}} \sum_{a\neq b} \frac{1}{(\eta_a+\eta_b)\sqrt{\eta_a}} \right),$$

where $[k] := \{1, \dots, k\}$ and $P_2(L)$ are the partitions of a set L into subsets of size 2.

Proof. For definiteness we prove the real symmetric case. Since the argument relies on counting pairings, the proof of the complex Hermitian case is very similar and we omit it. We have to compute

$$\begin{aligned}
&\mathbf{E}_1 \prod_{l=1}^{k} \left[\sum_{i_l \neq j_l} h_{i_l} G^{(l)}_{i_l j_l} h_{j_l} + \sum_{i_l} \left(h_{i_l}^2 - \frac{1}{N}\right) G^{(l)}_{i_l i_l} \right] \\
&\quad = \sum_{L\subset[k]} \mathbf{E}_1 \left[\left(\prod_{l\in L} \sum_{i_l\neq j_l} h_{i_l} G^{(l)}_{i_l j_l} h_{j_l} \right) \left(\prod_{l\notin L} \sum_{i_l} \left(h_{i_l}^2 - \frac{1}{N}\right) G^{(l)}_{i_l i_l} \right) \right],
\end{aligned}$$

where $[k] = \{1, \dots, k\}$ and $\mathbf{E}_1 = \mathbf{E}(\cdot|H^{(1)}) = \mathbf{E}(\cdot|\widehat{H})$ and we recall that $G^{(l)}$ is independent of h. We already know from eq. (6.30) and Lemma 6.3.6 that the leading order of this expression is at most $N^{-k/2}$. In order to have non-zero expectation we have to pair any h_{i_l} and h_{j_l} with at least some other h_{i_m} or h_{j_m}. An easy counting argument using the bound $G^{(l)}_{i_l j_l} \prec (N\eta_l)^{-1/2}$ shows that for any $L \subset [k]$ the corresponding L-term is at most of order

$$N^{-(k+1)/2} \prod_{l\in L} \eta_l^{-1/2}$$

whenever any three or more h_i's are paired. This already shows that we can restrict our attention to pairings and in particular odd moments asymptotically are of lower order.

Starting from some h_{i_l} with $l \notin L$ we have to pair it either to another h_{i_m} with $m \notin L$, or some h_{i_m} or h_{j_m} with $m \in L$. In the former case we have a closed pairing with expectation

$$\mathbf{E}_1 \left[(h_{i_l}^2 - 1/N)(h_{i_l}^2 - 1/N) G^{(l)}_{i_l i_l} G^{(m)}_{i_l i_l} \right] = \frac{\sigma_4 - 1}{N^2} G^{(l)}_{i_l i_l} G^{(m)}_{i_l i_l}.$$

In the latter case, say we paired h_{i_l} to h_{i_m}, we have to continue the pairing process by pairing h_{j_m} with another h_{i_k} or h_{j_k} with $k \in L$ etc., until we reach another h_{i_n} with $n \notin L$. This expression represents an open cycle as in (6.54) and is therefore subleading.

On the other hand, starting from some h_{i_l} or h_{j_l} with $l \in L$, and continue the pairings as in the previous paragraph until we pair to an h_{i_m} with $m \notin L$ which results in an open cycle as in (6.54) and is subleading. Therefore we only have to consider closed cycles of the pure L-type, from which, due to (6.53), only those of length 2 are leading. That means that pairing h_{i_l} to h_{i_m} automatically forces a pairing of h_{j_l} and h_{j_m}, and that a pairing of h_{i_l} to h_{j_m} automatically forces a pairing of h_{j_l} and h_{i_m}. These give the leading contribution of

$$\mathbf{E}_1\left[h_{i_l}G^{(l)}_{i_lj_l}h_{j_l}h_{j_l}G^{(m)}_{j_li_l}h_{i_l} + h_{i_l}G^{(l)}_{i_lj_l}h_{j_l}h_{i_l}G^{(m)}_{i_lj_l}h_{j_l}\right] = \frac{G^{(l)}_{i_lj_l}G^{(m)}_{i_lj_l}}{N^2}.$$

The above findings allow us to conclude that

$$\begin{aligned}
&N^{k/2}\,\mathbf{E}_1\left[\left(\prod_{l\in L}\sum_{i_l\neq j_l} h_{i_l}G^{(l)}_{i_lj_l}h_{j_l}\right)\left(\prod_{l\notin L}\sum_{i_l}\left(h_{i_l}^2-\frac{1}{N}\right)G^{(l)}_{i_li_l}\right)\right] \\
&\quad = N^{k/2}\,\mathbf{E}_1\left(\prod_{l\in L}\sum_{i_l\neq j_l} h_{i_l}G^{(l)}_{i_lj_l}h_{j_l}\right)\mathbf{E}_1\left(\prod_{l\notin L}\sum_{i_l}\left(h_{i_l}^2-\frac{1}{N}\right)G^{(l)}_{i_li_l}\right) + \mathcal{O}_{\prec}(\Psi) \\
&\quad = N^{k/2}\left(\sum_{\pi\in P_2(L)}\prod_{\{a,b\}\in\pi}\sum_{i\neq j}\frac{G^{(a)}_{ij}G^{(b)}_{ij}+G^{(a)}_{ij}G^{(b)}_{ji}}{N^2}\right) \\
&\qquad\quad \times\left(\sum_{\pi\in P_2([k]\setminus L)}\prod_{\{a,b\}\in\pi}\frac{\sigma_4-1}{N^2}\sum_i G^{(a)}_{ii}G^{(b)}_{ii}\right) + \mathcal{O}_{\prec}(\Psi) \\
&\quad = N^{k/2}\left(\sum_{\pi\in P_2(L)}\prod_{\{a,b\}\in\pi}\frac{2}{N}\frac{m(z_a)^2m(z_b)^2}{1-m(z_a)m(z_b)}\right) \\
&\qquad\quad \times\left(\sum_{\pi\in P_2([k]\setminus L)}\prod_{\{a,b\}\in\pi}\frac{\sigma_4-1}{N}m(z_a)m(z_b)\right) + \mathcal{O}_{\prec}(\Psi),
\end{aligned}$$

where in the last step we used Lemma 6.3.6 and we introduced the error term

$$\Psi = \frac{1}{\sqrt{N\eta_1\ldots\eta_k}}\sum_{a\neq b}\frac{1}{(\eta_a+\eta_b)\sqrt{\eta_a}}.$$

We now recognize the last expression as the sum over products of pairs of $\mathbf{E}[X(z_a)X(z_b)]$, completing the proof. □

We now have all ingredients to compute

$$\begin{aligned}
&\left(-\Im\int_{\mathbb{R}}\int_{\eta_0}^{10} g(z)\left[X(z)-\sqrt{N}h_{11}\right]\mathrm{d}\eta\,\mathrm{d}x\right)^k \\
&\quad = \sum_{j=0}^{k}\binom{k}{j}(\sqrt{N}h_{11})^{k-j}\left(\Im\int_{\mathbb{R}}\int_{\eta_0}^{10} g(z)\,\mathrm{d}\eta\,\mathrm{d}x\right)^{k-j}\left(-\Im\int_{\mathbb{R}}\int_{\eta_0}^{10} g(z)X(z)\,\mathrm{d}\eta\,\mathrm{d}x\right)^{j}.
\end{aligned}$$

Recall that h_{11} and X are independent. From Lemma 6.4.3 we can conclude that

$$\begin{aligned}
\mathbf{E}\left(-\Im\int_{\mathbb{R}}\int_{\eta_0}^{10} g(z)X(z)\,\mathrm{d}\eta\,\mathrm{d}x\right)^{j} &= \sum_{\pi\in P_2([j])}(2V_{f,1}+(\sigma_4-1)V_{f,2})^{j/2}+\mathcal{O}_{\prec}\left(N^{-1/6}\right) \\
&= (j-1)!!\,(2V_{f,1}+(\sigma_4-1)V_{f,2})^{j/2}+\mathcal{O}_{\prec}\left(N^{-1/6}\right)
\end{aligned}$$

for even j and

$$\mathbf{E}\left(-\Im\int_{\mathbb{R}}\int_{\eta_0}^{10} g(z)X(z)\,\mathrm{d}\eta\,\mathrm{d}x\right)^j = \mathcal{O}_\prec\left(N^{-1/6}\right)$$

for odd j. If h_{11} follows a normal distribution, then $\mathbf{E}\,h_{11}^{k-j} = (k-j-1)!!\,(s_{11}/N)^{(k-j)/2}$ whenever $k-j$ is even and $\mathbf{E}\,h_{11}^{k-j} = 0$, otherwise. Therefore, since

$$(j-1)!!(k-j-1)!!\binom{k}{j} = (k-1)!!\binom{k/2}{j/2}$$

for even j, k, we have that

$$\begin{aligned}\mathbf{E}&\left(-\Im\int_{\mathbb{R}}\int_{\eta_0}^{10} g(z)\left[X(z)-\sqrt{N}h_{11}\right]\mathrm{d}\eta\,\mathrm{d}x\right)^k\\ &= (k-1)!!\,[2V_{f,1}+(\sigma_4-1)V_{f,2}+s_{11}V_{f,3}]^{k/2}+\mathcal{O}_\prec\left(N^{-1/6}\right)\end{aligned} \tag{6.55}$$

whenever k is even and

$$\mathbf{E}\left(-\Im\int_{\mathbb{R}}\int_{\eta_0}^{10} g(z)\left[X(z)-\sqrt{N}h_{11}\right]\mathrm{d}\eta\,\mathrm{d}x\right)^k = \mathcal{O}_\prec\left(N^{-1/6}\right)$$

otherwise.

For the case of complex Hermitian H we can follow the same argument and ultimately find that eq. (6.55) becomes

$$\begin{aligned}\mathbf{E}&\left(-\Im\int_{\mathbb{R}}\int_{\eta_0}^{10} g(z)\left[X(z)-\sqrt{N}h_{11}\right]\mathrm{d}\eta\,\mathrm{d}x\right)^k\\ &= (k-1)!!\left[V_{f,1}+|\sigma_2|^2 V_{\sigma_2}+(\sigma_4-1)V_{f,2}+s_{11}V_{f,3}\right]^{k/2}+\mathcal{O}_\prec\left(N^{-1/6}\right).\end{aligned}$$

Finally, we remark that the same proof also works in the case of $\widetilde{f}_N$ and we basically only have to replace $V_{f,3}$ by $\widetilde{V}_{f,3}$.

6.A Comparison to Gaussian Free Field

In this section we investigate to what extent our main result on the Gaussian fluctuation of linear statistics of H and its minor $\widehat{H}$ is consistent with the Gaussian free field (GFF) limit proved in [34, 35, 128] for real symmetric matrices. In these papers the joint fluctuations of the spectral counting functions of minors were shown to converge to a GFF in the large N limit, assuming that the sizes of the minors asymptotically differed by cN. Our result corresponds to the difference of the linear statistics of two minors whose sizes differ only by one. The fluctuation is only of order $N^{-1/2}$ and it is not visible on the macroscopic scale studied in [34, 35, 128]. Nevertheless, one may *formally* apply these macroscopic result to our case. Here we show that this naive extension indeed provides the correct order of magnitude and also the correct variance of the fluctuations, but does not identify their precise distribution.

For comparability with [34, 35, 128] assume a constant variance on the diagonal and constant fourth moment on the off-diagonal, i.e., $\mathbf{E}\,h_{ii}^2 = \mathbf{E}\,h_{11}^2 = s_{11}/N$ and $\mathbf{E}\,h_{ij}^4 = \sigma_4/N^2$

for all $i \neq j$. First we recall the main result of [128] which is based on [34], where the corresponding formula was first proved for monomial test functions. Given an $N \times N$ Wigner matrix H, we denote the consecutive lower right minors by $H_n := (H_{jk})_{j,k=N-n+1}^{N}$. A special case of Theorem 2.2 of [128] then asserts that for any $f \in H^{5.5+\epsilon}(\mathbb{R})$ and for any $x, y \in (0,1]$, the covariance of linear statistics of two nested minors of size Nx and Ny is asymptotically given by

$$\begin{aligned} C_f(x,y) &:= \lim_{N\to\infty} \mathbf{Cov}(\operatorname{Tr} f(H_{[xN]}), \operatorname{Tr} f(H_{[yN]})) \qquad (6.56)\\ &= \frac{1}{\pi^2}\int_{\gamma_x}\int_{\gamma_y} f'\left(z+\frac{x}{z}\right) f'\left(w+\frac{y}{w}\right) \log\left|\frac{x\wedge y - zw}{x \wedge y - z\overline{w}}\right| \left(1-\frac{x}{z^2}\right)\left(1-\frac{y}{w^2}\right) \mathrm{d}w\,\mathrm{d}z\\ &\quad + \frac{s_{11}-2}{x\vee y}\left(\frac{1}{2\pi}\int_{-2\sqrt{x}}^{2\sqrt{x}} \frac{sf(s)}{\sqrt{4x-s^2}}\,\mathrm{d}s\right)\left(\frac{1}{2\pi}\int_{-2\sqrt{y}}^{2\sqrt{y}} \frac{tf(t)}{\sqrt{4y-t^2}}\,\mathrm{d}t\right)\\ &\quad + \frac{\sigma_4-3}{2(x\vee y)^2}\left(\int_{-2\sqrt{x}}^{2\sqrt{x}} \frac{2x-s^2}{\pi\sqrt{4x-s^2}} f(s)\,\mathrm{d}s\right)\left(\int_{-2\sqrt{y}}^{2\sqrt{y}} \frac{2y-t^2}{\pi\sqrt{4y-t^2}} f(t)\,\mathrm{d}t\right) \end{aligned}$$

where the γ_x denotes the contour $|z|^2 = x, \Im z > 0$ in counter-clockwise order.

Recalling our previous notation $H = H_N$ and $\widehat{H} = H_{N-1}$, in our Theorem 6.2.1 we derived a formula for the rescaled variance

$$\begin{aligned} D_{N,f} &:= N\,\mathbf{Var}[\operatorname{Tr} f(H_N) - \operatorname{Tr} f(H_{N-1})]\\ &= N\big[\mathbf{Cov}(\operatorname{Tr} f(H_N), \operatorname{Tr} f(H_N)) - \mathbf{Cov}(\operatorname{Tr} f(H_N), \operatorname{Tr} f(H_{N-1}))\\ &\qquad - \mathbf{Cov}(\operatorname{Tr} f(H_{N-1}), \operatorname{Tr} f(H_N)) + \mathbf{Cov}(\operatorname{Tr} f(H_{N-1}), \operatorname{Tr} f(H_{N-1}))\big], \end{aligned}$$

which corresponds to

$$N\Big[C_f(1,1) - C_f(1, 1-\frac{1}{N}) - C_f(1-\frac{1}{N}, 1-\frac{1}{N}) + C_f(1-\frac{1}{N}, 1-\frac{1}{N})\Big],$$

suggesting that we should compare our result to the limit

$$D_f := \lim_{\epsilon\to 0} \frac{C_f(1,1) - C_f(1,1-\epsilon) - C_f(1-\epsilon,1) - C_f(1-\epsilon,1-\epsilon)}{\epsilon}.$$

Note that this latter formula is the renormalized derivative of the Gaussian free field $\phi_x(f)$ with covariance $C_f(x,y)$ at $x = 1$:

$$D_f = \lim_{\epsilon\to 0} \mathbf{Var}\,\frac{\phi_1(f) - \phi_{1-\epsilon}(f)}{\sqrt{\epsilon}}.$$

In the following theorem we compare the field

$$\psi_x^{(N)}(f) := \operatorname{Tr} f(H_{[xN]}) - \mathbf{E}\operatorname{Tr} f(H_{[xN]})$$

defined by our linear eigenvalue statistics to the Gaussian free field $\phi_x(f)$.

Theorem 6.A.1. *Let H be real symmetric Wigner matrices satisfying the conditions from Theorem 6.2.1 and additionally assume that $\mathbf{E}\,h_{ii}^2 = \mathbf{E}\,h_{11}^2 = s_{11}/N$ and $\mathbf{E}\,h_{ij}^4 = \sigma_4/N^2$ for all $i \neq j$. Then for any $f \in H^2(\mathbb{R})$ the centered random variables*

$$X_f := \lim_{\epsilon\to 0}\frac{\phi_1(f) - \phi_{1-\epsilon}(f)}{\sqrt{\epsilon}} \qquad \text{and} \qquad Y_f := \lim_{N\to\infty}\frac{\psi_1^{(N)}(f) - \psi_{1-1/N}^{(N)}(f)}{\sqrt{1/N}}$$

are well defined (the limit is in distribution sense) and they have the same variance

$$\begin{aligned}\mathbf{E}\, X_f^2 = \mathbf{E}\, Y_f^2 = 2\int_{-2}^{2} f'(s)^2\rho(s)\,\mathrm{d}s + (\sigma_4-3)\left(\int_{-2}^{2} sf'(s)\rho(s)\,\mathrm{d}s\right)^2 \\ + (s_{11}-2)\left(\int_{-2}^{2} f'(s)\rho(s)\,\mathrm{d}s\right)^2 .\end{aligned} \tag{6.57}$$

Moreover, the distributions of X_f and Y_f agree if and only if h_{11} follows a Gaussian distribution.

Proof. The variance formula for Y_f follows immediately from Theorem 6.2.1.

In order to prove that X_f is well defined and follows a Gaussian distribution, it suffices to check that D_f is finite. To do so, we treat the three terms of $C_f(x,y)$ from (6.56) separately, which for convenience we call $C_f(x,y) = C_f^{(1)}(x,y) + C_f^{(2)}(x,y) + C_f^{(3)}(x,y)$. It is easy to check that

$$\begin{aligned}&\lim_{\epsilon\to 0}\frac{C_f^{(2)}(1,1) - C_f^{(2)}(1,1-\epsilon) - C_f^{(2)}(1-\epsilon,1) - C_f^{(2)}(1-\epsilon,1-\epsilon)}{\epsilon}\\ &\qquad = (s_{11}-2)\left(\int_{-2}^{2} f'(s)\rho(s)\,\mathrm{d}s\right)^2\end{aligned}$$

and that

$$\begin{aligned}&\lim_{\epsilon\to 0}\frac{C_f^{(3)}(1,1) - C_f^{(3)}(1,1-\epsilon) - C_f^{(3)}(1-\epsilon,1) - C_f^{(3)}(1-\epsilon,1-\epsilon)}{\epsilon}\\ &\qquad = (\sigma_4-3)\left(\int_{-2}^{2} sf'(s)\rho(s)\,\mathrm{d}s\right)^2 .\end{aligned}$$

For the computation of $C_f^{(1)}(x,y)$ we now substitute $z = \sqrt{x}e^{i\phi}$ and $w = \sqrt{y}e^{i\psi}$ with $\phi,\psi\in[0,\pi]$, so that

$$\begin{aligned}C_f^{(1)}(x,y) = \frac{4\sqrt{xy}}{\pi^2}\int_0^\pi\int_0^\pi & f'(2\sqrt{x}\cos\phi)f'(2\sqrt{y}\cos\psi)\\ &\times\log\left|\frac{x\wedge y - \sqrt{xy}e^{i(\phi+\psi)}}{x\wedge y - \sqrt{xy}e^{i(\phi-\psi)}}\right|\sin\phi\sin\psi\,\mathrm{d}\psi\,\mathrm{d}\phi\end{aligned}$$

and after a further substitution of $2\sqrt{x}\cos\phi = s$ and $2\sqrt{y}\cos\psi = t$ and simple algebraic manipulation we arrive at

$$C_f^{(1)}(x,y) = \frac{1}{\pi^2}\int_{-2\sqrt{x}}^{2\sqrt{x}}\int_{-2\sqrt{y}}^{2\sqrt{y}} f'(s)f'(t)\operatorname{arctanh}\frac{\sqrt{(4x-s^2)(4y-t^2)}}{2(x+y)-st}\,\mathrm{d}t\,\mathrm{d}s.$$

To keep the notation relatively short we now introduce

$$\begin{aligned}a_{x,y}(s,t) &:= \operatorname{arctanh}\frac{\sqrt{(4x-s^2)(4y-t^2)}}{2(x+y)-st}\\ &= \operatorname{arctanh}\sqrt{1-\frac{(x-y)^2+(t-s)(xt-ys)}{(x+y)^2-(x+y)st+s^2t^2/4}}\end{aligned}$$

and we claim that

$$\frac{a_{1,1}(s,t)-a_{1,1-\epsilon}(s,t)-a_{1-\epsilon,1}(s,t)+a_{1-\epsilon,1-\epsilon}(s,t)}{\epsilon}\approx\pi\delta(s-t)\sqrt{4-t^2}$$

for any fixed $s,t\in[-2,2]$ in the $\epsilon\to0$ limit. Firstly, one readily checks that when $|s-t|\gg\epsilon$, then

$$\lim_{\epsilon\to0}\frac{a_{1,1}(s,t)-a_{1,1-\epsilon}(s,t)-a_{1-\epsilon,1}(s,t)+a_{1-\epsilon,1-\epsilon}(s,t)}{\epsilon}=0.$$

Secondly, when $|x-y|\leq\epsilon$ and $|s-t|\leq M\epsilon$ for some large but fixed M, then a series expansion gives

$$\begin{aligned}a_{x,y}(s,t)=\log 2-\frac{1}{2}\log\frac{(x-y)^2+(t-s)(xt-ys)}{(x+y)^2-(x+y)st+s^2t^2/4}\\-\frac{1}{4}\frac{(x-y)^2+(t-s)(xt-ys)}{(x+y)^2-(x+y)st+s^2t^2/4}+\mathcal{O}\left(\epsilon^2\right),\end{aligned}$$

assuming, additionally, that $|s|\leq2\sqrt{x}(1-\delta)$, $|t|\leq2\sqrt{y}(1-\delta)$ with some fixed $\delta>0$. It can now be checked via an explicit integration that

$$\int_{|s-t|<M\epsilon}\frac{a_{1,1}(s,t)-a_{1,1-\epsilon}(s,t)-a_{1-\epsilon,1}(s,t)+a_{1-\epsilon,1-\epsilon}(s,t)}{\epsilon}\,\mathrm{d}s=\pi\sqrt{4-t^2}+\mathcal{O}\left(\epsilon\right)$$

for fixed t, proving the claim. We can conclude that

$$\begin{aligned}&\frac{C_f^{(1)}(1,1)-C_f^{(1)}(1,1-\epsilon)-C_f^{(1)}(1-\epsilon,1)-C_f^{(1)}(1-\epsilon,1-\epsilon)}{\epsilon}\\&=\frac{1}{\pi^2}\int_{-2}^{2}\int_{-2}^{2}f'(s)f'(t)\delta(s-t)\pi\sqrt{4-t^2}\,\mathrm{d}s\,\mathrm{d}t+\mathcal{O}\left(\epsilon\right)=2\int_{-2}^{2}f'(t)^2\rho(t)\,\mathrm{d}t+\mathcal{O}\left(\epsilon\right),\end{aligned}$$

where we used that $f'\in L^2$ and therefore the integral over the neglected area where $|s|>2\sqrt{x}(1-\delta)$ or $|t|>2\sqrt{y}(1-\delta)$ does not contribute to leading order. Thus

$$D_f=2\int_{-2}^{2}f'(s)^2\rho(s)\,\mathrm{d}s+(\sigma_4-3)\left(\int_{-2}^{2}sf'(s)\rho(s)\,\mathrm{d}s\right)^2+(s_{11}-2)\left(\int_{-2}^{2}f'(s)\rho(s)\,\mathrm{d}s\right)^2,$$

completing the proof of (6.57). In particular, the limit defining X_f exists and is Gaussian. Finally, the existence of the limit defining Y_f follows from the moment calculations in section 6.4 and assumption (6.6) on the moments of h_{11} that together also guarantee tightness. This completes the proof of the theorem. □

Fluctuations of Functions of Wigner Matrices 7

We show that matrix elements of functions of $N \times N$ Wigner matrices fluctuate on a scale of order $N^{-1/2}$ and we identify the limiting fluctuation. Our result holds for any function f of the matrix that has bounded variation thus considerably relaxing the regularity requirement imposed in [131, 137].

Published as L. Erdős and D. Schröder, *Fluctuations of functions of Wigner matrices*, Electron. Commun. Probab. **21**, Paper no. 86, 15 (2016), MR3600514.

7.1 Introduction

The density of states of an $N \times N$ Wigner random matrix $H = H^{(N)}$ converges to the Wigner semicircular law [176]. More precisely, for any continuous function $f\colon \mathbb{R} \to \mathbb{C}$

$$\lim_{N\to\infty} \frac{1}{N} \operatorname{Tr} f(H) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(\lambda_k) = \int f(x)\, \mu_{sc}(\mathrm{d}x) \tag{7.1}$$

where $\lambda_1, \dots, \lambda_N$ are the (real) eigenvalues of H and $\mu_{sc}(\mathrm{d}x) := \frac{1}{2\pi}\sqrt{(4-x^2)_+}\,\mathrm{d}x$.

It is well known that for regular functions f, the normalized linear eigenvalue statistics $\frac{1}{N} \operatorname{Tr} f(H)$ have an asymptotically Gaussian fluctuation on scale of order $1/N$, see, for example, [15, 132, 154, 24, 157, 162, 106] for different results in this direction, also for other random matrix ensembles. To our knowledge, this result under the weakest regularity condition on f was proved in [162]; for general Wigner matrices $f \in H^{1+\epsilon}$ was required, while for Wigner matrices with substantial GUE component $f \in H^{1/2+\epsilon}$ was sufficient. Notice that the order of the fluctuation $1/N$ is much smaller than $1/\sqrt{N}$ which would be predicted by the standard central limit theorem (CLT) if the eigenvalues were weakly dependent. The failure of CLT on scale $1/\sqrt{N}$ is a signature of the strong correlations among the eigenvalues.

In this paper we investigate the individual matrix elements of $f(H)$. We will show that the semicircle law (7.1) holds also for any diagonal matrix element $f(H)_{ii}$ and not only for their average, $\frac{1}{N} \operatorname{Tr} f(H)$; however, the corresponding fluctuation is much larger, it is on scale $1/\sqrt{N}$. Moreover, the limiting distribution of the rescaled fluctuation is not necessarily Gaussian; it also depends on the distribution of the matrix element h_{ii}. Similar fluctuation results hold for the off diagonal matrix elements $f(H)_{ij}$, $i \neq j$. For regularity condition, we merely assume that f is of bounded variation, $f \in BV$. We also prove an effective error bound of order $N^{-2/3}$ that we can improve to N^{-1} if $f' \in L^\infty$, i.e. we provide a two-term expansion for each matrix element of $f(H)$.

Similar results (with less precise error bounds) were obtained previously in [133] for Gaussian random matrices and in [147, 131, 137] for general Wigner matrices under the much stronger regularity assumptions that

$$\int_{\mathbb{R}} (1+|\xi|)^3 \left|\widehat{f}(\xi)\right| \mathrm{d}\xi < \infty \quad \text{or} \quad \int_{\mathbb{R}} (1+|\xi|)^{2s} \left|\widehat{f}(\xi)\right|^2 \mathrm{d}\xi < \infty \quad \text{for } s > 3, \tag{7.2}$$

where $\widehat{f}(\xi) := \int_{\mathbb{R}} e^{-i\xi x} f(x)\,\mathrm{d}x$. The main novelty of the current work is thus to relax these regularity conditions to $f \in BV$. In addition, [147, 131, 137] assumed that in the case of complex Hermitian matrices, the real and imaginary part of the entries have equal variance. Our approach does not require this technical assumption. We also refer to [131] where similar questions have been studied for more general statistics of the form $\mathrm{Tr}[f(H)A]$ for non-random matrices A under the fairly strong regularity condition $\int (1+|\xi|)^4 |\widehat{f}(\xi)|\,\mathrm{d}\xi < \infty$.

A special case of these questions is when the test function $f(x)$ is given by $\varphi_z(x) = (x-z)^{-1}$ for some complex parameter z in the upper half plane, $\eta := \Im z > 0$. In fact, for f which are analytic in a complex neighborhood of $[-2, 2]$, a simple contour integration shows that for the linear statistics it is sufficient to understand the resolvent of H, i.e., $\varphi_z(H) = (H-z)^{-1}$ for any fixed z in the upper half plane. If f is less regular, one may still express $f(H)$ as an integral of the resolvents over z, weighted by the $\partial_{\bar{z}}$-derivative of an almost analytic extension of f to the upper half plane (Helffer-Sjöstrand formula). In this case, the integration effectively involves the regime of z close to the real axis, so the resolvent $(H-z)^{-1}$ and its matrix elements need to be controlled even as $\eta \to 0$ simultaneously with $N \to \infty$. These results are commonly called *local semicircle laws*. They hold down to the optimal scale $\eta \gg 1/N$ with an optimal error bound of order $1/\sqrt{N\eta}$ for the individual matrix elements and a bound of order $1/N\eta$ for the normalized trace of the resolvent (see, e.g. [82]). With the help of the Helffer-Sjöstrand formula, more accurate local laws can be transformed to weaker regularity assumptions on the test function in the linear eigenvalue statistics, see [162]. In this paper we replace the Helffer-Sjöstrand formula by Pleijel's formula [148] that provides a more effective functional calculus for functions with low regularity.

A similar relation between regularity and local laws holds for individual matrix elements, $f(H)_{ii}$. Using the Schur complement formula one can relate $f(H)_{ii}$ to the *difference* of a linear statistics for H and for its minor $\widehat{H}$ obtained by removing the i-th row and column from H. In a recent paper [DS1] we investigated the fluctuations of this difference without directly connecting it to $f(H)_{ii}$. Applied to a special family of test function $f(x) = |x-a|$, the difference of linear statistics is closely related to the fluctuation of Kerov's interlacing sequences of the eigenvalues of H and its minor.

Motivated by this application, Sasha Sodin pointed out that this fluctuation can be related to the fluctuation of a single matrix element of the resolvent by the Markov correspondence, see [159] for details. It is therefore natural to ask if one could use the fluctuation result from [DS1] on the interlacing sequences to strengthen the existing results on the fluctuations of the matrix elements of the resolvent and hence of $f(H)$. In fact, not the result itself, but the core of the analysis in [DS1] can be applied; this is the content of the current paper. We thank Sasha for asking this question and calling our attention to the problem of fluctuation of the matrix elements of $f(H)$ and to the previous literature [133, 147, 131, 137]. Furthermore, he pointed out to us that the contour integral formula from Pleijel's paper [148] could potentially replace the Helffer-Sjöstrand formula in our argument to the end of further reducing the regularity assumptions on f. We are very grateful to him for this insightful idea that we believe will have further applications.

7.2 Main results

We consider complex Hermitian and real symmetric random $N\times N$ matrices $H = (h_{ij})_{i,j=1}^N$ with the entries being independent (up to the symmetry constraint $h_{ij} = \overline{h_{ji}}$) random variables satisfying

$$\mathbf{E}\, h_{ij} = 0, \quad \mathbf{E}\,|h_{ij}|^2 = \frac{s_{ij}}{N} \quad \text{and} \quad \mathbf{E}\,|h_{ij}|^p \leq \frac{\mu_p}{N^{p/2}} \tag{7.3}$$

for all i, j, p and some absolute constants μ_p. We assume that the matrix of variances is approximately stochastic, i.e.

$$\sum_j s_{ij} = N + \mathcal{O}\left(1\right)$$

to guarantee that the limiting density of states is the Wigner semicircular law.

To formulate the error bound concisely we introduce the following commonly used (see, e.g., [73]) notion of high probability bound.

Definition 7.2.1 (Stochastic Domination). *If*

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

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

$$\sup_{u\in U^{(N)}} \mathbf{P}\left[X^{(N)}(u) > N^\epsilon Y^{(N)}(u)\right] \leq N^{-D}$$

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

It can be checked (see [73, Lemma 4.4]) that $\prec$ satisfies the usual arithmetic properties, e.g. if $X_1 \prec Y_1$ and $X_2 \prec Y_2$, then also $X_1 + X_2 \prec Y_1 + Y_2$ and $X_1X_2 \prec Y_1Y_2$. We will say that a (sequence of) events $A = A^{(N)}$ holds with *overwhelming probability* if $\mathbf{P}(A^{(N)}) \geq 1 - N^{-D}$ for any $D > 0$ and $N \geq N_0(D)$. In particular, under the conditions (7.3), we have $h_{ij} \prec N^{-1/2}$ and $\max_k |\lambda_k| \leq 3$ with overwhelming probability.

We further introduce a notion quantifying the rate of weak convergence of distributions. We say that a sequence of random variables X_N *converges in distribution at a rate* $r(N)$ to X if for any $t \in \mathbb{R}$ it holds that

$$\mathbf{E}\, e^{itX_N} = \mathbf{E}^{itX} + \mathcal{O}_t\left(r(N)\right),$$

where we allow the coefficient of the rate to be t-dependent uniformly for $|t| \leq T$ for any fixed T. If X_N converges in distribution at a rate $r(N)$, we write

$$X_N \overset{d}{=} X + \mathcal{O}\left(r(N)\right).$$

In particular, this implies that

$$\mathbf{E}\,\Phi(X_N) = \mathbf{E}\,\Phi(X) + \mathcal{O}\left(r(N)\right)$$

for any analytic function Φ with compactly supported Fourier transform.

Our main result for the diagonal entries of $f(H)$ is summarized in the following theorem. By permutational symmetry there is no loss in generality in studying $f(H)_{11}$. By considering real and imaginary parts separately, from now on we always assume that f is real valued.

Theorem 7.2.2. *Let the Wigner matrix H satisfy (7.3), $s_{ij} = 1$ for $i \neq j$ and $s_{ii} \leq C$ for all i, $\mathbf{E}\,|h_{1j}|^4 = \sigma_4/N^2$ for $j = 2, \ldots, N$ and $\mathbf{E}\,h_{ij}^2 = \sigma_2/N$ with some $\sigma_2, \sigma_4 \in \mathbb{R}$. Moreover, let $f \in BV([-3,3])$ be some real-valued function of bounded variation and assume that $h_{11} \overset{d}{=} \xi_{11}/\sqrt{N}$ where ξ_{11} is an N-independent random variable. Then*

$$f(H)_{11} \overset{d}{=} \int f(x)\,\mu_{sc}(\mathrm{d}x) + \frac{\widehat{\Delta}_f + \xi_{11}\int f(x)x\,\mu_{sc}(\mathrm{d}x)}{\sqrt{N}} + \begin{cases} \mathcal{O}\left(N^{-1}\right) & \text{if } f' \in L^\infty, \\ \mathcal{O}\left(N^{-2/3}\right) & \text{else}, \end{cases} \tag{7.4}$$

where $\widehat{\Delta}_f$ is a centered Gaussian random variable of variance

$$\mathbf{E}\left(\widehat{\Delta}_f\right)^2 = V_{f,1} + V_{f,1}^{(\sigma_2)} - 2V_{f,2} - (1+\sigma_2)V_{f,3} + (\sigma_4 - 2 - \sigma_2^2)V_{f,4}, \tag{7.5}$$

and the $V_{f,i}$ and $V_{f,1}^{(\sigma_2)}$ are given by quadratic forms defined in (7.22).

More precisely, (7.4) means that, to leading order

$$f(H)_{11} = \int f(x)\,\mu_{sc}(\mathrm{d}x) + \mathcal{O}_\prec\left(N^{-1/2}\right)$$

and, weakly

$$T_f^{(N)} := \sqrt{N}\left[f(H)_{11} - \int f(x)\,\mu_{sc}(\mathrm{d}x)\right] - \xi_{11}\int f(x)x\,\mu_{sc}(\mathrm{d}x) \Rightarrow \widehat{\Delta}_f \tag{7.6}$$

at a speed

$$\mathbf{E}\left(T_f^{(N)}\right)^k = \mathbf{E}\,\widehat{\Delta}_f^k + \begin{cases} \mathcal{O}\left(\frac{C^k(k/2)!}{\sqrt{N}}\right) & \text{if } f' \in L^\infty, \\ \mathcal{O}\left(\frac{C^k(k/2)!}{N^{1/6}}\right) & \text{else} \end{cases}$$

for all k. The speed of convergence in the Lévy metric d_L is given by

$$d_L(T_f^{(N)}, \widehat{\Delta}_f) \leq C(f)\frac{\log\log N}{\sqrt{\log N}} \tag{7.7}$$

with some constant depending on f.

The corresponding result for the off diagonal terms is as follows.

Theorem 7.2.3. *Under the assumptions of Theorem 7.2.2,*

$$f(H)_{12} \overset{d}{=} \frac{1}{\sqrt{N}}\left[\widetilde{\Delta}_f + \xi_{12}\int f(x)x\,\mu_{sc}(\mathrm{d}x)\right] + \begin{cases}\mathcal{O}\left(N^{-1}\right) & \text{if } f' \in L^\infty,\\ \mathcal{O}\left(N^{-2/3}\right) & \text{else},\end{cases} \tag{7.8}$$

where $\widetilde{\Delta}_f$ is a centered complex Gaussian satisfying

$$\mathbf{E}\,\widetilde{\Delta}_f^2 = V_{f,1}^{(\sigma_2)} - V_{f,2} - \sigma_2 V_{f,3}, \quad \mathbf{E}\left|\widetilde{\Delta}_f\right|^2 = V_{f,1} - V_{f,2} - V_{f,3}.$$

and the $V_{f,i}$ and $V_{f,1}^{(\sigma_2)}$ are defined in (7.22).

More precisely, (7.8) *means that*

$$f(H)_{12} = \mathcal{O}_{\prec}\left(N^{-1/2}\right)$$

and, introducing the notation

$$S_f^{(N)} := \sqrt{N} f(H)_{12} - \xi_{12}\int f(x)x\,\mu_{sc}(\mathrm{d}x),$$

we have that

$$\mathbf{E}\left(S_f^{(N)}\right)^k\left(\overline{S_f^{(N)}}\right)^l = \mathbf{E}\,\widetilde{\Delta}_f^k\overline{\widetilde{\Delta}_f}^l + \begin{cases}\mathcal{O}\left(\frac{((k+l)/2)!}{\sqrt{N}}\right) & \text{if } f' \in L^\infty,\\ \mathcal{O}\left(\frac{((k+l)/2)!}{N^{1/6}}\right) & \text{else}\end{cases}$$

holds for all $k, l \in \mathbb{N}$. The analogues of (7.6) *and* (7.7) *also hold for $T_f^{(N)}$ replaced with $S_f^{(N)}$.*

The fluctuation results in Theorems 7.2.2 and 7.2.3 for test functions satisfying the stronger regularity assumption (7.2) and without explicit error terms have been proven in [131, 137]. We also remark that (7.6) implies the joint asymptotic normality of the fluctuations of $f(H^{(N)})_{11}$ for several test functions. More precisely, for any $f \in BV$ we define $T_f^{(N)}$ via (7.6). Then for any given functions $f_1, f_2, \dots, f_k \in BV$, the random k-vector

$$\left(T_{f_1}^{(N)}, T_{f_2}^{(N)}, \dots, T_{f_k}^{(N)}\right)$$

weakly converges to a Gaussian vector with covariance given via the variance (7.5) using the parallelogram identity. Similar result holds for the joint distribution of the off diagonal elements $f_k(H)_{12}$. One may specialize this result to the case when f is a characteristic function, i.e. we may define

$$T_x^{(N)} := T_{\mathbb{1}_{[-3,x]}}^{(N)}, \qquad x \in [-3, 3],$$

where $\mathbb{1}_{[a,b]}$ is the characteristic function of the interval $[a, b]$. Clearly, the finite dimensional marginals of the sequence of stochastic processes $\{T_x^{(N)}, x \in [-3, 3]\}$ are asymptotically Gaussian. The tightness remains an open question.

7.3 Pleijel's Inversion Formula

Our main tool relating $f(H)_{ij}$ to the resolvent $G = G(z) = (H-z)^{-1}$ is summarized in the following proposition. We formulate it for general probability measures μ supported on some $[-K, K]$ and their Stieltjes transform

$$m_\mu(z) = \int \frac{1}{\lambda - z}\,\mu(\mathrm{d}\lambda).$$

Later we will apply the proposition to $\mu = \rho_N$ and $\mu = \widetilde{\rho}_N$ with $\rho_N, \widetilde{\rho}_N$ being the spectral measures of typical diagonal and off-diagonal entries

$$\int f\,\mathrm{d}\rho_N = f(H)_{11}, \qquad \int f\,\mathrm{d}\widetilde{\rho}_N = f(H)_{12}.$$

Proposition 7.3.1. *Let $L > K > 0$ and let μ denote a probability measure which is supported on $[-K, K]$ and let $f \in BV([-L, L])$ be a function of bounded variation which is compactly supported in $[-L, L]$. Then*

$$\begin{aligned}\int f(\lambda)\,\mu(\mathrm{d}\lambda) &= \frac{1}{2\pi}\iint\limits_{I_{\eta_0}^M} m_\mu(x+i\eta)\,\mathrm{d}\eta\,\mathrm{d}f(x) + \frac{1}{\pi}\int_{-L}^{L} f(x)\Im m_\mu(x+Mi)\,\mathrm{d}x \qquad (7.9)\\ &\qquad + \mathcal{O}\Big(\eta_0 \left\|m_\mu(\cdot+i\eta_0)\right\|_{L^1(|\mathrm{d}f|)}\Big)\\ &= \frac{1}{2\pi}\iint\limits_{I_{\eta_0}^M} m_\mu(x+i\eta)\,\mathrm{d}\eta\,\mathrm{d}f(x) + \mathcal{O}\Big(\eta_0 \left\|m_\mu(\cdot+i\eta_0)\right\|_{L^1(|\mathrm{d}f|)} + \frac{1}{M}\left\|f\right\|_1\Big)\end{aligned}$$

holds for any $\eta_0, M > 0$ where $I_{\eta_0}^M := [-L, L] \times ([-M, M] \setminus [-\eta_0, \eta_0])$, $\|\cdot\|_1 = \|\cdot\|_{L^1(\mathrm{d}x)}$ and $\mathrm{d}f$ is understood as the (signed) Lebesgue–Stieltjes measure.

Before going into the proof, we present a special case of Proposition 7.3.1. If $f = \mathbb{1}_{[x,x']}$, then (7.9) can be written as the path integral

$$\mu([x, x']) = \frac{1}{2\pi i}\int_{\gamma(x,x')} m_\mu(z)\,\mathrm{d}z + \mathcal{O}\left(\eta_0[|m_\mu(x+i\eta_0)| + |m_\mu(x'+i\eta_0)|]\right),$$

where $\gamma(x, x')$ is the chain indicated in Figure (7.1(c)). We also want to remark that for our purposes (7.9) is favorable over the Helffer-Sjöstrand representation, as used in [DS1], since it requires considerably less regularity on f.

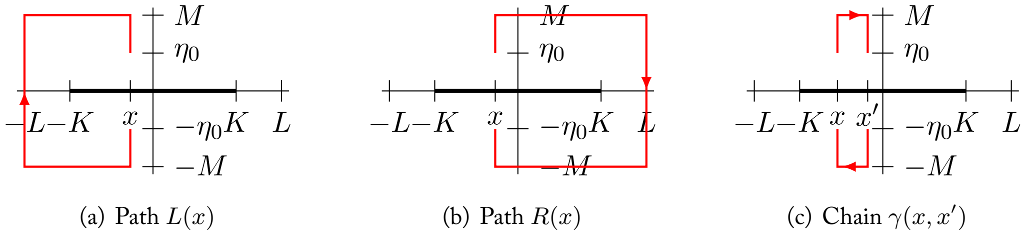
Proof of Proposition 7.3.1. From [148, Eq. (5)] we know that

$$\mu([-K, x)) = \frac{1}{2\pi i}\int_{L(x)} m_\mu(z)\,\mathrm{d}z + \frac{\eta}{\pi}\Re m_\mu(z_0) + \mathcal{O}\left(\eta \Im m_\mu(z_0)\right), \qquad (7.10)$$

where $L(x)$ is a directed path as indicated in Figure 7.1(a) and $z_0 = x + i\eta_0, \eta_0 > 0$.

By the definition of the Lebesgue–Stieltjes integral for functions of bounded variation we have that

$$\int f(\lambda)\,\mu(\mathrm{d}\lambda) = \int_{-L}^{L}\Big(\int \mathbb{1}(\lambda \geq x)\,\mu(\mathrm{d}\lambda)\Big)\,\mathrm{d}f(x) = \int_{-L}^{L}\mu([x, K])\,\mathrm{d}f(x).$$



(a) Path $L(x)$ (b) Path $R(x)$ (c) Chain $\gamma(x,x')$

Figure 7.1: Integration paths

By virtue of (7.10) we can write

$$\int f(\lambda)\,\mu(\mathrm{d}\lambda) = \frac{1}{\pi}\int_{-L}^{L}\left(\frac{1}{2i}\int_{R(x)} m_\mu(z)\,\mathrm{d}z\right)\mathrm{d}f(x) + \mathcal{O}\left(\eta_0\,\|m_\mu(\cdot+i\eta_0)\|_{L^1(|\mathrm{d}f|)}\right),$$

where $R(x)$ is the path indicated in Figure 7.1(b) and $|\mathrm{d}f|$ indicates the total variation measure of $\mathrm{d}f$. We then write out the inner integral as

$$\frac{1}{2i}\int_{R(x)} m_\mu(z)\,\mathrm{d}z = \int_{\eta_0}^{M}\Re m_\mu(x+i\eta)\,\mathrm{d}\eta + \int_{x}^{L}\Im m_\mu(y+iM)\,\mathrm{d}y \\ - \int_{0}^{M}\Re m_\mu(L+i\eta)\,\mathrm{d}\eta.$$

Since the last term is x-independent, it will vanish after integrating against $\mathrm{d}f$ since we assumed f to be compactly supported. For the second term we find

$$\int f(\lambda)\,\mu(\mathrm{d}\lambda) = \frac{1}{\pi}\int_{-L}^{L}\int_{\eta_0}^{M}\Re m_\mu(x+i\eta)\,\mathrm{d}\eta\,\mathrm{d}f(x) + \frac{1}{\pi}\int_{-L}^{L} f(x)\Im m_\mu(x+iM)\,\mathrm{d}x \\ + \mathcal{O}\left(\eta_0\,\|m_\mu(\cdot+i\eta_0)\|_{L^1(|\mathrm{d}f|)}\right).$$

Since $|\Im m_\mu(x+iM)| \le 1/M$ we thus have

$$\int f(\lambda)\,\mu(\mathrm{d}\lambda) = \frac{1}{\pi}\int_{-L}^{L}\int_{\eta_0}^{M}\Re m_\mu(x+i\eta)\,\mathrm{d}\eta\,\mathrm{d}f(x) \\ + \mathcal{O}\left(\eta_0\,\|m_\mu(\cdot+i\eta_0)\|_{L^1(|\mathrm{d}f|)} + \frac{1}{M}\|f\|_1\right)$$

for any $\eta_0, M > 0$. For applications it turns out to be favorable to get rid of the real part which we can by noting that $2\Re m_\mu(z) = m_\mu(z) + m_\mu(\overline{z})$ and therefore

$$\int f(\lambda)\,\mu(\mathrm{d}\lambda) = \frac{1}{2\pi}\iint_{I_{\eta_0}^M} m_\mu(x+i\eta)\,\mathrm{d}\eta\,\mathrm{d}f(x) + \mathcal{O}\left(\eta_0\,\|m_\mu(\cdot+i\eta_0)\|_{L^1(|\mathrm{d}f|)} + \frac{1}{M}\|f\|_1\right),$$

where we recall $I_{\eta_0}^M = [-L,L]\times([-M,M]\setminus[-\eta_0,\eta_0])$. □

We finally note that a variant of Proposition 7.3.1 could also be proven directly without appealing to the contour integration from [148]. The key computation in that direction is summarized in the following Lemma which we establish here for later convenience.

Lemma 7.3.2. *Let $f \in BV([-L,L])$ be compactly supported and let g be a function which is analytic away from the real axis and satisfies $g(\overline{z}) = \overline{g(z)}$. Then for any $\eta_0, M > 0$ we have that*

$$\frac{1}{2\pi}\iint\limits_{I_{\eta_0}^M} g(x+i\eta)\,\mathrm{d}\eta\,\mathrm{d}f(x) = \frac{1}{\pi}\int_{-L}^{L} f(x)\Im g(x+i\eta_0)\,\mathrm{d}x + \mathcal{O}\left(\|f\|_1 \max_{x\in[-L,L]} |g(x+iM)|\right).$$

Applying Lemma 7.3.2 to $g = m_\mu$ yields, modulo an error term,

$$\frac{1}{2\pi}\iint\limits_{I_{\eta_0}^M} m_\mu(x+i\eta)\,\mathrm{d}\eta\,\mathrm{d}f(x) \approx \int\int_{-L}^{L} f(x)\frac{1}{\pi}\frac{\eta_0}{(\lambda-x)+\eta_0^2}\,\mathrm{d}x\,\mu(\mathrm{d}\lambda)$$

and taking the limit $\eta_0 \to 0$ makes the inner integral tend to $f(\lambda)$ in L^1-sense. In this way we can establish a variant of Proposition 7.3.1, albeit with a weaker error estimate.

Proof of Lemma 7.3.2. This follows from the computation

$$\begin{aligned}\iint\limits_{I_{\eta_0}^M} g(x+i\eta)\,\mathrm{d}\eta\,\mathrm{d}f(x) &= -i\int_{\partial I_{\eta_0}^M} f(x)g(z)\,\mathrm{d}z = 2\int_{-3}^{3} f(x)\Im\left[g(x+i\eta) - g(x+iM)\right]\mathrm{d}x \\ &= 2\int_{-3}^{3} f(x)\Im g(x+i\eta_0)\,\mathrm{d}x + \mathcal{O}\left(\|f\|_1 \max_{x\in[-3,3]} |g(x+iM)|\right),\end{aligned}$$

where the first step follows from Stokes' or Green's Theorem. □

7.4 Diagonal entries

We first prove Theorem 7.2.2 about the diagonal entries of $f(H)$. The spectral measure corresponding to the $(1,1)$-matrix element, ρ_N defined as

$$\int f\,\mathrm{d}\rho_N = f(H)_{11}$$

is concentrated in $[-2.5, 2.5]$ with overwhelming probability. We can without loss of generality assume that f is compactly supported in $[-3,3]$ since smoothly cutting off f outside the spectrum does not change the result. Applying Proposition 7.3.1 to $\mu = \rho_N$ with $K = 2.5$, $L = 3$, we find that (using $z = x+i\eta$, $z_0 = x+i\eta_0$)

$$f(H)_{11} = \frac{1}{2\pi}\iint\limits_{I_{\eta_0}^M} G(z)_{11}\,\mathrm{d}\eta\,\mathrm{d}f(x) + \mathcal{O}_\prec\left(\eta_0\int |G(z_0)_{11}|\,\mathrm{d}f(x) + \frac{1}{M}\|f\|_1\right). \tag{7.11}$$

To analyze $G(z)_{11}$ we recall the Schur complement formula

$$G(z)_{11} = \frac{1}{h_{11} - z - \langle h, \widehat{G}(z)h\rangle}, \quad \text{where } H = \begin{pmatrix} h_{11} & h^* \\ h & \widehat{H}\end{pmatrix}, \qquad \widehat{G}(z) := (\widehat{H} - z)^{-1}.$$

To study the asymptotic behavior of $G(z)_{11}$ we rely on the local semicircle law in the averaged form (see [82] or [73, Theorem 2.3]) applied to the resolvent of the minor

$$\widehat{m}_N(z) = \frac{1}{N}\operatorname{Tr}\widehat{G}(z) = m(z) + \mathcal{O}_\prec\left(\frac{1}{N|\eta|}\right),$$

and its entry-wise form

$$G(z)_{ij} - \delta_{ij} m(z) \prec \frac{1}{\sqrt{N}\,|\eta|} \tag{7.12}$$

which both hold true for all $|\eta| = |\Im z| > \eta_0 \gg N^{-1}$. Here m denotes the Stieltjes transform of the semicircular distribution μ_{sc}, $m(z) := \int (\lambda - z)^{-1}\, \mu_{sc}(\mathrm{d}\lambda)$.

Since by (7.12),

$$\int |G(x+i\eta_0)_{11}|\,\mathrm{d}f(x) = \int |m(x+i\eta_0)|\,\mathrm{d}f(x) + \mathcal{O}_\prec\left(\int \left|\frac{1}{\sqrt{N\eta_0}}\right| \mathrm{d}f(x)\right) \prec \|\mathrm{d}f\|$$

for $\eta_0 \gg 1/N$, where $\|\mathrm{d}f\|$ is the total variation norm of the Lebesgue–Stieltjes measure $\mathrm{d}f$, we can write (7.11) as

$$f(H)_{11} = \frac{1}{2\pi} \iint\limits_{I_{\eta_0}^M} G(x+i\eta)_{11}\,\mathrm{d}\eta\,\mathrm{d}f(x) + \mathcal{O}_\prec\left(\eta_0 \|\mathrm{d}f\| + M^{-1}\|f\|_1\right).$$

In order to separate the leading order contribution from the fluctuation, we set

$$\Phi_N(z) = G(z)_{11} = \frac{1}{h_{11} - z - \langle h, \widehat{G}(z)h\rangle}, \qquad \widehat{\Phi}_N(z) = \frac{1}{-z - \widehat{m}_N(z)},$$

where $\widehat{m}_N(z) = \frac{1}{N}\operatorname{Tr}\widehat{G}(z)$ and observe that

$$\widehat{\Phi}_N(z) = \frac{1}{-z-m(z)} + \frac{\mathcal{O}_\prec\left(m(z) - \widehat{m}_N(z)\right)}{-z-m(z)} = m(z) + \mathcal{O}_\prec\left(\frac{1}{N\,|\eta|}\right) \tag{7.13}$$

and

$$\Phi_N(z) - \widehat{\Phi}_N(z) = m(z)^2\left[\langle h, \widehat{G}(z)h\rangle - \widehat{m}_N(z) - h_{11}\right] + \mathcal{O}_\prec\left(\frac{1}{N\,|\eta|}\right). \tag{7.14}$$

Thus $\widehat{\Phi}_N$ describes the leading order behavior, which is very close to a deterministic quantity, and the leading fluctuation is solely described by $\Phi_N - \widehat{\Phi}_N$. We then can write

$$f(H)_{11} = \Lambda_f^{(N)} + \frac{\Delta_f^{(N)}}{\sqrt{N}} + \mathcal{O}_\prec\left(\eta_0\|\mathrm{d}f\| + \frac{1}{M}\|f\|_1\right),$$

where

$$\Lambda_f^{(N)} := \frac{1}{2\pi}\iint\limits_{I_{\eta_0}^M} \widehat{\Phi}_N(z)\,\mathrm{d}\eta\,\mathrm{d}f(x) \quad \text{and} \quad \Delta_f^{(N)} := \frac{1}{2\pi}\iint\limits_{I_{\eta_0}^M} \sqrt{N}[\Phi_N - \widehat{\Phi}_N(z)]\,\mathrm{d}\eta\,\mathrm{d}f(x).$$

The reason for the normalization will become apparent later since in this way $\Delta_f^{(N)}$ is an object of order 1.

For the leading order term we use (7.13) and Proposition 7.3.1 to compute

$$\begin{aligned}\Lambda_f^{(N)} &= \frac{1}{2\pi}\int_{I_{\eta_0}^M} m(z)\,\mathrm{d}\eta\,\mathrm{d}f(x) + \mathcal{O}_\prec\left(\|\mathrm{d}f\|\int_{\eta_0}^M \frac{1}{N\eta}\,\mathrm{d}\eta\right) \\ &= \int f(x)\,\mu_{sc}(\mathrm{d}x) + \mathcal{O}_\prec\left(\left[\frac{|\log M| + |\log\eta_0|}{N} + \eta_0\right]\|\mathrm{d}f\| + \frac{1}{M}\|f\|_1\right).\end{aligned}$$

For the fluctuation we use (7.14) to compute

$$\begin{aligned}\Delta_f^{(N)} &= \frac{1}{2\pi}\int_{I_{\eta_0}^M} m(z)^2\sqrt{N}\left[\langle h,\widehat{G}(z)h\rangle - \widehat{m}_N(z) - h_{11}\right]\mathrm{d}\eta\,\mathrm{d}f(x)\\ &\quad + \mathcal{O}_\prec\left(\frac{|\log M| + |\log\eta|}{\sqrt{N}}\,\|\mathrm{d}f\|\right)\\ &= \widehat{\Delta}_f^{(N)} - \xi_{11}\frac{1}{2\pi}\int_{I_{\eta_0}^M} m(z)^2\,\mathrm{d}\eta\,\mathrm{d}f(x) + \mathcal{O}_\prec\left(\frac{|\log M| + |\log\eta|}{\sqrt{N}}\,\|\mathrm{d}f\|\right) \qquad (7.15)\\ &= \widehat{\Delta}_f^{(N)} + \xi_{11}\int f(x)x\,\mu_{sc}(\mathrm{d}x) + \mathcal{O}_\prec\left(\frac{|\log M| + |\log\eta|}{\sqrt{N}}\,\|\mathrm{d}f\| + \eta_0 + \frac{1}{M^2}\,\|f\|_1\right),\end{aligned}$$

where the last step followed from Lemma 7.3.2 and

$$\begin{aligned}\xi_{11} = \sqrt{N}h_{11}, \quad \widehat{\Delta}_f^{(N)} &:= \frac{1}{2\pi}\int_{I_{\eta_0}^M} m(z)^2X(z)\,\mathrm{d}\eta\,\mathrm{d}f(x),\\ X(z) = X^{(N)}(z) &= \langle h,\widehat{G}(z)h\rangle - \widehat{m}_N(z).\end{aligned}$$

We now concentrate on the computation of $\mathbf{E}\left(\widehat{\Delta}_f^{(N)}\right)^2$. We state the main estimate of $\mathbf{E}\,X(z)X(z')$ as a lemma.

Lemma 7.4.1. *Under the assumptions of Theorem 7.2.2 it holds that*

$$\begin{aligned}\mathbf{E}\,X(z)X(z') = \frac{m(z)^2m(z')^2}{1-m(z)m(z')} &+ \frac{\sigma_2^3m(z)^2m(z')^2}{1-\sigma_2m(z)m(z')}\\ &+ (\sigma_4-1)m(z)m(z') + \mathcal{O}_\prec\left(\frac{\Psi}{\sqrt{N}\Phi}\right),\end{aligned} \qquad (7.16)$$

where

$$\begin{aligned}\Psi &:= \frac{1}{\sqrt{|\eta\eta'|}}\left(\frac{1}{\sqrt{|\eta|}} + \frac{1}{\sqrt{|\eta'|}} + \frac{1}{\sqrt{N\,|\eta\eta'|}}\right)\\ \Phi &:= \mathbb{1}_{|x|,|x'|\le 2}\left(|\eta| + |\eta'| + |x-x'|^2\right) + \left[(|x|-2)_+ + (|x'|-2)_+\right]\end{aligned}$$

and $z = x + i\eta$, $z' = x' + i\eta'$.

We remark that in the $|x-x'|^2$ term in Φ could be replaced by $|x-x'|$ but we will not need this stronger bound here.

Proof of Lemma 7.4.1. From (6.29) we know that

$$\mathbf{E}\left[X(z)X(z')|\widehat{H}\right] = \frac{1}{N}\sum_{i\neq j}\left(\widehat{G}_{ij}\widehat{G}'_{ji} + \sigma_2^2\widehat{G}_{ij}\widehat{G}'_{ji}\right) + \frac{\sigma_4-1}{N}\sum_i \widehat{G}_{ii}\widehat{G}'_{ii} \qquad (7.17)$$

where, $\widehat{G}_{ij} := \widehat{G}(z)_{ij}$, $\widehat{G}'_{ij} := \widehat{G}(z')_{ij}$. The last term we directly estimate as

$$\frac{\sigma_4-1}{N}\sum_i \widehat{G}_{ii}\widehat{G}'_{ii} = (\sigma_4-1)m(z)m(z') + \mathcal{O}_\prec\left(\frac{1}{\sqrt{N\,|\eta|}} + \frac{1}{\sqrt{N\,|\eta'|}} + \frac{1}{N\sqrt{|\eta\eta'|}}\right). \qquad (7.18)$$

Furthermore, in Lemma 6.3.6 self-consistent equations for the first two terms on the rhs. of (7.17) were derived. We recall that

$$[1-m(z)m(z')]\frac{1}{N}\sum_{i\neq j}\widehat{G}_{ij}\widehat{G}'_{ji} = m(z)^2m(z')^2 + \mathcal{O}_\prec\left(\frac{\Psi}{\sqrt{N}}\right),$$
$$[1-\sigma_2 m(z)m(z')]\frac{1}{N}\sum_{i\neq j}\widehat{G}_{ij}\widehat{G}'_{ij} = \sigma_2 m(z)^2m(z')^2 + \mathcal{O}_\prec\left(\frac{\Psi}{\sqrt{N}}\right),$$

Using the straightforward inequality $|m(z)| \leq 1 - c\,|\eta|$, which holds for some small $c>0$ and z in the compact region $[-10,10]\times[-i,i]$, we find

$$|1-m(z)m(z')| \geq c(|\eta|+|\eta'|).$$

Since $|m|$ decays outside the spectrum $[-2,2]$ we have that $|m(z)|\leq 1-c'(|x|-2)_+$ for $|z|\leq 10$, and therefore

$$|1-m(z)m(z')| \geq c'(|x|-2)_+ + c'(|x'|-2)_+.$$

Moreover, in the remaining regime where both $|\eta|\,,|\eta'|\ll 1$ and $|x|\,,|x'|\leq 2$, it holds that

$$\begin{aligned}|1-m(z)m(z')| &\geq 1-\Re[m(z)m(z')] = 1-(\Re m(z))(\Re m(z'))+(\Im m(z))(\Im m(z'))\\ &\geq c''\Big(1-\frac{xx'}{4}\pm\frac{\sqrt{4-x^2}\sqrt{4-x'^2}}{4}\Big) \geq c''(x-x')^2,\end{aligned}$$

where the ± depends on the signs of η,η' and we allow for the constant c'' to change in the last inequality. This estimate follows from the explicit formula for $m(z)$. Putting these inequalities together, we therefore find a constant $C>0$ such that in the compact region $[-3,3]\times[-iM,iM]$ it holds that $C\,|1-m(z)m(z')|\geq\Phi$, from which we obtain

$$\begin{aligned}\frac{1}{N}\sum_{i\neq j}\widehat{G}_{ij}\widehat{G}'_{ji} &= \frac{m(z)^2m(z')^2}{1-m(z)m(z')} + \mathcal{O}_\prec\left(\frac{\Psi}{\sqrt{N}\Phi}\right),\\ \frac{1}{N}\sum_{i\neq j}\widehat{G}_{ij}\widehat{G}'_{ij} &= \frac{\sigma_2 m(z)^2m(z')^2}{1-\sigma_2 m(z)m(z')} + \mathcal{O}_\prec\left(\frac{\Psi}{\sqrt{N}\Phi}\right).\end{aligned} \tag{7.19}$$

Now (7.16) follows from combining (7.17), (7.18) and (7.19). □

Using Lemma 7.4.1 we then compute

$$\begin{aligned}\mathbf{E}\left(\widehat{\Delta}_f^{(N)}\right)^2 &= \frac{1}{(2\pi)^2}\iiiint_{I_{\eta_0}^M} m(z)^2m(z')^2\,\mathbf{E}\,X(z)X(z')\,\mathrm{d}\boldsymbol{\eta}\,\mathrm{d}f(\boldsymbol{x})\\ &= \frac{1}{(2\pi)^2}\iiiint_{I_{\eta_0}^M}\Big[\frac{m(z)^4m(z')^4}{1-m(z)m(z')}+\frac{\sigma_2^3 m(z)^4m(z')^4}{1-\sigma_2 m(z)m(z')}\\ &\quad+(\sigma_4-1)m(z)^3m(z')^3\Big]\,\mathrm{d}\boldsymbol{\eta}\,\mathrm{d}f(\boldsymbol{x}) + \mathcal{O}\left(\iiiint_{I_{\eta_0}^M}\frac{\Psi}{\sqrt{N}\Phi}\,\mathrm{d}\boldsymbol{\eta}\,\mathrm{d}f(\boldsymbol{x})\right),\end{aligned}$$

where $\mathrm{d}\boldsymbol{\eta}=\mathrm{d}\eta\,\mathrm{d}\eta'$ and $\mathrm{d}f(\boldsymbol{x})=\mathrm{d}f(x)\,\mathrm{d}f(x')$. To estimate the error term we have to compute

$$\iint_{-2}^{2}\iint_{\eta_0}^{M}\frac{1}{\eta+\eta'+|x-x'|^2}\frac{1}{\sqrt{\eta\eta'}}\left(\frac{1}{\sqrt{\eta}}+\frac{1}{\sqrt{\eta'}}+\frac{1}{\sqrt{N\eta\eta'}}\right)\mathrm{d}\boldsymbol{\eta}\,\mathrm{d}f(\boldsymbol{x})$$

and readily check that

$$\iiiint_{I_{\eta_0}^M} \frac{\Psi}{\sqrt{N}\Phi}\,\mathrm{d}\boldsymbol{\eta}\,\mathrm{d}f(\boldsymbol{x}) \prec \begin{cases} (|\log M| + |\log \eta_0|)/\sqrt{N} & \text{if } f' \text{ is bounded,} \\ (|\log M| + |\log \eta_0|)/\sqrt{N\eta_0} & \text{else.} \end{cases}$$

By using Lemma 7.3.2 and organizing the contributions from the boundary terms at η_0 and $-\eta_0$, we find that the leading order of $\mathbf{E}(\widehat{\Delta}_f^{(N)})^2$ becomes

$$\begin{aligned} &\frac{1}{2\pi^2}\Re \iint\limits_{-3}^{3} f(x)f(x')\Bigg(\left[\frac{m(z_0)^4 m(\overline{z_0'})^4}{1 - m(z_0)m(\overline{z_0'})} + \frac{\sigma_2^3 m(z_0)^4 m(\overline{z_0'})^4}{1 - \sigma_2 m(z_0)m(\overline{z_0'})} + (\sigma_4 - 1)m(z)^3 m(\overline{z_0'})^3\right] \\ &- \left[\frac{m(z_0)^4 m(z_0')^4}{1 - m(z_0)m(z_0')} + \frac{\sigma_2^3 m(z_0)^4 m(z_0')^4}{1 - \sigma_2 m(z_0)m(z_0')} + (\sigma_4 - 1)m(z)^3 m(z_0')^3\right]\Bigg)\mathrm{d}\boldsymbol{x} + \mathcal{O}_\prec\left(\frac{\|f\|_1}{M^3}\right), \end{aligned} \tag{7.20}$$

where $z_0 = x + i\eta_0$ and $z_0' = x' + i\eta_0$. Since

$$\frac{a^4}{1-a} = \frac{a}{1-a} - a - a^2 - a^3$$

and for any fixed $k \in \mathbb{N}$

$$\begin{aligned} &\frac{1}{2\pi^2}\Re \iint\limits_{-3}^{3} f(x)f(x')\left[m(z_0)^k m(\overline{z_0'})^k - m(z_0)^k m(z_0')^k\right]\mathrm{d}\boldsymbol{x} \\ &= \left(\frac{1}{\pi}\Im \int_{-2}^{2} f(x)m(x)^k\,\mathrm{d}x\right)^2 + \mathcal{O}_\prec(\eta_0) \end{aligned}$$

we can conclude that (7.20) becomes

$$\begin{aligned} &\frac{1}{2\pi^2}\Re \iint\limits_{-3}^{3} f(x)f(x')\left(\frac{m(z_0)m(\overline{z_0'})}{1 - m(z_0)m(\overline{z_0'})} - \frac{m(z_0)m(z_0')}{1 - m(z_0)m(z_0')}\right)\mathrm{d}\boldsymbol{x} \\ &+ \frac{1}{2\pi^2}\Re \iint\limits_{-3}^{3} f(x)f(x')\left(\frac{m(z_0)m(\overline{z_0'})}{1 - \sigma_2 m(z_0)m(\overline{z_0'})} - \frac{m(z_0)m(z_0')}{1 - \sigma_2 m(z_0)m(z_0')}\right)\mathrm{d}\boldsymbol{x} \\ &- 2\left(\frac{1}{\pi}\Im \int_{\mathbb{R}} f(x)m(x)\,\mathrm{d}x\right)^2 - (1 + \sigma_2)\left(\frac{1}{\pi}\Im \int_{\mathbb{R}} f(x)m(x)^2\,\mathrm{d}x\right)^2 \\ &+ (\sigma_4 - 2 - \sigma_2^2)\left(\frac{1}{\pi}\Im \int_{\mathbb{R}} f(x)m(x)^3\,\mathrm{d}x\right)^2 + \mathcal{O}\left(\frac{\|f\|_{L^1}}{M^3} + \eta_0\right). \end{aligned} \tag{7.21}$$

The first term of (7.21) was already computed in (6.39). The computation of the second term is very similar to the first one and the remaining terms are routine calculations. We arrive at

$$\begin{aligned} \mathbf{E}\left(\widehat{\Delta}_f^{(N)}\right)^2 &= V_{f,1} + V_{f,1}^{(\sigma_2)} - 2V_{f,2} - (1 + \sigma_2)V_{f,3} + (\sigma_4 - 2 - \sigma_2^2)V_{f,4} \\ &\quad + \mathcal{O}\left(\eta_0 + \frac{\|f\|_1}{M^3} + \frac{|\log M| + |\log \eta_0|}{\sqrt{N\eta_0}}\|\mathrm{d}f\|\right) \end{aligned}$$

in the general case and

$$\mathbf{E}\left(\widehat{\Delta}_f^{(N)}\right)^2 = V_{f,1} + V_{f,1}^{(\sigma_2)} - 2V_{f,2} - (1+\sigma_2)V_{f,3} + (\sigma_4 - 2 - \sigma_2^2)V_{f,4} + \mathcal{O}\left(\eta_0 + \frac{\|f\|_1}{M^3} + \frac{|\log M| + |\log \eta_0|}{\sqrt{N}} \|f'\|_{L^\infty}\right)$$

in the case of f with bounded derivative $f' \in L^\infty([-3,3])$, where

$$\begin{aligned} V_{f,1} &:= \int f(x)^2 \, \mu_{sc}(\mathrm{d}x), \\ V_{f,1}^{(\sigma_2)} &:= \iint \frac{f(x)f(y)(1-\sigma_2^2)}{1 - xy\sigma_2 + (x^2+y^2-2)\sigma_2^2 - xy\sigma_2^3 + \sigma_2^4} \, \mu_{sc}(\mathrm{d}x) \, \mu_{sc}(\mathrm{d}y) \\ V_{f,2} &:= \left(\int f(x) \, \mu_{sc}(\mathrm{d}x)\right)^2, \ V_{f,3} := \left(\int f(x) x \, \mu_{sc}(\mathrm{d}x)\right)^2, \\ V_{f,4} &:= \left(\int f(x)(x^2-1) \, \mu_{sc}(\mathrm{d}x)\right)^2. \end{aligned} \tag{7.22}$$

We note that $V_{f,1}^{(\sigma_2)}$ simplifies to $V_{f,1}^{(1)} = V_{f,1}$ and $V_{f,1}^{(0)} = V_{f,2}$ in the two important cases $\sigma_2 = 0, 1$.

We now choose $M = N$ and η_0 depending on the regularity of f. In the general case of $f \in BV([-3,3])$ it turns out that $\eta_0 = N^{-2/3}$ is optimal, whereas for f with bounded derivative, we can go all the way down to $\eta_0 = N^{-1+\epsilon}$ for any small $\epsilon > 0$. Thus

$$\mathbf{E}\left(\widehat{\Delta}_f^{(N)}\right)^2 = \mathbf{E}\left(\widehat{\Delta}_f\right)^2 + \begin{cases} \mathcal{O}_\prec\left(N^{-1/2}\right) & \text{if } f' \in L^\infty([-3,3]), \\ \mathcal{O}_\prec\left(N^{-1/6}\right) & \text{else.} \end{cases} \tag{7.23}$$

where $\widehat{\Delta}_f$ is a centered Gaussian of variance

$$\mathbf{E}\left(\widehat{\Delta}_f\right)^2 = V_{f,1} + V_{f,1}^{(\sigma_2)} - 2V_{f,2} - (1+\sigma_2)V_{f,3} + (\sigma_4 - 2 - \sigma_2^2)V_{f,4}.$$

For higher moments we recall the following Wick type factorization Lemma from [DS1].

Lemma 7.4.2. *For $k \geq 2$ and $z_1, \dots, z_k \in \mathbb{C}$ with $z_l = x_l \pm i\eta_l$ and $\eta_l > 0$ we have that*

$$\mathbf{E}[X(z_1) \dots X(z_k)] = \sum_{\pi \in P_2([k])} \prod_{\{a,b\} \in \pi} \mathbf{E}[X(z_a)X(z_b)] + \mathcal{O}_\prec\left(\frac{1}{\sqrt{N\boldsymbol{\eta}}} \sum_{a \neq b} \frac{1}{\sqrt{\eta_a}\Phi_{a,b}}\right), \tag{7.24}$$

where $[k] := \{1, \dots, k\}$, $\boldsymbol{\eta} = \eta_1 \dots \eta_k$, $P_2(L)$ are the partitions of a set L into subsets of size 2 and

$$\Phi_{a,b} := \mathbb{1}_{|x_a|,|x_b| \leq 2}\left(|\eta_a| + |\eta_b| + |x_a - x_b|^2\right) + \left[(|x_a| - 2)_+ + (|x_b| - 2)_+\right].$$

The error term in (7.24) is slightly stronger than that in [DS1] since the $\Phi_{a,b}$ includes a $|x_a - x_b|^2$. This strengthening follows along the lines of the original proof by using the more precise analysis of the self consistent equation outlined in Lemma 7.4.1. We check that integrating the error term from (7.24) over $(I_{\eta_0}^M)^k$, with η_0 being chosen as above according to the regularity of f, again gives asymptotically $N^{-1/2}$ in the case of bounded f' and $N^{-1/6}$ in the general case. By integrating the Wick type product and using (7.23) we therefore arrive at

$$\mathbf{E}\left(\widehat{\Delta}_f^{(N)}\right)^k = \mathbf{E}\left(\widehat{\Delta}_f\right)^k + \begin{cases} \mathcal{O}_\prec\left(N^{-1/2}\right) & \text{if } f' \in L^\infty([-3,3]), \\ \mathcal{O}_\prec\left(N^{-1/6}\right) & \text{else.} \end{cases}$$

We note that the error terms are implicitly k-dependent. By counting the number of pair partitions we find that, to the leading order in N, the implicit coefficients scale like $C^k(k/2)!$ with a constant depending on f.

Recalling (7.15) and the definition of $T_f^{(N)}$ from (7.6), we conclude that the overall fluctuations have moments

$$\mathbf{E}\left(T_f^{(N)}\right)^k = \mathbf{E}\left(\widehat{\Delta}_f\right)^k + \begin{cases} \mathcal{O}\left(C^k(k/2)!N^{-1/2}\right) & \text{if } f' \in L^\infty([-3,3]), \\ \mathcal{O}\left(C^k(k/2)!N^{-1/6}\right) & \text{else.} \end{cases} \tag{7.25}$$

Let $\phi_N(t)$ denote the characteristic function of $T_f^{(N)}$ and $\phi(t)$ the characteristic function of the Gaussian variable $\widehat{\Delta}_f$. Then the moment bound (7.25) implies that

$$|\phi_N(t) - \phi(t)| \le CN^{-1/6}te^{Ct^2}$$

with some constant C depending on f. Using the well-known bound (see, e.g., [173, Theorem 1.4.13.] and the references therein)

$$d_L(F,G) \le \frac{1}{\pi}\int_0^T |\phi_F(t) - \phi_G(t)| \frac{\mathrm{d}t}{t} + \frac{2e\log T}{T}$$

for any two distributions F and G with characteristic functions ϕ_F and ϕ_G, we immediately obtain (7.7) by choosing $T = c\sqrt{\log N}$. This completes the proof of Theorem 7.2.2.

7.5 Off-Diagonal Entries

For the decomposition

$$H = \begin{pmatrix} h_{11} & h_{12} & h_1^* \\ h_{21} & h_{22} & h_2^* \\ h_1 & h_2 & \widehat{H} \end{pmatrix}$$

we find from the Schur complement formula that

$$G(z)_{12} = -\frac{g_{12}}{g_{11}g_{22} - g_{12}g_{21}} = -m(z)^2 g_{12} + \mathcal{O}_\prec\left(\frac{1}{N\,|\eta|}\right),$$

where $g_{ij} := h_{ij} - \delta_{ij}z - \langle h_i, G(z)h_j\rangle$. We now set $Y(z) = Y^{(N)}(z) := \sqrt{N}\,\langle h_1, \widehat{G}(z)h_2\rangle$ and begin to compute (all summation indices run from 3 to N)

$$\begin{aligned}\mathbf{E}\left[Y(z)Y(z')|\widehat{H}\right] &= N\sum_{a,b,c,d}\mathbf{E}\left[h_{1a}\widehat{G}_{ab}h_{b2}h_{1c}\widehat{G}'_{cd}h_{d2}|\widehat{H}\right] \\ &= \frac{\sigma_2^2}{N}\sum_{a,b}\widehat{G}_{ab}\widehat{G}'_{ab} + \mathcal{O}_\prec\left(\frac{\Psi}{N}\right) = \frac{\sigma_2^2 m(z)m(z')}{1-\sigma_2 m(z)m(z')} + \mathcal{O}_\prec\left(\frac{\Psi}{\sqrt{N}\Phi}\right)\end{aligned}$$

and

$$\begin{aligned}\mathbf{E}\left[Y(z)\overline{Y(\overline{z'})}|\widehat{H}\right] &= N\sum_{a,b,c,d}\mathbf{E}\left[h_{1a}\widehat{G}_{ab}h_{b2}h_{2c}\widehat{G}'_{cd}h_{d1}|\widehat{H}\right]\\ &= \frac{1}{N}\sum_{a,b}\widehat{G}_{ab}\widehat{G}'_{ba} + \mathcal{O}_\prec\left(\frac{\Psi}{N}\right) = \frac{m(z)m(z')}{1-m(z)m(z')} + \mathcal{O}_\prec\left(\frac{\Psi}{\sqrt{N}\Phi}\right).\end{aligned}$$

For both estimates we made use of the fact the h_{ab} are centered and therefore have to appear at least twice to have non-zero expectation. The main contribution comes from the pairing $a = d, b = c$. Some exceptional pairings, such as the four-pairing $a = b = c = d$, were incorporated in the error term by their reduced combinatorics. From Proposition 7.3.1 we then find that

$$f(H)_{12} = \frac{1}{\pi}\iint_{I_{\eta_0}^M} m(z)^2\left[\langle h_1, \widehat{G}(z)h_2\rangle - h_{12}\right]\mathrm{d}\eta\,\mathrm{d}f(x) + \mathcal{O}_\prec\left(\frac{\|\mathrm{d}f\|}{N}\right).$$

For the second term it follows, just as before, that

$$\frac{1}{\pi}\iint_{I_{\eta_0}^M} m(z)^2 h_{12}\,\mathrm{d}\eta\,\mathrm{d}f(x) = h_{12}\int f(x)x\,\mu_{sc}(\mathrm{d}x) + \mathcal{O}_\prec(\eta_0).$$

For the first term we set

$$\widetilde{\Delta}_f^{(N)} := \iint_{I_{\eta_0}^M} Y(z)\,\mathrm{d}\eta\,\mathrm{d}f(x)$$

and following a computation very similar to that of $\widehat{\Delta}_F^{(N)}$ we arrive at

$$\mathbf{E}\left(\widetilde{\Delta}_f^{(N)}\right)^2 = V_{f,1}^{(\sigma_2)} - V_{f,2} - \sigma_2 V_{f,3} + \begin{cases}\mathcal{O}_\prec\left(N^{-1/2}\right) & \text{if } f' \in L^\infty([-3,3]),\\ \mathcal{O}_\prec\left(N^{-1/6}\right) & \text{else.}\end{cases}$$

Similarly we find that

$$\mathbf{E}\left|\widetilde{\Delta}_f^{(N)}\right|^2 = V_{f,1} - V_{f,2} - V_{f,3} + \begin{cases}\mathcal{O}_\prec\left(N^{-1/2}\right) & \text{if } f' \in L^\infty([-3,3]),\\ \mathcal{O}_\prec\left(N^{-1/6}\right) & \text{else.}\end{cases}.$$

Finally, due to a Wick type theorem for $Y(z)$ which is proved along the lines of Lemma 7.4.2 we arrive at

$$\mathbf{E}\left(S_f^{(N)}\right)^k\left(\overline{S_f^{(N)}}\right)^l = \mathbf{E}\left(\widetilde{\Delta}_f\right)^k\left(\overline{\widetilde{\Delta}_f}\right)^l + \begin{cases}\mathcal{O}_\prec\left(N^{-1/2}\right) & \text{if } f' \in L^\infty([-3,3]),\\ \mathcal{O}_\prec\left(N^{-1/6}\right) & \text{else,}\end{cases}$$

where $\widetilde{\Delta}_f$ is a centered complex Gaussian such that

$$\mathbf{E}\,\widetilde{\Delta}_f^2 = V_{f,1}^{(\sigma_2)} - V_{f,2} - \sigma_2 V_{f,3}, \quad \mathbf{E}\left|\widetilde{\Delta}_f\right|^2 = V_{f,1} - V_{f,2} - V_{f,3}.$$

We have proven Theorem 7.2.3.

References

[1] A. Adhikari and Z. Che, *The edge universality of correlated matrices*, preprint (2017), arXiv:1712.04889.

[2] B. Adlam and Z. Che, *Spectral statistics of sparse random graphs with a general degree distribution*, preprint (2015), arXiv:1509.03368.

[3] M. Adler, M. Cafasso, and P. van Moerbeke, *From the Pearcey to the Airy process*, Electron. J. Probab. **16**, no. 36, 1048–1064 (2011), MR2820069.

[4] M. Adler, P. L. Ferrari, and P. van Moerbeke, *Airy processes with wanderers and new universality classes*, Ann. Probab. **38**, 714–769 (2010), MR2642890.

[5] M. Adler and P. van Moerbeke, *PDEs for the Gaussian ensemble with external source and the Pearcey distribution*, Comm. Pure Appl. Math. **60**, 1261–1292 (2007), MR2337504.

[6] O. H. Ajanki, L. Erdős, and T. Krüger, *Local spectral statistics of Gaussian matrices with correlated entries*, J. Stat. Phys. **163**, 280–302 (2016), MR3478311.

[7] O. H. Ajanki, L. Erdős, and T. Krüger, *Quadratic vector equations on complex upper half-plane*, preprint (2015), arXiv:1506.05095.

[8] O. H. Ajanki, L. Erdős, and T. Krüger, *Stability of the matrix Dyson equation and random matrices with correlations*, Probab. Theory Related Fields **173**, 293–373 (2019), MR3916109.

[9] O. H. Ajanki, L. Erdős, and T. Krüger, *Universality for general Wigner-type matrices*, Probab. Theory Related Fields **169**, 667–727 (2017), MR3719056.

[10] O. Ajanki, L. Erdős, and T. Krüger, *Singularities of solutions to quadratic vector equations on the complex upper half-plane*, Comm. Pure Appl. Math. **70**, 1672–1705 (2017), MR3684307.

[11] J. Alt, *The local semicircle law for random matrices with a fourfold symmetry*, J. Math. Phys. **56**, 103301, 20 (2015), MR3406427.

[12] J. Alt, L. Erdős, and T. Krüger, *The Dyson equation with linear self-energy: Spectral bands, edges and cusps*, preprint (2018), arXiv:1804.07752.

[13] J. Alt, L. Erdős, T. Krüger, and Y. Nemish, *Location of the spectrum of Kronecker random matrices*, preprint (2017), arXiv:1706.08343.

[14] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, Vol. 118, Cambridge Studies in Advanced Mathematics (Cambridge University Press, Cambridge, 2010), pp. xiv+492, MR2760897.

[15] G. W. Anderson and O. Zeitouni, *A CLT for a band matrix model*, Probab. Theory Related Fields **134**, 283–338 (2006), MR2222385.

[16] G. W. Anderson and O. Zeitouni, *A law of large numbers for finite-range dependent random matrices*, Comm. Pure Appl. Math. **61**, 1118–1154 (2008), MR2417889.

[17] A. Auffinger, G. Ben Arous, and S. Péché, *Poisson convergence for the largest eigenvalues of heavy tailed random matrices*, Ann. Inst. Henri Poincaré Probab. Stat. **45**, 589–610 (2009), MR2548495.

[18] Z. D. Bai and J. W. Silverstein, *Exact separation of eigenvalues of large-dimensional sample covariance matrices*, Ann. Probab. **27**, 1536–1555 (1999), MR1733159.

[19] Z. D. Bai and Y. Q. Yin, *Convergence to the semicircle law*, Ann. Probab. **16**, 863–875 (1988), MR929083.

[20] J. Baik, P. Deift, and K. Johansson, *On the distribution of the length of the second row of a Young diagram under Plancherel measure*, Geom. Funct. Anal. **10**, 702–731 (2000), MR1791137.

[21] J. Baik, T. Kriecherbauer, K. T.-R. McLaughlin, and P. D. Miller, *Discrete orthogonal polynomials*, Vol. 164, Annals of Mathematics Studies, Asymptotics and applications (Princeton University Press, Princeton, NJ, 2007), pp. viii+170, MR2283089.

[22] J. Baik, P. Deift, and K. Johansson, *On the distribution of the length of the longest increasing subsequence of random permutations*, J. Amer. Math. Soc. **12**, 1119–1178 (1999), MR1682248.

[23] 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**, 2700–2726 (2015), MR3332852.

[24] Z. Bao, G. Pan, and W. Zhou, *Central limit theorem for partial linear eigenvalue statistics of Wigner matrices*, J. Stat. Phys. **150**, 88–129 (2013), MR3018879.

[25] Y. Baryshnikov, *GUEs and queues*, Probab. Theory Related Fields **119**, 256–274 (2001), MR1818248.

[26] R. Bauerschmidt, J. Huang, A. Knowles, and H.-T. Yau, *Bulk eigenvalue statistics for random regular graphs*, Ann. Probab. **45**, 3626–3663 (2017), MR3729611.

[27] R. Bauerschmidt, A. Knowles, and H.-T. Yau, *Local semicircle law for random regular graphs*, Comm. Pure Appl. Math. **70**, 1898–1960 (2017), MR3688032.

[28] F. Bekerman, A. Figalli, and A. Guionnet, *Transport maps for β-matrix models and universality*, Comm. Math. Phys. **338**, 589–619 (2015), MR3351052.

[29] F. Bekerman, *Transport maps for β-matrix models in the multi-cut regime*, Random Matrices Theory Appl. **7**, 1750013, 36 (2018), MR3756421.

[30] R. Bhatia, *Matrix analysis*, Vol. 169, Graduate Texts in Mathematics (Springer-Verlag, New York, 1997), pp. xii+347, MR1477662.

[31] P. Biane, *On the free convolution with a semi-circular distribution*, Indiana Univ. Math. J. **46**, 705–718 (1997), MR1488333.

[32] O. Bohigas, M.-J. Giannoni, and C. Schmit, *Characterization of chaotic quantum spectra and universality of level fluctuation laws*, Phys. Rev. Lett. **52**, 1–4 (1984), MR730191.

[33] G. Bonnet, F. David, and B. Eynard, *Breakdown of universality in multi-cut matrix models*, J. Phys. A **33**, 6739–6768 (2000), MR1790279.

[34] A. Borodin, *CLT for spectra of submatrices of Wigner random matrices*, Mosc. Math. J. **14**, 29–38, 170 (2014), MR3221945.

[35] A. Borodin, "CLT for spectra of submatrices of Wigner random matrices, II: Stochastic evolution", in *Random matrix theory, interacting particle systems, and integrable systems*, Vol. 65, Math. Sci. Res. Inst. Publ. (Cambridge Univ. Press, New York, 2014), pp. 57–69, MR3380682.

[36] A. Borodin and V. Gorin, *General β-Jacobi corners process and the Gaussian free field*, Comm. Pure Appl. Math. **68**, 1774–1844 (2015), MR3385342.

[37] A. Borodin and V. Gorin, "Lectures on integrable probability", in *Probability and statistical physics in St. Petersburg*, Vol. 91, Proc. Sympos. Pure Math. (Amer. Math. Soc., Providence, RI, 2016), pp. 155–214, MR3526828.

[38] A. Borodin, A. Okounkov, and G. Olshanski, *Asymptotics of Plancherel measures for symmetric groups*, J. Amer. Math. Soc. **13**, 481–515 (2000), MR1758751.

[39] G. Borot and A. Guionnet, *Asymptotic expansion of beta matrix models in the multi-cut regime*, preprint (2013), arXiv:1303.1045.

[40] P. Bourgade and H.-T. Yau, *The eigenvector moment flow and local quantum unique ergodicity*, Comm. Math. Phys. **350**, 231–278 (2017), MR3606475.

[41] P. Bourgade, L. Erdős, and H.-T. Yau, *Edge universality of beta ensembles*, Comm. Math. Phys. **332**, 261–353 (2014), MR3253704.

[42] P. Bourgade, L. Erdős, and H.-T. Yau, *Universality of general β-ensembles*, Duke Math. J. **163**, 1127–1190 (2014), MR3192527.

[43] P. Bourgade, L. Erdős, H.-T. Yau, and J. Yin, *Fixed energy universality for generalized Wigner matrices*, Comm. Pure Appl. Math. **69**, 1815–1881 (2016), MR3541852.

[44] P. Bourgade, L. Erdős, H.-T. Yau, and J. Yin, *Universality for a class of random band matrices*, Adv. Theor. Math. Phys. **21**, 739–800 (2017), MR3695802.

[45] P. Bourgade, H.-T. Yau, and J. Yin, *Local circular law for random matrices*, Probab. Theory Related Fields **159**, 545–595 (2014), MR3230002.

[46] P. Bourgade, H.-T. Yau, and J. Yin, *Random band matrices in the delocalized phase, I: Quantum unique ergodicity and universality*, preprint (2018), arXiv:1807.01559.

[47] A. Boutet de Monvel, A. Khorunzhy, and V. Vasilchuk, *Limiting eigenvalue distribution of random matrices with correlated entries*, Markov Process. Related Fields **2**, 607–636 (1996), MR1431189.

[48] R. C. Bradley, *Basic properties of strong mixing conditions. A survey and some open questions*, Probab. Surv. **2**, Update of, and a supplement to, the 1986 original, 107–144 (2005), MR2178042.

[49] E. Brézin and S. Hikami, *Level spacing of random matrices in an external source*, Phys. Rev. E (3) **58**, 7176–7185 (1998), MR1662382.

[50] E. Brézin and S. Hikami, *Universal singularity at the closure of a gap in a random matrix theory*, Phys. Rev. E (3) **57**, 4140–4149 (1998), MR1618958.

[51] A. Bufetov, *Kerov's interlacing sequences and random matrices*, J. Math. Phys. **54**, 113302, 10 (2013), MR3137038.

[52] M. Capitaine and S. Péché, *Fluctuations at the edges of the spectrum of the full rank deformed GUE*, Probab. Theory Related Fields **165**, 117–161 (2016), MR3500269.

[53] G. Casati, I. Guarneri, F. Izrailev, and R. Scharf, *Scaling behavior of localization in quantum chaos*, Phys. Rev. Lett. **64**, 5–8 (1990), PMID10041259.

[54] G. Casati, L. Molinari, and F. Izrailev, *Scaling properties of band random matrices*, Phys. Rev. Lett. **64**, 1851–1854 (1990), MR1046365.

[55] Z. Che, *Universality of random matrices with correlated entries*, Electron. J. Probab. **22**, Paper No. 30, 38 (2017), MR3629874.

[56] W. Choi, C. Yin, I. Hooper, W. Barnes, and J. Bertolotti, *Absence of anderson localization in certain random lattices*, Physical review. E **96**, 022122 (2017), PMID28950489.

[57] T. Claeys, A. B. J. Kuijlaars, K. Liechty, and D. Wang, *Propagation of singular behavior for Gaussian perturbations of random matrices*, Comm. Math. Phys. **362**, 1–54 (2018), MR3833603.

[58] T. Claeys, T. Neuschel, and M. Venker, *Boundaries of sine kernel universality for gaussian perturbations of Hermitian matrices*, preprint (2017), arXiv:1712.08432.

[59] R. Couillet and M. Debbah, *Random matrix methods for wireless communications* (Cambridge University Press, Cambridge, 2011), pp. xxii+539, MR2884783.

[60] P. A. Deift, *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach*, Vol. 3, Courant Lecture Notes in Mathematics (New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 1999), pp. viii+273, MR1677884.

[61] P. Deift, T. Kriecherbauer, and K. T.-R. McLaughlin, *New results on the equilibrium measure for logarithmic potentials in the presence of an external field*, J. Approx. Theory **95**, 388–475 (1998), MR1657691.

[62] P. Deift, T. Kriecherbauer, K. T.-R. McLaughlin, S. Venakides, and X. Zhou, *Uniform asymptotics for polynomials orthogonal with respect to varying exponential weights and applications to universality questions in random matrix theory*, Comm. Pure Appl. Math. **52**, 1335–1425 (1999), MR1702716.

[63] P. Deift and D. Gioev, *Random matrix theory: invariant ensembles and universality*, Vol. 18, Courant Lecture Notes in Mathematics (Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2009), pp. x+217, MR2514781.

[64] P. Deift and D. Gioev, *Universality at the edge of the spectrum for unitary, orthogonal, and symplectic ensembles of random matrices*, Comm. Pure Appl. Math. **60**, 867–910 (2007), MR2306224.

[65] 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), MR0337229.

[66] E. Duse, K. Johansson, and A. Metcalfe, *The cusp-Airy process*, Electron. J. Probab. **21**, Paper No. 57, 50 (2016), MR3546394.

[67] F. J. Dyson, *A Brownian-motion model for the eigenvalues of a random matrix*, J. Mathematical Phys. **3**, 1191–1198 (1962), MR0148397.

[68] A. Edelman and N. R. Rao, *Random matrix theory*, Acta Numer. **14**, 233–297 (2005), MR2168344.

[69] K. Efetov, *Supersymmetry in disorder and chaos* (Cambridge University Press, Cambridge, 1997), pp. xiv+441, MR1628498.

[70] L. Erdős, A. Knowles, and H.-T. Yau, *Averaging fluctuations in resolvents of random band matrices*, Ann. Henri Poincaré **14**, 1837–1926 (2013), MR3119922.

[71] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Spectral statistics of Erdős-Rényi graphs I: Local semicircle law*, Ann. Probab. **41**, 2279–2375 (2013), MR3098073.

[72] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Spectral statistics of Erdős-Rényi Graphs II: Eigenvalue spacing and the extreme eigenvalues*, Comm. Math. Phys. **314**, 587–640 (2012), MR2964770.

[73] 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**, no. 59, 58 (2013), MR3068390.

[74] L. Erdős, S. Péché, J. A. Ramírez, B. Schlein, and H.-T. Yau, *Bulk universality for Wigner matrices*, Comm. Pure Appl. Math. **63**, 895–925 (2010), MR2662426.

[75] L. Erdős, B. Schlein, and H.-T. Yau, *Universality of random matrices and local relaxation flow*, Invent. Math. **185**, 75–119 (2011), MR2810797.

[76] 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. Henri Poincaré Probab. Stat. **48**, 1–46 (2012), MR2919197.

[77] L. Erdős and K. Schnelli, *Universality for random matrix flows with time-dependent density*, Ann. Inst. Henri Poincaré Probab. Stat. **53**, 1606–1656 (2017), MR3729630.

[78] L. Erdős and H.-T. Yau, *A dynamical approach to random matrix theory*, Vol. 28, Courant Lecture Notes in Mathematics (Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2017), pp. ix+226, MR3699468.

[79] L. Erdős and H.-T. Yau, *Gap universality of generalized Wigner and β-ensembles*, J. Eur. Math. Soc. (JEMS) **17**, 1927–2036 (2015), MR3372074.

[80] L. Erdős and H.-T. Yau, *Universality of local spectral statistics of random matrices*, Bull. Amer. Math. Soc. (N.S.) **49**, 377–414 (2012), MR2917064.

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

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

[83] L. Erdős, H.-T. Yau, and J. Yin, *Universality for generalized Wigner matrices with Bernoulli distribution*, J. Comb. **2**, 15–81 (2011), MR2847916.

[84] P. Erdős and A. Hajnal, *On chromatic number of graphs and set-systems*, Acta Math. Acad. Sci. Hungar **17**, 61–99 (1966), MR0193025.

[85] P. J. Forrester, *The spectrum edge of random matrix ensembles*, Nuclear Phys. B **402**, 709–728 (1993), MR1236195.

[86] W. Fulton, *Young tableaux*, Vol. 35, London Mathematical Society Student Texts, With applications to representation theory and geometry (Cambridge University Press, Cambridge, 1997), pp. x+260, MR1464693.

[87] D. Geudens and L. Zhang, *Transitions between critical kernels: From the tacnode kernel and critical kernel in the two-matrix model to the Pearcey kernel*, Int. Math. Res. Not. IMRN, 5733–5782 (2015), MR3384456.

[88] L. Giraitis and D. Surgailis, "Multivariate Appell polynomials and the central limit theorem", in *Dependence in probability and statistics (Oberwolfach, 1985)*, Vol. 11, Progr. Probab. Statist. (Birkhäuser Boston, Boston, MA, 1986), pp. 21–71, MR0899984.

[89] V. L. Girko, *Asymptotics of the distribution of the spectrum of random matrices*, Uspekhi Mat. Nauk **44**, 7–34, 256 (1989), MR1023102.

[90] V. L. Girko, *The circular law*, Teor. Veroyatnost. i Primenen. **29**, 669–679 (1984), MR773436.

[91] V. L. Girko, *Theory of stochastic canonical equations. Vol. I*, Vol. 535, Mathematics and its Applications (Kluwer Academic Publishers, Dordrecht, 2001), pp. xxiv+497, MR1887675.

[92] V. Gorin and L. Zhang, *Interlacing adjacent levels of β-Jacobi corners processes*, Probab. Theory Related Fields **172**, 915–981 (2018), MR3877550.

[93] A. Guionnet and J. Huang, *Rigidity and edge universality of discrete β-ensembles*, preprint (2017), arXiv:1705.05527.

[94] W. Hachem, P. Loubaton, and J. Najim, *The empirical eigenvalue distribution of a Gram matrix: From independence to stationarity*, Markov Process. Related Fields **11**, 629–648 (2005), MR2191967.

[95] W. Hachem, A. Hardy, and J. Najim, "A survey on the eigenvalues local behavior of large complex correlated Wishart matrices", in *Modélisation Aléatoire et Statistique—Journées MAS 2014*, Vol. 51, ESAIM Proc. Surveys (EDP Sci., Les Ulis, 2015), pp. 150–174, MR3440796.

[96] W. Hachem, A. Hardy, and J. Najim, *Large complex correlated Wishart matrices: Fluctuations and asymptotic independence at the edges*, Ann. Probab. **44**, 2264–2348 (2016), MR3502605.

[97] W. Hachem, A. Hardy, and J. Najim, *Large complex correlated Wishart matrices: The Pearcey kernel and expansion at the hard edge*, Electron. J. Probab. **21**, Paper No. 1, 36 (2016), MR3485343.

[98] Y. He and A. Knowles, *Mesoscopic eigenvalue statistics of Wigner matrices*, Ann. Appl. Probab. **27**, 1510–1550 (2017), MR3678478.

[99] Y. He, A. Knowles, and M. Marcozzi, *Local law and complete eigenvector delocalization for supercritical erdős-rényi graphs*, preprint (2018), arXiv:1808.09437.

[100] Y. He, A. Knowles, and R. Rosenthal, *Isotropic self-consistent equations for mean-field random matrices*, Probab. Theory Related Fields **171**, 203–249 (2018), MR3800833.

[101] B. Helffer and J. Sjöstrand, "Équation de Schrödinger avec champ magnétique et équation de Harper", in *Schrödinger operators (Sønderborg, 1988)*, Vol. 345, Lecture Notes in Phys. (Springer, Berlin, 1989), pp. 118–197, MR1037319.

[102] J. W. Helton, R. Rashidi Far, and R. Speicher, *Operator-valued semicircular elements: Solving a quadratic matrix equation with positivity constraints*, Int. Math. Res. Not. IMRN, Art. ID rnm086, 15 (2007), MR2376207.

[103] J. Huang, B. Landon, and H.-T. Yau, *Bulk universality of sparse random matrices*, J. Math. Phys. **56**, 123301, 19 (2015), MR3429490.

[104] J. Huang, B. Landon, and H.-T. Yau, *Transition from Tracy-Widom to Gaussian fluctuations of extremal eigenvalues of sparse Erdős-Rényi graphs*, preprint (2017), arXiv:1712.03936.

[105] V. Ivanov and G. Olshanski, "Kerov's central limit theorem for the Plancherel measure on Young diagrams", in *Symmetric functions 2001: Surveys of developments and perspectives*, Vol. 74, NATO Sci. Ser. II Math. Phys. Chem. (Kluwer Acad. Publ., Dordrecht, 2002), pp. 93–151, MR2059361.

[106] I. Jana, K. Saha, and A. Soshnikov, *Fluctuations of linear eigenvalue statistics of random band matrices*, Theory Probab. Appl. **60**, 407–443 (2016), MR3568789.

[107] K. Johansson, *From Gumbel to Tracy-Widom*, Probab. Theory Related Fields **138**, 75–112 (2007), MR2288065.

[108] K. Johansson, *Discrete orthogonal polynomial ensembles and the Plancherel measure*, Ann. of Math. (2) **153**, 259–296 (2001), MR1826414.

[109] K. Johansson, *Universality of the local spacing distribution in certain ensembles of Hermitian Wigner matrices*, Comm. Math. Phys. **215**, 683–705 (2001), MR1810949.

[110] K. Johansson and E. Nordenstam, *Eigenvalues of GUE minors*, Electron. J. Probab. **11**, no. 50, 1342–1371 (2006), MR2268547.

[111] T. Johnson and S. Pal, *Cycles and eigenvalues of sequentially growing random regular graphs*, Ann. Probab. **42**, 1396–1437 (2014), MR3262482.

[112] V. Kargin, *Limit theorems for linear eigenvalue statistics of overlapping matrices*, Electron. J. Probab. **20**, Paper No. 121, 30 (2015), MR3425541.

[113] J. Keating, "The Riemann zeta-function and quantum chaology", in *Quantum chaos (Varenna, 1991)*, Proc. Internat. School of Phys. Enrico Fermi, CXIX (North-Holland, Amsterdam, 1993), pp. 145–185, MR1246830.

[114] S. V. Kerov, *Asymptotics of the separation of roots of orthogonal polynomials*, Algebra i Analiz **5**, 68–86 (1993), MR1263315.

[115] S. V. Kerov, *Transition probabilities of continual Young diagrams and the Markov moment problem*, Funktsional. Anal. i Prilozhen. **27**, 32–49, 96 (1993), MR1251166.

[116] A. M. Khorunzhy, B. A. Khoruzhenko, and L. A. Pastur, *Asymptotic properties of large random matrices with independent entries*, J. Math. Phys. **37**, 5033–5060 (1996), MR1411619.

[117] A. Knowles and J. Yin, *Anisotropic local laws for random matrices*, Probab. Theory Related Fields **169**, 257–352 (2017), MR3704770.

[118] A. Knowles and J. Yin, *Eigenvector distribution of Wigner matrices*, Probab. Theory Related Fields **155**, 543–582 (2013), MR3034787.

[119] M. Krishnapur, B. Rider, and B. Virág, *Universality of the stochastic Airy operator*, Comm. Pure Appl. Math. **69**, 145–199 (2016), MR3433632.

[120] B. Landon, P. Sosoe, and H.-T. Yau, *Fixed energy universality for Dyson Brownian motion*, preprint (2016), arXiv:1609.09011.

[121] B. Landon and H.-T. Yau, *Convergence of local statistics of Dyson Brownian motion*, Comm. Math. Phys. **355**, 949–1000 (2017), MR3687212.

[122] B. Landon and H.-T. Yau, *Edge statistics of Dyson Brownian motion*, preprint (2017), arXiv:1712.03881.

[123] J. O. Lee and K. Schnelli, *Edge universality for deformed Wigner matrices*, Rev. Math. Phys. **27**, 1550018, 94 (2015), MR3405746.

[124] J. O. Lee and K. Schnelli, *Local law and Tracy-Widom limit for sparse random matrices*, Probab. Theory Related Fields **171**, 543–616 (2018), MR3800840.

[125] J. O. Lee and K. Schnelli, *Tracy-Widom distribution for the largest eigenvalue of real sample covariance matrices with general population*, Ann. Appl. Probab. **26**, 3786–3839 (2016), MR3582818.

[126] J. O. Lee, K. Schnelli, B. Stetler, and H.-T. Yau, *Bulk universality for deformed Wigner matrices*, Ann. Probab. **44**, 2349–2425 (2016), MR3502606.

[127] J. O. Lee and J. Yin, *A necessary and sufficient condition for edge universality of Wigner matrices*, Duke Math. J. **163**, 117–173 (2014), MR3161313.

[128] L. Li, M. Reed, and A. Soshnikov, *Central limit theorem for linear eigenvalue statistics for submatrices of Wigner random matrices*, preprint (2015), arXiv:1504.05933.

[129] D. R. Lick and A. T. White, *k-degenerate graphs*, Canad. J. Math. **22**, 1082–1096 (1970), MR0266812.

[130] B. F. Logan and L. A. Shepp, *A variational problem for random Young tableaux*, Advances in Math. **26**, 206–222 (1977), MR1417317.

[131] A. Lytova, *On non-Gaussian limiting laws for certain statistics of Wigner matrices*, Zh. Mat. Fiz. Anal. Geom. **9**, 536–581, 611, 615 (2013), MR3155024.

[132] A. Lytova and L. Pastur, *Central limit theorem for linear eigenvalue statistics of random matrices with independent entries*, Ann. Probab. **37**, 1778–1840 (2009), MR2561434.

[133] A. Lytova and L. Pastur, *Fluctuations of matrix elements of regular functions of Gaussian random matrices*, J. Stat. Phys. **134**, 147–159 (2009), MR2489497.

[134] P. McCullagh, *Tensor notation and cumulants of polynomials*, Biometrika **71**, 461–476 (1984), MR775392.

[135] M. L. Mehta, *Random matrices and the statistical theory of energy levels* (Academic Press, New York-London, 1967), pp. x+259, MR0220494.

[136]M. L. Mehta and M. Gaudin, *On the density of eigenvalues of a random matrix*, Nuclear Phys. **18**, 420–427 (1960), MR0112895.

[137]S. O'Rourke, D. Renfrew, and A. Soshnikov, *On fluctuations of matrix entries of regular functions of Wigner matrices with non-identically distributed entries*, J. Theoret. Probab. **26**, 750–780 (2013), MR3090549.

[138]S. O'Rourke and V. Vu, *Universality of local eigenvalue statistics in random matrices with external source*, Random Matrices Theory Appl. **3**, 1450005, 37 (2014), MR3208886.

[139]A. Okounkov, *Random matrices and random permutations*, Internat. Math. Res. Notices, 1043–1095 (2000), MR1802530.

[140]A. Okounkov and N. Reshetikhin, *Random skew plane partitions and the Pearcey process*, Comm. Math. Phys. **269**, 571–609 (2007), MR2276355.

[141]G. I. Olshanski, ed., *Kirillov's seminar on representation theory*, Vol. 181, American Mathematical Society Translations, Series 2, Advances in the Mathematical Sciences, 35 (American Mathematical Society, Providence, RI, 1998), pp. xiv+271, MR1618767.

[142]L. Pastur, *Limiting laws of linear eigenvalue statistics for Hermitian matrix models*, J. Math. Phys. **47**, 103303, 22 (2006), MR2268864.

[143]L. A. Pastur, *Spectra of random selfadjoint operators*, Uspehi Mat. Nauk **28**, 3–64 (1973), MR0406251.

[144]L. Pastur and M. Shcherbina, *Bulk universality and related properties of Hermitian matrix models*, J. Stat. Phys. **130**, 205–250 (2008), MR2375744.

[145]L. Pastur and M. Shcherbina, *On the edge universality of the local eigenvalue statistics of matrix models*, Mat. Fiz. Anal. Geom. **10**, 335–365 (2003), MR2012268.

[146]T. Pearcey, *The structure of an electromagnetic field in the neighbourhood of a cusp of a caustic*, Philos. Mag. (7) **37**, 311–317 (1946), MR0020857.

[147]A. Pizzo, D. Renfrew, and A. Soshnikov, *Fluctuations of matrix entries of regular functions of Wigner matrices*, J. Stat. Phys. **146**, 550–591 (2012), MR2880032.

[148]Å. Pleijel, *On a theorem by P. Malliavin*, Israel J. Math. **1**, 166–168 (1963), MR0167751.

[149]M. Prähofer and H. Spohn, *Universal distributions for growth processes in 1+1 dimensions and random matrices*, Physical review letters **84**, 4882–5 (2000), PMID10990822.

[150]K. Rajan and L. Abbott, *Eigenvalue spectra of random matrices for neural networks*, Phys. Rev. Lett. **97**, 188104 (2006), PMID17155583.

[151]R. Rashidi Far, T. Oraby, W. Bryc, and R. Speicher, *On slow-fading MIMO systems with nonseparable correlation*, IEEE Trans. Inform. Theory **54**, 544–553 (2008), MR2444540.

[152]J. Schenker, *Eigenvector localization for random band matrices with power law band width*, Comm. Math. Phys. **290**, 1065–1097 (2009), MR2525652.

[153]J. H. Schenker and H. Schulz-Baldes, *Semicircle law and freeness for random matrices with symmetries or correlations*, Math. Res. Lett. **12**, 531–542 (2005), MR2155229.

[154]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, 197, 199 (2011), MR2829615.

[155]M. Shcherbina, *Change of variables as a method to study general β-models: Bulk universality*, J. Math. Phys. **55**, 043504, 23 (2014), MR3390602.

[156]M. Shcherbina, *Edge universality for orthogonal ensembles of random matrices*, J. Stat. Phys. **136**, 35–50 (2009), MR2525225.

[157]M. Shcherbina, *Fluctuations of linear eigenvalue statistics of β matrix models in the multi-cut regime*, J. Stat. Phys. **151**, 1004–1034 (2013), MR3063494.

[158]T. Shcherbina, *On universality of local edge regime for the deformed Gaussian unitary ensemble*, J. Stat. Phys. **143**, 455–481 (2011), MR2799948.

[159]S. Sodin, *Fluctuations of interlacing sequences*, Zh. Mat. Fiz. Anal. Geom. **13**, 364–401 (2017), MR3733197.

[160]S. Sodin, *The spectral edge of some random band matrices*, Ann. of Math. (2) **172**, 2223–2251 (2010), MR2726110.

[161]A. Soshnikov, *Universality at the edge of the spectrum in Wigner random matrices*, Comm. Math. Phys. **207**, 697–733 (1999), MR1727234.

[162]P. Sosoe and P. Wong, *Regularity conditions in the CLT for linear eigenvalue statistics of Wigner matrices*, Adv. Math. **249**, 37–87 (2013), MR3116567.

[163]T. P. Speed, *Cumulants and partition lattices*, Austral. J. Statist. **25**, 378–388 (1983), MR725217.

[164]T. Spencer, "SUSY statistical mechanics and random band matrices", in *Quantum many body systems*, Vol. 2051, Lecture Notes in Math. (Springer, Heidelberg, 2012), pp. 125–177, MR2953867.

[165]E. M. Stein and G. Weiss, *Fractional integrals on n-dimensional Euclidean space*, J. Math. Mech. **7**, 503–514 (1958), MR0098285.

[166] T. Tao and V. Vu, *Random matrices: universality of ESDs and the circular law*, Ann. Probab. **38**, With an appendix by Manjunath Krishnapur, 2023–2065 (2010), MR2722794.

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

[168] T. Tao and V. Vu, *Random matrices: Universality of local eigenvalue statistics up to the edge*, Comm. Math. Phys. **298**, 549–572 (2010), MR2669449.

[169] T. Tao and V. Vu, *Random matrices: universality of local spectral statistics of non-Hermitian matrices*, Ann. Probab. **43**, 782–874 (2015), MR3306005.

[170] C. A. Tracy and H. Widom, *Level-spacing distributions and the Airy kernel*, Comm. Math. Phys. **159**, 151–174 (1994), MR1257246.

[171] C. A. Tracy and H. Widom, *On orthogonal and symplectic matrix ensembles*, Comm. Math. Phys. **177**, 727–754 (1996), MR1385083.

[172] C. A. Tracy and H. Widom, *The Pearcey process*, Comm. Math. Phys. **263**, 381–400 (2006), MR2207649.

[173] N. G. Ushakov, *Selected topics in characteristic functions*, Modern Probability and Statistics (VSP, Utrecht, 1999), pp. x+355, MR1745554.

[174] B. Valkó and B. Virág, *Continuum limits of random matrices and the Brownian carousel*, Invent. Math. **177**, 463–508 (2009), MR2534097.

[175] A. M. Veršik and S. V. Kerov, *Asymptotic behavior of the Plancherel measure of the symmetric group and the limit form of Young tableaux*, Dokl. Akad. Nauk SSSR **233**, 1024–1027 (1977), MR0480398.

[176] E. P. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, Ann. of Math. (2) **62**, 548–564 (1955), MR0077805.

[177] E. P. Wigner, *On the distribution of the roots of certain symmetric matrices*, Ann. of Math. (2) **67**, 325–327 (1958), MR0095527.

[178] J. Wishart, *The generalised product moment distribution in samples from a normal multivariate population*, Biometrika **20A**, 32–52 (1928).

[179] J. Yin, *The local circular law III: general case*, Probab. Theory Related Fields **160**, 679–732 (2014), MR3278919.