Lecture Notes on Zigzag Strategy

László Erdős*
lerdos@ist.ac.at

Abstract. We give a short pedagogical introduction to a new dynamical way to prove multi-resolvent local laws for a large class of random matrices. This zigzag strategy starts with an initial estimate at spectral parameters far away from the spectrum and then it consists of a tandem of two different stochastic flows (the zig and the zag steps) that gradually approach the spectrum. For simplicity of notation, we consider the single resolvent local law in the bulk for Wigner matrices for most of the presentation, but we also comment on the more general situations. We also mention selected applications.
This lecture notes contains the extended material of the author's four hour course at the Summer School on Interacting Particle Systems and Random Matrices at the Rényi Institute, Budapest, June 16–20, 2025. Suggestions on the material and the structure of the lectures, as well as corrections to these notes by Giorgio Cipolloni, Oleksii Kolupaiev and Joscha Henheik are gratefully acknowledged.

Date: June 24, 2025
Keywords and phrases: Characteristic flow, Ornstein-Uhlenbeck process, local law, Green function comparison theorem,
2020 Mathematics Subject Classification: 60B20, 15B52

1. Introduction to multi-resolvent local laws

1.1. **Ensembles.** We consider large $N \times N$ random self-adjoint matrices $H = H^*$, typically scaled such that $\|H\| \sim 1$ even as N increases[1]. We consider both symmetry classes, i.e. when H is real symmetric or complex Hermitian. We impose conditions on the distribution of the matrix elements directly, unlike for invariant ensembles that are characterized by a probability density $\sim e^{-TrV(H)}\mathrm{d}H$ for some real valued function V, where $\mathrm{d}H$ is the flat Lebesgue measure on the space of real symmetric or complex Hermitian matrices. The primary example is the Wigner ensemble (or Wigner matrices) whose matrix elements are centered i.i.d. (up to the symmetry constraint – see the precise Definition 2.1 later). The normalization

$$\mathbb{E}|h_{ij}|^2 = \frac{1}{N}$$

is chosen to keep the spectrum O(1) even as $N \to \infty$. We also consider various generalizations of the Wigner matrix, here is non-exhaustive list:

(i) **Deformed Wigner matrices**: $H = W + D$, where W is Wigner and $D = D^*$ is a deterministic matrix that is bounded (D can be thought of $\mathbb{E}H$ if we think of dropping the centredness condition in the definition of Wigner matrices).

(ii) **Wigner type matrices**: Keep centredness and independence, but drop identical distribution, i.e. assume

$$\mathbb{E}h_{ij} = 0, \qquad S_{ij} := \mathbb{E}|h_{ij}|^2$$

where the *matrix of variances* S satisfies the mean-field (or *flatness*) condition

$$\frac{c}{N} \le S_{ij} \le \frac{C}{N} \tag{1.1}$$

with two positive constants[2]. Once naturally combine Wigner type with deformation to yield **deformed Wigner type** ensemble.

(iii) **Correlated matrices**: We may also drop the independence condition and replace it with a nontrival covariance structure of h_{ij}, encoded in the *self-energy* operator $\mathcal{S} : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ defined by

$$\mathcal{S}\big[R\big] := \mathbb{E}HRH \tag{1.2}$$

for any deterministic matrix R. Besides assuming the flatness condition (analogue of (1.1)),

$$c\langle R\rangle \le \mathcal{S}\big[R\big] \le C\langle R\rangle, \qquad \text{for any deterministic } R \ge 0 \text{ matrix},$$

we assume further conditions on the decay of correlations, e.g.

$$\mathbb{E}h_{ij}h_{ab} \lesssim \frac{1}{\big[1 + \mathrm{dist}(ij, ab)\big]^p}$$

*Institute of Science and Technology Austria, Am Campus 1, 3400 Klosterneuburg, Austria. Supported by the ERC Advanced Grant "RMTBeyond" No. 101020331.

[1] All statements are understood for sufficiently large N.

[2] For certain results the upper bound is sufficient.

for some power $p > 2$ (to guarantee summability), where $\mathrm{dist}(ij, ab) := \min\{|i-a|+|j-b|, |i-b|+|j-a|\}$. Some further conditions on higher order covariances are also required. One can again add deformation.

(iv) **Hermitization of a non-Hermitian random matrix**. Let X be an $N \times N$ random matrix without symmetry constraint (e.g. all x_{ij} are i.i.d. random varianbles). The lack of self-adjointness makes studying X much harder. One often considers its **Hermitization**, which is a $(2N) \times (2N)$ matrix with an additional spectral parameter $z \in \mathbb{C}$:

$$H^z := \begin{pmatrix} 0 & X - z \\ X^* - \bar{z} & 0 \end{pmatrix}. \tag{1.3}$$

Technically it is a special Wigner type matrix (without the lower bound in (1.1)) with a 2×2 block structure and with a constant deformation in the two off-diagonal blocks. The significance of H^z comes from **Girko's formula** that allows one to compute linear statistics of the spectrum of X (a non-Hermitian problem) in terms of H^z (Hermitian problem):

$$\sum_{\sigma \in \mathrm{Spec} H} f(\sigma) = \frac{1}{4\pi} \int_{\mathbb{C}} \Delta f(z) \log |\det H^z| \mathrm{d}^2 z = -\frac{1}{4\pi} \int_{\mathbb{C}} \Delta f(z) \int_0^\infty \Im \mathrm{Tr} G^z(i\eta) \mathrm{d}\eta \, \mathrm{d}^2 z, \tag{1.4}$$

for any nice test function f, where $G^z(w) := (H^z - w)^{-1}$ is the resolvent of H^z.

For most of our presentation, we will work with Wigner matrices, but most results extend to these generalisations with nontrivial additional efforts.

Notations and conventions. We use the notation $[N]$ to represent the index set $\{1, \dots, N\}$. The letters a, b, j, and k are used to denote integer indices, while sometimes α (with various subscripts) denotes elements of $[N]^2$. All unrestricted summations of the form $\sum_a$ and $\sum_\alpha$ are understood to run over $a \in [N]$ and $\alpha \in [N]^2$, respectively.

We denote vectors in $\mathbb{C}^{N \times N}$ using boldface letters, e.g., $\boldsymbol{x}$. The scalar product on $\mathbb{C}^N$ is defined by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \sum_{j=1}^N \overline{x_j} y_j$, and the corresponding Euclidean norm is denoted by $\|\boldsymbol{x}\| := \langle \boldsymbol{x}, \boldsymbol{x} \rangle^{1/2}$.

Matrices are denoted by capital letters. Unless explicitly stated otherwise, all matrices we consider are $N \times N$. For a matrix $A \in \mathbb{C}^{N \times N}$, the angle brackets $\langle A \rangle := N^{-1} \mathrm{Tr}[A]$ denote its normalized trace. We use the following notations for the matrix norms:

$$\|A\| := \sup_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\|, \quad \|A\|_{\mathrm{hs}} := \big\langle |A|^2 \big\rangle^{1/2},$$

where $|A|^2 := AA^*$. Furthermore, for any $a \in [N]$ and vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, we use the following notation:

$$A_{\boldsymbol{xy}} := \langle \boldsymbol{x}, A\boldsymbol{y} \rangle, \quad A_{\boldsymbol{x}a} := \langle \boldsymbol{x}, A\boldsymbol{e}_a \rangle, \quad A_{a\boldsymbol{y}} := \langle \boldsymbol{e}_a, A\boldsymbol{y} \rangle,$$

where $\boldsymbol{e}_a$ is the standard a-th basis vector of $\mathbb{C}^N$. We denote the complex upper half-plane by $\mathbb{H}$, that is, $\mathbb{H} := \{z \in \mathbb{C} : \Im z > 0\}$, and its closure by $\overline{\mathbb{H}} := \mathbb{H} \cup \mathbb{R}$.

We use c and C to denote unspecified, positive constants—small and large, respectively—that are independent of N and may change from line to line. Various tolerance exponents are denoted by Greek letters such as ε, ξ, δ. The notation $\xi \ll \varepsilon$ means that there exists a small absolute constant $c > 0$ such that $\xi \le c\varepsilon$.

For two positive quantities $\mathcal{X}$ and $\mathcal{Y}$, we write $\mathcal{X} \lesssim \mathcal{Y}$ if there exists a constant $C > 0$ that depends only on the *model parameters*, such that $\mathcal{X} \le C\mathcal{Y}$. We use the notation $\mathcal{X} \sim \mathcal{Y}$ if both $\mathcal{X} \lesssim \mathcal{Y}$ and $\mathcal{Y} \lesssim \mathcal{X}$ hold. For an arbitrary quantity $\mathcal{X}$ and a positive quantity $\mathcal{Y}$, we use the notation $\mathcal{X} = \mathcal{O}(\mathcal{Y})$ to indicate that $|\mathcal{X}| \lesssim \mathcal{Y}$.

Let $\Omega := \{\Omega^{(N)}(u) \,|\, N \in \mathbb{N}, u \in \mathcal{U}^{(N)}\}$ be a family of events depending on N and possibly on a parameter u that varies over some parameter set $\mathcal{U}^{(N)}$. We say that Ω holds *with very high probability* (w.v.h.p.) uniformly in $u \in \mathcal{U}^{(N)}$ if, for any $D > 0$,

$$\sup_{u \in \mathcal{U}^{(N)}} \mathbb{P}\big[\Omega^{(N)}(u)\big] \ge 1 - N^{-D},$$

for any $N \ge N_0(D)$. We often discard the explicit dependence of $\Omega^{(N)}$ and $\mathcal{U}^{(N)}$ on N, and simply refer to Ω as a very-high-probability event. A bound is said to hold w.v.h.p. if it holds on a very-high-probability event.

1.2. **Single resolvent and its approximations.** Let H be a self-adjoint matrix and $z \in \mathbb{C} \setminus \mathbb{R}$ be a spectral parameter. We define its resolvent

$$G(z) := (H - z)^{-1}.$$

Note that we have the trivial norm bound

$$\|G(z)\| \le \frac{1}{\eta}, \quad \eta := |\Im z|,$$

that blows up as η, the Im-part of the spectral parameter approaches to zero.

The really interesting regime is $z = E + i\eta$, when E is in the limiting spectrum of H and $\eta \ll 1$, since then the resolvent $G(z)$ carries local information about the spectrum of H near E in a window of order η. For example,

$$\langle \Im G(z)\rangle = \frac{1}{N}\mathrm{Tr}\Im G(z) = \frac{1}{N}\sum_{\alpha=1}^{N}\frac{\eta}{(\lambda_\alpha - E)^2+\eta^2}, \tag{1.5}$$

where $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_N$ are the eigenvalues of H. Note that this sum is concentrated on eigenvalues $|\lambda_\alpha - E| \lesssim \eta$ and all others are suppressed. Ideally, one would like to understand $\langle \Im G(z)\rangle$ with η comparable (or even smaller) than the typical level spacing (distance between neighboring eigenvalues), which, in our standard normalization, is $\eta \sim 1/N$ in the bulk spectrum. So in all discussions below we implicitly assume that $\eta \lesssim 1$ and in most cases we think of $\eta \ll 1$.

A remarkable property of the resolvent of many random matrices is that it satisfies a type of *law of large numbers (LLN)*, i.e. it tends to concentrate around a deterministic matrix $M = M(z)$ that is typically bounded uniformly in $z \in \mathbb{C}\setminus\mathbb{R}$ (i.e. it remains bounded even as $\eta \to 0$). This holds in weak sense, both in *averaged* and *isotropic* form, more precisely we expect

$$\langle (G-M)B\rangle \lesssim \frac{1}{N\eta}, \qquad \langle \boldsymbol{x}, (G-M)\boldsymbol{y}\rangle \lesssim \frac{1}{\sqrt{N\eta}}, \qquad \text{w.v.h.p.} \tag{1.6}$$

for any deterministic bounded matrix B and deterministic unit vectors $\boldsymbol{x}, \boldsymbol{y}$. Estimates of this type are called **local laws**[3]. The word *deterministic* is essential here; it is trivial to see that if $\boldsymbol{x} = \boldsymbol{y}$ were the (random) eigenvectors of H with eigenvalue near E, then $\langle \boldsymbol{x}, G\boldsymbol{x}\rangle \sim 1/\eta$ would blow up as $\eta \ll 1$. The bounds (1.6), together with the information that M is bounded, show a concentration of G around M in the regime where $\eta \gg 1/N$, i.e. when η is just above the eigenvalue spacing. Indeed, this is an optimal range for LLN to hold; it is easy to see from (1.5) that $\langle G\rangle$ becomes a genuinely fluctuating object once $\eta \sim 1/N$ or smaller.

The deterministic approximation M is model dependent. In the simplest Wigner case, $M(z) = m(z)\cdot I$ where $m(z)$ is the Stieltjes transform of the semicircle law, see (2.2) later. For more complicated ensembles, M is the unique solution of the *Matrix Dyson equation (MDE)*, see (2.8) later; for example if $H = W + D$ is a deformed Wigner matrix W with deformation $D = D^*$, then

$$-\frac{1}{M} = z - D + \langle M\rangle, \qquad \big(\Im M(z)\big)(\Im z) > 0.$$

The detailed stochastic analysis of G and $G - M$ relies on a good understanding of M itself, and the properties of the solution to the MDE have been thoroughly investigated. In these notes we will not focus on these deterministic analyses, especially since we are mostly concerned with the Wigner case, where $m(z)$ has a simple explicit formula.

As (1.5) indicates, single resolvent local laws are suitable to get the eigenvalue density on any *mesoscopic* scales, i.e. on scales $\eta \gg 1/N$ that contains more than $O(1)$ eigenvalues. For example, as a corollary of the average law (1.6) with $B = I$ is the *eigenvalue rigidity*, stating that

$$|\lambda_\alpha - \gamma_\alpha| \lesssim \frac{N^\xi}{N}, \qquad \text{w.v.h.p.} \tag{1.7}$$

for any $\xi > 0$, for any *bulk*[4] index $\alpha \in [\epsilon N, (1-\epsilon)N]$, where γ_α is the α-th quantile of the semicircle law, defined by

$$\int_{-2}^{\gamma_\alpha}\frac{1}{2\pi}\sqrt{4-x^2}\,\mathrm{d}x = \frac{\alpha}{N}.$$

Another important corollary of the isotropic local law is the *eigenfunction delocalization*, i.e.

$$|\langle \boldsymbol{x}, \boldsymbol{u}_\alpha\rangle|^2 \lesssim \frac{N^\xi}{N}, \qquad \text{w.v.h.p.} \tag{1.8}$$

for any normalized eigenvector $\boldsymbol{u}_\alpha$, i.e., $H\boldsymbol{u}_\alpha = \lambda_\alpha \boldsymbol{u}_\alpha$, and any normalized deterministic vector $\boldsymbol{x}$. In particular, each coordinate is bounded by $|\boldsymbol{u}_\alpha(i)| \lesssim N^{-1/2+\xi}$, showing that no eigenfunction can be supported only on a small fraction of the available N sites. The bound (1.8) follows from the spectral theorem and the boundedness of $(\Im G(\gamma_\alpha + i\eta))_{\boldsymbol{x}\boldsymbol{x}}$:

$$|\langle \boldsymbol{x}, \boldsymbol{u}_\alpha\rangle|^2 \lesssim \sum_{\alpha=1}^{N}|\langle \boldsymbol{x}, \boldsymbol{u}_\alpha\rangle|^2\frac{\eta^2}{(\lambda_\alpha-\gamma_\alpha)^2+\eta^2} = \eta(\Im G(\gamma_\alpha+i\eta))_{\boldsymbol{x}\boldsymbol{x}} \lesssim \eta,$$

for $\eta \sim N^{-1+\xi}$, using the rigidity (1.7) in the first inequality, and isotropic local law in the second.

[3] The word *local* refers to the typical situation when $\eta \ll 1$, i.e. $\Im G$ is concentrated on a local part of the spectrum. Statements of the form (1.6) for $\eta \sim 1$ are often called **global laws**.

[4] Similar results hold for all indices taking into account that the eigenvalue spacing increases near the edges.

1.3. **Longer resolvent chains and their local laws.** Not every relevant question can be answered by a single resolvent local law, sometimes we need information on products of resolvents, possibly alternating with observables of the form

$$G_{[1,k]} := G_1 A_1 G_2 A_2 \dots A_{k-1} G_k, \tag{1.9}$$

where A_i are deterministic matrices (observables) and G_i are resolvents possible at different spectral parameters $G_i = G(z_i)$, but sometimes even for different matrices $G_i = (H_i - z_i)^{-1}$, where the different matrices H_i are closely related (e.g. they are different deformations of the same random matrix).

Information about eigenvalues and individual eigenvectors are typically accessible via single resolvents, but correlations between different eigenvalues or eigenvectors often involve more than one resolvent. For example, general quadratic forms $\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle$ with different eigenvectors, $\boldsymbol{u}_i, \boldsymbol{u}_j$, $i \neq j$, require at least two resolvents, see (1.23)–(1.24) later. In Section 1.5 we give more applications.

The key fact about the chains (1.9) is that their deterministic approximation $M_{[1,k]} = M(\boldsymbol{z}, \boldsymbol{A})$ (depending on all spectral parameters and observables) is not a simple product of the approximations of single resolvents. E.g. even if $G_i \approx M_i$ by single resolvent local law (1.6), it is not true in general that, e.g. $G_1 A G_2 \approx M_1 A M_2$; the truth is more complicated. For example, if $G_i = (H_i - z_i)^{-1}$ and $H_i = W + D_i$ is deformed Wigner matrix W, then

$$G_1 A G_2 \approx M_{[1,2]} = M_1 A M_2 + \frac{\langle M_1 A M_2\rangle}{1 - \langle M_1 M_2\rangle} M_1 M_2. \tag{1.10}$$

If $D_i = 0$, i.e. we consider the simplest case $G_i = (H - z_i)^{-1}$, and H is a Wigner matrix, then we have

$$G_1 A G_2 \approx M_{[1,2]} = m_1 m_2 A + m_1 m_2 \frac{m_1 m_2}{1 - m_1 m_2}\langle A\rangle, \qquad m_i = m(z_i). \tag{1.11}$$

In general $M_{[1,k]}$ can be obtained by a recursive formula on the length k, in some special cases a closed form exists, see, e.g. [15, Theorem 3.4] for Wigner matrices. In general, they satisfy the bound

$$\|M_{[1,k]}\| \lesssim \frac{1}{\eta^{k-1}} \prod_{i=1}^{k-1} \|A_i\|, \tag{1.12}$$

where $\eta := \min \eta_i = \min |\Im z_i|$. The multi-resolvent averaged and isotropic analogues of the single resolvent local laws (1.6) are[5]:

$$\big\langle (G_{[1,k]} - M_{[1,k]}) A_k \big\rangle \leq \frac{1}{N\eta} \frac{1}{\eta^{k-1}} \prod_{i=1}^{k} \|A_i\|, \qquad \text{w.v.h.p.} \tag{1.13}$$

$$(G_{[1,k]} - M_{[1,k]})_{\boldsymbol{xy}} \leq \frac{1}{\sqrt{N\eta}} \frac{1}{\eta^{k-1}} \|\boldsymbol{x}\| \|\boldsymbol{y}\| \prod_{i=1}^{k-1} \|A_i\|, \qquad \text{w.v.h.p.} \tag{1.14}$$

where $\eta = \min \eta_i = \min |\Im z_i|$. The averaged law is by a factor $1/N\eta$ better than the corresponding M-bound from (1.12), by the isotropic law this factor is $1/\sqrt{N\eta}$.

1.4. **Improvements of the multi-resolvent local laws.** The multiresolvent local law stated in (1.13)–(1.14) is its crudest form. It is optimal in its full generality, but it has several refinements which under additional conditions give stronger results. Here we list a few of them.

1.4.1. *Regular observables.* It turns out that if some of the observables A_i are traceless, $\langle A_i\rangle = 0$, then this fact reduces both the size of $M_{[1,k]}$ and the corresponding estimate in the local law. The general $\sqrt{\eta}$ *rule* summarizes this as follows [14]: Suppose that n out of the k matrices $A_1, A_2, \dots A_k$ are traceless, then (1.13) is modified to

$$\big\langle (G_{[1,k]} - M_{[1,k]}) A_k \big\rangle \leq \frac{1}{N\eta} \frac{1}{\eta^{k-1-n/2}} \prod_{i=1}^{k} \|A_i\| \qquad \text{w.v.h.p.} \tag{1.15}$$

and if n out of the $k-1$ matrices $A_1, A_2, \dots A_{k-1}$ are traceless, then (1.14) is modified to

$$(G_{[1,k]} - M_{[1,k]})_{\boldsymbol{xy}} \leq \frac{1}{\sqrt{N\eta}} \frac{1}{\eta^{k-1-n/2}} \|\boldsymbol{x}\| \|\boldsymbol{y}\| \prod_{i=1}^{k-1} \|A_i\| \qquad \text{w.v.h.p.} \tag{1.16}$$

[5] We often ignore absolute values in the LHS when we estimate sizes.

i.e. both estimates improve by a factor $\sqrt{\eta}$ for each traceless observables. The improvement in the bound on M is similar, but it comes with integer powers of M:

$$\|M_{[1,k]}\| \lesssim \frac{1}{\eta^{k-1-\lceil n/2\rceil}} \prod_{i=1}^{k-1} \|A_i\|, \tag{1.17}$$

where $\lceil x\rceil$ is the upper integer part of x. For example, from the explicit formula (1.11) it is clear that $\|M_{[1,2]}\| \lesssim 1$ if $\langle A\rangle = 0$, but it can be of order $1/\eta$ otherwise. This is because traceless matrices are orthogonal to a one-dimensional subspace in the Hilbert space of matrices, namely to the constant matrices. For the purpose of G_1AG_2 only constant matrices A behave badly, because they are in the eigenspace with a very bad (small) eigenvalue of the *two-body stability operator* to be discussed later in general. For Wigner matrices this operator is just multiplication by $1-m_1m_2$ which can be of order η if $z_1 = \bar{z}_2$. The fact that the approximation $M_{[1,2]}$ to G_1G_2 can be of order $1/\eta$ (for $A = I$) is due inverting the stability operator on its own subspace. Observables orthogonal to this bad direction (also called *regular observables*) will not face this magnification.

For more complicated ensembles, the regular observables A are not directly characterized by the traceless property, but by its analogue, namely that A are orthogonal to the bad direction of the stability operator. Mean field condition usually guarantees that there is only one bad direction. For example, for deformed Wigner matrices, (1.10) shows that the condition $\langle M_1AM_2\rangle = 0$ for regularity replaces $\langle A\rangle = 0$. This concept in general is energy (z_1, z_2)-dependent. The $\sqrt{\eta}$ rule still holds if n counts the number of regular observables in this generalized sense.

1.4.2. *Behavior at small density: edge es cusp.* So far we focused on the bulk regime where the density of states is of O(1). Near the edges (and the cusps) the density $\rho(x)$ is reduced. This leads to two opposite effects: on the one hand the size of certain M terms get reduced, on the other hand the stability of the equation deteriorates. In some cases these two effects balance each other, for example (1.6) hold in the same form (and optimal) also in the edge regime. In some other cases the ρ improvement is explicit, for example for Wigner matrices and traceless A we have

$$\langle (G-m)A\rangle \lesssim \frac{\sqrt{\rho}}{N\sqrt{\eta}}, \qquad \langle A\rangle = 0, \tag{1.18}$$

i.e. on top of the $\sqrt{\eta}$ gain, we also gain a $\sqrt{\rho}$ factor.

There is also a ρ improvement if we replace some G with $\Im G$ in a chain with traceless observables, see [21] for detailed results.

1.4.3. *Hilbert-Schmidt and Schatten norm of the observables.* So far we assumed that the observables A_i are bounded in operator norm. Actually the natural norm is the (sometimes much smaller) normalized HS norm: $\|A\|_{hs} := \langle A^2\rangle^{1/2}$. Indeed, this can be done and in the Wigner case we have [21] that if all A_i's are traceless, then (1.15)–(1.16) hold with all $\|A_i\|$ replaced with the typically much smaller $\|A_i\|_{hs}$-norm. For example, the precise norm dependence in (1.18) is

$$\langle (G-m)A\rangle \lesssim \frac{\sqrt{\rho}}{N\sqrt{\eta}}\|A\|_{hs}, \qquad \langle A\rangle = 0. \tag{1.19}$$

In fact, one can make an even more refined analysis when some HS norm is replaced with other Schatten norms. For example, we have [22]

$$\left|\langle GAGA\rangle - m^2\langle A^2\rangle\right| \lesssim \frac{\langle |A|^2\rangle}{N\eta} + \frac{\langle |A|^4\rangle^{1/2}}{N\sqrt{\eta}}. \tag{1.20}$$

1.5. **Applications of multi-resolvent local laws.** Here are several examples, where longer chains (1.9) with $k \geq 2$ are necessary. Most important is the $k = 2$ case, but as we will see in their proofs, often one has to consider an entire hierarchy of chains of different length even if eventually one is interested only in the physically most relevant $k = 1, 2$ chains.

1.5.1. *Eigenstate Thermalisation Hypothesis (ETH).* (Also known in mathematics literature as *Quantum Unique Ergodicity (QUE)*) asserts that for normalized eigenvectors of Wigner matrices [13]

$$\left|\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle - \delta_{ij}\langle A\rangle\right| \leq \frac{N^\xi}{\sqrt{N}} \qquad \text{w.v.h.p.} \tag{1.21}$$

for any deterministic bounded observable A. Here $\langle \boldsymbol{u}_i, A\boldsymbol{u}_j\rangle$ is called the *eigenvector overlap* (on the observable A) and such quadratic forms have clear quantum mechanical interpretation (cf. with standard quantum ergodicity theorems, like Shnirelman's theorem [48], stating that high energy eigenfunctions of the Laplacian on a surface with negative curvature are uniformly distributed on the phase space).

Set $\mathring{A} := A - \langle A\rangle I$ to be the traceless part of A. The ETH is equivalent to

$$\big|\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_j\rangle\big|^2 \le \frac{N^\xi}{N} \qquad \text{w.v.h.p.} \tag{1.22}$$

Even for $i = j$ (diagonal overlap), we cannot conclude from the spectral theorem

$$\langle \Im G(z)\mathring{A}\rangle = \frac{1}{N}\sum_{i=1}^N \langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle \frac{\eta}{(\lambda_i-\gamma_i)^2+\eta^2}, \qquad z=\gamma_i+i\eta, \quad \eta\sim N^{-1+\xi}, \tag{1.23}$$

since $\langle \boldsymbol{u}_i, \mathring{A}\boldsymbol{u}_i\rangle$ has no sign. The correct quantity to look at is

$$\langle \Im G(z)\mathring{A}\Im G(z')\mathring{A}\rangle = \frac{1}{N}\sum_{i',j'=1}^N \big|\langle \boldsymbol{u}_{i'}, \mathring{A}\boldsymbol{u}_{j'}\rangle\big|^2 \frac{\eta}{(\lambda_{i'}-\gamma_i)^2+\eta^2}\frac{\eta}{(\lambda_{j'}-\gamma_j)^2+\eta^2}, \quad z=\gamma_i+i\eta, \quad z'=\gamma_j+i\eta, \quad \eta\sim N^{-1+\xi}. \tag{1.24}$$

Indeed, if we manage to prove that $\langle \Im G(z)\mathring{A}\Im G(z')\mathring{A}\rangle \lesssim 1$ with very high probability, uniformly in z, z' with $\Im z, \Im z' \sim N^{-1+\xi}$, then (1.22) follows from (1.24).

1.5.2. *Thermalisation.* Imagine that H is the Hamiltonian of a quantum system and $A, B, \dots$ are deterministic observables. Let

$$A(t) = e^{-itH}Ae^{itH}$$

be the Heisenberg (quantum) time evolution of A. Typical thermalisation questions asks if and how much $A(t)$ and B become orthogonal (independent) at large time. The answer is [15]

$$\langle A(t)B\rangle = \langle A\rangle\langle B\rangle + \theta(t)^2\frac{\langle \mathring{A}\mathring{B}\rangle}{t^3} + O\Big(\frac{t^2}{N}\Big) \qquad \text{w.v.h.p.} \tag{1.25}$$

for any bounded A, B, where $\theta(t) := J_1(2t)\sqrt{t}$ is an O(1) oscillatory function (here J_1 is the first Bessel function of the first kind). Similar results can be derived for more than two observables, for example for three observables and two different times t, s with $t \ge s \gg 1$, $t - s \gg 1$ we have

$$\begin{aligned}\langle A(t)B(s)C\rangle = \langle A\rangle\langle B\rangle\langle C\rangle + \theta(s)^2\frac{\langle A\rangle\langle \mathring{B}\mathring{C}\rangle}{s^3} + \theta(t)^2\frac{\langle B\rangle\langle \mathring{A}\mathring{C}\rangle}{t^3} + \theta(t-s)^2\frac{\langle C\rangle\langle \mathring{A}\mathring{B}\rangle}{(t-s)^3} \\ + \theta(s)\theta(t)\theta(t-s)\frac{\langle \mathring{A}\mathring{B}\mathring{C}\rangle}{s^{3/2}t^{3/2}(t-s)^{3/2}} + O\Big(\frac{t^3}{N}\Big) \qquad \text{w.v.h.p.}\end{aligned} \tag{1.26}$$

A related object is the *out-of-time-ordered correlator (OTOC)*, extensively studied in the quantum chaos physics literature, it is defined by

$$\mathcal{C}_{A,B}(t) := \frac{1}{2}\langle \big|[A(t), B]\big|^2\rangle$$

for any two observables A, B. Similarly to $\langle A(t)B\rangle$, it also expresses how much mixing happens in the system [22].

All these questions go back to multi-resolvent local laws by expressing the unitary time evolution via contour integral of the resolvent:

$$e^{itH} = \frac{1}{2\pi i}\oint_\gamma \frac{e^{itz}}{H-z}\mathrm{d}z = \frac{1}{2\pi i}\oint_\gamma e^{itz}G(z)\mathrm{d}z,$$

where γ encircles the spectrum of H. For example

$$\langle A(t)B\rangle = -\frac{1}{4\pi^2}\oint_\gamma\oint_\gamma e^{itz}e^{-itz'}\langle G(z)AG(z')B\rangle \mathrm{d}z\mathrm{d}z'.$$

If we find a deterministic approximation $M = M(A, B, z, z')$ to $G(z)AG(z')B$, then we can compute the leading term by explicit contour integration.

1.5.3. *Eigenvector overlaps (Hermitian case).* Consider a Wigner matrix W with two different deformations

$$H_1 = W + D_1, \qquad H_2 = W + D_2,$$

where D_1, D_2 are deterministic (hermitian) matrices with zero trace $\langle D_i\rangle = 0$, for simplicity. Let $\lambda_i^{(\ell)}$ and $\boldsymbol{u}_i^{(\ell)}$, $\ell = 1, 2$, be the corresponding system of eigenvalues/eigenvectors i.e.

$$H_\ell \boldsymbol{u}_i^{(\ell)} = \lambda_i^{(\ell)}\boldsymbol{u}_i^{(\ell)}, \qquad \ell = 1, 2.$$

If $D_1 = D_2$, then the eigenfunction overlap is trivial, $\langle \boldsymbol{u}_i^{(1)}, \boldsymbol{u}_j^{(2)} \rangle = \delta_{ij}$. For $D_1 \neq D_2$ we have the bound [16]

$$\big|\langle \boldsymbol{u}_i^{(1)}, \boldsymbol{u}_j^{(2)} \rangle\big|^2 \lesssim \frac{N^\xi}{N} \frac{1}{\langle (D_1 - D_2)^2 \rangle + |\lambda_i^{(1)} - \lambda_j^{(2)}|^2 + \dots} \qquad \text{w.v.h.p.} \tag{1.27}$$

(where the additional positive terms are ignored for this presentation, these contain some linear combination of D's and λ's). This shows that eigenvectors can become decorrelated in two ways: either their energies are at distance or their deformations are far away in Hilbert-Schmidt norm sense.

Similarly to (1.24), a good upper bound on the overlap $\big|\langle \boldsymbol{u}_i^{(1)}, \boldsymbol{u}_j^{(2)} \rangle\big|^2$ is accessible via a two-resolvent local law of the form

$$\big|\langle \boldsymbol{u}_i^{(1)}, \boldsymbol{u}_j^{(2)} \rangle\big|^2 \leq \eta \langle \Im G^{(1)}(\gamma_i^{(1)} + i\eta) \Im G^{(2)}(\gamma_j^{(2)} + i\eta) \rangle, \qquad \eta \sim N^{-1+\xi},$$

where $G^{(\ell)}$ is the resolvent of $H^{(\ell)}$.

The result (1.27) is essentially used in our papers on the *decorrelation transition* [5] and the *Law of Fractional Logarithm* [6] in the Wigner minor process. Here we just list the results for completeness:

Decorrelation transition. Let $W = W^{(N)}$ be an $N \times N$ Wigner matrix and let $W^{(N-1)}, W^{(N-2)}, \dots$ be its upper left corner minors. The eigenvalues of $W^{(N-k)}$ are denoted by

$$\lambda_1^{(N-k)} \geq \lambda_2^{(N-k)} \geq \dots \geq \lambda_{N-k}^{(N-k)}$$

listed in decreasing order. By interlacing of the eigenvalues of any Hermitian matrix and its those of its minor (of dimension one less) we have

$$\lambda_1^{(N)} \geq \lambda_1^{(N-1)} \geq \lambda_2^{(N)} \geq \lambda_2^{(N-1)} \geq \lambda_3^{(N)} \geq \dots$$

It is well known that $\lambda_1^{(N-1)}$ and $\lambda_1^{(N)}$ are strongly correlated, in fact $\lambda_1^{(N)} - \lambda_1^{(N-1)} \sim 1/N$ despite that interlacing in principle provides an interval $[\lambda_2^{(N)}, \lambda_1^{(N)}]$ of length $\sim N^{-2/3}$ for $\lambda_1^{(N-1)}$. So $\lambda_1^{(N-1)}$ sticks to $\lambda_1^{(N)}$, then $\lambda_1^{(N-2)}$ sticks to $\lambda_1^{(N-1)}$ etc. so eventually $\lambda_1^{(N-k)}$ sticks to $\lambda_1^{(N)}$ if k is not too big. At which k, i.e. at which level of subminors does this strong correlation start weaken?

Motivated by a similar study in the GUE case by Forrester and Nagao [35], we recently proved [5] the following phase transition picture:

(i) [Subcritical regime] If $k \ll N^{2/3}$, then $\lambda_1^{(N-k)}$ is still close to $\lambda_1^{(N)}$ at a distance $\sqrt{k}/N \ll N^{-2/3}$ and their difference with a natural shift is asymptotically normal:

$$\lambda_1^{(N)} - \lambda_1^{(N-k)} - k/N \approx \mathcal{N}(0, \sqrt{k}/N).$$

The corresponding eigenvectors are almost parallel.

(ii) [Critical regime] If $k \sim N^{2/3}$, then $\lambda_1^{(N)}$ and $\lambda_1^{(N-k)}$ have a universal, nontrivial joint correlation function that was explicitly identified by Forrester and Nagao in the GUE case.

(iii) [Supercritical regime] If $k \gg N^{2/3}$, then $\lambda_1^{(N)}$ and $\lambda_1^{(N-k)}$ are asymptotically independent (i.e. their appropriately rescaled versions follow two independent Tracy-Widom distribution). The corresponding eigenvectors are essentially orthogonal.

Law of Fractional Logarithm. Let $(x_{ij})_{i,j \in \mathbf{N}}$ be a double infinite array of i.i.d. random variables with the Hermitian symmetry, $x_{ij} = \bar{x}_{ji}$, with $\mathbb{E}x_{ij} = 0$, $\mathbb{E}|x_{ij}|^2 = 1$ (and $\mathbb{E}x_{ij}^2 = 0$ in the complex case). Let $X^{(N)} = (x_{ij})_{i,j=1}^N$ be its $N \times N$ upper left minor and define

$$W^{(N)} := \frac{1}{\sqrt{N}} X^{(N)},$$

which is a Wigner matrix with standard normalisation. Note $W^{(N)}$'s for different N's are strongly correlated, they are essentially minors of each other (up to natural normalisation), so this sequence of random matrices is called the *Wigner minor process*. Let $\lambda_1^{(N)}$ be the largest eigenvalue of $W^{(N)}$ and let

$$\widetilde{\lambda}_1^{(N)} := N^{2/3}\big(\lambda_1^{(N)} - 2\big)$$

be its appropriate rescaling. Inspired by the question and a similar result for the Gaussian (GUE) case by Paquette and Zeitouni [44] and some further more precise results [8] still for GUE, we proved [6] the following general form of what Paquette and Zeitouni called[6] the *Law of Fractional Logarithm*: for any Wigner minor process (without Gaussian

[6] Compare it with the standard *Law of Iterated Logarithm* for sums of independent random variables – this classical result inspired the current terminology.

assumption) almost surely we have

$$\liminf_{N\to\infty} \frac{\widetilde{\lambda}_1^{(N)}}{(\log N)^{1/3}} = -\Big(\frac{8}{\beta}\Big)^{1/3}, \quad \text{and} \quad \limsup_{N\to\infty} \frac{\widetilde{\lambda}_1^{(N)}}{(\log N)^{2/3}} = \Big(\frac{1}{2\beta}\Big)^{2/3},$$

where, as usual, $\beta = 1$ is the real symmetric case and $\beta = 2$ is the complex hermitian case. Note the strong asymmetry of the lower and upper tail results, the $\log N$-scaling is different. This is caused by the strong asymmetry of the two tails of the Tracy-Widom distribution.

1.5.4. *Eigenvector overlaps (non-Hermitian case).* Similar question on eigenvector overlaps arises about the Hermitization (1.3) of an iid. matrix X at two different spectral parameters z_1, z_2. Let $\boldsymbol{w}_j^z \in \mathbb{C}^{2N}$ be the eigenvectors of H^z, in particular this means that if we write $\boldsymbol{w}_j^z = (\boldsymbol{u}_j^z, \boldsymbol{v}_j^z)$, then $\boldsymbol{u}_j^z, \boldsymbol{v}_j^z \in \mathbb{C}^N$ are the left and right singular vectors of $X - z$. Then we have [19]

$$|\langle \boldsymbol{w}_i^{z_1}, \boldsymbol{w}_j^{z_2}\rangle|^2 \le \frac{1}{N}\frac{1}{|z_1 - z_2|^2 + N^{-1}}, \quad i,j \le N^\epsilon \qquad \text{w.v.h.p.} \tag{1.28}$$

which we could translate into overlaps of the non-Hermitian eigenfunctions. Indeed, if $\{\sigma_i\}_{i=1}^N$ are the eigenvalues of X and $\{\boldsymbol{r}_i\}, \{\boldsymbol{l}_i\}$ are the corresponding right and left eigenvectors, $X\boldsymbol{r}_i = \sigma_i \boldsymbol{r}_i$, $\boldsymbol{l}_i^t X = \sigma_i \boldsymbol{l}_i^t$ with the usual biorthogonal normalization, $\langle \bar{\boldsymbol{l}}_i^t, \boldsymbol{r}_j\rangle = \delta_{ij}$, then we have [19]

$$\frac{\mathcal{O}_{ij}}{\sqrt{\mathcal{O}_{ii}\mathcal{O}_{jj}}} \le \frac{|\langle \boldsymbol{r}_i, \boldsymbol{r}_j\rangle|^2}{\|\boldsymbol{r}_i\|^2\|\boldsymbol{r}_j\|^2} + \frac{|\langle \boldsymbol{l}_i, \boldsymbol{l}_j\rangle|^2}{\|\boldsymbol{l}_i\|^2\|\boldsymbol{l}_j\|^2} \le \frac{1}{N}\frac{1}{|\sigma_i - \sigma_j|^2 + N^{-1}}, \qquad \text{w.v.h.p.}$$

where

$$\mathcal{O}_{ij} := \langle \boldsymbol{l}_i, \boldsymbol{l}_j\rangle\langle \boldsymbol{r}_i, \boldsymbol{r}_j\rangle$$

is the standard *non-Hermitian eigenvector overlap*.

A weaker form of the overlap of singular vectors (1.28) was essentially used in the DBM analysis when we proved the macroscopic and mesoscopic CLT for linear statistics of the eigenvalues of X [17, 24], as well as in the universal Gumbel distribution for the rightmost eigenvalue and for the spectral radius of X in [18, 26].

Gumbel distribution of the rightmost eigenvalue. As a representative example, we explain the result for the rightmost eigenvalue. Let X be an $N \times N$ complex i.i.d. random matrix, i.e. its matrix elements are i.i.d. with normalisation $\mathbb{E}x_{ij} = 0$, $\mathbb{E}|x_{ij}|^2 = \frac{1}{N}$ but no symmetry constraint. Let σ_j, $j = 1, 2, \dots, N$ be the eigenvalues of X. They obey the (global) *circular law*, i.e.

$$\frac{1}{N}\sum_j f(\sigma_j) = \frac{1}{\pi}\int f(z)\mathrm{d}^2 z + O(N^\xi/N)$$

with very high probability, for a smooth N-independent test function f (there are also local or mesoscopic versions). It is also known that

$$\max_j |\sigma_j| \le 1 + \frac{N^\xi}{\sqrt{N}}, \qquad \text{w.v.h.p.}$$

The problem is to identify more precisely the behavior of $\max_j \Re\sigma_j$. One motivation for this question comes studying the standard linear system of ODE's with random coefficients

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{v}(t) = -(I + gX)\boldsymbol{v}(t), \qquad \boldsymbol{v}(0) = \boldsymbol{v}_0,$$

where g is a tunable coupling constant. This was the starting point in R. May's famous article [43] about stability of complex systems; this simple ODE is used in many phenomenological models in population dynamics and neuroscience. The question is tune g appropriately so that the solution $\boldsymbol{v}(t)$ remains roughly O(1) as time $t \to \infty$. Recall that the solution of a linear ODE tends to behave exponentially for large time with a rate proportional to the largest real part among all eigenvalues, so stability requires that $\max \Re\mathrm{Spec}(I + gX) = 0$.

Our result is the following distributional limit [18]:

$$\sqrt{4\gamma_N N}\Big[\max_j \Re\sigma_j - 1 - \sqrt{\frac{\gamma_N}{4N}}\Big] \Longrightarrow G,$$

where G is standard Gumbel random variable, i.e. its distribution function is $\mathbb{P}(G \le x) = \exp(-e^{-x})$ and

$$\gamma_N := \frac{1}{2}\Big[\log N - 5\log\log N - \log(2\pi^4)\Big].$$

There is a similar result (with slightly different γ_N) for $\max|\sigma_j|$, i.e. the spectral radius of X and we resolved the question for both symmetry classes. We also have a statement that the few rightmost eigenvalues form a Poisson point process (with correct rescalings).

How is this result related to local laws? We look at linear statistics $\frac{1}{N}\sum_j f(\sigma_j)$ with a carefully chosen test function f that is supported in a regime where the rightmost eigenvalue is supported (this is an anisotropic elongated rectangle around $z=1$). We use Girko's formula (1.4). We need to compute (among other things) the variance (or only the second moment) of this linear statistics, i.e.

$$\mathbb{E}\Big|\frac{1}{N}\sum_j f(\sigma_j)\Big|^2 = \frac{1}{16\pi^2}\iint_{\mathbb{C}} \Delta f(z)\Delta f(z')\iint_0^\infty \mathbb{E}\langle \Im G^z(i\eta)\rangle\langle \Im G^{z'}(i\eta')\rangle \mathrm{d}\eta\mathrm{d}\eta'\,\mathrm{d}^2z\mathrm{d}^2z'.$$

So we need to study the correlation of $\langle \Im G^z(i\eta)\rangle$ and $\langle \Im G^{z'}(i\eta')\rangle$ for all regimes of η. In the mesoscopic regime[7], $\eta \gg 1/N^{3/4}$, this question boils down to a two resolvent local law $\langle \Im G^z(i\eta)\Im G^{z'}(i\eta')\rangle$, and it is especially important to extract a decay in this correlation as $z-z'$ gets larger. Considering H^z from (1.3), this corresponds to a decay in the difference of the two deformations similarly to (1.27).

In the microscopic regime $\eta \sim N^{-3/4}$ we need to use DBM methods, i.e. we follow how the eigenvalues[8] λ^z and $\lambda^{z'}$ of $H^z, H^{z'}$, respectively, behave under the Dyson Brownian motion. The idea is that want to prove that λ^z and $\lambda^{z'}$ are essentially independent if $|z-z'| \gg N^{-1/2}$. We cannot do it directly, but we run a DBM for relatively short time, but long enough so that the initial conditions are forgotten. By a coupling argument we could show this independence if the driving Brownian motions were independent. In fact, they are not exactly independent, but their correlations are given by the overlap of the corresponding singular vectors (1.28) (which was proven by a two resolvent local law). Since this overlap is small if $|z-z'| \gg N^{-1/2}$, we can prove the asymptotic independence of λ^z and $\lambda^{z'}$ after some time. Finally, we use a *Green function comparison (GFT)* to remove the Gaussian component.

1.5.5. *Random band matrices (RBM).* So far we discussed *mean field models*, characterized by the property that each entry h_{ij} of the random matrix H is roughly of the same order $1/\sqrt{N}$. In the corresponding quantum mechanical system, where H is considered the Hamilton operator, this means that the quantum transition from any state i to any state j has roughly the same amplitude. In particular, the underlying state space $\{1,2,\dots,N\}$ is essentially zero dimensional. More sophisticated models involve a nontrivial spatial structure imposed on $\{1,2,\dots,N\}$. The simplest case is when $N=L^d$, with $d\geq 1$ being the physical dimension, and the state space is the discrete d-dimensional torus $\mathbf{T}_L^d$ with linear length L. We set a new physical parameter, W, the *band width* and we assume that for any $i,j\in\mathbf{T}_L^d$ the matrix element H_{ij} is nonzero only if $\mathrm{dist}(i,j)\lesssim W$, where dist is the natural (periodic) distance on the torus.

The main interest in random band matrices stems from the fact that by changing the band width, RBM's naturally interpolate between the *random Schrödinger operators* ($W\sim 1$) and the mean field RMT models ($W\sim L$). Accordingly, RBM's are expected to exhibit the *Anderson metal-insulator phase transition*, similarly to random Schrödinger operators. This asserts that the system has a localized and a delocalized phase, depending on the choice of the parameters. In the localized phase the eigenvectors are essentially supported on a set that is much smaller than the entire state space. They typically have a fast decay on a scale ℓ, called *localization length*, that is much smaller than the linear size of the system, $\ell \ll L$. Moreover, the local spectral statistics is uncorrelated, Poissonian. In contrast, in the delocalized phase, the localization length $\ell\sim L$ and the local eigenvalue statistics is the standard Wigner-Dyson statistics from mean field random matrice.

For example, in $d=1$ dimensions, the phase transition occurs at $W\sim\sqrt{N}$; for $W\ll\sqrt{N}$ we have localisation (see [47, 27, 12, 46, 37]), while for $W\gg\sqrt{N}$ we are in the fully delocalized phase, [52, 33]. These two regimes require very different strategies. Localization results typically go back to the basic ideas of the localization proofs for random Schrödinger operators, i.e. Fröchlich-Spencer multi-scale method [36] or Aizenman-Molchanov fractional moment method [4]. Delocalization results rely on extensions of mean field random matrix methods to the non-mean-field situation, see e.g. [31, 7, 11, 10, 51, 28] for earlier results on delocalisation in the regime $W\gg N^\alpha$ with non-optimal α (larger than 1/2). Very recently, the new dynamical methods have been applied to prove delocalization for RBM, first in [28] for $W\gg N^{8/11}$ and then in the entire regime $W\gg N^{1/2}$ in [52, 33]. The paper [52] used essentially the zig flow for proving the corresponding averaged multi-resolvent local law for the Gaussian case for a special class of random block band matrices. The full zig-zag strategy was implemented later in [33] to handle general distributions, variance profile and it obtained the general averaged and isotropic local laws with estimates that optimally incorporated the spatial dependence.

[7] The unusual power 3/4 comes from the fact that if $|z|\approx 1$, we are near the spectral edge of X, this corresponds to a cusp singularity in the spectrum of H^z at the origin

[8] Actually the really important ones are the eigenvalues close to 0, these influence $\Im G^z(i\eta)$ the most.

2. Single resolvent local law for Wigner matrices

Definition 2.1 (Wigner matrix)**.** *Let $H = H^*$ an $N \times N$ Hermitian (complex Hermitian, $\beta = 2$, or real symmetric, $\beta = 1$) random matrix whose matrix elements $\{h_{ab} : 1 \le a \le b \le N\}$ are independent random varianbles. The diagonal and off-diagonal elements have common distributions, respectively, with the normalisation*

$$h_{ab} \stackrel{\mathrm{d}}{=} \frac{1}{\sqrt{N}}\chi_{\mathrm{od}}, \quad 1 \le a < b \le N, \quad h_{aa} \stackrel{\mathrm{d}}{=} \frac{1}{\sqrt{N}}\chi_{\mathrm{d}}, \quad 1 \le a \le N,$$

where χ_{d} is a real random variable and χ_{od} is real or complex, depending on the symmetry class, with the conditions

$$\mathbb{E}\chi_{\mathrm{od}} = \mathbb{E}\chi_{\mathrm{d}} = 0, \qquad \mathbb{E}|\chi_{\mathrm{od}}|^2 = 1, \quad \mathbb{E}|\chi_{\mathrm{d}}|^2 = 1/\beta. \tag{2.1}$$

For simplicity, in the complex Hermitian case ($\beta = 2$) we assume that $\mathbb{E}\chi_{\mathrm{od}}^2 = 0$. Finally, we assume moment conditions (for simplicity, all moments), i.e. that for any[9] *$p \in \mathbf{N}$ there exists a C_p (N-independent) such that*

$$\mathbb{E}|\chi_{\mathrm{od}}|^p + \mathbb{E}|\chi_{\mathrm{d}}|^p \le C_p.$$

Random matrix ensembles satisfying these conditions will be called **Wigner matrices.**

The constants C_p will be considered as *model parameters*, meaning that all implicit constants in the entire discussion may depend on them. All statements are understood for *large enough* N, meaning that they hold if $N \ge N_0$ where the threshold N_0 may depend on the model parameters plus some other natural parameters that we will specify in the concrete statements.

The normalization of the matrix elements is chosen in such a way that $\|H\|$ remains of order one uniformly in N. One way to see it is to compute

$$\frac{1}{N}\sum_i \mathbb{E}\lambda_i^2 = \frac{1}{N}\mathbb{E}\mathrm{Tr}H^2 = \frac{1}{N}\sum_{ab}\mathbb{E}|h_{ab}|^2 = 1,$$

where λ_i, $i = 1, 2, \dots, N$ are the eigenvalues of H.

Denote by $m(z) = m_{sc}(z)$ the Stieltjes transform of the Wigner semicircle law $\rho_{sc}(x) = \frac{1}{2\pi}\sqrt{(4-x^2)_+}$:

$$m(z) = \int_{\mathbb{R}} \frac{\rho(x)\mathrm{d}x}{x - z}, \qquad z \in \mathbb{C} \setminus [-2, 2]. \tag{2.2}$$

It satisfies the quadratic equation

$$-\frac{1}{m} = m + z, \qquad (\Im m)(\Im z) > 0, \tag{2.3}$$

where the side condition makes the choice unique (we usually work in the upper half plane, $\eta = \Im z > 0$). By the inverse Stieltjes transform we have

$$\rho_{sc}(x) = \frac{1}{\pi}\lim_{\eta \to 0+} \Im m(x + i\eta).$$

We often extend ρ to the upper half plane

$$\rho(z) := \frac{1}{\pi}\Im m(z), \qquad \Im z > 0.$$

Simple elementary calculations show the following basic properties (with $\eta := |\Im z|$)

$$|m(z)| < 1, \qquad 1 - |m|^2 \sim \eta, \qquad |1 - m^2| \gtrsim \sqrt{\kappa}, \quad \kappa(z) := \min\{|\Re z - 2|,\ |\Re z + 2|\}. \tag{2.4}$$

Theorem 2.2 (Local Law in the Bulk)**.** *For any fixed (small) $\kappa > 0$, $\xi > 0$ and $\epsilon \gg \xi$, we have the* **averaged local law**

$$\langle (G(z) - m(z))B\rangle \le \frac{N^\xi}{N\eta}\|B\|, \qquad \eta = |\Im z| \ge N^{-1+\epsilon}, \tag{2.5}$$

for any deterministic matrix B, and **isotropic local law**

$$\langle \boldsymbol{x}, (G(z) - m(z))\boldsymbol{y}\rangle \le \frac{N^\xi}{N\eta}\|\boldsymbol{x}\|\|\boldsymbol{y}\|, \qquad \eta = |\Im z| \ge N^{-1+\epsilon}, \tag{2.6}$$

for any deterministic vectors $\boldsymbol{x}, \boldsymbol{y}$. These relations hold with very high probability, i.e. with probability larger than $1 - N^{-D}$ for any fixed (large) exponent D, if $N \ge N_0(\kappa, \xi, D)$ is large enough.

[9] In practice finitely many moments are sufficient.

Note that from $H = H^*$ we automatically have

$$\|G(z)\| \le \frac{1}{\eta}, \tag{2.7}$$

but in the interesting small η regime this bound does not properly reflect the true size of G. More precisely, as a norm bound it is optimal (if $z = \lambda + i\eta$ where λ is an eigenvalue with normalized eigenvector $\boldsymbol{u}$, $H\boldsymbol{u} = \lambda\boldsymbol{u}$, then $\langle \boldsymbol{u}, G\boldsymbol{u}\rangle = 1/\eta$), but if measured in some weaker sense, it turns out that G behaves more like an $O(1)$ random variable. In fact, $G(z)$ concentrates around the constant matrix $m(z)\cdot I$ if tested against deterministic test matrices B or test vectors $\boldsymbol{x}, \boldsymbol{y}$. Local laws (2.5)–(2.6) are precise statements of this concentration phenomenon.

We formulate all results for the simplest Wigner case in the bulk, but everything can be extended to much more general mean field matrix models. For example, we can consider *deformed* models, $H = A + W$, where A is a deterministic matrix, W is random, $\mathbb{E}W = 0$, and we can also have nontrivial variance structure of still independent matrix elements, $\mathbb{E}|w_{ij}|^2 =: S_{ij}$ with a matrix S such that $S_{ij} \lesssim 1/N$ (mean field condition), or even we can have correlations among the matrix elements (with some conditions). In this most general case, the analogue of (2.3) is the **Matrix Dyson Equation (MDE)**:

$$-\frac{1}{M} = z - A + \mathcal{S}[M], \qquad M = M(z), \quad (\Im M(z))\Im z > 0, \tag{2.8}$$

where $\mathcal{S} = \mathcal{S}_W$ is the *self-energy* operator of W, defined by

$$\mathcal{S}[R] := \mathbb{E}\, WRW \tag{2.9}$$

for any deterministic matrix R. The local law holds in great generality, just m in (2.5)–(2.6) is replaced with M, the matrix-valued deterministic approximation of the resolvent.

3. Proof of the single resolvent local law with zigzag strategy

We will now present a modern proof of the local law, Theorem 2.2, via the **zigzag strategy.** Historically, single G local law has been proven well before the zigzag strategy with other methods, but the new zigzag method is especially powerful and easily generalizes to multi-resolvent situations.

3.1. Overview of the zigzag strategy.

The zigzag strategy consists of three steps:

(1) Initial global bound at $\eta \sim 1$.
(2) The zig step, where the spectral parameter z_t is moved by a specific ODE, the *characteristic flow*, and at the same time, H evolves by an Ornstein-Uhlenbeck (OU) process. The characteristic flow reduces η, so we approach to the more local regime but at the expense of adding a Gaussian component (along the OU flow).
(3) The zag step, where we keep the spectral parameter fixed and use *Green function comparison (GFT)* argument to remove the Gaussian component.

Since the GFT is effective only if a small Gaussian component needs to be removed, we cannot go from the global scale $\eta \sim 1$ to the fully local scale $\eta \sim N^{-1+\delta}$ in one go; we need to iterate the zig and zag steps several times in tandem, see Fig. 1 later. At the end of Section 2, we will comment on an apparent mystery, namely that zig adds a Gaussian component and zag removes it, so it sounds somewhat tautological.

We remark, however, that the *three step strategy* used in the proof of Wigner-Dyson universality, see e.g. [34], operates with a somewhat similar idea on a very high level. There, after establishing an a priori bound (first step: local law and rigidity), the main (second) step is the proof of the robust emergence of the Wigner-Dyson statistics along the Dyson Brownian Motion (DBM), at the expense of adding a small Gaussian component. Finally, a GFT argument removes this Gaussian component. There is a competition between the two steps: longer the DBM runs (larger Gaussian component is created), the emergence of the universal statistics becomes more obvious; on the other hand, removal of a larger Gaussian component by GFT is harder. In the universality proof there is a big room: in the bulk regime (under the usual normalisation where the eigenvalue spacing is of order $1/N$) DBM should run for a time $t \gg 1/N$, i.e. add a Gaussian component of size a bit larger than $1/N$. Fortunately, the GFT can remove much larger Gaussian components; its dynamical version (following resolvents along the OU flow) can remove a Gaussian component of order $N^{-1/2-\epsilon}$, while the more sophisticated one-by-one Lindeberg replacement strategy (originally introduced in [50] in the context of random matrices) can even remove an order $N^{-\epsilon}$ Gaussian component.

Notice that in both proofs the more important part is the one that adds a Gaussian component; the key feature emerges in this step. Indeed, in the zigzag strategy, this step reduces $\Im z$, while in the DBM proof of universality, this step generates the universality itself; the price in both cases is the added Gaussian component. The GFT step in both proofs plays a secondary role to remove this component with a very different argument.

Going back to zigzag, before more details on each step, we mention previous references about the zigzag strategy. The first version of the characteristic flow appeared in a paper by Pastur [45], who related the complex Burgers equation to the evolution of the resolvent along the OU flow. The idea was later revived by Lee and Schnelli in [42] to prove edge universality for deformed Wigner matrices and then by von Soosten and Warzel in [49] where it was used to prove a local law, albeit a non-optimal one. The full power of the characteristic flow method has gradually been realized in the context of Dyson Brownian motion in [39, 1, 2, 3], and later for matrix models, e.g., in [9, 41, 40]. Its combination with a GFT argument to prove specific multi-resolvent local laws—that is, the full zigzag strategy—first appeared in [25] in the setting of Hermitization of non-Hermitian matrices, where the characteristics were explicitly used to resolve the instability problem. The full dynamical version of the master inequalities, systematically controlling the fluctuations of resolvent chains of arbitrary length, was introduced in [21]. There, it was used to study Wigner matrices—a setting where the averaged master inequalities remain self-consistent, forgoing the analysis of isotropic chains[10].

The first use of isotropic master inequalities in tandem with the averaged ones, both in their dynamical form, appeared first in [22], which also introduced the term *zigzag*. Since then, the full power of the zigzag method was successfully exploited in several models and regimes that were previously inaccessible due to instabilities [20, 26, 23, 32, 30]. By now, the zigzag strategy has proved to be a very powerful and robust tool for proving local laws across a wide class of mean-field matrix model and most recently even for random band matrices [33].

We remark that the zigzag strategy for the simplest Wigner case for a single resolvent in the bulk has never been written up in any paper, so the presentation I give here does not have a direct reference in the literature. The closest is [30], where the more complicated correlated case was discussed and focusing on the cusp regime. One may follow this paper with trivial correlation, i.e. $\mathcal{S}[R] = \langle R\rangle$, and ignore all ϱ factors indicating the dependence on the density (since in the bulk $\varrho \sim 1$).

Now we discuss all these three steps. The most important one is the zig step since here the key point, the reduction of η happens, so we start with this step.

3.2. **Zig step.** We define two flows, the first is a simple ODE on $\mathbb{C}$, it is called the *characteristic flow* and is given by

$$\mathrm{d}z_t = -\frac{1}{2}z_t\mathrm{d}t - m(z_t)\mathrm{d}t, \qquad z_{t=0} = z_0. \tag{3.1}$$

The second is a stochastic flow on the space of matrices. In our concrete Wigner example this will be the OU flow

$$\mathrm{d}H_t = -\frac{1}{2}H_t\mathrm{d}t + \frac{1}{\sqrt{N}}\mathrm{d}\mathcal{B}_t, \qquad H_{t=0} = H_0, \tag{3.2}$$

where $\mathcal{B}_t$ is a standard matrix-valued Brownian motion. But the precise form of the flow is not important, i.e. one could use the simple DBM as well, $\mathrm{d}H_t = N^{-1/2}\mathrm{d}\mathcal{B}_t$, see e.g. [28, 52], but then (3.1) needs a bit modification to achieve the same key cancellation, see (3.6) below. In any case, denote the stochastic flow (time dependent stochastic semigroup) by $\mathfrak{F}^t_{\mathrm{zig}}$, i.e.

$$H_t = \mathfrak{F}^t_{\mathrm{zig}}\big[H_0\big]. \tag{3.3}$$

Set

$$G_t = G_t(z_t) := \frac{1}{H_t - z_t}.$$

If we change only z_t:

$$\mathrm{d}G(z_t) = (\partial_t z_t)G(z_t)^2\mathrm{d}t. \tag{3.4}$$

If we change only $G = G_t$, by Ito calculus,

$$\begin{aligned}\mathrm{d}G =& -G(\mathrm{d}H)G + \langle G\rangle G^2\mathrm{d}t\\ =&\Big(\frac{1}{2}GHG + \langle G\rangle G^2\Big)\mathrm{d}t - \frac{1}{\sqrt{N}}G(\mathrm{d}\mathcal{B})G\\ =&\left[\frac{1}{2}G + \Big(\langle G\rangle + \frac{z}{2}\Big)G^2\right]\mathrm{d}t - \frac{1}{\sqrt{N}}G(\mathrm{d}\mathcal{B})G.\end{aligned} \tag{3.5}$$

Here, in the first line, we considered G as a function of N^2 random variables h_{ab}, $a, b = 1, \dots, N$ and wrote

$$\mathrm{d}G = \sum_{ab}(\partial_{ab}G)\mathrm{d}h_{ab} + \frac{1}{2}\sum_{ab}\sum_{cd}(\partial_{ab}\partial_{cd}G)\mathrm{d}h_{ab}\mathrm{d}h_{cd} = -G(\mathrm{d}H)G + \frac{1}{2}\,2G(\mathrm{d}H)G(\mathrm{d}H)G,$$

[10]Technically speaking, isotropic local laws were also derived in [21] and used in the GFT step, but, due to the control of traceless observables in the optimal Hilbert–Schmidt norm, the isotropic laws followed directly from the averaged ones.

where $\partial_{ab} = \partial_{h_{ab}}$, and we used

$$(\partial_{ab} G)_{cd} = -G_{ca} G_{bd}.$$

Then we used that

$$(\mathrm{d}H) G (\mathrm{d}H) = \langle G \rangle \mathrm{d}t$$

(in the complex Hermitian case, where $\mathrm{d}h_{ab} \mathrm{d}h_{cd} = \frac{1}{N} \delta_{ad} \delta_{bc} \mathrm{d}t$) giving

$$\frac{1}{2}\, 2G(\mathrm{d}H)G(\mathrm{d}H)G = G^2 \langle G \rangle.$$

In general, here we have the self-energy operator (2.9), i.e we get $(\mathrm{d}H)G(\mathrm{d}H) = \mathcal{S}[G]\mathrm{d}t$.

In the second line of (3.5) we used (3.2) and to arrive at the third line, we wrote out $GH = I + zG$.

Now if we combine (3.4)–(3.5), for the evolution of $G_t(z_t)$ we have

$$\mathrm{d}G_t(z_t) = \left[\frac{1}{2} G_t(z_t) + \Big(\partial_t z_t + \langle G_t(z_t) \rangle + \frac{z_t}{2} \Big) G_t(z_t)^2 \right] \mathrm{d}t - \frac{1}{\sqrt{N}} G_t(z_t) (\mathrm{d}\mathcal{B}) G_t(z_t). \tag{3.6}$$

Now we choose z_t such that the leading term $m(z_t)$ of $\langle G \rangle$ in the big bracket is cancelled, this is exactly the characteristic equation (3.1), the result is

$$\mathrm{d}G_t(z_t) = \left[\frac{1}{2} G_t(z_t) + \langle G_t(z_t) - m(z_t) \rangle G_t(z_t)^2 \right] \mathrm{d}t - \frac{1}{\sqrt{N}} G_t(z_t) (\mathrm{d}\mathcal{B}) G_t(z_t). \tag{3.7}$$

We are really interested in the fluctation, i.e. $G_t(z_t) - m(z_t)$, so we also compute from (3.1)

$$\frac{\mathrm{d}}{\mathrm{d}t} m(z_t) = -m'(z_t) \Big(\frac{1}{2} z_t + m(z_t) \Big).$$

Recall that m satisfies $-1/m = z + m$, i.e., $-1 = mz + m^2$ so after differentiating

$$0 = m + m' z + 2mm' \qquad \Longrightarrow \qquad m'(z + 2m) = -m,$$

and thus we have

$$\frac{\mathrm{d}}{\mathrm{d}t} m(z_t) = \frac{m(z_t)}{2} \qquad \Longrightarrow \qquad m(z_t) = e^{t/2} m(z_0). \tag{3.8}$$

Putting (3.8) and (3.7) together, we have

$$\mathrm{d}\big(G_t(z_t) - m(z_t) \big) = \left[\frac{1}{2} \big(G_t(z_t) - m(z_t) \big) + \langle G_t(z_t) - m(z_t) \rangle G_t(z_t)^2 \right] \mathrm{d}t - \frac{1}{\sqrt{N}} G_t(z_t) (\mathrm{d}\mathcal{B}) G_t(z_t). \tag{3.9}$$

We will now focus on the average local law (2.5), and for simplicity set $B = I$. Using $G - m := G_t(z_t) - m(z_t)$, $G = G_t(z_t)$ for short, we have

$$\mathrm{d}\langle G - m \rangle = \left[\frac{1}{2} \langle G - m \rangle + \langle G - m \rangle \langle G^2 \rangle \right] \mathrm{d}t - \frac{1}{\sqrt{N}} \langle G (\mathrm{d}\mathcal{B}) G \rangle. \tag{3.10}$$

Before we continue, let's analyse a bit the sizes of various terms. First, notice that by taking the imaginary part of (3.1), we have

$$\frac{\mathrm{d}\eta_t}{\mathrm{d}t} = -\frac{1}{2} \eta_t - \Im m(z_t), \tag{3.11}$$

and as long as we are in the bulk regime, $\Im m(z_t) \sim 1$, thus

$$\frac{\mathrm{d}\eta_t}{\mathrm{d}t} \sim -1.$$

So $t \to \eta_t$ decreases roughly linearly, i.e.

$$\eta_t \approx c_t (T_* - t) \tag{3.12}$$

with some function $c_t \sim 1$, where T_* is the final time when $\eta_{T_*} = 0$. Technically, we will run the process up to time $T \sim T_* - N^{-1+\delta}$ since we are interested in final $\eta_T \sim N^{-1+\delta}$.

Making the ansatz that $\langle G - m \rangle \sim 1/N\eta_t$, for the LHS of (3.10) we have

$$\mathrm{d}\langle G - m \rangle \sim \mathrm{d} \frac{1}{N\eta_t} \sim \frac{1}{N\eta_t^2}.$$

On the RHS, the bigger term is the one with $\langle G^2\rangle$, its rough size[11] is bounded by

$$|\langle G^2\rangle| \le \langle |G|^2\rangle = \frac{1}{\eta}\langle \Im G\rangle \sim \frac{1}{\eta}$$

using the *Ward identity*, $GG^* = \frac{1}{\eta}\Im G$ which holds for the resolvent of any self-adjoint matrix H. Thus we have

$$\left[\frac{1}{2}\langle G-m\rangle + \langle G-m\rangle\langle G^2\rangle\right] \sim \frac{1}{N\eta_t^2},$$

i.e. the first term in the RHS of (3.10) is consistent with the LHS. Note that this consistency is achieved only because we carefully choose $\partial_t z_t$ which reduced the corresponding factor

$$\left(\partial_t z_t + \langle G_t(z_t)\rangle + \frac{z_t}{2}\right) \sim 1 \tag{3.13}$$

in (3.6) to a much smaller term $\langle G-m\rangle \sim 1/(N\eta)$.

Finally, we discuss the martingale term in (3.10). As usual, we compute its quadratic variation process:

$$\left[\frac{1}{\sqrt{N}}\langle G(\mathrm{d}\mathcal{B})G\rangle, \frac{1}{\sqrt{N}}\langle G(\mathrm{d}\mathcal{B})G\rangle\right] = \frac{1}{N^3}\Big[\sum_{ij}(G^2)_{ij}\mathrm{d}\mathcal{B}_{ji}, \sum_{kl}(G^2)_{kl}\mathrm{d}\mathcal{B}_{lk}\Big] = \frac{1}{N^3}\sum_{ij}|(G^2)_{ij}|^2\mathrm{d}t. \tag{3.14}$$

Here in the last step we assumed, for simplicity, that we look at the complex Hermitian symmetry class, which means $[\mathrm{d}\mathcal{B}_{ji}, \mathrm{d}\mathcal{B}_{lk}] = \delta_{il}\delta_{jk}\mathrm{d}t$. Using again the Ward identity, we have

$$\frac{1}{N^3}\sum_{ij}|(G^2)_{ij}|^2 = \frac{1}{N^2}\langle |G|^4\rangle \le \frac{1}{N^2\eta^2}\langle |G|^2\rangle = \frac{1}{N^2\eta^3}\langle \Im G\rangle \sim \frac{1}{N^2\eta^3}. \tag{3.15}$$

To get the effect of the martingale term in (3.10) we need to integrate the quadratic variation and take the square root, roughly speaking

$$\sqrt{\int_0^t \frac{1}{N^2\eta_s^3}\mathrm{d}s} \sim \sqrt{\frac{1}{N^2\eta_t^2}} = \frac{1}{N\eta_t},$$

so this is also consistent with the target bound $\langle G-m\rangle \sim 1/(N\eta)$.

Now we discuss how to make this argument more rigorous. Define the spectral domains

$$\mathcal{D}_\gamma := \{z = E+i\eta \in \mathbb{C}\,;\ \eta \ge N^{-1+\gamma}, |E| \le 2-\kappa\} \tag{3.16}$$

for any $0 < \gamma \le 1$. We prove the following

Proposition 3.1 (Zig step). *Fix two exponents $\gamma_1 < \gamma_0$. Suppose that the local law*

$$|\langle G(z_0) - m(z_0)\rangle| \le \frac{N^{\xi/2}}{N\eta_0}, \qquad z_0 \in \mathcal{D}_{\gamma_0}, \tag{3.17}$$

holds with very high probability for any bulk spectral parameter z_0 with $\eta_0 = \Im z_0 \gtrsim N^{-1+\gamma_0}$. Let z_t satisfy (3.1) and G_t be the resolvent of the H_t from the OU flow (3.2). Then for H_t we have the local law up to times t

$$|\langle G_t(z_t) - m(z_t)\rangle| \le \frac{N^{\xi}}{N\eta_t}, \qquad \forall t, \Im z_t = \eta_t \ge N^{-1+\gamma_1} \tag{3.18}$$

with very high probability.

One could use this zig proposition in one step, i.e. choosing $\gamma_0 = 1$ and $\gamma_1 = \delta$, i.e. propagate the global law, (3.17) for $\gamma_0 = 1$, down to the smallest scale. The global law will be proven in a separate Section 3.4, but notice that this is expected to be an easier task than the local law since for $\eta \sim 1$ there is no discrepancy between the crude norm bound (2.7) and the more refined bounds in weaker sense.

In the actual proof we will do zig and zag steps alternatingly several times, this is because the zag step cannot remove too big Gaussian component in one go. So we will use Proposition 3.1 in such a way that we will define a sequence of thresholds $N^{-j\delta}$, $j = 1, 2, \dots, 1/\delta$, and assuming (3.17) for $\gamma_0 = 1 - j\delta$, we will conclude (3.18) for $\gamma_1 = 1-(j+1)\delta$.

[11] We note that the true size is

$$\langle G^2\rangle \sim 1 + \frac{1}{N\eta^2},$$

but this requires a local law for G^2 which would make the argument circular if we tried to use it for a proof.

Proof of Proposition 3.1. We fix a small exponent $\xi > 0$ and define the stopping time

$$\tau := \inf\Big\{ s > 0 \,;\, \big|\langle G_s(z_s) - m(z_s)\rangle\big| > \frac{N^\xi}{N\eta_s}\Big\} \wedge T. \tag{3.19}$$

We are given a final z with $\eta = \Im z$ where we want to prove (2.5). Choose an initial condition $z = z_0$ with $\Im z \geq N^{-1+\gamma_0}$ such that after time $T \sim N^{-1+\gamma_0}$ the solution of (3.1) is $z_T = z$ (simple ODE theory shows that it is possible). We want to show that $\tau = T$ with very high probability.

First, we need to show that $\tau > 0$, by continuity this follows from

$$|\langle G(z_0) - m(z_0)\rangle| \leq \frac{N^{\xi/2}}{N\eta_0}, \qquad \eta_0 = \Im z_0 \geq N^{-1+\gamma_0}. \tag{3.20}$$

Second, in the main part of the proof, we need to show that τ cannot be smaller than T, i.e. the estimate $\langle G - m\rangle \lesssim N^\xi/N\eta$ is consistently maintained along the flow (this is basically what our earlier heuristics showed). Here we can use the basic information from the definition of τ that

$$|X_s| \leq \frac{N^\xi}{N\eta_s}, \qquad \forall s \leq \tau,$$

where we set

$$X_s := \langle G_s(z_s) - m(z_s)\rangle$$

for brevity.

The structure of (3.10) is the following:

$$\mathrm{d}X_s = \mathcal{E}_1(s) X_s \mathrm{d}s + \mathrm{d}\mathcal{E}_2(s). \tag{3.21}$$

For the coeffient of the linear term (the *generator*) we have

$$|\mathcal{E}_1(s)| \leq \frac{1}{2} + \frac{\langle \Im G_s\rangle}{\eta_s} \leq \frac{1}{2} + \frac{\Im m_s}{\eta_s}(1 + o(1)), \qquad s \leq \tau. \tag{3.22}$$

Here we used the definition of the stopping time and $s \leq \tau$ to control

$$\langle \Im G_s\rangle = \Im m_s + \langle \Im(G - m)_s\rangle \leq \Im m_s + \frac{N^\xi}{N\eta_s} \leq \Im m_s(1 + o(1)), \qquad s \leq \tau,$$

and that $N^\xi/N\eta \leq N^{\xi-\delta} \ll 1$.

For the martingale term, by (3.15) and the stopping time again, we have

$$\big[\mathrm{d}\mathcal{E}_2(s), \mathrm{d}\mathcal{E}_2(s)\big] \leq \frac{\langle \Im G_s\rangle}{N^2\eta_s^3}\mathrm{d}s \lesssim \frac{1}{N^2\eta_s^3}\mathrm{d}s, \qquad s \leq \tau. \tag{3.23}$$

By the BDG (Burkholder-Davis-Gundy) inequality,

$$\max_{s\leq t\wedge\tau}\Big|\int_0^s \mathrm{d}\mathcal{E}_2(r)\Big| \lesssim \left(\mathbb{E}\Big|\int_0^{t\wedge\tau} \mathrm{d}\mathcal{E}_2(s)\Big|^2\right)^{1/2} = \left(\int_0^{t\wedge\tau} \big[\mathrm{d}\mathcal{E}_2(s), \mathrm{d}\mathcal{E}_2(s)\big]\right)^{1/2} \lesssim \left(\int_0^{t\wedge\tau} \frac{1}{N^2\eta_s^3}\mathrm{d}s\right)^{1/2} \lesssim \frac{1}{N\eta_{t\wedge\tau}}$$

with very high probability. Here we used again that

$$\int_0^t \frac{\mathrm{d}s}{\eta_s^\alpha} \lesssim \frac{1}{\eta_t^{\alpha-1}}, \qquad \alpha > 1,$$

that is based upon (3.12).

We now solve (3.21) by (stochastic) Gromwall inequality up to time $t \wedge \tau$ with any $t \leq T$. Let $\mathcal{P}_{s,t}$ be the propagator (here it is just a scalar):

$$\mathcal{P}_{s,t} := \exp\Big(\int_s^t \mathcal{E}_1(r)\mathrm{d}r\Big),$$

then

$$X_t = \mathcal{P}_{0,t}X_0 + \int_0^t \mathcal{P}_{s,t}\mathrm{d}\mathcal{E}_2(s),$$

i.e. using that $\mathcal{P}_{s,t}$ is positive, we have

$$|X_t| \lesssim \mathcal{P}_{0,t}|X_0| + \Big|\int_0^t \mathcal{P}_{s,t}\mathrm{d}\mathcal{E}_2(s)\Big|. \tag{3.24}$$

We estimate the last term by the BDG inequality and (3.23) in the last step, for any $t \le \tau$:

(3.25)
$$\Big|\int_0^t \mathcal{P}_{s,t}\mathrm{d}\mathcal{E}_2(s)\Big| \lesssim \left(\mathbb{E}\Big|\int_0^t \mathcal{P}_{s,t}\mathrm{d}\mathcal{E}_2(s)\Big|^2\right)^{1/2} = \left(\int_0^t\int_0^t \mathcal{P}_{s,t}\mathcal{P}_{s',t}[\mathrm{d}\mathcal{E}_2(s),\mathrm{d}\mathcal{E}_2(s')]\right)^{1/2} = \left(\int_0^t (\mathcal{P}_{s,t})^2\frac{1}{N^2\eta_s^3}\mathrm{d}s\right)^{1/2}$$

with very high probability.

Now we estimate the propagator for any $s \le t \le \tau$, using (3.22):

$$\mathcal{P}_{s,t} \le \exp\Big((1+o(1))\int_s^t \Big(\frac{1}{2}+\frac{\Im m_r}{\eta_r}\Big)\mathrm{d}r\Big).$$

From (3.11) we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\log\eta_t) = -\frac{1}{2}-\frac{\Im m_t}{\eta_t} \qquad \Longrightarrow \qquad \int_s^t\Big(\frac{1}{2}+\frac{\Im m_r}{\eta_r}\Big)\mathrm{d}r = -\log(\eta_t/\eta_s),$$

thus we have

$$\mathcal{P}_{s,t} \lesssim \frac{\eta_s}{\eta_t} \tag{3.26}$$

(ignoring the irrelevant $1+o(1)$ factor).

Using $|X_0| \le N^{\xi/2}/N\eta_0$ from (3.20), plugging (3.26) into (3.24) with (3.25), we have

$$|X_t| \lesssim \frac{\eta_0}{\eta_t}\frac{N^{\xi/2}}{N\eta_0} + \left(\int_0^t\Big(\frac{\eta_s}{\eta_t}\Big)^2\frac{1}{N^2\eta_s^3}\mathrm{d}s\right)^{1/2} \lesssim \frac{N^{\xi/2}}{N\eta_t} \tag{3.27}$$

for any $t \le \tau$ with very high probability using that

$$\int_0^t \frac{1}{\eta_s}\mathrm{d}s \le \log N, \qquad t \le T.$$

The bound (3.27) together with the continuity of X_t in time and the definition of the stopping time τ (note the factor $N^{\xi/2} \ll N^{\xi}$!) shows that $\tau = T$. This completes the proof of Proposition 3.1. □

3.3. **Zag step.** In this zag step we remove the Gaussian component introduced in the zig step. Unlike in the zig step, here the spectral parameter z is fixed and thus we will often omit if from the notation.

Let $\mathfrak{F}^t_{\mathrm{zag}}$ be the flow of an OU process, i.e.

$$H_t = \mathfrak{F}^t_{\mathrm{zag}}\big[H_0\big], \tag{3.28}$$

where H_t is the solution to

$$\mathrm{d}H_t = -\frac{1}{2}H_t\mathrm{d}t + \frac{1}{\sqrt{N}}\mathrm{d}\mathcal{B}_t, \qquad H_{t=0} = H_0. \tag{3.29}$$

Let H_t again[12] solve the OU equation (3.2). We run this OU in a time interval $s \in [0, s_{\mathrm{final}}]$. Recall the spectral domains

$$\mathcal{D}_\gamma := \{z = E+i\eta \in \mathbb{C}\ ;\ \eta \ge N^{-1+\gamma}\}.$$

The following is the main proposition of the zag step.

[12]We remark that in general the OU flow in the zig and zag steps are different, here for the simplest Wigner case they happen to be the same. In general, the zig flow uses the standard OU flow (3.2) that adds a pure GUE/GOE component. In contrast, the zag flow solves a modified OU process

$$\mathrm{d}H_t = -\frac{1}{2}(H_t - \mathbb{E}H_t)\mathrm{d}t + \Sigma_{H_0}^{1/2}[\mathrm{d}\mathcal{B}_t], \tag{3.30}$$

where Σ_H is the covariance tensor of any random matrix H, defined for any fixed deterministic matrix R by

$$\Sigma[R] = \Sigma_H[R] := \mathbb{E}\Big[\mathrm{Tr}\big[(H-\mathbb{E}H)R\big](H-\mathbb{E}H)\Big].$$

It is easy to see that Σ is positive (as a self-adjoint map on the space of $N\times N$ matrices equipped with the Hilbert-Schmidt scalar product), and $\Sigma^{1/2}$ is its operator square root. This modified OU process is designed in such a way that the covariance tensor Σ_{H_t}, hence the self-energy operator $\mathcal{S}_{H_t}$ remain constant along the time evolution. Recall that $\mathcal{S}_H[R] = \mathbb{E}HRH$, i.e. $\mathcal{S}_H$ and Σ_H are not the same, but they contain the same information. In the complex Wigner case

$$\mathcal{S}[R] = \langle R\rangle, \qquad \Sigma[R] = \frac{1}{N}R,$$

i.e. $\Sigma = \frac{1}{N}\cdot I$, which means that (3.30) becomes (3.29), i.e. the standard OU process.

In general, we have quite a freedom to choose the zig flow (if the characteristic flow (3.1) adjusted), while the zag flow is quite rigid, it must be chosen such that covariance tensors of the matrix elements remain invariant along the flow.

Proposition 3.2. *Fix two (small) exponents $\delta \ll \xi$. Fix constants $0 < \gamma_0 < 1$ and $\gamma_1 \geq \gamma_0 - \delta$. Suppose we have the bounds*

$$\left|(G_s(z))_{\boldsymbol{uv}}\right| \lesssim 1 \tag{3.31}$$

hold with very high probability, uniformly in $s \in [0, s_{\mathrm{final}}]$, in $z \in \mathcal{D}_{\gamma_0}$ and in deterministic unit vectors $\boldsymbol{u}, \boldsymbol{v}$. Assume the resolvents G_s satisfy the local laws (2.5)–(2.6) (with ξ as tolerance exponent) for any $z \in \mathcal{D}_{\gamma_1}$ at the final time $s = s_{\mathrm{final}}$. Then G_s satisfies the local laws uniformly for all $s \in [0, s_{\mathrm{final}}]$ (with a slightly larger ξ exponent that we will ignore in this presentation, but the increase of ξ is tiny compared with δ).

We start with two preparatory steps. First, we use Ito's formula that asserts for any $f \in C^2$ function that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}f(H_t) = -\frac{1}{2}\mathbb{E}\sum_{\alpha} h_\alpha(t)(\partial_\alpha f)(H_t) + \frac{1}{2N}\sum_{\alpha,\beta}\kappa_t(\alpha,\beta)\mathbb{E}(\partial_\alpha\partial_\beta f)(H_t)\,, \tag{3.32}$$

where $\kappa_t(\alpha,\beta)$ denotes the (normalized) second order cumulant of the rescaled matrix entries $\sqrt{N}h_\alpha(t)$ and $\sqrt{N}h_\beta(t)$, i.e. $\kappa_t(\alpha,\beta) = N\mathbb{E}h_\alpha(t)h_\beta(t)$ since $\mathbb{E}h = 0$, and $\partial_\alpha = \partial_{h_\alpha}$ is the partial derivative wrt to the α-th matrix element. The first summand on the rhs. of (3.32) can now be further treated by cumulant expansion, which is first key ingredient for our proof. Below we give precise references, but in the separate Section 5 we give a pedagogical introduction to it.

Proposition 3.3 (Multivariate cumulant expansion; cf. Proposition 3.2 in [29] and Lemma 3.1 in [38])**.** *Let $f : \mathbb{R}^{N\times N} \to \mathbb{C}$ be a L times differentiable function with bounded derivatives. Then, for any double index $\alpha_0 = (a,b) \in [N]^2$ it holds that*

$$\mathbb{E}h_{\alpha_0}f(H) = \sum_{k=0}^{L-1}\sum_{\boldsymbol{\alpha}\in\{ab,ba\}^k}\frac{\kappa(\alpha_0,\boldsymbol{\alpha})}{N^{(k+1)/2}k!}\mathbb{E}(\partial_{\boldsymbol{\alpha}}f)(H) + \Omega_L(f,\alpha_0), \tag{3.33}$$

where $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_k)$ and $\partial_{\boldsymbol{\alpha}} = \partial_{h_{\alpha_1}}...\partial_{h_{\alpha_k}}$ for $k \geq 1$, and for $k = 0$ is considered as the function f itself. For completeness, we mention that the error term in (3.33) satisfies the following bound (for all practical purposes one can just ignore this term)

$$\left|\Omega_L(f,\alpha_0)\right| \lesssim \frac{C_L}{N^{(L+1)/2}}\sum_{\boldsymbol{\alpha}\in\{ab,ba\}^L}\sup_{\lambda\in[0,1]}\left(\mathbb{E}\left|(\partial_{\boldsymbol{\alpha}}f)(\lambda H|_{\{ab,ba\}} + H|_{[N]^2\setminus\{ab,ba\}})\right|^2\right)^{1/2}, \tag{3.34}$$

for some constant $C_L > 0$ depending only on L. The notation $H|_{\mathcal{N}}$ for $\mathcal{N} \subset [N]^2$ in (3.34) refers to the matrix which equals H at all entries $\alpha \in \mathcal{N}$ and is zero otherwise.

Note that the $k = 1$ term in the expansion of the first summand on the rhs. of (3.32) exactly cancels the second summand on the rhs. of (3.32). For Proposition 3.3 being practically applicable we need to control (i) every order of the expansion, and (ii) the truncation error term Ω. Ignore the error term for this presentation.

Lemma 3.4 (Monotonicity estimate)**.** *Fix a constant $0 < \gamma_0 \leq 1$ and assume that the very-high-probability bound (3.31) holds uniformly in $z \in \mathcal{D}_{\gamma_0}$ and $s \in [0, s_{\mathrm{final}}]$, for any deterministic $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^N$ with $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$.*

Fix $\gamma_1 \geq \gamma_0 - \delta$. Then, we have

$$|G^s(E + \mathrm{i}\eta_1)_{\boldsymbol{uv}}| \lesssim \frac{\eta_0}{\eta_1}\,, \tag{3.35}$$

with very high probability, uniformly in $z := E + \mathrm{i}\eta_1 \in \mathcal{D}_{\gamma_1}$ for any time $s \in [0, s_{\mathrm{final}}]$, and for any deterministic vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^N$ with $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$.

Proof of Lemma 3.4. The map $\eta \mapsto \eta^2/(x^2+\eta^2)$ is increasing in $\eta > 0$ for any $x \in \mathbb{R}$, hence it follows by spectral decomposition of $\Im G$ that

$$\eta_1\Im G(E + \mathrm{i}\eta_1) \leq \eta_0\Im G(E + \mathrm{i}\eta_0), \tag{3.36}$$

in the sense of quadratic forms.

Using the Schwarz inequality and the Ward identity, we deduce that for all $0 < \eta < \eta_0$,

$$\left|\frac{\mathrm{d}}{\mathrm{d}\eta}\big(G(E+\mathrm{i}\eta)\big)_{\boldsymbol{uv}}\right| \lesssim \frac{\left|\big(\Im G(E+\mathrm{i}\eta)\big)_{\boldsymbol{uu}}\big(\Im G(E+\mathrm{i}\eta)\big)_{\boldsymbol{vv}}\right|^{1/2}}{\eta} \lesssim \frac{\eta_0}{\eta^2}, \tag{3.37}$$

where in the second step we used the monotonicity of the map $\eta \mapsto \eta\Im G(E+\mathrm{i}\eta)$, and the bound (3.35). Integrating the bound (3.37) from η_1 to η_0, we obtain

$$\left|\big(G(E+\mathrm{i}\eta_1)\big)_{\boldsymbol{uv}}\right| \lesssim \left|\big(G(E+\mathrm{i}\eta_0)\big)_{\boldsymbol{uv}}\right| + \frac{\eta_0}{\eta_1} \lesssim \frac{\eta_0}{\eta_1}. \tag{3.38}$$

□

The isotropic part of Proposition 3.2 will be concluded in a self-contained way, based entirely on the *isotropic Gronwall estimate* in Proposition 3.5 below. We will explain this in detail. Then the isotropic local law serves as an input for the similar average analysis that we skip.

Proof of isotropic part of Proposition 3.2. The key step is the following Gronwall estimate:

Proposition 3.5 (Isotropic Gronwall estimate)**.** *Assume the conditions of Proposition 3.2. Fix $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^N$ of bounded norm, $z := E + \mathrm{i}\eta_1 \in \mathcal{D}_{\gamma_1}$ and $\eta_0 \geq N^{-1+\gamma_0} \vee \eta_1$ such that $\eta_0/\eta_1 \leq N^{\delta}$. Set $s_{\mathrm{final}} \lesssim \eta_0$ and $\eta_0 \lesssim T \sim N^{-\xi/100}$ and we assume $\delta \ll \xi$. For $s \in [0, s_{\mathrm{final}}]$, define*

$$S_s := \big(G_s(E+\mathrm{i}\eta_1) - M(E+\mathrm{i}\eta_1)\big)_{\boldsymbol{xy}}\,. \tag{3.39}$$

Then, for any (large) even $p \in \mathbf{N}$, it holds that

$$\left|\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}|S_s|^p\right| \lesssim \left(1 + \frac{N^{10\delta}}{\sqrt{\eta_0}}\right)\big[\mathbb{E}|S_s|^p + (\Psi(\eta_1))^p\big], \qquad \Psi(\eta) := \frac{1}{\sqrt{N\eta}}, \tag{3.40}$$

uniformly in $s \in [0, s_{\mathrm{final}}]$, bounded $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^N$, and $z \in \mathcal{D}_{\gamma_1}$.

By Gronwall's lemma, uniformly in $s \in [0, s_{\mathrm{final}}]$, from (3.40) we find that

$$\begin{aligned}\mathbb{E}|S_s|^p &\lesssim \exp\left(\left(1 + \frac{N^{10\delta}}{\sqrt{\eta_0}}\right)(s_{\mathrm{final}} - s)\right)\big[\mathbb{E}|S_{s_{\mathrm{final}}}|^p + (\Psi(\eta_1))^p\big] \\ &\lesssim \mathbb{E}|S_{s_{\mathrm{final}}}|^p + (\Psi(\eta_1))^p\,.\end{aligned} \tag{3.41}$$

Here we used that $s_{\mathrm{final}} \lesssim \eta_0$, so the exponent is bounded by $N^{10\delta}\sqrt{\eta_0} \leq N^{10\delta}T \leq N^{10\delta-\xi/100} \leq 1$ if $\delta \ll \xi$. We point out that in (3.41), we use the final value rather than the initial value, as is more customary in a typical Gronwall argument, since in the zigzag strategy, illustrated in Figure 1, the endpoint of the flow is the known object.

To estimate $\mathbb{E}|S_{s_{\mathrm{final}}}|^p$, recall that the resolvent G^s satisfies the isotropic local law from the zig step Proposition 3.1 at $s = s_{\mathrm{final}}$. Therefore, since p in (3.41) was arbitrary, we find that

$$\left|\big(G_s(z) - M(z)\big)_{\boldsymbol{xy}}\right| \leq N^{\xi}\sqrt{\frac{1}{N\eta_1}}, \tag{3.42}$$

uniformly in $z := E + \mathrm{i}\eta_1 \in \mathcal{D}_{\gamma_1}$, $s \in [0, s_{\mathrm{final}}]$, and bounded $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^N$, with very high probability.

This proves the isotropic part of Proposition 3.2 the average part is similar. □

Proof of Proposition 3.5. Using (3.32), we have

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}|S_s|^p = -\frac{1}{2}\mathbb{E}\sum_{\alpha_1} h_{\alpha_1}(s)(\partial_{\alpha_1}|S_s|^p) + \frac{1}{2N}\sum_{\alpha_1,\alpha_2}\kappa_s(\alpha_1,\alpha_2)\mathbb{E}\big[\partial_{\alpha_1}\partial_{\alpha_2}|S_s|^p\big]. \tag{3.43}$$

The first term on the rhs. of (3.43) can now be expanded by means of Proposition 3.3:

$$\mathbb{E}\big[h_{\alpha_1}(s)(\partial_{\alpha_1}|S_s|^p)\big] = \sum_{k=0}^{L-1}\sum_{\boldsymbol{\alpha}\in\{\alpha_1,\alpha_1^t\}^k}\frac{\kappa_s(\alpha_1,\boldsymbol{\alpha})}{N^{(k+1)/2}\,k!}\mathbb{E}\big[\partial_{\alpha_1}\partial_{\boldsymbol{\alpha}}|S_s|^p\big] + \Omega_L\,. \tag{3.44}$$

The $k = 0$ is zero since the first cumulants vanish. The $k = 1$ term cancels the second term on the rhs. of (3.43), so the sum effectively starts from $k \geq 2$. If L is large enough, the error Ω_L is negligible. We thus find

$$\left|\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}|S_s|^p\right| \lesssim \left|\sum_{k=2}^{L-1}\sum_{\alpha_1}\sum_{\boldsymbol{\alpha}\in\{\alpha_1,\alpha_1^t\}^k}\frac{\kappa_s(\alpha_1,\boldsymbol{\alpha})}{N^{(k+1)/2}\,k!}\mathbb{E}\big[\partial_{\alpha_1}\partial_{\boldsymbol{\alpha}}|S_s|^p\big]\right|. \tag{3.45}$$

The most critical is the $k = 2$ (third order cumulant) term:

$$\left|N^{-3/2}\sum_{\alpha_1,\alpha_2,\alpha_3}\kappa(\alpha_1,\alpha_2,\alpha_3)\mathbb{E}\big[\partial_{\alpha_1}\partial_{\alpha_2}\partial_{\alpha_3}|S|^p\big]\right|. \tag{3.46}$$

The cumulants $\kappa(\alpha_1,\alpha_2,\alpha_3)$ are nonzero (order 1) only if all $\alpha_1,\alpha_2,\alpha_3$ are the same up to transposition. Distributing the derivatives according to the Leibniz rule, we need to estimate various terms of the forms $(\partial_\alpha^3 S)|S|^{p-1}$, $(\partial_\alpha S)(\partial_\alpha^2 S)|S|^{p-2}$, and $(\partial_\alpha S)^3|S|^{p-3}$.

We estimate the first type of terms, the others are similar. The cumulants $\kappa(\alpha_1,\alpha_2,\alpha_3)$ are nonzero (order 1) only if all $\alpha_1,\alpha_2,\alpha_3$ are the same up to transposition (i.e. if $\alpha_1=(a,b)$, then both α_2,α_3 are either (a,b) or (b,a)) so when computing $\partial_\alpha^3 S=\partial_\alpha^3 G_{\boldsymbol{xy}}$, we have a few types of terms

$$G_{\boldsymbol{x}a}G_{bb}G_{aa}G_{b\boldsymbol{y}},\quad G_{\boldsymbol{x}a}G_{ba}G_{bb}G_{a\boldsymbol{y}},\quad G_{\boldsymbol{x}a}G_{ba}G_{ba}G_{b\boldsymbol{y}}.$$

Focus on the first one, the others are similar, its contribution to (3.46) is estimated

$$\begin{aligned}
N^{-3/2}|S|^{p-1}\sum_{ab}\big|G_{\boldsymbol{x}a}G_{bb}G_{aa}G_{b\boldsymbol{y}}\big| &\le N^{-3/2}N^{2\delta}|S|^{p-1}\Big(\sum_a|G_{\boldsymbol{x}a}|\Big)\Big(\sum_b|G_{b\boldsymbol{y}}|\Big)\\
&\le N^{-3/2}N^{2\delta}|S|^{p-1}N\Big(\sum_a|G_{\boldsymbol{x}a}|^2\Big)^{1/2}\Big(\sum_b|G_{b\boldsymbol{y}}|^2\Big)^{1/2}\\
&\le N^{-1/2}N^{2\delta}|S|^{p-1}\eta_1^{-1}\Big(\Im G_{\boldsymbol{xx}}\Im G_{\boldsymbol{yy}}\Big)^{1/2}\le N^{1/2+3\delta}\frac{1}{N\eta_1}|S|^{p-1}\\
&\le N^{1/2+4\delta}\Psi(\eta_0)\big[|S|^p+\Psi(\eta_1)^p\big],
\end{aligned}\tag{3.47}$$

where we estimated $|G_{bb}|,|G_{aa}|\le N^\delta$ using Lemma 3.4 (recall $\eta_0/\eta_1\le N^\delta$), and finally we used Schwarz inequality and Ward identity. In the last step we wrote

$$\frac{1}{N\eta_1}|S|^{p-1}=\Psi(\eta_1)^2|S|^{p-1}\le N^{\delta/2}\Psi(\eta_0)\Psi(\eta_1)|S|^{p-1}\lesssim N^{\delta/2}\Psi(\eta_0)\big[|S|^p+\Psi(\eta_1)^p\big]$$

using Young's inequality. Similar estimates hold for all other terms, thus we conclude the following bound for the third cumulant term:

$$\Big|N^{-3/2}\sum_{\alpha_1,\alpha_2,\alpha_3}\kappa(\alpha_1,\alpha_2,\alpha_3)\mathbb{E}\big[\partial_{\alpha_1}\partial_{\alpha_2}\partial_{\alpha_3}|S|^p\big]\Big|\lesssim N^{1/2+4\delta}\Psi(\eta_0)\big[\mathbb{E}|S|^p+\Psi(\eta_1)^p\big]\qquad\text{w.v.h.p.}$$

The estimate of the higher order cumulant terms (with $n=k+1\ge 4$)

$$\Big|N^{-n/2}\sum_{\alpha_1,\dots,\alpha_n}\kappa(\alpha_1,\dots,\alpha_n)\mathbb{E}\big[\partial_{\alpha_1}\dots\partial_{\alpha_n}|S|^p\big]\Big|.\tag{3.48}$$

are similar, in fact a bit easier and give the same estimate. The reason why they are easier is that as n increases, the prefactor $N^{-n/2}$ gets better, but the summation is still effectively over N^2 indices. Higher derivatives yield more G factors, but each can be estimated by $|G_{ba}|\le N^\delta$ using Lemma 3.4. Since every derivative produces an extra G factor, we pay a price of N^δ, but we gain a factor $N^{-1/2}$ from $N^{-n/2}$.

Writing these estimates back to (3.45), this gives

$$\Big|\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}|S_s|^p\Big|\lesssim N^{1/2+4\delta}\Psi(\eta_0)\big[\mathbb{E}|S_s|^p+\Psi(\eta_1)^p\big]\qquad\text{w.v.h.p.}$$

which then gives (3.40). This finishes the proof of Proposition 3.5. □

3.4. **Global law, initial condition.** In this section we prove Theorem 2.2 in the global regime, i.e. where $\eta\sim 1$. For simplicity, we consider only this regime, although in the applications we will need it for a little bit smaller $\eta\ge N^{-\xi/100}$, but it is clear from the proof that given the tolerance exponents N^ξ in (2.5)–(2.6), this is affordable.

Step 1. We define a type of *renormalisation* which we will call *underline* operation. It is defined as follows: for any smooth matrix valued function $f:\mathbb{C}^{N\times N}\to\mathbb{C}^{N\times N}$ we define

$$\underline{Hf(H)}:=Hf(H)-\widetilde{\mathbb{E}}\widetilde{H}(\partial_{\widetilde{H}}f)(H),\tag{3.49}$$

where $\widetilde{H}$ denotes an independent copy of H and

$$(\partial_{\widetilde{H}}f)(H)=\lim_{\epsilon\to 0}\frac{1}{\epsilon}\big[f(H+\epsilon\widetilde{H})-f(H)\big]$$

is the directional derivative, in particular,

$$\widetilde{\mathbb{E}}\big[\widetilde{H}(\partial_{\widetilde{H}}f)(H)\big]_{xy}=\widetilde{\mathbb{E}}\sum_a\widetilde{h}_{xa}\big[(\partial_{\widetilde{H}}f)(H)\big]_{ay}=\widetilde{\mathbb{E}}\sum_a\widetilde{h}_{xa}\sum_{ij}\big[\partial_{h_{ij}}f(H)\big]_{ay}\big]\widetilde{h}_{ij}=\frac{1}{N}\sum_a\big[\partial_{h_{ax}}f(H)\big]_{ay},\tag{3.50}$$

where in the last step we used that we are in the complex Wigner case, so $\widetilde{\mathbb{E}}\widetilde{h}_{xa}\widetilde{h}_{ij}=\frac{1}{N}\delta_{xj}\delta_{ai}$.

For example, if $f(H) = (H-z)^{-1}$ is the resolvent, then $\partial_{\widetilde{H}} G = -G\widetilde{H}G$ and thus

$$\underline{HG} = HG + \widetilde{\mathbb{E}}\widetilde{H}G\widetilde{H}G = HG + \mathcal{S}[G]G = HG + \langle G\rangle G. \tag{3.51}$$

The underline is designed in such a way that

$$\mathbb{E}\underline{Hf(H)} = 0 \quad \text{if } H \text{ is Gaussian}, \tag{3.52}$$

in particular

$$\mathbb{E}\underline{HG} = \mathbb{E}\big[HG + \langle G\rangle G\big] = 0 \tag{3.53}$$

in the Gaussian case.

To see (3.52), we use cumulant expansion (3.33), and the fact that for Gaussian variables all cumulants or order three or higher vanish as we explain in Section 5:

$$\mathbb{E}(Hf(H))_{xy} = \sum_a \mathbb{E}h_{xa} f(H)_{ay} = \frac{1}{N}\sum_a \mathbb{E}[\partial_{h_{ax}} f(H)]_{ay}. \tag{3.54}$$

Note that the first cumulant is zero since $\mathbb{E}H = 0$ and only the second cumulant, i.e. $k=1$ in (3.33) matters, $\alpha_0 = (xa)$, and the only nonzero term in the sum is $\boldsymbol{\alpha} = (ax)$ (since we assumed $\mathbb{E}\chi_{\mathrm{od}}^2 = 0$ and we work in the complex case for simplicity). Combining (3.54) and (3.50) with the definition of the underline (3.49), we obtain (3.52).

Step 2. Now we derive the basic equation

$$G - m = -m\underline{HG} + m\langle G-m\rangle G \tag{3.55}$$

that expresses $G-m$ in terms of an underline. The proof follows from the resolvent identity written as $HG = zG + I$, multiplied by m and expressing $mz = -1-m^2$ from (2.3):

$$mHG = -G - m^2 G + m \quad \Longrightarrow \quad G - m = -mHG - m^2 G,$$

from which (3.55) follows by (3.51).

Taking the average trace of (3.55) and writing $G = m + G - m$, we have

$$(1-m^2)\langle G-m\rangle = -m\langle \underline{HG}\rangle + m\langle G-m\rangle^2. \tag{3.56}$$

This equation shows the basic structure: the quantity $X := \langle G-m\rangle$ to be estimated satisfies a quadratic equation plus a fluctuating underline term that is small, very roughly

$$X \approx \text{small} + X^2.$$

This will guarantee that X is small as long as $X \ll 1$. Before that, however, the left hand side of (3.56) contains a *stability factor* $1-m^2$ which may be small. This is what happens when $z \approx \pm 2$, but away from the spectral edges this factor is harmless, see (2.4). In this presentation we will stay in the bulk spectrum.

Step 3. To verify that the underline term is indeed small, we compute its high $2p$-moments. The goal is to show that

$$\mathbb{E}\big|\langle \underline{HG}\rangle\big|^{2p} \lesssim \frac{C_\epsilon}{N^{2p}} + \epsilon \mathbb{E}\big|\langle G-m\rangle\big|^{2p} \tag{3.57}$$

for any (small) ϵ with some (large) constant C_ϵ.

We use the identity (from taking the trace of (3.55))

$$\langle \underline{HG}\rangle = -\frac{1}{m}\big(1 - m\langle G\rangle\big)\langle G-m\rangle$$

to write

$$\mathbb{E}\big|\langle \underline{HG}\rangle\big|^{2p} = \frac{1}{|m|^{2p-1}}\mathbb{E}\left[\langle \underline{HG}\rangle\Big(\big(1-m\langle G\rangle\big)\langle G-m\rangle\Big)^{p-1}\overline{\Big(\big(1-m\langle G\rangle\big)\langle G-m\rangle\Big)}^{p}\right]$$

and we perform cumulant expansion in the first term.

We do $p=1$ for simplicity, the general case is similar. Ignore the harmless prefactor $1/|m|^{2p-1}$. Denote

$$g(H) := \big(1-\bar m\langle G^*\rangle\big)\langle G^* - \bar m\rangle$$

for brevity, then using (3.51) we have

$$\mathbb{E}\Big[\langle \underline{HG}\rangle\big(1-\bar m\langle G^*\rangle\big)\langle G^*-\bar m\rangle\Big] = \frac{1}{N}\mathbb{E}\sum_{ab}\Big(h_{ab}G_{ba} + \frac{1}{N}G_{aa}G_{bb}\Big)g(H)$$

$$= \frac{1}{N}\sum_{ab}\frac{1}{N}G_{ba}\partial_{ba}g(H) + \frac{1}{N}\sum_{ab}\sum_{k\ge 2}\frac{\kappa_{k+1}}{N^{(k+1)/2}k!}\partial_{ab}^k\Big[G_{ba}g(H)\Big]. \tag{3.58}$$

In the second step we used the cumulant expansion (3.33) for the function $f(H) = G_{ba}g(H)$. Note that the second cumulant term $k = 1$ is computed separately. When the derivative $\partial_{ba}\big[G_{ba}g(H)\big]$ hits G_{ba} by Leibniz rule, then it exactly cancels the renormalization term $\frac{1}{N}G_{aa}G_{bb}$. Indeed, we have $\partial_{ba}G_{ba} = -G_{bb}G_{aa}$ by the general differentiation rule of the resolvent

$$\partial_{xy}G_{ab} = -G_{ax}G_{yb},$$

and the corresponding cumulant $\kappa_{1,1} := \kappa(\chi_{\mathrm{od}}, \overline{\chi_{\mathrm{od}}}) = 1$ from $\mathbb{E}|\chi_{\mathrm{od}}|^2 = 1$ in (2.1). We also note that the last term in (3.58) is a written a bit schematically, instead of a single k we also have a summation over k_1, k_2 with $k = k_1 + k_2$, with $k_2 \geq 1$, and instead of ∂_{ab}^k we have $\partial_{ab}^{k_1}\partial_{ba}^{k_2}$. But these subtleties will not change the estimates below.

We start with the first term in (3.58). This is expected to be small since $g(H)$ depends on H only through an averaged trace, so we expect it to depend little on an individual matrix element h_{ab}. We will just compute the term where ∂_{ab} hits $\langle G^* - \bar{m}\rangle$, the other is similar. So we consider

$$\left|\frac{1}{N^2}\sum_{ab} G_{ba}\big(1 - \bar{m}\langle G^*\rangle\big)\partial_{ba}\langle G^* - \bar{m}\rangle\right| \lesssim \frac{1}{N^3}\Big|\sum_{abc} G_{ba}G^*_{cb}G^*_{ac}\Big| = \frac{1}{N^3}\Big|\mathrm{Tr}\big(G^*\big)^2 G\Big| \lesssim \frac{1}{N^2},$$

where we used that $|1 - \bar{m}\langle G^*\rangle| \lesssim 1$ since $\|G\| \lesssim 1$ as we are in the global regime, $\eta \gtrsim 1$. This estimate is exactly consistent with our target (3.57) since $p = 1$.

Now we consider the second term in (3.58). The worst is the $k = 2$ case (smallest power of $1/N$). There are several subcases, depending on how many derivatives hit G_{ba} and $g(H)$. Since $g(H)$ is an averaged object, we expect that the most critical is when all derivatives hit G_{ba}. In this case we have (bound all κ's by a constant)

$$\lesssim \frac{1}{N}\left|\sum_{ab}\frac{1}{N^{3/2}}\Big[\partial_{ba}\big(\partial_{ba} + \partial_{ab}\big)G_{ba}\Big]g(H)\right| \lesssim \frac{1}{N^{5/2}}\sum_{ab}|G_{ab}||g(H)|.$$

We used that in all possible combination of the derivatives on G_{ba}, we will get at least one off-diagonal resolvent. The others we estimated by $\|G\| \lesssim 1$. Using a Schwarz inequality, we continue

$$\frac{1}{N^{5/2}}\sum_{ab}|G_{ab}||g(H)| \leq \frac{1}{N^{5/2}}N\Big(\sum_{ab}|G_{ab}|^2\Big)^{1/2}|g(H)| = \frac{1}{N^{3/2}}\Big(\sum_{a}[GG^*]_{aa}\Big)^{1/2}|g(H)| \lesssim \frac{1}{N}|g(H)|.$$

We also compute another case, when one derivative in the last term of (3.58) hits $g(H)$, say $\partial_{ab}g(H)$. The other derivative may hit G_{ba} and notice that offdiagonality may not be guaranteed, indeed $\partial_{ba}G_{ba} = -G_{bb}G_{aa}$. So we will not gain from that, but we gain from $\partial_{ab}g(H)$. Again, differentiating $g(H)$ yields two terms, but both of them are averaged trace, effectively we have

$$\partial_{ab}\langle G^*\rangle = \frac{1}{N}\sum_{c} G^*_{ca}G^*_{bc}.$$

For example, we have a representative term of size

$$\lesssim \frac{1}{N}\sum_{ab}\frac{1}{N^{3/2}}|G_{aa}G_{bb}|\frac{1}{N}\Big|\sum_{c} G^*_{ca}G^*_{bc}\Big| = \frac{1}{N}\sum_{ab}\frac{1}{N^{5/2}}|[(G^*)^2]_{ab}| \leq \frac{1}{N^{5/2}}\Big(\sum_{a}(|G|^4)_{aa}\Big)^{1/2} \lesssim \frac{1}{N^2}.$$

Note that we did not take the absolute value inside the last summation $\sum_c$ in order to perform it and in the last step we again used a Schwarz and Ward to sum up the offdiagonal terms effectively.

In summary, we proved that

$$\mathbb{E}\big|\langle \underline{HG}\rangle\big|^2 \lesssim \frac{1}{N^2} + \frac{1}{N}\mathbb{E}|g(H)| \lesssim \frac{1}{N^2} + \frac{1}{N}\mathbb{E}|\langle G - m\rangle| \leq \frac{C_\epsilon}{N^2} + \epsilon\mathbb{E}|\langle G - m\rangle|^2,$$

(with $C_\epsilon \sim 1/\epsilon$) which proves the $p = 1$ version of (3.57). The higher p-moments are similar, this proves (3.57).

Step 4. Now we complete the proof of the global law. Taking high moment of (3.56) and recall that $1 - m^2 \sim 1$, we get

$$\mathbb{E}|\langle G - m\rangle|^{2p} \lesssim \mathbb{E}\big|\langle \underline{HG}\rangle\big|^{2p} + \mathbb{E}|\langle G - m\rangle|^{4p} \leq \frac{C_p}{N^{2p}} + \mathbb{E}|\langle G - m\rangle|^{4p}, \tag{3.59}$$

where the $\epsilon\mathbb{E}|\langle G - m\rangle|^{2p}$ from (3.57) could be absorbed in the left hand side.

Now we use a simple continuity argument in η. For large $\eta \geq 100$, we have

$$|\langle G - m\rangle| \leq \|G\| + |m| \leq \frac{2}{\eta} \leq \frac{2}{100},$$

therefore (3.59) implies

$$\mathbb{E}|\langle G-m\rangle|^{2p} \le \frac{C_p'}{N^{2p}}, \qquad \eta \ge 100. \tag{3.60}$$

Applying this for $p \to 2p$, in particular, the term $\mathbb{E}|\langle G-m\rangle|^{4p}$ is much smaller than the leading term C_p/N^{2p}. Since $|\langle G-m\rangle|$ is continuous in η, this means that η can be reduced and (3.60) remains true. The precise argument is a bit more tedious, it relies on the basic idea that there is gap in the values of $|\langle G-m\rangle|$; by Markov inequality we see that (3.59) implies that for any $\eta \gtrsim 1$

$$|\langle G-m\rangle| \le N^{-\epsilon} \quad \text{w.v.h.p.} \quad \Longrightarrow \quad |\langle G-m\rangle| \lesssim N^{-1+\xi} \quad \text{w.v.h.p.}$$

where *w.v.h.p.* stands for *with very high probability*, i.e. an event that holds with probability $1-N^{-D}$ for any D if $N \ge N_0(D)$ is large enough. In other words, $|\langle G-m\rangle|$ has a large gap in the set of its values. Since $|\langle G-m\rangle| \le N^{-\epsilon}$ for $\eta \ge 100$ and it is continuous in η, it cannot jump the big gap, so $|\langle G-m\rangle| \le N^{-\epsilon}$ remains true for any $\eta \gtrsim 1$. This proves (2.5) for $B=I$ in the global regime. The case of general B and the isotropic bound (2.6) are similar.

3.5. **Putting together the zigzag strategy. Proof of Theorem 2.2.** Recall that we have two small exponents: ϵ is the lower threshold for $\eta \ge N^{-1+\epsilon}$ and ξ is the final tolerance exponent in the local law (in the tolerance factors N^{ξ}). We always assume that $0 < \xi \ll \varepsilon$ and we will choose a *zigzag step size* δ as $0 < \delta \ll \xi$. Fix a total time

$$T := N^{-\xi/100}, \tag{3.61}$$

We will use the global law at level $\eta \sim T \sim N^{-\xi/100}$, this is where our zigzag starts.

For the terminal time T chosen as in (3.61), let K be the smallest integer such that $N^{-K\delta}T \le N^{-1+\varepsilon}$, and define a sequence of times $\{t_k\}_{k=0}^K$ as

$$t_0 := 0, \quad t_k := T - N^{-k\delta}T, \quad k \in \{1,\dots,K-1\}, \quad t_K := T. \tag{3.62}$$

Let $\{\Delta t_k\}_{k=1}^K$ denote the difference sequence of $\{t_k\}_{k=0}^K$, that is

$$\Delta t_k := t_k - t_{k-1}, \quad k \in \{1,\dots,K\}. \tag{3.63}$$

Given the target random matrix ensemble H, we construct two sequences of random matrices, $\{H_k\}_{k=0}^K$ and $\{H^k\}_{k=1}^K$ recursively by[13]

$$H_K := H, \quad H^k := \mathfrak{F}_{\mathrm{zag}}^{\Delta t_k}[H_k], \quad H_{k-1} := \left(\mathfrak{F}_{\mathrm{zig}}^{\Delta t_k}\right)^{-1}[H^k], \quad k \in \{1,\dots,K\}, \tag{3.68}$$

In the special Wigner case $\mathfrak{F}_{\mathrm{zig}}^t = \mathfrak{F}_{\mathrm{zag}}^t$, so

$$H_0 = H_1 = \ldots = H_K = H \quad \text{original matrix}$$

[13] Strictly speaking the inverse of $\mathfrak{F}_{\mathrm{zig}}^{\Delta t_k}$ does not make sense, so by the second relation in (3.64) we mean to construct a matrix H_{k-1} such that in distribution $\mathfrak{F}_{\mathrm{zig}}^{\Delta t_k}[H_{k-1}] = H^k$. More generally, when the two flows $\mathfrak{F}_{\mathrm{zig}}^t, \mathfrak{F}_{\mathrm{zag}}^t$ are not the same. Recall from Footnote 12 that $\mathfrak{F}_{\mathrm{zig}}^t$ is the standard OU flow (3.2), while $\mathfrak{F}_{\mathrm{zag}}^t$ is the solution of (3.30)), then we have

$$H_K := H, \quad H^k := \mathfrak{F}_{\mathrm{zag}}^{s(\Delta t_k)}[H_k], \quad H_{k-1} := \mathfrak{H}_{c,\Delta t_k}(H_k), \quad k \in \{1,\dots,K\}, \tag{3.64}$$

where $2c$ is a lower bound in the *fullness* assumption on the matrix distribution

$$N\,\mathbb{E}\big[|\mathrm{Tr}[(H-\mathbb{E}H)X]|^2\big] \ge 2c\,\mathrm{Tr}[X^2], \tag{3.65}$$

for any deterministic matrix X of the same symmetry class as H (real symmetric or complex Hermitian). This will guarantee that the covariance matrix is lower bounded, $\Sigma_t \ge c$, for all $t \in [0,T]$. Furthermore, $\mathfrak{H}_{c,t}$ is defined by the relation

$$\mathfrak{F}_{\mathrm{zig}}^t[\mathfrak{H}_{c,t}(H)] \overset{d}{=} \mathfrak{F}_{\mathrm{zag}}^{s(t)}[H], \quad 0 \le t \le -\log(1-c). \tag{3.66}$$

where the function $s(t) \equiv s_c(t)$ is defined as

$$s(t) \equiv s_c(t) := \log c - \log\big(c-1+\mathrm{e}^{-t}\big), \tag{3.67}$$

It is a separate (easy) lemma to show that $\mathfrak{H}_{c,t}$ exists. It follows by a simple backward inductive argument starting at $k=K$ that the covariance tensor of both H_k and H^k is given by Σ_{t_k}, hence by $\Sigma_t \ge c$, H_{k-1} is well-defined.

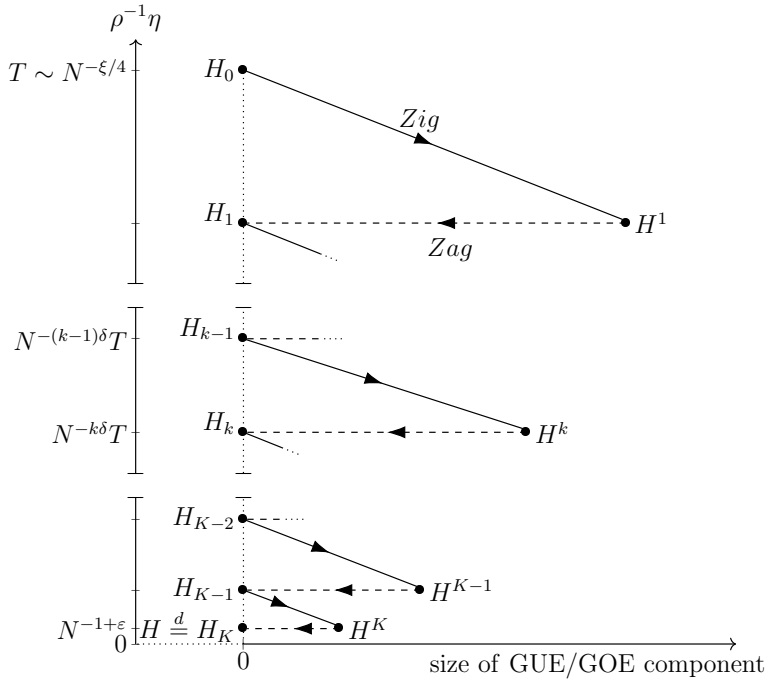

FIGURE 1. Schematic representation of the Zigzag induction. The random matrices H_k, H^k, as defined in (3.68), are situated within an abstract coordinate system. The horizontal axis represents the size of the Gaussian component, while the vertical axis indicates the lower bound on η in the domains, where we prove the local laws in Theorem 2.2. Solid arrows denote applications of Proposition 3.1 (referred to as *Zig* steps, in which we reduce η at the cost of introducing a Gaussian component), and dashed arrows indicate applications of Proposition 3.2 (*Zag* steps, in which we keep the spectral parameter fixed and remove the previously introduced Gaussian component).

Having seen the zig and zag steps, one may wonder how they work. Eventually, the zig step adds a Gaussian component, the zag step removes it, it looks a bit tautology, especially in the Wigner case, when the two OU processes happen to be the same. Around (3.6)–(3.7) we argued that without choosing z_t quite specifically, we would not have a cancellation and the G^2 term were too big. In particular, this were the case if we chose $z_t = z$ (time independent z), like we do in zag. So how could this work? The point is that both zig and zag carry a key cancellation, but they are quite different (the zag cancellation happens in the second order ($k = 1$) cumulant term in (3.44)). This is because we handle the term that is linear in H from the Ito calculus quite differently in these two procedures. In the zig case, this is the GHG term in (3.5) and we use the resolvent identity $HG = zG + I$ to remove H and aim at a self-consistent equation containing only G. In contrast, in the zag step, the analogous term is $h_{\alpha_1}(s)(\partial_{\alpha_1}|S_s|^p)$ in (3.43) and here we do a cumulant expansion. As we mentioned earlier, the zig step is more novel and in some sense more essential, since this is where the key reduction of $\Im z$ happens. In particular if we applied the theory for purely Gaussian ensembles, then the zag step is trivial (no need to perform it), but the zig step is still nontrivial.

4. MULTI-RESOLVENT LOCAL LAWS WITH ZIGZAG

The zigzag strategy is especially suitable for proving multi-resolvent local laws (1.13)–(1.14) for the chain (1.9). The basic idea is the same, but there are several additional complications.

Similarly to the material presented in Section 2 for the single resolvent case, the multi-resolvent local law for the simplest Wigner case in the bulk has never been written up in the literature. The material I present here is a simplified version of the paper [23], where Wigner matrices with different deformations were considered. Taking the trivial $D_1 = D_2 = 0$ deformations on that paper, we arrive the pure Wigner case.

First, we need a good understanding of the corresponding $M_{[1,k]}$ term, in particular we need (1.12). This is not trivial despite the explicit recursive formula for $M_{[1,k]}$ because that formula contains many cancellations. The main point is that the recursion is nonlinear and it contains repeated inverses of the *stability operator* which is

$$\mathcal{B}_{12} = I - M_1 \mathcal{S}[\cdot] M_2, \tag{4.1}$$

for a general correlated model. For deformed Wigner matrices, it simplifies to $1 - M_1\langle\cdot\rangle M_2$, see (1.10), and even simpler, for Wigner matrices it is just $1 - m_1 m_2\langle\cdot\rangle$, see (1.11). The inverses of these operators are all bounded by $1/\eta$ but if one just counts the number of inverses of $\mathcal{B}_{12}$ in the recursive formula, then one obtains about twice as many $1/\eta$ factor than the real size (1.11), indicating a lot of cancellation. This can be done fully algebraically in case of Wigner matrices [15] for any k. For more complicated ensembles, it was done by ad hoc methods. A new dynamical method that works even for random band matrices was developed in [33].

Second, we need a global law whose proof is similar to the one presented for the single G case, with many more terms to consider.

Third, we need to develop an entire *hierarchy of master inequalities* that control the fluctuations of longer and longer chains. In short, even if we are interested in, say $k = 2$-chains, we need to control longer chains as well, but maybe not with the optimal precision. The reason is in the structure of the zig-equation, see (3.10) for $k = 1$. The RHS of this equation involves longer chains, partly in the linear term and more importantly in the quadratic variation of the martingale term (3.14), which involves four resolvents. In this simple case we used Ward identities and trivial norm bounds $|G|^2 \leq 1/\eta^2$ (see (3.15)) to reduce these longer chains back to the original single resolvent object, but in many cases this is too crude. For example, if we are interested in $\langle GAGA \rangle$ with traceless observable A and we want to gain the $\sqrt{\eta}$ factor from each A, then Ward identity or norm bounds destroy this gain. Below we sketch the setup of the dynamical proof of the $k = 2$-chain (for general observables) in case of deformed Wigner matrices[14], following Section 5 of [23].

Let R_1, R_2 be deterministic observables and $G_j := G_j(z_j(t))$, $j = 1, 2$, two time dependent resolvents with moving spectral parameters along the characteristic flow. We want to control the evolution of

$$X = X_t^{R_1,R_2} := \Big\langle \big[G_1(z_1(t))R_1G_2(z_2(t)) - M_{12,t}^{R_1}\big]R_2 \Big\rangle$$

for a fixed choice $R_1 := A_1$ and $R_2 = A_2$, where A_1, A_2 are given observables in the original chain (1.9). We will need to consider not only $X_t^{A_1,A_2}$, but also $X_t^{A_1,I}$, X_t^{I,A_2}, and $X_t^{I,I}$, this is why we introduced the more general letters R_j, but eventually we will always consider $R_1, R_2 \in \{I, A_1, A_2, A_1^*, A_2^*\}$.

The analogue of (3.10) for the corresponding time dependent two chain

$$X = X_t^{R_1,R_2} := \Big\langle \big[G_1(z_1(t))R_1G_2(z_2(t)) - M_{12,t}^{R_1}\big]R_2 \Big\rangle$$

is the following

$$\mathrm{d}X_t^{R_1,R_2} = \big(1 + (2 - k(R_1,R_2))\langle M_{12,t}^I\rangle\big)\, g_t^{R_1,R_2}\mathrm{d}t + \mathrm{d}\mathcal{E}_t + \mathcal{F}_t\mathrm{d}t\,. \tag{4.2}$$

Here we denoted

$$k(R_1,\dots,R_m) := \#\{j \in [1,m] : R_j \neq I\} \tag{4.3}$$

for deterministic $R_1, \dots, R_m \in \mathbb{C}^{N\times N}$.

The *martingale term* is given by

$$\mathrm{d}\mathcal{E}_t = \frac{1}{\sqrt{N}} \sum_{a,b=1}^{N} \partial_{ab}\langle G_{1,t}R_1G_{2,t}R_2\rangle \mathrm{d}B_{ab}. \tag{4.4}$$

The *forcing term* is decomposed into $\mathcal{F}_t = \mathrm{Lin}_t + \mathrm{Err}_t$, where the *linear term* and *error term* are given by

$$\begin{aligned} \mathrm{Lin}_t &= k(R_1)\langle M_{12,t}^{R_1}\rangle X_t^{I,R_2} + k(R_2)\langle M_{21,t}^{R_2}\rangle X_t^{R_1,I}, \\ \mathrm{Err}_t &= X_t^{I,R_2}X_t^{R_1,I} + \langle G_{1,t} - M_{1,t}\rangle\langle G_{1,t}^2R_1G_{2,t}R_2\rangle + \langle G_{2,t} - M_{2,t}\rangle\langle G_{1,t}R_1G_{2,t}^2R_2\rangle\,, \end{aligned} \tag{4.5}$$

respectively.

We define a stopping time, analogous to (3.19). Take any small ξ_0, ξ_1, ξ_2 such that $\xi_0 < \xi_1/2 < \xi_2/4$ and define the stopping time

$$\begin{aligned} \tau^{R_1,R_2} &:= \inf\{s \in [0,T] : \max_{z_{j,0}\in\Omega_0^j} \alpha_s^{-1}\,|g_s^{R_1,R_2}| \geq N^{2\xi_{k(R_1,R_2)}}\}, \\ \tau &:= \min\{\tau^{R_1,R_2} :\, R_1, R_2 \in \mathfrak{S}\}\,, \quad \text{with} \quad \mathfrak{S} := \{I, B_1, B_1^*, B_2, B_2^*\}\,, \end{aligned} \tag{4.6}$$

where we introduced the shorthand notation

$$\alpha_t := \frac{1}{N\eta_{1,t}\eta_{2,t}}$$

for the target bound, and Ω_0^j is a bulk spectral domain, essentially $\mathcal{D}_{\gamma=1}$ from (3.16). Note that here we will prove a slightly stronger bound than (1.13) since the control parameter α follows the two η's separately, instead of estimating everything in terms of $\eta_t := \min\{\eta_{1,t}, \eta_{2,t}\}$.

Now we control various terms in the RHS of the equation (4.2) in terms of X_s, $s \leq \tau$, controlled by the stopping time and in terms of single resolvent averaged local laws like $\langle G - M\rangle$ for which we already have optimal estimate of order $1/N\eta$.

[14] It might look pedagogically more advisable to consider an even simpler case, the pure Wigner matrices, however then M is just a scalar and some structure is not visible.

There are two problematic terms. First, the term $\langle G_{1,t}^2 R_1 G_{2,t} R_2\rangle$ in Err_t has three resolvents, so apparently does not fit the scheme. But it is a special three resolvent chain, as it contains G^2. So we can use an integral representation to reduce it to a single G-object, more generally

$$G_1(z_1)G_2(z_2) = \frac{1}{2\pi i}\oint \frac{G(\zeta)}{(\zeta - z_1)(\zeta - z_2)}\mathrm{d}\zeta$$

This approach has the disadvantage that the spectral domain needs to be changed. Alternatively, we could do a Schwarz as follows $\langle G_1^2 R_1 G_2 R_2\rangle$ via a Cauchy-Schwarz inequality followed by a Ward identity:

$$|\langle G_1^2 R_1 G_2 R_2\rangle| \le \frac{\langle \Im G_1\rangle^{1/2}\langle \Im G_1 R_1 G_2 R_2 R_2^* G_2^* R_1^*\rangle^{1/2}}{\eta_1} \le \frac{\langle \Im G_1\rangle \|R_1 G_2 R_2 R_2^* G_2^* R_1^*\|^{1/2}}{\eta_1} \le \frac{1}{\eta_1\eta_2}, \tag{4.7}$$

where in the middle step we also used $\langle AB\rangle \le \langle A\rangle\|B\|$ for any positive matrix $A \ge 0$.

The other problematic term is the quadratic variation of (4.4)

$$\mathrm{QV}\big[g_t^{R_1,R_2}\big] := \frac{1}{N}\sum_{a,b=1}^{N} |\partial_{ab}\langle G_{1,t} R_1 G_{2,t} R_2\rangle|^2 . \tag{4.8}$$

After computing the derivatives, we get a chain with six resolvents, but two Ward identities immediately reduce it to four resolvents of the form

$$\langle \Im G_1 R_1 G_2 R_2 \Im G_1 R_2^* G_2^* R_1^*\rangle \le N\langle \Im G_1 R_2^* | G_2 | R_2\rangle\langle \Im G_1 R_1 | G_2 | R_1^*\rangle. \tag{4.9}$$

The inequality can be obtained by spectral decomposition and a smart Schwarz inequality (see (5.27) of [23]). This (4.9) is one version of various *reduction inequalities*[15] that estimate longer chain in terms of shorter ones, usually with a loss (here it is the additional N factor instead of a $1/\eta$ factor; effectively reduction inequality loses a big factor $N\eta$). This is the price to truncate the hierarchy, i.e. to stop the cascade that the natural equation of a chain of length k involves (in the quadratic variation) a chain of length $2k$ etc. This phenomenon is very well known in many-body theory: we have a structure analogous to the BBGKY hierarchy. To turn this, potentially infinite, hierarchy into a rigorous mathematical tool, we need to truncate it. This is what the reduction inequalities are doing. It turns out that the price $N\eta$ is affordable and plugging all these estimates (up to $s \le \tau$) into (4.2), we see that after integrating back by Gronwall's argument, the RHS gives a better control than the threshold in the stopping time, hence $\tau = T_{max}$.

It is remarkable that the most critical linear propagators, $\langle M_{12,t}^{R_1}\rangle$ for $R_1 = I$ satisfy the consistent upper bound (Lemma 5.2 of [23]), namely

$$\int_s^t 2\Re\langle M_{12,t}^I\rangle \mathrm{d}r \le \log\frac{\eta_{1,s}\eta_{2,s}}{\eta_{1,t}\eta_{2,t}} \tag{4.10}$$

in complete analogy with (3.26). After exponentiating, this propagates the bound $1/(N\eta_1\eta_2)$ consistently from s to t. This is a very general structural property of the zigzag proof; the paper [23] demonstrated it for deformed Wigner case, but it holds for more general mean field models as well.

The proof for regular observables is similar.

Fourth, we again need a zag step, which again has many more terms to handle, but from one point of view it is conceptually easier for longer chains: the single G bounds are already available, so they do not need to be imported from a larger η-level using the monotonicity lemma, Lemma 3.4.

5. Auxiliary results

5.1. **Cumulants and cumulant expansion formula.** Let h be a real random variable, suppose all its moments $m_k := \mathbb{E}h^k$ have a control of the form

$$\mathbb{E}|h|^k \le C^k k!, \quad \forall k \in \mathbf{N}, \tag{5.1}$$

for some constant C. In particular, the *moment generating function* and its Taylor series converges for small enough $|t| < 1/C$:

$$\mathbb{E}\big[e^{th}\big] = \sum_{k=0}^{\infty}\frac{t^k}{k!}m_k. \tag{5.2}$$

[15] Also note that $|G|$ appears instead of G or G^*, but there is a standard integral formula to express $|G(E+i\eta)|$ in terms of integrating $\Im G(E+is)$ over $s \ge \eta$.

The cumulants κ_k of h are certain functions of the moments that can be defined, for example, by their *cumulant generating function*

$$\chi(t) = \chi_h(t) := \log \mathbb{E}\big[e^{th}\big] =: \sum_{k=0}^{\infty} \frac{t^k}{k!}\kappa_k \tag{5.3}$$

at least for small enough t (it is easy to see that for small enough t we have $\mathbb{E}\big[e^{th}\big] \approx 1$, so we can take its logarithm). Here are the first few relations and their inverse relations

$$\begin{aligned}
m_1 =& \kappa_1 \\
m_2 =& \kappa_2 + \kappa_1^2 \\
m_3 =& \kappa_3 + 3\kappa_2\kappa_1 + \kappa_1^3 \\
m_4 =& \kappa_4 + 4\kappa_3\kappa_1 + 3\kappa_2^2 + 6\kappa_2\kappa_1^2 + \kappa_1^4 \\
&\text{... etc} \\
\kappa_1 =& m_1 \\
\kappa_2 =& m_2 - m_1^2 \\
\kappa_3 =& m_3 - 3m_2m_1 + 2m_1^3 \\
\kappa_4 =& m_4 - 4m_3m_1 - 3m_2^2 + 12m_2m_1^2 - 6m_1^4 \\
&\text{... etc.}
\end{aligned} \tag{5.4}$$

The triangular structure guarantees that the cumulants up the order k can be obtained from the moments up to order k and vice versa. Note that the second cumulant κ_2 is the variance.

Among several other reasons, the main significance of the cumulants is that they give a very simple characterisation of the Gaussian random variables, indeed

$$h \text{ is Gaussian} \quad \Longrightarrow \quad \kappa_k = 0 \text{ for } k \geq 3,$$

i.e. that all higher order cumulants vanish. This can be directly seen from the moment generating function of the standard Gaussian h

$$\mathbb{E}e^{th} = e^{t^2/2}$$

and their trivial transformations.

Now we state the *cumulant expansion formula*: for any smooth function f (with some growth control at infinity that makes the sums below convergent), we have

$$\mathbb{E}hf(h) = \sum_{k=0}^{\infty} \frac{\kappa_{k+1}}{k!}\mathbb{E}f^{(k)}(h). \tag{5.5}$$

For the proof, we start with an identity; we claim that

$$m_n = \sum_{k=1}^{n} \binom{n-1}{k-1} \kappa_k m_{n-k}. \tag{5.6}$$

To see this, we note that from (5.2) and (5.3)

$$m_n = \partial_t^n e^{\chi(t)}\Big|_{t=0}, \qquad \kappa_n = \partial_t^n \chi(t)\Big|_{t=0}.$$

On the other hand, by Leibniz rule

$$\partial_t^n e^{\chi(t)} = \partial_t^{n-1}\Big(\chi'(t)e^{\chi(t)}\Big) = \sum_{k=1}^{n} \binom{n-1}{k-1} \big(\partial_t^k \chi(t)\big)\big(\partial_t^{n-k} e^{\chi(t)}\big),$$

and evaluating this identity at $t = 0$, we get (5.6).

Now we prove (5.5). We check it for monomials, $f(h) = h^q$, then it follows for polynomials and then by density argument for any smooth function (as stated in Proposition 3.3). Indeed, using (5.6) we have

$$\mathbb{E}hf(h) = \mathbb{E}h^{q+1} = \sum_{k=1}^{q+1} \binom{q}{k-1} \kappa_k \mathbb{E}h^{q+1-k} = \sum_{k=0}^{q} \binom{q}{k} \kappa_{k+1}\mathbb{E}h^{q-k} = \sum_{k=0}^{q} \frac{\kappa_{k+1}}{k!}\mathbb{E}f^{(k)}(h),$$

since $f^{(k)}(h) = q(q-1)\dots(q-k+1)h^{q-k}$.

We stated everything for one single random variable, but all these concepts directly extend to a family $\boldsymbol{h} = (h_1, h_2, \dots h_p)$ random variables, we can talk about their joint moments and joint cumulants. These are labelled with $\boldsymbol{k} = (k_1, k_2, \dots k_p)$, i.e. by p-tuples of natural numbers. The only difference is that the corresponding generating functions will be functions of p variables $\boldsymbol{t} = (t_1, t_2, \dots, t_p)$ as

$$\mathbb{E}\big[e^{\boldsymbol{t}\cdot\boldsymbol{h}}\big] = \sum_{k=0}^{\infty} \frac{\boldsymbol{t}^{\boldsymbol{k}}}{\boldsymbol{k}!} m_{\boldsymbol{k}}, \qquad \chi(\boldsymbol{t}) := \log \mathbb{E}\big[e^{\boldsymbol{t}\cdot\boldsymbol{h}}\big] =: \sum_{\boldsymbol{k}=0}^{\infty} \frac{\boldsymbol{t}^{\boldsymbol{k}}}{\boldsymbol{k}!} \kappa_{\boldsymbol{k}},$$

where $\boldsymbol{k}! = k_1!k_2! \dots k_p!$.

REFERENCES

[1] A. Adhikari and J. Huang. Dyson Brownian motion for general β and potential at the edge. In: *Probability Theory and Related Fields* 178 (2020), pp. 893–950. DOI: 10.1007/s00440-020-00992-9.

[2] A. Adhikari and B. Landon. Local law and rigidity for unitary Brownian motion. In: *Probab. Theory Relat. Fields.* 187 (2023), pp. 753–815. DOI: 10.1007/s00440-023-01230-8.

[3] A. Aggarwal and J. Huang. Edge rigidity of Dyson Brownian motion with general initial data. In: *Electronic Journal of Probability* 29 (2024), pp. 1–62. DOI: 10.1214/24-EJP1178.

[4] M. Aizenman and S. Molchanov. Localization at large disorder and at extreme energies: An elementary derivations. In: *Communications in Mathematical Physics* 157 (1993), pp. 245–278.

[5] Z. Bao, G. Cipolloni, L. Erdős, J. Henheik, and O. Kolupaiev. Decorrelation transition in the Wigner minor process. In: *arXiv: 2503.06549* (2025).

[6] Z. Bao, G. Cipolloni, L. Erdős, J. Henheik, and O. Kolupaiev. Law of fractional logarithm for random matrices. In: *arxiv:2503.18922* (2025).

[7] Z. Bao and L. Erdős. Delocalization for a class of random block band matrices. In: *Probability Theory and Related Fields* 167 (2017), pp. 673–776. DOI: 10.1007/s00440-015-0692-y.

[8] J. Baslingker, R. Basu, S. Bhattacharjee, and M. Krishnapur. The Paquette-Zeitouni law of fractional logarithms for the GUE minor process. In: *preprint, arXiv: 2410.11836* (2024).

[9] P. Bourgade. Extreme gaps between eigenvalues of Wigner matrices. In: *J. Eur. Math. Soc.* 24 (2021), pp. 2823–2873. DOI: 10.4171/JEMS/1141.

[10] P. Bourgade, F. Yang, H.-T. Yau, and J. Yin. Random band matrices in the delocalized phase, II: Generalized resolvent estimates. In: *Journal of Statistical Physics* 174 (2019), pp. 1189–1221. DOI: 10.1007/s10955-019-02229-z.

[11] P. Bourgade, H.-T. Yau, and J. Yin. Random Band Matrices in the Delocalized Phase I: Quantum Unique Ergodicity and Universality. In: *Communications on Pure and Applied Mathematics* 73 (2020), pp. 1526–1596. DOI: 10.1002/cpa.21895.

[12] N. Chen and C. K. Smart. Random band matrix localization by scalar fluctuations. 2022. arXiv:2206.06439.

[13] G. Cipolloni, L. Erdős, and D. Schröder. Eigenstate Thermalization Hypothesis for Wigner matrices. In: *Commun. Math. Phys.* 388 (2021), pp. 1005–1048. DOI: 10.1007/s00220-021-04239-z.

[14] G. Cipolloni, L. Erdős, and D. Schröder. Optimal multi-resolvent local laws for Wigner matrices. In: *Electron. J. Probab.* 27 (2022), pp. 1–38. DOI: 10.1214/22-EJP838.

[15] G. Cipolloni, L. Erdős, and D. Schröder. Thermalisation for Wigner matrices. In: *J. Funct. Anal.* 282 (2022), p. 109394. DOI: 10.1016/j.jfa.2022.109394.

[16] G. Cipolloni, L. Erdős, J. Henheik, and O. Kolupaiev. Eigenvector decorrelation for random matrices. In: *arxiv:2410.10718* (2024).

[17] G. Cipolloni, L. Erdős, and D. Schröder. Central limit theorem for linear eigenvalue statistics of non-Hermitian random matrices. In: *Comm. Pure Appl. Math.* 76 (2023), pp. 899–1136.

[18] G. Cipolloni, L. Erdős, D. Schröder, and Y. Xu. On the rightmost eigenvalue of non-Hermitian random matrices. In: *Ann. Probab.* 51 (2023), pp. 2192–2242.

[19] G. Cipolloni, L. Erdős, and Y. Xu. Optimal decay of eigenvector overlap for non-Hermitian random matrices. In: *arxiv:2411.16572* (2024).

[20] G. Cipolloni, L. Erdős, J. Henheik, and D. Schröder. Optimal lower bound on eigenvector overlaps for non-Hermitian random matrices. In: *Journal of Functional Analysis* (2024). DOI: 10.1016/j.jfa.2024.110495.

[21] G. Cipolloni, L. Erdős, and J. Henheik. Eigenstate thermalisation at the edge for Wigner matrices. 2023. arXiv:2309.05488.

[22] G. Cipolloni, L. Erdős, and J. Henheik. Out-of-time-ordered correlators for Wigner matrices. In: *Advances in Theoretical and Mathematical Physics* 28 (2024), pp. 2025–2083. DOI: 10.4310/ATMP.241031013250.

[23] G. Cipolloni, L. Erdős, J. Henheik, and O. Kolupaiev. Eigenvector decorrelation for random matrices. 2024. arXiv:2410.10718.

[24] G. Cipolloni, L. Erdős, and D. Schröder. Mesoscopic central limit theorem for non-Hermitian random matrices. In: *Probab. Theory Relat. Fields.* 188 (2024), pp. 1131–1182. DOI: 10.1007/s00440-023-01229-1.

[25] G. Cipolloni, L. Erdős, and D. Schröder. Mesoscopic central limit theorem for non-Hermitian random matrices. In: *Probability Theory and Related Fields* 188 (2024), pp. 1131–1182. DOI: 10.1007/s00440-023-01229-1.

[26] G. Cipolloni, L. Erdős, and Y. Xu. Universality of extremal eigenvalues of large random matrices. In: (2023). arXiv:2312.08325.

[27] G. Cipolloni, R. Peled, J. Schenker, and J. Shapiro. Dynamical Localization for Random Band Matrices Up to $W \ll N^{1/4}$. In: *Communications in Mathematical Physics* 405 (2024), p. 82. DOI: 10.1007/s00220-024-04948-1.

[28] S. Dubova and K. Yang. Quantum diffusion and delocalization in one-dimensional band matrices via the flow method. 2024. arXiv:2412.15207.

[29] L. Erdős, T. Krüger, and D. Schröder. Random matrices with slow correlation decay. In: *Forum Math. Sigma* 7 (2019), E8. DOI: 10.1017/fms.2019.2.

[30] L. Erdős, J. Henheik, and V. Riabov. Cusp universality for correlated random matrices. 2024. arXiv:2410.06813.

[31] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Delocalization and diffusion profile for random band matrices. In: *Communications in Mathematical Physics* 323 (2013), pp. 367–416. DOI: 10.1007/s00220-013-1773-3.

[32] L. Erdős and V. Riabov. Eigenstate thermalization hypothesis for Wigner-type matrices. In: *Communications in Mathematical Physics* 405 (2024), p. 282. DOI: 10.1007/s00220-024-05143-y.

[33] L. Erdős and V. Riabov. The zigzag strategy for random band matrices. In: *arvix:2506.06441* (2025).

[34] L. Erdős and H.-T. Yau. A dynamical approach to random matrix theory. Vol. 28. American Mathematical Soc., 2017. DOI: 10.1090/cln/028.

[35] P. J. Forrester and T. Nagao. Determinantal correlations for classical projection processes. In: *J. Stat. Mech.: Theory and Experiment* 2011 (2011), P08011.

[36] J. Fröhlich and T. Spencer. Absence of diffusion in the Anderson tight binding model for large disorder or low energy. In: *Communications in Mathematical Physics* 88 (1983), pp. 151–184. DOI: 10.1007/BF01209475.

[37] M. Goldstein. Fluctuations and localization length for random band GOE matrix. 2022. arXiv:2210.04346.

[38] Y. He and A. Knowles. Mesoscopic eigenvalue statistics of Wigner matrices. In: *Ann. Appl. Probab.* 27 (2017), pp. 1510–1550. DOI: 10.1214/16-AAP1237.

[39] J. Huang and B. Landon. Rigidity and a mesoscopic central limit theorem for Dyson Brownian motion for general β and potentials. In: *Probab. Theory Relat. Fields.* 175 (2019), pp. 209–253. DOI: 10.1007/s00440-018-0889-y.

[40] B. Landon, P. Lopatto, and P. Sosoe. Single eigenvalue fluctuations of general Wigner-type matrices. In: *Probab. Theory Relat. Fields.* 188 (2024), pp. 1–62. DOI: 10.1007/s00440-022-01181-6.

[41] B. Landon and P. Sosoe. Almost-optimal bulk regularity conditions in the CLT for Wigner matrices. 2022. arXiv:2204.03419.

[42] J. O. Lee and K. Schnelli. Edge universality for deformed Wigner matrices. In: *Rev. Math. Phys.* 27 (2015), p. 1550018. DOI: 10.1142/S0129055X1550018X.

[43] R May. Will a Large Complex System be Stable? In: *Nature* 238 (1972), pp. 413–414.

[44] E. Paquette and O. Zeitouni. Extremal eigenvalue correlations in the GUE minor process and a law of fractional logarithm. In: *Ann. Probab.* 45 (2017), pp. 4112–4166. ISSN: 0091-1798,2168-894X. DOI: 10.1214/16-AOP1161.

[45] L. A. Pastur. On the spectrum of random matrices. In: *Teoreticheskaya i Matematicheskaya Fizika* 10 (1972), pp. 102–112. DOI: 10.1007/BF01035768.

[46] R. Peled, J. Schenker, M. Shamis, and S. Sodin. On the Wegner orbital model. In: *International Mathematics Research Notices* (2019), pp. 1030–1058. DOI: 10.1093/imrn/rnx145.

[47] J. Schenker. Eigenvector localization for random band matrices with power law band width. In: *Communications in Mathematical Physics* 290 (2009), pp. 1065–1097. DOI: 10.1007/s00220-009-0798-0.

[48] A. I. Snirelman. Ergodic properties of eigenfunctions. In: *Uspehi Mat. Nauk* 29 (1974), pp. 181–182. ISSN: 0042-1316.

[49] P. von Soosten and S. Warzel. Random characteristics for Wigner matrices. In: *Electronic Communications in Probability* 24 (2019), pp. 1 –12. DOI: 10.1214/19-ECP278.

[50] T. Tao and V. Vu. Random matrices: Universality of local eigenvalue statistics. In: *Acta Math.* 206 (2011), pp. 127–204. DOI: 10.1007/s11511-011-0061-3.

[51] F. Yang and J. Yin. Random band matrices in the delocalized phase, III: averaging fluctuations. In: *Probability Theory and Related Fields* 179 (2021), pp. 451–540. DOI: 10.1007/s00440-020-01013-5.

[52] H.-T. Yau and J. Yin. Delocalization of One-Dimensional Random Band Matrices. 2025. arXiv:2501.01718.