ON THE RIGHTMOST EIGENVALUE OF NON-HERMITIAN RANDOM MATRICES

BY GIORGIO CIPOLLONI[1,a], LÁSZLÓ ERDŐS[2,b],
DOMINIK SCHRÖDER[3,d] AND YUANYUAN XU[2,c]

[1]*Princeton Center for Theoretical Science, Princeton University ,* [a]*gc4233@princeton.edu*

[2]*Institute of Science and Technology Austria ,* [b]*laszlo.erdoes@ist.ac.at;* [c]*yuanyuan.xu@ist.ac.at*

[3]*Institute for Theoretical Studies, ETH Zurich ,* [d]*dschroeder@ethz.ch*

We establish a precise three-term asymptotic expansion, with an optimal estimate of the error term, for the rightmost eigenvalue of an $n \times n$ random matrix with independent identically distributed complex entries as n tends to infinity. All terms in the expansion are universal.

1. Introduction. Large random matrices are frequently used to model complex systems of many degrees of freedom. Quantum Hamiltonians are naturally self-adjoint, so their conventional random matrix models are Hermitian; the most common example is Wigner matrices. Beyond quantum mechanics, random matrices often appear without any symmetry condition in natural phenomenological models. For example, the time evolution of many interacting agents $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)$ may be described by a linear system of differential equations of the form

$$(1.1) \qquad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}(t) = X\boldsymbol{u}(t).$$

Lacking any specific knowledge about how u_j precisely influences the evolution of u_i, the simplest phenomenological model assumes that the coefficient matrix X is random. Despite its simplicity, since the ground-breaking paper of May [57], this model has been extensively used to describe the evolution of complex systems both in theoretical neuroscience, see e.g. [61, 65] and in mathematical ecology, e.g. [3, 4], see also a recent comprehensive review [5]. The problem is often presented in the form [65]

$$(1.2) \qquad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}(t) = (-I + gX)\boldsymbol{u}(t),$$

where the identity matrix stands for a natural exponential decay at unit rate and the coupling constant $g > 0$ explicitly expresses the strength of the random couplings. The main question is to tune g so that the system is stable in the sense that it is neither exponentially decaying nor exponentially increasing. The maximal growth rate of the solution of (1.2) is determined by the maximal real part of the spectrum of the coefficient matrix $-I + gX$. This motivates the main task of this paper: to understand very accurately the real part of the rightmost eigenvalue of a large non-Hermitian random matrix.

We remark that a similar optimal stability question of (1.2) for uniformly random initial data $\boldsymbol{u}(0)$ and when the size of $\boldsymbol{u}(t)$ measured in ℓ^2-sense has been answered in [20, 58] when the coefficients x_{ij} are all Gaussian and in [31, 32] for more general distributions even beyond the i.i.d. case. The rightmost eigenvalue studied in the current paper is relevant when we consider the worst-case scenario, i.e. when we measure $\boldsymbol{u}(t)$ in maximum norm and we take the supremum over all initial data $\boldsymbol{u}(0)$ (see Corollary 2.4 below).

To be more specific, we consider $n \times n$ random matrices X with independent, identically distributed (i.i.d.) matrix elements, called the *i.i.d. matrix ensemble*. This is the non-Hermitian counterpart of the Wigner ensemble. We choose the normalization such that $x_{ij} \overset{\mathrm{d}}{=} n^{-1/2}\chi$, for all i, j, where χ is a fixed complex centred random variable with $\boldsymbol{E}\,|\chi|^2 = 1$ and $\boldsymbol{E}\,\chi^2 = 0$. This normalization guarantees that the spectrum of X remains essentially within the unit disk, uniformly in n. We claim our result and present the proof only for the complex case but our basic method works for the real case as well. We will comment on the necessary modifications that we do not carry out in this paper for brevity.

More precisely, the celebrated *Girko's circular law*, proven in increasing generality in [10, 43, 66], asserts that the eigenvalue density of X is uniform on the unit disk of the complex plane. Furthermore, there are no outlier eigenvalues far away since the spectral radius $\rho(X)$ converges to 1, see [11, 15, 16, 41]. A speed of convergence of order $n^{-1/2+\epsilon}$, for any small $\epsilon > 0$, with very high probability was recently established in [6]. Nevertheless some extremal eigenvalues do lie outside of the unit disk, hence $\max \operatorname{Re}\operatorname{Spec}(X)$ is slightly larger than one. It is well known that eigenvalues genuinely fluctuate on scale $n^{-1/2}$ near the boundary of the unit disk, in fact the local eigenvalue statistics in this regime is universal [23]. Therefore we know that

$$n^{-1/2} \ll \max \operatorname{Re}\operatorname{Spec}(X) - 1 \ll n^{-1/2+\epsilon}$$

holds for any $\epsilon > 0$ with very high probability and our goal is to find a more accurate asymptotics. We remark that this natural question was posed in the first version on the arXiv of [16] in Section 1.1.8. Beforehand, a leading order large deviation principle was established for $\max \operatorname{Re}\operatorname{Spec}(X)$ even for the more general elliptic ensemble in [12] and the refined asymptotics was mentioned as an open question.

For the complex Gaussian case, i.e., when χ is a standard complex random variable (*Ginibre ensemble*), the eigenvalues form a determinantal process with an explicit correlation kernel computed first by Ginibre [42]. Based upon these formulas in our companion paper [28] we recently gave a new short proof of the exact asymptotics:
(1.3)
$$\max \operatorname{Re}\operatorname{Spec}(X) \overset{\mathrm{d}}{=} 1 + \sqrt{\frac{\gamma_n}{4n}} + \frac{1}{\sqrt{4n\gamma_n}}\mathfrak{G}_n, \qquad \gamma_n := \frac{\log n - 5\log\log n - \log(2\pi^4)}{2},$$

where the random variable $\mathfrak{G}_n$ converges to a Gumbel distribution

$$\lim_{n\to\infty} \boldsymbol{P}(\mathfrak{G}_n \leq t) = \exp\left(-e^{-t}\right)$$

for any fixed $t \in \mathbb{R}$ with an effective error term. We also proved a similar result for the real Ginibre ensemble. These results without error term were first proven by Bender [14] in the complex case and by Akemann and Phillips [2] in the real case even for the more involved Gaussian elliptic ensembles where a sophisticated saddle point analysis for the correlation kernel was necessary, while our proof in [28] is elementary. In particular, we obtained that typically

$$(1.4) \qquad \max \operatorname{Re}\operatorname{Spec}(X) = 1 + \sqrt{\frac{\log n - 5\log\log n - \log(2\pi^4)}{8n}} + O\left(\frac{1}{\sqrt{n\log n}}\right)$$

for the Ginibre ensemble. In this paper we prove (1.4) for any i.i.d. matrix ensemble, in particular we show that the three-term asymptotics is universal in the sense that it is independent of the single entry distribution χ. The Gumbel fluctuation is also expected to be universal; this is an open problem that we leave to future work. However, our result (1.4) already implies the tightness of $\mathfrak{G}_n$ in (1.3) even for the i.i.d. case, see Remark 2.5 below.
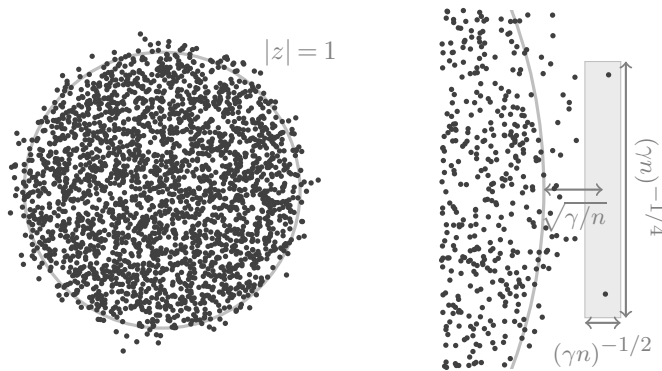
FIG 1. *The figure shows the eigenvalues of complex Ginibre matrices. The eigenvalues for the left figures have been computed for* 50 *independent Ginibre matrices of size* 50×50, *while for the right figure* 100 *independent matrices of size* 100×100 *have been sampled. The gray box on the right right hand side indicates the high-probability location of the eigenvalue with the largest real part.*

Extreme value statistics for independent random variables are fully described by the Fisher-Tippett-Gnedenko theorem, with the Gumbel distribution being one of the three main universal laws. The three-term asymptotics for the scaling factor is also common, for example the maximum of n standard Gaussian variables is given by

$$\sqrt{2\log n} - \frac{\log\log n + \log(4\pi)}{2\sqrt{2\log n}} + \frac{1}{\sqrt{2\log n}}\mathfrak{G}_n,$$

where $\mathfrak{G}_n$ is again asymptotically Gumbel. While such precise asymptotics in the independent case is a fairly simple exercise by the tail asymptotics of the individual random variables, it is remarkable that such precision can be maintained in certain weakly correlated situations[1] such as $\max\operatorname{Re}\operatorname{Spec}(X)$. The intuition is that effectively only few rightmost eigenvalues compete for $\max\operatorname{Re}\operatorname{Spec}(X)$ and these eigenvalues are typically far away from each other, well beyond their correlation length of order $n^{-1/2}$, hence they asymptotically form a Poisson point process and are essentially independent. While this scenario could be directly verified for the Ginibre ensemble [28] (and even for the more general Gaussian elliptic ensemble [2, 14], as well as for the chiral two-matrix model with complex entries [1]), its validity for a general i.i.d. matrix is a highly nontrivial fact since explicit formulas for the eigenvalue correlation functions are lacking.

We remark that starting with the ground-breaking paper of Fyodorov, Hiary and Keating [38], see also [39], similar three-term asymptotics have recently been investigated for extreme values of characteristic polynomials of various random matrix ensembles, e.g. [40, 7, 60, 21, 51], as well as for the Riemann ζ-function, e.g. [62, 8, 59, 44, 9]. In a different context, the largest eigenvalues of minors have recently been shown to follow Gumbel distribution as well [37].

We now briefly comment on the novelties of our method to prove (1.4); more details will be given in Sections 2.2 and 2.3. The standard method to extend any result on local eigenvalue statistics from the Gaussian case to a random matrix with a general entry distribution is the *Green function comparison theorem (GFT)* going back to [35], see also the related *Four moment theorem* of Tao and Vu [67]. Direct application of GFT in the bulk spectral regime for

---

[1]Another such situation, introduced first in [48], is the transition between Gumbel and Tracy-Widom distributions for GUE with an independent sizeable random deformation.

Hermitian matrices typically assumes that the third and fourth moments of the entry distribution also (almost) match, and a more sophisticated dynamical approach relying on the Dyson Brownian motion is necessary to remove these restrictive matching conditions [34, 17, 52]. At the edge regime, however, two matching moments are sufficient for GFT [36]. Alternatively, at the edges the Green functions can be controlled along the Ornstein-Uhlenbeck (OU) matrix flow for a very long time which allows one to compare a general matrix with a Gaussian one directly. This flow idea was first used in [53, 54] (see also [64, 63]) in the Hermitian context to investigate the Tracy-Widom edge universality and later in [23] for the non-Hermitian situation in the edge regime of the circular law. In the latter case first one translates the non-Hermitian eigenvalue problem to a Hermitian one via Girko's formula

$$(1.5) \qquad \sum_{\sigma \in \mathrm{Spec}(X)} f(\sigma) = -\frac{1}{4\pi} \int_{\mathbb{C}} \Delta f(z) A(z) \, \mathrm{d}^2 z, \qquad A(z) := \int_0^{\infty} \mathrm{Im} \, \mathrm{Tr} \, G^z(\mathrm{i}\eta) \, \mathrm{d}\eta,$$

and then performs the analysis for a continuous family of Hermitized resolvents, parametrized by an additional spectral parameter $z \in \mathbb{C}$, given by

$$(1.6) \qquad G^z(w) := (H^z - w)^{-1}, \qquad H^z := \begin{pmatrix} 0 & X - z \\ X^* - \overline{z} & 0 \end{pmatrix}, \qquad w \in \mathbb{C} \setminus \mathbb{R}.$$

Customarily one performs a cumulant expansion (see e.g. in [49, 19, 56, 45, 33] for the random matrix context) for the time derivative of $F(\mathrm{Tr} \, G_t(w))$, where F is a smooth test function and $G_t(w)$ is the Green function at time t. The spectral parameter w is chosen sufficiently close to the real axis to detect individual eigenvalues, i.e. $\eta := \mathrm{Im} \, w$ is smaller than the typical eigenvalue spacing, e.g. $\eta \ll n^{-2/3}$ in the Hermitian edge regime and $\eta \ll n^{-3/4}$ in the non-Hermitian edge regime where the spectral density of H^z develops a cusp singularity at zero. Typically the first and second order terms in the cumulant expansion are automatically cancelled by the choice of the OU process, the third and fourth order terms require careful estimates, while terms with higher order cumulants can be estimated quite crudely. The estimates are done via the *optimal local laws* that identify the leading deterministic term of the Green functions plus an error term. In the edge regimes where $\eta := \mathrm{Im} \, w$ is typically much larger than $1/n$, the cumulant expansion can be *iterated*: in every step one may gain an additional factor $\psi := 1/(n\eta) \ll 1$ in the error terms from the so-called *un-matched indices* (Definition 4.4), while the leading deterministic terms can be computed and the first non-vanishing one gives the final size. We need to exploit an explicit cancellation of the leading term after the z-integration in (1.5), forcing us to expand beyond the usual order. The iterative cumulant expansion has been systematically developed in [64, 63] after several previous works using the iterative gain from un-matched indices [36, 29] combined with cancellations of leading deterministic terms in certain situations [55, 46, 47].

The main difference between the current work and all previous applications of sophisticated cumulant expansions along a GFT proof is that now we work in a very *atypical* regime which means that all natural a priori estimates from local laws are off by a large factor (of size $n^{1/4}$). Indeed, due to the curvature of the unit circle near 1, the eigenvalues that may typically contribute to $\max \mathrm{Re} \, \mathrm{Spec}(X)$ are located in an elongated vertical box of size $n^{-1/2} \times n^{-1/4}$ (modulo logarithmic factors) with center around $1 + \sqrt{\gamma_n/4n}$ and this box contains typically finitely many (independently of n) eigenvalues, see Fig. 1 – this was proven for the Gaussian case in [28]. Therefore to obtain a lower and upper bound on $\max \mathrm{Re} \, \mathrm{Spec}(X)$ we will need to use (1.5) for a smooth test functions f supported on such box and we need to control (1.5) in expectation and variance sense.

If f in (1.5) is a smooth function supported in this box then $\int_{\mathbb{C}} |\Delta f(z)| \, \mathrm{d}^2 z \sim n^{1/4}$ is unusually large due to *strong anisotropy* of the box. The typical size of the fluctuation of $A(z)$ by local law is $\int_0^{\infty} \eta^{-1} \, \mathrm{d}\eta \sim O(1)$ (ignoring logarithmic singularities). Thus the naive

size of the fluctuation in the rhs. of Girko's formula in (1.5) is of order $n^{1/4}$ for a quantity that is only of order one by its lhs. This overestimate has a drastic effect on the usual cumulant expansions. Higher order terms in the cumulant expansion of $F(\int_{\mathbb{C}} \Delta f(z) A(z) \, \mathrm{d}^2 z)$ will involve higher powers of the problematic quantity $\int (\Delta f) A$ whose a priori size we do not control effectively. For smooth and bounded test functions F the standard iterative cumulant expansion, similar to [64, 63], is still effective but only if $n^{1/4} \psi = n^{1/4}/(n\eta) \ll 1$, i.e. in the regime where $\eta \gg n^{-3/4}$. We circumvent this difficulty by considering only the expectation and the variance of $\int (\Delta f) A$ instead of a general test function F which has the advantage that the Taylor expansion of F stops at first or second order. This restricted choice of F is the main reason why our current result is able to control only the size of $\max \operatorname{Re} \operatorname{Spec}(X)$ in (1.4) but not yet its Gumbel fluctuation.

The complementary $\eta \lesssim n^{-3/4}$ regime is not accessible by robust expansions. In fact, the regime $\eta \ll n^{-3/4}$ is dominated by the smallest (in modulus) eigenvalue λ^z of H^z (equivalently, the lowest singular value of $X - z$), especially by its lower tail behaviour. Two independent special effects need to be exploited simultaneously. First, there is a level repulsion between λ^z and $-\lambda^z$ (since the spectrum of H^z is symmetric to the origin, hence $-\lambda^z$ is also an eigenvalue), which effectively suppresses the event that $|\lambda^z|$ is much smaller than its natural scale $n^{-3/4}$. Second, the density of non-Hermitian eigenvalues (of X) are suppressed by a factor $e^{-n(|z|-1)^2/2}$ in the regime where $|z| \geq 1$, expressing the very strong concentration of the spectral radius near 1. Heuristically, this gives an additional small factor of order $e^{-n(|z|-1)^2/2}$ for the lower tail of λ^z as well. However, note that having a non-Hermitian eigenvalue extremely close to z and λ^z being unusually small is not an effectively controllable relation, even though $z \in \operatorname{Spec}(X)$ is equivalent to $\lambda^z = 0$. Both effects are extremely delicate and cannot be obtained directly for a general i.i.d. ensemble, but they can be extracted from the corresponding Ginibre ensemble via explicit calculations. We therefore first establish a very accurate lower tail estimate on λ^z in the Ginibre case (see Proposition 2.7 below), then via a separate GFT argument we transfer its consequence to the i.i.d. ensemble by obtaining an improved bound on $\boldsymbol{E} \operatorname{Im} \operatorname{Tr} G^z(\mathrm{i}\eta)$ (see Proposition 3.6 below). Up to an intermediate cutoff scale $\eta \ll n^{-7/8}$ this bound is sufficient to overcome the $n^{1/4}$ loss from $\int |\Delta f|$, ensuring that this small-η regime is negligible in the expectation sense and proving that only the $\eta \gtrsim n^{-7/8}$ regime matters in (1.5).

Finally, we revisit the iterative cumulant expansion for the large-η regime and use that we are interested only in the expectation and variance instead of a general F. This means that the factor $\int |\Delta f| \sim n^{1/4}$ occurs at most twice which can still be compensated by the improved estimate on $\boldsymbol{E} \operatorname{Im} \operatorname{Tr} G^z(\mathrm{i}\eta)$ in the cumulant expansions. For the comparison argument we also need a variance bound on the large-η regime for the Gaussian case, which is not available directly, but which we deduce indirectly from the variance of the lhs. of (1.5) (that is available via the Ginibre kernels [28]) and from the vanishing variance of the small-η regime. The optimal variance bound for the entire $\eta \ll n^{-3/4}$ regime would require the precise correlation between λ^z and $\lambda^{z'}$ for two different spectral parameters z, z' – an information that is not available even in the Gaussian case. Nevertheless, the suboptimal estimate, ignoring the decorrelation for large $z - z'$, is still sufficient for our smaller regime $\eta \ll n^{-7/8}$. This is another independent reason for choosing the threshold $n^{-7/8}$ for splitting the η-integral in (1.5).

Note that we use the $n^{-7/8}$ threshold to distinguish between the negligible small-η regime and the large-η regime requiring separate GFT comparisons, although the natural cutoff threshold should have been at $n^{-3/4}$ (the threshold $n^{-7/8}$ is only a technical choice).

In summary, our proof is much more involved than a typical direct iterative GFT argument and requires to choose an "unnatural" threshold $n^{-7/8}$ due to two main reasons: (i) we do not have almost matching a priori bounds due the anisotropy of the regime we consider, and (ii)

the necessary direct information on the correlation between λ^z and $\lambda^{z'}$ is lacking even in the Gaussian case.

Notations and conventions. We introduce some notations we use throughout the paper. For integers $k, l \in \mathbb{N}$ with $k \leq l$ we use the notation $[\![k, l]\!] := \{k, k+1, \ldots, l\}$. For positive quantities f, g we write $f \lesssim g$ and $f \sim g$ if $f \leq Cg$ or $cg \leq f \leq Cg$, respectively, for some constants $c, C > 0$ which depend only on the constants appearing in (2.1). For n-dependent positive sequences $f = f_n, g = g_n$ we also introduce $f \ll g$ indicating that $f_n = o(g_n)$. We denote vectors by bold-faced lower case Roman letters $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^k$, for some $k \in \mathbb{N}$. Vector and matrix norms, $\|\boldsymbol{x}\|$ and $\|A\|$, indicate the usual Euclidean norm and the corresponding induced matrix norm. For any $2n \times 2n$ matrix A we use the notation $\langle A \rangle := (2n)^{-1} \mathrm{Tr} A$ to denote the normalized trace of A. Moreover, for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$ and matrices $A, B \in \mathbb{C}^{2n \times 2n}$ we define

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \sum \overline{x}_i y_i, \qquad \langle A, B \rangle := \langle A^* B \rangle.$$

Moreover, we use $\Delta = 4\partial_z \partial_{\bar{z}}$ to denote the usual Laplacian and $\mathrm{d}^2 z$ denotes the Lebesgue measure on $\mathbb{C}$. For any function $h : \mathbb{C} \to \mathbb{C}$, we define the L^p-norm by $\|h\|_p^p := \int_{\mathbb{C}} |h(z)|^p \, \mathrm{d}^2 z$.

We will use the concept of "with very high probability" meaning that for any fixed $D > 0$ the probability of the event is bigger than $1 - n^{-D}$ if $n \geq n_0(D)$. Moreover, we use the convention that $\xi > 0$ denotes an arbitrary small constant which is independent of n. Finally, we introduce the notion of *stochastic domination* (see e.g. [30]): given two families of non-negative random variables

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

indexed by n (and possibly some parameter u in some parameter space $U^{(n)}$), we say that X is stochastically dominated by Y, if for all $\xi, D > 0$ we have

$$(1.7) \qquad \sup_{u \in U^{(n)}} \mathbb{P} \left[X^{(n)}(u) > n^\xi Y^{(n)}(u) \right] \leq n^{-D}$$

for large enough $n \geq n_0(\xi, D)$. In this case we use the notation $X \prec Y$ or $X = O_\prec(Y)$. We also use the convention that $\xi > 0$ denotes an arbitrary small constant which is independent of n. We often use the notation $\prec$ also for deterministic quantities, then the probability in (1.7) is zero for any $\xi > 0$ and sufficiently large n.

2. Main result and the proof ingredients. In this section we first formulate our main result precisely for the complex symmetry class. Then in Section 2.2 we collect some key ingredients of the proof from the literature: the local circular law, the strong concentration of the spectral radius, Girko's formula, the local law for the Hermitized matrix H^z, and most importantly we present a new lower tail bound on the smallest (in modulus) eigenvalue of H^z (Proposition 2.7 with its proof presented in the appendix). In the brief Section 2.3 we informally explain the main strategy that will be formalized in Section 3. Finally, in Section 2.4 we comment on the extension of our argument to the real symmetry class.

2.1. *Statement of the main result.* We consider $n \times n$ matrices X with independent identically distributed (i.i.d.) entries $x_{ab} \overset{\mathrm{d}}{=} n^{-1/2}\chi$. On the n-independent random variable χ we make the following assumption:

ASSUMPTION 2.1.   *We assume that $\boldsymbol{E}\chi = 0$, $\boldsymbol{E}|\chi|^2 = 1$; additionally in the complex case we also assume that $\boldsymbol{E}\chi^2 = 0$. Furthermore, for any $p \in \mathbb{N}$ we assume that there exists constants $C_p > 0$ such that*

$$(2.1) \qquad\qquad\qquad\qquad \boldsymbol{E}\left|\chi^p\right| \le C_p.$$

Moreover, we assume that there exists $\alpha, \beta > 0$ such that the probability density of χ, denoted by g, satisfies

$$(2.2) \qquad\qquad g \in L^{1+\alpha}(\mathbb{F}), \quad \|g\|_{1+\alpha} \le n^\beta, \qquad \mathbb{F} = \mathbb{R} \text{ or } \mathbb{C}.$$

REMARK 2.2.   *The condition on the density (2.2) is used only to control the unlikely event that there is a tiny singular value of $X - z$ in a very simple way; see (3.41) below. We make this assumption only to simplify the presentation of the proof. It can easily be removed with a separate argument from in [68, Section 6.1] (see also a slightly streamlined version in [50, Section 2.2]) as follows. To deal with a random matrix X failing to satisfy (2.2), one may add a tiny independent Gaussian component $n^{-\gamma}X_{\text{Gauss}}$ to X for some large fixed $\gamma > 0$ to achieve a probability density for the entries that satisfies (2.2), hence our main results hold for $X + n^{-\gamma}X_{\text{Gauss}}$. This tiny component $n^{-\gamma}X_{\text{Gauss}}$ can then be removed by using the proof of [68, Theorem 23] (or its refinement [50, Lemma 4]) that combines a sampling idea with the standard moment matching technique[2]. We will not present the details here since they are fairly standard and they are independent of our main arguments.*

Let $\{\sigma_i\}_{i \in [\![1,n]\!]}$ be the eigenvalues of X, then our main result is the following:

THEOREM 2.3.   *Let X be an $n \times n$ matrix satisfying Assumption 2.1 in the complex case, and define*

$$(2.3) \qquad\qquad\qquad \gamma_n := \frac{\log n - 5\log\log n - \log(2\pi^4)}{2}.$$

Then

$$(2.4) \qquad\qquad \lim_{n\to\infty} \boldsymbol{P}\left(\left|\max_{i\in[\![1,n]\!]}\operatorname{Re}\sigma_i - 1 - \sqrt{\frac{\gamma_n}{4n}}\right| \ge \frac{C_n}{\sqrt{4n\gamma_n}}\right) = 0,$$

for any sequence[3] $C_n \to \infty$.

We remark that Theorem 2.3 with the same proof holds if instead of the eigenvalue with the largest real part we consider the largest eigenvalue in any given direction, i.e. (2.4) holds for $\max_i \operatorname{Re}(e^{\mathrm{i}\theta}\sigma_i)$ with any fixed $\theta \in \mathbb{R}$, but for simplicity we consider the $\theta = 0$ case. We also note that the matrix elements of X do not necessarily have to be identically distributed. Our proof works with minor modifications as long as all the entries $\chi_{ab} = \sqrt{n}x_{ab}$ satisfy Assumption 2.1 uniformly for any a, b, but for simplicity we consider the i.i.d. case.

We stated our main Theorem 2.3 only for the complex case. The same result holds for the real case but we give a complete proof only for the complex case; we explain the reason in Section 2.4.

Our precise estimate on $\max_i \operatorname{Re}\sigma_i$ has the following immediate corollary on the solution to (1.2) with an i.i.d. matrix X.

---

[2]This theorem shows that if the first four moments of the single entry distributions of two ensembles X and X' coincide, then their microscopic local statistics are close with an effective error. With sufficiently high moment matching a straightforward modification of the proof of [68, Theorem 23] yields much finer error estimates up to any polynomial order in $1/n$ which can be used to offset the anisotropic loss of our test function f. The same conclusion also holds if the moments do not exactly match, but they differ only up to an order $n^{-\gamma}$.

[3]Following our proof we may obtain an effective control on the probability in (2.4) of order $O(C_n^{-\tau} + n^{-\tau})$ for some fixed $\tau > 0$.

COROLLARY 2.4. *Let $\boldsymbol{u}(t)$ be the solution to (1.2) with deterministic initial condition $\boldsymbol{u}(0)$ and with coupling constant $g = g_n$. Then for any sequence $C_n \to \infty$, the following statements hold with probability tending to one as $n \to \infty$:*

(i) *If $g \leq 1 - \sqrt{\frac{\gamma_n}{4n}} - \frac{C_n}{\sqrt{4n\gamma_n}}$, then*

$$\limsup_{t \to \infty} \frac{\max_i |u_i(t)|}{\max_i |u_i(0)|} = 0.$$

(ii) *If $g \geq 1 - \sqrt{\frac{\gamma_n}{4n}} + \frac{C_n}{\sqrt{4n\gamma_n}}$, then*

$$\limsup_{t \to \infty} \frac{\max_i |u_i(t)|}{\max_i |u_i(0)|} = \infty.$$

This corollary shows that to prevent the decay or blow-up of the solution for arbitrary long time, i.e. to remain in the so-called *stable regime* in many applications, it is *necessary* to fine tune the coupling constant very accurately. Our main theorem can also be used for *sufficient* conditions for stability up to a certain time scale of order $\sqrt{4n/\gamma_n}$ but we refrain from formalizing such statement. We note that the maximum norm on $\boldsymbol{u}$ and the deterministic initial condition indicate that we considered the worst-case scenario. As we mentioned in the introduction, the problem with random initial data and measuring stability in ℓ^2-sense has been investigated earlier and gives a somewhat different optimal tuning for g.

REMARK 2.5. *Theorem 2.3 implies that the sequence of random variables*

$$\mathfrak{G}_n := \sqrt{4n\gamma_n} \left[\max_{i \in [\![1,n]\!]} \operatorname{Re} \sigma_i - 1 - \sqrt{\frac{\gamma_n}{4n}} \right]$$

is tight, hence it has subsequential limits by Prokhorov's theorem. The limit is conjectured to be unique and to be the standard Gumbel distribution with distribution function $F(x) = \exp(-e^{-x})$ in the complex case and $F(x) = \exp(-\frac{1}{2}e^{-x})$ in the real case. For the Ginibre ensembles this conjecture was recently proven in [28].

2.2. *Proof ingredients.* First we recall two earlier results that locate $\max_{i \in [\![1,n]\!]} \operatorname{Re} \sigma_i$ on a cruder scale than our eventual target precision. The local circular law [18] implies that for any fixed $\tau > 0$ there is an eigenvalue in the rectangle

$$(2.5) \qquad \Omega_0 := \left[1 - \frac{n^\tau}{\sqrt{n}}, \ 1 + \frac{n^\tau}{\sqrt{n}} \right] \times \left[-\frac{n^{\tau/2}}{n^{1/4}}, \ \frac{n^{\tau/2}}{n^{1/4}} \right],$$

with very high probability, using the curvature of the boundary. In particular, this shows that typically $\max_{[\![1,n]\!]} \operatorname{Re} \sigma_i \geq 1 - n^{-1/2+\tau}$. Furthermore, by [6, Theorem 2.1] we have a strong concentration estimate for the spectral radius $\rho(X) = \max_{i \in [\![1,n]\!]} |\sigma_i|$:

$$(2.6) \qquad |\rho(X) - 1| \leq \frac{n^\tau}{\sqrt{n}}$$

for any $\tau > 0$, with very high probability. In particular, fixing a small $\tau > 0$, (2.5) and (2.6) imply that the rightmost eigenvalue is located in Ω_0, see Fig. 2.

Next we will show that with vanishing probability,

$$(2.7) \qquad \left| \max_{i \in [\![1,n]\!]} \operatorname{Re} \sigma_i - 1 - \sqrt{\frac{\gamma_n}{4n}} \right| \geq \frac{C_n}{\sqrt{4n\gamma_n}},$$
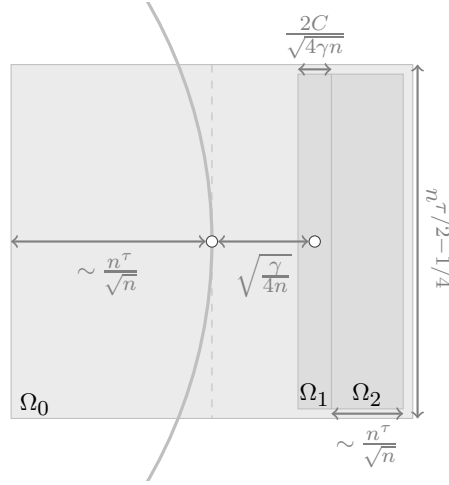
FIG 2. *The figure shows the domains* $\Omega_0, \Omega_1, \Omega_2$. *The set* Ω_1 *is chosen such that the number of eigenvalues therein are approximately Poisson-distributed with parameter* $2\sinh(C_n) \sim e^{C_n}$, *while in* Ω_2 *no eigenvalues are expected with high probability.*

with γ_n from (2.3) and C_n being any sequence $C_n \to \infty$. To see this, we define the following two sub-rectangles of Ω_0 with the same height:

$$(2.8) \qquad \Omega_1 := \left[1 + \sqrt{\frac{\gamma_n}{4n}} - \frac{C_n}{\sqrt{4n\gamma_n}},\ 1 + \sqrt{\frac{\gamma_n}{4n}} + \frac{C_n}{\sqrt{4n\gamma_n}}\right] \times \left[-\frac{n^{\tau/2}}{n^{1/4}},\ \frac{n^{\tau/2}}{n^{1/4}}\right],$$

$$(2.9) \qquad \Omega_2 := \left[1 + \sqrt{\frac{\gamma_n}{4n}} + \frac{C_n}{\sqrt{4n\gamma_n}},\ 1 + \frac{n^{\tau}}{\sqrt{n}}\right] \times \left[-\frac{n^{\tau/2}}{n^{1/4}},\ \frac{n^{\tau/2}}{n^{1/4}}\right],$$

see Fig. 2. From now on without loss of generality we may assume $1 \ll C_n \ll (\log n)^{1/2}$.

To prove the upper bound in (2.7), it suffices to show that

$$(2.10) \qquad \boldsymbol{P}(\#\{\sigma_i \in \Omega_2\} \geq 1) \leq \boldsymbol{E}[\#\{\sigma_i \in \Omega_2\}] = o(1),$$

here by $o(1)$ we denote a quantity that goes to zero as $n \to \infty$. To prove the matching lower bound in (2.7), $\boldsymbol{P}(\#\{\sigma_i \in \Omega_1\} = 0) = o(1)$, we need not only the expectation bound

$$(2.11) \qquad \boldsymbol{E}[\#\{\sigma_i \in \Omega_1\}] \geq c_0,$$

for some n-independent constant $c_0 > 0$, but also the concentration bound

$$(2.12) \qquad \boldsymbol{E}\left|\#\{\sigma_i \in \Omega_1\} - \boldsymbol{E}[\#\{\sigma_i \in \Omega_1\}]\right| = o(1)\,\boldsymbol{E}[\#\{\sigma_i \in \Omega_1\}].$$

More precisely, using the Markov inequality in combination with (2.11) and (2.12), we get

(2.13)
$$\boldsymbol{P}\left(\#\{\sigma_i \in \Omega_1\} = 0\right) \leq \boldsymbol{P}\left(\left|\#\{\sigma_i \in \Omega_1\} - \boldsymbol{E}[\#\{\sigma_i \in \Omega_1\}]\right| \geq \frac{\boldsymbol{E}[\#\{\sigma_i \in \Omega_1\}]}{2}\right) = o(1).$$

Now we have reduced the proof of (2.7) to proving the expectation bounds in (2.10), (2.11) and the concentration bound in (2.12). Their proof consist of two main steps. We first use the explicit formulae for the eigenvalue correlation functions of the Ginibre ensemble to show that (2.10)-(2.12) hold true for the Gaussian case. In fact, we will need a slightly modified version of these estimates where the counting functions are replaced by a smooth test functions supported on the corresponding Ω domains, see Lemma 3.1 below, whose proof

is an easy consequence of our estimates on the Ginibre correlation kernel from [28]. In the second step we then use the Green function comparison theorem (GFT) to extend the above estimates to general i.i.d. matrices. In the rest of this section we now introduce some tools for this second step and explain the strategy.

To perform a GFT analysis we rely on Girko's Hermitization formula [43] in the form introduced by Tao and Vu [68]:

(2.14)
$$\sum_{i=1}^{n} f(\sigma_i) = -\frac{1}{4\pi} \int_{\mathbb{C}} \Delta f(z) \int_0^T \operatorname{Im} \operatorname{Tr} G^z(\mathrm{i}\eta) \, \mathrm{d}\eta \, \mathrm{d}^2 z + \frac{1}{4\pi} \int_{\mathbb{C}} \Delta f(z) \log |\det(H^z - \mathrm{i}T)| \, \mathrm{d}^2 z,$$

for any $T > 0$ and for any compactly supported smooth test function $f \in C_c^2(\mathbb{C})$. Here we recall the definition of the $2n \times 2n$ Hermitian matrix H^z and its resolvent G^z from (1.6):

(2.15) $\qquad H^z := \begin{pmatrix} 0 & X - z \\ X^* - \overline{z} & 0 \end{pmatrix}, \qquad G^z(w) := (H^z - w)^{-1}, \quad w \in \mathbb{C} \setminus \mathbb{R}, \ z \in \mathbb{C}.$

The 2×2 block structure of H^z induces a symmetric spectrum around zero, i.e. the eigenvalues of H^z are $\{\lambda_{\pm i}^z\}_{i \in [\![1,n]\!]}$ (labelled in a non-decreasing order) with $\lambda_{-i}^z = -\lambda_i^z$ for $i \in [\![1,n]\!]$. Note that $\{\lambda_i^z\}_{i \in [\![1,n]\!]}$ exactly coincide with the singular values of $X - z$. As a consequence of the spectral symmetry of H^z, we find that

$$G_{vv}^z(\mathrm{i}\eta) = \mathrm{i} \operatorname{Im} G_{vv}^z(\mathrm{i}\eta), \quad \operatorname{Im} G_{vv}^z(\mathrm{i}\eta) > 0, \qquad v \in [2n], \quad \eta > 0.$$

Our fundamental input, the *local law for G^z* stated below in Theorem 2.6, asserts that as $n \to \infty$ the resolvent G^z becomes approximately deterministic. Its deterministic approximation is given by

(2.16)
$$M^z(\mathrm{i}\eta) = \begin{pmatrix} m^z(\mathrm{i}\eta) & \mathfrak{m}^z(\mathrm{i}\eta) \\ \overline{\mathfrak{m}^z}(\mathrm{i}\eta) & m^z(\mathrm{i}\eta) \end{pmatrix},$$

where m^z is the unique solution of the scalar equation

(2.17) $\qquad -\frac{1}{m^z(w)} = w + m^z(w) - \frac{|z|^2}{w + m^z(w)}, \quad \text{with} \quad \operatorname{Im}[m^z(w)]\operatorname{Im} w > 0,$

and

(2.18) $\qquad \mathfrak{m}^z(\mathrm{i}\eta) := -z u^z(\mathrm{i}\eta), \quad u^z(\mathrm{i}\eta) := \frac{\operatorname{Im} m^z(\mathrm{i}\eta)}{\eta + \operatorname{Im} m^z(\mathrm{i}\eta)}.$

By taking the real part of (2.17) it readily follows that on the imaginary axis m^z is purely imaginary, hence $m^z(\mathrm{i}\eta) = \mathrm{i} \operatorname{Im} m^z(\mathrm{i}\eta)$ (which also implies that $u^z(\mathrm{i}\eta)$ is real). In addition, by [6, Lemma 3.3] we have that

(2.19) $\qquad \operatorname{Im} m^z(\mathrm{i}\eta) \sim \begin{cases} \frac{\eta}{|1-|z|^2|+\eta^{2/3}}, & |z| > 1 \\ \eta^{1/3} + |1 - |z|^2|^{1/2}, & |z| \le 1 \end{cases}, \qquad 0 \le \eta \le 1.$

With these notations we have the local law for the resolvent G^z on the imaginary axis:

THEOREM 2.6 (Theorem 5.2 [6], Proposition 1 [23]). *For any deterministic vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^{2n}$ and matrix $A \in \mathbb{C}^{2n \times 2n}$, for any z with $-Cn^{-1/2} \lesssim |z| - 1 \le n^{-1/2+\tau}$ and $n^{-1} \le \eta \le 1$, we have*

(2.20) $\qquad \left| \langle \boldsymbol{x}, (G^z(\mathrm{i}\eta) - M^z(\mathrm{i}\eta))\boldsymbol{y} \rangle \right| \prec \|\boldsymbol{x}\| \|\boldsymbol{y}\| \left(\frac{1}{n^{1/2}\eta^{1/3}} + \frac{1}{n\eta} \right),$

(2.21) $\qquad \left| \langle A(G^z(\mathrm{i}\eta) - M^z(\mathrm{i}\eta)) \rangle \right| \prec \frac{\|A\|}{n\eta}.$

To apply Girko's formula (2.14) for the proof of (2.10)-(2.12) we need to regularize the indicator function on the corresponding $\Omega = \Omega_1, \Omega_2$ domains, and we also split the η-integration into two regimes which require different analysis. Hence, with some properly chosen smooth cut-off function f (see (3.2)–(3.6) below), we have

(2.22)

$$\#\{\sigma_i \in \Omega\} \approx \sum_{i=1}^n f(\sigma_i) \approx -\frac{1}{4\pi} \int_{\mathbb{C}} \Delta f(z) \left(\int_0^{\eta_0} + \int_{\eta_0}^T \right) \operatorname{Im} \operatorname{Tr} G^z(\mathrm{i}\eta) \, \mathrm{d}\eta \, \mathrm{d}^2 z =: I_0^{\eta_0} + I_{\eta_0}^T,$$

with $T := n^{100}$ and η_0 is an intermediate cutoff level $n^{-7/8-\tau}$ with the fixed $\tau > 0$ from (2.5). The small-η regime, $I_0^{\eta_0}$, and the large-η regime, $I_{\eta_0}^T$, are analysed separately.

For very small η, the local law (Theorem 2.6) does not effectively control the resolvent $G^z(\mathrm{i}\eta)$ as it is dominated by the smallest (in modulus) eigenvalue λ_1^z of H^z (equivalently, λ_1^z is the smallest singular value of $X - z$). We need a separate lower tail estimate for it, which is done again in two steps: first for Ginibre matrices and then we extend it to i.i.d. matrices via GFT. Besides the level repulsion effect, this accurate estimate also contains an additional small factor due to the fact that z is far outside of the unit disk, although this second effect is needed only for the Ginibre ensemble in this paper.

We now state the precise lower tail result for the Ginibre ensemble. Its proof, which is given in the appendix, relies on the explicit formula for the correlation functions of the eigenvalues of $(X - z)^*(X - z)$ from [13], or, alternatively, on the supersymmetric representation of its resolvent from [22]. In the sequel we denote by $\boldsymbol{P}^{\mathrm{Gin}}$, $\boldsymbol{E}^{\mathrm{Gin}}$ and $\boldsymbol{Var}^{\mathrm{Gin}}$ the corresponding probability, expectation and variance.

PROPOSITION 2.7. *Fix $\delta := |z|^2 - 1$ with $n^{-1/2} \ll \delta \ll 1$ and let λ_1^z be the smallest singular value of $X - z$, where X is a complex Ginibre matrix. Then there exists a constant $C > 0$, independent of n and δ, such that for any $y \leq C/(n\delta^2)$ we have the following lower tail bound*

(2.23)
$$\boldsymbol{P}^{\mathrm{Gin}}\left(\lambda_1^z \leq y\delta^{3/2} \right) \lesssim y^2 (n\delta^2)^{4/3} e^{-n\delta^2(1+O(\delta))/2}.$$

Recall that $\frac{1}{\pi} \operatorname{Im} m^z(x + \mathrm{i}0)$, the self-consistent density of states of H^z, has a gap of size $4\delta^3/27$ close to 0 justifying the $\delta^{3/2}$ scaling in (2.23) (see the paragraph below [22, Eq. (18a)], we remark that in [22] we defined δ with the opposite sign).

2.3. *Sketch of the proof of (2.10)-(2.12).* Having introduced the necessary ingredients, we briefly summarize the strategy to prove (2.10)-(2.12) for i.i.d. matrices. These steps will be outlined more precisely in Section 3 after the necessary cutoff functions are introduced.

The exponential factor in (2.23), obtained for $|z| > 1$ away from the boundary, ensures that with our choice $\eta_0 \ll n^{-7/8}$, $\boldsymbol{E}^{\mathrm{Gin}}\left| I_0^{\eta_0} \right|^2$ is negligible, which implies that the main contribution to both the expectation and the variance of (2.22) comes from the large-η regime, $I_{\eta_0}^T$, at least for the Ginibre matrices. Next we will use GFT arguments to extend these Ginibre estimates to generic i.i.d. matrices. We first show that $\boldsymbol{E}\left| I_0^{\eta_0} \right|$ is negligible also for i.i.d. matrices using the bound on the resolvent for Ginibre ensemble implied by Proposition 2.7 together with a GFT argument. Then we will consider the large-η regime and use another GFT to show that $\boldsymbol{E}[I_{\eta_0}^T]$ and $\boldsymbol{Var}[I_{\eta_0}^T]$ have the same bound as their Ginibre counterparts (modulo a negligible error). Finally, we need a bound on $\boldsymbol{Var}^{\mathrm{Gin}}[I_{\eta_0}^T]$, which is not accessible directly, but can be deduced indirectly from $\boldsymbol{Var}^{\mathrm{Gin}}(I_0^{\eta_0} + I_{\eta_0}^T) \approx \boldsymbol{Var}^{\mathrm{Gin}}(\#\{\sigma_i \in \Omega\})$ and using that $\boldsymbol{E}^{\mathrm{Gin}}\left| I_0^{\eta_0} \right|^2$ is negligible.

2.4. *Comments on the real symmetry class.* We stated and proved our main result only for the complex case; the proof for the real case would require two changes. First, the precise form of the cumulant expansions behind the GFT arguments slightly depends on the symmetry class: the real case yields some extra terms that can be treated routinely (see e.g. [24, 64] for an analogous extension). Second, we prove the lower tail bound in Proposition 2.7 only for the complex case since our detailed proof relies on the formula from [13] that has no analogue for the real case. The alternative supersymmetric method has both complex and real versions, the latter being considerably more involved. For the sake of brevity, we show how the complex version can also be used to prove Proposition 2.7 and we omit the more cumbersome details of the real case, however we have no doubt that this analysis is feasible based upon our experience from [22, 26]. Precise tail bounds in both symmetry classes for the $|z| \leq 1 + O(n^{-1/2})$ regime have already been proven in [22, Corollary 2.4] (see also [26, 27]). In the real case the factor y^2 in (2.23) is replaced with $(y^2 + y \exp\left(-\frac{n}{2}(\operatorname{Im} z)^2\right))$ indicating a weaker level repulsion near the real axis. This weaker estimate however does not affect the usage of Proposition 2.7 in the main body of our proof since it is effective only for a small regime of the z parameter.

The rest of the paper is organized as follows. In Section 3, we prove our main theorem using the GFTs for the two different η regimes as an input. Next, we prove the GFT used for the small-η integral, i.e. Proposition 3.6 in Section 4, and then in Section 5 we present the proof of Proposition 3.8 used for the large-η integral.

3. Proof of Theorem 2.3. To use Girko's formula for proving (2.10)-(2.12), we need to first regularize the indicator functions. For the domains Ω_k, $k = 1, 2$, given in (2.8) and (2.9), we may choose two smooth cut-off functions f_k^- and f_k^+ which are supported on a slightly smaller domain $\Omega_k^- \subset \Omega_k$ and a slightly larger domain $\Omega_k^+ \supset \Omega_k$ respectively, such that

$$(3.1) \qquad \sum_{i=1}^n f_k^-(\sigma_i) \leq \#\{\sigma_i \in \Omega_k\} \leq \sum_{i=1}^n f_k^+(\sigma_i), \qquad k = 1, 2.$$

In fact, for $k = 1$ we will only need the lower bound, while for $k = 2$ we need only the upper bound, so we will define only f_1^- and f_2^+, the other two cut-off functions are not used in our proof.

More precisely, for the domain Ω_1 given in (2.8), we choose the lower bound cut-off function

$$(3.2) \qquad f_1^-(z) := g_1^-(x)h_1^-(y), \qquad z = x + \mathrm{i}y \in \mathbb{C},$$

where $g_1^-(x) \in [0, 1]$ and $h_1^-(y) \in [0, 1]$ are smooth functions given by

$$(3.3) \qquad g_1^-(x) = \begin{cases} 1, & |x - L| \leq 4l_n/5, \\ 0, & |x - L| \geq l_n, \end{cases}, \qquad h_1^-(y) = \begin{cases} 1, & |y| \leq 4h_n/5, \\ 0, & |y| \geq h_n. \end{cases}$$

Here, we used the shorthand notations

$$(3.4) \qquad L := 1 + \sqrt{\frac{\gamma_n}{4n}}, \qquad l_n := \frac{C_n}{\sqrt{4n\gamma_n}}, \qquad h_n := n^{-1/4 + \tau/2},$$

with $1 \ll C_n \ll \sqrt{\log n}$. Additionally, g_1^-, h_1^- are defined so that their second derivatives can be bounded by

$$(3.5) \qquad \|(g_1^-)''\|_1 \lesssim l_n^{-1}, \qquad \|(h_1^-)''\|_1 \lesssim h_n^{-1}.$$

For the spectral domain Ω_2 given in (2.10), we similarly choose $f_2^+(x+iy) := g_2^+(x)h_2^+(y)$ where g_2^+ is supported on the regime enlarging the x-domain of Ω_2 by $l_n/5$ and h_2^+ is supported on the regime enlarging the y-domain of Ω_2 by $h_n/5$ (c.f., (3.3)), such that

$$(3.6) \qquad \|(g_2^+)''\|_1 \lesssim l_n^{-1}, \qquad \|(h_2^+)''\|_1 \lesssim h_n^{-1}.$$

We are now ready to present the proof of our main result.

PROOF OF THEOREM 2.3. With the cut-off functions f_1^- and f_2^+ as above we have (3.1) and from (3.5) and (3.6), we also have

$$(3.7) \qquad \|\Delta f_1^-\|_1, \ \|\Delta f_2^+\|_1 \lesssim \frac{h_n}{l_n} \lesssim n^{\frac{1}{4}+\frac{\tau}{2}}\sqrt{\log n}.$$

We will show the following results in expectation

$$(3.8) \qquad \boldsymbol{E}\Big[\sum_{i=1}^n f_2^+(\sigma_i)\Big] = o(1); \qquad \boldsymbol{E}\Big[\sum_{i=1}^n f_1^-(\sigma_i)\Big] \geq c_0,$$

for some constant $c_0 > 0$, and the concentration result

$$(3.9) \qquad \boldsymbol{E}\Big|\sum_{i=1}^n f_1^-(\sigma_i) - \boldsymbol{E}\Big[\sum_{i=1}^n f_1^-(\sigma_i)\Big]\Big| = o(1)\Big(\boldsymbol{E}\Big[\sum_{i=1}^n f_1^-(\sigma_i)\Big]\Big).$$

These two key results easily imply Theorem 2.3. More precisely, combining the second inequality in (3.1) for $k=2$ with the first estimate in (3.8), we have

$$(3.10) \qquad \boldsymbol{P}(\#\{\sigma_i \in \Omega_2\} \geq 1) \leq \boldsymbol{P}\Big(\sum_{i=1}^n f_2^+(\sigma_i) \geq 1\Big) \leq \boldsymbol{E}\Big[\sum_{i=1}^n f_2^+(\sigma_i)\Big] = o(1).$$

Moreover, using the first inequality in (3.1) for $k=1$, the Markov inequality in combination with (3.8) and (3.9), we have

$$\boldsymbol{P}\left(\#\{\sigma_i \in \Omega_1\} = 0\right) \leq \boldsymbol{P}\Big(\sum_{i=1}^n f_1^-(\sigma_i) = 0\Big)$$

$$(3.11) \qquad \leq \boldsymbol{P}\left(\Big|\sum_{i=1}^n f_1^-(\sigma_i) - \boldsymbol{E}\Big[\sum_{i=1}^n f_1^-(\sigma_i)\Big]\Big| \geq \frac{\boldsymbol{E}\Big[\sum_{i=1}^n f_1^-(\sigma_i)\Big]}{2}\right) = o(1).$$

This shows that the key bounds (3.8) and (3.9) are indeed sufficient for the proof of Theorem 2.3.

The first step to prove these key bounds is to use the explicit formula for the eigenvalue correlation functions of the complex Ginibre ensemble to show that (3.8) and (3.9) hold true for the Gaussian case.

LEMMA 3.1. *For the complex Ginibre ensemble, we have*

$$(3.12) \qquad \boldsymbol{E}^{\text{Gin}}\Big[\sum_{i=1}^n f_2^+(\sigma_i)\Big] \lesssim e^{-4C_n/5}; \qquad \boldsymbol{E}^{\text{Gin}}\Big[\sum_{i=1}^n f_1^-(\sigma_i)\Big] \gtrsim \sinh(4C_n/5),$$

with any $C_n \ll \sqrt{\log n}$ and

$$(3.13) \qquad \boldsymbol{Var}^{\text{Gin}}\Big[\sum_{i=1}^n f_1^-(\sigma_i)\Big] \lesssim e^{-2C_n/5}\Big(\boldsymbol{E}^{\text{Gin}}\Big[\sum_{i=1}^n f_1^-(\sigma_i)\Big]\Big)^2.$$

The proof of Lemma 3.1 is given in Section 3.4. The next step is to extend this lemma to generic i.i.d. matrices. From now on, we use f to denote either of the smooth cut-off functions f_1^-, f_2^+ for simplicity of notation. Lacking explicit formula for eigenvalue correlations for i.i.d. matrices, we first link the linear statistics in Lemma 3.1 to the Green function of H^z using Girko's formula in (2.14). Choosing T sufficiently large, e.g. $T = n^{100}$, we next show that the last term in (2.14) is very small with very high probability. Note that

$$\log|\det(H^z - \mathrm{i}T)| = 2n\log T + \sum_j \log\left(1 + \left(\frac{\lambda_j^z}{T}\right)^2\right) = 2n\log T + O\left(\frac{\mathrm{Tr}(H^z)^2}{T^2}\right)$$

$$(3.14) \qquad\qquad = 2n\log T + O_\prec\left(\frac{n^2}{T^2}\right),$$

where we used that $|x_{ij}| \prec n^{-1/2}$ from (2.1). Using the L^1 norm of Δf in (3.7), we have

$$(3.15) \qquad \left|\frac{1}{4\pi}\int_{\mathbb{C}} \Delta f(z)\log|\det(H^z - \mathrm{i}T)|\,\mathrm{d}^2 z\right| = O_\prec(n^{-100}).$$

Therefore, we have

$$\sum_{i=1}^n f(\sigma_i) = -\frac{1}{4\pi}\int_{\mathbb{C}} \Delta f(z)\left(\int_0^{\eta_0} + \int_{\eta_0}^T\right)\mathrm{Im}\,\mathrm{Tr}\,G^z(\mathrm{i}\eta)\,\mathrm{d}\eta\,\mathrm{d}^2 z + O_\prec(n^{-100})$$

$$(3.16) \qquad\qquad =: I_0^{\eta_0}(f) + I_{\eta_0}^T(f) + O_\prec(n^{-100}),$$

where we split the η integral into the two parts at the truncation level $\eta_0 := n^{-7/8-\tau}$, where $\tau > 0$ is the fixed small number from (2.5). Then the proof of (3.8) and (3.9) is reduced to studying $I_0^{\eta_0}(f)$ and $I_{\eta_0}^T(f)$ respectively. The idea is to first estimate $I_0^{\eta_0}(f)$ and $I_{\eta_0}^T(f)$ for the Ginibre ensemble and then extend these estimates to i.i.d. matrices using GFT arguments respectively.

Next we outline the three main steps of the rest of the proof, the precise details will be given in the following three subsections, respectively.

Step 1. For the Ginibre ensemble, we use the explicit lower tail bound for the smallest eigenvalue λ_i^z in Proposition 2.7 to show that (see Lemma 3.4 below)

$$(3.17) \qquad \mathbf{E}^{\mathrm{Gin}}\left[\left|I_0^{\eta_0}(f)\right|^2\right] = o(1), \qquad f = f_1^- \text{ or } f_2^+.$$

Combining this with Lemma 3.1, we have

$$(3.18) \qquad \mathbf{E}^{\mathrm{Gin}}[I_{\eta_0}^T(f_2^+)] = o(1), \qquad \mathbf{E}^{\mathrm{Gin}}[I_{\eta_0}^T(f_1^-)] \geq c,$$

for some constant $c > 0$, and

$$(3.19) \qquad \mathbf{E}^{\mathrm{Gin}}\left(I_{\eta_0}^T(f_1^-) - \mathbf{E}^{\mathrm{Gin}}[I_{\eta_0}^T(f_1^-)]\right)^2 = o(1)\left(\mathbf{E}^{\mathrm{Gin}}[I_{\eta_0}^T(f_1^-)]\right)^2.$$

Step 2. We next use a GFT argument (see Proposition 3.7 below) together with the corresponding estimate of the resolvent (see Lemma 3.2 below) for the Ginibre ensemble to show

$$(3.20) \qquad \mathbf{E}\,|I_0^{\eta_0}(f)| = o(1), \qquad f = f_1^- \text{ or } f_2^+.$$

This directly implies that

$$\mathbf{E}\left[\sum_{i=1}^n f(\sigma_i)\right] = \mathbf{E}[I_{\eta_0}^T(f)] + o(1),$$

$$(3.21) \qquad \mathbf{E}\left|\sum_{i=1}^n f(\sigma_i) - \mathbf{E}\left[\sum_{i=1}^n f(\sigma_i)\right]\right| = \mathbf{E}\left|I_{\eta_0}^T(f) - \mathbf{E}[I_{\eta_0}^T(f)]\right| + o(1).$$

Step 3. For the large–η integral $I_{\eta_0}^T$, we use another GFT argument (see Proposition 3.8 below) to extend the corresponding Ginibre estimates (3.18)–(3.19) obtained in Step 1 to i.i.d. matrices, *i.e.,*

$$\boldsymbol{E}\left[I_{\eta_0}^T(f_2^+)\right] = o(1); \qquad \boldsymbol{E}\left[I_{\eta_0}^T(f_1^-)\right] \geq c,$$

(3.22) $$\boldsymbol{E}\left|I_{\eta_0}^T(f_1^-) - \boldsymbol{E}\left[I_{\eta_0}^T(f_1^-)\right]\right|^2 = o(1)\left(\boldsymbol{E}\left[I_{\eta_0}^T(f_1^-)\right]\right)^2.$$

The variance bound (3.22) directly implies

$$\boldsymbol{E}\left|I_{\eta_0}^T(f_1^-) - \boldsymbol{E}[I_{\eta_0}^T(f_1^-)]\right| = o(1)\left|\boldsymbol{E}[I_{\eta_0}^T(f_1^-)]\right|.$$

Combining this with (3.21) for $f = f_1^-$, we proved the concentration estimate in (3.9). It is also straightforward to prove the expectation estimates in (3.8) using the first line of (3.21) and the corresponding expectation estimates in (3.22). Hence we completed the proof of Theorem 2.3. □

Notice that our strategy follows a somewhat unconventional indirect route. Typical proofs based upon the Green function comparison method assume that all necessary information is available for the Gaussian model. This does not quite hold in our case; Step 1 above is not a purely explicit calculation. While the statistics of the eigenvalues of the Ginibre ensemble are fully available via explicit formulas in both symmetry classes, less is known about the eigenvalues of H^z. For a fixed z, their correlation functions are known, at least in the complex case, but no explicit formula is available for their statistics for different z's. Note that the variance of $I_{\eta_0}^T(f)$ in Step 1 involves the correlation between eigenvalues of H^z and $H^{z'}$ for two different z, z' and the necessary estimate requires this correlation to decay when $z - z'$ are far away. While it is plausible that the local spectral statistics of H^z and $H^{z'}$, especially their lowest eigenvalues, are independent whenever $|z - z'| \gg n^{-1/2}$, this has only been shown [24, 25] in the regime of their typical behavior; now we would need such information in the atypical lower tail regime. Lacking such decorrelation bound, the variance of $I_{\eta_0}^T(f)$ is controlled indirectly as

$$\boldsymbol{Var}^{\mathrm{Gin}}(I_{\eta_0}^T(f)) \approx \boldsymbol{Var}^{\mathrm{Gin}}\left(I_0^{\eta_0}(f) + I_{\eta_0}^T(f)\right),$$

as long as $\boldsymbol{Var}^{\mathrm{Gin}}(I_0^{\eta_0}(f)) \ll \boldsymbol{Var}^{\mathrm{Gin}}\left(I_0^{\eta_0}(f) + I_{\eta_0}^T(f)\right)$. Using (3.16), the explicit Ginibre eigenvalue statistics gives the control for $\boldsymbol{Var}^{\mathrm{Gin}}\left(I_0^{\eta_0}(f) + I_{\eta_0}^T(f)\right)$. We cannot control $\boldsymbol{Var}^{\mathrm{Gin}}(I_0^{\eta_0}(f))$ optimally, but we chose the threshold η_0 sufficiently small that this variance is negligible by (3.17) and thus no effective decorrelation estimate is necessary.

We now explain in details how Steps 1-3 are proven.

3.1. *Ginibre estimate for $I_0^{\eta_0}(f)$ and $I_{\eta_0}^T(f)$.* Using the explicit lower tail estimate of the smallest eigenvalue of H^z in Proposition 2.7 with X being the complex Ginibre ensemble, we obtain the following improved estimate:

LEMMA 3.2. *Fix $\delta = |z|^2 - 1$ with $n^{-1/2} \ll \delta \ll 1$. Then for any $\eta \leq C/(n\delta^{1/2})$, it holds*

$$\boldsymbol{E}^{\mathrm{Gin}}\left[\mathrm{Im}\langle G^z(\mathrm{i}\eta)\rangle\right] \lesssim \frac{\eta}{\delta} + n^{1+\xi}\eta\delta,$$

(3.23) $$\boldsymbol{E}^{\mathrm{Gin}}\left[\left(\mathrm{Im}\langle G^z(\mathrm{i}\eta)\rangle\right)^2\right] \lesssim \frac{n^\xi}{n^{2/3}\delta^{1/3}}e^{-n\delta^2(1+O(\delta))/2} + \frac{\eta^2}{\delta^2} + n^{2+\xi}\eta^2\delta^2,$$

where $\xi > 0$ is an arbitrary small constant.

REMARK 3.3. *The exponential factor in (2.23) for the smallest eigenvalue λ_1^z does not manifest in the first moment estimate on the resolvent since the main contribution comes from larger eigenvalues. For the second moment, however, the lowest eigenvalue plays the leading role.*

The proof of Lemma 3.2 will be given in Section 3.4. Therefore from Lemma 3.2, for any $z \in \mathrm{supp}(f_1^-) \cup \mathrm{supp}(f_2^+)$ and $\eta \leq n^{-3/4-\epsilon}$, we have

$$(3.24) \quad \boldsymbol{E}^{\mathrm{Gin}}\left[\mathrm{Im}\langle G^z(\mathrm{i}\eta)\rangle\right] \lesssim n^\xi \sqrt{n}\eta, \qquad \boldsymbol{E}^{\mathrm{Gin}}\left[\left(\mathrm{Im}\langle G^z(\mathrm{i}\eta)\rangle\right)^2\right] \lesssim n^\xi\left(n\eta^2 + n^{-3/4}\right).$$

Thus we can prove that $I_0^{\eta_0}(f)$ is negligible in the second moment sense:

LEMMA 3.4. *With $\eta_0 = n^{-7/8-\tau}$ and $f = f_1^-$ or f_2^+, we have*

$$(3.25) \qquad\qquad\qquad \boldsymbol{E}^{\mathrm{Gin}}|I_0^{\eta_0}(f)|^2 = O_\prec(n^{-\tau/2}).$$

The proof of Lemma 3.4 is given in Section 3.4. Combining Lemma 3.4 with Lemma 3.1 and (3.16), we have proved (3.18) and (3.19).

Next, we will use the Green function comparison to extend these Ginibre estimates to generic i.i.d. matrices satisfying Assumption 2.1.

3.2. *Green function comparison for $I_0^{\eta_0}(f)$.* In this part, we will show that $I_0^{\eta_0}(f)$ is negligible in the first absolute moment for generic i.i.d. matrices.

LEMMA 3.5. *With $\eta_0 = n^{-7/8-\tau}$, for any $f = f_1^-$ or f_2^+, we have*

$$(3.26) \qquad \boldsymbol{E}\,|I_0^{\eta_0}(f)| = \boldsymbol{E}\left|\frac{n}{4\pi}\int_{\mathbb{C}} \Delta f(z) \int_0^{\eta_0} \mathrm{Im}\langle G^z(\mathrm{i}\eta)\rangle\,\mathrm{d}\eta\,\mathrm{d}^2z\right| = O(n^{-\tau}).$$

We will postpone the proof of Lemma 3.5 to Section 3.4. The proof of Lemma 3.5 crucially relies on the following improved estimate of the resolvent at the intermediate level $\eta = \eta_0$ using the monotonicity of $\mathrm{Im}\,\mathrm{Tr}\,G(\mathrm{i}\eta)$.

PROPOSITION 3.6. *Let X be an i.i.d. complex matrix satisfying Assumption 2.1. For any small $\epsilon > 0$, then for any $n^{-1+\epsilon} \leq \eta \leq n^{-3/4-\epsilon}$ and $-Cn^{-1/2} \lesssim |z|-1 \leq n^{-1/2+\tau}$, we have*

$$(3.27) \qquad\qquad \boldsymbol{E}\left[\mathrm{Im}\langle G^z(\mathrm{i}\eta)\rangle\right] \prec n^{1/2}\eta + \frac{1}{n^{5/2}\eta^2} + \frac{1}{n^5\eta^5} + n^{-1}.$$

Proposition 3.6 is a direct result of the Ginibre estimate in Lemma 3.2 and the following Green function comparison which will be proved in Section 4.

PROPOSITION 3.7. *For any small $\epsilon > 0$, for any $n^{-1+\epsilon} \leq \eta \leq n^{-3/4-\epsilon}$ and $-Cn^{-1/2} \lesssim |z|-1 \leq n^{-1/2+\tau}$, we have*

$$(3.28) \qquad \left|\boldsymbol{E}[\langle G^z(\mathrm{i}\eta)\rangle] - \boldsymbol{E}^{\mathrm{Gin}}[\langle G^z(\mathrm{i}\eta)\rangle]\right| \prec \frac{1}{n^{5/2}\eta^2} + \frac{1}{n^5\eta^5} + n^{-1}.$$

Therefore, combining Lemma 3.5 with (3.16), for any $f = f_1^-$ or f_2^+, we proved (3.21) in the effective form

$$\boldsymbol{E}\left[\sum_{i=1}^n f(\sigma_i)\right] = \boldsymbol{E}[I_{\eta_0}^T(f)] + O(n^{-\tau}),$$

$$(3.29) \qquad \boldsymbol{E}\left|\sum_{i=1}^n f(\sigma_i) - \boldsymbol{E}\left[\sum_{i=1}^n f(\sigma_i)\right]\right| = \boldsymbol{E}\left|I_{\eta_0}^T(f) - \boldsymbol{E}[I_{\eta_0}^T(f)]\right| + O(n^{-\tau}).$$

Hence to prove the target estimates in (3.8) and (3.9), it suffices to study $I_{\eta_0}^T(f)$.

3.3. *Green function comparison for $I_{\eta_0}^T(f)$.*   In this part, we will extend the estimates of $I_{\eta_0}^T(f)$ in (3.18) and (3.19) from Gaussian matrices to generic i.i.d. matrices. It then suffices to establish the following Green function comparison for $I_{\eta_0}^T(f)$, whose proof will be presented in Section 5.

PROPOSITION 3.8.   *For any $f = f_1^-$ or f_2^+, there exist some constants $c_1, c_2 > 0$ such that*

$$\left| \boldsymbol{E}[I_{\eta_0}^T(f)] - \boldsymbol{E}^{\mathrm{Gin}}[I_{\eta_0}^T(f)] \right| = O(n^{-c_1}), \tag{3.30}$$

and

$$\left| \boldsymbol{E}\left(I_{\eta_0}^T(f) - \boldsymbol{E}[I_{\eta_0}^T(f)]\right)^2 - \boldsymbol{E}^{\mathrm{Gin}}\left(I_{\eta_0}^T(f) - \boldsymbol{E}^{\mathrm{Gin}}[I_{\eta_0}^T(f)]\right)^2 \right| = O(n^{-c_2}). \tag{3.31}$$

Combining the Ginibre estimates (3.18)–(3.19) with the GFT estimates (3.30)–(3.31), we obtain the second line of (3.22)

$$\boldsymbol{E}\left(I_{\eta_0}^T(f_1^-) - \boldsymbol{E}[I_{\eta_0}^T(f_1^-)]\right)^2 = o(1)\left(\boldsymbol{E}[I_{\eta_0}^T(f_1^-)] + O(n^{-c_1})\right)^2 + O(n^{-c_2}) = o(1)\left(\boldsymbol{E}[I_{\eta_0}^T(f_1^-)]\right)^2.$$

The first line of (3.22) is obtained similarly.

3.4. *Proof of some lemmas.*   We now give the proofs of Lemmas 3.1, 3.2, 3.4 and 3.5.

PROOF OF LEMMA 3.1.   We first recall that the expectation and variance of the linear statistics of a general test function f can be expressed as

$$\begin{aligned}
\boldsymbol{E}^{\mathrm{Gin}} \sum_i f(\sigma_i) &= \int f(z)\widetilde{K}_n(z,z)\,\mathrm{d}^2 z \\
\boldsymbol{Var}^{\mathrm{Gin}} \sum_i f(\sigma_i) &= \int f(z)^2 \widetilde{K}_n(z,z)\,\mathrm{d}^2 z - \iint f(z)f(w)\big|\widetilde{K}_n(z,w)\big|^2\,\mathrm{d}^2 z\,\mathrm{d}^2 w
\end{aligned} \tag{3.32}$$

in terms of the kernel $\widetilde{K}_n(z,w)$ from [28]. Using the kernel asymptotics from [28, Lemma 6] we obtain

$$\int_{L-t/\sqrt{4\gamma n}}^{L+t/\sqrt{4\gamma n}} \int_{-s/(\gamma n)^{1/4}}^{s/(\gamma n)^{1/4}} \widetilde{K}_n(z,z)\,\mathrm{d}\operatorname{Im} z\,\mathrm{d}\operatorname{Re} z = 2\operatorname{erf}(s)\sinh(t)\left(1 + O\Big(\frac{\log\log n + t^2 + s^4}{\log n}\Big)\right) \tag{3.33}$$

for $|t| + s^2 \le \sqrt{\log n}/10$, while for any $t > 0$ we have the bounds

$$\int_{L-t/\sqrt{4\gamma n}}^{L+t/\sqrt{4\gamma n}} \left(\int_{-\infty}^{\sqrt{2t}/(\gamma n)^{1/4}} + \int_{\sqrt{2t}/(\gamma n)^{1/4}}^{\infty}\right)\widetilde{K}_n(z,z)\,\mathrm{d}\operatorname{Im} z\,\mathrm{d}\operatorname{Re} z \lesssim e^{-t/4} \tag{3.34}$$

$$\int_{L+t/\sqrt{4\gamma n}}^{\infty} \int_{\mathbb{R}} \widetilde{K}_n(z,z)\,\mathrm{d}\operatorname{Im} z\,\mathrm{d}\operatorname{Re} z \lesssim e^{-t/4}.$$

From (3.32)–(3.34) we immediately conclude

$$\boldsymbol{E}^{\mathrm{Gin}} \sum_i f_1^-(\sigma_i) \gtrsim \sinh\Big(\frac{4C_n}{5}\Big) \tag{3.35}$$

and

$$(3.36) \qquad \boldsymbol{Var}^{\mathrm{Gin}} \sum_i f_1^-(\sigma_i) \le \int f(z)^2 \widetilde{K}_n(z,z)\,\mathrm{d}^2 z \lesssim \sinh\left(\frac{6C_n}{5}\right)$$

recalling that $C_n \ll \sqrt{\log n}$, proving (3.13). Finally, the first bound in (3.12) follows directly from (3.34). □

PROOF OF LEMMA 3.2. We reformulate (2.23) as follows: for any $\eta \le \widetilde{\eta} := C/(n\delta^{1/2})$,

$$(3.37) \qquad \boldsymbol{P}^{\mathrm{Gin}}\left(\lambda_1^z \le \eta\right) \lesssim \frac{n^{4/3}\eta^2}{\delta^{1/3}} e^{-n\delta^2(1+O(\delta))/2}.$$

Using spectral decomposition of H^z and the spectrum symmetry, we write

$$\boldsymbol{E}^{\mathrm{Gin}}[\mathrm{Im}\langle G^z(\mathrm{i}\eta)\rangle] = \frac{1}{n}\sum_{k=0}^{K_0} \boldsymbol{E}^{\mathrm{Gin}}\left[\sum_{3^k\eta \le \lambda_i < 3^{k+1}\eta} \frac{\eta}{(\lambda_i^z)^2 + \eta^2}\right] + \frac{1}{n}\boldsymbol{E}^{\mathrm{Gin}}\left[\sum_{\lambda_i^z \ge \widetilde{\eta}} \frac{\eta}{(\lambda_i^z)^2 + \eta^2}\right]$$

$$(3.38) \qquad \lesssim \frac{n^\xi}{n}\sum_{k=0}^{K_0} \frac{1}{3^{2k}\eta + \eta} \boldsymbol{P}^{\mathrm{Gin}}\left(\lambda_1^z \le 3^{k+1}\eta\right) + \frac{1}{n}\boldsymbol{E}^{\mathrm{Gin}}\left[\sum_{\lambda_i^z \ge \widetilde{\eta}} \frac{\eta}{(\lambda_i^z)^2 + \eta^2}\right],$$

with $K_0 = \lceil\log_3(\widetilde{\eta}/\eta)\rceil = O(\log n)$, where we used the rigidity of eigenvalues and $\xi > 0$ is an arbitrary small number. The second term in (3.38) can be bounded effectively using the local law in (2.21) and (2.19), *i.e.,*

$$(3.39)$$
$$\frac{1}{n}\boldsymbol{E}^{\mathrm{Gin}}\left[\sum_{\lambda_i^z \ge \widetilde{\eta}} \frac{\eta}{(\lambda_i^z)^2 + \eta^2}\right] = \frac{\eta}{\widetilde{\eta}}\boldsymbol{E}^{\mathrm{Gin}}\left[\frac{\widetilde{\eta}}{n}\sum_{\lambda_j \ge \widetilde{\eta}} \frac{1}{\lambda_j^2 + \widetilde{\eta}^2}\right] \lesssim \frac{\eta}{\widetilde{\eta}}\boldsymbol{E}^{\mathrm{Gin}}[\mathrm{Im}\langle G^z(\mathrm{i}\widetilde{\eta})\rangle] \lesssim \frac{\eta}{\delta} + n^{1+\xi}\eta\delta,$$

where $\xi > 0$ is an arbitary small number. Using (3.37), the first term in (3.38) can be bounded by

$$\frac{n^\xi}{n}\sum_{k=0}^{K_0} \frac{1}{3^{2k}\eta + \eta} \boldsymbol{P}^{\mathrm{Gin}}\left(\lambda_1^z \le 3^{k+1}\eta\right) \lesssim \frac{n^\xi}{n}\sum_{k=0}^{K_0} \frac{1}{(3^{2k}+1)\eta} \frac{3^{2k+2}n^{4/3}\eta^2}{\delta^{1/3}} \lesssim n^\xi \log n \frac{n^{1/3}\eta}{\delta^{1/3}}.$$

Note that we did not use the exponential factor in (3.37) yet since this bound is already much smaller compared to the second term estimate in (3.39) for $\delta \gg n^{-1/2}$. Therefore, we finished the proof of the first moment estimate in (3.23).

Similarly, for the second moments, we have

$$\boldsymbol{E}^{\mathrm{Gin}}[(\mathrm{Im}\langle G^z(\mathrm{i}\eta)\rangle)^2] = \frac{1}{n^2}\boldsymbol{E}^{\mathrm{Gin}}\left[\left(\sum_{k=0}^{K_0}\sum_{3^k\eta \le \lambda_i^z < 3^{k+1}\eta} \frac{\eta}{(\lambda_i^z)^2 + \eta^2} + \sum_{\lambda_i^z \ge \widetilde{\eta}} \frac{\eta}{(\lambda_i^z)^2 + \eta^2}\right)^2\right]$$

$$\lesssim \log n \frac{n^\xi}{n^2}\sum_{k=0}^{K_0} \frac{1}{(3^{2k}+1)^2\eta^2} \boldsymbol{P}^{\mathrm{Gin}}\left(\lambda_1^z \le 3^{k+1}\eta\right) + \frac{\eta^2}{n^2}\boldsymbol{E}^{\mathrm{Gin}}\left[\left(\sum_i \frac{1}{(\lambda_i^z)^2 + \widetilde{\eta}^2}\right)^2\right]$$

$$(3.40) \qquad \lesssim \log n \frac{n^\xi}{n^{2/3}\delta^{1/3}} e^{-n\delta^2(1+O(\delta))/2} + \frac{\eta^2}{\delta^2} + n^{2+\xi}\eta^2\delta^2,$$

with $\xi > 0$ being any arbitrary small number, where we used the tail bound in (3.37) and the local law in (2.21). This finishes the proof of Lemma 3.2. □

Using Lemma 3.2, we will prove Lemma 3.4 and 3.5. Since the proof of these two lemmas is similar, we present only the detailed proof of Lemma 3.5.

PROOF OF LEMMAS 3.4 AND 3.5. We start with showing that the regime $\eta \in [0, n^{-l}]$, for some very large $l > 0$ is negligible, exactly as it was done in [6]. By a direct computation,

$$\int_0^{n^{-l}} \operatorname{Im} \operatorname{Tr} G(\mathrm{i}\eta) \, \mathrm{d}\eta = \frac{1}{2} \left(\sum_{|\lambda_i| \lesssim n^{-l}} + \sum_{|\lambda_i| \gtrsim n^{-l}} \right) \log \left(1 + \frac{n^{-2l}}{\lambda_i^2} \right).$$

The second sum can easily be estimated using Lemma 3.2 or Proposition 3.6, *i.e.,*

$$\boldsymbol{E} \left[\sum_{|\lambda_i| \gtrsim n^{-l}} \log \left(1 + \frac{n^{-2l}}{\lambda_i^2} \right) \right] \lesssim (\log n) \, \boldsymbol{E} \, |\{ i : |\lambda_i| \leq \eta_1 \}| + \frac{n^{1-2l}}{\eta_1^2} \lesssim n^{-1/4 - 100\tau} + n^{-2l+3},$$

where we chose $\eta_1 := n^{-7/8 - 100\tau}$. To estimate the first sum, we recall [6, Proposition 5.7], *i.e.,* under the density condition (2.2), there exists $C_\alpha > 0$ such that

$$(3.41) \qquad \mathbb{P} \left(\min_{i=-n}^{n} |\lambda_i^z| \leq \frac{u}{n} \right) \leq C_\alpha u^{\frac{2\alpha}{1+\alpha}} n^{\beta+1}, \qquad z \in \mathbb{C}, u > 0,$$

with α, β given in (2.2). Then following [6][Eq. (5.34)-(5.35)], we have

$$\boldsymbol{E} \left[\sum_{|\lambda_i| \lesssim n^{-l}} \log \left(1 + \frac{n^{-2l}}{\lambda_i^2} \right) \right] \lesssim n \, \boldsymbol{E} \left[|\log \lambda_1^z| \mathbb{1}_{\lambda_1 \lesssim n^{-l}} \right] \lesssim n^{-10},$$

with l large enough depending on α, β. Combining this bound with the L^1-norm of Δf in (3.7), we conclude that the very tiny regime $\eta \in [0, n^{-l}]$ is negligible.

Hence, it is enough to estimate the contribution to $I_0^{\eta_0}(f)$ of the remaining η-integral over $[n^{-l}, \eta_0]$. Using that $\eta \to \eta \operatorname{Im} \operatorname{Tr} G^z(\mathrm{i}\eta)$ is increasing in $\eta \geq 0$, we have

$$\left| \int_{\mathbb{C}} \Delta f(z) \, \boldsymbol{E} \left[\int_{n^{-l}}^{\eta_0} \operatorname{Im} \operatorname{Tr} G^z(\mathrm{i}\eta) \, \mathrm{d}\eta \right] \mathrm{d}^2 z \right| \leq \int_{\mathbb{C}} |\Delta f(z)| \int_{n^{-l}}^{\eta_0} \frac{\eta_0}{\eta} \, \boldsymbol{E} \left[\operatorname{Im} \operatorname{Tr} G^z(\mathrm{i}\eta_0) \right] \mathrm{d}\eta \, \mathrm{d}^2 z$$

$$(3.42) \qquad\qquad \leq C(\log n) n \eta_0 \int_{\mathbb{C}} |\Delta f(z)| \, \boldsymbol{E} \left[\operatorname{Im} \langle G^z(\mathrm{i}\eta_0) \rangle \right] \mathrm{d}^2 z.$$

Using the estimate in (3.27) with $\eta_0 = n^{-7/8-\tau}$ *i.e.,*

$$(3.43) \qquad\qquad \boldsymbol{E}[\operatorname{Im} \langle G^z(\mathrm{i}\eta_0) \rangle] \prec n^{1/2} \eta_0 + n^{-5/8 + 5\tau} \lesssim n^{-3/8 - \tau},$$

together with the L^1 norm of Δf in (3.7), by (3.42) we obtain that

$$(3.44) \qquad\qquad \left| \int_{\mathbb{C}} \Delta f(z) \, \boldsymbol{E} \left[\int_{n^{-l}}^{\eta_0} \operatorname{Im} \operatorname{Tr} G^z(\mathrm{i}\eta) \, \mathrm{d}\eta \right] \mathrm{d}^2 z \right| = O(n^{-\tau}).$$

We hence finish the proof of Lemma 3.5. Lemma 3.4 can be proven similarly using the second estimate in (3.24). $\qquad \square$

4. Green function comparison for resolvents: Proof of Proposition 3.7.

Before starting with the proof of Proposition 3.7 we introduce some notations which we will use throughout this section.

NOTATION 4.1. We use lower case letters to denote the indices taking values in $[\![1, n]\!]$ and upper case letters to denote the indices taking values in $[\![n+1, 2n]\!]$. We also use calligraphic letters $\mathfrak{u}, \mathfrak{v}$ to denote the indices ranging fully from 1 to $2n$.

For any index $\mathfrak{v} \in [\![1, 2n]\!]$, the conjugate of $\mathfrak{v}$, denoted by $\operatorname{conj}(\mathfrak{v})$, is given by $\operatorname{conj}(\mathfrak{v}) \in [\![1, 2n]\!]$ and it is such that $|\operatorname{conj}(\mathfrak{v}) - \mathfrak{v}| = n$. In particular, for an index $a \in [\![1, n]\!]$, we define

its index conjugate $\mathrm{conj}(a) = \bar{a} := a+n$, and for an index $B \in [\![n+1, 2n]\!]$ we define its index conjugate $\mathrm{conj}(B) = \underline{B} := B - n$. With a slight abuse of terminology, we say two indices *coincide* if either they are equal or one is equal to the conjugate of the other one. For instance, we say $a \in [\![1, n]\!]$ coincides with the index $B \in [\![n+1, 2n]\!]$ if $a = \underline{B}$ (or equivalently $B = \bar{a}$). We also say that a collection of indices are *distinct* if there is no index coincidence among them (in the sense explained above).

Moreover, we often use generic letters x and y to denote the (first) row and the (second) column index of a Green function entry. In this context the lower case letters x, y do not indicate that they take values in $[\![1, n]\!]$. We say that a Green function entry G_{xy} is *diagonal* if $x = y$ or $x = \mathrm{conj}(y)$; otherwise we say that G_{xy} is off-diagonal.

We now explain this terminology with an example:

$$(4.1) \qquad \frac{1}{n^2} \sum_{a=1}^{n} \sum_{B=n+1}^{2n} G_{aB} G_{B\bar{a}} = \frac{1}{n^2} \sum_{a=1}^{n} \sum_{B \neq \bar{a}} G_{aB} G_{B\bar{a}} + \frac{1}{n^2} \sum_{a=1}^{n} G_{a\bar{a}} G_{\bar{a}\bar{a}},$$

where we split the summation into two parts: 1) the two summation indices a and B are distinct; 2) there is an index coincidence $B = \bar{a}$ (or equivalently $a = \underline{B}$) in the summation. In the first term the two Green function entries are off-diagonal, while in the second term the two Green function entries are diagonal.

We prove Proposition 3.7 via a continuous interpolating flow. Given the initial ensemble H^z in (2.15), we consider the matrix flow

$$(4.2) \qquad \mathrm{d}H_t^z = -\frac{1}{2}(H_t^z + Z)\,\mathrm{d}t + \frac{1}{\sqrt{n}}\,\mathrm{d}\mathscr{B}_t, \quad Z = \begin{pmatrix} 0 & zI \\ \bar{z}I & 0 \end{pmatrix}, \quad \mathscr{B}_t = \begin{pmatrix} 0 & B_t \\ B_t^* & 0 \end{pmatrix}$$

where B_t is an $n \times n$ matrix with independent standard complex valued Brownian motion entries. The matrix flow H_t^z interpolates between the initial matrix H^z in (2.15) and an independent matrix as in (2.15) with X being replaced with an independent complex Ginibre ensemble.

The Green function of the time dependent matrix H_t^z is denoted by G_t^z. Since the flow in (4.2) is stochastically Hölder continuous in time, the local law for the Green function in Theorem 2.6 also holds true for the time dependent Green function G_t^z simultaneously for all $t \geq 0$ by a grid argument, together with the Hölder regularity of G_t^z for $0 \leq t \leq n^{100}$ and a simple perturbation argument for $t \geq n^{100}$. More precisely for $n^{-1+\epsilon} \leq \eta \leq n^{-3/4-\epsilon}$ and $-Cn^{-1/2} \lesssim |z| - 1 \leq n^{-1/2+\tau}$, it holds uniformly

$$(4.3) \qquad \sup_{t \geq 0} \max_{1 \leq \mathfrak{v}, \mathfrak{u} \leq 2n} \left\{ \left| \left(G_t^z(\mathrm{i}\eta)\right)_{\mathfrak{u}\mathfrak{v}} - m^z \delta_{\mathfrak{v}=\mathfrak{u}} - \mathfrak{m}^z \delta_{|\mathfrak{v}-\mathfrak{u}|=n} \right| \right\} \prec \Psi := \frac{1}{n\eta},$$

where $m^z, \mathfrak{m}^z$ are given in (2.18) and (2.19), and they are such that

$$(4.4) \qquad m^z \equiv m^z(\mathrm{i}\eta) = O(\Psi), \qquad \mathfrak{m}^z \equiv \mathfrak{m}^z(\mathrm{i}\eta) = O(1).$$

Note that in our range of parameters, the small deterministic term $m^z \delta_{\mathfrak{v}=\mathfrak{u}}$ may be included in the error in (4.3). We remark that the diagonal Green function entry $G_{\mathfrak{u}\mathfrak{u}}$ as well as the off-diagonal Green function entries are bounded by Ψ with very high probability, while the other kind of diagonal Green function entry $G_{\mathfrak{u},\mathrm{conj}(\mathfrak{u})}$ can only be bounded by 1 with very high probability.

In the following, we often drop the dependence on t and η and set $G^z \equiv G_t^z(\mathrm{i}\eta)$ for notational simiplicity. Without specfic mentioning, all the estimates in this section hold true uniformly for any $-Cn^{-1/2} \lesssim |z| - 1 \leq n^{-1/2+\tau}$, $n^{-1+\epsilon} \leq \eta \leq n^{-3/4-\epsilon}$ and $t \geq 0$.

PROOF OF PROPOSITION 3.7. In order to prove Proposition 3.7, it suffices to show the following estimate on the time derivative of $\langle G_t^z(i\eta) \rangle$:

LEMMA 4.2. *For any $n^{-1+\epsilon} \le \eta \le n^{-3/4-\epsilon}$, $-Cn^{-1/2} \lesssim |z| - 1 \le n^{-1/2+\tau}$ and $t \ge 0$, it holds*

$$(4.5) \qquad \left| \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{E}[\langle G_t^z(i\eta) \rangle] \right| = O_\prec(n^{-1/2}\Psi^2 + \Psi^5 + n^{-1}).$$

Integrating (4.5) over $t \in [0, t_0]$ with $t_0 = 100 \log n$ we obtain

$$(4.6) \qquad \left| \boldsymbol{E}[\langle G_0^z(i\eta) \rangle] - \boldsymbol{E}[\langle G_{t_0}^z(i\eta) \rangle] \right| = O_\prec\left(\log n(n^{-1/2}\Psi^2 + \Psi^5 + n^{-1}) \right).$$

Note that H_t^z in (4.2) is given as in (2.15) with X being replaced with the time dependent matrix

$$X_t \overset{\mathrm{d}}{=} e^{-\frac{t}{2}}X + \sqrt{1 - e^{-t}}\mathrm{Gin}(\mathbb{C}), \qquad t \ge 0,$$

where $X_\infty \overset{\mathrm{d}}{=} \mathrm{Gin}(\mathbb{C})$ is the complex Ginibre ensemble which is independent of X. Then we have

$$(4.7) \qquad \|G_{t_0}^z(i\eta) - G_\infty^z(i\eta)\| \le \|G_{t_0}^z\| \|G_\infty^z\| \|X_{t_0} - X_\infty\| \prec n^{-48},$$

where we used that $\|G^z(i\eta)\| \le \eta^{-1} \le n$ and that $|X_{ij}| \prec n^{-1/2}$ from the assumption in (2.1). Combining (4.6) with (4.7) we conclude the proof of Proposition 3.7. $\qquad\square$

We now present the proof of Lemma 4.2.

4.1. *Proof of Lemma 4.2.* Recall the matrix flow in (4.2) with complex-valued X, we set

$$(4.8) \qquad W \equiv W_t = H_t^z + Z = \begin{pmatrix} 0 & X_t \\ X_t^* & 0 \end{pmatrix}.$$

Then $\mathrm{d}H_t^z = -\frac{1}{2}W_t\,\mathrm{d}t + \frac{1}{\sqrt{n}}\mathrm{d}\mathscr{B}_t$. Applying Ito's formula and setting $\partial/\partial w_{aB} = \partial/\partial h_{aB}$, we have

$$\mathrm{d}\langle G_t^z \rangle = \sum_{a,B} \frac{\partial \langle G_t^z \rangle}{\partial h_{aB}} \mathrm{d}h_{aB} + \sum_{a,B} \frac{\partial \langle G_t^z \rangle}{\partial \overline{h}_{aB}} \mathrm{d}\overline{h}_{aB}$$

$$+ \frac{1}{2} \sum_{a,B} \frac{\partial^2 \langle G_t^z \rangle}{\partial h_{aB}\partial \overline{h}_{aB}} \mathrm{d}h_{aB}\,\mathrm{d}\overline{h}_{aB} + \frac{1}{2} \sum_{a,B} \frac{\partial^2 \langle G_t^z \rangle}{\partial \overline{h}_{aB}\partial h_{aB}} \mathrm{d}\overline{h}_{aB}\,\mathrm{d}h_{aB}$$

$$= \left(-\frac{1}{2} \sum_{a,B} w_{aB} \frac{\partial \langle G_t^z \rangle}{\partial w_{aB}} - \frac{1}{2} \sum_{a,B} \overline{w}_{aB} \frac{\partial \langle G_t^z \rangle}{\partial \overline{w}_{aB}} + \frac{1}{n} \sum_{a,B} \frac{\partial^2 \langle G_t^z \rangle}{\partial w_{aB}\partial \overline{w}_{aB}} \right) \mathrm{d}t$$

$$(4.9) \qquad + \frac{1}{\sqrt{n}} \sum_{a,B} \frac{\partial \langle G_t^z \rangle}{\partial w_{aB}} \mathrm{d}(\mathscr{B}_t)_{aB} + \frac{1}{\sqrt{n}} \sum_{a,B} \frac{\partial \langle G_t^z \rangle}{\partial \overline{w}_{aB}} \mathrm{d}\overline{(\mathscr{B}_t)}_{aB}.$$

Note that the expectation of the martingale term on the last line of (4.9) vanishes. It then suffices to study the expectation of the remaining terms. We note that

$$(4.10) \qquad \langle G^z \rangle = \frac{1}{n} \sum_{v=1}^{n} G_{vv}^z = \frac{1}{n} \sum_{V=n+1}^{2n} G_{VV}^z = \frac{1}{2n} \sum_{\mathfrak{v}=1}^{2n} G_{\mathfrak{v}\mathfrak{v}}^z,$$

which follows from the spectral symmetry induced by the 2×2 block matrix in (2.15). Performing the cumulant expansion formula on $\{w_{aB}\}$ and $\{\overline{w_{aB}}\}$ on the right side of (4.9) (see e.g. [45, Lemma 7.1]), we observe the precise cancellations of the second order terms, and the summation below starts from the third order terms *i.e.*, $p + q + 1 = 3$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\boldsymbol{E}[\langle G_t^z \rangle] = \boldsymbol{E}\left[\frac{\mathrm{d}\langle G_t^z \rangle}{\mathrm{d}t}\right] = -\frac{1}{2n}\sum_{v,a=1}^{n}\sum_{B=n+1}^{2n}\left(\sum_{p+q+1=3}^{K_0}\frac{c_{aB}^{(p+1,q)}}{p!q!n^{\frac{p+q+1}{2}}}\,\boldsymbol{E}\left[\frac{\partial^{p+q+1}G_{vv}^z}{\partial w_{aB}^{p+1}\partial \overline{w_{aB}}^q}\right]\right)$$

$$(4.11) \qquad -\frac{1}{2n}\sum_{v,a=1}^{n}\sum_{B=n+1}^{2n}\left(\sum_{p+q+1=3}^{K_0}\frac{c_{aB}^{(q,p+1)}}{p!q!n^{\frac{p+q+1}{2}}}\,\boldsymbol{E}\left[\frac{\partial^{p+q+1}G_{vv}^z}{\partial \overline{w_{aB}}^{p+1}\partial w_{aB}^q}\right]\right) + O_\prec(n^{-\frac{K_0}{2}+2}),$$

where $c_{aB}^{(p,q)}$ are the (p,q)-cumulants of the normalized complex-valued entries $\sqrt{n}w_{aB}$, with $c_{aB}^{(p,q)} = c_{Ba}^{(q,p)}$ from the complex symmetry, and we omit their dependence on t for simplicity. The last error stems from truncating the cumulant expansions at a sufficiently large K_0-th order, say $K_0 = 100$, using the local law in (4.3) and the finite moment condition in (2.1); see also [23] for a similar truncation argument.

For simplicity we assume i.i.d. entries of X in the our model, thus the cumulants are independent of the indices, $c_{aB}^{(p,q)} = c^{(p,q)}$. We next consider only the first line of (4.11), *i.e.*,

$$(4.12) \qquad \sum_{p+q+1=3}^{K_0} L_{p+1,q}^z := \sum_{p+q+1=3}^{K_0}\frac{c^{(p+1,q)}}{2p!q!}\left(\frac{1}{n^{\frac{p+q+3}{2}}}\sum_{v,a,B}\boldsymbol{E}\left[\frac{\partial^{p+q+1}G_{vv}^z}{\partial w_{aB}^{p+1}\partial \overline{w_{aB}}^q}\right]\right),$$

and the second line of (4.11) is exactly the same as (4.12) by interchanging a with B.

Using the following differentiation rules for any $1 \le \mathfrak{u}, \mathfrak{v} \le 2n$

$$(4.13) \qquad \frac{\partial G_{\mathfrak{u}\mathfrak{v}}^z}{\partial w_{aB}} = -G_{\mathfrak{u}a}^z G_{B\mathfrak{v}}^z, \qquad \frac{\partial G_{\mathfrak{u}\mathfrak{v}}^z}{\partial \overline{w_{aB}}} = -G_{\mathfrak{u}B}^z G_{a\mathfrak{v}}^z,$$

each term $L_{p+1,q}^z$ in (4.12) can be written as a linear combination of products of $p + q + 2$ Green function entries of the form

$$(4.14) \qquad \frac{1}{n^{\frac{p+q+3}{2}}}\sum_{v=1}^{n}\sum_{a=1}^{n}\sum_{B=n+1}^{2n}\boldsymbol{E}\left[\prod_{i=1}^{p+q+2}G_{x_iy_i}^z\right].$$

Here x_i, y_i denote generic row and column indices of G^z, respectively, to which we assign actual summation indices v, a, B, depending on the precise structure of the corresponding term dictated by (4.12)–(4.13). The assignment will be denoted by the symbol $\equiv$, *e.g.*, $x_i \equiv a$, $y_i \equiv B$ means that the generic factor $G_{x_iy_i}^z$ is replaced with the actual G_{aB}^z in (4.14). Note that both lower and upper case summation indices can be assigned to the generic x, y indices. The assignments that appear from (4.12)–(4.13) have the following properties: $x_1 \equiv v$, $y_{p+q+2} \equiv v$, and all the other indices $x_i, y_i \equiv$ either a or B such that

$$(4.15) \qquad \#\{x_i \equiv a\} = \#\{y_i \equiv B\} = q, \qquad \#\{x_i \equiv B\} = \#\{y_i \equiv a\} = p+1.$$

From the local law in (4.3) and (4.4), we have $|G_{aa}^z|, |G_{BB}^z|, |G_{aB}^z|, |G_{Ba}^z| \prec \Psi$ unless $a = \underline{B}$. If we restrict to the summation when all the indices are distinct in (4.14) (*i.e.*, $a \ne \underline{B}$, $v \ne a$, and $v \ne \underline{B}$), then the product of $p + q + 2$ Green function entries in (4.14) can be bounded by Ψ^{p+q+2}. For the remaining summation when there is some index coincidence (e.g. $a = \underline{B}$), we gain a factor n^{-1} since the number of free summation indices is reduced by one. Therefore, we obtain the following so-called *naive estimate*

$$(4.16) \qquad |L_{p+1,q}^z| \prec n^{-\frac{p+q-3}{2}}\left(\Psi^{p+q+2} + n^{-1}\right).$$

Thus for any $p + q + 1 \geq 4$, we have

$$(4.17) \qquad\qquad |L^z_{p+1,q}| \prec \Psi^5 + n^{-1}.$$

However the naive estimate in (4.16) for $p + q + 1 = 3$ is not sufficiently fine to prove Lemma 4.2.

Next we focus on proving an improved estimate for these third order terms, *i.e.,* proving

$$(4.18) \qquad\qquad \sum_{p+q+1=3} |L^z_{p+1,q}| = O_{\prec}\big(n^{-1/2}\Psi^2 + n^{-1}\big).$$

By direct computations, the third order terms $L^z_{p+1,q}$ with $p + q + 1 = 3$ in (4.12) are linear combinations of the following terms

$$\frac{\sqrt{n}}{n^3} \sum_{v,a,B} \boldsymbol{E}[G^z_{va}G^z_{BB}G^z_{aa}G^z_{Bv}], \quad \frac{\sqrt{n}}{n^3} \sum_{v,a,B} \boldsymbol{E}[G^z_{va}G^z_{Ba}G^z_{Ba}G^z_{Bv}],$$

$$(4.19) \qquad \frac{\sqrt{n}}{n^3} \sum_{v,a,B} \boldsymbol{E}[G^z_{va}G^z_{BB}G^z_{aB}G^z_{av}], \quad \frac{\sqrt{n}}{n^3} \sum_{v,a,B} \boldsymbol{E}[G^z_{vB}G^z_{aB}G^z_{aa}G^z_{Bv}],$$

as well as the other terms with the index a and B interchanged. As explained below (4.15), we split the threefold summations in (4.19) into the following three cases (recall the concept of coinciding and distinct indices from Notation 4.1):

1) all three summation indices coincide in the summation (*i.e.,* $v = a = \underline{B}$): the resulting sum in (4.19) can be bounded by $O_{\prec}(n^{-3/2})$ using that $|G_{\mathfrak{uv}}| \prec 1$ from (4.3), which is small enough to prove (4.18);
2) exactly two of the summation indices coincide (*e.g.,* $a = \underline{B} \neq v$ or $a = v \neq \underline{B}$): the resulting sum in (4.19) can be bounded by $O_{\prec}(n^{-1/2}\Psi^2)$ using the local law in (4.3) and (4.4), which is also sufficient to prove (4.18);
3) all three summation indices are distinct (*i.e.,* $a \neq \underline{B} \neq v$): using the local law in (4.3) and (4.4) naively, the resulting term in (4.19) can be bounded by $O_{\prec}(\sqrt{n}\Psi^4)$, which is however far from the truth. We observe from (4.19) that these third order terms have indices a and B that both appear three times as a first and as a second index of a G-factor. A somewhat more complicated version of this feature (see the concept of unmatched indices in Definition 4.4 later) allows us to improve the bound on them.

Next, we will discuss in details for the third order terms from (4.19) in Case 3), *i.e.,* with the summation restriction of all indices different, $a \neq \underline{B} \neq v$. We first introduce the shifted version of the Green function

$$(4.20) \qquad\qquad \widehat{G^z} := G^z - M^z = O_{\prec}(\Psi), \qquad M^z = \begin{pmatrix} m^z & \mathfrak{m}^z \\ \overline{\mathfrak{m}^z} & m^z \end{pmatrix},$$

with m^z and $\mathfrak{m}^z$ given in (2.18). The shifted version $\widehat{G^z}$ differs from G^z only for the diagonal entries, *i.e.,* $G_{xy} = \widehat{G_{xy}}$ unless $x = y$ or $x = \mathrm{conj}(y)$. Then the first term among the third order terms in (4.19) with $a \neq \underline{B} \neq v$ can be written as (omitting the factor $\sqrt{n}$)

$$\frac{1}{n^3} \sum_{a\neq\underline{B}\neq v} \boldsymbol{E}[G^z_{va}G^z_{BB}G^z_{aa}G^z_{Bv}] = \frac{1}{n^3} \sum_{a\neq\underline{B}\neq v} \boldsymbol{E}[\widehat{G^z_{va}}\widehat{G^z_{BB}}\widehat{G^z_{aa}}\widehat{G^z_{Bv}}] + \frac{m^z}{n^3} \sum_{a\neq\underline{B}\neq v} \boldsymbol{E}[\widehat{G^z_{va}}\widehat{G^z_{BB}}\widehat{G^z_{Bv}}]$$

$$(4.21) \qquad\qquad + \frac{m^z}{n^3} \sum_{a\neq\underline{B}\neq v} \boldsymbol{E}[\widehat{G^z_{va}}\widehat{G^z_{aa}}\widehat{G^z_{Bv}}] + \frac{(m^z)^2}{n^3} \sum_{a\neq\underline{B}\neq v} \boldsymbol{E}[\widehat{G^z_{va}}\widehat{G^z_{Bv}}].$$

Note that the terms on the right side above are averaged products of shifted Green function entries of the form defined in (4.23) below. Moreover, these terms are unmatched since the index a (or B) appears odd number times in the product of Green function entries which clearly does not satisfy the match condition in (4.25); see Defintion 4.4 below. In fact, any term in (4.19) with the restriction $a \neq \underline{B} \neq v$ can be written as a linear combination of unmatched terms of the form in (4.23) as in (4.21) with a factor $\sqrt{n}$. Using Proposition 4.5 below in combination with additional contributions from Case 1) and 2) with the index coincidences, we have obtained the improved estimate for the third order terms in (4.18).

Combining (4.11), (4.17) and (4.18), we finish the proof of Lemma 4.2. $\qquad\square$

Before giving the formal definition of unmatched indices (and unmatched terms) to study the third order terms in *e.g.,* (4.21) from Case 3) systematically, we first set some notational conventions.

For any fixed $l_1, l_2 \in \mathbb{N}$, we use $\mathscr{I}_{l_1,l_2}$ to denote a set of l_1 lower case letters and l_2 upper case letters, *e.g.,* the set may contain lower case letters a, v and upper case letter B as in (4.21). In general, we may write $\mathscr{I}_{l_1,l_2} := \{v_j\}_{j=1}^{l_1} \cup \{V_j\}_{j=1}^{l_2}$. Each element in $\mathscr{I}_{l_1,l_2}$ will represent a summation index and the font type of each letter indicates the range of the summation for that index; as before, the lower case letters v_j run from 1 to n, and the upper case letters V_j run from $n+1$ to $2n$. We denote the sum over these $l := l_1 + l_2$ summation indices (indicated by $\mathscr{I}_{l_1,l_2}$) by

$$\sum_{\mathscr{I}_{l_1,l_2}} = \sum_{v_1, \cdots v_{l_1}, V_1, \cdots, V_{l_2}} := \sum_{v_1=1}^{n} \cdots \sum_{v_{l_1}=1}^{n} \sum_{V_1=n+1}^{2n} \cdots \sum_{V_{l_2}=n+1}^{2n}.$$

We also introduce a partial summation restricted to distinct indices,

$$(4.22) \qquad \sum_{\mathscr{I}_{l_1,l_2}}^{*} := \sum_{v_1, \cdots v_{l_1}, V_1, \cdots, V_{l_2}} \Big(\prod_{j \neq j'}^{l_1} \delta_{v_j \neq v_{j'}} \Big) \Big(\prod_{j \neq j'}^{l_2} \delta_{V_j \neq V_{j'}} \Big) \Big(\prod_{j=1}^{l_1} \prod_{j'=1}^{l_2} \delta_{v_j \neq \underline{V_{j'}}} \Big),$$

i.e., each summation index in $\mathscr{I}_{l_1,l_2}$ is different from all the other indices and their conjugates.

DEFINITION 4.3. *Given $l_1, l_2 \in \mathbb{N}$ and a collection of lower and upper case summation indices $\mathscr{I}_{l_1,l_2} = \{v_j\}_{j=1}^{l_1} \cup \{V_j\}_{j=1}^{l_2}$, we consider a product of d generic shifted Green function entries $\widehat{G_{x_1y_1}^z G_{x_2y_2}^z} \cdots \widehat{G_{x_dy_d}^z}$ and assign a summation index v_j, V_j or their conjugates $\overline{v_j}, \underline{V_j}$ to each generic index x_i, y_i (e.g., $x_1 \equiv v_2, y_1 \equiv \underline{V_5}, x_2 \equiv \overline{v_3}, y_2 \equiv V_5$, etc.). A term of the form*

$$(4.23) \qquad \frac{1}{n^l} \sum_{\mathscr{I}_{l_1,l_2}}^{*} \widehat{G_{x_1y_1}^z G_{x_2y_2}^z} \cdots \widehat{G_{x_dy_d}^z} = \frac{1}{n^l} \sum_{\mathscr{I}_{l_1,l_2}}^{*} \prod_{i=1}^{d} \widehat{G_{x_iy_i}^z}, \qquad l = l_1 + l_2,$$

of degree d with a concretely specified assignment is denoted by P_d. The collection of the terms of the form in (4.23) with degree d is denoted by $\mathscr{P}_d$.

Given a term $P_d \in \mathscr{P}_d$ in (4.23), the local law in (4.3) yields a naive bound using power counting, *i.e.,* for any $P_d \in \mathscr{P}_d$,

$$(4.24) \qquad |P_d| \prec \Psi^d, \qquad n^{-1/4+\epsilon} \leq \Psi = (n\eta)^{-1} \leq n^{-\epsilon}.$$

We now give the formal definition of the (un)matched terms of the form in (4.23).

DEFINITION 4.4 ((Un)matched terms in $\mathscr{P}_d$). *Given a term $P_d \in \mathscr{P}_d$ in (4.23), we say that a lower case index $v_j \in \mathscr{I}_{l_1,l_2}$ is matched if the number of assignments of v_j and its*

conjugate $\overline{v_j}$ to a row index in the product agrees with their number of assignments to a column index, i.e.,

$$(4.25) \qquad \#\{i : x_i \equiv v_j\} + \#\{i : x_i \equiv \overline{v_j}\} = \#\{i : y_i \equiv v_j\} + \#\{i : y_i \equiv \overline{v_j}\}.$$

Otherwise, we say that v_j is an unmatched index. For instance, looking at the two terms,

$$\frac{1}{n^2} \sum_{a,b}^{*} \widehat{G_{ab}}\widehat{G_{ab}}\widehat{G_{aa}}\widehat{G_{BB}}, \qquad\qquad \frac{1}{n^2} \sum_{a,b}^{*} \widehat{G_{ab}}\widehat{G_{ba}}\widehat{G_{aa}}\widehat{G_{bb}},$$

both a and b are unmatched indices in the first term, while they are matched indices in the second term.

Similarly, we say that an upper case index $V_j \in \mathscr{I}_{l_1,l_2}$ is matched if

$$(4.26) \qquad \#\{i : x_i \equiv V_j\} + \#\{i : x_i \equiv \underline{V_j}\} = \#\{i : y_i \equiv V_j\} + \#\{i : y_i \equiv \underline{V_j}\}.$$

Otherwise, V_j is an unmatched index.

If all the summation indices in $\mathscr{I}_{l_1,l_2}$ are matched, then P_d is a matched term. Otherwise, if there exists at least one unmatched index, P_d is an unmatched term. If a term P_d is unmatched, we indicate this fact by denoting it by P_d^o. The collection of the unmatched terms of the form in (4.23) with degree d is denoted by $\mathscr{P}_d^o \subset \mathscr{P}_d$.

From Definition 4.4, the terms on the right side of (4.21) with $v \neq a \neq \underline{B}$ are indeed unmatched terms of the form in (4.23), where both the index a and B are unmatched while the index v is matched. Moreover, we give additional examples of unmatched terms below

$$(4.27) \qquad \frac{1}{n^3} \sum_{v,a,B}^{*} \boldsymbol{E}[\widehat{G_{v\bar{a}}^z}\widehat{G_{\underline{B}v}^z}], \qquad \frac{1}{n^2} \sum_{a,B}^{*} \boldsymbol{E}[\widehat{G_{\bar{a}a}^z}\widehat{G_{BB}^z}\widehat{G_{a\underline{B}}^z}], \qquad \frac{1}{n^2} \sum_{a,B}^{*} \boldsymbol{E}[\widehat{G_{B\bar{a}}^z}\widehat{G_{B\bar{a}}^z}\widehat{G_{aB}^z}].$$

PROPOSITION 4.5. *Given an unmatched term P_d^o of the form in (4.23) with fixed $d \geq 1$, we have*

$$\boldsymbol{E}[P_d^o] = O_{\prec}(n^{-3/2}).$$

REMARK 4.6. *The above estimate is much smaller than the naive size in (4.24) either when d is small, say $1 \leq d \leq 5$, or when η is close to $n^{-1+\epsilon}$. For a general unmatched term P_d^o, the estimate $O_{\prec}(n^{-3/2})$ is sharp due to some matched terms of order $n^{-3/2}$ stemming from third order terms in the cumulant expansions with an index coincidence; see (4.57) below.*

REMARK 4.7. *The statement of Proposition 4.5 holds true even when the parameters z of the shifted Green function entries in the product in (4.23) have different values. We also remark that the proof of Proposition 4.5 is not sensitive to the fact that m^z given in (2.19) is small, in fact the argument works as long as $m^z = O(1)$.*

The rest of Section 4 is devoted to proving Proposition 4.5. The proof relies on iterative cumulant expansions for the unmatched indices in products of resolvents. Before we dive into the formal proof of Proposition 4.5, we start with expanding a concrete example of unmatched term to explain the one-step improvement mechanism (essentially gaining an additional small factor Ψ) in Section 4.2. The reader experienced with cumulant expansions may skip Section 4.2. In Section 4.3, we state in Lemma 4.8 the full version of the improvement mechanism for a general unmatched term, and subsequently use Lemma 4.8 iteratively to prove Proposition 4.5. Finally we present the complete proof of Lemma 4.8 for a general unmatched term in Section 4.4.

4.2. *Expansion mechanism: an example.* In this subsection we consider a concrete example of an unmatched term in (4.23) with degree three and $x_1 \equiv a, y_1 \equiv B, x_2 \equiv B, y_2 \equiv a, x_3 \equiv \bar{a}, y_3 \equiv B$, *i.e.*,

$$(4.28) \qquad P_3^o = \frac{1}{n^2} \sum_{a,B}^{*} \widehat{G_{aB}^z} \widehat{G_{Ba}^z} \widehat{G_{\bar{a}B}^z},$$

whose naive estimate is $O_\prec(\Psi^3)$ from (4.24). We will show how to improve this naive estimate using cumulant expansions essentially by an additional small factor Ψ.

For the term P_3^o in (4.28) with an unmatched index a which appears twice as a row and once as a column in the product of resolvents, we aim to expand using the unmatched $x_1 \equiv a$ to show that

$$(4.29) \qquad \boldsymbol{E}[P_3^o] = O_\prec\big(\Psi^4 + n^{-1}\Psi + n^{-1/2}\Psi^3 + n^{-3/2}\big).$$

We will see that the terms that contribute the first three error terms with Ψ-factors in (4.29) still have unmatched indices. So we can continue expanding these unmatched terms to get an arbitrary number of Ψ-improvements and ending up with the final estimate $O_\prec(n^{-3/2})$ given in Proposition 4.5. The corresponding iteration scheme will be presented directly in full generality for any unmatched term in (4.23) in the next subsection.

Recall the following identity from [25, Eq. (5.2)]

$$(4.30) \qquad \widehat{G^z} = -M^z \underline{WG^z} + \langle \widehat{G^z} \rangle M^z G^z,$$

where $M^z = M^z(\mathrm{i}\eta)$ is the deterministic matrix given in (2.16) and $\langle G^z \rangle$ is given in (4.10). The underline notation $\underline{WG^z}$ is defined as follows. For a function $f(W)$ of the random matrix W given in (4.8), we define

$$(4.31) \qquad \underline{Wf(W)} := Wf(W) - \widetilde{\boldsymbol{E}}\widetilde{W}(\partial_{\widetilde{W}} f)(W),$$

where $\widetilde{W}$ is an independent of W defined as in (4.8) with X_t being replaced by a complex Ginibre ensemble. Here $\partial_{\widetilde{W}}$ denoted the directional derivative in the direction $\widetilde{W}$, the expectation in (4.31) is with respect to this matrix.

Applying the identity in (4.30) on the first Green function entry $\widehat{G_{aB}^z}$ in (4.28) and performing cumulant expansion formula on the resulting $\underline{WG^z}$ given in (4.31), we have

$$\boldsymbol{E}[P_3^o] = -\frac{m^z}{n^3} \sum_{a,B}^{*} \sum_{J} \boldsymbol{E}\Big[\frac{\partial \widehat{G_{Ba}^z} \widehat{G_{\bar{a}B}^z}}{\partial w_{Ja}} G_{JB}^z\Big] + \frac{m^z}{n^2} \sum_{a,B}^{*} \boldsymbol{E}\Big[G_{aB}^z \widehat{G_{Ba}^z} \widehat{G_{\bar{a}B}^z} \langle \widehat{G^z} \rangle\Big]$$

$$- \frac{\mathfrak{m}^z}{n^3} \sum_{a,B}^{*} \sum_{j} \boldsymbol{E}\Big[\frac{\partial \widehat{G_{Ba}^z} \widehat{G_{\bar{a}B}^z}}{\partial w_{j\bar{a}}} G_{jB}^z\Big] + \frac{\mathfrak{m}^z}{n^2} \sum_{a,B}^{*} \boldsymbol{E}\Big[G_{\bar{a}B}^z \widehat{G_{Ba}^z} \widehat{G_{\bar{a}B}^z} \langle \widehat{G^z} \rangle\Big]$$

$$- \frac{m^z}{n^2} \sum_{p+q+1 \geq 3} \frac{c^{(p+1,q)}}{p!q!n^{\frac{p+q+1}{2}}} \Big(\sum_{a,B}^{*} \sum_{J} \boldsymbol{E}\Big[\frac{\partial^{p+q} \widehat{G_{Ba}^z} \widehat{G_{\bar{a}B}^z} G_{JB}^z}{\partial w_{aJ}^p \partial w_{Ja}^q}\Big]\Big)$$

$$(4.32) \qquad - \frac{\mathfrak{m}^z}{n^2} \sum_{p+q+1 \geq 3} \frac{c^{(q,p+1)}}{p!q!n^{\frac{p+q+1}{2}}} \Big(\sum_{a,B}^{*} \sum_{j} \boldsymbol{E}\Big[\frac{\partial^{p+q} \widehat{G_{Ba}^z} \widehat{G_{\bar{a}B}^z} G_{jB}^z}{\partial w_{\bar{a}j}^p \partial w_{j\bar{a}}^q}\Big]\Big),$$

with $c^{(p,q)}$ being the (p,q)-th cumulants of the normalized i.i.d. entries $\sqrt{n}w_{aB}$. We first look at the third order terms with $p + q + 1 = 3$ in (4.32). By direct computations using the differentiation rule in (4.13), since J or j is a fresh index appearing three times, the number of resulting (shifted) off-diagonal entries remains at least d with unmatched J or j. From the

local law in (4.3), these third order terms can be bounded by $O_{\prec}(n^{-1/2}\Psi^3 + n^{-3/2})$, where the last error $n^{-3/2}$ stems from the existence of index coincidences, *e.g.,* $J = B$ or $j = \underline{B}$. Similarly, the fourth order terms with $p+q+1=4$ can be bounded by $O_{\prec}(n^{-1}\Psi^3 + n^{-2})$ and note that the index J or j could be matched (for $p=1, q=2$), however the index a remains unmatched. We can truncate the expansion at the fourth order with an error $O_{\prec}(n^{-3/2})$ using (4.3). Thus the higher order terms (*i.e.,* the last two lines of (4.32)) can be bounded by

$$(4.33) \qquad -\frac{m^z}{n^2} \sum_{p+q+1 \geq 3} (\cdots) - \frac{\mathfrak{m}^z}{n^2} \sum_{p+q+1 \geq 3} (\cdots) = O_{\prec}(n^{-1/2}\Psi^3 + n^{-3/2}).$$

We next focus on the second order terms, *i.e.,* the first two lines of (4.32). We start with the first term on the right side of (4.32). After direct computations, we split the summation over the fresh index $J \in [\![n+1, 2n]\!]$ into the following three cases, *i.e.,*

$$-\frac{m^z}{n^3} \sum_{a,B}^{*} \sum_{J} \boldsymbol{E}\left[\frac{\partial \widehat{G^z_{Ba}} \widehat{G^z_{\bar{a}B}}}{\partial w_{Ja}} G^z_{JB}\right] = \frac{m^z}{n^3} \sum_{a,B}^{*} \sum_{J} \boldsymbol{E}\left[\left(G^z_{BJ} G^z_{aa} \widehat{G^z_{\bar{a}B}} + \widehat{G^z_{Ba}} G^z_{\bar{a}J} G^z_{aB}\right) G^z_{JB}\right]$$

$$(4.34) \qquad =: \left(\frac{m^z}{n^3} \sum_{a,B,J}^{*} + \frac{m^z}{n^3} \sum_{a,B}^{*} \sum_{J} \delta_{J\bar{a}} + \frac{m^z}{n^3} \sum_{a,B}^{*} \sum_{J} \delta_{JB}\right) \boldsymbol{E}[(\cdots)].$$

We first consider the last two cases with index coincidence $J = \bar{a}$ or $J = B$. We will create diagonal entries with $J = \bar{a}$ or $J = B$ and as the result the number of off-diagonal entries will be reduced to at least one. Using (4.3), the last two cases in (4.34) can be bounded by

$$(4.35) \qquad \frac{m^z}{n^3} \sum_{a,B}^{*} \sum_{J} \delta_{J\bar{a}} \boldsymbol{E}[(\cdots)] + \frac{m^z}{n^3} \sum_{a,B}^{*} \sum_{J} \delta_{JB} \boldsymbol{E}[(\cdots)] = O_{\prec}(n^{-1}\Psi).$$

We remark that we did not use the smallness of m^z given in (4.4), since this smallness is not essential for the Ψ-improvement. For the first case in (4.34) with $a \neq \underline{B} \neq \underline{J}$, we transform the resulting terms into the form in (4.23), *i.e.,* write the Green function entries with their shifted versions using (4.20). In particular, the diagonal entry G^z_{aa}, from acting ∂w_{Ja} on $\widehat{G^z_{Ba}}$, will be replaced with $m^z + \widehat{G^z_{aa}}$. Then

$$\frac{m^z}{n^3} \sum_{a,B,J}^{*} \boldsymbol{E}[(\cdots)] = \frac{(m^z)^2}{n^3} \sum_{a,B,J}^{*} \boldsymbol{E}\left[\widehat{G^z_{JB}} \widehat{G^z_{BJ}} \widehat{G^z_{\bar{a}B}}\right]$$

$$(4.36) \qquad + \frac{m^z}{n^3} \sum_{a,B,J}^{*} \boldsymbol{E}\left[\widehat{G^z_{BJ}} \widehat{G^z_{aa}} \widehat{G^z_{\bar{a}B}} \widehat{G^z_{JB}}\right] + \frac{m^z}{n^3} \sum_{a,B,J}^{*} \boldsymbol{E}\left[\widehat{G^z_{Ba}} \widehat{G^z_{\bar{a}J}} \widehat{G^z_{aB}} \widehat{G^z_{JB}}\right].$$

We note that the terms in the second line of (4.36) have degree being increased to four to accommodate a pair of the fresh index J, hence can be bounded by $O_{\prec}(\Psi^4)$ from (4.24). Therefore, combining (4.34)-(4.36) the first term on the right side of (4.32) can be estimated as

(4.37)
$$-\frac{m^z}{n^3} \sum_{a,B}^{*} \sum_{J} \boldsymbol{E}\left[\frac{\partial \widehat{G^z_{Ba}} \widehat{G^z_{\bar{a}B}}}{\partial w_{Ja}} G^z_{JB}\right] = \frac{(m^z)^2}{n^3} \sum_{a,B,J}^{*} \boldsymbol{E}\left[\widehat{G^z_{JB}} \widehat{G^z_{BJ}} \widehat{G^z_{\bar{a}B}}\right] + O_{\prec}(\Psi^4 + n^{-1}\Psi),$$

where the first error Ψ^4 is from the second order terms with higher degrees and $n^{-1}\Psi$ is from the second order terms with the index coincidences in (4.35).

The third term on the right side of (4.32) can be estimated similarly, *i.e.,*

$$-\frac{\mathfrak{m}^z}{n^3}\sum_{a,B}^{*}\sum_{j}\boldsymbol{E}\left[\frac{\partial\widehat{G_{Ba}^z}\widehat{G_{\bar{a}B}^z}}{\partial w_{j\bar{a}}}G_{jB}^z\right]=\frac{\mathfrak{m}^z}{n^3}\sum_{a,B}^{*}\sum_{j}\boldsymbol{E}\left[(G_{Bj}^zG_{\bar{a}a}^z\widehat{G_{\bar{a}B}^z}+\widehat{G_{Ba}^z}G_{\bar{a}j}^zG_{\bar{a}B}^z)G_{jB}^z\right]$$

(4.38)
$$=\frac{|\mathfrak{m}^z|^2}{n^3}\sum_{a,B,j}^{*}\boldsymbol{E}\left[\widehat{G_{jB}^z}\widehat{G_{Bj}^z}\widehat{G_{\bar{a}B}^z}\right]+O_{\prec}(\Psi^4+n^{-1}\Psi).$$

It remains to estimate the second and fourth term on the right side of (4.32). Since we restrict to $a\neq\underline{B}$ (equivalent to $\bar{a}\neq B$) in the summation $\sum_{a,B}^{*}$, we write $G_{aB}^z=\widehat{G_{aB}^z}$ and $G_{\bar{a}B}^z=\widehat{G_{\bar{a}B}^z}$. We also write out $\langle\widehat{G^z}\rangle$ as in (4.10) and clearly these two terms are of the form in (4.23) with degree increased to four and the index a remains unmatched. In particular, these two terms gain additional Ψ-factor from $\langle\widehat{G^z}\rangle$ and thus can be bounded by $O_{\prec}(\Psi^4)$.

Combining (4.32), (4.33), (4.37) and (4.38), we conclude

$$\boldsymbol{E}[P_3^o]=\frac{(m^z)^2}{n^3}\sum_{a,B,J}^{*}\boldsymbol{E}\left[\widehat{G_{JB}^z}\widehat{G_{BJ}^z}\widehat{G_{\bar{a}B}^z}\right]+\frac{|\mathfrak{m}^z|^2}{n^3}\sum_{a,B,j}^{*}\boldsymbol{E}\left[\widehat{G_{jB}^z}\widehat{G_{Bj}^z}\widehat{G_{\bar{a}B}^z}\right]$$

(4.39)
$$+O_{\prec}(\Psi^4+n^{-1}\Psi+n^{-1/2}\Psi^3+n^{-3/2}),$$

where the first two errors are from the second order terms with higher degrees (*e.g.,* in the second line of (4.36)) and with the index coincidences (*e.g.,* with $J=\bar{a}$ or $J=B$ in (4.35)), respectively, and the last two errors are from the higher order terms in (4.33). Most importantly, for these leading terms of degree three appearing on the first line of (4.39), we have replaced one pair of the index a of the original term P_3^o in (4.28) with a fresh index J or j. We now introduce a notation for such index replacement, *i.e.,* if

$$P_3^o=\frac{1}{n^2}\sum_{a,B}^{*}\widehat{G_{aB}^z}\widehat{G_{Ba}^z}\widehat{G_{\bar{a}B}^z},$$

which is a term of the form in (4.23) with $x_1\equiv a,y_1\equiv B,x_2\equiv B,y_2\equiv a,x_3\equiv\bar{a},y_3\equiv B$, then we define

(4.40)
$$P_3^o(x_1,y_2\to J):=\frac{1}{n^3}\sum_{a,B,J}^{*}\widehat{G_{JB}^z}\widehat{G_{BJ}^z}\widehat{G_{\bar{a}B}^z};\qquad P_3^o(x_1,y_2\to j):=\frac{1}{n^3}\sum_{a,B,j}^{*}\widehat{G_{jB}^z}\widehat{G_{Bj}^z}\widehat{G_{\bar{a}B}^z},$$

where j and J are 'symbolic' lower and upper case letters indicating the range of the new summation index. Using these notations, the expansion in (4.39) can be written for short as

$$\boldsymbol{E}[P_3^o]=(m^z)^2\,\boldsymbol{E}[P_3^o(x_1,y_2\to J)]+|\mathfrak{m}^z|^2\,\boldsymbol{E}[P_3^o(x_1,y_2\to j)]$$

(4.41)
$$+O_{\prec}(\Psi^4+n^{-1}\Psi+n^{-1/2}\Psi^3+n^{-3/2}).$$

Notice that in the two explicit third order terms the number of assignments of the unmatched index a after replacement has been reduced by two to one, in fact it appears as its conjugation $\bar{a}$ with $x_3\equiv\bar{a}$; see (4.40). The good news is that the index a (in fact $\bar{a}$) remains unmatched, thus we can further expand these leading terms using $x_3\equiv\bar{a}$ to gain the Ψ-improvement.

We will look at only the second leading term on the right side of (4.41), and the first one can be estimated similarly (actually more easily if we take $m^z=O(\Psi)$ into consideration).

Omitting the factor $|\mathfrak{m}^z|^2 \sim 1$ and expanding the Green function entry $\widehat{G^z_{\bar{a}B}}$, we obtain as in (4.32),

$$\boldsymbol{E}[P^o_3(x_1, y_2 \to j)] = \frac{1}{n^3} \sum^*_{a,B,j} \boldsymbol{E}\left[\widehat{G^z_{jB}}\widehat{G^z_{Bj}}\widehat{G^z_{\bar{a}B}}\right]$$

$$= -\frac{m^z}{n^4} \sum^*_{a,B,j} \sum_{j'} \boldsymbol{E}\left[\frac{\partial \widehat{G^z_{Bj}}\widehat{G^z_{jB}}}{\partial w_{j'\bar{a}}} G^z_{j'B}\right] + \frac{m^z}{n^3} \sum^*_{a,B,j} \boldsymbol{E}\left[G^z_{\bar{a}B}\widehat{G^z_{Bj}}\widehat{G^z_{jB}}\langle\widehat{G^z}\rangle\right]$$

$$- \frac{\overline{\mathfrak{m}^z}}{n^4} \sum^*_{a,B,j} \sum_{J'} \boldsymbol{E}\left[\frac{\partial \widehat{G^z_{Bj}}\widehat{G^z_{jB}}}{\partial w_{J'a}} G^z_{J'B}\right] + \frac{\overline{\mathfrak{m}^z}}{n^3} \sum^*_{a,B,j} \boldsymbol{E}\left[G^z_{aB}\widehat{G^z_{Bj}}\widehat{G^z_{jB}}\langle\widehat{G^z}\rangle\right]$$

$$(4.42) \qquad\qquad + O_\prec(n^{-1/2}\Psi^3 + n^{-3/2}).$$

where the last error is from higher order terms as in (4.33). Since the index a or its conjugate $\bar{a}$ no longer appears in the remaining product of Green function entries, we gain additional Ψ from one more off-diagonal Green function entry or a factor $\langle\widehat{G^z}\rangle$ on the right side of (4.42), plus an error $O_\prec(n^{-1}\Psi^2)$ from the index coincidences, *e.g.*, $J' = B$ or $J' = \bar{j}$. Therefore, since the number of assignments of the unmatched index a after replacement has been reduced to one, we obtain the improved estimate

$$(4.43) \qquad \boldsymbol{E}[P^o_3(x_1, y_2 \to j)] = O_\prec(\Psi^4 + n^{-1}\Psi^2 + n^{-1/2}\Psi^3 + n^{-3/2}).$$

The same upper bound also applies to $\boldsymbol{E}[P^o_3(x_1, y_2 \to J)]$.

Therefore, combining (4.41) and (4.43), we have improved the naive estimate (4.24) of $\boldsymbol{E}[P^o_3]$ to the better bound in (4.29).

4.3. *Expansion mechanism: general case and proof of Proposition 4.5.* Given any unmatched term $P^o_d \in \mathscr{P}^o_d$ in (4.23), from Definition 4.4, there must exist a lower case index $v_j \in \mathscr{I}_{l_1,l_2}$ or an upper case index $V_j \in \mathscr{I}_{l_1,l_2}$ such that this index (or its conjugation) is assigned to more row indices of Green function entries in the product than column indices. For notational simplicity, we may denote this special unmatched index by $a \in [\![1,n]\!]$ and $B \in [\![n+1, 2n]\!]$, respectively.

We will first consider the formal case with an unmatched index $a \in [\![1,n]\!]$ satisying

$$(4.44) \qquad \#\{i : x_i \equiv a\} + \#\{i : x_i \equiv \bar{a}\} > \#\{i : y_i \equiv a\} + \#\{i : y_i \equiv \bar{a}\},$$

and the latter case with $B \in [\![n+1, 2n]\!]$ will follow similarly. Then there exists an off-diagonal Green function entry $G_{x_i y_i}$ with $x_i \equiv a$ and $y_i \not\equiv a, \bar{a}$. Without loss of generality we may assume that this is the first Green function factor, i.e. we set $x_1 \equiv a$ and $y_1 \not\equiv a, \bar{a}$. We will denote this term by $P^o_d(x_1 \equiv a)$ to emphasize that we will expand it using the unmatched index $x_1 \equiv a$. Then we have the following estimate whose proof will be given in the next subsection:

LEMMA 4.8. *Let $P^o_d \in \mathscr{P}^o_d$ be a given term with an unmatched index a satisfying (4.44) and without loss of generality assigned to x_1, i.e., $P^o_d = P^o_d(x_1 \equiv a)$. Let*

$$k^{(r)}_a := \#\{i : x_i \equiv a, \bar{a}\}; \qquad k^{(c)}_a := \#\{i : y_i \equiv a, \bar{a}\}$$

denote the number of $a/\bar{a}$-assignments as a row or a column index of the Green function entries, respectively, such that $k^{(r)}_a > k^{(c)}_a$. Then there exist finite (bounded by some constant depending only on d) subsets

$$(4.45) \qquad \mathscr{A}^o_d \subset \mathscr{P}^o_d, \quad \mathscr{A}^o_{>d} \subset \mathscr{P}^o_{d+1}, \quad \mathscr{B}^o_{\geq d} \subset \bigcup_{d' \geq d} \mathscr{P}^o_{d'}, \quad \mathscr{C}^o_{\geq d-2} \subset \bigcup_{d' \geq d-2} \mathscr{P}^o_{d'}$$

with the property that the number of $a/\bar{a}$-assignments as a row or column index in all elements of $\mathscr{A}_d^o$ is reduced to $k_a^{(r)} - 1$ and $k_a^{(c)} - 1$, respectively, so that we have the bound

$$\left| \boldsymbol{E}[P_d^o(x_1 \equiv a)] \right| \lesssim \sum_{P_{d'}^o \in \mathscr{A}_d^o} \left| \boldsymbol{E}[P_{d'}^o] \right| + \sum_{P_{d'}^o \in \mathscr{A}_{>d}^o} \left| \boldsymbol{E}[P_{d'}^o] \right|$$

(4.46)
$$+ \frac{1}{\sqrt{n}} \sum_{P_{d'}^o \in \mathscr{B}_{\geq d}^o} \left| \boldsymbol{E}[P_{d'}^o] \right| + \frac{1}{n} \sum_{P_{d'}^o \in \mathscr{C}_{\geq d-2}^o} \left| \boldsymbol{E}[P_{d'}^o] \right| + O_\prec(n^{-3/2}),$$

here d' denotes a degree compatible with (4.45). In particular, if $k_a^{(c)} = 0$, then $\mathscr{A}_d^o$ is an empty set.

The precise structure of the terms in the rhs. of (4.46) is irrelevant, hence we do not follow them explicitly, we will only need a few properties. Note that all terms in the rhs. of (4.46) remain unmatched; this key feature will allow us to iterate this estimate. We now briefly explain the origin and the main features of each sums and show that every term in the rhs. is "better" in a certain sense than the initial term.

The set $\mathscr{A}_d^o$ contains four types of terms (if exist) of degree d obtained by index replacements defined in (4.40). They can be written explicitly as

$$(m^z)^2 \sum_{i \geq 2:y_i \equiv a} \boldsymbol{E}\left[P_d^o(x_1, y_i \to J)\right] + m^z \mathfrak{m}^z \sum_{i \geq 2:y_i \equiv \bar{a}} \boldsymbol{E}\left[P_d^o(x_1, y_i \to J)\right]$$

(4.47)
$$+ m^z \mathfrak{m}^z \sum_{i \geq 2:y_i \equiv \bar{a}} \boldsymbol{E}\left[P_d^o(x_1, y_i \to j)\right] + |\mathfrak{m}^z|^2 \sum_{i \geq 2:y_i \equiv a} \boldsymbol{E}\left[P_d^o(x_1, y_i \to j)\right],$$

although the only important fact is that the number of $a/\bar{a}$-indices is reduced by two (*i.e.,* one from the row and one from the column) compared with the initial term $P_d^o(x_1 \equiv a)$. These are the generalisations of the first two terms in the rhs. of (4.39) for the concrete example.

The set $\mathscr{A}_{>d}^o$ corresponds to the second order terms with higher degrees, *e.g.,* in the second line in (4.36); their degree is increased by one compared to the original term.

The set $\mathscr{B}_{\geq d}^o$ comes from the third order cumulant expansion, indicated by the additional $1/\sqrt{n}$ prefactor (see the last two lines of (4.32) with $p + q + 1 = 3$). The number of off-diagonal Green function entries remains at least d and we gained $1/\sqrt{n}$ from the third order cumulants.

Finally, the set $\mathscr{C}_{\geq d-2}^o$ coming with a prefactor $1/n$ has two very different sources. On the one hand, it comes from the fourth order cumulant expansion carrying an extra $1/n$ and the number of off-diagonal Green function entries remains at least d. On the other hand, in the second order cumulant expansion the fresh index J or j may coincide with an old index creating a diagonal term. Each diagonal term has to be re-written, e.g., as $G_{aa}^z = m^z + \widehat{G}_{aa}^z$, and thus the term carrying m^z "loses" a G-factor. Thus the degree may be reduced by two from these diagonal elements; see *e.g.,* (4.34) with $J = B$. In this case the $1/n$ comes from the reduced number of summation indices.

For definiteness, we stated and explained Lemma 4.8 for the lower case index a, the modifications for the upper case index B are very minor. In the latter case we may set $x_1 \equiv B, y_1 \not\equiv B, \underline{B}$ and denote the term by $P_d^o(x_1 \equiv B)$. This term can be expanded using the unmatched index $x_1 \equiv B$. The abstract bound (4.46), with the index a replaced with B, remains unchanged, only the (irrelevant) explicit formula changes: J and j are interchanged within both lines of (4.47).

We are now ready to prove Proposition 4.5 by iteratively invoking Lemma 4.8 for an unmatched lower case index $a \in [\![1, n]\!]$ and the analogous formula for $B \in [\![n + 1, 2n]\!]$. The

proof in fact relies on iterations on two different levels: the first level uses Lemma 4.8 to gain one Ψ-factor improvement as explained in the previous Section 4.2; the second level is to iterate this one-step Ψ-improvement to an arbitrary power of Ψ until it becomes negligible and only the $O_{\prec}(n^{-3/2})$ error survives.

PROOF OF PROPOSITION 4.5. Given an unmatched term P_d^o in (4.23), without loss of generality, we may assume that there exists an unmatched index $a \in [\![1, n]\!]$ satisfying the assignment condition in (4.44) and denote this term by $P_d^o(x_1 \equiv a)$. The case with $B \in [\![n+1, 2n]\!]$ follows similarly.

We define the number of the assignments of the index a and $\bar{a}$ to a row and column index of the Green function entries in the product, *i.e.,*

$$(4.48) \qquad k_0^{(r)} = \#\{i : x_i \equiv a\} + \#\{i : x_i \equiv \bar{a}\}, \quad k_0^{(c)} = \#\{i : y_i \equiv a\} + \#\{i : y_i \equiv \bar{a}\},$$

with $k_0^{(r)} > k_0^{(c)}$ from (4.44), here we use the subscript 0 to indicate this quantity is applied to the original term before iterations.

Applying Lemma 4.8, if $k_0^{(c)} = 0$, then the first type of subset $\mathscr{A}_d^o$ in (4.45) is empty. However if $k_0^{(c)} \geq 1$, we need to repeatedly invoke Lemma 4.8 to eliminate resulting terms in non-empty $\mathscr{A}_d^o$. This is our first-level iteration procedure. In the first step, using Lemma 4.8, we have

$$\left| \boldsymbol{E}[P_d^o(x_1 \equiv a)] \right| \lesssim \sum_{P_{d'}^o \in \mathscr{A}_{d,1}^o} \left| \boldsymbol{E}[P_{d'}^o] \right| + \sum_{P_{d'}^o \in \mathscr{A}_{>d,1}^o} \left| \boldsymbol{E}[P_{d'}^o] \right|$$

$$(4.49) \qquad\qquad + \frac{1}{\sqrt{n}} \sum_{P_{d'}^o \in \mathscr{B}_{\geq d,1}^o} \left| \boldsymbol{E}[P_{d'}^o] \right| + \frac{1}{n} \sum_{P_{d'}^o \in \mathscr{C}_{\geq d-2,1}^o} \left| \boldsymbol{E}[P_{d'}^o] \right| + O_{\prec}(n^{-3/2}),$$

where we use the subscript 1 in the four types of subsets to indicate the iteration step, and each term in the first subset $\mathscr{A}_{d,1}^o$ still has the unmatched index a satisfying (4.44), with *c.f.,* (4.48),

$$k_1^{(r)} = k_0^{(r)} - 1, \qquad k_1^{(c)} = k_0^{(c)} - 1, \qquad k_1^{(r)} > k_1^{(c)}.$$

Hence we can further apply Lemma 4.8 on these leading terms of degree d in $\mathscr{A}_{d,1}^o$. In general, in the s-th iteration step, we have

$$\left| \boldsymbol{E}[P_d^o(x_1 \equiv a)] \right| \lesssim \sum_{P_{d'}^o \in \mathscr{A}_{d,s}^o} \left| \boldsymbol{E}[P_{d'}^o] \right| + \sum_{P_{d'}^o \in \mathscr{A}_{>d,s}^o} \left| \boldsymbol{E}[P_{d'}^o] \right|$$

$$(4.50) \qquad\qquad + \frac{1}{\sqrt{n}} \sum_{P_{d'}^o \in \mathscr{B}_{\geq d,s}^o} \left| \boldsymbol{E}[P_{d'}^o] \right| + \frac{1}{n} \sum_{P_{d'}^o \in \mathscr{C}_{\geq d-2,s}^o} \left| \boldsymbol{E}[P_{d'}^o] \right| + O_{\prec}(n^{-3/2}),$$

where each term in the first subset $\mathscr{A}_{d,s}^o$ (if exists) satisfies

$$k_s^{(r)} = k_0^{(r)} - s, \qquad k_s^{(c)} = k_0^{(c)} - s, \qquad k_s^{(r)} > k_s^{(c)}.$$

We stop the iterations at step $s = k_0^{(c)} + 1$ so that the resulting subset $\mathscr{A}_{d,s}^o$ is empty, we hence obtain the following estimate for $P_d^o = P_d^o(x_1 \equiv a)$;

$$\left| \boldsymbol{E}[P_d^o] \right| \lesssim \sum_{P_{d'}^o \in \mathscr{A}_{>d,*}^o} \left| \boldsymbol{E}[P_{d'}^o] \right| + \frac{1}{\sqrt{n}} \sum_{P_{d'}^o \in \mathscr{B}_{\geq d,*}^o} \left| \boldsymbol{E}[P_{d'}^o] \right|$$

$$(4.51) \qquad\qquad\qquad\qquad + \frac{1}{n} \sum_{P_{d'}^o \in \mathscr{C}_{\geq d-2,*}^o} \left| \boldsymbol{E}[P_{d'}^o] \right| + O_{\prec}(n^{-3/2}),$$

where the three subsets, $\mathscr{A}^o_{>d,*}, \mathscr{B}^o_{\geq d,*}$ and $\mathscr{C}^o_{\geq d-2,*}$ are defined as the union of the corresponding subsets in (4.45) generated in the above s-iterations. Their precise form is irrelevant, beyond their degree what matters is that their cardinality can be bounded by some constant only depending on d.

In this way, we have improved the estimate essentially by an additional small factor $\Psi = (n\eta)^{-1}$ from (4.24), *i.e.*, the first group of terms on the right side of (4.51) has degree at least $d + 1$, while the remaining terms gain extra $n^{-1/2}$ or n^{-1} from the prefactors. The above iteration procedure generalizes the Ψ-improvement mechanism explained in the previous subsection for a concrete example. Moreover, we obtain a similar formula in (4.51) for $P^o_d = P^o_d(x_1 \equiv B)$ using the analogous version of Lemma 4.8 for the upper case index B.

Next, we will perform our second-level iteration, *i.e.*, iterating the Ψ-improvement mechanism stated in (4.51) to increase the degree further. We note that the resulting terms on the right side of (4.51) remain unmatched either with one Ψ-improvement from $\mathscr{A}^o_{>d,*}$ (*i.e.*, with higher degrees) or with the gain from the prefactor $1/\sqrt{n}$ or $1/n$. Iterating (4.51) for $D - d$ times with a large fixed $D > 0$ chosen later, we have

$$\big| \boldsymbol{E}[P^o_d] \big| \lesssim \sum_{P^o_{d'} \in \mathscr{A}^o_{>D,*}} \big| \boldsymbol{E}[P^o_{d'}] \big| + \frac{1}{\sqrt{n}} \sum_{P^o_{d'} \in \mathscr{B}^o_{\geq D,*}} \big| \boldsymbol{E}[P^o_{d'}] \big|$$

$$(4.52) \qquad\qquad\qquad + \frac{1}{n} \sum_{P^o_{d'} \in \mathscr{C}^o_{\geq D-2,*}} \big| \boldsymbol{E}[P^o_{d'}] \big| + O_\prec(n^{-3/2}),$$

where the sets $\mathscr{A}^o_{>D,*}$, $\mathscr{B}^o_{\geq D,*}$ and $\mathscr{C}^o_{\geq D-2,*}$ denote the union of the corresponding sets in (4.51) generated in the second-level iterations, whose cardinality can be bounded by a constant depending only on d and D. Using the naive estimate in (4.24), we have

$$(4.53) \qquad \boldsymbol{E}[P^o_d] = O_\prec\big(\Psi^D + n^{-1/2}\Psi^{D-1} + n^{-1}\Psi^{D-3} + n^{-3/2}\big) = O_\prec(n^{-3/2} + \Psi^D).$$

For any $\eta \geq n^{-1+\epsilon}$ with a fixed small $\epsilon > 0$, we choose $D = \lfloor 2/\epsilon \rfloor$ so that $\Psi^D \lesssim n^{-3/2}$. In particular, if we choose $\eta = \eta_0 = n^{-7/8-\tau}$ (in fact, used to prove Lemma 3.5), we can choose smaller $D = \lceil \frac{12}{1-8\tau} \rceil$ so that $\Psi^D \lesssim n^{-3/2}$. This completes the proof of Proposition 4.5. $\qquad\square$

4.4. *Proof of Lemma 4.8.* Let $P^o_d(x_1 \equiv a)$ be a given term in $\mathscr{P}^o_d$ with an unmatched index a satisfying (4.44) and without loss of generality $x_1 \equiv a$, $y_1 \neq a, \bar{a}$. Using the identity in (4.30) on the first Green function factor $\widehat{G^z_{ay_1}}$ and performing the cumulant expansions as in (4.32), we have

$$\boldsymbol{E}[P^o_d(x_1 \equiv a)] = -\frac{\mathfrak{m}^z}{n^{l+1}} \sum_{\mathscr{I}_{l_1,l_2}}^{*} \sum_J \boldsymbol{E}\left[\frac{\partial \prod_{i=2}^d \widehat{G^z_{x_i y_i}}}{\partial w_{Ja}} G^z_{Jy_1} \right] + \frac{\mathfrak{m}^z}{n^l} \sum_{\mathscr{I}_{l_1,l_2}}^{*} \boldsymbol{E}\left[G^z_{ay_1} \prod_{i=2}^d \widehat{G^z_{x_i y_i}} \langle \widehat{G^z} \rangle \right]$$

$$- \frac{\mathfrak{m}^z}{n^{l+1}} \sum_{\mathscr{I}_{l_1,l_2}}^{*} \sum_j \boldsymbol{E}\left[\frac{\partial \prod_{i=2}^d \widehat{G^z_{x_i y_i}}}{\partial w_{j\bar{a}}} G^z_{jy_1} \right] + \frac{\mathfrak{m}^z}{n^l} \sum_{\mathscr{I}_{l_1,l_2}}^{*} \boldsymbol{E}\left[G^z_{\bar{a}y_1} \prod_{i=2}^d \widehat{G^z_{x_i y_i}} \langle \widehat{G^z} \rangle \right]$$

$$(4.54) \qquad + \sum_{p+q+1 \geq 3} \left(H^{(1)}_{p+1,q} + H^{(2)}_{q,p+1} \right),$$

where $H^{(1)}_{p+1,q}$ and $H^{(2)}_{q,p+1}$ are the higher order terms given by

$$H^{(1)}_{p+1,q} := -\frac{\mathfrak{m}^z c^{(p+1,q)}}{p!q!n^{\frac{p+q+1}{2}+l}} \sum_{\mathscr{I}_{l_1,l_2}}^{*} \sum_J \boldsymbol{E}\left[\frac{\partial^{p+q} \prod_{i=2}^d \widehat{G^z_{x_i y_i}} G^z_{Jy_1}}{\partial w^p_{aJ} \partial w^q_{Ja}} \right];$$

$$(4.55) \qquad H_{q,p+1}^{(2)} = -\frac{\mathfrak{m}^z c^{(q,p+1)}}{p!q!n^{\frac{p+q+1}{2}+l}} \sum_{\mathscr{I}_{l_1,l_2}}^{*} \sum_{j} \boldsymbol{E}\left[\frac{\partial^{p+q} \prod_{i=2}^{d} \widehat{G_{x_i y_i}^z} G_{jy_1}^z}{\partial w_{\bar{a}j}^p \partial w_{j\bar{a}}^q}\right],$$

with $c^{(p,q)}$ the (p,q)-th cumulants of the normalized i.i.d. entries $\sqrt{n} w_{aB}$. We will only estimate $H_{p+1,q}^{(1)}$ in (4.55) and $H_{p+1,q}^{(2)}$ can be treated similarly. We remark that the smallness of m^z in (4.4) will not be used in the proof. We now split of $H_{p+1,q}^{(1)}$ in (4.55) into the following two parts, *i.e.*, J is distinct from or coinciding with the old indices in $\mathscr{I}_{l_1,l_2} = \{v_k\}_{k=1}^{l_1} \cup \{V_k\}_{k=1}^{l_2}$ (omitting the irrelevant prefactor $c^{(p+1,q)}/p!q!$),

$$H_{p+1,q}^{(1)} = -\frac{m^z}{n^{\frac{p+q+1}{2}+l}} \sum_{\mathscr{I}_{l_1,l_2},J}^{*} \boldsymbol{E}[(\cdots)] - \frac{m^z}{n^{\frac{p+q+1}{2}+l}} \sum_{\mathscr{I}_{l_1,l_2}}^{*} \sum_{J} \left(\sum_{k=1}^{l_1} \delta_{J\overline{v_k}} + \sum_{k=1}^{l_2} \delta_{JV_k}\right) \boldsymbol{E}[(\cdots)]$$

$$(4.56)$$

$$=: -\frac{m^z}{n^{\frac{p+q-1}{2}}} \left(\frac{1}{n^{l+1}} \sum_{\mathscr{I}_{l_1,l_2+1}}^{*} \boldsymbol{E}[(\cdots)]\right) - \frac{m^z}{n^{\frac{p+q+1}{2}}} \left(\frac{1}{n^l} \sum_{\mathscr{I}_{l_1,l_2}}^{*} \sum_{J} \delta_{J\in\mathscr{I}_{l_1,l_2}} \boldsymbol{E}[(\cdots)]\right),$$

where the notation $\sum^{*}$ is given in (4.22) indicating that all the summation indices are distinct, and we use the short hand notation $\delta_{J\in\mathscr{I}_{l_1,l_2}}$ to indicate the part with index coincidence.

We first estimate the second part in (4.56) with index coincidences $J \in \mathscr{I}_{l_1,l_2}$. Using that $|G_{\mathfrak{uv}}| \prec 1$ from (4.3) naively, we gain an additional n^{-1} from the summation and have

$$(4.57) \qquad \left|\frac{m^z}{n^{\frac{p+q+1}{2}}} \left(\frac{1}{n^l} \sum_{\mathscr{I}_{l_1,l_2}}^{*} \sum_{J} \delta_{J\in\mathscr{I}_{l_1,l_2}} \boldsymbol{E}[(\cdots)]\right)\right| = O_{\prec}(n^{-\frac{p+q+1}{2}}).$$

We remark that for $p+q+1 = 3$, the error $n^{-3/2}$ is sharp in general. By setting $J = \overline{v_k}$ or $J = V_k$, the terms in (4.55) might switch to matched terms; see *e.g.,* the last two lines of (4.32) with $J = B$ or $j = \underline{B}$.

Next, the first part in (4.56) with distinct summation indices can be written as a linear combination of averaged products of shifted Green function entries of the form in (4.23) with an additional factor $n^{-\frac{p+q-1}{2}}$. Since J is a fresh index, the number of (shifted) off-diagonal Green function entries remains at least d in the product. If $q \neq p+1$, then from (4.13) and Definition 4.4, the fresh index J becomes an unmatched index. Otherwise if $q = p+1$, the index J is matched, but the index a remains unmatched using (4.13) and Definition 4.4. Thus the first part of (4.56) yields a collection of unmatched terms of the form in (4.23) with degrees at least d and with an additional factor $n^{-\frac{p+q-1}{2}}$. Similar estimates also apply to $H_{p+1,q}^{(2)}$ in (4.55).

Therefore for the third and fourth order terms with $p+q+1 = 3, 4$ in (4.55), we denote by $\mathscr{B}_{\geq d}^o \subset \bigcup_{d' \geq d} \mathscr{P}_{d'}^o$ and $\mathscr{C}_{\geq d}^o \subset \bigcup_{d' \geq d} \mathscr{P}_{d'}^o$, respectively, the set of the resulting unmatched terms of degrees at least d. With these notations and combining with (4.57), we write for short that

$$\sum_{p+q+1=3} \left(H_{p+1,q}^{(1)} + H_{q,p+1}^{(2)}\right) = \frac{1}{\sqrt{n}} \sum_{P_{d'}^o \in \mathscr{B}_{\geq d}^o} \boldsymbol{E}[P_{d'}^o] + O_{\prec}(n^{-3/2});$$

$$(4.58) \qquad \sum_{p+q+1=4} \left(H_{p+1,q}^{(1)} + H_{q,p+1}^{(2)}\right) = \frac{1}{n} \sum_{P_{d'}^o \in \mathscr{C}_{\geq d}^o} \boldsymbol{E}[P_{d'}^o] + O_{\prec}(n^{-2}),$$

and we truncate the cumulant expansion at the fourth order with an error $O_{\prec}(n^{-3/2})$ using (4.3).

We next estimate the second order terms, *i.e.,* the first two lines of (4.54). Writing $\langle \widehat{G^z} \rangle$ as in (4.10) and $G^z_{ay_1} = \widehat{G^z_{ay_1}}$ and $G^z_{\bar{a}y_1} = \widehat{G^z_{\bar{a}y_1}}$ since $y_1 \not\equiv a$ and $\bar{a}$, the second and fourth term on the right side of (4.54) are of the form in (4.23) and the degrees of these terms are increased to $d+1$. For the first term on the right side of (4.54), we split the summation into two parts as in (4.56). By direct computations, the part with index coincidences $J \in \mathscr{I}_{l_1,l_2}$ is given by

$$-\frac{m^z}{n^{l+1}} \sum_{\mathscr{I}_{l_1,l_2}}^* \boldsymbol{E}\Big[\frac{\partial \prod_{i=2}^d \widehat{G^z_{x_iy_i}}}{\partial w_{Ja}} G^z_{Jy_1} \delta_{J \in \mathscr{I}_{l_1,l_2}} \Big] = \frac{m^z}{n^{l+1}} \sum_{\mathscr{I}_{l_1,l_2}}^* \boldsymbol{E}\Big[\sum_{p=2}^d \big(\prod_{i=2}^{(p)} \widehat{G^z_{x_iy_i}} \big) G^z_{x_pJ} G^z_{ay_p} G^z_{Jy_1} \delta_{J \in \mathscr{I}_{l_1,l_2}} \Big].$$

If we set $J = \bar{a}$, then the index a remains unmatched since the index J appeared once as a row and once as a column of the Green function entries in the product. Otherwise if $J \in \mathscr{I}_{l_1,l_2} \setminus \{a\}$, then the index a obviously remains unmatched. After transforming the Green function entries into their shifted versions using *e.g.,* $G^z_{aa} = m^z + \widehat{G^z}_{aa}$, the degrees of the resulting terms might be decreased to $d-2$ when all the entries $G^z_{x_pJ}, G^z_{ay_p}, G^z_{Jy_1}$ are diagonal; see (4.35) for a concrete example. Thus we obtain a collection of unmatched terms of degrees at least $d-2$ with a factor n^{-1} from the index coincidence. Together with the subset $\mathscr{C}^o_{\geq d}$ in (4.58) with the same prefactor $1/n$ from the fourth order cumulant expansion, we denote by $\mathscr{C}^o_{\geq d-2}$ the union of these unmatched terms with degrees at least $d-2$, *i.e.,* we write them together for short as

(4.59)
$$\frac{1}{n} \sum_{P^o_{d'} \in \mathscr{C}^o_{\geq d-2}} \boldsymbol{E}[P^o_{d'}].$$

For the remaining summation with J distinct from the indices in $\mathscr{I}_{l_1,l_2}$, writing $G^z_{Jy_1} = \widehat{G^z_{Jy_1}}$ and $G^z_{x_pJ} = \widehat{G^z_{x_pJ}}$, we have

$$-\frac{m^z}{n^{l+1}} \sum_{\mathscr{I}_{l_1,l_2},J}^* \boldsymbol{E}\Big[\frac{\partial \prod_{i=2}^d \widehat{G^z_{x_iy_i}}}{\partial w_{Ja}} G^z_{Jy_1} \Big] = \frac{m^z}{n^{l+1}} \sum_{\mathscr{I}_{l_1,l_2},J}^* \boldsymbol{E}\Big[\sum_{p=2}^d \big(\prod_{i=2}^{(p)} \widehat{G^z_{x_iy_i}} \big) \widehat{G^z_{x_pJ}} G^z_{ay_p} \widehat{G^z_{Jy_1}} \Big].$$

If $y_p \not\equiv a, \bar{a}$, then $G^z_{ay_p}$ from acting ∂w_{Ja} on G_{ay_p} is an extra off-diagonal entry and the degree is thus increased to $d+1$. Otherwise if there exists some $y_p \equiv a$ or $\bar{a}$, then the resulting diagonal entry G_{aa} or $G_{a\bar{a}}$ which will be replaced with the deterministic function m^z or $\mathfrak{m}^z$. In both cases, the index a remains unmatched. Then we have

(4.60)
$$-\frac{m^z}{n^{l+1}} \sum_{\mathscr{I}_{l_1,l_2},J}^* \boldsymbol{E}\Big[\frac{\partial \prod_{i=2}^d \widehat{G^z_{x_iy_i}}}{\partial w_{Ja}} G^z_{Jy_1} \Big] = (m^z)^2 \sum_{i \geq 2: y_i \equiv a} \boldsymbol{E}[P^o_d(x_1, y_i \to J)]$$
$$+ m^z \mathfrak{m}^z \sum_{i \geq 2: y_i \equiv \bar{a}} \boldsymbol{E}[P^o_d(x_1, y_i \to J)] + \sum_{P^o_{d'} \in \mathscr{P}^o_{d+1}} \boldsymbol{E}[P^o_{d'}],$$

where $P^o_d(x_1, y_i \to J)$ given in (4.40) denotes a term obtained from the original term P^o_d with the row index $x_1 \equiv a$ and column index $y_i \equiv a$ or $\bar{a}$ of the Green function entries being replaced with a fresh (averaged) summation index J, and with a slight abuse of notations, the last sum denotes a specific linear combination of unmatched terms with higher degree $d+1$.

Similarly we estimate the third term on the right side of (4.54). For the cases with index coincidences $\delta_{j \in \mathscr{I}_{l_1,l_2}}$, we obtain unmatched terms with degree at least $d-2$ and with a factor n^{-1} which will be added to (4.59). For the cases with distinct summation indices, we have

c.f., (4.60)

$$-\frac{\mathfrak{m}^z}{n^{l+1}} \sum_{\mathscr{I}_{l_1,l_2},j}^{*} \boldsymbol{E}\left[\frac{\partial \prod_{i=2}^{d}\widehat{G_{x_iy_i}^z}}{\partial w_{j\bar{a}}}G_{jy_1}^z\right] = |\mathfrak{m}^z|^2 \sum_{i\geq 2:y_i\equiv a}\boldsymbol{E}[P_d^o(x_1,y_i\to j)]$$

$$(4.61) \qquad\qquad + m^z\mathfrak{m}^z \sum_{i\geq 2:y_i\equiv\bar{a}}\boldsymbol{E}[P_d^o(x_1,y_i\to j)] + \sum_{P_{d'}^o\in\mathscr{P}_{d+1}^o}\boldsymbol{E}[P_{d'}^o].$$

The collection of all the unmatched terms with higher degree $d+1$ in both (4.60) and (4.61) is denoted by $\mathscr{A}_{>d}^o$. Moreover, the collection of the leading terms of degree d (if exists) defined by index replacements in both (4.60) and (4.61) is denoted by $\mathscr{A}_d^o$. We note that, for any term in $\mathscr{A}_d^o$, from the index replacement defined in (4.40), the number of $a/\bar{a}$-assignments as a row and column index of the Green function entries has been reduced by one, respectively.

To sum up, with the above notations, combining (4.54), (4.58), (4.59), (4.60) and (4.61), we have proved (4.46) and hence finish the proof of Lemma 4.8. $\qquad\square$

REMARK 4.9. *Though here we present only the expansions starting from an off-diagonal Green function entry $\widehat{G_{ay_1}^z}$, a similar expansion also holds true if we start from a diagonal entry $\widehat{G_{aa}^z}$. We remark that the above expansion is not unique since it depends on the choice of the Green function entry to start performing expansions. The proof of Proposition 4.5, however, does not rely on the uniqueness of the expansions.*

5. Green function comparison in Girko's formula: Proof of Proposition 3.8. Recall the matrix flow in (4.2). To prove Proposition 3.8, it then suffices to show the following:

LEMMA 5.1. *Set $\eta_0 = n^{-7/8-\tau}$ with a small fixed $\tau > 0$ from (2.5) and $T = n^{100}$. Let $f = f_1^-$ or f_2^+. Then there exists some constant $c > 0$ such that*

$$(5.1) \qquad \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{C}}\Delta f(z)\,\boldsymbol{E}\left[\int_{\eta_0}^{T}\mathrm{Im}\,\mathrm{Tr}\,G_t^z(\mathrm{i}\eta)\,\mathrm{d}\eta\right]\mathrm{d}^2 z = O(n^{-c}),$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{C}}\int_{\mathbb{C}}\Delta f(z_1)\Delta f(z_2)\,\boldsymbol{E}\left[\int_{\eta_0}^{T}\int_{\eta_0}^{T}\Big((1-\boldsymbol{E})\,\mathrm{Im}\,\mathrm{Tr}\,G_t^{z_1}(\mathrm{i}\eta_1)\Big)\times\right.$$

$$(5.2) \qquad \left.\Big((1-\boldsymbol{E})\,\mathrm{Im}\,\mathrm{Tr}\,G_t^{z_2}(\mathrm{i}\eta_2)\Big)\,\mathrm{d}\eta_1\,\mathrm{d}\eta_2\,\mathrm{d}^2 z_1\,\mathrm{d}^2 z_2\right] = O(n^{-c}).$$

PROOF OF PROPOSITION 3.8. Integrating the bounds from Lemma 5.1 over $t\in[0,100\log n]$ and using standard perturbation theory as in (4.7) we conclude the proof of Proposition 3.8. $\qquad\square$

5.1. *Expectation estimate: Proof of (5.1).* We introduce the short-hand notation

$$(5.3) \qquad \mathscr{F}_t^z := \int_{\eta_0}^{T}\mathrm{Im}\,\mathrm{Tr}\,G_t^z(\mathrm{i}\eta)\,\mathrm{d}\eta = -\mathrm{i}\int_{\eta_0}^{T}\mathrm{Tr}\,G_t^z(\mathrm{i}\eta)\,\mathrm{d}\eta,$$

and we will prove a slightly stronger estimate than needed in (5.1), *i.e.,*

$$(5.4) \qquad \left|\int_{\mathbb{C}}\Delta f(z)\Big(\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{E}\left[\mathscr{F}_t^z\right]\Big)\mathrm{d}^2 z\right| = O(n^{-1/4+5\tau}).$$

Using the L^1 norm of Δf in (3.7), it suffices to show

$$(5.5) \qquad \left| \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{E}\left[\mathscr{F}_t^z\right] \right| = O_\prec(n^{-1/2+4\tau}).$$

Recall the matrix flow in (4.2). Using Ito's formula and performing the cumulant expansion on the expectation of the drift term, we obtain the analogue of (4.11),

$$\frac{\mathrm{d}\boldsymbol{E}\left[\mathscr{F}_t^z\right]}{\mathrm{d}t} = -\frac{1}{2} \sum_{a=1}^n \sum_{B=n+1}^{2n} \left(\sum_{p+q+1=3}^{K_0} \frac{c^{(p+1,q)}}{p!q!n^{\frac{p+q+1}{2}}} \boldsymbol{E}\left[\frac{\partial^{p+q+1}\mathscr{F}_t^z}{\partial w_{aB}^{p+1}\partial \overline{w}_{aB}^q} \right] \right)$$

$$(5.6) \qquad -\frac{1}{2} \sum_{a=1}^n \sum_{B=n+1}^{2n} \left(\sum_{p+q+1=3}^{K_0} \frac{c^{(q,p+1)}}{p!q!n^{\frac{p+q+1}{2}}} \boldsymbol{E}\left[\frac{\partial^{p+q+1}\mathscr{F}_t^z}{\partial \overline{w}_{aB}^{p+1}\partial w_{aB}^q} \right] \right) + O_\prec(n^{-\frac{K_0}{2}+2}),$$

with $K_0 = 100$ and $c^{(p,q)}$ the (p,q)-th cumulants of the normalized i.i.d. entries $\sqrt{n}w_{aB}$. It then suffices to consider the first line of (5.6) to show

$$(5.7)$$

$$\sum_{p+q+1=3}^{K_0} K_{p+1,q}^z := \sum_{p+q+1=3}^{K_0} \frac{c^{(p+1,q)}}{2p!q!} \left(\frac{1}{n^{\frac{p+q+1}{2}}} \sum_{a,B} \boldsymbol{E}\left[\frac{\partial^{p+q+1}\mathscr{F}_t^z}{\partial w_{aB}^{p+1}\partial \overline{w}_{aB}^q} \right] \right) = O_\prec(n^{-1/2+4\tau}),$$

and the second line of (5.6) is the same with a and B interchanged.

Recall the differentiation rule in (4.13). We further have

$$\frac{\partial \mathscr{F}_t^z}{\partial w_{aB}} = \mathrm{i} \int_{\eta_0}^T \sum_{\mathfrak{v}=1}^{2n} \left(G_{\mathfrak{v}a}^z(\mathrm{i}\eta)G_{B\mathfrak{v}}^z(\mathrm{i}\eta) \right) \mathrm{d}\eta = \mathrm{i} \int_{\eta_0}^T \left((G^z)^2(\mathrm{i}\eta) \right)_{Ba} \mathrm{d}\eta$$

$$(5.8) \qquad\qquad = G_{Ba}^z(\mathrm{i}T) - G_{Ba}^z(\mathrm{i}\eta_0) = -G_{Ba}^z(\mathrm{i}\eta_0) + O(n^{-100}),$$

where we used that $(G^2)(\mathrm{i}\eta) = -\mathrm{i}\frac{\mathrm{d}G(\mathrm{i}\eta)}{\mathrm{d}\eta}$, the deterministic norm bound $\|G(\mathrm{i}T)\| \le T^{-1}$ with $T = n^{100}$. By direct computations using (4.13) and (5.8), each term $K_{p+1,q}^z$ given in (5.7) is a linear combination of products of $p+q+1$ Green function entries of the following form

$$(5.9) \qquad \frac{1}{n^{\frac{p+q+1}{2}}} \sum_{a,B} \boldsymbol{E}\left[\prod_{i=1}^{p+q+1} G_{x_i,y_i}^z(\mathrm{i}\eta_0) + O(n^{-100}) \right] = O_\prec\left(n^{-\frac{p+q-3}{2}}(\Psi^{p+q+1} + n^{-1}) \right)$$

where $x_i, y_i \equiv a$ or B satisfying the assignment condition in (4.15), and the last estimate follows from the local law in (4.3) and (4.4) with $\Psi = (n\eta_0)^{-1} = n^{-1/8+\tau}$. We remark that the error term n^{-1} is from the cases with index coincidence, *i.e.*, $a = \underline{B}$. In particular, we have from (5.9) that

$$(5.10) \qquad |K_{p+1,q}^z| \prec n^{-1/2+4\tau}, \qquad p+q+1 \ge 4,$$

which is enough to prove (5.5) except for the third order terms.

We next improve the estimate for these third order terms in (5.7) with $p+q+1=3$. Transforming the Green function entries in these terms to their shifted versions by (4.20), these third order terms with the summation restriction $a \ne \underline{B}$ are of the form in (4.23) with a factor $n^{1/2}$ and with unmatched indices a and B from Definition 4.4. Using Proposition 4.5, these unmatched terms with a factor $\sqrt{n}$ can be bounded by $O_\prec(n^{-1})$. For the remaining summation with the index coincidence $a = \underline{B}$, they can be bounded by $O_\prec(n^{-1/2})$ using that $|G_{\mathfrak{u}\mathfrak{v}}| \prec 1$ from (4.3). Therefore, the third order terms in (5.7) can be bounded by

$$(5.11) \qquad \sum_{p+q+1=3} K_{p+1,q}^z = O_\prec(n^{-1/2}).$$

Combining (5.6), (5.10) and (5.11), we have proved the expectation estimate in (5.5).

5.2. *Variance estimate: Proof of (5.2).* We start with the short-hand notation, $j = 1, 2$

$$(5.12) \qquad \widehat{\mathscr{F}_t^{z_j}} := \mathscr{F}_t^{z_j} - \boldsymbol{E}[\mathscr{F}_t^{z_j}] = -\mathrm{i} \int_{\eta_0}^T \Big(\operatorname{Tr} G_t^{z_j}(\mathrm{i}\eta) - \boldsymbol{E}\left[\operatorname{Tr} G_t^{z_j}(\mathrm{i}\eta) \right] \Big) \mathrm{d}\eta \prec 1,$$

where the last estimate follows from the local law in (4.3). We will prove a slightly stronger estimate than needed in (5.2), i.e. we will prove

$$(5.13) \qquad \int_{\mathbb{C}} \int_{\mathbb{C}} \Delta f(z_1) \Delta f(z_2) \Big(\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{E}\left[\widehat{\mathscr{F}_t^{z_1}} \widehat{\mathscr{F}_t^{z_2}} \right] \Big) \mathrm{d}^2 z_1 \, \mathrm{d}^2 z_2 = O(n^{-1/8+3\tau}).$$

Using the L^1 norm of Δf in (3.7), it suffices to prove that

$$(5.14) \qquad \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{E}\left[\widehat{\mathscr{F}_t^{z_1}} \widehat{\mathscr{F}_t^{z_2}} \right] = \text{M-terms}(z_1, z_2) + O(n^{-5/8+\tau}),$$

where M-terms(z_1, z_2) is a deterministic function satisfying the following integral condition

$$(5.15) \qquad \left| \int_{\mathbb{C}} \int_{\mathbb{C}} \Delta f(z_1) \Delta f(z_2) \, \text{M-terms}(z_1, z_2) \, \mathrm{d}^2 z_1 \, \mathrm{d}^2 z_2 \right| \ll n^{-1/8+3\tau}.$$

Now we focus on proving (5.14). Using Ito's formula and performing the cumulant expansion on the drift term, we obtain the analogue of (5.6)

$$\frac{\mathrm{d} \boldsymbol{E}\left[\widehat{\mathscr{F}_t^{z_1}} \widehat{\mathscr{F}_t^{z_2}} \right]}{\mathrm{d}t} = -\frac{1}{2} \sum_{a=1}^n \sum_{B=n+1}^{2n} \left(\sum_{p+q+1=3}^{K_0} \frac{c^{(p+1,q)}}{p!q!n^{\frac{p+q+1}{2}}} \boldsymbol{E}\left[\frac{\partial^{p+q+1} \widehat{\mathscr{F}_t^{z_1}} \widehat{\mathscr{F}_t^{z_2}}}{\partial w_{aB}^{p+1} \partial \overline{w_{aB}}^q} \right] \right)$$

$$(5.16) \qquad\qquad -\frac{1}{2} \sum_{a=1}^n \sum_{B=n+1}^{2n} \left(\sum_{p+q+1=3}^{K_0} \frac{c^{(q,p+1)}}{p!q!n^{\frac{p+q+1}{2}}} \boldsymbol{E}\left[\frac{\partial^{p+q+1} \widehat{\mathscr{F}_t^{z_1}} \widehat{\mathscr{F}_t^{z_2}}}{\partial \overline{w_{aB}}^{p+1} \partial w_{aB}^q} \right] \right) + O_\prec(n^{-\frac{K_0}{2}+2}),$$

with $K_0 = 100$ and $c^{(p,q)}$ the (p, q)-th cumulants of the normalized i.i.d. entries $\sqrt{n} w_{aB}$. It then suffices to consider the first line of (5.16) to show

$$(5.17) \qquad \sum_{p+q+1=3}^{K_0} \mathscr{K}_{p+1,q}^{z_1,z_2} := \sum_{p+q+1=3}^{K_0} \frac{c^{(p+1,q)}}{2p!q!} \left(\frac{1}{n^{\frac{p+q+1}{2}}} \sum_{a,B} \boldsymbol{E}\left[\frac{\partial^{p+q+1} \widehat{\mathscr{F}_t^{z_1}} \widehat{\mathscr{F}_t^{z_2}}}{\partial w_{aB}^{p+1} \partial \overline{w_{aB}}^q} \right] \right),$$

and the second line of (5.16) is the same with a and B interchanged.

Using the differentiation rules in (4.13) and (5.8), each term $\mathscr{K}_{p+1,q}^{z_1,z_2}$ in (5.17) is a linear combination of products of $p + q + 1$ Green function entries (either G^{z_1} or G^{z_2}) with a possible factor $\widehat{\mathscr{F}_t^{z_1}}$ or $\widehat{\mathscr{F}_t^{z_2}}$, i.e., in the following general form

$$(5.18) \qquad \frac{1}{n^{\frac{p+q+1}{2}}} \sum_{a,B} \boldsymbol{E}\left[\big(\widehat{\mathscr{F}_t^{z^{(0)}}} \big) \prod_{i=1}^{p+q+1} G_{x_i,y_i}^{z^{(i)}}(\mathrm{i}\eta_0) + O(n^{-100}) \right],$$

with $\{z^{(i)} \in \mathbb{C}\}_{i=0}^{p+q+1}$ being either z_1 or z_2, and $x_i, y_i \equiv a$ or B satisfying the assignment condition in (4.15). Using the local law in (4.3), (4.4) and that $|\mathscr{F}_t^z| \prec 1$, we have the following naive bound

$$(5.19) \qquad |\mathscr{K}_{p+1,q}^{z_1,z_2}| = O_\prec\big(n^{-\frac{p+q-3}{2}} (\Psi^{p+q+1} + n^{-1}) \big), \qquad \Psi = n^{-1/8+\tau},$$

where the error n^{-1} is from the cases with index coincidence, *i.e.,* $a = \underline{B}$. In particular we have

$$(5.20) \qquad |\mathscr{K}_{p+1,q}^{z_1,z_2}| \prec n^{-9/8+5\tau}, \qquad p + q + 1 \geq 5,$$

which is sufficiently small to prove (5.14) except for the third and fourth order terms with $p+q+1 = 3, 4$.

We next estimate the third and fourth order terms more carefully. In the following, we will drop the argument $i\eta_0$ with $\eta_0 = n^{-7/8-\tau}$ for notational simplicity and ignoring the error $O(n^{-100})$ in (5.18).

5.2.1. *Third order terms.* By direct computations using (4.13) and (5.8), the third order terms $\mathscr{K}_{p+1,q}^{z_1,z_2}$ in (5.17) with $p+q+1 = 3$ are given by linear combinations of the following terms (ignoring the irrelevant $c^{(p,q)}$ coefficients)

$$\frac{1}{n^{3/2}} \sum_{a,B} \boldsymbol{E}\left[\widehat{\mathscr{F}_t^{z_2}} G_{aa}^{z_1} G_{BB}^{z_1} G_{aB}^{z_1} \right], \qquad \frac{1}{n^{3/2}} \sum_{a,B} \boldsymbol{E}\left[\widehat{\mathscr{F}_t^{z_2}} G_{aB}^{z_1} G_{aB}^{z_1} G_{aB}^{z_1} \right],$$

(5.21)
$$\frac{1}{n^{3/2}} \sum_{a,B} \boldsymbol{E}\left[G_{aB}^{z_1} G_{aa}^{z_2} G_{BB}^{z_2} \right], \quad \frac{1}{n^{3/2}} \sum_{a,B} \boldsymbol{E}\left[G_{aB}^{z_1} G_{aB}^{z_2} G_{aB}^{z_2} \right], \quad \frac{1}{n^{3/2}} \sum_{a,B} \boldsymbol{E}\left[G_{aB}^{z_1} G_{Ba}^{z_2} G_{Ba}^{z_2} \right],$$

together with the other terms by interchanging a with B and z_1 with z_2.

We first consider the terms in (5.21) with the index coincidence $B = \bar{a}$ in the summations. The resulting terms except from the last two terms in (5.21) can be bounded by $O_\prec(n^{-1/2}\Psi) = O_\prec(n^{-5/8+\tau})$ using the local law in (4.3), (4.4) and that $\widehat{\mathscr{F}_t^z}$ is centered. For the last two terms in (5.21) consisting of factors G_{aB} and G_{Ba} only, we have from the local law in (4.3) that

$$\frac{1}{n^{3/2}} \sum_{a=\underline{B}} \boldsymbol{E}\left[G_{aB}^{z_1} G_{aB}^{z_2} G_{aB}^{z_2} \right] = \frac{1}{\sqrt{n}} \mathfrak{m}^{z_1} \left(\mathfrak{m}^{z_2} \right)^2 + O_\prec(n^{-5/8+\tau}),$$

(5.22)
$$\frac{1}{n^{3/2}} \sum_{a=\underline{B}} \boldsymbol{E}\left[G_{aB}^{z_1} G_{Ba}^{z_2} G_{Ba}^{z_2} \right] = \frac{1}{\sqrt{n}} \mathfrak{m}^{z_1} \left(\overline{\mathfrak{m}^{z_2}} \right)^2 + O_\prec(n^{-5/8+\tau}),$$

where the leading deterministic functions can only be bounded by $O(n^{-1/2})$. However from (2.18) and (2.19), for any $z \in \text{supp}(f_1^-) \cup \text{supp}(f_2^+)$, we have

(5.23) $\quad \mathfrak{m}^z = -z + \dfrac{z\eta}{\eta + \operatorname{Im} \mathfrak{m}^z} = -z + O\left(|z|(|1-|z|^2| + \eta^{2/3}) \right) = -z + O(n^{-1/2+\tau}).$

Hence the leading deterministic terms in (5.22) satisfy the intergral condition in (5.15), *i.e.*, for simplicity, we only consider the first term in (5.22),

$$\int_{\mathbb{C}} \int_{\mathbb{C}} \Delta f(z_1) \Delta f(z_2) \left(\frac{1}{\sqrt{n}} \mathfrak{m}^{z_1} (\mathfrak{m}^{z_2})^2 \right) d^2 z_1 \, d^2 z_2$$

$$= \frac{1}{\sqrt{n}} \left(\int_{\mathbb{C}} \Delta f(z_1) \left(-z_1 + O(n^{-1/2+\tau}) \right) d^2 z_1 \right) \left(\int_{\mathbb{C}} \Delta f(z_2) \left((z_2)^2 + O(n^{-1/2+\tau}) \right) d^2 z_2 \right)$$

(5.24)
$$= O(n^{-1+4\tau}),$$

where we used (5.23), the L^1 norm of Δf in (3.7), and that both z and $(z)^2$ are harmonic functions.

Next we study the remaining summations with the restriction $B \neq \bar{a}$ in (5.21). In contrast to the form of averaged products of shifted Green function entries in (4.23), we introduce a

slightly different abstract form to adapt this notation to the terms in (5.21), *i.e.,* with a possibe factor $\widehat{\mathscr{F}_t^{z_1}}$ or $\widehat{\mathscr{F}_t^{z_2}}$,

$$(5.25) \qquad (\widehat{\mathscr{F}_t^{z^{(0)}}})\frac{1}{n^l} \sum_{\mathscr{I}_{l_1,l_2}}^{*} \prod_{i=1}^{d} \widehat{G_{x_i,y_i}^{z^{(i)}}}, \qquad l = l_1 + l_2,$$

with each $\{z^{(i)} \in \mathbb{C}\}_{i=0}^{p+q+1}$ being either z_1 or z_2, where the restricted sum $\sum_{\mathscr{I}_{l_1,l_2}}^{*}$ is defined in (4.22), and we assign a summation index v_j or $V_j \in \mathscr{I}_{l_1,l_2}$ or their conjugates $\overline{v_j}, V_j$ to each row index x_i and column index y_i of the shifted Green function entries in the product. We also define unmatched indices and unmatched terms of the form in (5.25) as in Definition 4.4. Since the proof of Proposition 4.5 is not sensitive to the modifications in the abstract form, the statement still holds true for the general form in (5.25). We omit the proof details.

Provided the assignment condition in (4.15) with $p + q + 1 = 3$, all the third order terms in (5.21) with restricted summations $B \neq \bar{a}$ can be tranformed by (4.20) to linear combinations of unmatched terms of the form in (5.25) with a factor $\sqrt{n}$. Thus by analogous Proposition 4.5 for general unmatched term in (5.25), we have

$$(5.26) \qquad \sum_{p+q+1=3} \mathscr{K}_{p+1,q}^{z_1,z_2}\Big|_{a\neq \underline{B}} = O_{\prec}(n^{-1}).$$

Therefore, combining (5.22) and (5.26), we have

$$(5.27) \qquad \sum_{p+q+1=3} \mathscr{K}_{p+1,q}^{z_1,z_2} = \text{M-terms}(z_1,z_2) + O_{\prec}(n^{-5/8+\tau}),$$

where the function M-terms(z_1, z_2) is a linear combination of leading deterministic functions in (5.22) which satisfy the integral condition in (5.15).

5.2.2. *Fourth order terms.* By direct computations using (4.13) and (5.8), the fourth order terms $\mathscr{K}_{p+1,q}^{z_1,z_2}$ in (5.17) with $p + q + 1 = 4$ are averaged products of Green function entries satisfying the assignment condition in (4.15). From Definition 4.4, these fourth order terms with restricted summations $a \neq \underline{B}$ are unmatched unless $p = 1$ and $q = 2$. Then by Proposition 4.5, we have

$$(5.28) \qquad \sum_{p+q+1=4;p+1\neq q} \mathscr{K}_{p+1,q}^{z_1,z_2} = \sum_{p+q+1=4;p+1\neq q} \mathscr{K}_{p+1,q}^{z_1,z_2}\Big|_{a\neq \underline{B}} + O_{\prec}(n^{-1}) = O_{\prec}(n^{-1}),$$

where the error $O_{\prec}(n^{-1})$ comes from the cases with index coincidence, *i.e.,* $a = \underline{B}$.

We next estimate the remaining term $\mathscr{K}_{2,2}^{z_1,z_2}$ for $p = 1$ and $q = 2$ in (5.17). By direct computations, $\mathscr{K}_{2,2}^{z_1,z_2}$ is a linear combination of the following terms

$$\frac{1}{n^2}\sum_{a,B} \boldsymbol{E}\left[\widehat{\mathscr{F}_t^{z_2}}(G_{BB}^{z_1}G_{aa}^{z_1})^2\right], \qquad \frac{1}{n^2}\sum_{a,B}\boldsymbol{E}\left[\widehat{\mathscr{F}_t^{z_2}}G_{aa}^{z_1}G_{BB}^{z_1}G_{aB}^{z_1}G_{Ba}^{z_1}\right],$$

(5.29)
$$\frac{1}{n^2}\sum_{a,B}\boldsymbol{E}\left[G_{aa}^{z_1}G_{BB}^{z_1}G_{aa}^{z_2}G_{BB}^{z_2}\right], \quad \frac{1}{n^2}\sum_{a,B}\boldsymbol{E}\left[G_{aB}^{z_1}G_{Ba}^{z_1}G_{aa}^{z_2}G_{BB}^{z_2}\right], \quad \frac{1}{n^2}\sum_{a,B}\boldsymbol{E}\left[G_{aB}^{z_1}G_{aB}^{z_1}G_{Ba}^{z_2}G_{Ba}^{z_2}\right]$$

as well as the other terms by interchanging a with B and z_1 with z_2. Those terms in (5.29) containing G_{aa} or G_{BB} in the product of Green function entries can be bounded using the

improved estimate of resolvent in (3.27) and the local law in (4.3) and (4.4), *e.g.,* the first term in (5.29) is bounded by

(5.30)
$$\frac{1}{n^2}\sum_{a,B}\boldsymbol{E}\left[\widehat{\mathscr{F}_t^{z_2}}(G_{BB}^{z_1}G_{aa}^{z_1})^2\right] \prec \Psi^3\frac{1}{n}\sum_a\boldsymbol{E}\left[|G_{aa}^{z_1}|\right] = \Psi^3\,\boldsymbol{E}[\mathrm{Im}\langle G^{z_1}\rangle] = O_{\prec}(n^{-3/4+2\tau}),$$

where we used the estimate in (3.27) with $\eta = n^{-7/8-\tau}$. The last term in (5.29) is bounded similarly using the Ward identity, *i.e.,*

(5.31)
$$\frac{1}{n^2}\sum_{a,B}\boldsymbol{E}\left[G_{aB}^{z_1}G_{aB}^{z_1}G_{Ba}^{z_2}G_{Ba}^{z_2}\right] \prec \frac{1}{n^2}\Psi^2\sum_{a,B}\boldsymbol{E}[|G_{aB}^{z_1}|^2] \leq \Psi^2\frac{\boldsymbol{E}[\mathrm{Im}\langle G^{z_1}\rangle]}{n\eta} = O_{\prec}(n^{-3/4+2\tau}).$$

Combining these bounds with (5.28), we conclude that

(5.32)
$$\sum_{p+q+1=4}\mathscr{K}_{p+1,q}^{z_1,z_2} = O_{\prec}(n^{-3/4+2\tau}).$$

Therefore, using (5.16), (5.20), (5.27) and (5.32), we prove (5.14) and (5.15), hence finish the proof of the variance estimate in (5.2).

APPENDIX: PROOFS OF PROPOSITION 2.7

In this appendix we prove a lower tail bound on the smallest eigenvalue of

$$Y^z := (X - z)^*(X - z),$$

which can also be viewed as the square of the smallest singular value of $X - z$ or as the smallest (in modulus) eigenvalue of H^z, for a standard complex Ginibre matrix X. Recall that the parameter $\delta := |z|^2 - 1$ monitors the distance of z from the unit circle. We point out that in earlier papers [23, 22, 26, 27] we defined δ with an opposite sign (i.e. $\delta := 1 - |z|^2$) because in those works we were primarily focused on the regime where $|z| \leq 1$. Proposition 2.7 in the current paper our focus is on the regime $|z| > 1$ so δ is positive with the new definition.

A simple redefinition of the variable y shows that (2.23) is equivalent to

(A.33)
$$\boldsymbol{P}^{\mathrm{Gin}}\left((\lambda_1^z)^2 \leq \frac{x}{n^2\delta}\right) \lesssim \frac{x}{(n\delta^2)^{2/3}}e^{-n\delta^2(1+O(\delta))/2}, \qquad 0 \leq x \leq C.$$

We point out that the $(n\delta^2)^{-2/3}$ prefactor in (A.33) is not optimal, but it is sufficient for our purposes. To make the presentation clearer here we present only the proof of the simpler version (A.33), while the following remark explains the possible improvements.

REMARK A.2. *First, the bound* (A.33) *should hold all the way up to* $x \leq c(n\delta^2)^2$ *with some small constant c, corresponding to the fact that* (2.23) *should hold up to* $y \leq c$, *i.e. for an entire regime comparable with the gap size of order* δ^3 *in the spectrum of* Y^z. *Second, we can improve the bound* (A.33) *to*

(A.34)
$$\boldsymbol{P}^{\mathrm{Gin}}\left((\lambda_1^z)^2 \leq \frac{x}{n^2\delta}\right) \lesssim \frac{x}{n\delta^2}e^{-n\delta^2(1+O(\delta))/2}, \qquad 0 \leq x \leq \frac{C}{(n\delta^2)^2},$$

by exploiting an extra improvement choosing a different contour along the proof (see Remark A.5 below for a detailed explanation). A simple asymptotic expansion indicates that the bound (A.34) *is actually optimal.*

REMARK A.3.   *In Proposition 2.7 (and in Remark A.2) we presented the bound on $(\lambda_1^z)^2$ for $n^{-1/2} \ll \delta \ll 1$ to make our presentation more concise. However, a similar analysis gives an analogous bound for $\delta \sim n^{-1/2}$ and $\delta \sim 1$ as well (see also [22, Section 5.2] for the case $\delta \sim n^{-1/2}$).*

This rest of this section is devoted to the proof of Proposition 2.7 in the form of (A.33). We present two arguments. Our first proof with all details relies on an explicit formula for the eigenvalue correlation kernel for Y^z from [13]. This approach is fairly elementary but it works only for the complex symmetry class. An alternative method is based upon the supersymmetric (SUSY) representation for the resolvent in [22] which also has a version for the real symmetry class. We sketch the rigorous argument for the simpler complex case and we comment on the considerably more cumbersome details of the real case. Note that (A.33) is formulated for the complex case, the factor x is expected to be replaced with $\sqrt{x} + x \exp\left(-\frac{n}{2}(\operatorname{Im} z)^2\right)$ as it was the case in [22, Corollary 2.4] for the $|z| \leq 1$ regime (see also [27, 26]).

FIRST PROOF OF PROPOSITION 2.7.   By [13, Theorem 7.1] the correlation kernel for Y^z is given by (to make the notations consistent we set the dimension $N \equiv n$)
(A.35)
$$K_n(u,v) = \frac{n^3}{\mathrm{i}\pi} \int_\Gamma \mathrm{d}\zeta \int_\gamma \mathrm{d}w\, e^{n[f(w)-f(\zeta)]} K_B(2n\zeta\sqrt{u}, 2nw\sqrt{v})\zeta w\left(1 - \frac{|z|^2}{(|z|^2 - w^2)(|z|^2 - \zeta^2)}\right),$$

where Γ is any contour symmetric around 0 which encircles $\pm|z|^2$, γ is the imaginary axis positively oriented $0 \to +\infty$, $0 \to -\infty$, and

(A.36)
$$f(w) := w^2 + \log(|z|^2 - w^2).$$

Here, for any $x, y \in \mathbf{C}$, the kernel K_B is defined by

(A.37)
$$K_B(x,y) = \frac{xI_0'(x)I_0(y) - yI_0'(y)I_0(x)}{x^2 - y^2},$$

with $I_0(x)$ being the zeroth modified Bessel function:

(A.38)
$$I_0(x) := \frac{1}{\pi}\int_0^\pi e^{x\cos\theta}\,\mathrm{d}\theta.$$

Note that

$$K_B(x,y) = K_B(x,-y) = K_B(-x,y) = K_B(-x,-y)$$

as a consequence of I_0, I_0' being even and odd functions, respectively.

We are interested in the case when $|z|^2 = 1 + \delta$, with $1 \gg \delta \gg n^{-1/2}$, and $u = v$. In this case the formula (A.35) reduces to

(A.39)
$$K_n(u,u) = \frac{2n^3}{\mathrm{i}\pi} \int_\Gamma \mathrm{d}\zeta \int_{\widehat{\gamma}} \mathrm{d}w\, e^{n[f(w)-f(\zeta)]} K_B(2n\zeta\sqrt{u}, 2nw\sqrt{u})\zeta w\left(1 - \frac{1+\delta}{(1+\delta - w^2)(1+\delta - \zeta^2)}\right),$$
$$f(w) = w^2 + \log(1 + \delta - w^2),$$

with $\widehat{\gamma}$ being the line $0 \to \mathrm{i}\infty$. We point out that here we used the symmetry with respect to the variable w of the integrand in (A.35) to replace the contour γ by $\widehat{\gamma}$.

The main technical result is the following lemma:

LEMMA A.4. *For any $n^{-1/2} \ll \delta \ll 1$ and $u \lesssim 1/(n^2\delta)$ it holds*

(A.40) $$K_n(u,u) \lesssim n^{4/3}\delta^{-1/3}e^{-n\delta^2(1+O(\delta))/2}.$$

Hence, for any $0 \leq x \lesssim 1$ we compute

$$\boldsymbol{P}^{\mathrm{Gin}}\left((\lambda_1^z)^2 \leq \frac{x}{n^2\delta}\right) \leq \int_0^{x/(n^2\delta)} K(\lambda,\lambda)\,\mathrm{d}\lambda \lesssim \frac{x}{(n\delta^2)^{2/3}}e^{-n\delta^2(1+O(\delta))/2},$$

which concludes the proof of (A.33), hence Proposition 2.7. $\qquad\square$

We now conclude this section with the proof of Lemma A.4.

PROOF OF LEMMA A.4. By explicit computations we get

$$f'(w) = 2w\left(1 - \frac{1}{1+\delta-w^2}\right).$$

We thus find that the saddles of f, i.e. the solutions of $f'(w_*) = 0$, are given by $w_* \in \{0, \pm\sqrt{\delta}\}$. Additionally, by Taylor expansion, we get

(A.41) $$f(\zeta) = \delta - \frac{\delta^2}{2} + \delta\zeta^2 - \frac{\zeta^4}{2} + O(\delta^3 + |\zeta|^6).$$

Step 1: Deformation of the contours. We parametrize the $\widehat{\gamma}$-contour as $w = \mathrm{i}s$, with $s \geq 0$, then

$$f(\mathrm{i}s) = -s^2 + \log(1+\delta+s^2).$$

Note that by (A.41) it follows

(A.42) $$\mathrm{Re}[f(\mathrm{i}s)] = f(\mathrm{i}s) = \delta - \frac{\delta^2}{2} - \delta s^2 - \frac{s^4}{2} + O(\delta^3 + s^6).$$

Additionally, simple calculus shows that the map

(A.43) $$s \mapsto \mathrm{Re}\,f(\mathrm{i}s)$$

is decreasing for $s \geq 0$. In particular, this implies $f(\mathrm{i}s) \leq f(0) = \log(1+\delta)$ for any $s \geq 0$.

We choose the contour Γ to consist of two disjoint closed curves around $\sqrt{1+\delta}$ and $-\sqrt{1+\delta}$, respectively. We focus on the contour encircling $\sqrt{1+\delta}$, the other one can be handled in the same way, hence we neglect it from the discussion. Next, we parametrize the part of the Γ-contour lying on the region $\mathrm{Re}\,\zeta > 0$ as $\zeta = \sqrt{\delta} + t \pm \mathrm{i}t$, with $t \geq 0$. The curve may be closed with a circular arc $|\zeta| = R$ with some very large R, this regime of integration is negligible; for practical purposes we consider $R = \infty$. Note that by (A.41) we get

(A.44) $$\mathrm{Re}\,f(\sqrt{\delta} + t \pm \mathrm{i}t) = \delta + 2t^4 + 4\sqrt{\delta}t^3 + O(\delta^3 + t^6).$$

Additionally, by an elementary calculation, we have that

(A.45) $$t \mapsto \mathrm{Re}\,f(\sqrt{\delta} + t \pm \mathrm{i}t) = \delta + 2\sqrt{\delta}t + \frac{1}{2}\log\left[1 - 4\sqrt{\delta}t + 8t^2\delta + 4t^4 + 8\sqrt{\delta}t^3\right]$$

is a strictly increasing function on $t \geq 0$, as a consequence of

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Re}\,f(\sqrt{\delta} + t \pm \mathrm{i}t) \geq 0, \qquad t \geq 0;$$

the equality holds only for $t = 0$.

Finally, along the chosen contours we compute

$$1 - \frac{1+\delta}{(1+\delta-w^2)(1+\delta-\zeta^2)}$$

(A.46)
$$= \frac{-(1+\delta)(2\sqrt{\delta}t \pm 2\mathrm{i}(\sqrt{\delta}+t)t + s^2) - (\delta + 2\sqrt{\delta}t \pm 2\mathrm{i}(\sqrt{\delta}+t)t)s^2}{(1+\delta+s^2)(1 - 2\sqrt{\delta}t - \pm 2\mathrm{i}t(\sqrt{\delta}+t))}.$$

Step 2. Estimates of the integrals along the contours. We split the analysis into four regimes: $(s,t) \in [0, K\sqrt{\delta}]^2$, $(s,t) \in [0, K\sqrt{\delta}] \times (K\sqrt{\delta}, +\infty)$, $(s,t) \in (K\sqrt{\delta}, +\infty) \times [0, K\sqrt{\delta}]$, $(s,t) \in (K\sqrt{\delta}, +\infty)^2$. Here $K > 0$ is a large constant independent of n and δ that we will choose shortly.

Regime $(s,t) \in [0, K\sqrt{\delta}]^2$**:** In this regime we start with the following expansion for the kernel K_B:

(A.47)
$$K_B(x,y) = \frac{1}{2\pi^2} + O(|x|^2 + |y|^2),$$

which follows by standard asymptotic of modified Bessel functions for $|x|, |y| \lesssim 1$.

Since in our regime $u \lesssim 1/(n^2\delta)$ and we have

$$\left| 1 - \frac{1+\delta}{(1+\delta-w^2)(1+\delta-\zeta^2)} \right| \lesssim \sqrt{\delta}t + s^2$$

for $s, t \leq K\sqrt{\delta}$, by (A.46), together with $w = \mathrm{i}s$ and $\zeta = \sqrt{\delta} + t \pm \mathrm{i}t$ we find that for $(s,t) \in [0, K\sqrt{\delta}]^2$ it holds

(A.48)

$$K_n(u,u) \lesssim n^3\delta \left(\int_0^{K\sqrt{\delta}} s e^{n\delta - n\frac{\delta^2 + s^4 + 2\delta s^2}{2} + O(n\delta^3)} \, \mathrm{d}s \right) \left(\int_0^{K\sqrt{\delta}} t e^{-n\delta - 2nt^4 - 4n\sqrt{\delta}t^3 + O(n\delta^3)} \, \mathrm{d}t \right)$$

$$+ n^3\sqrt{\delta}e^{-n\delta} \int_0^{K\sqrt{\delta}} s^3 e^{n\delta - n\frac{\delta^2 + s^4 + 2\delta s^2}{2} + O(n\delta^3)} \, \mathrm{d}s$$

$$\lesssim n^3\delta e^{-n\delta^2(1+O(\delta)/2)} \left(\int_0^{K\sqrt{\delta}} s e^{-n\delta s^2} \, \mathrm{d}s \right) \left(\int_0^{K\sqrt{\delta}} t e^{-n\sqrt{\delta}t^3} \, \mathrm{d}t \right)$$

$$+ n^3\sqrt{\delta} \int_0^{K\sqrt{\delta}} s^3 e^{-n\delta s^2} \, \mathrm{d}s$$

$$\lesssim n^{4/3}\delta^{-1/3} e^{-n\delta^2(1+O(\delta))/2}.$$

where we also used (A.42) and (A.44).

REMARK A.5. *The improved bound (A.34) (compared to (A.33)) can by achieved by choosing the $\widehat{\gamma}$-contour as in Step 1, i.e. $w = \mathrm{i}s$, and the Γ-contour to be any admissible contour which is given by $\zeta = \sqrt{\delta} + t \pm \mathrm{i}ct$, with some $1 < c \leq 2$, for $t \ll \sqrt{\delta}$ and by $\zeta = \sqrt{\delta} + t \pm \mathrm{i}t$ for $t \gg \sqrt{\delta}$. In particular an additional gain of a factor $(n\delta^2)^{-1/3}$ is due to the fact that for this new ζ-contour the expansion in (A.44) is replaced by*

$$\mathrm{Re}\, f(\sqrt{\delta} + t \pm \mathrm{i}ct) = \delta + 2(c^2-1)t^2\delta + t^4\left(3c^2 - \frac{1}{2} - \frac{c^4}{2}\right) + 2(3c^2-1)\sqrt{\delta}t^3 + O(\delta^3 + t^6).$$

In particular, the term $2(c^2-1)t^2\delta$, which non vanishes only for $c > 1$, in the exponent ensures an additional gain $(n\delta^2)^{-1/3}$ compared to (A.48) where we only gained using the smaller (for $t \ll \sqrt{\delta}$) factor $2(3c^2-1)\sqrt{\delta}t^3$.

Regime $(s,t) \in [0, K\sqrt{\delta}] \times (K\sqrt{\delta}, +\infty)$**:** We start with the bound

$$|K_B(x,y)| \lesssim e^{|y|}, \tag{A.49}$$

for $|x| \lesssim 1$. We remark that a similar bound holds for $|y| \lesssim 1$ after replacing y by x in the r.h.s. of (A.49).

Define

$$g(t) := \mathrm{Re}\, f(\sqrt{\delta} + t \pm \mathrm{i}t) - \delta, \tag{A.50}$$

then, by (A.45), it follows that $t \mapsto g(t)$ is strictly increasing on $t \geq 0$. Hence, using (A.43) together with $f(0) = \log(1+\delta)$, we get

$$e^{n[f(w)-f(\zeta)]} \leq e^{-ng(t)} \leq e^{-Kn\delta^2/4} e^{-ng(t)/2},$$

where we used (A.45) to estimate one of the two $e^{-ng(t)/2}$ factors.

Then, using that

$$|1 + \delta - (\sqrt{\delta} + t \pm \mathrm{i}t)^2|^2 \gtrsim 1 - 4\sqrt{\delta}t + 8t^2\delta + 4t^4 + 8\sqrt{\delta}t^3 \geq \frac{1}{2},$$

we readily conclude

$$K_n(u,u) \lesssim n^3 \delta K^2 e^{-Kn\delta^2/4} \int_{K\sqrt{\delta}}^{\infty} e^{-ng(t)/2} e^{t/\sqrt{\delta}} t^3 \, \mathrm{d}t \lesssim e^{-Kn\delta^2/8}, \tag{A.51}$$

where we used (A.49), (A.50), and that

$$\left| \frac{\delta - (\zeta^2 + w^2) + \zeta^2 w^2}{(1+\delta-w^2)(1+\delta-\zeta^2)} \right| \lesssim t^2$$

uniformly in t in this regime. To ensure that the error term in (A.51) is smaller than our goal in (A.40) we choose $K \geq 5$.

Regimes $(s,t) \in (K\sqrt{\delta}, +\infty) \times [0, K\sqrt{\delta}]$ **and** $(s,t) \in (K\sqrt{\delta}, +\infty)^2$**:** Given the bound

$$|K_B(x,y)| \lesssim e^{|x|+|y|},$$

and using the monotonicity properties (A.43),(A.45) of $\mathrm{Re}\, f$ along the contours chosen in Step 1, the analysis of these regimes is analogous to the regime $(s,t) \in [0, K\sqrt{\delta}] \times (K\sqrt{\delta}, +\infty)$ and so omitted. In particular, the contribution of both these regimes is bounded as in (A.51). Combining this fact with (A.48) and (A.51) we conclude the proof of (A.40). $\qquad\square$

Next we sketch the alternative proof relying on SUSY.

SECOND PROOF OF PROPOSITION 2.7. The starting point is the following contour integral representation of the trace of the resolvent of $Y^z = (X - z)^*(X - z)$ for a complex Ginibre matrix X at any spectral parameter $w = E + \mathrm{i}\epsilon$, with $E \in \mathbb{R}$, $\epsilon > 0$, (see [22, Eq. (28)]):

$$\boldsymbol{E}^{\mathrm{Gin}} \mathrm{Tr}\, \frac{1}{Y^z - w} = \frac{n^2}{2\pi\mathrm{i}} \int_0^{\mathrm{i}\infty} \mathrm{d}x \oint \mathrm{d}y\, e^{-nf(x)+nf(y)} y \cdot G(x,y),$$

$$G(x,y) := \frac{1}{xy} - \frac{1}{(1+x)(1+y)}\left[1 + \frac{|z|^2}{1+x} + \frac{|z|^2}{1+y}\right], \tag{A.52}$$

$$f(x) := \log \frac{1+x}{x} - \frac{|z|^2}{1+x} - wx,$$

where the x-integration is over the half imaginary axis and the y-integration is over a positively oriented circle around the origin that does not enclose -1. Since the integrand is analytic away from 0 and -1, the integration contours can be freely deformed away from these two singularities. We need to investigate the imaginary part of (A.52) in the regime where

$$(A.53) \qquad 0 \le E^{1/3} \ll n^{-1/2} \ll \delta = |z|^2 - 1, \qquad \epsilon = +0,$$

to detect the density $\rho(E)$ of eigenvalues $(\lambda^z)^2$ of Y^z at $E \ll n^{-3/2}$ that would directly imply (2.23). Here ϵ is an infinitesimally small positive regularization parameter, its only role is to specify in which direction the x-contour goes out to infinity.

Typically, the large n asymptotics of such contour integral is obtained by saddle point analysis. Both contours are deformed through the saddle point x_* of f, defined by $f'(x_*) = 0$, where a second order Taylor expansion is performed both for f and G and the main contribution comes from the value of these functions and their derivatives at the saddle. The exponential factors cancel and the result is typically polynomial in n. Among others, this strategy is followed in our analysis in [22] for the regime $\delta = |z|^2 - 1 < 0$, where we found that the saddle has a non-zero imaginary part. The current regime (A.53) behaves quite differently since now $E \ll \delta^3$ lies outside of the support of $\frac{1}{\pi} \operatorname{Im} m^z(x + i0)$ (see (2.17)), which implies that the relevant saddle x_* is on the positive real axis, in fact by a simple calculation we have[4]

$$(A.54) \qquad x_* = \delta^{-1}\big[1 + (E/\delta^3) + O(E/\delta^3)^2\big]$$

for the unique positive solution to $f'(x_*) = 0$.

The spectral density at E is given by
$$(A.55)$$
$$\rho(E) := \boldsymbol{E}^{\mathrm{Gin}} \frac{1}{\pi} \operatorname{Im} \operatorname{Tr} \frac{1}{Y^z - E - i0} = \frac{1}{2\pi i} \boldsymbol{E}^{\mathrm{Gin}} \Big[\operatorname{Tr} \frac{1}{Y^z - E - i0} - \operatorname{Tr} \frac{1}{Y^z - E + i0} \Big],$$

i.e. we need to evaluate the difference of two copies of (A.52) with an opposite sign in front of the regularization ϵ. Note that G and large part of f is independent of ϵ, this parameter appears only as a $\pm i\epsilon x$ term in $f(x)$ and is relevant only for the non-compact x-integration as ϵ is infinitesimally small. We deform the x-contour to $\gamma_\pm := [0, a] \cup [a, a \pm i\infty]$, where $a > x_*$ is a large real parameter, i.e. we first go from the origin along the real axis to a and then we move vertically up or down depending on the sign in front of $\epsilon = +0$ in (A.55). When taking the difference in (A.55), the contributions of the x-integration from the horizontal segment $[0, a]$ exactly cancel. The only contribution comes from the opposite vertical x-integration regimes, that can now be estimated separately, yielding the bound

$$(A.56) \qquad \rho(E) \lesssim n^2 \Big| \int_a^{a+i\infty} \mathrm{d}x \oint \mathrm{d}y \, e^{-nf(x)+nf(y)} y \cdot G(x, y) \Big|$$

that we need to estimate in the regime $E \lesssim n^{-2}\delta^{-1}$ in order to prove (A.33). We choose $a := (nE)^{-1}$ and note that $a \gg \delta^{-1}$ since $n\delta^2 \gg 1$, i.e. $a \gg x_*$ by using (A.54). Thus by deforming the y contour to pass through the saddle x_*, the two contours will not intersect, analogously to the situation in Step 1 of the previous proof.

The rest of the computation is a standard saddle point analysis for the y-integration. Using (A.54), in our regime of parameters we have

$$f(x_*) = -\frac{\delta^2}{2}(1 + O(\kappa)), \quad f''(x_*) = 3\delta^4(1 + O(\kappa)), \quad x_* \cdot G(x, x_*) = \frac{\delta^2}{x}\Big[1 + O\Big(\frac{1}{\delta|x|}\Big)\Big]$$

---

[4]In the first displayed formula in Section 6.2 of [22] we erroneously claimed that $x_* \approx 3\delta^{-1}/2$ in the regime $E \ll \delta^3$, the correct behavior is $x_* \approx \delta^{-1}$. This wrong factor does not influence the arguments in [22].

uniformly, whenever $x = a + \mathrm{i}t$, $t \in [0, \infty)$, with a small parameter $\kappa := \delta + 1/(n\delta^2) \ll 1$. This yields

$$\rho(E) \lesssim \frac{n^2}{\sqrt{nf''(x_*)}} e^{nf(x_*)} \Big| \int_a^{a+\mathrm{i}\infty} \mathrm{d}x \, e^{-nf(x)} x_* \cdot G(x, x_*) \Big|$$

(A.57)

$$\lesssim n^{3/2} e^{-n\delta^2(1+O(\delta))/2} \Big| \int_a^{a+\mathrm{i}\infty} \mathrm{d}x \, \frac{e^{-nf(x)}}{x} \Big[1 + O\Big(\frac{1}{\delta|x|}\Big) \Big] \Big|,$$

assuming for the moment that the main contribution comes from the y-region around the saddle.

In the large x regime, where $|x| = |a + \mathrm{i}t| \gg \delta^{-1}$ we have the expansion

(A.58)
$$f(x) = -\frac{\delta}{1+x} - Ex + O(|x|^{-2}).$$

Note that

$$-n \operatorname{Re} f(a + \mathrm{i}t) \leq \frac{n\delta}{|a + \mathrm{i}t|} + nEa \lesssim 1, \qquad t \in [0, \infty),$$

therefore the error terms in the integrand can be handled trivially and we have

(A.59) $\Big| \int_a^{a+\mathrm{i}\infty} \mathrm{d}x \, \frac{e^{-nf(x)}}{x} \Big[1 + O\Big(\frac{1}{\delta|x|}\Big) \Big] \Big| \lesssim \Big| \int_0^\infty \mathrm{d}t \, \frac{e^{\mathrm{i}nEt}}{a + \mathrm{i}t} \Big| + \Big| \int_0^\infty \mathrm{d}t \, \frac{\delta^{-1} + n\delta}{|a + \mathrm{i}t|^2} \Big| \lesssim 1$

using $n\delta/a = n^2 E\delta \lesssim 1$. This yields $\rho(E) \lesssim n^{3/2} e^{-n\delta^2(1+O(\delta))/2}$ in the regime $E \lesssim n^{-2}\delta^{-1}$, which gives (A.33) with a slightly weaker $(n\delta^2)^{-1/2}$ prefactor instead of $(n\delta^2)^{-2/3}$.

Finally, the y-integration in the regime away from the saddle is estimated by using monotonicity of $\operatorname{Re} f(y)$ along an appropriate contour found by plotting the level sets of $\operatorname{Re} f$. We omit these uninteresting details. $\qquad\square$

Compared with (A.52), for the real case an analogous but more involved representation formula holds, see [22, Eq. (34)–(36)]. It carries an additional integration variable $\tau \in [0, 1]$ related to the nontrivial dependence on $\operatorname{Im} z$. The phase function $f(y)$ involving the integration variable on the compact contour in (A.52) is also present in the real case; this gives the critical $e^{-n\delta^2/2}$ factor exactly as in the complex case. The analogue of the phase function $f(x)$ for the non-compact integration (called g in [22]) will now depend on the additional parameter τ, but for the relevant regime of very large $|x|$ its asymptotic expansion is similar to $f(x)$ in (A.58). Both phase functions depend trivially on ϵ, hence we have exactly the same cancellation effect in (A.55) as in the complex case, thus we indeed need to consider only the large $|x|$ regime. The precise estimates analogous to (A.57)–(A.59) and the control of the regimes far away from the saddle are more cumbersome and we do not pursue them in this paper.

REFERENCES

[1] AKEMANN, G. and BENDER, M. (2010). Interpolation between Airy and Poisson statistics for unitary chiral non-Hermitian random matrix ensembles. *J. Math. Phys.* **51** 103524, 21. https://doi.org/10.1063/1.3496899 MR2761338

[2] AKEMANN, G. and PHILLIPS, M. J. (2014). The interpolating Airy kernels for the $\beta = 1$ and $\beta = 4$ elliptic Ginibre ensembles. *J. Stat. Phys.* **155** 421–465. https://doi.org/10.1007/s10955-014-0962-6 MR3192169

[3] ALJADEFF, J., STERN, M. and SHARPEE, T. (2015). Transition to chaos in random networks with cell-type-specific connectivity. *Phys. Rev. Lett.* **114** 088101. https://doi.org/10.1103/PhysRevLett.114.088101

[4] ALLESINA, S., GRILLI, J., BARABÁS, G., TANG, S., ALJADEFF, J. and MARITAN, A. (2015). Predicting the stability of large structured food webs. *Nat. Commun.* **6** 7842. https://doi.org/10.1038/ncomms8842

[5] ALLESINA, S. and TANG, S. (2015). The stability–complexity relationship at age 40: a random matrix perspective. *Popul. Ecol.* **57** 63–75. https://doi.org/10.1007/s10144-014-0471-0

[6] ALT, J., ERDŐS, L. and KRÜGER, T. (2018). Local inhomogeneous circular law. *Ann. Appl. Probab.* **28** 148–203. https://doi.org/10.1214/17-AAP1302 MR3770875

[7] ARGUIN, L.-P., BELIUS, D. and BOURGADE, P. (2017). Maximum of the characteristic polynomial of random unitary matrices. *Comm. Math. Phys.* **349** 703–751. https://doi.org/10.1007/s00220-016-2740-6 MR3594368

[8] ARGUIN, L.-P., BELIUS, D., BOURGADE, P., RADZIWIŁŁ, M. and SOUNDARARAJAN, K. (2019). Maximum of the Riemann zeta function on a short interval of the critical line. *Comm. Pure Appl. Math.* **72** 500–535. https://doi.org/10.1002/cpa.21791 MR3911893

[9] ARGUIN, L.-P., BOURGADE, P. and RADZIWIŁŁ, M. (2020). The Fyodorov–Hiary–Keating Conjecture. I. *arXiv preprint*. https://doi.org/10.38550/arXiv.2007.00988

[10] BAI, Z. D. (1997). Circular law. *Ann. Probab.* **25** 494–529. https://doi.org/10.1214/aop/1024404298 MR1428519

[11] BAI, Z. D. and YIN, Y. Q. (1986). Limiting behavior of the norm of products of random matrices and two problems of Geman-Hwang. *Probab. Theory Related Fields* **73** 555–569. https://doi.org/10.1007/BF00324852 MR863545

[12] BEN AROUS, G., FYODOROV, Y. V. and KHORUZHENKO, B. A. (2021). Counting equilibria of large complex systems by instability index. *Proc. Natl. Acad. Sci. USA* **118** Paper No. e2023719118, 8. https://doi.org/10.1073/pnas.2023719118 MR4305622

[13] BEN AROUS, G. and PÉCHÉ, S. (2005). Universality of local eigenvalue statistics for some sample covariance matrices. *Comm. Pure Appl. Math.* **58** 1316–1357. https://doi.org/10.1002/cpa.20070 MR2162782

[14] BENDER, M. (2010). Edge scaling limits for a family of non-Hermitian random matrix ensembles. *Probab. Theory Related Fields* **147** 241–271. https://doi.org/10.1007/s00440-009-0207-9 MR2594353

[15] BORDENAVE, C., CAPUTO, P., CHAFAÏ, D. and TIKHOMIROV, K. (2018). On the spectral radius of a random matrix: an upper bound without fourth moment. *Ann. Probab.* **46** 2268–2286. https://doi.org/10.1214/17-AOP1228 MR3813992

[16] BORDENAVE, C., CHAFAÏ, D. and GARCÍA-ZELADA, D. (2022). Convergence of the spectral radius of a random matrix through its characteristic polynomial. *Probab. Theory Related Fields* **182** 1163–1181. https://doi.org/10.1007/s00440-021-01079-9 MR4408512

[17] BOURGADE, P. (2022). Extreme gaps between eigenvalues of Wigner matrices. *J. Eur. Math. Soc. (JEMS)* **24** 2823–2873. https://doi.org/10.4171/jems/1141 MR4416591

[18] BOURGADE, P., YAU, H.-T. and YIN, J. (2014). The local circular law II: the edge case. *Probab. Theory Related Fields* **159** 619–660. https://doi.org/10.1007/s00440-013-0516-x MR3230004

[19] BOUTET DE MONVEL, A. and KHORUNZHY, A. (1999). Asymptotic distribution of smoothed eigenvalue density. II. Wigner random matrices. *Random Oper. Stochastic Equations* **7** 149–168. https://doi.org/10.1515/rose.1999.7.2.149 MR1689027

[20] CHALKER, J. T. and MEHLIG, B. (1998). Eigenvector Statistics in Non-Hermitian Random Matrix Ensembles. *Phys. Rev. Lett.* **81** 3367–3370. https://doi.org/10.1103/PhysRevLett.81.3367

[21] CHHAIBI, R., MADAULE, T. and NAJNUDEL, J. (2018). On the maximum of the CβE field. *Duke Math. J.* **167** 2243–2345. https://doi.org/10.1215/00127094-2018-0016 MR3848391

[22] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2020). Optimal lower bound on the least singular value of the shifted Ginibre ensemble. *Probab. Math. Phys.* **1** 101–146. https://doi.org/10.2140/pmp.2020.1.101 MR4408004

[23] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2021). Edge universality for non-Hermitian random matrices. *Probab. Theory Related Fields* **179** 1–28. https://doi.org/10.1007/s00440-020-01003-7 MR4221653

[24] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2021). Fluctuation around the circular law for random matrices with real entries. *Electron. J. Probab.* **26** Paper No. 24, 61. https://doi.org/10.1214/21-EJP591 MR4235475

[25] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2021). Central Limit Theorem for Linear Eigenvalue Statistics of Non-Hermitian Random Matrices. *Commun. Pure Appl. Math.* https://doi.org/10.1002/cpa.22028

[26] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2022). Density of small singular values of the shifted real Ginibre ensemble. *Ann. Henri Poincaré* **23** 3981–4002. https://doi.org/10.1007/s00023-022-01188-8 MR4496598

[27] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2022). On the condition number of the shifted real Ginibre ensemble. *SIAM J. Matrix Anal. Appl.* **43** 1469–1487. https://doi.org/10.1137/21M1424408 MR4474380

[28] CIPOLLONI, G., ERDŐS, L., SCHRÖDER, D. and XU, Y. (2022). Directional extremal statistics for Ginibre eigenvalues. *J. Math. Phys.* **63** Paper No. 103303, 11. https://doi.org/10.1063/5.0104290 MR4496015

[29] ERDŐS, L., KNOWLES, A. and YAU, H.-T. (2013). Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré* **14** 1837–1926. https://doi.org/10.1007/s00023-013-0235-y MR3119922

[30] ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2013). The local semicircle law for a general class of random matrices. *Electron. J. Probab.* **18** no. 59, 58. https://doi.org/10.1214/EJP.v18-2473 MR3068390

[31] ERDŐS, L., KRÜGER, T. and RENFREW, D. (2018). Power law decay for systems of randomly coupled differential equations. *SIAM J. Math. Anal.* **50** 3271–3290. https://doi.org/10.1137/17M1143125 MR3816180

[32] ERDŐS, L., KRÜGER, T. and RENFREW, D. (2019). Randomly coupled differential equations with elliptic correlations. *To appear in Ann. Appl. Probab.* https://doi.org/10.48550/arXiv.1908.05178

[33] ERDŐS, L., KRÜGER, T. and SCHRÖDER, D. (2019). Random matrices with slow correlation decay. *Forum Math. Sigma* **7** Paper No. e8, 89. https://doi.org/10.1017/fms.2019.2 MR3941370

[34] ERDŐS, L. and YAU, H.-T. (2017). *A dynamical approach to random matrix theory*. *Courant Lecture Notes in Mathematics* **28**. Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI. MR3699468

[35] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Bulk universality for generalized Wigner matrices. *Probab. Theory Related Fields* **154** 341–407. https://doi.org/10.1007/s00440-011-0390-3 MR2981427

[36] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.* **229** 1435–1515. https://doi.org/10.1016/j.aim.2011.12.010 MR2871147

[37] FENG, R., TIAN, G., WEI, D. and YAO, D. (2022). Principal minors of Gaussian orthogonal ensemble. *arXiv preprint.* https://doi.org/10.48550/arXiv.2205.05732

[38] FYODOROV, Y. V., HIARY, G. A. and KEATING, J. P. (2012). Freezing transition, characteristic polynomials of random matrices, and the Riemann zeta function. *Phys Rev Lett.* **108** 170601. https://doi.org/10.1103/PhysRevLett.108.170601

[39] FYODOROV, Y. V. and KEATING, J. P. (2014). Freezing transitions and extreme values: random matrix theory, and disordered landscapes. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **372** 20120503, 32. https://doi.org/10.1098/rsta.2012.0503 MR3151088

[40] FYODOROV, Y. V. and SIMM, N. J. (2016). On the distribution of the maximum value of the characteristic polynomial of GUE random matrices. *Nonlinearity* **29** 2837–2855. https://doi.org/10.1088/0951-7715/29/9/2837 MR3544809

[41] GEMAN, S. (1986). The spectral radius of large random matrices. *Ann. Probab.* **14** 1318–1328. MR866352

[42] GINIBRE, J. (1965). Statistical ensembles of complex, quaternion, and real matrices. *J. Mathematical Phys.* **6** 440–449. https://doi.org/10.1063/1.1704292 MR173726

[43] GIRKO, V. L. (1984). The circular law. *Teor. Veroyatnost. i Primenen.* **29** 669–679. MR773436

[44] HARPER, A. J. (2019). On the partition function of the Riemann zeta function, and the Fyodorov–Hiary–Keating conjecture. *arXiv preprint.* https://doi.org/10.48550/arXiv.1906.05783

[45] HE, Y. and KNOWLES, A. (2017). Mesoscopic eigenvalue statistics of Wigner matrices. *Ann. Appl. Probab.* **27** 1510–1550. https://doi.org/10.1214/16-AAP1237 MR3678478

[46] HE, Y. and KNOWLES, A. (2021). Fluctuations of extreme eigenvalues of sparse Erdős-Rényi graphs. *Probab. Theory Related Fields* **180** 985–1056. https://doi.org/10.1007/s00440-021-01054-4 MR4288336

[47] HUANG, J., LANDON, B. and YAU, H.-T. (2020). Transition from Tracy-Widom to Gaussian fluctuations of extremal eigenvalues of sparse Erdős-Rényi graphs. *Ann. Probab.* **48** 916–962. https://doi.org/10.1214/19-AOP1378 MR4089498

[48] JOHANSSON, K. (2007). From Gumbel to Tracy-Widom. *Probab. Theory Related Fields* **138** 75–112. https://doi.org/10.1007/s00440-006-0012-7 MR2288065

[49] KHORUNZHY, A. M., KHORUZHENKO, B. A. and PASTUR, L. A. (1996). Asymptotic properties of large random matrices with independent entries. *J. Math. Phys.* **37** 5033–5060. https://doi.org/10.1063/1.531589 MR1411619

[50] KOPEL, P. (2015). Linear statistics of non-Hermitian matrices matching the real or complex Ginibre ensemble to four moments. *arXiv preprint*. https://doi.org/10.48550/arXiv.1510.02987

[51] LAMBERT, G. (2020). Maximum of the characteristic polynomial of the Ginibre ensemble. *Comm. Math. Phys.* **378** 943–985. https://doi.org/10.1007/s00220-020-03813-1 MR4134939

[52] LANDON, B., SOSOE, P. and YAU, H.-T. (2019). Fixed energy universality of Dyson Brownian motion. *Adv. Math.* **346** 1137–1332. https://doi.org/10.1016/j.aim.2019.02.010 MR3914908

[53] LEE, J. O. and SCHNELLI, K. (2015). Edge universality for deformed Wigner matrices. *Rev. Math. Phys.* **27** 1550018, 94. https://doi.org/10.1142/S0129055X1550018X MR3405746

[54] LEE, J. O. and SCHNELLI, K. (2016). Tracy-Widom distribution for the largest eigenvalue of real sample covariance matrices with general population. *Ann. Appl. Probab.* **26** 3786–3839. https://doi.org/10.1214/16-AAP1193 MR3582818

[55] LEE, J. O. and SCHNELLI, K. (2018). Local law and Tracy-Widom limit for sparse random matrices. *Probab. Theory Related Fields* **171** 543–616. https://doi.org/10.1007/s00440-017-0787-8 MR3800840

[56] LYTOVA, A. and PASTUR, L. (2009). Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *Ann. Probab.* **37** 1778–1840. https://doi.org/10.1214/09-AOP452 MR2561434

[57] MAY, R. M. (1972). Will a large complex system be stable? *Nature* **238** 413–4. https://doi.org/10.1038/238413a0

[58] MEHLIG, B. and CHALKER, J. T. (2000). Statistical properties of eigenvectors in non-Hermitian Gaussian random matrix ensembles. *J. Math. Phys.* **41** 3233–3256. https://doi.org/10.1063/1.533302 MR1755501

[59] NAJNUDEL, J. (2018). On the extreme values of the Riemann zeta function on random intervals of the critical line. *Probab. Theory Related Fields* **172** 387–452. https://doi.org/10.1007/s00440-017-0812-y MR3851835

[60] PAQUETTE, E. and ZEITOUNI, O. (2018). The maximum of the CUE field. *Int. Math. Res. Not. IMRN* **16** 5028–5119. https://doi.org/10.1093/imrn/rnx033 MR3848227

[61] RAJAN, K. and ABBOTT, L. F. (2006). Eigenvalue spectra of random matrices for neural networks. *Phys. Rev. Lett.* **97** 188104. https://doi.org/10.1103/PhysRevLett.97.188104

[62] SAKSMAN, E. and WEBB, C. (2020). The Riemann zeta function and Gaussian multiplicative chaos: statistics on the critical line. *Ann. Probab.* **48** 2680–2754. https://doi.org/10.1214/20-AOP1433 MR4164452

[63] SCHNELLI, K. and XU, Y. (2021). Convergence rate to the Tracy–Widom laws for the largest eigenvalue of sample covariance matrices. *arXiv preprint*. https://doi.org/10.48550/arXiv.2108.02728

[64] SCHNELLI, K. and XU, Y. (2022). Convergence Rate to the Tracy–Widom Laws for the Largest Eigenvalue of Wigner Matrices. *Commun. Math. Phys.* **393** 839–907. https://doi.org/10.1007/s00220-022-04377-y

[65] SOMPOLINSKY, H., CRISANTI, A. and SOMMERS, H. J. (1988). Chaos in random neural networks. *Phys. Rev. Lett.* **61** 259–262. https://doi.org/10.1103/PhysRevLett.61.259 MR949871

[66] TAO, T. and VU, V. (2008). Random matrices: the circular law. *Commun. Contemp. Math.* **10** 261–307. https://doi.org/10.1142/S0219199708002788 MR2409368

[67] TAO, T. and VU, V. (2011). Random matrices: universality of local eigenvalue statistics. *Acta Math.* **206** 127–204. https://doi.org/10.1007/s11511-011-0061-3 MR2784665

[68] TAO, T. and VU, V. (2015). Random matrices: universality of local spectral statistics of non-Hermitian matrices. *Ann. Probab.* **43** 782–874. https://doi.org/10.1214/13-AOP876 MR3306005