Electron. J. Probab. **19** (2014), no. 33, 1–53.
ISSN: 1083-6489 DOI: 10.1214/EJP.v19-3054

Isotropic local laws for sample covariance and generalized Wigner matrices*

Alex Bloemendal† László Erdős‡ Antti Knowles§
Horng-Tzer Yau† Jun Yin¶

Abstract

We consider sample covariance matrices of the form X^*X, where X is an $M \times N$ matrix with independent random entries. We prove the *isotropic local Marchenko-Pastur law*, i.e. we prove that the resolvent $(X^*X - z)^{-1}$ converges to a multiple of the identity in the sense of quadratic forms. More precisely, we establish sharp high-probability bounds on the quantity $\langle \mathbf{v}, (X^*X - z)^{-1}\mathbf{w}\rangle - \langle \mathbf{v}, \mathbf{w}\rangle m(z)$, where m is the Stieltjes transform of the Marchenko-Pastur law and $\mathbf{v}, \mathbf{w} \in \mathbb{C}^N$. We require the logarithms of the dimensions M and N to be comparable. Our result holds down to scales $\operatorname{Im} z \geqslant N^{-1+\varepsilon}$ and throughout the entire spectrum away from 0. We also prove analogous results for generalized Wigner matrices.

Keywords: Sample covariance matrix; isotropic local law; eigenvalue rigidity; delocalization.
AMS MSC 2010: 15B52, 82B44.
Submitted to EJP on October 3, 2013, final version accepted on February 22, 2014.

1 Introduction

The empirical density of eigenvalues of large $N \times N$ random matrices typically converges to a deterministic limiting law. For Wigner matrices this law is the celebrated Wigner semicircle law [22] and for sample covariance matrices it is the Marchenko-Pastur law [20]. Under some additional moment conditions this convergence also holds in very small spectral windows, all the way down to the scale of the eigenvalue spacing. In this paper we normalize the matrix so that the support of its spectrum remains bounded as N tends to infinity. In particular, the typical eigenvalue spacing is of order $1/N$ away from the spectral edges. The empirical eigenvalue density is conveniently, and commonly, studied via its Stieltjes transform – the normalized trace of the resolvent, $\frac{1}{N}\operatorname{Tr}(H - z)^{-1}$, where $z = E + \mathrm{i}\eta$ is the spectral parameter with positive imaginary

*Support: SFB-TR 12 German Research Council Grant ; ERC Advanced Grant RANMAT 338804 ; Swiss National Science Foundation grant 144662 ; NSF Grant DMS-1307444 and Simons investigator fellowship.
†Harvard University, USA. E-mail: alexb@math.harvard.edu,htyau@math.harvard.edu
‡IST, Austria. E-mail: lerdos@ist.ac.at. On leave from Institute of Mathematics, University of Munich.
§ETH Zürich, Switzerland. E-mail: knowles@math.ethz.ch
¶University of Wisconsin, USA. E-mail: yin@math.wisc.edu

part η. Understanding the eigenvalue density on small scales of order η around a fixed value $E \in \mathbb{R}$ is roughly equivalent to understanding its Stieltjes transform with spectral parameter $z = E + \mathrm{i}\eta$. The smallest scale on which a deterministic limit is expected to emerge is $\eta \gg N^{-1}$; below this scale the empirical eigenvalue density remains a fluctuating object even in the limit of large N, driven by the fluctuations of individual eigenvalues. We remark that a local law on the optimal scale $1/N$ (up to logarithmic corrections) was first obtained in [11].

In recent years there has been substantial progress in understanding the local versions of the semicircle and the Marchenko-Pastur laws (see [9, 5] for an overview and detailed references). This research was originally motivated by the Wigner-Dyson-Mehta universality conjecture for the local spectral statistics of random matrices. The celebrated sine kernel universality and related results for other symmetry classes concern higher-order correlation functions, and not just the eigenvalue density. Moreover, they pertain to scales of order $1/N$, smaller than the scales on which local laws hold. Nevertheless, local laws (with precise error bounds) are essential ingredients for proving universality. In particular, one of their consequences, the precise localization of the eigenvalues (called *rigidity bounds*), has played a fundamental role in the relaxation flow analysis of the Dyson Brownian Motion, which has led to the proof of the Wigner-Dyson-Mehta universality conjecture for all symmetry classes [12, 13].

The basic approach behind the proofs of local laws is the analysis of a self-consistent equation for the Stieltjes transform, a scalar equation which controls the trace of the resolvent (and hence the empirical eigenvalue density). A vector self-consistent equation for the diagonal resolvent matrix entries, $[(H - z)^{-1}]_{ii}$, was introduced in [15]. Later, a matrix self-consistent equation was derived in [7]. Such self-consistent equations provide *entrywise* control of the resolvent and not only its trace. This latter fact has proved a key ingredient in the *Green function comparison method* (introduced in [15] and extended to the spectral edge in [16]), which allows the comparison of local statistics via moment matching even below the scale of eigenvalue spacing.

In this paper we are concerned with *isotropic local laws*, in which the control of the matrix entries $[(H - z)^{-1}]_{ij}$ is generalized to a control of quantities of the form $\langle \mathbf{v}, (H - z)^{-1}\mathbf{w} \rangle$, where $\mathbf{v}, \mathbf{w} \in \mathbb{C}^N$ are deterministic vectors. This may be interpreted as basis-independent control on the resolvent. The fact that the matrix entries are independent distinguishes the standard basis of $\mathbb{C}^N$ in the analysis of the resolvent. Unless the entries of H are Gaussian, this independence of the matrix entries is destroyed after a change of basis, and the isotropic law is a nontrivial generalization of the entrywise law. The first isotropic local law was proved in [18], where it was established for Wigner matrices.

The main motivation for isotropic local laws is the study of *deformed matrix ensembles*. A simple example is the sum $H + A$ of a Wigner matrix H and a deterministic finite-rank matrix A. As it turns out, a powerful means to study the eigenvalues and eigenvectors of such deformed matrices is to derive large deviation bounds and central limit theorems for quantities of the form $\langle \mathbf{v}, (H - z)^{-1}\mathbf{w} \rangle$, where $\mathbf{v}$ and $\mathbf{w}$ are eigenvectors of A. Deformed matrix ensembles are known to exhibit rather intricate spectra, depending on the spectrum of A. In particular, the spectrum of $H + A$ may contain *outliers* – lone eigenvalues separated from the bulk spectrum. The creation or annihilation of an outlier occurs at a sharp transition when an eigenvalue of A crosses a critical value. This transition is often referred to as the *BBP transition* and was first established in [1] for unitary matrices and extended in [4, 3] to other symmetry classes. Similarly to the above deformed Wigner matrices, one may introduce a class of deformed sample covariance matrices, commonly referred to as *spiked population models* [17], which describe populations with nontrivial correlations (or "spikes").

The isotropic local laws established in this paper serve as a key input in establishing detailed results about the eigenvalues and eigenvectors of deformed matrix models. These include:

(a) A complete picture of the distribution of outlier eigenvalues/eigenvectors, as well as the non-outlier eigenvalues/eigenvectors near the spectral edge.

(b) An investigation of the BBP transition using that, thanks to the optimality of the high-probability bounds in the local laws, the results of (a) extend even to the case when some eigenvalues of A are very close to the critical value.

This programme for the eigenvalues of deformed Wigner matrices was carried out in [18, 19]. In the upcoming paper [2], we shall carry out this programme for the eigenvectors of spiked population models.

In this paper we prove the isotropic Marchenko-Pastur law for sample covariance matrices as well as the isotropic semicircle law for generalized Wigner matrices. Our proofs are based on a novel method, which is considerably more robust than that of [18]. Both proofs (the one from [18] and the one presented here) crucially rely on the entrywise local law as input, but follow completely different approaches to obtain the isotropic law from the entrywise one. The basic idea of the proof in [18] is to use the Green function comparison method to compare the resolvent of a given Wigner matrix to the resolvent of a Gaussian random matrix, for which the isotropic law is a trivial corollary of the entrywise one (by basis transformation). Owing to various moment matching conditions imposed by the Green function comparison, the result of [18] required the variances of all matrix entries to coincide and, for results in the bulk spectrum, the third moments to vanish. In contrast, our current approach does not rely on Green function comparison. Instead, it consists of a precise analysis of the cancellation of fluctuations in Green functions. We use a graphical expansion method inspired by techniques recently developed in [6] to control fluctuations in Green functions of random band matrices.

Our first main result is the isotropic local Marchenko-Pastur law for sample covariance matrices $H = X^*X$, where X is an $M \times N$ matrix. We allow the dimensions of X to differ wildly: we only assume that $\log N \asymp \log M$. In particular, the aspect ratio $\phi = M/N$ – a key parameter in the Marchenko-Pastur law – may scale as a power of N. Our entrywise law (required as input for the proof of the isotropic law) is a generalization of the one given in [21]. In addition to generalizing the proof of [21], we simplify and streamline it, so as to obtain a short and self-contained proof.

Our second main result is the isotropic local semicircle law for generalized Wigner matrices. This extends the isotropic law of [18] from Wigner matrices to generalized Wigner matrices, in which the variances of the matrix entries need not coincide. It also dispenses with the third moment assumption of [18] mentioned previously. In fact, our proof applies to even more general matrix models, provided that an entrywise law has been established. As an application of the isotropic laws, we also prove a basis-independent version of eigenvector delocalization for both sample covariance and generalized Wigner matrices.

We conclude with an outline of the paper. In Section 2 we define our models and state our results, first for sample covariance matrices (Section 2.1) and then for generalized Wigner matrices (Section 2.2). The rest of the paper is devoted to the proofs. Since they are very similar for sample covariance matrices and generalized Wigner matrices, we only give the details for sample covariance matrices. Thus, Sections 3–6 are devoted to the proof of the isotropic Marchenko-Pastur law for sample covariance matrices; in Section 7, we describe how to modify the arguments to prove the isotropic semicircle law for generalized Wigner matrices. Section 3 collects some basic identities

and estimates that we shall use throughout the proofs. In Section 4 we prove the entry-wise local Marchenko-Pastur law, generalizing the results of [21]. The main argument and the bulk of the proof, i.e. the proof of the isotropic law, is given in Section 5. For a sketch of the argument we refer to Section 5.3. Finally, in Section 6 we draw some simple consequences from the isotropic law: optimal control outside of the spectrum and isotropic delocalization bounds.

Conventions

We use C to denote a generic large positive constant, which may depend on some fixed parameters and whose value may change from one expression to the next. Similarly, we use c to denote a generic small positive constant. For two positive quantities A_N and B_N depending on N we use the notation $A_N \asymp B_N$ to mean $C^{-1}A_N \leqslant B_N \leqslant CA_N$ for some positive constant C.

2 Results

2.1 Sample covariance matrix

Let X be an $M \times N$ matrix whose entries $X_{i\mu}$ are independent complex-valued random variables satisfying

$$\mathbb{E}X_{i\mu} \;=\; 0\,, \qquad \mathbb{E}|X_{i\mu}|^2 \;=\; \frac{1}{\sqrt{NM}}\,. \tag{2.1}$$

We shall study the $N \times N$ matrix X^*X; hence we regard N as the fundamental large parameter, and write $M \equiv M_N$. Our results also apply to the matrix XX^* provided one replaces $N \leftrightarrow M$. See Remark 2.11 below for more details.

We always assume that M and N satisfy the bounds

$$N^{1/C} \;\leqslant\; M \;\leqslant\; N^C \tag{2.2}$$

for some positive constant C. We define the ratio

$$\phi \;=\; \phi_N \;:=\; \frac{M}{N}\,,$$

which may depend on N. Here, and throughout the following, in order to unclutter notation we omit the argument N in quantities, such as X and ϕ, that depend on it.

We make the following technical assumption on the tails of the entries of X. We assume that, for all $p \in \mathbb{N}$, the random variables $(NM)^{1/4}X_{i\mu}$ have a uniformly bounded p-th moment: there is a constant C_p such that

$$\mathbb{E}\big|(NM)^{1/4}X_{i\mu}\big|^p \;\leqslant\; C_p\,. \tag{2.3}$$

It is well known that the empirical distribution of the eigenvalues of the $N\times N$ matrix X^*X has the same asymptotics as the *Marchenko-Pastur* law[20]

$$\varrho_\phi(\mathrm{d}x) \;:=\; \frac{\sqrt{\phi}}{2\pi}\sqrt{\frac{[(x-\gamma_-)(\gamma_+-x)]_+}{x^2}}\,\mathrm{d}x + (1-\phi)_+\,\delta(\mathrm{d}x)\,, \tag{2.4}$$

where we defined

$$\gamma_\pm \;:=\; \sqrt{\phi} + \frac{1}{\sqrt{\phi}} \pm 2 \tag{2.5}$$

to be the edges of the limiting spectrum. Note that (2.4) is normalized so that its integral is equal to one. The Stieltjes transform of the Marchenko-Pastur law (2.4) is

$$m_\phi(z) \;:=\; \int \frac{\varrho_\phi(\mathrm{d}x)}{x-z} \;=\; \frac{\phi^{1/2} - \phi^{-1/2} - z + \mathrm{i}\sqrt{(z-\gamma_-)(\gamma_+-z)}}{2\,\phi^{-1/2}\,z}\,, \tag{2.6}$$

where the square root is chosen so that m_ϕ is holomorphic in the upper half-plane and satisfies $m_\phi(z) \to 0$ as $z \to \infty$. The function $m_\phi = m_\phi(z)$ is also characterized as the unique solution of the equation

$$m + \frac{1}{z + z\phi^{-1/2}m - (\phi^{1/2} - \phi^{-1/2})} \;=\; 0 \tag{2.7}$$

satisfying $\operatorname{Im} m(z) > 0$ for $\operatorname{Im} z > 0$. The formulas (2.4)–(2.7) were originally derived for the case when $\phi = M/N$ is independent of N (or, more precisely, when ϕ has a limit in $(0, \infty)$ as $N \to \infty$). Our results allow ϕ to depend on N under the constraint (2.2), so that m_ϕ and ϱ_ϕ may also depend on N through ϕ.

Throughout the following we use a spectral parameter

$$z \;=\; E + \mathrm{i}\eta\,,$$

with $\eta > 0$, as the argument of Stieltjes transforms and resolvents. Define the resolvent

$$R(z) \;:=\; (X^*X - z)^{-1}\,. \tag{2.8}$$

For $z \in \mathbb{C}$, define $\kappa := \kappa(z)$ to be the distance of $E = \operatorname{Re} z$ to the spectral edges $\gamma_\pm$, i.e.

$$\kappa \;:=\; \min\bigl\{|\gamma_+ - E|\,,\, |\gamma_- - E|\bigr\}\,. \tag{2.9}$$

The following notion of a high-probability bound was introduced in [6], and has been subsequently used in a number of works on random matrix theory. It provides a simple way of systematizing and making precise statements of the form "ξ is bounded with high probability by ζ up to small powers of N".

Definition 2.1 (Stochastic domination)**.** *Let*

$$\xi = \bigl(\xi^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}\bigr)\,, \qquad \zeta = \bigl(\zeta^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}\bigr)$$

be two families of nonnegative random variables, where $U^{(N)}$ is a possibly N-dependent parameter set. We say that ξ is stochastically dominated by ζ, uniformly in u, *if for all (small) $\varepsilon > 0$ and (large) $D > 0$ we have*

$$\sup_{u \in U^{(N)}} \mathbb{P}\Bigl[\xi^{(N)}(u) > N^\varepsilon \zeta^{(N)}(u)\Bigr] \;\leqslant\; N^{-D}$$

for large enough $N \geqslant N_0(\varepsilon, D)$. Throughout this paper the stochastic domination will always be uniform in all parameters (such as matrix indices and the spectral parameter z) that are not explicitly fixed. Note that $N_0(\varepsilon, D)$ may depend on the constants from (2.2) *and* (2.3) *as well as any constants fixed in the assumptions of our main results. If ξ is stochastically dominated by ζ, uniformly in u, we use the notation $\xi \prec \zeta$. Moreover, if for some complex family ξ we have $|\xi| \prec \zeta$ we also write $\xi = O_\prec(\zeta)$.*

Remark 2.2. *Because of* (2.2)*, all (or some) factors of N in Definition* (2.1) *could be replaced with M without changing the definition of stochastic domination.*

Sometimes we shall need the following notion of high probability.

Definition 2.3. *An N-dependent event $\Xi \equiv \Xi_N$ holds with* high probability *if $1 - \mathbf{1}(\Xi) \prec 0$ (or, equivalently, if $1 \prec \mathbf{1}(\Xi)$).*

We introduce the quantity

$$K \;\equiv\; K_N \;:=\; \min\{M, N\}\,, \tag{2.10}$$

which is the number of nontrivial (i.e. nonzero) eigenvalues of X^*X; the remaining $N-K$ eigenvalues of X^*X are zero. (Note that the K nontrivial eigenvalues of X^*X coincide with those of XX^*.) Fix a (small) $\omega \in (0,1)$ and define the domain

$$\mathbf{S} \equiv \mathbf{S}(\omega, K) := \left\{ z = E + \mathrm{i}\eta \in \mathbb{C} : \kappa \leqslant \omega^{-1}, K^{-1+\omega} \leqslant \eta \leqslant \omega^{-1}, |z| \geqslant \omega \right\}. \tag{2.11}$$

Throughout the following we regard ω as fixed once and for all, and do not track the dependence of constants on ω.

Theorem 2.4 (Isotropic local Marchenko-Pastur law)**.** *Suppose that* (2.1)*,* (2.2)*, and* (2.3) *hold. Then*

$$\left| \langle \mathbf{v}, R(z)\mathbf{w} \rangle - m_\phi(z) \langle \mathbf{v}, \mathbf{w} \rangle \right| \prec \sqrt{\frac{\operatorname{Im} m_\phi(z)}{N\eta}} + \frac{1}{N\eta} \tag{2.12}$$

uniformly in $z \in \mathbf{S}$ and any deterministic unit vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^N$. Moreover, we have

$$\left| \frac{1}{N} \operatorname{Tr} R(z) - m_\phi(z) \right| \prec \frac{1}{N\eta} \tag{2.13}$$

uniformly in $z \in \mathbf{S}$.

Beyond the support of the limiting spectrum, one has stronger control all the way down to the real axis. For fixed (small) $\omega \in (0,1)$ define the region

$$\begin{aligned} \widetilde{\mathbf{S}} &\equiv \widetilde{\mathbf{S}}(\omega, K) \\ &:= \left\{ z = E + \mathrm{i}\eta \in \mathbb{C} : E \notin [\gamma_-, \gamma_+], K^{-2/3+\omega} \leqslant \kappa \leqslant \omega^{-1}, |z| \geqslant \omega, 0 < \eta \leqslant \omega^{-1} \right\} \end{aligned} \tag{2.14}$$

of spectral parameters separated from the asymptotic spectrum by $K^{-2/3+\omega}$, which may have an arbitrarily small positive imaginary part η.

Theorem 2.5 (Isotropic local Marchenko-Pastur law outside the spectrum)**.** *Suppose that* (2.1)*,* (2.2)*, and* (2.3) *hold. Then*

$$\left| \langle \mathbf{v}, R(z)\mathbf{w} \rangle - m_\phi(z) \langle \mathbf{v}, \mathbf{w} \rangle \right| \prec \sqrt{\frac{\operatorname{Im} m_\phi(z)}{N\eta}} \asymp \frac{1}{1+\phi^{-1}} (\kappa + \eta)^{-1/4} K^{-1/2} \tag{2.15}$$

uniformly in $z \in \widetilde{\mathbf{S}}$ and any deterministic unit vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^N$.

Remark 2.6. *All probabilistic estimates* (2.12)*–*(2.15) *of Theorems 2.4 and 2.5 may be strengthened to hold simultaneously for all $z \in \mathbf{S}$ and for all $z \in \widetilde{\mathbf{S}}$, respectively. For instance,* (2.12) *may be strengthened to*

$$\mathbb{P}\left[\bigcap_{z \in \mathbf{S}} \left\{ \left| \langle \mathbf{v}, R(z)\mathbf{w} \rangle - m_\phi(z) \langle \mathbf{v}, \mathbf{w} \rangle \right| \leqslant N^\varepsilon \left(\sqrt{\frac{\operatorname{Im} m_\phi(z)}{N\eta}} + \frac{1}{N\eta} \right) \right\} \right] \geqslant 1 - N^{-D},$$

for all $\varepsilon > 0$, $D > 0$, and $N \geqslant N_0(\varepsilon, D)$.

In the case of Theorem 2.5 this generalization is an immediate consequence of its proof, and in the case of Theorem 2.4 it follows from a simple lattice argument combined with the Lipschitz continuity of R and m_ϕ on $\mathbf{S}$. See e.g. [10, Corollary 3.19] for the details.

Remark 2.7. *The right-hand side of* (2.15) *is stable under the limit $\eta \to 0$, and may therefore be extended to $\eta = 0$. Recalling the previous remark, we conclude that* (2.15) *also holds for $\eta = 0$.*

The next results are on the nontrivial eigenvalues of X^*X as well as the corresponding eigenvectors. As remarked above, the matrix X^*X has K nontrivial eigenvalues, which we order according to $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_K$. Let $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(K)} \in \mathbb{C}^N$ be the normalized eigenvectors of X^*X associated with the nontrivial eigenvalues $\lambda_1, \ldots, \lambda_K$.

Theorem 2.8 (Isotropic delocalization)**.** *Suppose that* (2.1), (2.2), *and* (2.3) *hold. Then for any $\varepsilon > 0$ we have the bound*

$$|\langle \mathbf{u}^{(\alpha)}, \mathbf{v} \rangle|^2 \;\prec\; N^{-1} \tag{2.16}$$

uniformly for $\alpha \leqslant (1-\varepsilon)K$ and all normalized $\mathbf{v} \in \mathbb{C}^N$. If in addition $|\phi - 1| \geqslant c$ for some constant $c > 0$, then (2.16) *holds uniformly for all $\alpha \leqslant K$.*

Remark 2.9. *Isotropic delocalization bounds in particular imply that the entries $u_i^{(\alpha)}$ of the eigenvectors $\mathbf{u}^{(\alpha)}$ are strongly oscillating in the sense that $\sum_{i=1}^N |u_i^{(\alpha)}| \succ N^{1/2}$ but $\left|\sum_{i=1}^N u_i^{(\alpha)}\right| \prec 1$. To see this, choose $\mathbf{v} = \mathbf{e}_i$ in* (2.16)*, which implies $|u_i^{(\alpha)}| \prec N^{-1/2}$, from which the first estimate follows using $\sum_{i=1}^N |u_i^{(\alpha)}|^2 = 1$. On the other hand, choosing $\mathbf{v} = N^{-1/2}(1, 1, \ldots, 1)$ in* (2.16) *yields the second estimate. Note that, if $\mathbf{u} = (u_1, \ldots, u_N)$ is uniformly distributed on the unit sphere $\mathbb{S}^{N-1}$, the high-probability bounds $\sum_{i=1}^N |u_i| \succ N^{1/2}$ and $|\sum_{i=1}^N u_i| \prec 1$ are sharp (in terms of the power of N on the right-hand side).*

The following result is on the rigidity of the nontrivial eigenvalues of X^*X, which coincide with the nontrivial eigenvalues of XX^*. Let $\gamma_1 \geqslant \gamma_2 \geqslant \cdots \geqslant \gamma_K$ be the *classical eigenvalue locations according to ϱ_ϕ* (see (2.4)), defined through

$$\int_{\gamma_\alpha}^{\infty} \varrho_\phi(\mathrm{d}x) \;=\; \frac{\alpha}{N}\,. \tag{2.17}$$

Theorem 2.10 (Eigenvalue rigidity)**.** *Fix a (small) $\omega \in (0,1)$ and suppose that* (2.1), (2.2), *and* (2.3) *hold. Then*

$$\left|\lambda_\alpha - \gamma_\alpha\right| \;\prec\; \alpha^{-1/3} K^{-2/3} \tag{2.18}$$

uniformly for all $\alpha \in \{1, \ldots, [(1-\omega)K]\}$. If in addition $|\phi - 1| \geqslant c$ for some constant $c > 0$ then

$$\left|\lambda_\alpha - \gamma_\alpha\right| \;\prec\; (K + 1 - \alpha)^{-1/3} K^{-2/3}, \tag{2.19}$$

uniformly for all $\alpha \in \{[K/2], \ldots, K\}$.

Remark 2.11. *We stated our results for the matrix X^*X, but they may be easily applied to the matrix XX^* as well. Indeed, Theorems 2.4, 2.5, 2.8, and 2.10 remain valid after the following changes: $X \mapsto X^*$, $M \mapsto N$, $N \mapsto M$, and $\phi \mapsto \phi^{-1}$. (In the case of Theorem 2.10 these changes leave the statement unchanged.) Note that the empirical distribution of the eigenvalues of XX^* has the same asymptotics as $\varrho_{\phi^{-1}}(\mathrm{d}x)$, whose Stieltjes transform is*

$$m_{\phi^{-1}}(z) \;=\; \frac{1}{\phi}\left(m_\phi(z) + \frac{1-\phi}{z}\right). \tag{2.20}$$

2.2 Generalized Wigner matrix

Let $H = H^*$ be an $N \times N$ Hermitian matrix whose entries H_{ij} are independent complex-valued random variables for $i \leqslant j$. We always assume that entries are centred, i.e. $\mathbb{E}H_{ij} = 0$. Moreover, we assume that the variances

$$S_{ij} \;:=\; \mathbb{E}|H_{ij}|^2 \tag{2.21}$$

satisfy

$$C^{-1} \leqslant NS_{ij} \leqslant C, \qquad \sum_j S_{ij} = 1, \tag{2.22}$$

for some constant $C > 0$. We assume that all moments of the entries of $\sqrt{N}H$ are finite in the sense that for all $p \in \mathbb{N}$ there exists a constant C_p such that

$$\mathbb{E}\left|\sqrt{N}H_{ij}\right|^p \leqslant C_p \tag{2.23}$$

for all N, i, and j.

Let

$$\varrho(\mathrm{d}x) := \frac{1}{2\pi}\sqrt{(4-x^2)_+}\,\mathrm{d}x$$

denote the semicircle law, and

$$m(z) := \int \frac{\varrho(\mathrm{d}x)}{x-z} = \frac{-z+\sqrt{z^2-4}}{2} \tag{2.24}$$

its Stieltjes transform; here we chose the square root so that m is holomorphic in the upper half-plane and satisfies $m(z) \to 0$ as $z \to \infty$. Note that $m = m(z)$ is also characterized as the unique solution of

$$m + \frac{1}{z+m} = 0 \tag{2.25}$$

that satisfies $\operatorname{Im} m > 0$ for $\eta > 0$. Let

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

be the resolvent of H.

Fix a (small) $\omega \in (0,1)$ and define

$$\mathbf{S}_W \equiv \mathbf{S}_W(\omega, N) := \left\{z = E + \mathrm{i}\eta \in \mathbb{C} : |E| \leqslant \omega^{-1},\, N^{-1+\omega} \leqslant \eta \leqslant \omega^{-1}\right\}. \tag{2.26}$$

The subscript W in $\mathbf{S}_W$ stands for Wigner, and is added to distinguish this domain from the one defined in (2.11).

Theorem 2.12 (Isotropic local semicircle law)**.** *Suppose that* (2.22) *and* (2.23) *hold. Then*

$$\left|\langle \mathbf{v}, G(z)\mathbf{w}\rangle - \langle \mathbf{v}, \mathbf{w}\rangle m(z)\right| \prec \sqrt{\frac{\operatorname{Im} m(z)}{N\eta}} + \frac{1}{N\eta} \tag{2.27}$$

uniformly in $z \in \mathbf{S}_W$ and any deterministic unit vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^N$.

Theorem 2.12 is the isotropic generalization of the following result, proved in [9]. A similar result first appeared in [16].

Theorem 2.13 (Local semicircle law, [9, 16])**.** *Suppose that* (2.22) *and* (2.23) *hold. Then*

$$\left|G_{ij}(z) - \delta_{ij} m(z)\right| \prec \sqrt{\frac{\operatorname{Im} m(z)}{N\eta}} + \frac{1}{N\eta} \tag{2.28}$$

uniformly in $z \in \mathbf{S}_W$ and $i, j = 1, \ldots, N$.

The proof of the isotropic law (2.27) uses the regular, entrywise, law from Theorem 2.13 as input, in which $\mathbf{v}$ and $\mathbf{w}$ are taken to be parallel to the coordinate axes. Assuming the entrywise law (2.28) has been established, the proof of (2.27) is very robust, and holds under more general assumptions than (2.22). For instance, we have the following result.

Theorem 2.14. *Let $\widetilde{\mathbf{S}} \subset \mathbf{S}_W$ be an N-dependent spectral domain, $\widetilde{m}(z)$ a deterministic function on $\widetilde{\mathbf{S}}$ satisfying $c \leqslant |\widetilde{m}(z)| \leqslant C$ for $z \in \widetilde{\mathbf{S}}$, and $\widetilde{\Psi}(z)$ a deterministic control parameter satisfying $cN^{-1} \leqslant \widetilde{\Psi}(z) \leqslant N^{-c}$ for $z \in \widetilde{\mathbf{S}}$ and some constant $c > 0$. Suppose that*

$$|G_{ij}(z) - \delta_{ij}\widetilde{m}(z)| \;\prec\; \widetilde{\Psi}(z)$$

uniformly in $z \in \widetilde{\mathbf{S}}$. Suppose that the entries of H satisfy (2.23) *and the variances* (2.21) *of H satisfy $S_{ij} \leqslant CN^{-1}$ (which replaces the stronger assumption* (2.22)*). Then we have*

$$\big|\langle \mathbf{v}, G(z)\mathbf{w}\rangle - \langle \mathbf{v}, \mathbf{w}\rangle \widetilde{m}(z)\big| \;\prec\; \widetilde{\Psi}(z) \tag{2.29}$$

uniformly in $z \in \widetilde{\mathbf{S}}$ and any deterministic unit vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^N$.

The proof of Theorem 2.14 is the same as that of Theorem 2.12. Below we give the proof for Theorem 2.12, which can be trivially adapted to yield Theorem 2.14.

Combining Theorem 2.14 with the isotropic local semicircle law from [9], we may for instance obtain an isotropic local semicircle law for matrices where the lower bound of (2.22) is relaxed, so that some matrix entries may vanish.

Beyond the support of the limiting spectrum $[-2, 2]$, the statement of Theorem 2.12 may be improved to a bound that is stable all the way down to the real axis. For fixed (small) $\omega \in (0, 1)$ define the region

$$\widetilde{\mathbf{S}}_W \;\equiv\; \widetilde{\mathbf{S}}_W(\omega, N) \;:=\; \big\{z = E + \mathrm{i}\eta \in \mathbb{C} : 2 + N^{-2/3+\omega} \leqslant |E| \leqslant \omega^{-1}\,,\, 0 < \eta \leqslant \omega^{-1}\big\} \tag{2.30}$$

of spectral parameters separated from the asymptotic spectrum by $N^{-2/3+\omega}$, which may have an arbitrarily small positive imaginary part η.

Theorem 2.15 (Isotropic local semicircle law outside the spectrum)**.** *Suppose that* (2.22) *and* (2.23) *hold. Then*

$$\big|\langle \mathbf{v}, G(z)\mathbf{w}\rangle - m(z)\langle \mathbf{v}, \mathbf{w}\rangle\big| \;\prec\; \sqrt{\frac{\operatorname{Im} m(z)}{N\eta}} \tag{2.31}$$

uniformly in $z \in \widetilde{\mathbf{S}}_W$ and any deterministic unit vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^N$.

The statements in Theorems 2.12 and 2.15 can also be strengthened to simultaneously apply for all $z \in \mathbf{S}_W$ and $z \in \widetilde{\mathbf{S}}_W$, respectively; see Remark 2.6.

Let $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(N)}$ denote the normalized eigenvectors of H associated with the eigenvalues $\lambda_1, \dots, \lambda_N$.

Theorem 2.16 (Isotropic delocalization)**.** *Suppose that* (2.22) *and* (2.23) *hold. Then*

$$|\langle \mathbf{u}^{(\alpha)}, \mathbf{v}\rangle|^2 \;\prec\; N^{-1}$$

uniformly for all $\alpha = 1, \dots, N$ and all deterministic unit vectors $\mathbf{v} \in \mathbb{C}^N$.

Finally, in analogy to Theorem 2.10, we record the following rigidity result, which is a trivial consequence of [9, Theorem 7.6] with $X = CN^{-2/3}$ and $Y = CN^{-1}$; see also [16, Theorem 2.2]. Write $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_N$ for the eigenvalues of H, and let $\gamma_1 \geqslant \gamma_2 \geqslant \cdots \geqslant \gamma_N$ be their *classical locations according to* ϱ, defined through

$$\int_{\gamma_\alpha}^{\infty} \varrho(\mathrm{d}x) \;=\; \frac{\alpha}{N}\,. \tag{2.32}$$

Then we have

$$|\lambda_\alpha - \gamma_\alpha| \;\prec\; \big(\min\{\alpha, N+1-\alpha\}\big)^{-1/3} N^{-2/3} \tag{2.33}$$

for all $\alpha = 1, \dots, N$.

3 Preliminaries

The rest of this paper is devoted to the proofs of our main results. They are similar for sample covariance matrices and generalized Wigner matrices, and in Sections 3–6 we give the argument for sample covariance matrices (hence proving the results of Section 2.1). How to modify these arguments to generalized Wigner matrices (and hence prove the results of Section 2.2) is explained in Section 7. We choose to present our method in the context of sample covariance matrices mainly for two reasons. First, we take this opportunity to give a version of the entrywise local law (Section 4) – required as input for the proof of the isotropic law – which is more general and has a simpler proof than the local law previously established in [21]. Second, the proof of the isotropic law in the case of sample covariance matrices is conceptually slightly clearer due to a natural splitting of summation indices into two categories (which we distinguish by the use of Latin and Greek letters); this splitting is an essential structure behind our proof in Section 5, and is also used in the case of generalized Wigner matrices, in which case it is however purely artificial.

We now move on to the proofs. In order to unclutter notation, we shall often omit the argument z from quantities that depend on it. Thus, we for instance often write G instead of $G(z)$. We put the arguments z back when needed, typically if we are working with several different spectral parameters z.

3.1 Basic tools

We begin by recording some basic large deviations estimates. We consider complex-valued random variables ξ satisfying

$$\mathbb{E}\xi \;=\; 0\,, \qquad \mathbb{E}|\xi|^2 \;=\; 1\,, \qquad (\mathbb{E}|\xi|^p)^{1/p} \;\leqslant\; C_p \tag{3.1}$$

for all $p \in \mathbb{N}$ and some constants C_p.

Lemma 3.1 (Large deviation bounds)**.** *Let $\big(\xi_i^{(N)}\big)$ and $\big(\zeta_i^{(N)}\big)$ be independent families of random variables and $\big(a_{ij}^{(N)}\big)$ and $\big(b_i^{(N)}\big)$ be deterministic; here $N \in \mathbb{N}$ and $i, j = 1, \ldots, N$. Suppose that all entries $\xi_i^{(N)}$ and $\zeta_i^{(N)}$ are independent and satisfy* (3.1)*. Then we have the bounds*

$$\sum_i b_i \xi_i \;\prec\; \bigg(\sum_i |b_i|^2\bigg)^{1/2}, \tag{3.2}$$

$$\sum_{i,j} a_{ij} \xi_i \zeta_j \;\prec\; \bigg(\sum_{i,j} |a_{ij}|^2\bigg)^{1/2}, \tag{3.3}$$

$$\sum_{i \neq j} a_{ij} \xi_i \xi_j \;\prec\; \bigg(\sum_{i \neq j} |a_{ij}|^2\bigg)^{1/2}. \tag{3.4}$$

If the coefficients $a_{ij}^{(N)}$ and $b_i^{(N)}$ depend on an additional parameter u, then all of these estimates are uniform in u (see Definition 2.1), i.e. the threshold $N_0 = N_0(\varepsilon, D)$ in the definition of $\prec$ depends only on the family C_p from (3.1)*; in particular, N_0 does not depend on u.*

Proof. These estimates are an immediate consequence of Lemmas B.2, B.3, and B.4 in [7]. See also Theorem C.1 in [9]. □

The following lemma collects basic algebraic properties of stochastic domination $\prec$. We shall use them tacitly throughout the following.

Lemma 3.2. *1. Suppose that $\xi(u,v) \prec \zeta(u,v)$ uniformly in $u \in U$ and $v \in V$. If $|V| \leqslant N^C$ for some constant C then*

$$\sum_{v\in V} \xi(u,v) \;\prec\; \sum_{v\in V} \zeta(u,v)$$

uniformly in u.

2. *Suppose that $\xi_1(u) \prec \zeta_1(u)$ uniformly in u and $\xi_2(u) \prec \zeta_2(u)$ uniformly in u. Then*

$$\xi_1(u)\xi_2(u) \;\prec\; \zeta_1(u)\zeta_2(u)$$

uniformly in u.

3. *Suppose that $\Psi(u) \geqslant N^{-C}$ is deterministic and $\xi(u)$ is a nonnegative random variable satisfying $\mathbb{E}\xi(u)^2 \leqslant N^C$ for all u. Then, provided that $\xi(u) \prec \Psi(u)$ uniformly in u, we have*

$$\mathbb{E}\xi(u) \;\prec\; \Psi(u)$$

uniformly in u.

Proof. The claims (i) and (ii) follow from a simple union bound. For (iii), pick $\varepsilon > 0$ and assume to simplify notation that ξ and Ψ do not depend on u. Then

$$\begin{aligned}\mathbb{E}\xi \;&=\; \mathbb{E}\xi\mathbf{1}(\xi \leqslant N^\varepsilon\Psi) + \mathbb{E}\xi\mathbf{1}(\xi > N^\varepsilon\Psi)\\ &\leqslant\; N^\varepsilon\Psi + \sqrt{\mathbb{E}\xi^2}\sqrt{\mathbb{P}(\xi > N^\varepsilon\Psi)} \;\leqslant\; N^\varepsilon\Psi + N^{C/2-D/2},\end{aligned}$$

for arbitrary $D > 0$. The claim (iii) therefore follows by choosing $D \geqslant 3C$. □

Next, we give some basic facts about the Stieltjes transform m_ϕ of the Marchenko-Pastur law defined in (2.6). They have an especially simple form in the case $\phi \geqslant 1$; the complementary case $\phi < 1$ can be easily handled using (2.20). Recall the definition (2.9) of κ. We record the following elementary properties of m_ϕ, which may be proved e.g. by starting from the explicit form (2.6).

Lemma 3.3. *For $z \in \mathbf{S}$ and $\phi \geqslant 1$ we have*

$$|m_\phi(z)| \;\asymp\; 1\,, \qquad |1 - m_\phi(z)^2| \;\asymp\; \sqrt{\kappa+\eta}\,, \tag{3.5}$$

and

$$\operatorname{Im} m_\phi(z) \;\asymp\; \begin{cases} \sqrt{\kappa+\eta} & \text{if } E \in [\gamma_-,\gamma_+] \\ \frac{\eta}{\sqrt{\kappa+\eta}} & \text{if } E \notin [\gamma_-,\gamma_+]\,. \end{cases} \tag{3.6}$$

(All implicit constants depend on ω in the definition (2.11) *of* $\mathbf{S}$*.)*

ě The basic object is the $M \times N$ matrix X. We use the indices $i, j = 1, \dots, M$ to denote the rows of X and $\mu, \nu = 1, \dots, N$ to denote its columns. Unrestricted summations over Latin indices always run over the set $\{1, 2, \dots, M\}$ while Greek indices run over $\{1, 2, \dots, N\}$.

In addition to the resolvent R from (2.8), we shall need another resolvent, G:

$$G(z) \;:=\; (XX^* - z)^{-1}\,, \qquad R(z) \;:=\; (X^*X - z)^{-1}\,.$$

Although our main results only pertain to R, the resolvent G will play a crucial role in the proofs, in which we consider both X^*X and XX^* in tandem. In the following formulas the spectral parameter z plays no explicit role, and we therefore omit it from the notation, as explained at the beginning of this section.

Remark 3.4. *More abstractly, we may introduce two sets of indices, $I_{\mathrm{population}}$ and I_{sample}, such that $I_{\mathrm{population}}$ indexes the entries of G and I_{sample} the entries of R. Elements of $I_{\mathrm{population}}$ are always denoted by Latin letters and elements of I_{sample} by Greek letters. These two sets are to be viewed as distinct. For convenience of notation, we shall always use the customary representations $I_{\mathrm{population}} := \{1, \dots, M\}$ and $I_{\mathrm{sample}} := \{1, \dots, N\}$, bearing in mind that neither should be viewed as a subset of the other. This terminology originates from the statistical application of sample covariance matrices. The idea is that we are observing the statistics of a population of size M by making N independent measurements ("samples") of the population. Each observation is a column of X. Hence the population index i labels the rows of X and the sample index μ the columns of X.*

Definition 3.5 (Removing rows)**.** *For $T \subset \{1, \dots, M\}$ define*

$$(X^{(T)})_{i\mu} \;:=\; \mathbf{1}(i \notin T) X_{i\mu}\,.$$

Moreover, for $i, j \notin T$ we define the resolvents entries

$$G^{(T)}_{ij}(z) \;:=\; \big(X^{(T)}(X^{(T)})^* - z\big)^{-1}_{ij}\,, \qquad R^{(T)}_{\mu\nu}(z) \;:=\; \big((X^{(T)})^* X^{(T)} - z\big)^{-1}_{\mu\nu}\,.$$

When $T = \{a\}$, we abbreviate $(\{a\})$ by (a) in the above definitions; similarly, we write (ab) instead of $(\{a, b\})$.

We shall use the following expansion formulas for G. They are elementary consequences of Schur's complement formula; see e.g. Lemma 4.2 of [15] and Lemma 6.10 of [8] for proofs of similar identities.

Lemma 3.6 (Resolvent identities for G)**.** *For $i, j, k \notin T$ and $i, j \neq k$ we have*

$$G^{(T)}_{ij} \;=\; G^{(Tk)}_{ij} + \frac{G^{(T)}_{ik} G^{(T)}_{kj}}{G^{(T)}_{kk}}\,, \qquad \frac{1}{G^{(T)}_{ii}} = \frac{1}{G^{(Tk)}_{ii}} - \frac{G^{(T)}_{ik} G^{(T)}_{ki}}{G^{(T)}_{ii} G^{(Tk)}_{ii} G^{(T)}_{kk}}\,. \tag{3.7}$$

For $i \notin T$ we have

$$\frac{1}{G^{(T)}_{ii}} \;=\; -z - z \sum_{\mu,\nu} X_{i\mu} R^{(Ti)}_{\mu\nu} X^*_{\nu i}\,. \tag{3.8}$$

Moreover, for $i, j \notin T$ and $i \neq j$ we have

$$G^{(T)}_{ij} \;=\; z G^{(T)}_{ii} G^{(iT)}_{jj} \sum_{\mu,\nu} X_{i\mu} R^{(Tij)}_{\mu\nu} X^*_{\nu j}\,. \tag{3.9}$$

Definition 3.5 and Lemma 3.6 have the following analogues for removing columns and identities for R.

Definition 3.7 (Removing columns)**.** *For $T \subset \{1, \dots, N\}$ define*

$$(X^{[T]})_{i\mu} \;:=\; \mathbf{1}(\mu \notin T) X_{i\mu}\,.$$

Moreover, for $\mu, \nu \notin T$ we define the resolvents entries

$$G^{[T]}_{ij}(z) \;:=\; \big(X^{[T]}(X^{[T]})^* - z\big)^{-1}_{ij}\,, \qquad R^{[T]}_{\mu\nu}(z) \;:=\; \big((X^{[T]})^* X^{[T]} - z\big)^{-1}_{\mu\nu}\,.$$

When $T = \{\mu\}$, we abbreviate $[\{\mu\}]$ by $[\mu]$ in the above definitions; similarly, we write $[\mu\nu]$ instead of $[\{\mu, \nu\}]$.

We use the following expansion formulas for R, which are analogoues to those of Lemma 3.6.

Lemma 3.8 (Resolvent identities for R)**.** *For $\mu, \nu, \rho \notin T$ and $\mu, \nu \neq \rho$ we have*

$$R^{[T]}_{\mu\nu} \;=\; R^{[T\rho]}_{\mu\nu} + \frac{R^{[T]}_{\mu\rho} R^{[T]}_{\rho\nu}}{R^{[T]}_{\rho\rho}}\,, \qquad \frac{1}{R^{[T]}_{\mu\mu}} \;=\; \frac{1}{R^{[T\rho]}_{\mu\mu}} - \frac{R^{[T]}_{\mu\rho} R^{[T]}_{\rho\mu}}{R^{[T]}_{\mu\mu} R^{[T\rho]}_{\mu\mu} R^{[T]}_{\rho\rho}}\,. \tag{3.10}$$

For $\mu \notin T$ we have

$$\frac{1}{R^{[T]}_{\mu\mu}} \;=\; -z - z \sum_{i,j} X^*_{\mu i} G^{[T\mu]}_{ij} X_{j\mu}\,. \tag{3.11}$$

Moreover, for $\mu, \nu \notin T$ and $\mu \neq \nu$ we have

$$R^{[T]}_{\mu\nu} \;=\; z R^{[T]}_{\mu\mu} R^{[\mu T]}_{\nu\nu} \sum_{i,j} X^*_{\mu i} G^{[T\mu\nu]}_{ij} X_{j\nu}\,. \tag{3.12}$$

Here we draw attention to a fine notational distinction. Parentheses $(\cdot)$ in $X^{(T)}$ indicate removal of rows (indexed by Latin letters), while square brackets $[\cdot]$ in $X^{[T]}$ indicate removal of columns (indexed by Greek letters). In particular, this notation makes it unambiguous whether T is required to be a subset of $\{1, \dots, M\}$ or of $\{1, \dots, N\}$.

The following lemma is an immediate consequence of the fact that for $\phi \geqslant 1$ the spectrum of XX^* is equal to the spectrum of X^*X plus $M - N$ zero eigenvalues. (A similar result holds for for $\phi \leqslant 1$, and if X is replaced with $X^{[T]}$ or $X^{(U)}$.)

Lemma 3.9. *Let $T \subset \{1, \dots, N\}$ and $U \subset \{1, \dots, M\}$. Then we have*

$$\operatorname{Tr} R^{[T]} - \operatorname{Tr} G^{[T]} \;=\; \frac{M - (N - |T|)}{z}$$

as well as

$$\operatorname{Tr} R^{(U)} - \operatorname{Tr} G^{(U)} \;=\; \frac{(M - |U|) - N}{z}$$

In particular, we find

$$\frac{1}{M} \operatorname{Tr} G \;=\; \frac{1}{\phi} \frac{1}{N} \operatorname{Tr} R + \frac{1}{\phi} \frac{1 - \phi}{z}\,, \tag{3.13}$$

in agreement with (2.20) and the heuristics $M^{-1} \operatorname{Tr} G \sim m_{\phi^{-1}}$ and $N^{-1} \operatorname{Tr} R \sim m_\phi$.

The following lemma is an easy consequence of the well-known interlacing property of the eigenvalues of XX^* and $X^{(i)}(X^{(i)})^*$, as well as the eigenvalues of X^*X and $(X^{[\mu]})^* X^{[\mu]}$.

Lemma 3.10 (Eigenvalue interlacing)**.** *For any $T \subset \{1, \dots, N\}$ and $U \subset \{1, \dots, M\}$, there exists a constant C, depending only on $|T|$ and $|U|$, such that*

$$\left|\operatorname{Tr} R^{[T]} - \operatorname{Tr} R\right| \;\leqslant\; C\eta^{-1}\,, \qquad \left|\operatorname{Tr} R^{(U)} - \operatorname{Tr} R\right| \;\leqslant\; C\eta^{-1}\,.$$

Finally, we record the fundamental identity

$$\sum_j \left|G^{[T]}_{ij}\right|^2 \;=\; \frac{1}{\eta} \operatorname{Im} G^{[T]}_{ii}\,, \tag{3.14}$$

which follows easily by spectral decomposition of $G^{[T]}$.

3.2 Reduction to the case $\phi \geqslant 1$

We shall prove Theorems 2.4 and 2.5 by restricting ourselves to the case $\phi \geqslant 1$ but considering both X^*X and XX^* simultaneously. In this short section we give the details of this reduction. Define the control parameter

$$\Psi(z) \;\equiv\; \Psi_\phi(z) \;:=\; \sqrt{\frac{\operatorname{Im} m_\phi(z)}{N\eta}} + \frac{1}{N\eta}\,. \tag{3.15}$$

We shall in fact prove the following. Recall the definitions (2.11) of $\mathbf{S}$ and (2.14) and of $\widetilde{\mathbf{S}}$.

Theorem 3.11. *Suppose that* (2.1), (2.2), (2.3), *and $\phi \geqslant 1$ hold. Then*

$$\big|\langle \mathbf{v}, G(z)\mathbf{w}\rangle - m_{\phi^{-1}}(z)\langle \mathbf{v}, \mathbf{w}\rangle\big| \;\prec\; \frac{1}{\phi}\,\Psi(z) \tag{3.16}$$

uniformly in $z \in \mathbf{S}$ and any deterministic unit vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^M$. Similarly,

$$\big|\langle \mathbf{v}, R(z)\mathbf{w}\rangle - m_{\phi}(z)\langle \mathbf{v}, \mathbf{w}\rangle\big| \;\prec\; \Psi(z) \tag{3.17}$$

uniformly in $z \in \mathbf{S}$ and any deterministic unit vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^N$. Moreover, we have

$$\left|\frac{1}{N}\operatorname{Tr} R(z) - m_\phi(z)\right| \;\prec\; \frac{1}{N\eta}\,, \qquad \left|\frac{1}{M}\operatorname{Tr} G(z) - m_{\phi^{-1}}(z)\right| \;\prec\; \frac{1}{M\eta} \tag{3.18}$$

uniformly in $z \in \mathbf{S}$.

Theorem 3.12. *Suppose that* (2.1), (2.2), (2.3), *and $\phi \geqslant 1$ hold. Then*

$$\big|\langle \mathbf{v}, G(z)\mathbf{w}\rangle - m_{\phi^{-1}}(z)\langle \mathbf{v}, \mathbf{w}\rangle\big| \;\prec\; \frac{1}{\phi}\sqrt{\frac{\operatorname{Im} m_\phi(z)}{N\eta}} \tag{3.19}$$

uniformly in $z \in \widetilde{\mathbf{S}}$ and any deterministic unit vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^M$. Similarly,

$$\big|\langle \mathbf{v}, R(z)\mathbf{w}\rangle - m_{\phi}(z)\langle \mathbf{v}, \mathbf{w}\rangle\big| \;\prec\; \sqrt{\frac{\operatorname{Im} m_\phi(z)}{N\eta}} \tag{3.20}$$

uniformly in $z \in \widetilde{\mathbf{S}}$ and any deterministic unit vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^N$.

Let $\phi \geqslant 1$, i.e. $N \leqslant M$. Recall that $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(N)} \in \mathbb{C}^N$ denote the normalized eigenvectors of X^*X associated with the nontrivial eigenvalues $\lambda_1, \dots, \lambda_N$, and let $\widetilde{\mathbf{u}}^{(1)}, \dots, \widetilde{\mathbf{u}}^{(N)} \in \mathbb{C}^M$ denote the normalized eigenvectors of XX^* associated with the same eigenvalues $\lambda_1, \dots, \lambda_N$.

Theorem 3.13. *Suppose that* (2.1), (2.2), (2.3), *and $\phi \geqslant 1$ hold. For any $\varepsilon > 0$ we have the bounds*

$$|\langle \mathbf{u}^{(\alpha)}, \mathbf{v}\rangle|^2 \;\prec\; N^{-1}\,, \qquad |\langle \widetilde{\mathbf{u}}^{(\alpha)}, \mathbf{w}\rangle|^2 \;\prec\; M^{-1} \tag{3.21}$$

uniformly for $\alpha \leqslant (1-\varepsilon)N$ and all normalized $\mathbf{v} \in \mathbb{C}^N$ and $\mathbf{w} \in \mathbb{C}^M$. If in addition $\phi \geqslant 1 + c$ for some constant $c > 0$, then (3.21) *holds uniformly for all $\alpha \leqslant N$.*

Theorems 2.4, 2.5, and 2.8 are easy consequences of Theorems 3.11, 3.12, and 3.13 respectively, combined with the observation that

$$\operatorname{Im} m_{\phi^{-1}}(z) \;\asymp\; \frac{1}{\phi}\operatorname{Im} m_\phi(z) \tag{3.22}$$

for $z \in \mathbf{S}$; in addition, the asymptotic equivalence in (2.15) follows from (3.6) and (3.22). The estimate (3.22) itself can be proved by noting that (2.20) implies

$$\operatorname{Im} m_{\phi^{-1}}(z) \;=\; \frac{1}{\phi}\left(\operatorname{Im} m_\phi(z) + \frac{\phi-1}{|z|^2}\eta\right).$$

Since $|z|^2 \asymp \phi$ for $z \in \mathbf{S}$, (3.22) for $\phi \geqslant 1$ follows from Lemma 3.3. Replacing ϕ with ϕ^{-1} in (3.22), we conclude that (3.22) holds for all ϕ.

What remains therefore is to prove Theorems 2.10, 3.11, 3.12, and 3.13. We shall prove Theorem 2.10 in Section 4.3, Theorem 3.11 in Section 5, and Theorems 3.12 and 3.13 in Section 6.

4 The entrywise local Marchenko-Pastur law

In this section we prove a entrywise version of Theorem 3.11, in which the vectors $\mathbf{v}$ and $\mathbf{w}$ from (3.17) and (3.16) are assumed to lie in the direction of a coordinate axis. A similar result was previously proved in [21, Theorem 3.1]. Recall the definition of Ψ from (3.15).

Theorem 4.1 (Entrywise local Marchenko-Pastur law)**.** *Suppose that* (2.1)*,* (2.2)*,* (2.3)*, and $\phi \geqslant 1$ hold. Then*

$$\left|R_{\mu\nu}(z) - \delta_{\mu\nu} m_\phi(z)\right| \;\prec\; \Psi(z)\,, \tag{4.1}$$

uniformly in $z \in \mathbf{S}$ and $\mu, \nu \in \{1, \dots, N\}$. Similarly,

$$\left|G_{ij}(z) - \delta_{ij} m_{\phi^{-1}}(z)\right| \;\prec\; \frac{1}{\phi}\Psi(z)\,, \tag{4.2}$$

uniformly in $z \in \mathbf{S}$ and $i, j \in \{1, \dots, M\}$. Moreover, (3.18) *holds uniformly in $z \in \mathbf{S}$.*

Theorem 4.1 differs from Theorem 3.1 in [21] in the following two ways.

1. The restriction $1 \leqslant \phi \leqslant C$ in [21] is relaxed to $1 \leqslant \phi \leqslant N^C$ (and hence, as explained in Section 3.2, to $N^{-C} \leqslant \phi \leqslant N^C$).
2. The uniform subexponential decay assumption of [21] is relaxed to (2.3). On the other hand, thanks to the stronger subexponential decay assumption the statement of Theorem 3.1 of [21] is slightly stronger than Theorem 4.1: in Theorem 3.1 of [21], the error bounds N^ε in the definition of $\prec$ are replaced with $(\log N)^{C \log\log N}$.

The difference (ii) given above is technical and amounts to using Lemma 3.1, which is tailored for random variables satisfying (2.3), for the large deviation estimates. We remark that all of the arguments of the current paper may be translated to the setup of [21], explained in (ii) above, by modifying the definition of $\prec$. The essence of the proofs remains unchanged; the only nontrivial difference is that in Section 5 we have to control moments whose power depends weakly on N; this entails keeping track of some basic combinatorial bounds. We do not pursue this modification any further.

The difference (i) is more substantial, and requires to keep track of the ϕ-dependence of all appropriately rescaled quantities throughout the proof. In addition, we take this opportunity to simplify and streamline the argument from [21]. This provides a short and self-contained proof of Theorem 4.1, up to a fluctuation averaging result, Lemma 4.9 below, which was proved in the current simple and general form in [9].

4.1 A weak local Marchenko-Pastur law

We begin with the proof of (4.1) and (3.18). For the following it will be convenient to use the rescaled spectral parameters

$$\widetilde{z} \;:=\; \phi^{-1/2} z\,, \qquad \widehat{z} \;:=\; z - \phi^{1/2} + \phi^{-1/2}\,. \tag{4.3}$$

Using $\widetilde{z}$ and $\widehat{z}$ we may write the the defining equation (2.7) of m_ϕ as

$$m(z) + \frac{1}{\widehat{z} + \widetilde{z} m(z)} \;=\; 0\,. \tag{4.4}$$

From the definition (2.11) of $\mathbf{S}$, we find

$$|\widetilde{z}| \;\asymp\; 1\,, \qquad |\widehat{z}| \;\leqslant\; C \tag{4.5}$$

for all $z \in \mathbf{S}$. We remark that, as in [21], the Stieltjes transform m_ϕ satisfies $|m_\phi(z)| \asymp 1$ for $z \in \mathbf{S}$; see (3.5).

We define the z-dependent random control parameters

$$\Lambda(z) \;:=\; \max_{\mu,\nu} |R_{\mu\nu}(z) - \delta_{\mu\nu} m_\phi(z)|\,, \tag{4.6a}$$

$$\Lambda_o(z) \;:=\; \max_{\mu \neq \nu} |R_{\mu\nu}(z)|\,, \tag{4.6b}$$

$$\Theta(z) \;:=\; |m_R(z) - m_\phi(z)|\,, \tag{4.6c}$$

where we defined the Stieltjes transform of the empirical density of $X^* X$,

$$m_R(z) \;:=\; \frac{1}{N} \operatorname{Tr} R(z)\,.$$

The goal of this subsection is to prove the following weaker variant of Theorem 4.1.

Proposition 4.2. *We have $\Lambda \prec (N\eta)^{-1/4}$ uniformly in $z \in \mathbf{S}$.*

The rest of this subsection is devoted to the proof of Proposition 4.2. We begin by introducing the basic z-dependent event

$$\Xi(z) \;:=\; \big\{ \Lambda(z) \leqslant (\log N)^{-1} \big\}\,.$$

Lemma 4.3. *For any $\ell \in \mathbb{N}$ there exists a constant $C \equiv C_\ell$ such that for $z \in \mathbf{S}$, all $T \subset \{1, 2, \ldots, N\}$ satisfying $|T| \leqslant \ell$, and all $\mu, \nu \notin T$ we have*

$$\mathbf{1}(\Xi) \big| R^{[T]}_{\mu\nu} - R_{\mu\nu} \big| \;\leqslant\; C \Lambda_o^2 \tag{4.7}$$

and

$$\mathbf{1}(\Xi) C^{-1} \;\leqslant\; \mathbf{1}(\Xi) |R^{[T]}_{\mu\mu}| \;\leqslant\; C \tag{4.8}$$

for large enough N depending on ℓ.

Proof. The proof is a simple induction argument using (3.10) and the bound $|m_\phi| \geqslant c$ from (3.5). We omit the details. □

As in the works [16, 21], the main idea of the proof is to derive a self-consistent equation for $m_R = \frac{1}{N} \sum_\mu R_{\mu\mu}$ using the resolvent identity (3.11). To that end, we introduce the conditional expectation

$$\mathbb{E}^{[\mu]}(\cdot) \;:=\; \mathbb{E}(\,\cdot\,|X^{[\mu]})\,, \tag{4.9}$$

i.e. the partial expectation in the randomness of the μ-th column of X. We define

$$Z_\mu \;:=\; \big(1-\mathbb{E}^{[\mu]}\big) z \sum_{i,j} X^*_{\mu i} G^{[\mu]}_{ij} X_{j\mu} \;=\; z \sum_{i,j} X^*_{\mu i} G^{[\mu]}_{ij} X_{j\mu} - \frac{\widetilde{z}}{N} \operatorname{Tr} G^{[\mu]}\,, \tag{4.10}$$

where in the last step we used (2.1) and (4.3). Using (3.11) with $T = \emptyset$, Lemma 3.9, and (4.3), we find

$$\frac{1}{R_{\mu\mu}} \;=\; -z - \frac{\widetilde{z}}{N} \operatorname{Tr} G^{[\mu]} - Z_\mu \;=\; -\widehat{z} - \frac{\widetilde{z}}{N} \operatorname{Tr} R^{[\mu]} - Z_\mu - \frac{1}{\sqrt{\phi} N}\,. \tag{4.11}$$

The following lemma contains the key estimates needed to control the error terms Z_μ and Λ_o. The errors are controlled using of the (random) control parameter

$$\Psi_\Theta \;:=\; \sqrt{\frac{\operatorname{Im} m_\phi + \Theta}{N\eta}}\,, \tag{4.12}$$

whose analogue in the context of Wigner matrices first appeared in [16].

Lemma 4.4. *For $z \in \mathbf{S}$ we have*

$$\mathbf{1}(\Xi)\big(|Z_\mu| + \Lambda_o\big) \;\prec\; \Psi_\Theta \tag{4.13}$$

as well as

$$\mathbf{1}\big(\eta \geqslant (\log N)^{-1}\big)\,\big(|Z_\mu| + \Lambda_o\big) \;\prec\; \Psi_\Theta\,. \tag{4.14}$$

Proof. The proof is very similar to that of Theorems 6.8 and 6.9 of [21]. We consequently only give the details for the estimate of Λ_o; the estimate of Z_μ is similar.

For $\mu \neq \nu$ we use (3.12) with $T = \emptyset$ to expand $R_{\mu\nu}$. Conditioning on $X^{[\mu\nu]}$ and invoking (3.3) yields

$$|R_{\mu\nu}| \;\prec\; \big|R_{\mu\mu} R^{[\mu]}_{\nu\nu}\big|\, \frac{|\widetilde{z}|}{N} \sqrt{\sum_{i,j} \big|G^{[\mu\nu]}_{ij}\big|^2}\,. \tag{4.15}$$

On the event Ξ, we estimate the right-hand side using

$$\begin{aligned}
\mathbf{1}(\Xi) \frac{|\widetilde{z}|}{N} \sqrt{\sum_{i,j} \big|G^{[\mu\nu]}_{ij}\big|^2} \;&=\; \mathbf{1}(\Xi) \frac{|\widetilde{z}|}{N} \sqrt{\frac{\operatorname{Im} \operatorname{Tr} G^{[\mu\nu]}}{\eta}} \\
&\leqslant\; \mathbf{1}(\Xi) \frac{C}{N} \sqrt{\frac{\operatorname{Im} \operatorname{Tr} R^{[\mu\nu]} - ((\phi-1)N+2) \operatorname{Im} z^{-1}}{\eta}} \\
&\leqslant\; C\mathbf{1}(\Xi) \sqrt{\frac{\operatorname{Im} m_\phi + \Theta + \Lambda_o^2}{N\eta} + \frac{1}{N}} \\
&\leqslant\; C\mathbf{1}(\Xi) \sqrt{\frac{\operatorname{Im} m_\phi + \Theta + \Lambda_o^2}{N\eta}}\,,
\end{aligned} \tag{4.16}$$

where the first step follows from (3.14), the second from Lemma 3.9, the third from $\operatorname{Im} z^{-1} = -\eta |z|^{-2} \geqslant -C\eta/\phi$ and (4.7), and the fourth from the fact that $\operatorname{Im} m_\phi \geqslant c\eta$ by (3.6).

Recalling (3.5), we have therefore proved that

$$\mathbf{1}(\Xi)\Lambda_o \;\prec\; \mathbf{1}(\Xi)\big(\Psi_\Theta + (N\eta)^{-1/2}\Lambda_o\big)\,. \tag{4.17}$$

Since $(N\eta)^{-1/2} \leqslant N^{-\omega/2}$ on $\mathbf{S}$, we find

$$\mathbf{1}(\Xi)\Lambda_o \;\prec\; \mathbf{1}(\Xi)\Psi_\Theta\,,$$

which, together with the analogous bound for Z_μ, concludes the proof of (4.13).

In order to prove the estimate $\Lambda_o \prec \Psi_\Theta$ from (4.14) for $\eta \geqslant (\log N)^{-1}$, we proceed similarly. From (4.15) and the trivial deterministic bound $\left|R_{\mu\mu}R^{[\mu]}_{\nu\nu}\right| \leqslant \eta^{-2} \leqslant (\log N)^2$ we get

$$|R_{\mu\nu}| \;\prec\; \frac{1}{N}\sqrt{\frac{\operatorname{Im}\operatorname{Tr} G^{[\mu\nu]}}{\eta}} \;=\; \frac{1}{N}\sqrt{\frac{\operatorname{Im}\operatorname{Tr} R^{[\mu\nu]} - ((\phi-1)N+2)\operatorname{Im} z^{-1}}{\eta}} \leqslant\; C\sqrt{\frac{\operatorname{Im} m_\phi + \Theta}{N\eta} + \frac{1}{(N\eta)^2}}\,,$$

where the estimate is similar to (4.16), except that in the last step we use Lemma 3.10 to estimate $\operatorname{Tr} R^{\mu\nu} - \operatorname{Tr} R$. Since $\eta \geqslant (\log N)^{-1}$, we easily find that $|R_{\mu\nu}| \prec \Psi_\Theta$. This concludes the proof. □

As in [21, Equation (6.13)], in order to analyse the stability of the equation (4.4) we introduce the operation $\mathcal{D}$ on functions $u : \mathbf{S} \to \mathbb{C}$, defined through

$$\mathcal{D}(u)(z) \;:=\; \frac{1}{u(z)} + \widetilde{z}u(z) + \widehat{z}\,. \tag{4.18}$$

Note that, by (4.4), the function m_ϕ satisfies $\mathcal{D}(m_\phi) = 0$.

Next, we derive a stability result for $\mathcal{D}^{-1}$. Roughly, we prove that if $\mathcal{D}(u)$ is small then u is close to m_ϕ. Note that this result is entirely deterministic. It relies on a discrete continuity argument, whose essence is the existence of a sufficiently large gap between the two solutions of $\mathcal{D}(\cdot) = 0$. Once this gap is established, then, together with the fact that u is close to m_ϕ for large η, we may conclude that u is close to m_ϕ for smaller η as well. We use a discrete version of a continuity argument (as opposed to a continuous one used e.g. in [21]), which allows us to bypass several technical issues when applying it to estimating the random quantity $|m_R - m_\phi|$. For more details of this application, see the explanation following (4.34).

For $z \in \mathbf{S}$ introduce the discrete set

$$L(z) \;:=\; \{z\} \cup \left\{w \in \mathbf{S} : \operatorname{Re} w = \operatorname{Re} z\,,\, \operatorname{Im} w \in [\operatorname{Im} z, 1] \cap (N^{-5}\mathbb{N})\right\}.$$

Thus, if $\operatorname{Im} z \geqslant 1$ then $L(z) = \{z\}$ and if $\operatorname{Im} z \leqslant 1$ then $L(z)$ is a one-dimensional lattice with spacing N^{-5} plus the point z. Clearly, we have the bound

$$|L(z)| \;\leqslant\; N^5\,. \tag{4.19}$$

Lemma 4.5 (Stability of $\mathcal{D}^{-1}$)**.** *There exists a constant $\varepsilon > 0$ such that the following holds. Suppose that $\delta : \mathbf{S} \to \mathbb{C}$ satisfies $N^{-2} \leqslant \delta(z) \leqslant \varepsilon$ for $z \in \mathbf{S}$ and that δ is Lipschitz continuous with Lipschitz constant N. Suppose moreover that for each fixed E, the function $\eta \mapsto \delta(E+\mathrm{i}\eta)$ is nonincreasing for $\eta > 0$. Suppose that $u : \mathbf{S} \to \mathbb{C}$ is the Stieltjes transform of a probability measure. Let $z \in \mathbf{S}$, and suppose that for all $w \in L(z)$ we have*

$$\left|\mathcal{D}(u)(w)\right| \;\leqslant\; \delta(w)\,.$$

Then we have

$$|u(z) - m_\phi(z)| \;\leqslant\; \frac{C\delta(z)}{\sqrt{\kappa + \eta + \delta(z)}}\,.$$

for some constant C independent of z and N.

Proof. Let u be as in Lemma 4.5, and abbreviate $\mathcal{R} := \mathcal{D}(u)$. Hence, by assumption on u, we have $|\mathcal{R}| \leqslant \delta$. We introduce $u_1 \equiv u_1^{\mathcal{R}}$ and $u_2 \equiv u_2^{\mathcal{R}}$ by setting $u_1 := u$ and defining u_2 as the other solution of the quadratic equation $\mathcal{D}(u) = \mathcal{R}$. Note that each u_i is continuous. Explicitly, for $|\mathcal{R}| \leqslant 1/2$ we get

$$u_{1,2} \;=\; \frac{\mathcal{R} - \widehat{z} \pm \mathrm{i}\sqrt{(z-\lambda_{-,\mathcal{R}})(\lambda_{+,\mathcal{R}}-z)}}{2\widetilde{z}}\,, \qquad \lambda_{\pm,\mathcal{R}} \;:=\; \phi^{1/2}+\phi^{-1/2}+\mathcal{R}\pm 2\sqrt{1+\phi^{-1/2}\mathcal{R}}\,, \tag{4.20}$$

where the square root in $\sqrt{1+\phi^{-1/2}\mathcal{R}}$ is the principal branch. (Note that the sign $\pm$ in the expression for $u_{1,2}$ bears no relation to the indices $1, 2$, since we have not even specified which complex square root we take.) In particular, for $\mathcal{R} = 0$ we have $\lambda_{\pm,\mathcal{R}=0} = \gamma_\pm$, defined in (2.6). Observe that for any complex square root $\sqrt{\cdot}$ and $w, \zeta \in \mathbb{C}$ we have $|\sqrt{w+\zeta} - \sqrt{w}| \leqslant (|w|+|\zeta|)^{-1/2}|\zeta|$ or $|\sqrt{w+\zeta} + \sqrt{w}| \leqslant (|w|+|\zeta|)^{-1/2}|\zeta|$. We use these formulas to compare (4.20) with a small $\mathcal{R}$ with (4.20) with $\mathcal{R} = 0$. Thus we conclude from (4.20) and (4.5) that for $i = 1$ or for $i = 2$ we have

$$|u_i - m_\phi| \;\leqslant\; \frac{C_0|\mathcal{R}|}{\sqrt{\kappa+\eta+|\mathcal{R}|}} \tag{4.21}$$

for some constant $C_0 \geqslant 2$. What remains is to show that (4.21) holds for $i = 1$. We shall prove this using a continuity argument.

Note first that (4.20) and (4.5) yield

$$C_1^{-1}\sqrt{(\kappa+\eta-|\mathcal{R}|)_+} \;\leqslant\; |u_1 - u_2| \;\leqslant\; C_1\sqrt{\kappa+\eta+|\mathcal{R}|} \tag{4.22}$$

for some constant $C_1 \geqslant 1$.

Now consider $z = \mathrm{i}$. Clearly, for $\mathcal{R}(\mathrm{i}) = 0$ we have $u_1^0(\mathrm{i}) = m_\phi(\mathrm{i})$. Note that by the lower bound of (4.22) the two roots $u_1^{\mathcal{R}}(\mathrm{i})$ and $u_2^{\mathcal{R}}(\mathrm{i})$ are distinct, and they are continuous in $\mathcal{R}$. Therefore there is an $\varepsilon \in (0, 1/2]$ such that for $|\mathcal{R}(\mathrm{i})| \leqslant \varepsilon$ we have, after possibly increasing C_0, that

$$|u_1 - m_\phi| \;\leqslant\; C_0|\mathcal{R}| \tag{4.23}$$

at $z = \mathrm{i}$. Next, we note that (4.21) and (4.22) imply, for any z with $\operatorname{Im} z \geqslant 1$, that $|u_i - m_\phi| \leqslant C_0|\mathcal{R}|$ for some $i \in \{1, 2\}$, and that $|u_1 - u_2| \geqslant (2C_1)^{-1}$. Hence, requiring that $\varepsilon \leqslant (8C_0C_1)^{-1}$ we find from (4.23) with $z = \mathrm{i}$ and using the continuity of u_1 that (4.23) holds provided $\operatorname{Im} z \geqslant 1$.

Next, for arbitrary $z \in \mathbf{S}$ with $\operatorname{Im} z < 1$ we consider two cases, depending on whether

$$\frac{C_0\delta}{\sqrt{\kappa+\eta+\delta}} \;\leqslant\; \frac{1}{4C_1}\sqrt{(\kappa+\eta-\delta)_+} \tag{4.24}$$

holds or not. If (4.24) does not hold, then we have $\kappa + \eta \leqslant 4C_0C_1\delta$, so that (4.21), $|\mathcal{R}| \leqslant \delta$, and the upper bound of (4.22) imply

$$|u_1 - m_\phi| \;\leqslant\; \frac{C_0|\mathcal{R}|}{\sqrt{\kappa+\eta+\delta}} + C_1\sqrt{\kappa+\eta+|\mathcal{R}|} \;\leqslant\; C\sqrt{\delta} \;\leqslant\; \frac{C\delta}{\sqrt{\kappa+\eta+\delta}}\,.$$

What remains is the case where (4.24) holds. We use a continuity argument along the set $L(z)$, which we parametrize as $L(z) = \{z_0, \ldots, z_L\}$, where $\operatorname{Im} z_0 = 1$, $z_L = z$, and $\operatorname{Im} z_{l+1} < \operatorname{Im} z_l$. Note that $|z_{l+1} - z_l| \leqslant N^{-5}$. By assumption, $|\mathcal{R}| \leqslant \delta$ at each $z_l \in L(z)$, so that (4.21) and (4.22) yield

$$\exists\, i = 1, 2 \;:\; |u_i - m_\phi| \;\leqslant\; \frac{C_0\delta}{\sqrt{\kappa+\eta+\delta}}\,, \qquad C_1^{-1}\sqrt{(\kappa+\eta-\delta)_+} \;\leqslant\; |u_1 - u_2| \tag{4.25}$$

at each $z_l \in L(z)$. (Here the quantities $\kappa \equiv \kappa(z_l)$, $\eta \equiv \eta(z_l)$, and $\delta \equiv \delta(z_l)$ are understood as functions of the spectral parameters z_l.) Moreover, since (4.24) holds at $z = z_L$, by

the monotonicity assumption on δ we find that (4.24) holds for all $z_l \in L(z)$. We now prove that

$$|u_1(z_l) - m_\phi(z_l)| \;\leqslant\; \frac{C_0\delta}{\sqrt{\kappa+\eta+\delta}}\bigg|_{z_l} \tag{4.26}$$

for all $l = 1, \dots, L$ by induction on l. For $l = 0$ the bound (4.26) is simply (4.23) proved above. Suppose therefore that (4.26) holds for some l. Since u_1 and m_ϕ are Lipschitz continuous with Lipschitz constant N, we get

$$\begin{aligned} |u_1(z_{l+1}) - m_\phi(z_{l+1})| \;&\leqslant\; 2NN^{-5} + \big|u_1(z_l) - m_\phi(z_l)\big| \;\leqslant\; 2N^{-4} + \frac{C_0\delta}{\sqrt{\kappa+\eta+\delta}}\bigg|_{z_l} \\ &\leqslant\; 2N^{-4} + CN^2N^{-5} + \frac{C_0\delta}{\sqrt{\kappa+\eta+\delta}}\bigg|_{z_{l+1}} \;\leqslant\; \frac{2C_0\delta}{\sqrt{\kappa+\eta+\delta}}\bigg|_{z_{l+1}}, \end{aligned} \tag{4.27}$$

where in the second step we used the induction assumption, in the third step the Lipschitz continuity of δ and the bound $\eta \geqslant N^{-1}$, and in the last step the bounds $\delta \geqslant N^{-2}$ and $\kappa + \eta + \delta \leqslant C$. Next, recalling (4.24), it is easy to deduce (4.26) with l replaced by $l+1$, using the bounds (4.25) and (4.27). This concludes the proof. □

We may now combine the probabilistic estimates from Lemma 4.4 with the stability of $\mathcal{D}^{-1}$ from Lemma 4.5 to get the following result for $\eta \geqslant 1$, which will be used as the starting estimate in the bootstrapping in η.

Lemma 4.6. *We have $\Lambda \prec N^{-1/4}$ uniformly in $z \in \mathbf{S}$ satisfying $\operatorname{Im} z \geqslant 1$.*

Proof. Let $z \in \mathbf{S}$ with $\operatorname{Im} z \geqslant 1$. From (4.11) and the estimate on Z_μ from (4.14) we find

$$R_{\mu\mu} \;=\; \frac{1}{-\widehat{z} - \widetilde{z}m_R + O_\prec(\Psi_\Theta + N^{-1})} \;=\; \frac{1}{-\widehat{z} - \widetilde{z}m_R + O_\prec(N^{-1/2})}\,, \tag{4.28}$$

where in the last step we used that $\Psi_\Theta = O(N^{-1/2})$ since $\eta \geqslant 1$ and $\operatorname{Im} m_\phi + \Theta = O(1)$, as follows from (3.5) and the trivial bound $|m_R| \leqslant C$. Taking the average over μ yields $1/m_R = -\widehat{z} - \widetilde{z}m_R + O_\prec(N^{-1/2})$, i.e. $|\mathcal{D}(m_R)| \prec N^{-1/2}$; see (4.18). Since $L(z) = \{z\}$, we therefore get from Lemma 4.5 that $|m_R - m_\phi| \prec N^{-1/4}$. Returning to (4.28) and recalling (4.4) and (3.5), we get $|R_{\mu\mu} - m_\phi| \prec N^{-1/4}$. Together with the estimate on Λ_o from (4.14), we therefore get $\Lambda \prec N^{-1/4}$ uniformly in $z \in \mathbf{S}$ satisfying $\operatorname{Im} z \geqslant 1$. □

Next, we plug the estimates from Lemma 4.4 into (4.11) in order to obtain estimates on m_R. The summation in $m_R = \frac{1}{N}\sum_\mu R_{\mu\mu}$ will give rise to an error term of the form

$$[Z] \;:=\; \frac{1}{N}\sum_\mu Z_\mu\,. \tag{4.29}$$

For the proof of Proposition 4.2, it will be enough to estimate $|[Z]| \leqslant \max_\mu |Z_\mu|$, but for the eventual proof of Theorem 4.1, we shall need to exploit cancellation in the averaging in $[Z]$. Bearing this in mind, we state our estimates in terms of $[Z]$ to avoid repeating the following argument in Section 4.2.

Lemma 4.7. *We have*

$$\mathbf{1}(\Xi)|R_{\mu\mu} - m_R| \;\prec\; \Psi_\Theta \tag{4.30}$$

uniformly in μ and $z \in \mathbf{S}$, as well as

$$\mathbf{1}(\Xi)\mathcal{D}(m_R) \;=\; \mathbf{1}(\Xi)\big(-[Z] + O_\prec(\Psi_\Theta^2)\big) \tag{4.31}$$

uniformly in $z \in \mathbf{S}$.

Proof. From (4.11), (4.7), and (4.13) we get

$$\mathbf{1}(\Xi)\frac{1}{R_{\mu\mu}} \;=\; \mathbf{1}(\Xi)\Big(-\widehat{z}-\widetilde{z}m_R - Z_\mu + O_\prec(\Psi_\Theta^2)\Big)\,, \tag{4.32}$$

where we absorbed the error term N^{-1} on the right-hand side of (4.11) into Ψ_Θ^2 using (3.6). Thus, using (4.13) we get

$$\mathbf{1}(\Xi)(R_{\mu\mu}-R_{\nu\nu}) \;=\; \mathbf{1}(\Xi)R_{\mu\mu}R_{\nu\nu}O_\prec(\Psi_\Theta) \;=\; O_\prec(\Psi_\Theta)\,.$$

Hence (4.30) follows. Next, expanding $R_{\mu\mu} = m_R + (R_{\mu\mu}-m_R)$ yields

$$\frac{1}{R_{\mu\mu}} \;=\; \frac{1}{m_R} - \frac{1}{m_R^2}(R_{\mu\mu}-m_R) + \frac{1}{m_R^2}(R_{\mu\mu}-m_R)^2\frac{1}{R_{\mu\mu}}\,.$$

After taking the average $[\cdot] = \frac{1}{N}\sum_\mu \cdot$ the second term on the right-hand side vanishes. Taking the average of (4.32) therefore yields, using (4.30) and (4.3),

$$\mathbf{1}(\Xi)\bigg(\frac{1}{m_R}+O_\prec(\Psi_\Theta^2)\bigg) \;=\; \mathbf{1}(\Xi)\Big(-\widehat{z}-\widetilde{z}m_R-[Z]+O_\prec(\Psi_\Theta^2)\Big)\,,$$

from which the claim follows. □

From (4.31) and (4.13) we get

$$\mathbf{1}(\Xi)|\mathcal{D}(m_R)| \;\prec\; \mathbf{1}(\Xi)\Psi_\Theta \;\leqslant\; C(N\eta)^{-1/2} \tag{4.33}$$

uniformly in $\mathbf{S}$. In order to conclude the proof of Proposition 4.2, we use a continuity argument. The main ingredients are (4.33), Lemma 4.5, Lemma 4.6. Choose $\varepsilon < \omega/4$ and an arbitrary $D > 0$. It is convenient to introduce the random function

$$v(z) \;:=\; \max_{w\in L(z)} \Lambda(w)(N\operatorname{Im} w)^{1/4}\,.$$

Our goal is to prove that with high probability there is a gap in the range of v, i.e.

$$\mathbb{P}\big(v(z)\leqslant N^\varepsilon\,,\, v(z) > N^{\varepsilon/2}\big) \;\leqslant\; N^{-D+5} \tag{4.34}$$

for all $z\in\mathbf{S}$ and large enough $N\geqslant N_0(\varepsilon,D)$. This equation says that with high probability the range of v has a gap: it cannot take values in the interval $(N^{\varepsilon/2}, N^\varepsilon]$.

The basic idea behind the proof of (4.34) is to use the deterministic result from Lemma 4.5 to propagate smallness of the random variable $\Lambda(z)$ from large values of η to smaller values of η. Since we are dealing with random variables, one has to keep track of probabilities of exceptional events. To that end, we only work on a discrete set of values of η, which allows us to control the exceptional probabilities by a simple union bound. We remark that the first instance of such a *stochastic continuity argument* combined with stability of a self-consistent equation was given in [11] in the context of Wigner matrices. Over the years it has been improved through several papers in the context of Wigner matrices [9, 15, 16] as well as in the context of sample covariance matrices [13, 21].

Next, we prove (4.34). Since $\{v(z)\leqslant N^\varepsilon\}\subset \Xi(z)\cap\Xi(w)$ for all $z\in\mathbf{S}$ and $w\in L(z)$, we find that (4.33) implies for all $z\in\mathbf{S}$ and $w\in L(z)$ that

$$\mathbb{P}\Big(v(z)\leqslant N^\varepsilon\,,\, |\mathcal{D}(m_R)(w)|(N\operatorname{Im} w)^{1/2} > N^{\varepsilon/2}\Big) \;\leqslant\; N^{-D}$$

for large enough $N\geqslant N_0(\varepsilon,D)$ (independent of z and w). Using (4.19) and a union bound, we therefore get

$$\mathbb{P}\Big(v(z)\leqslant N^\varepsilon\,,\, \max_{w\in L(z)}|\mathcal{D}(m_R)(w)|(N\operatorname{Im} w)^{1/2} > N^{\varepsilon/2}\Big) \;\leqslant\; N^{-D+5}\,.$$

Next, we use Lemma 4.5 with $u = m_R$ and $\delta = N^{\varepsilon/2}(N\eta)^{-1/2}$ to get

$$\mathbb{P}\Big(v(z) \leqslant N^{\varepsilon}\,,\ \max_{w\in L(z)} \Theta(w)(N\operatorname{Im} w)^{1/4} > N^{\varepsilon/4}\Big) \ \leqslant\ N^{-D+5}\,.$$

(Here we used the trivial observation that the conclusion of Lemma 4.5 is valid not only at z but in the whole set $L(z)$.) Using (4.13) and (4.30) we therefore get (4.34).

We conclude the proof of Proposition 4.2 by combining (4.34) and Lemma 4.6 with a continuity argument, similar to the proof of [9, Proposition 5.3]. We choose a lattice $\Delta \subset \mathbf{S}$ such that $|\Delta| \leqslant N^{10}$ and for each $z \in \mathbf{S}$ there exists a $w \in \Delta$ satisfying $|z - w| \leqslant N^{-4}$. Then (4.34) combined with a union bound yields

$$\mathbb{P}\big(\exists w \in \Delta : v(w) \in (N^{\varepsilon/2}, N^{\varepsilon}]\big) \ \leqslant\ N^{-D+15}\,. \tag{4.35}$$

From the definitions of Λ and $\mathbf{S}$ we find that v is Lipschitz continuous on $\mathbf{S}$, with Lipschitz constant N^2. Hence (4.35) and the definition of Δ imply

$$\mathbb{P}\big(\exists z \in \mathbf{S} : v(z) \in (2N^{\varepsilon/2}, 2N^{\varepsilon}/2]\big) \ \leqslant\ N^{-D+15}\,. \tag{4.36}$$

By Lemma 4.6, we have

$$\mathbb{P}\big(v(z) > N^{\varepsilon/2}\big) \ \leqslant\ N^{-D+15} \tag{4.37}$$

for some (in fact any) $z \in \mathbf{S}$ satisfying $\operatorname{Im} z \geqslant 1$. It is not hard to infer from (4.36) and (4.37) that

$$\mathbb{P}\Big(\max_{z\in\mathbf{S}} v(z) > 2N^{\varepsilon/2}\Big) \ \leqslant\ 2N^{-D+15}\,. \tag{4.38}$$

Since ε can be made arbitrarily small and D arbitrarily large, Proposition 4.2 follows from (4.38).

4.2 Fluctuation averaging and conclusion of the proof of Theorem 4.1

In order to improve the negative power of $(N\eta)$ in Proposition 4.2, and hence prove the optimal bound in Theorem 4.1, we shall use the following result iteratively. Recall the definition of Θ from (4.6) and definition of $[Z]$ from (4.29).

Lemma 4.8. *Let $\tau \in (0,1)$ and suppose that we have*

$$\Theta \ \prec\ (N\eta)^{-\tau} \tag{4.39}$$

uniformly in $z \in \mathbf{S}$. Then we have

$$|[Z]| \ \prec\ \frac{\operatorname{Im} m_\phi + (N\eta)^{-\tau}}{N\eta} \tag{4.40}$$

uniformly in $z \in \mathbf{S}$.

In order to prove Lemma 4.8, we invoke the following fluctuation averaging result. We remark that the fluctuation averaging mechanism was first exploited in [14]. Here we use the result from [9], where a general version with a streamlined proof was given. Recall the definition of the partial expectation $\mathbb{E}^{[\mu]}$ from (4.9).

Lemma 4.9 (Fluctuation averaging [9])**.** *Suppose that Φ and Φ_o are positive, N-dependent, deterministic functions on $\mathbf{S}$ satisfying $N^{-1/2} \leqslant \Phi, \Phi_o \leqslant N^{-c}$ for some constant $c > 0$. Suppose moreover that $\Lambda \prec \Phi$ and $\Lambda_o \prec \Phi_o$. Then*

$$\frac{1}{N}\sum_{\mu}\big(1 - \mathbb{E}^{[\mu]}\big)\frac{1}{R_{\mu\mu}} \ =\ O_{\prec}(\Phi_o^2)\,. \tag{4.41}$$

Proof. This result was given in a slightly different context in Theorem 4.7 in [9]. However, it is a triviality that the proof of Theorem 4.7 in [9] carries over word for word, provided one replaces $G_{ij}^{(T)}$ there with $R_{\mu\nu}^{[T]}$; see Remark B.3 in [9]. The proof relies only on the identity (3.10), which is the analogue of Equation (4.6) in [9]. □

Remark 4.10. *The conclusion of Lemma 4.9 remains true under somewhat more general hypotheses, whereby* Λ *is not required to be small. Indeed,* (4.41) *holds provided that* Φ_o *is as in Lemma 4.9 and that*

$$\left|\frac{1}{R_{\mu\mu}}\right| \prec 1\,, \qquad \left|\left(1-\mathbb{E}^{[\mu]}\right)\frac{1}{R_{\mu\mu}}\right| \prec \Phi_o\,, \qquad \Lambda_o \prec \Phi_o\,.$$

The proof is the same as that of Theorem 4.7 in [9].

Proof of Lemma 4.8. We apply Lemma 4.9 to

$$-Z_\mu \;=\; \left(1-\mathbb{E}^{[\mu]}\right)\frac{1}{R_{\mu\mu}}\,, \tag{4.42}$$

where we used (3.11). From Proposition 4.2 we get $1 \prec \mathbf{1}(\Xi)$, i.e. Ξ holds with high probability. Therefore (4.13) yields $\Lambda_o \prec \Psi_\Theta$. Using this bound for Λ_o and Proposition 4.2 again to estimate $\Lambda \prec (N\eta)^{-1/4}$, we therefore get

$$\Lambda_o \;\prec\; \Phi_o\,, \qquad \Lambda \;\prec\; \Phi\,, \qquad \Phi_o \;:=\; \sqrt{\frac{\operatorname{Im} m_\phi + (N\eta)^{-\tau}}{N\eta}}\,, \qquad \Phi \;:=\; (N\eta)^{-1/4}\,.$$

Using (3.6), it is easy to check that these definitions of Φ_o and Φ satisfy the assumptions of Lemma 4.9. Hence the claim follows from Lemma 4.9 and (4.42). □

Now suppose that $\Theta \prec (N\eta)^{-\tau}$. From Lemma 4.8, the fact that $1-\mathbf{1}(\Xi) \prec 0$ from Proposition 4.2, Lemma 4.39, and (4.31), we find

$$|\mathcal{D}(m_R)| \;\prec\; \frac{\operatorname{Im} m_\phi + (N\eta)^{-\tau}}{N\eta}$$

uniformly in $z \in \mathbf{S}$. Using (4.19) and a simple union bound, we may invoke Lemma 4.5 to get

$$\Theta \;\prec\; \frac{\operatorname{Im} m_\phi}{N\eta}\frac{1}{\sqrt{\kappa+\eta}} + \sqrt{\frac{(N\eta)^{-\tau}}{N\eta}} \;\leqslant\; \frac{C}{N\eta} + (N\eta)^{-1/2-\tau/2} \;\leqslant\; C(N\eta)^{-1/2-\tau/2}\,,$$

where in the second step we used (3.6). Summarizing, we have proved the self-improving estimate

$$\Theta \;\prec\; (N\eta)^{-\tau} \qquad\Longrightarrow\qquad \Theta \;\prec\; (N\eta)^{-1/2-\tau/2}\,. \tag{4.43}$$

From Proposition 4.2 we know that $\Theta \prec (N\eta)^{-1/4}$. Thus, for any $\varepsilon > 0$, we iterate (4.43) an order C_ε times to get $\Theta \prec (N\eta)^{-1+\varepsilon}$. This concludes the proof of the first bound of (3.18). The second bound of (3.18) follows from the first one and the identity (3.13).

Next, (4.1) follows from (3.18) and $1-\mathbf{1}(\Xi) \prec 0$, combined with (4.30) and (4.13).

What remains is the proof of (4.2). To that end, in analogy to the partial expectation $\mathbb{E}^{[\mu]}$ defined above, we define $\mathbb{E}^{(i)}(\cdot) := \mathbb{E}(\cdot|X^{(i)})$. Introducing $1 = \mathbb{E}^{(i)} + \left(1-\mathbb{E}^{(i)}\right)$ into the right-hand side of (3.8) yields

$$\frac{1}{G_{ii}} \;=\; -z - z\sum_{\mu,\nu} X_{i\mu}R^{(i)}_{\mu\nu}X^*_{\nu i} \;=\; -z - \frac{\widetilde{z}}{N}\sum_\mu R^{(i)}_{\mu\mu} - (1-\mathbb{E}^i)z\sum_{\mu,\nu} X_{i\mu}R^{(i)}_{\mu\nu}X^*_{\nu i}\,.$$

Using Lemma 3.10, we rewrite the sum in the second term according to $N^{-1}\sum_\mu R^{(i)}_{\mu\mu} = m_R + O\big((N\eta)^{-1}\big)$. Moreover, the third term is estimated exactly as Z_μ in Lemma 4.4, using Lemma 3.1. Putting everything together yields

$$G_{ii} \;=\; \frac{1}{-z - \widetilde{z} m_R + O_\prec(\Psi_\Theta)} \;=\; \frac{1}{1/m_{\phi^{-1}} + O_\prec(\Theta + \Psi_\Theta)}\,,$$

as follows after some elementary algebra using (4.4). Moreover, using (3.5) it is not hard to see that $m_{\phi^{-1}} \asymp \phi^{-1/2}$ on the domain $\mathbf{S}$. This yields $G_{ii} = m_{\phi^{-1}} + O(\phi^{-1}\Psi)$, where we used $\Theta \prec (N\eta)^{-1}$ to estimate $\Theta + \Psi_\Theta \prec \Psi$. This concludes the proof of (4.2) for $i = j$.

In particular, $|G_{ii}| \prec \phi^{-1/2}$. The same argument applied to the matrix $X^{(j)}$ instead of X yields $|G^{(j)}_{ii}| \prec \phi^{-1/2}$. Thus we get from (3.9) that for $i \neq j$ we have

$$|G_{ij}| \;\prec\; \phi^{-1}\bigg|\sum_{\mu,\nu} X_{i\mu} R^{(ij)}_{\mu\nu} X^*_{\nu j}\bigg| \;\prec\; \phi^{-1}\Psi\,,$$

where the last step follows using (3.3), exactly as in the proof of Lemma 4.4, and (4.1). This concludes the proof of (4.2), and hence of Theorem 4.1.

4.3 Proof of Theorem 2.10

The proof of Theorem 2.10 is similar to that of Theorem 2.2 in [16] and Theorem 3.3 in [21]. We therefore only sketch the argument. First we observe that, since the nontrivial eigenvalues $\lambda_1, \dots, \lambda_K$ of X^*X and XX^* coincide and

$$N\int_\gamma^\infty \varrho_\phi(\mathrm{d}x) \;=\; M\int_\gamma^\infty \varrho_{\phi^{-1}}(\mathrm{d}x)$$

for all $\gamma > 0$, it suffices to prove Theorem 2.10 for $\phi \geqslant 1$, i.e. $K = N$.

Define the normalized counting functions

$$n_\phi(E_1, E_2) \;:=\; \int_{E_1}^{E_2} \varrho_\phi(\mathrm{d}x)\,, \qquad n(E_1, E_2) \;:=\; \frac{1}{N}\big|\{\alpha : E_1 \leqslant \lambda_\alpha \leqslant E_2\}\big|\,.$$

The proof relies on the following key estimates.

Lemma 4.11. *We have*

$$\big|n(E_1, E_2) - n_\phi(E_1, E_2)\big| \;\prec\; \frac{1}{N} \tag{4.44}$$

uniformly for any E_1 and E_2 satisfying $E_1 + \mathrm{i} \in \mathbf{S}$ and $E_2 + \mathrm{i} \in \mathbf{S}$. Moreover, we have

$$|\lambda_1 - \gamma_+| \;\prec\; N^{-2/3}\,. \tag{4.45}$$

Finally, if $\phi \geqslant 1 + c$ for some constant $c > 0$, then

$$|\lambda_N - \gamma_-| \;\prec\; N^{-2/3}\,. \tag{4.46}$$

Starting from Lemma 4.11, the proof of Theorem 2.10 is elementary. (The details are given e.g. on the last page of Section 5 in [16].)

Proof of Lemma 4.11. The estimate (4.44) is a standard consequence of (3.18), using Helffer-Sjöstrand functional calculus; see e.g. [16, Section 5].

What remains is the proof of (4.45) and (4.46). Here the argument from [21, Section 8] applies with trivial modifications. The key inputs in our case are (4.44), Lemma 4.5, (4.40), and Lemma 4.9 combined with Remark 4.10. We omit further details. □

5 The isotropic law: proof of Theorem 3.11

In this section we complete the proof of Theorem 3.11. Since (3.18) was proved in Section 4, we only need to prove (3.16) and (3.17). For definiteness, we give the details of the proof of (3.16); the proof of (3.17) is very similar, and the required modifications are outlined at the end of Section 5.15 below.

5.1 Rescaling

It is convenient to introduce the rescaled quantities

$$\widetilde{G}(z) \;:=\; \phi^{1/2}\, G(z)\,, \qquad \widetilde{z} \;:=\; \phi^{-1/2} z\,.$$

The reason for this scaling is that for $z \in \mathbf{S}$ the diagonal entries of $\widetilde{G}$ and $\widetilde{z}$ are of order one (See (4.5) as well as (5.2) and (5.3) below). Note that all formulas from Lemma 3.6 hold after the replacement $(z, G) \mapsto (\widetilde{z}, \widetilde{G})$.

We also introduce the rescaled quantity

$$\widetilde{m}_\phi \;:=\; \phi^{1/2} m_{\phi^{-1}} \;=\; \phi^{-1/2}\biggl(m_\phi + \frac{1-\phi}{z}\biggr)\,. \tag{5.1}$$

The motivation behind this definition is that

$$|\widetilde{m}_\phi| \;\asymp\; 1 \tag{5.2}$$

for $z \in \mathbf{S}$, as can be easily seen using (3.5). (Recall that $\phi \geqslant 1$ by assumption.) The following result is an immediate corollary of Theorem 4.1. Recall the definition of Ψ from (3.15).

Lemma 5.1. *In* $\mathbf{S}$ *we have*

$$\bigl|\widetilde{G}_{ij} - \delta_{ij}\widetilde{m}_\phi\bigr| \;\prec\; \phi^{-1/2}\Psi\,, \tag{5.3}$$

$$|R_{\mu\nu} - \delta_{\mu\nu} m_\phi| \;\prec\; \Psi\,. \tag{5.4}$$

Lemma 5.2. *Fix* $\ell \in \mathbb{N}$. *Then we have, uniformly in* $\mathbf{S}$ *and for* $|T| \leqslant \ell$ *and* $i, j \notin T$,

$$\bigl|\widetilde{G}^{(T)}_{ij} - \delta_{ij}\widetilde{m}_\phi\bigr| \;\prec\; \phi^{-1/2}\Psi \tag{5.5}$$

as well as

$$\bigl|\widetilde{G}^{(T)}_{ii}\bigr| \;\prec\; 1\,, \qquad \biggl|\frac{1}{\widetilde{G}^{(T)}_{ii}}\biggr| \;\prec\; 1\,.$$

Proof. From (5.3) and (5.2) we easily find

$$\bigl|\widetilde{G}_{ij} - \delta_{ij}\widetilde{m}_\phi\bigr| \;\prec\; \phi^{-1/2}\Psi\,, \qquad \bigl|\widetilde{G}_{ii}\bigr| \;\prec\; 1\,, \qquad \bigl|1/\widetilde{G}_{ii}\bigr| \;\prec\; 1\,.$$

The statement for general T satisfying $|T| \leqslant \ell$ then follows easily by induction on the size of T, using the identity (3.7) and the fact that $\phi^{-1/2}\Psi \leqslant 1$. □

5.2 Reduction to off-diagonal entries

By linearity and polarization, in order to prove (3.16) it suffices to prove that

$$\bigl|\langle \mathbf{v}, G\mathbf{v}\rangle - \phi^{-1/2}\widetilde{m}_\phi\bigr| \;\prec\; \phi^{-1}\Psi$$

for deterministic unit vectors $\mathbf{v}$. All of our estimates will be trivially uniform in the unit vector $\mathbf{v}$ and $z \in \mathbf{S}$, and we shall not mention this uniformity any more. Thus, for the following we fix a deterministic unit vector $\mathbf{v} \in \mathbb{C}^M$.

We write

$$\langle \mathbf{v}, G\mathbf{v}\rangle - \phi^{-1/2}\widetilde{m}_\phi \;=\; \sum_a |v_a|^2 (G_{aa} - \phi^{-1/2}\widetilde{m}_\phi) + \mathcal{Z}\,,$$

where we defined

$$\mathcal{Z} \;:=\; \sum_{a\neq b} \overline{v}_a G_{ab} v_b \;=\; \phi^{-1/2} \sum_{a\neq b} \overline{v}_a \widetilde{G}_{ab} v_b\,. \tag{5.6}$$

By (4.2) we have

$$\left| \sum_a |v_a|^2 (G_{aa} - \phi^{-1/2}\widetilde{m}_\phi) \right| \;\prec\; \phi^{-1}\Psi\,.$$

Hence it suffices to prove that

$$|\mathcal{Z}| \;\prec\; \phi^{-1}\Psi\,. \tag{5.7}$$

The rest of this section is devoted to the proof of (5.7).

5.3 Sketch of the proof

The basic reason why (5.7) holds is that G_{ab} can be expanded, to leading order, as a sum of independent random variables using the identity (3.9). To simplify the presentation in this sketch, we set $M = N$, so that $\phi = 1$ and the rescalings from Section 5.1 indicated by a tilde are trivial. Hence we drop all tildes. From (3.9) we get

$$\sum_{a\neq b} \overline{v}_a G_{ab} v_b \;=\; z \sum_{a\neq b} G_{aa} G^{(a)}_{bb} \sum_{\mu,\nu} \overline{v}_a X_{a\mu} R^{(ab)}_{\mu\nu} X^*_{\nu b} v_b\,. \tag{5.8}$$

If we could replace the diagonal entries by the deterministic value m_ϕ, it would suffice to estimate the sum $\sum_{a\neq b}\sum_{\mu,\nu} \overline{v}_a X_{a\mu} R^{(ab)}_{\mu\nu} X^*_{\nu b} v_b$. By the independence of the entries of X we have, using (3.3),

$$\begin{aligned}\left| \sum_{a\neq b} \sum_{\mu\nu} \overline{v}_a X_{a\mu} R^{(ab)}_{\mu,\nu} X^*_{\nu b} v_b \right| \;&\prec\; \left(\frac{1}{N^2} \sum_{a\neq b} \sum_{\mu,\nu} |v_a|^2 |v_b|^2 \left| R^{(ab)}_{\mu\nu} \right|^2 \right)^{1/2} \\ &=\; \left(\frac{1}{N^2\eta} \sum_{a\neq b} |v_a|^2 |v_b|^2 \operatorname{Im}\operatorname{Tr} R^{(ab)} \right)^{1/2} \;\prec\; \Psi\,,\end{aligned}$$

where we used the analogue of (3.14) for R, (4.1), (3.10), and the normalization of $\mathbf{v}$. Hence, if we could ignore the error arising from the approximation $G_{aa} \approx m_\phi$, the proof of Theorem 3.11 would be very simple.

The error made in the approximation $G_{aa} \approx m_\phi$ is of order Ψ by (5.3), so that the corresponding error term on the right-hand side of (5.8) may be bounded using (3.3) by

$$O_\prec(\Psi) \sum_{a\neq b} |v_a||v_b| \left| \sum_{\mu\nu} X_{a\mu} R^{(ab)}_{\mu\nu} X^*_{\nu b} \right| \;\prec\; \Psi^2 \sum_{a\neq b} |v_a||v_b| \;\leqslant\; \Psi^2 \|\mathbf{v}\|_1^2\,.$$

However, the vector $\mathbf{v}$ is normalized not in ℓ^1 but in ℓ^2. In general, all that can be said about its ℓ^1-norm is $\|\mathbf{v}\|_1 \leqslant M^{1/2}\|\mathbf{v}\|_2 = M^{1/2}$. This estimate is sharp if $\mathbf{v}$ is delocalized, i.e. if the entries of $\mathbf{v}$ have size of order $M^{-1/2}$. The ℓ^1- and ℓ^2-norms of $\mathbf{v}$ are of the same order precisely when only a finite number of entries of $\mathbf{v}$ are nonzero, in which case Theorem 3.11 is anyway a trivial consequence of Theorem 4.1.

We conclude that the simple replacement of G_{aa} with its deterministic approximation in (5.8) is not affordable. Not only the leading term but also every error term has to be expanded in the entries of X. This expansion is most effectively controlled if performed within a high-moment estimate. Thus, for large and even p we shall estimate

$$\mathbb{E} \left| \sum_{a\neq b} \overline{v}_a G_{ab} v_b \right|^p \;=\; \mathbb{E} \sum_{a_1\neq b_1} \cdots \sum_{a_p\neq b_p} \prod_{i=1}^p \overline{v}_{a_i} G_{a_i b_i} v_{b_i}\,. \tag{5.9}$$

(To simplify notation we drop the unimportant complex conjugations on $p/2$ factors.) We shall show that the expectation forces many indices of the leading-order terms to coincide, at least in pairs, so that eventually every v_a appears at least to the second power, which consistently leads to estimates in terms of the ℓ^2-norm of $\mathbf{v}$. Any index that remains single gives rise to a small factor $M^{-1/2}$ which counteracts the large factor $\|\mathbf{v}\|_1 \leqslant M^{1/2}$. The trivial bound (arising from estimating each entry $|v_a|$ by 1 and the summation over a and b by M^2) is affordable only at a very high order, when the number of factors $\Psi \leqslant N^{-\omega/2}$ that have been generated is sufficient to compensate the loss from the trivial bound. This idea will be used to stop the expansion after a sufficiently large, but finite, number of steps.

Before explaining the general strategy, we sketch a second moment calculation. First, we write

$$\mathbb{E}\left|\sum_{a\neq b}\overline{v}_a G_{ab} v_b\right|^2 \;=\; \mathbb{E}\sum_{a\neq b}\sum_{c\neq d}\overline{v}_a G_{ab} v_b\;\overline{v}_c G^*_{cd} v_d\,. \tag{5.10}$$

Using (3.7), we *maximally expand* all resolvent entries in the indices a, b, c, d. This means that we use (3.7) repeatedly until each term in the expansion is independent of all Latin indices that do not explicitly appear among its lower indices; here an entry is independent of an index if that index is an upper index of the entry. This generates a series of *maximally expanded terms*, whereby a resolvent entry is by definition maximally expanded if we cannot add to it upper indices from the set a, b, c, d by using the identity (3.7). In other words, $G^{(T)}_{ij}$ is maximally expanded if and only if $T = \{a, b, c, d\} \setminus \{i, j\}$.

To illustrate this procedure, we assume temporarily that a, b, c, d are all distinct, and write, using (3.7),

$$G_{ab} \;=\; G^{(c)}_{ab} + \frac{G_{ac}G_{cb}}{G_{cc}} \;=\; G^{(cd)}_{ab} + \frac{G_{ac}G_{cb}}{G_{cc}} + \frac{G^{(c)}_{ad}G^{(c)}_{db}}{G^{(c)}_{dd}}\,. \tag{5.11}$$

Here the first term is maximally expanded, but the second and third are not; we therefore continue to expand them in a similar fashion by applying (3.7) to each resolvent entry. In general, this procedure does not terminate, but it does generate finitely many maximally expanded terms with no more than a fixed number, say ℓ, of off-diagonal resolvent entries, in addition to finitely many terms that are not maximally expanded but contain more than ℓ off-diagonal entries. By choosing ℓ large enough, these latter terms may be estimated trivially. We therefore focus on the maximally expanded terms, and we write

$$G_{ab} \;=\; G^{(cd)}_{ab} + \frac{G^{(bd)}_{ac}G^{(ad)}_{cb}}{G^{(abd)}_{cc}} + \frac{G^{(bc)}_{ad}G^{(ac)}_{db}}{G^{(abc)}_{dd}} + \dots\,.$$

We get a similar expression for G^*_{cd}. We plug both of these expansions into (5.10) and multiply out the product. The leading term is

$$\mathbb{E}\sum_{a\neq b}\sum_{c\neq d}\overline{v}_a G^{(cd)}_{ab} v_b\,\overline{v}_c G^{*(ab)}_{cd} v_d\,.$$

We now expand both resolvent entries using (3.9), which gives

$$\begin{aligned}&\mathbb{E}\sum_{a\neq b}\sum_{c\neq d}\overline{v}_a G^{(cd)}_{ab} v_b\,\overline{v}_c G^{*(ab)}_{cd} v_d\\ &\;=\;\sum_{a\neq b}\sum_{c\neq d}\sum_{\mu,\nu,\alpha,\beta}\overline{v}_a v_b \overline{v}_c v_d\,\mathbb{E}\Big(G^{(cd)}_{aa}G^{(acd)}_{bb}X_{a\mu}R^{(abcd)}_{\mu\nu}X^*_{\nu b}G^{*(ab)}_{cc}G^{*(abc)}_{dd}X_{c\alpha}R^{*(abcd)}_{\alpha\beta}X^*_{\beta d}\Big)\end{aligned}$$

The goal is to use the expectation to get a pairing (or a more general partition) of the entries of X. In order to do that, we shall require all terms that are not entries of

X to be independent of the randomness in the rows a, b, c, d of X. While the entries of R satisfy this condition, the entries of G do not. We shall hence have to perform a further expansion on them using the identities (3.7) and (3.9). In fact, these two types of expansions will have to be performed in tandem, using a two-step recursive procedure. The main reason behind this is that even if all entries of G are maximally expanded, each application of (3.9) produces a diagonal entry that is not maximally expanded; for such terms the expansion using (3.7) has to be repeated. For the purposes of this sketch, however, we omit the details of the further expansion of the entries of G, and replace them with their deterministic leading order, m_ϕ (see (5.3)). This approximation gives

$$\mathbb{E}\sum_{a\neq b}\sum_{c\neq d}\overline{v}_a G^{(cd)}_{ab} v_b\,\overline{v}_c G^{*(ab)}_{cd} v_d \\ \approx\ |m_\phi|^4\sum_{a\neq b,c\neq d}\overline{v}_a v_b\overline{v}_c v_d\sum_{\mu,\nu,\alpha,\beta}\mathbb{E}\Big(X_{a\mu}R^{(abcd)}_{\mu\nu}X^*_{\nu b}X_{c\alpha}R^{*(abcd)}_{\alpha\beta}X^*_{\beta d}\Big)\,. \quad (5.12)$$

Since all entries of R are independent of all entries of X, we can compute the expectation with respect to the rows a, b, c, d. Note that the only possible pairing is $a = d$, $\mu = \beta$, $b = c$, and $\nu = \alpha$. This results in the expression

$$\frac{|m_\phi|^4}{N^2}\sum_{a\neq b}|v_a|^2|v_b|^2\mathbb{E}\sum_{\mu\nu}\big|R^{(abcd)}_{\mu\nu}\big|^2\ \approx\ \frac{1}{N\eta}\sum_{a\neq b}|v_a|^2|v_b|^2\frac{1}{N}\mathbb{E}\operatorname{Tr}\operatorname{Im}R^{(abcd)}\ \approx\ \frac{\operatorname{Im}m_\phi}{N\eta}\ \sim\ \Psi^2\,,$$

where we used that $\mathbf{v}$ is ℓ^2-normalized.

This calculation, while giving the right order, was in fact an oversimplification, since the expansion (5.11) required that $b \neq c$ and $d \neq a$. The correct argument requires first a decomposition of the summation over a, b, c, d into a finite number of terms, indexed by the partitions of four elements, according to the coincidences among these four indices. Thus, we write

$$\sum_{a\neq b,c\neq d}\ =\ \sum_{a,b,c,d}^{*}+\sum_{a=c,b,d}^{*}+\sum_{a=d,b,c}^{*}+\sum_{b=c,a,d}^{*}+\sum_{b=d,a,c}^{*}+\sum_{a=c,b=d}^{*}+\sum_{a=d,b=c}^{*}\,, \quad (5.13)$$

where a star over the summation indicates that all summation indices that are not explicitly equal to each other have to be distinct. The above calculation leading to (5.12) is valid for the first summation of (5.13), whose contribution (up to leading order) is zero, since the only possible pairing contradicts the condition that the indices a, b, c, d be all distinct. It is not too hard to see that, among the sums in (5.13), only the last one gives a nonzero contribution (up to leading order), and it is, going back to (5.10), equal to

$$\mathbb{E}\sum_{a\neq b}|v_a|^2|v_b|^2|G_{ab}|^2\ \prec\ \Psi^2\,;$$

here we used the bound (5.3). Notice that taking the expectation forced us to chose the pairing $a = d$, $b = c$ to get a non-zero term. This example provides a glimpse into the mechanism that guarantees that the ℓ^2-norm of $\mathbf{v}$ appears.

Next, we consider a subleading term from the first summation in (5.13), which has three off-diagonal entries:

$$\sum_{a,b,c,d}^{*}\overline{v}_a v_b\overline{v}_c v_d\frac{G^{(bd)}_{ac}G^{(ad)}_{cb}}{G^{(abd)}_{cc}}G^{*(ab)}_{cd}\,.$$

We proceed as before, expanding all off-diagonal entries of G using (3.9). Up to leading order, we get

$$\sum_{a,b,c,d}^{*} \overline{v}_a v_b \overline{v}_c v_d \sum_{\mu,\nu,\alpha,\beta,\gamma,\delta} \mathbb{E}\Big[\big(X_{a\mu} R_{\mu\nu}^{(abcd)} X_{\nu c}^*\big)\big(X_{c\alpha} R_{\alpha\beta}^{(abcd)} X_{\beta b}^*\big)\big(X_{c\gamma} R_{\gamma\delta}^{*(abcd)} X_{\delta d}^*\big)\Big]$$

The expectation again renders this term zero if a, b, c, d are distinct.

Based on these preliminary heuristics, we outline the main steps in estimating a high moment of $\mathcal{Z}$.

Step 1. Partition the indices in (5.9) according to their coincidence structure: indices in the same block of the partition are required to coincide and indices in different blocks are required to be distinct. This leads to a reduced family, T, of distinct indices.

Step 2. Make all entries of G maximally expanded by repeatedly applying the identity (3.7). Roughly, this entails adding upper indices from the family T to each entry of G using the identity (3.7). We stop the iteration if either (3.7) cannot be applied to any entry of G or we have generated a sufficiently large number of off-diagonal entries of G.

Step 3. Apply (3.9) to each maximally expanded off-diagonal entry of G. This yields factors of the form $\sum_{\mu,\nu} X_{a\mu} R_{\mu\nu}^{(T)} X_{\nu b}^*$ with $a, b \in T$ and $R^{(T)}$ is independent of all entries of X by construction. In addition, this application of (3.9) produces new diagonal entries of G that are not maximally expanded.

Step 4. Repeat Steps 2 and 3 recursively in tandem until we only have a sum of terms whose factors consist of maximally expanded diagonal entries of G, entries of $R^{(T)}$, and entries of X from the rows indexed by T.

Step 5. Apply (3.8) to each maximally expanded diagonal entry of G. We end up with factors consisting only of entries of $R^{(T)}$ and entries of X from the rows indexed by T.

Step 6. Using the fact that all entries of R are independent of all entries of X, take a partial expectation over the rows of X indexed by the set T; this only involves the entries of X. Only those terms give a nonzero contribution whose Greek indices have substantial coincidences.

Step 7. For entropy reasons, the leading-order term arises from the smallest number of constraints among the summation vertices that still results in a nonzero contribution. This corresponds to a pairing both among the Greek and the Latin indices. This naturally leads to estimates in terms to the ℓ^2-norm of $\mathbf{v}$.

Step 8. Observe that if a Latin index i remained single in the partitioning of Step 1 (so that the corresponding weight factor will involve the ℓ^1-norm $\sum_i |v_i|$) then, by a simple parity argument, the number of appearances of the index i will remain odd along the expansion of Steps 2–5. This forces us to take at least a third (but in fact at least a fifth) moment of some entry $X_{i\mu}$, which reduces the combinatorics of the summations compared with the fully paired situation from Step 7. This combinatorial gain offsets the factor $M^{1/2}$ lost in taking the ℓ^1-norm of $\mathbf{v}$.

Steps 1–6 require a careful expansion algorithm and a meticulous bookkeeping of the resulting terms. We shall develop a graphical language that encodes the resulting

monomials. Expansion steps will be recorded via operations on graphs such as merging certain vertices or replacing some vertex or edge by a small subgraph. Several ingredients of the graphical representation and the concept of graph operations are inspired by tools from [6] developed for random band matrices. Once the appropriate graphical language is in place and the expansion algorithm has been constructed, the observations in Steps 7 and 8 will yield the desired estimate by a power counting coupled with a parity argument.

5.4 The p-th moment of $\mathcal{Z}$ and introduction of graphs

We shall estimate $\mathcal{Z}$ with high probability by estimating its p-th moment for a large but fixed p. It is convenient to rename the summation variables in the definition of $\mathcal{Z}$ as $(a, b) = (b_1, b_2)$. Let p be an even integer and write

$$\mathbb{E}|\mathcal{Z}|^p \;=\; \phi^{-p/2}\,\mathbb{E}\sum_{b_{11}\neq b_{12}}\cdots\sum_{b_{p1}\neq b_{p2}}\left(\prod_{k=1}^{p/2}\overline{v}_{b_{k1}}\widetilde{G}_{b_{k1}b_{k2}}v_{b_{k2}}\right)\left(\prod_{k=p/2+1}^{p}\overline{v}_{b_{k1}}\widetilde{G}^*_{b_{k1}b_{k2}}v_{b_{k2}}\right), \tag{5.14}$$

where we recall the definition of $\mathcal{Z}$ from (5.6).

We begin by partitioning the summation according to the coincidences among the indices $\mathbf{b} = (b_{kr} : 1 \leqslant k \leqslant p\,,\, r = 1, 2)$. Denote by $\mathcal{P}(\mathbf{b})$ the partition of $\{1, \ldots, p\} \times \{1, 2\}$ defined by the equivalence relation $(k, r) \sim (l, s)$ if and only if $b_{kr} = b_{ls}$. We define $\mathfrak{P}_p$ as the set of partitions P of $\{1, \ldots, p\} \times \{1, 2\}$ such that, for all $k = 1, \ldots, p$, the elements $(k, 1)$ and $(k, 2)$ are not in the same block of P. Hence we may rewrite (5.14) as

$$\mathbb{E}|\mathcal{Z}|^p \;=\; \phi^{-p/2}\,\mathbb{E}\sum_{P\in\mathfrak{P}_p}\sum_{\mathbf{b}}\mathbf{1}(\mathcal{P}(\mathbf{b}) = P)\left(\prod_{k=1}^{p/2}\overline{v}_{b_{k1}}\widetilde{G}_{b_{k1}b_{k2}}v_{b_{k2}}\right)\left(\prod_{k=p/2+1}^{p}\overline{v}_{b_{k1}}\widetilde{G}^*_{b_{k1}b_{k2}}v_{b_{k2}}\right). \tag{5.15}$$

We shall perform the summation by first fixing the partition $P \in \mathfrak{P}_p$ and by deriving an upper bound that is uniform in P; at the very end we shall sum trivially over $P \in \mathfrak{P}_p$.

In order to handle expressions of the form (5.15), as well as more complicated ones required in later stages of the proof, we shall need to develop a graphical notation. The basic idea is to associate matrix indices with vertices and resolvent entries with edges. The following definition introduces graphs suitable for our purposes.

Definition 5.3 (Graphs)**.** *By a* graph *we mean a finite, directed, edge-coloured, multigraph*

$$\Gamma \;=\; \big(V(\Gamma), E(\Gamma), \xi(\Gamma)\big) \;\equiv\; (V, E, \xi)\,.$$

Here V is a finite set of vertices, E a finite set of directed edges, and ξ is a "colouring of E", i.e. a mapping from E to some finite set of colours. The graph Γ may have multiple edges and loops. More precisely, E is some finite set with maps $\alpha, \beta : E \to V$; here $\alpha(e)$ and $\beta(e)$ represent the source and target vertices of the edge $e \in E$. We denote by $\deg_\Gamma(i)$ the degree of the vertex $i \in V(\Gamma)$.

We may now express the right-hand side of (5.15) using graphs. Fix the partition $P \in \mathfrak{P}_p$. We associate a graph $\Delta \equiv \Delta(P)$ with P as follows. The vertex set $V(\Delta)$ is given by the blocks of P, i.e. $V(\Delta) = P$. The set of colours, i.e. the range of ξ, is $\{G, G^*\}$ (we emphasize that these two colours are simply formal symbols whose name is supposed to evoke their meaning). The set of edges $E(\Delta)$ is parametrized as follows by the resolvent entries on the right-hand side of (5.15). Each resolvent entry $G^{\#}_{b_{k1}b_{k2}}$ gives rise to an edge $e \in E(\Delta)$ with colour $\xi(e) = G$ if $\#$ is nothing and $\xi(e) = G^*$ if $\#$ is $*$. The source vertex $\alpha(e)$ of this edge is the unique block of P satisfying $(k, 1) \in \alpha(e)$, and its target vertex $\beta(e)$ the unique block of P satisfying $(k, 2) \in \beta(e)$. Figure 1 illustrates

the construction of $\Delta(P)$, where the two different types of line correspond to the two colours G, G^*. The graph Δ has no loops.

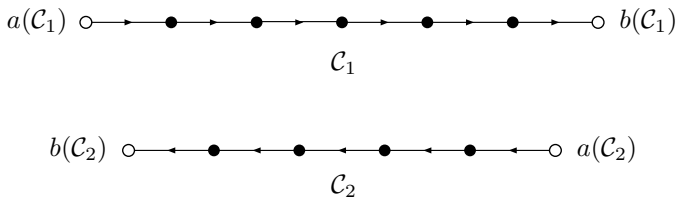

Figure 1: The construction of Δ. Here we took $p = 6$. On the left we give a graphical representation of the right-hand side of (5.15); the vertices (k, r) are labelled by $k = 1, \dots, 6$ and $r = 1, 2$; each edge is associated with an entry of $\widetilde{G}$ (drawn with a solid line) or $\widetilde{G}^*$ (drawn with a dashed line); the partition P is drawn using grey regions representing the blocks of P. On the right we draw $\Delta(P)$, with $|V(\Delta)| = |P| = 7$ vertices. The partition depicted here corresponds to the index coincidences $b_{11} = b_{21}$, $b_{12} = b_{22}$, $b_{31} = b_{42}$, $b_{32} = b_{41} = b_{52}$; apart from these constraints, all indices are distinct in the summation over $\mathbf{b}$ in (5.15). The index associated with block $i \in V(\Delta)$ is denoted by a_i, so that $a_i = b_{kr}$ for any (k, r) in the block i of the partition P.

Using the graph $\Delta \equiv \Delta(P)$ we may rewrite the right-hand side of (5.15). Each vertex $i \in V(\Delta)$, associated with a block of P, is assigned a summation index $a_i \in \{1, 2, \dots, M\}$, and we write $\mathbf{a} = (a_i)_{i \in V(\Delta)}$. The indicator function on the right-hand side of (5.15) translates to the condition that $a_i \neq a_j$ for $i \neq j$ (where $i, j \in V(\Delta)$). We use the notation $\sum^*$ to denote summation subject to this condition (distinct summation indices). Thus we may rewrite (5.15) as

$$\mathbb{E}|\mathcal{Z}|^p \;=\; \phi^{-p/2} \sum_{P \in \mathfrak{P}_p} Y(\Delta(P))\,, \tag{5.16}$$

where we defined

$$Y(\Delta) \;:=\; \sum_{\mathbf{a}}^{*} w_{\mathbf{a}}(\Delta)\, \mathbb{E}\mathcal{A}_{\mathbf{a}}(\Delta) \tag{5.17}$$

using the abbreviations

$$w_{\mathbf{a}}(\Delta) \;:=\; \prod_{e \in E(\Delta)} \overline{v}_{a_{\alpha(e)}} v_{a_{\beta(e)}}\,, \qquad \mathcal{A}_{\mathbf{a}}(\Delta) \;:=\; \prod_{e:\xi(e)=G} \widetilde{G}_{a_{\alpha(e)} a_{\beta(e)}} \prod_{e:\xi(e)=G^*} \widetilde{G}^*_{a_{\alpha(e)} a_{\beta(e)}}\,. \tag{5.18}$$

The function $w_{\mathbf{a}}(\Delta)$ has the interpretation of a deterministic (complex) weight for the summation over $\mathbf{a}$; it satisfies the basic estimate

$$|w_{\mathbf{a}}(\Delta)| \;\leqslant\; \prod_{i \in V(\Delta)} |v_{a_i}|^{\deg_\Delta(i)}\,. \tag{5.19}$$

We record the following basic properties of Δ.

- $|E(\Delta)| = p$.
- Δ has no loops, i.e. $\alpha(e) \neq \beta(e)$ for all $e \in E(\Delta)$.
- $1 \leqslant |V(\Delta)| \leqslant 2p$.

Our first goal is to use the expansion formulas (3.7)–(3.9) to express $\mathcal{A}_{\mathbf{a}}(\Delta)$ as a sum of monomials involving only entries of X and R, so that no entries of G remain. The entries of R and X will be independent by construction, which will make the evaluation

of the expectation possible. The result will be given in Proposition 5.10 below, which expresses $Y(\Delta)$ as a sum of terms associated with graphs, which are themselves conveniently indexed using a finite binary tree, denoted by $\mathcal{T}$. To bookkeep this expansion, we shall need a more general class of graphs than Δ.

5.5 Generalized colours and encoding

For the following we fix $p \in 2\mathbb{N}$ and a partition $P \in \mathfrak{P}_p$, and set $\Delta = \Delta(P)$. We shall develop an expansion scheme for monomials of type $\mathcal{A}_{\mathbf{a}}(\Delta)$. A fundamental notion in our expansion is that of *maximally expanded entries of* $\widetilde{G}$, given in Definition 5.4 below.

We shall need to enlarge the set of colours of edges, so as to be able to encode entries of not only $\widetilde{G}$ and $\widetilde{G}^*$, but also entries of R, R^*, X, and X^*; in addition, we shall need to encode diagonal entries of $\widetilde{G}$ and $\widetilde{G}^*$ that are in the denominator, as in the formulas (3.7), as well as to keep track of upper indices. We need all of these factors, since our expansion relies on a repeated application of the identities (3.7), (3.8), and (3.9).

In order to define the graphs precisely, we consider graphs Γ satisfying Definition 5.3 whose vertex set $V(\Gamma) = V_b(\Gamma) \cup V_w(\Gamma)$ is partitioned into *black vertices* $V_b(\Gamma)$ and *white vertices* $V_w(\Gamma)$. Informally, black vertices are incident to edges encoding entries of $\widetilde{G}$ and $\widetilde{G}^*$, and white vertices to edges encoding entries of $\widetilde{R}$ and $\widetilde{R}^*$. In other words, *black vertices* are associated with *Latin* summation indices in the *population space* $\{1, \dots, M\}$; *white vertices* are associated with *Greek* summation indices in the *sample space* $\{1, \dots, N\}$. See Remark 3.4. We shall only consider graphs Γ satisfying

$$V_b(\Gamma) \;=\; V(\Delta)\,, \tag{5.20}$$

an assumption we make throughout the following. This means that only new Greek summation indices but no new Latin indices are generated, corresponding to the repeated applications of (3.8) and (3.9). In particular, the vertex colouring for our graph is very simple: the vertices of Δ are black and all other vertices are white.

As our set of colours we choose

$$\left\{\xi = (\xi_1, \xi_2, \xi_3) : \xi_1 \in \{G, G^*, R, R^*, X, X^*\}\,,\, \xi_2 \in \{+, -\}\,,\, \xi_3 \subset V_b(\Gamma)\right\}. \tag{5.21}$$

Note that these colours are to be interpreted merely as list of formal symbols; the choice of their names is supposed to evoke their meaning. The component ξ_1 determines whether the edge encodes an entry of $\widetilde{G}$ (corresponding to $\xi_1 = G$), of $\widetilde{G}^*$ (corresponding to $\xi_1 = G^*$), of R (corresponding to $\xi_1 = R$), of R^* (corresponding to $\xi_1 = R^*$), of X (corresponding to $\xi_1 = X$), or of X^* (corresponding to $\xi_1 = X^*$). The component ξ_2 determines whether the entry is in the numerator (corresponding to $\xi_2 = +$) or in the denominator (corresponding to $\xi_2 = -$). Finally, the component ξ_3 is used to keep track of the upper indices of entries of $\widetilde{G}$ and $\widetilde{G}^*$, which we shall set to be $\mathbf{a}_{\xi_3} := \{a_i : i \in \xi_3\}$. The entries of R and R^* also have upper indices, but they always carry the maximal set $\mathbf{a}_b$ of upper indices, i.e. they always appear in the form $R^{(\mathbf{a}_b)}$ and $R^{*(\mathbf{a}_b)}$. Hence, upper indices need not be tracked for the entries of R and R^*, and for them we set $\xi_3(e) = \emptyset$. Let Γ be a graph with colour set (5.21).

We now list some properties of all graphs we shall consider. To that end, we call $e \in E(\Gamma)$ a *G-edge* if $\xi_1(e) \in \{G, G^*\}$, an *R-edge* if $\xi_1(e) \in \{R, R^*\}$, and an *X-edge* if $\xi_1(e) \in \{X, X^*\}$.

1. If e is a G-edge then $\alpha(e), \beta(e) \in V_b(\Gamma)$.
2. If e is an R-edge then $\alpha(e), \beta(e) \in V_w(\Gamma)$.
3. If $\xi_1(e) = X$ then $\alpha(e) \in V_b(\Gamma)$ and $\beta(e) \in V_w(\Gamma)$.

4. If $\xi_1(e) = X^*$ then $\alpha(e) \in V_w(\Gamma)$ and $\beta(e) \in V_b(\Gamma)$.
5. If $\xi_2(e) = -$ then $\xi_1(e) \in \{G, G^*\}$ and $\alpha(e) = \beta(e)$.
6. If $\xi_3(e) \neq \emptyset$ then $\xi_1(e) \in \{G, G^*\}$ and $\xi_3(e) \subset V_b(\Gamma) \setminus \{\alpha(e), \beta(e)\}$.

Properties (i)–(iv) are straightforward compatibility conditions which are obvious in light of the type of matrix entry that the edge e encodes. Property (v) states that only diagonal entries of $\widetilde{G}$ and $\widetilde{G}^*$ may be in the denominator. Property (vi) states that only entries of G or $\widetilde{G}$ may have a (nontrivial) upper index and the lower indices of an entry of $\widetilde{G}$ or $\widetilde{G}^*$ may not coincide with its upper indices (by definition of minors).

In order to give a precise definition of the monomial encoded by a coloured edge, and hence of a graph Γ, it is convenient to split the vertex indices as $\mathbf{a} = (a_i)_{i \in V(\Gamma)} = (\mathbf{a}_b, \mathbf{a}_w)$, where

$$\mathbf{a}_b := (a_i)_{i \in V_b(\Gamma)} \in \{1, \dots, M\}^{|V_b(\Gamma)|}, \qquad \mathbf{a}_w := (a_i)_{i \in V_w(\Gamma)} \in \{1, \dots, N\}^{|V_w(\Gamma)|}.$$

Under the former convention, indices assigned to black vertices (elements of $\{1, \dots, N\}$) were Latin letters, while indices assigned to white vertices (elements of $\{1, \dots, M\}$) were Greek letters. In the above expression, all indices assigned to vertices of $V(\Gamma)$ (the indices of $\mathbf{a}$) are also denoted by Latin letters i. This notation is independent of the previous convention: we simply do not have a third alphabet available. We always assume that the indices $\mathbf{a}_b$ are distinct; we impose no constraints on the indices $\mathbf{a}_w$. For the following definitions we fix a collection of vertex indices $\mathbf{a}$. At the end of the proof, we shall sum over $\mathbf{a}_b$ under the constraint that the indices of $\mathbf{a}_b$ be distinct.

For $e \in E(\Gamma)$ with $\xi_1(e) \in \{G, G^*\}$ we define the *resolvent entry encoded by* e *in* Γ as

$$\mathcal{A}_{\mathbf{a}}(e, \Gamma) := \begin{cases} \widetilde{G}^{(\mathbf{a}_{\xi_3(e)})}_{a_{\alpha(e)} a_{\beta(e)}} & \text{if } \xi_1(e) = G \text{ and } \xi_2(e) = + \\ \widetilde{G}^{*(\mathbf{a}_{\xi_3(e)})}_{a_{\alpha(e)} a_{\beta(e)}} & \text{if } \xi_1(e) = G^* \text{ and } \xi_2(e) = + \\ 1/\widetilde{G}^{(\mathbf{a}_{\xi_3(e)})}_{a_{\alpha(e)} a_{\beta(e)}} & \text{if } \xi_1(e) = G \text{ and } \xi_2(e) = - \\ 1/\widetilde{G}^{*(\mathbf{a}_{\xi_3(e)})}_{a_{\alpha(e)} a_{\beta(e)}} & \text{if } \xi_1(e) = G^* \text{ and } \xi_2(e) = -. \end{cases} \tag{5.22}$$

When drawing graphs, we represent a black vertex as a black dot and a white vertex as a white dot. An edge with $\xi_1 = G$ is represented as a solid directed line joining two black dots, and an edge with $\xi_1 = G^*$ as a dashed directed line joining two black dots. If $\xi_2 = -$ we indicate this by decorating the edge with a white diamond (not to be confused with a white dot). Notice that such edges are always loops, according to property (v). Sometimes we also indicate the component $\xi_3(e)$ in our graphs, simply by writing it next to the edge e. See Figure 2 for our graphical conventions when depicting edges with $\xi_1 \in \{G, G^*\}$.

For the other edges, $e \in E(\Gamma)$ with $\xi_1(e) \in \{R, R^*, X, X^*\}$, we set

$$\mathcal{A}_{\mathbf{a}}(e, \Gamma) := \begin{cases} R^{(\mathbf{a}_b)}_{a_{\alpha(e)} a_{\beta(e)}} & \text{if } \xi_1(e) = R \\ R^{*(\mathbf{a}_b)}_{a_{\alpha(e)} a_{\beta(e)}} & \text{if } \xi_1(e) = R^* \\ X_{a_{\alpha(e)} a_{\beta(e)}} & \text{if } \xi_1(e) = X \\ X^*_{a_{\alpha(e)} a_{\beta(e)}} & \text{if } \xi_1(e) = X^*. \end{cases}$$

When drawing graphs, we represent an edge with $\xi_1 = R$ as a solid directed line joining two white vertices, an edge with $\xi_1 = R^*$ as a dashed directed line joining two white vertices, an edge with $\xi_1 = X$ as a dotted directed line from a black to a white vertex, and an edge with $\xi_1 = X^*$ as a dotted directed line from a white to a black vertex. Note that we use the same line style to draw X- and X^*-edges, since the orientation of the edge together with the vertex colouring distinguishes them uniquely. See Figure 3 for an illustration of these conventions, and Figure 6 for an illustration of (5.28).

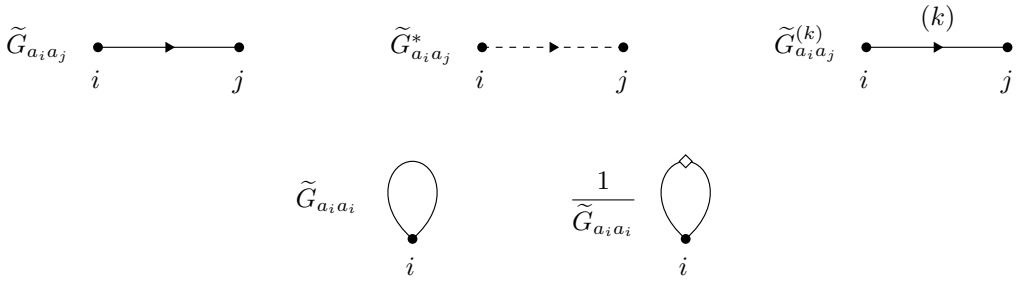

Figure 2: The graphical conventions for off-diagonal and diagonal edges. Here we draw the case $\xi_1 = G$ (solid lines encoding entries of $\widetilde{G}$); if $\xi_1 = G^*$ (encoding entries of $\widetilde{G}^*$) we use dashed lines, but the pictures are otherwise identical. The case $\xi_2 = +$ (encoding resolvent entries in the numerator) is drawn without any decorations; if $\xi_2 = -$ (encoding resolvent entries in the denominator) we indicate this with with a white diamond attached to the edge. Note that, since $\xi_2 = -$ only for diagonal entries (encoded by loops), the orientation of the edge is immaterial and the arrow therefore superfluous.

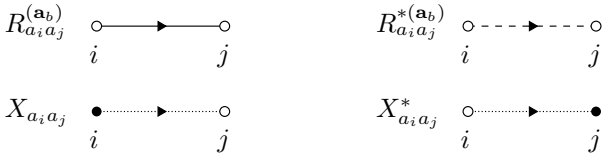

Figure 3: The graphical conventions for entries of $R^{(\mathbf{a}_b)}$ (corresponding to $\xi_1 = R$), $R^{*(\mathbf{a}_b)}$ (corresponding to $\xi_1 = R^*$), X (corresponding to $\xi_1 = X$), and X^* (corresponding to $\xi_1 = X^*$).

Having defined $\mathcal{A}_{\mathbf{a}}(e, \Gamma)$ for an arbitrary graph Γ with colour set (5.21) and $e \in E(\Gamma)$, we define the *monomial encoded by* Γ,

$$\mathcal{A}_{\mathbf{a}}(\Gamma) \;:=\; \prod_{e \in E(\Gamma)} \mathcal{A}_{\mathbf{a}}(e, \Gamma)\,. \tag{5.23}$$

Note that (5.23) extends (5.18). At this point we introduce a convention that will simplify notation throughout the proof. We allow the monomial $\mathcal{A}_{\mathbf{a}}(\Gamma)$ to be multiplied by a deterministic function of z that is bounded, i.e. in general we replace (5.23) with

$$\mathcal{A}_{\mathbf{a}}(\Gamma) \;:=\; u(\Gamma) \prod_{e \in E(\Gamma)} \mathcal{A}_{\mathbf{a}}(e, \Gamma)\,, \tag{5.24}$$

where $u(\Gamma)$ is some deterministic function of z satisfying $|u(\Gamma, z)| \leqslant C_\Gamma$ for $z \in \mathbf{S}$. This will allow us to forget signs and various factors of $\widetilde{z}$ and m_ϕ that are generated along the expansion. The functions $u(\Gamma)$ could be easily tracked throughout the proof, but all that we need to know about them is that they satisfy the conditions listed after (5.24). Not tracking the precise form of these prefactors is sufficient for our purposes, since after completing the graphical expansion we shall estimate each graph individually, without making use of further cancellations among different graphs.

5.6 R-groups

We define an *R-group* to be an induced subgraph of Γ consisting of three edges, e_1, e_2, e_3, such that e_1 and e_3 are X-edges and e_2 is an R-edge, and they form a chain

in the sense that $\beta(e_1) = \alpha(e_2)$, $\beta(e_2) = \alpha(e_3)$, and both of these vertices have degree two. We call e_2 the *centre* of the R-group and define $A(e_2) := \alpha(e_1)$ and $B(e_2) := \beta(e_3)$. If $A(e_2) = B(e_2)$ we call the R-group *diagonal*; otherwise we call it *off-diagonal*. See Figure 4 for an illustration. We require that our graphs Γ satisfy the following property.

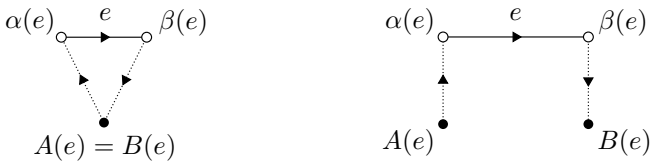

Figure 4: A diagonal R-group (left) and an off-diagonal R-group (right). We label the centre of the group by e.

(vii) Each X-edge and R-edge of Γ belongs to some R-group of Γ. In particular, all white vertices have degree two, and an R-group is uniquely determined by its centre.

The R-groups constitute graphical representations of the monomials on the right-hand sides of (3.8) and (3.9). It is important to stress that there is no restriction on possible coincidences among the white-vertex indices $(a_i)_{i\in V_w}$; this means that even if two Greek summation indices arising from two different applications (3.8) or (3.9) coincide, they will nevertheless be encoded by distinct white vertices. This allows us to keep the graphical structure involving R and X edges very simple.

Note that the initial graph $\Delta = \Delta(P)$ trivially satisfies the properties (i)–(vii).

5.7 Maximally expanded edges and sketch of the expansion

The following definition introduces a notion that underlies our entire expansion. Note that it only applies to G-edges.

Definition 5.4. *The G-edge $e \in E(\Gamma)$ is* maximally expanded *if $\xi_3(e) = V_b(\Gamma)\backslash\{\alpha(e), \beta(e)\}$. If e is maximally expanded then we also call the entry encoded by it, $\mathcal{A}_{\mathbf{a}}(e, \Gamma)$,* maximally expanded.

For instance, if e encodes an entry of the form $\widetilde{G}^{(T)}_{a_i a_j}$ then this entry is maximally expanded if and only if $T = \mathbf{a}_b \setminus \{a_i, a_j\}$. The idea behind this definition is that a maximally expanded entry has as many upper indices from the set $\mathbf{a}_b$ as possible.

We conclude this section with an outline of the expansion algorithm that will ultimately yield a family of graphs, whose contributions can be explicitly estimated and whose encoded monomials sum up to the monomial encoded by $\Delta = \Delta(P)$ from Section 5.4. The goal of the expansion is to get rid of all G-edges, by replacing them with R-groups. Of course, this replacement has to be done in such a manner that the original monomial $\mathcal{A}_{\mathbf{a}}(\Delta)$ can be expressed as a sum of the monomials encoded by the new graphs. Having done the expansion, we shall be able to exploit the fact that the R-entries and the X-entries are independent. This independence originates from the upper indices i and j in the entries of R in (3.8) and (3.9). It allows us to take the expectation in the X-variables. Combined with sufficient information about the graphs generated by the expansion, this yields a reduction in the summation that is sufficient to complete the proof.

The expansion relies of three main operations:

(a) make one of the G-entries maximally expanded by adding upper indices using the identity (3.7);

(b) expand all off-diagonal maximally expanded G-entries of in terms of X using the identity (3.9);

(c) expand all diagonal maximally expanded G-entries in terms of X using (3.8).

We shall implement each ingredient by a graph surgery procedure. Operation (a) is the subject of Section 5.8; it creates two new graphs, $\tau_0(\Gamma)$ and $\tau_1(\Gamma)$, from an initial graph Γ. Operation (b) is the subject of Section 5.9; it creates one new graph, $\rho(\Gamma)$, from an initial graph Γ. As it turns out, Operations (a) and (b) have to be performed in tandem using a coupled recursion, described by a tree $\mathcal{T}$, which is the subject of Section 5.10. Once this recursion has terminated, Operation (c) may be performed (see Section 5.11).

5.8 Operation (a): construction of the graphs $\tau_0(\Gamma)$ and $\tau_1(\Gamma)$

In order to avoid trivial ambiguities, we choose and fix an arbitrary ordering of the vertices $V(\Delta)$ and of the family of resolvent entries $(\widetilde{G}^{(T)}_{ab} : a, b \notin T)$. Hence we may speak of the first vertex of $V(\Delta)$ and the first factor of a monomial in the entries $\widetilde{G}^{(T)}_{ab}$.

We now describe Operation (a) of the expansion. It relies on the identities

$$\widetilde{G}^{(T)}_{ab} \;=\; \widetilde{G}^{(Tc)}_{ab} + \frac{\widetilde{G}^{(T)}_{ac}\widetilde{G}^{(T)}_{cb}}{\widetilde{G}^{(T)}_{cc}}\,, \qquad \frac{1}{\widetilde{G}^{(T)}_{aa}} \;=\; \frac{1}{\widetilde{G}^{(Tc)}_{aa}} - \frac{\widetilde{G}^{(T)}_{ac}\widetilde{G}^{(T)}_{ca}}{\widetilde{G}^{(T)}_{aa}\widetilde{G}^{(Tc)}_{aa}\widetilde{G}^{(T)}_{cc}}\,, \tag{5.25}$$

which follow immediately from (3.7); here $a, b, c \in \mathbf{a}_b \setminus T$ and $a, b \neq c$. The same identities hold for $\widetilde{G}^*$. The basic idea is to take some graph Γ with at least one G-entry that is not maximally expanded, to pick the first such G-entry, and to apply the first identity of (5.25) if this entry is in the numerator and the second identity if this entry is in the denominator. By Definition 5.4, if the G-entry is not maximally expanded, there is a $c \in \mathbf{a}_b$ such that (5.25) may be applied. The right-hand sides of (5.25) consist of two terms: the first one has one additional upper index, and the second one at least one additional off-diagonal G-entry. These two terms can be described by two new graphs, derived from Γ, denoted by $\tau_0(\Gamma)$ and $\tau_1(\Gamma)$. The graph $\tau_0(\Gamma)$ is almost identical to Γ, except that the edge corresponding to the selected entry $\widetilde{G}^{(T)}_{ab}$ receives an additional upper index c, so that the upper indices of the chosen entry are changed as $T \to (Tc)$. The graph $\tau_1(\Gamma)$ also differs from Γ only locally: the single edge of $\widetilde{G}^{(T)}_{ab}$ is replaced by two edges and loop with a diamond.

We now give the precise definition of Operation (a). Take a graph Γ that has a G-edge that is not maximally expanded. We shall define two new graphs, $\tau_0(\Gamma)$ and $\tau_1(\Gamma)$ as follows. Let e be the first[1] G-edge of Γ that is not maximally expanded, and let i be the first vertex of $V_b(\Gamma) \setminus \big(\xi_3(e) \cup \{\alpha(e), \beta(e)\}\big)$; note that by assumption on Γ and e this set of vertices is not empty. We now apply (5.25) to the entry $\mathcal{A}_{\mathbf{a}}(e, \Gamma)$. We set $a = a_{\alpha(e)}$, $b = a_{\beta(e)}$, $c = a_i$, and $T = \mathbf{a}_{\xi_3(e)}$ in (5.25), and express $\mathcal{A}_{\mathbf{a}}(e, \Gamma)$ as a sum of two terms given by the right-hand sides of (5.25); we use the first identity of (5.25) if $\xi_2(e) = +$ and the second if $\xi_2(e) = -$. This results in a splitting of the whole monomial into a sum of two monomials,

$$\mathcal{A}_{\mathbf{a}}(\Gamma) \;=\; \mathcal{A}_{0,\mathbf{a}}(\Gamma) + \mathcal{A}_{1,\mathbf{a}}(\Gamma)\,,$$

in self-explanatory notation. By definition, $\tau_0(\Gamma)$ is the graph that encodes $\mathcal{A}_{0,\mathbf{a}}(\Gamma)$ and $\tau_1(\Gamma)$ the graph that encodes $\mathcal{A}_{1,\mathbf{a}}(\Gamma)$. Hence, by definition, we have

$$\mathcal{A}_{\mathbf{a}}(\Gamma) \;=\; \mathcal{A}_{\mathbf{a}}(\tau_0(\Gamma)) + \mathcal{A}_{\mathbf{a}}(\tau_1(\Gamma))\,. \tag{5.26}$$

[1] Recall that we fixed an arbitrary ordering of the resolvent entries of G, which induces an ordering on the edges of Γ via the map $e \mapsto \mathcal{A}_{\mathbf{a}}(e, \Gamma)$.

Moreover, it follows immediately that the maps τ_0 and τ_1 do not change the vertices, so that we have

$$V_b(\tau_0(\Gamma)) \;=\; V_b(\tau_1(\Gamma)) \;=\; V_b(\Gamma)\,, \qquad V_w(\tau_0(\Gamma)) \;=\; V_w(\tau_1(\Gamma)) \;=\; V_w(\Gamma)\,. \tag{5.27}$$

The procedure $\Gamma \mapsto (\tau_0(\Gamma), \tau_1(\Gamma))$ may also be explicitly described on the level graphs alone, but we shall neither need nor do this. Instead, we give a graphical depiction of this process in Figure 5.

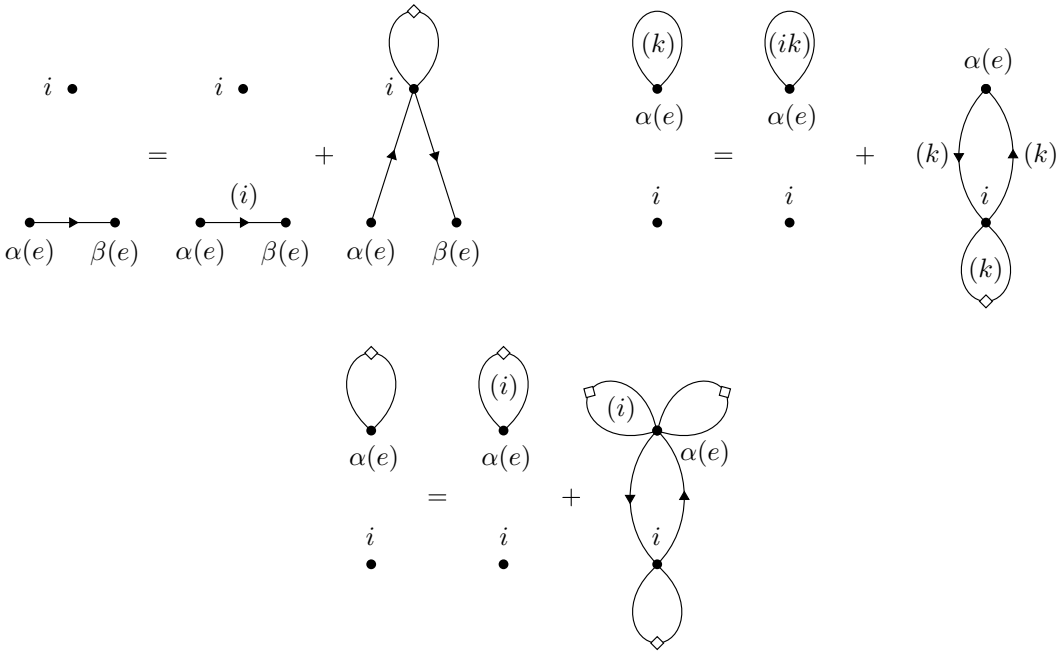

Figure 5: A graphical depiction of the splitting $\Gamma \mapsto (\tau_0(\Gamma), \tau_1(\Gamma))$ arising from (5.25). We only draw the edge e and the vertices $\alpha(e)$, $\beta(e)$, and i. All other edges of Γ are left unchanged by the operation, and are not drawn. The set ξ_3 is indicated in parentheses next to each edge, provided it is not empty. The first graph depicts the operation for the case $\alpha(e) \neq \beta(e)$ (encoding an off-diagonal entry), the second for the case $\alpha(e) = \beta(e)$ and $\xi_2(e) = +$ (encoding a diagonal entry in the numerator), and the third for the case $\alpha(e) = \beta(e)$ and $\xi_2(e) = -$ (encoding a diagonal entry in the denominator). The first graph on the right-hand side in each identity encodes $\tau_0(\Gamma)$ and the second $\tau_1(\Gamma)$. Recall that the graphs do not track irrelevant signs according to the convention made around (5.24).

The following result is trivial.

Lemma 5.5. *If Γ satisfies the properties (i)–(vii) from Sections 5.5 and 5.6 then so do $\tau_0(\Gamma)$ and $\tau_1(\Gamma)$.*

5.9 Operation (b): construction of the graph $\rho(\Gamma)$

In this section we give the second operation, (b), outlined in Section 5.7. The idea is that Operation (a) from Section 5.5 generates off-diagonal G-entries that are maximally expanded. They in turn will have to be expanded further using (3.9), so as to extract their explicit X-dependence. Roughly, the map ρ replaces each maximally expanded off-diagonal G-edge by an off-diagonal R-group.

It will be convenient to have a shorthand for a maximally expanded entry of $\widetilde{G}$. To that end, we define, for $a, b \in \mathbf{a}_b$, the maximally expanded entry

$$\widehat{G}_{ab} \;:=\; \widetilde{G}_{ab}^{(\mathbf{a}_b \setminus \{a,b\})} .$$

Using (3.9) we may write, for $a \neq b$,

$$\begin{aligned} \widehat{G}_{ab} \;&=\; \widetilde{z} \widetilde{G}_{aa}^{(\mathbf{a}_b \setminus \{a,b\})} \widehat{G}_{bb} \sum_{\mu,\nu} X_{a\mu} R_{\mu\nu}^{(\mathbf{a}_b)} X_{\nu b}^* \\ \widehat{G}_{ab}^* \;&=\; \widetilde{z}^* \widetilde{G}_{aa}^{*(\mathbf{a}_b \setminus \{a,b\})} \widehat{G}_{bb}^* \sum_{\mu,\nu} X_{a\mu} R_{\mu\nu}^{*(\mathbf{a}_b)} X_{\nu b}^* , \end{aligned} \tag{5.28}$$

where $\widetilde{z}^*$ denotes the complex conjugate of $\widetilde{z}$. Note that the first diagonal term on the right-hand side is not maximally expanded (while the second one is).

The identity (5.28) may also be formulated in terms of graphs. We denote by $\rho(\Gamma)$ the graph encoding the monomial obtained from $\mathcal{A}_{\mathbf{a}}(\Gamma)$ by applying the identity (5.28) to each maximally expanded off-diagonal G-entry of Γ. This replacement can be done in any order. By definition of $\rho(\Gamma)$, we have

$$\sum_{\mathbf{a}_w} \mathcal{A}_{\mathbf{a}_b, \mathbf{a}_w}(\Gamma) \;=\; \sum_{\mathbf{a}_w} \mathcal{A}_{\mathbf{a}_b, \mathbf{a}_w}(\rho(\Gamma)) . \tag{5.29}$$

Note that both sides depend on $\mathbf{a}_b$. Each application of (5.28) adds two white vertices, so that in general $V_w(\rho(\Gamma)) \supset V_w(\Gamma)$. In particular, in (5.29) we slightly abuse notation by using the symbol $\mathbf{a}_w$ for different families on the left- and right-hand sides. The point is that we always perform an unrestricted summation of the Greek indices associated with the white vertices of the graph. However, the black vertices are left unchanged, so that we have

$$V_b(\rho(\Gamma)) \;=\; V_b(\Gamma) . \tag{5.30}$$

Like τ_0 and τ_1, the map ρ may be explicitly defined on the level of graphs, which we shall however not do in order to avoid unnecessary and heavy notation. See Figure 6 for an illustration of ρ.

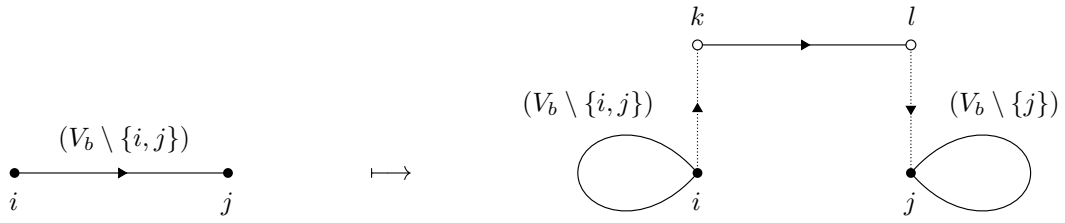

Figure 6: A graphical depiction of the map ρ resulting from applications of (5.28). For simplicity, we draw a graph with a single edge. The indices a, b, μ, ν of (5.28) are associated with the vertices i, j, k, l, so that we have $a = a_i$, $b = a_j$, $\mu = a_k$, and $\nu = a_l$. In the picture we abbreviated $V_b = V_b(\Gamma)$. Note that $V_b(\cdot)$ remains unchanged under ρ while $V_w(\cdot)$ is increased by the addition of two new white vertices, k, l. The prefactor $\widetilde{z}$ is omitted from the graphical representation.

The following result is an immediate corollary of the definition of ρ.

Lemma 5.6. *If Γ satisfies the properties (i)–(vii) from Sections 5.5 and 5.6 then so does $\rho(\Gamma)$.*

5.10 Constructing the tree $\mathcal{T}$: recursion using (a) and (b)

We now apply Operations (a) and (b) alternately and recursively to the graph $\Delta = \Delta(P)$. We start by applying Operation (a) to the graph Δ; the two new graphs thus produced may have newly created maximally expanded off-diagonal G-entries. We then apply ρ to these edges. Along the procedure we get new R-groups and additional diagonal entries, some which may not be maximally expanded. We then repeat the cycle: apply Operation (a) and then Operation (b). For some graphs the procedure stops because all G-edges have become maximally expanded. For some other graphs, the algorithm would continue indefinitely, since Operation (b) keeps on producing diagonal G-entries that are not maximally expanded. We shall however show that in such graphs the number of off-diagonal G-edges and R-groups increases as the algorithm is run. Since both of these objects are small, after the accumulation of a sufficiently large number of them we can stop the recursion and estimate such terms brutally. In summary, the end result will be a family of graphs encoding monomials in the entries of $R^{(\mathbf{a}_b)}, R^{*(\mathbf{a}_b)}, X, X^*$ as well as diagonal entries of $\widehat{G}, \widehat{G}^*$. In addition, by a brutal truncation in this procedure, the algorithm yields terms that do not satisfy this property, but contain a large enough number of off-diagonal G-edges and R-groups to be negligible.

The algorithm generates a family of graphs Θ_σ which are indexed by finite binary strings σ, or, equivalently, by vertices of a rooted binary tree $\mathcal{T} = (V(\mathcal{T}), E(\mathcal{T}))$. We start the algorithm with $\Theta_\emptyset := \Delta$, corresponding to the empty string or the root of the tree. The tree is constructed recursively according to

$$\Theta_0 := \rho(\tau_0(\Delta)), \qquad \Theta_1 := \rho(\tau_1(\Delta)), \qquad \Theta_{00} := \rho(\tau_0(\Theta_0)),$$
$$\Theta_{10} := \rho(\tau_1(\Theta_0)), \qquad \Theta_{01} := \rho(\tau_0(\Theta_1)),$$

and so on, until a stopping rule is satisfied (see Definition 5.7 below). See Figure 7 for an illustration of the resulting tree.

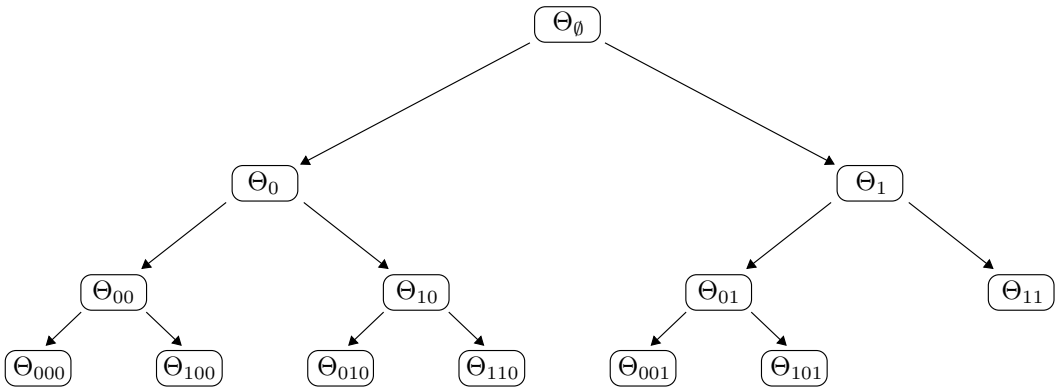

Figure 7: The tree $\mathcal{T}$ whose vertices are binary strings σ. The root is the empty string $\emptyset$. Each vertex of $\sigma \in V(\mathcal{T})$ encodes a graph Θ_σ. The graph associated with the two children of a vertex σ are obtained from Θ_σ using the maps τ_0, τ_1, and ρ. More precisely, an arrow towards the left corresponds to the map $\rho \circ \tau_0$ and an arrow towards the right to the map $\rho \circ \tau_1$. In this example, the graph Θ_{11} satisfies the stopping rule from Definition 5.7, and is therefore a leaf of $\mathcal{T}$.

We use the notation $i\sigma$, for $i = 0, 1$, to denote the binary string σ to which i has been appended on the left. The children in $\mathcal{T}$ of the vertex $\sigma \in V(\mathcal{T})$ are 0σ and 1σ. The precise construction of Θ_σ and the binary tree $\mathcal{T}$ is as follows. Let $\ell > 0$ be a cutoff

to be chosen later (see (5.39) below); it will be used as a threshold for the stopping rule which ensures that the tree $\mathcal{T}$ is finite. Let $d(\Gamma)$ denote the number of off-diagonal G-edges plus the number of off-diagonal R-groups of Γ, i.e.

$$d(\Gamma) \;:=\; \sum_{e\in E(\Gamma)} \Big(\mathbf{1}(\xi_1(e) \in \{G, G^*\}) \mathbf{1}\big(\alpha(e) \neq \beta(e)\big) + \mathbf{1}\big(\xi_1(e) \in \{R, R^*\}\big) \mathbf{1}\big(A(e) \neq B(e)\big) \Big) . \tag{5.31}$$

The construction of the tree $\mathcal{T}$ relies on the following stopping rule.

Definition 5.7 (Stopping rule)**.** *We say that a graph Γ* satisfies the stopping rule *if $d(\Gamma) \geqslant \ell$ or if all G-edges of Γ are maximally expanded.*

The tree $\mathcal{T}$, along with the graphs $(\Theta_\sigma)_{\sigma\in V(\mathcal{T})}$, is constructed recursively from the trivial tree, consistsing of the single vertex $\emptyset$ with $\Theta_\emptyset = \Delta$, as follows. Let σ be a leaf of the tree such that Θ_σ does not satisfy the stopping rule. We add the children of σ, i.e. 0σ and 1σ, to the tree, and set

$$\Theta_{0\sigma} \;:=\; \rho(\tau_0(\Theta_\sigma)) , \qquad \Theta_{1\sigma} \;:=\; \rho(\tau_1(\Theta_\sigma)) .$$

We continue this recursion on each leaf until all leaves satisfy the stopping rule from Definition 5.7. By Lemma 5.9 below, the resulting tree $\mathcal{T}$ is finite, i.e. the recursion terminates after a finite number of steps.

Lemma 5.8. *The graphs Θ_σ have the following two properties. First,*

$$V_b(\Theta_{0\sigma}) \;=\; V_b(\Theta_{1\sigma}) \;=\; V_b(\Theta_\sigma) . \tag{5.32}$$

In particular, the set of black vertices remains unchanged throughout the recursion: $V_b(\Theta_\sigma) = V_b(\Delta)$. Second,

$$\sum_{\mathbf{a}_w} \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Theta_{0\sigma}) + \sum_{\mathbf{a}_w} \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Theta_{1\sigma}) \;=\; \sum_{\mathbf{a}_w} \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Theta_\sigma) . \tag{5.33}$$

Note that both sides depend on $\mathbf{a}_b$, and we slightly abuse notation as explained after (5.29)*.*

Moreover, each Θ_σ for $\sigma \in V(\mathcal{T})$ satisfies the properties (i)–(vii) from Sections 5.5 and 5.6.

Proof. The identity (5.32) follows immediately from (5.27) and (5.30). Similarly, (5.33) follows from (5.26) and (5.29). The final statement follows immediately from Lemmas 5.5 and 5.6. □

The interpretation of (5.33) is that the value of any graph Θ_σ is equal to the sum of the values of its two children, $\Theta_{0\sigma}$ and $\Theta_{1\sigma}$.

The following estimate ensures that the tree $\mathcal{T}$ is finite, i.e. that the expansion procedure does not produce and infinite sequence of graphs whose value $d(\cdot)$ remains below ℓ indefinitely.

Lemma 5.9. *The tree $\mathcal{T}$ has depth at most $2p(p+6\ell)$ and consequently at most $2^{2p(p+6\ell)}$ vertices.*

Proof. Observe that τ_0 and ρ leave the function d defined in (5.31) invariant: $d(\tau_0(\Gamma)) = d(\Gamma)$, $d(\rho(\Gamma)) = d(\Gamma)$. Moreover, τ_1 increases d by at least one (for a diagonal entry the increase is two): $d(\tau_1(\Gamma)) \geqslant d(\Gamma) + 1$. We conclude that, by the stopping rule from Definition 5.7, any string σ of the tree contains at most ℓ ones, i.e. that τ_1 has been applied at most ℓ times.

Next, let $f = f(\Gamma)$ denote the number of G-edges minus the number of R-edges in the graph Γ. It follows immediately that f is left invariant by τ_0 and ρ, and is increased by at most 4 by τ_1: $f(\tau_0(\Gamma)) = f(\rho(\Gamma)) = f(\Gamma)$ and $f(\tau_1(\Gamma)) \leqslant f(\Gamma)+4$. Since in the initial graph there is no R-edge, so that $f(\Delta) = |E(\Delta)|$, we conclude that $f(\Theta_\sigma) \leqslant |E(\Delta)|+4\ell = p+4\ell$ for all $\sigma \in V(\mathcal{T})$. By Definition 5.7, the number of R-edges is bounded by ℓ. (Note that only off-diagonal R-groups have been created along the procedure, so that the number of R-edges is the same as the number of off-diagonal R groups. Diagonal R-groups will appear in later in Section 5.12). Hence we conclude that the number of G-edges of any Θ_σ is bounded by $p + 5\ell$.

In order to estimate the number of zeros in the string σ, we note that, since each G-entry can have at most $|V(\Delta)| \leqslant 2p$ upper indices, the total number of upper indices in all the G-entries of $\mathcal{A}_{\mathbf{a}}(\Theta_\sigma)$ is bounded by $2p(p + 5\ell)$. We conclude by noting that τ_1 and ρ do not decrease the total number of upper indices in the G-entries, while τ_0 increases this number by one. Hence the total number of zeros in any string σ is bounded by $2p(p + 5\ell)$. Thus, the total length of σ is bounded by $2p(p + 5\ell) + \ell \leqslant 2p(p + 6\ell)$. This concludes the proof. □

Next, we express $Y(\Delta)$ from (5.17) in terms of the graphs we just introduced. By Lemma 5.8 and the fact that $V_b(\Delta) = V(\Delta)$, we have for all $\sigma \in V(\mathcal{T})$ that

$$V_b(\Theta_\sigma) \;=\; V(\Delta). \tag{5.34}$$

Let $L(\mathcal{T}) \subset V(\mathcal{T})$ denote the leaves of $\mathcal{T}$. The identity (5.33) states that if $\sigma \in V(\mathcal{T})$ is not a leaf of $\mathcal{T}$, we may replace the value of Θ_σ by the sum of the values of its two children. Starting from the root $\emptyset$ and the graph $\Theta_\emptyset = \Delta$, we may propagate this identity recursively from the root down to the leaves. We conclude that

$$\mathcal{A}_{\mathbf{a}_b}(\Delta) \;=\; \sum_{\sigma \in L(\mathcal{T})} \sum_{\mathbf{a}_w} \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Theta_\sigma)\,. \tag{5.35}$$

Recalling the definition (5.17) of $Y(\Delta)$, we get the following result.

Proposition 5.10. *The quantity $Y(\Delta)$ defined in* (5.17) *may be written in terms of the tree $\mathcal{T}$ as*

$$Y(\Delta) \;=\; \sum_{\sigma \in L(\mathcal{T})} \sum_{\mathbf{a}_b}^{*} w_{\mathbf{a}_b}(\Delta) \sum_{\mathbf{a}_w} \mathbb{E}\,\mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Theta_\sigma)\,. \tag{5.36}$$

For the following we partition $L(\mathcal{T}) = L_0(\mathcal{T}) \cup L_1(\mathcal{T})$ into the *trivial leaves* $L_0(\mathcal{T})$ and the *nontrivial leaves* $L_1(\mathcal{T})$. By definition, the trivial leaves of $\mathcal{T}$ are those $\sigma \in V(\mathcal{T})$ satisfying $d(\Theta_\sigma) \geqslant \ell$. We shall estimate the contribution of the trivial leaves brutally in Section 5.11 below, using the fact that they contain a large enough number of small factors.

By Definition 5.7, if $\sigma \in L_1(\mathcal{T})$ is a nontrivial leaf then all G-edges of Θ_σ are diagonal and maximally expanded. The estimate of the nontrivial leaves will be performed in Sections 5.12–5.14.

5.11 The trivial leaves

In this section we estimate the contribution of Θ_σ for a trivial leaf $\sigma \in L_0(\mathcal{T})$. Thus, fix $\sigma \in L_0(\mathcal{T})$. From (5.28) and Lemma 5.2 we get for $a \neq b$

$$\sum_{\mu,\nu} X_{a\mu} R^{(\mathbf{a}_b)}_{\mu\nu} X^*_{\nu b} \;\prec\; \phi^{-1/2}\Psi\,, \qquad \sum_{\mu,\nu} X_{a\mu} R^{*(\mathbf{a}_b)}_{\mu\nu} X^*_{\nu b} \;\prec\; \phi^{-1/2}\Psi\,. \tag{5.37}$$

We therefore conclude that each off-diagonal R-group of Γ yields a contribution of size $O_\prec(\phi^{-1/2}\Psi)$ after summation over the indices associated with the vertices incident to

its centre. Moreover, by definition of $\mathcal{T}$, each R-group of Θ_σ is off-diagonal. In addition, each off-diagonal G-edge yields a contribution of size $\phi^{-1/2}\Psi$ by Lemma 5.2. Thus we get, summing out all indices associated with white vertices (i.e. inner vertices of R-groups),

$$\sum_{\mathbf{a}_w} \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Theta_\sigma) \;\prec\; (\phi^{-1/2}\Psi)^{d(\Theta_\sigma)} \;\leqslant\; (\phi^{-1/2}\Psi)^\ell \,.$$

Hence the contribution of Θ_σ to the right-hand side of (5.36) may be bounded by

$$\sum_{\mathbf{a}_b}^{*} w_{\mathbf{a}_b}(\Delta) \sum_{\mathbf{a}_w} \mathbb{E}\, \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Theta_\sigma) \;\prec\; M^{2p}(\phi^{-1/2}\Psi)^\ell \,,$$

where we estimated the summation over $\mathbf{a}_b$ by M^{2p} using the trivial bound $|w_{\mathbf{a}_b}(\Delta)| \leqslant 1$ (from (5.19) and $\|\mathbf{v}\|_2 = 1$). In the last step we used Lemma 3.2 (i) and (iii). The assumption $\mathbb{E}Z^2 \leqslant N^C$ of Lemma 3.2 (iii) for the random variable $Z = \left|\sum_{\mathbf{a}_w} \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Theta_\sigma)\right|$ follows from the following lemma combined with Hölder's inequality, and from the fact that the number of white vertices of Θ_σ is independent of N, so that the sum $\sum_{\mathbf{a}_w}$ contains $O(N^C)$ terms.

Lemma 5.11. *For any p there exists a constant C_p such that for any graph Γ and any $e \in E(\Gamma)$ we have*

$$\mathbb{E}\left|\mathcal{A}_{\mathbf{a}}(e,\Gamma)\right|^p \;\leqslant\; M^{C_p} \,.$$

Proof. The cases $\xi_1(e) \in \{G, G^*, R, R^*\}$ and $\xi_2(e) = +$ are dealt with the deterministic estimates

$$\left|\widetilde{G}^{(T)}_{ij}\right| \;\leqslant\; N\phi^{1/2} \;\leqslant\; M^2 \,, \qquad \left|R^{(T)}_{ij}\right| \;\leqslant\; N \leqslant M \,,$$

which follows from $|G^{(T)}_{ij}|, |R^{(T)}_{ij}| \leqslant \eta^{-1} \leqslant N$. The cases $\xi_1(e) \in \{X, X^*\}$ follow immediately from (2.3). Finally, the cases $\xi_1(e) \in \{G, G^*\}$ and $\xi_2(e) = -$ follow easily from (3.8). □

Using Lemma 5.9, we therefore conclude that the contribution of all trivial leaves to the right-hand side of (5.36) is bounded by

$$\sum_{\sigma \in L_0(\mathcal{T})} \sum_{\mathbf{a}_b}^{*} w_{\mathbf{a}_b}(\Delta) \sum_{\mathbf{a}_w} \mathbb{E}\, \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Theta_\sigma) \;\prec\; C_{p,\ell}\, M^{2p}(\phi^{-1/2}\Psi)^\ell \;\leqslant\; C_{p,\omega}(\phi^{-1}\Psi)^p \,, \tag{5.38}$$

where $C_{p,\ell} = 2^{2p(p+6\ell)}$ estimates the number of vertices in $\mathcal{T}$ (see Lemma 5.9). The last step holds provided we choose

$$\ell \;:=\; \left(\frac{8}{\omega} + 2\right) p \,. \tag{5.39}$$

Here we used the bound $\Psi \leqslant CN^{-\omega/2}$, which follows from the definitions (3.15), (2.11), and (3.5).

5.12 The nontrivial leaves I: Operation (c)

From now on we focus on the nontrivial leaves, $\sigma \in L_1(\mathcal{T})$. Our goal is to prove the following estimate, which is analogous to (5.38). Its proof will be the content of this and the two following subsections, and will be completed at the end of Section 5.14.

Proposition 5.12. *We have the bound*

$$\sum_{\sigma \in L_1(\mathcal{T})} \sum_{\mathbf{a}_b}^{*} w_{\mathbf{a}_b}(\Delta) \sum_{\mathbf{a}_w} \mathbb{E}\, \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Theta_\sigma) \;\prec\; C_{p,\omega}(\phi^{-1}\Psi)^p \,.$$

By definition of $L_1(\mathcal{T})$, all G-edges of Θ_σ are diagonal and maximally expanded for any $\sigma \in L_1(\mathcal{T})$. The first step behind the proof of Proposition 5.12 uses Operation (c) from Section 5.5, i.e. expanding all diagonal G-entries of $\mathcal{A}_{\mathbf{a}}(\Theta_\sigma)$ using (3.8). Roughly, this amounts to replacing diagonal G-edges by (a collection of) diagonal R-groups. More precisely, for entries in the denominator we use the identity

$$\frac{1}{\widehat{G}_{aa}} \;=\; -\widetilde{z} - \widetilde{z}\sum_{\mu,\nu} X_{a\mu} R^{(\mathbf{a}_b)}_{\mu\nu} X^*_{\nu a}\,. \tag{5.40}$$

In order to handle entries in the numerator, we rewrite this identity in the form

$$\frac{1}{\widehat{G}_{aa}} \;=\; \frac{1}{\widetilde{m}_\phi} - \left(\widetilde{z}\sum_{\mu,\nu} X_{i\mu} R^{(\mathbf{a}_b)}_{\mu\nu} X^*_{\nu a} - \widetilde{z}\phi^{-1/2} m_\phi\right), \tag{5.41}$$

where used the definition (5.1) of $\widetilde{m}_\phi$ and that m_ϕ satisfies the identity (2.7). From Lemma 5.2 and (5.2) we get $1/\widehat{G}_{aa} - 1/\widetilde{m}_\phi \prec \phi^{-1/2}\Psi$. Thus we may expand the inverse of (5.41) up to order ℓ:

$$\widehat{G}_{aa} \;=\; \sum_{k=0}^{\ell-1} \widetilde{m}_\phi^{k+1} \left(\widetilde{z}\sum_{\mu,\nu} X_{i\mu} R^{(\mathbf{a}_b)}_{\mu\nu} X^*_{\nu a} - \widetilde{z}\phi^{-1/2} m_\phi\right)^k + O_\prec\big((\phi^{-1/2}\Psi)^\ell\big)\,. \tag{5.42}$$

This is our main expansion for the diagonal G-entries in the numerator. Both formulas (5.40) and (5.42) have trivial analogues for the Hermitian conjugate $\widehat{G}^*_{aa}$.

Recall that all G-entries of $\mathcal{A}_{\mathbf{a}}(\Theta_\sigma)$ are diagonal and maximally expanded. We apply (5.40) or (5.42) to each G-entry of $\mathcal{A}_{\mathbf{a}}(\Theta_\sigma)$, and multiply everything out. The result may be written in terms graphs as

$$\sum_{\mathbf{a}_w} \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Theta_\sigma) \;=\; \sum_{\Gamma\in\mathfrak{G}(\Theta_\sigma)} \sum_{\mathbf{a}_w} \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Gamma) + O_\prec\big((\phi^{-1/2}\Psi)^\ell\big)\,, \tag{5.43}$$

where the error term $O_\prec\big((\phi^{-1/2}\Psi)^\ell\big)$ contains all terms containing at least one error term from the expansion (5.42). The sum on the right-hand side of (5.43) consists of monomials in the entries of $R^{(\mathbf{a}_b)}$, $R^{*(\mathbf{a}_b)}$, X, and X^* (note that entries of $\widehat{G}$ and $\widehat{G}^*$ no longer appear), and can hence be encoded using a family graphs which we call $\mathfrak{G}(\Theta_\sigma)$. By construction, the family $\mathfrak{G}(\Theta_\sigma)$ is finite. (In fact, it satisfies $|\mathfrak{G}(\Theta_\sigma)| \leqslant \ell^{6\ell}$, where we used that the number of G-entries of $\mathcal{A}_{\mathbf{a}}(\Theta_\sigma)$ to which (5.40) or (5.42) are applied is bounded by $p + 5\ell \leqslant 6\ell$; see the proof of Lemma 5.9.)

Exactly as in Section 5.11, we may brutally estimate the contribution of the rest term on the right-hand side of (5.43) by

$$\sum_{\mathbf{a}_b}^{*} w_{\mathbf{a}_b}(\Delta) O_\prec\big((\phi^{-1/2}\Psi)^\ell\big) \;\prec\; C_{p,\omega}(\phi^{-1}\Psi)^p$$

with ℓ defined in (5.39); we omit the details.

Hence, in order to complete the proof of Proposition 5.12, it suffices to prove that for all $\sigma \in L_1(\mathcal{T})$ and all $\Gamma \in \mathfrak{G}(\Theta_\sigma)$ we have

$$\sum_{\mathbf{a}_b}^{*} w_{\mathbf{a}_b}(\Delta) \sum_{\mathbf{a}_w} \mathbb{E}\mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Gamma) \;\prec\; C_{p,\omega}(\phi^{-1}\Psi)^p\,. \tag{5.44}$$

As before, the map $\Theta_\sigma \mapsto \mathfrak{G}(\Theta_\sigma)$ may be explicitly given on the level of graphs, but we refrain from doing so. Instead, we illustrate this process for some simple cases in Figure 8.

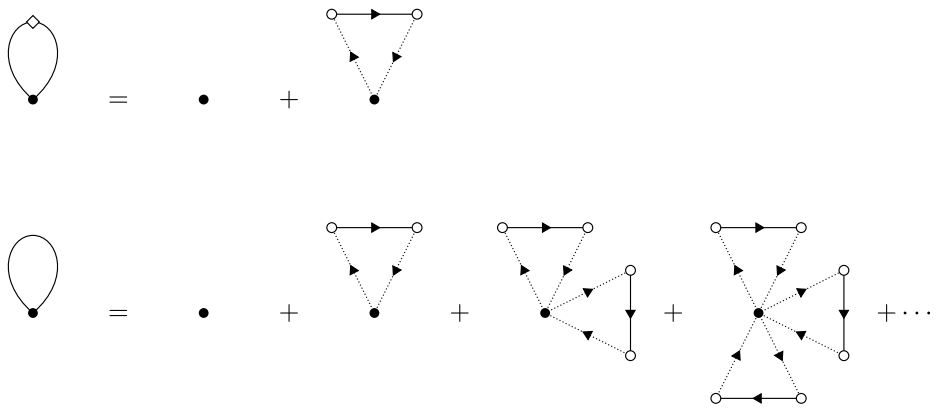

Figure 8: A graphical depiction of the identities (5.40) and (5.42) that generate $\mathfrak{G}(\Theta_\sigma)$ from Θ_σ. A G-edge encoding an entry in the denominator is replaced by either nothing (leaving just the vertex) or a diagonal R-group. A G-edge encoding an entry in the numerator is replaced by either nothing or up to $\ell - 1$ diagonal R-groups.

5.13 The nontrivial leaves II: taking the expectation

Let us now consider a nontrivial leaf $\sigma \in L_1(\mathcal{T})$. By definition of $L_1(\mathcal{T})$, all G-edges of Θ_σ are diagonal and maximally expanded. Therefore, any $\Gamma \in \mathfrak{G}(\Theta_\sigma)$ does not contain any G-edges. This was the goal of the expansion generated by Operations (a)–(c). Hence, each $\Gamma \in \mathfrak{G}(\Theta_\sigma)$ consists solely of R-groups.

Let $\sigma \in L_1(\mathcal{T})$ and $\Gamma \in \mathfrak{G}(\Theta_\sigma)$. Fix the summation indices $\mathbf{a}_b$, and recall that $a_i \neq a_j$ for $i, j \in V_b(\Gamma)$ and $i \neq j$. By definition of $R^{(\mathbf{a}_b)}$, the $|V_b(\Gamma)| + 1$ families $\big(R^{(\mathbf{a}_b)}_{\mu\nu}\big)^N_{\mu,\nu=1}$ and $(X_{a_i\mu})^N_{\mu=1}$, $i \in V_b(\Gamma)$, are independent. Therefore we may take the expectation of the R-entries and the X-entries separately. The expectation of the X-entries may be kept track of using partitions, very much like in Section 5.4, except in this case the partition is on the white vertices. In fact, the combinatorics here are much simpler, since two white vertices may only be in the same block of the partition if they are adjacent to a common black vertex. Indeed, the (Latin) indices associated with two different black vertices are different, so that the two entries of X encoded by two X-edges incident to two different black vertices are independent, since $X_{a\mu}$ and $X_{b\nu}$ are independent if $a \neq b$ for all μ and ν (even if $\mu = \nu$). The precise definition is the following.

We recall from Property (vii) in Section 5.6 that each white vertex $j \in V_w(\Gamma)$ is adjacent in Γ to a unique black vertex $\pi(j) \equiv \pi_\Gamma(j)$. For each $i \in V_b(\Gamma)$ we introduce a partition ζ_i of the subset of white vertices $\pi^{-1}(\{i\})$, and constrain the values of the indices $(a_j : \pi(j) = i)$ to be compatible with ζ_i. On the level of graphs, such a partition amounts to merging vertices in $\pi^{-1}(\{i\})$. Abbreviate $\zeta = (\zeta_i)_{i \in V_b(\Gamma)}$, and denote by Γ_ζ the graph obtained from Γ by merging, for each $i \in V_b(\Gamma)$, the vertices adjacent to i according to ζ_i. Note that, like Γ, each Γ_ζ satisfies the properties (i)–(vi) from Section 5.5, but, unlike Γ, in general Γ_ζ does not satisfy the property (vii) from Section 5.6. See Figure 9 for an illustration of the mapping $\Gamma \mapsto \Gamma_\zeta$.

Define the indicator function

$$\chi_{\mathbf{a}_w}(\Gamma) := \prod_{i \in V_b(\Gamma)} \mathbf{1}\big(a_j \neq a_{j'} \text{ for } j, j' \in \pi_\Gamma^{-1}(\{i\}) \text{ and } j \neq j'\big)\,,$$

which constrains the summation indices associated with different white vertices adjacent to the same black vertex to have different values. By definition of Γ_ζ, we therefore

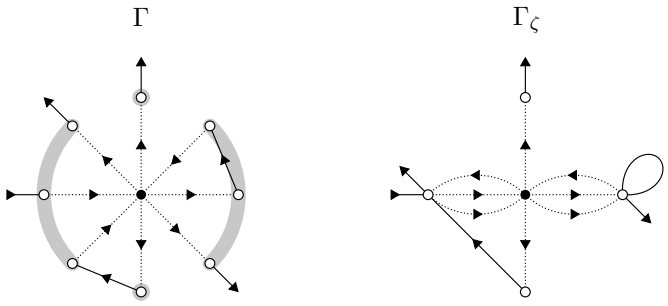

Figure 9: The process $\Gamma \mapsto \Gamma_\zeta$. Since this operation is local at each black vertex, we only draw the neighbourhood of (more precisely the unit ball around) a selected black vertex $i \in V_b(\Gamma)$. The depicted black vertex is part of two diagonal R-blocks and four off-diagonal R-blocks; the latter ones are not drawn completely. The blocks of the partition ζ_i are drawn in grey. On the right we draw the corresponding neighbourhood of Γ_ζ.

have

$$\sum_{\mathbf{a}_w} \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Gamma) \;=\; \sum_{\zeta} \sum_{\mathbf{a}_w} \chi_{\mathbf{a}_w}(\Gamma_\zeta) \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Gamma_\zeta)\,,$$

where the sum ranges over all families of partitions $\zeta = (\zeta_i)_{i \in V_b(\Gamma)}$. As before, the summation of the white indices $\mathbf{a}_w$ runs over different sets on the left- and the right-hand sides, owing to the merging white vertices. Taking the expectation yields

$$\sum_{\mathbf{a}_w} \mathbb{E}\mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(\Gamma) \;=\; \sum_{\zeta} \sum_{\mathbf{a}_w} \chi_{\mathbf{a}_w}(\Gamma_\zeta) \Bigg(\mathbb{E} \prod_{e \in E_R(\Gamma_\zeta)} \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(e,\Gamma_\zeta) \Bigg) W_{\mathbf{a}_b,\mathbf{a}_w}(\Gamma_\zeta)\,, \tag{5.45}$$

where we set

$$W_{\mathbf{a}_b,\mathbf{a}_w}(\Gamma_\zeta) \;:=\; \prod_{i \in V_b(\Gamma_\zeta)} \Bigg(\mathbb{E} \prod_{e \in E_i(\Gamma_\zeta)} \mathcal{A}_{\mathbf{a}_b,\mathbf{a}_w}(e,\Gamma_\zeta) \Bigg)$$

and abbreviated $E_R(\cdot)$ for the set of R-edges and $E_i(\cdot)$ for the set of X-edges incident to i. Here we used the independence described above.

Since $\mathbb{E}X_{a\mu} = 0$, we immediately get that $W_{\mathbf{a}_b,\mathbf{a}_w}(\Gamma_\zeta) = 0$ unless, for each $i \in V_b(\Gamma)$, each block of ζ_i has size at least two. By (2.3) we get in fact that

$$\big|W_{\mathbf{a}_b,\mathbf{a}_w}(\Gamma_\zeta)\big| \;\leqslant\; C_\Gamma (NM)^{-|V_w(\Gamma)|/4} \prod_{i \in V_b(\Gamma)} \mathbf{1}\big(\text{each block of } \zeta_i \text{ has size at least two}\big)\,. \tag{5.46}$$

The following result is the main power counting estimate for $W_{\mathbf{a}_b,\mathbf{a}_w}$. It shows that each black vertex of degree one in Δ (corresponding to Latin indices that remained unpaired in the partition (5.15)) results in an extra factor $M^{-1/2}$. This will balance the passage from ℓ^1- to ℓ^2-norm of $\mathbf{v}$, as explained in Section 5.3.

Note that by definition of Γ_ζ we have $V_b(\Gamma_\zeta) = V_b(\Gamma)$ and $\deg_\Gamma(i) = \deg_{\Gamma_\zeta}(i)$ for all $i \in V_b(\Gamma)$. For the following we therefore drop the argument of V_b. Define the subset

$$V_b^* \;:=\; \big\{ i \in V_b : \deg_\Delta(i) = 1 \big\}\,.$$

For $i \in V_b(\Gamma)$ let $n_\zeta(i)$ denote the number of vertices of Γ_ζ adjacent to i (these are all white since there are no G-edges in Γ_ζ, which are the only edges that join two black vertices).

Lemma 5.13. *We have the bound*

$$\left|W_{\mathbf{a}_b,\mathbf{a}_w}(\Gamma_\zeta)\right| \;\leqslant\; C_\Gamma\,\phi^{-p/2}\prod_{i\in V_b} N^{-n_\zeta(i)}\prod_{i\in V_b^*} M^{-1/2}\,.$$

Proof. Recalling (5.46), we assume without loss of generality that, for each $i\in V_b$, eack block of ζ_i has size at least two; in particular, we assume that for each $i\in V_b$ we have $\deg_\Gamma(i)\geqslant 2$. From (5.46) we get

$$\begin{aligned}\left|W_{\mathbf{a}_b,\mathbf{a}_w}(\Gamma_\zeta)\right| &\leqslant C_\Gamma(NM)^{-|V_w(\Gamma)|/4}\\ &= C_\Gamma\prod_{i\in V_b\setminus V_b^*}(NM)^{-\deg_\Gamma(i)/4}\prod_{i\in V_b^*}(NM)^{-\deg_\Gamma(i)/4}\\ &= C_\Gamma\prod_{i\in V_b\setminus V_b^*}\left(N^{-\deg_\Gamma(i)/2}\phi^{-\deg_\Gamma(i)/4}\right)\\ &\quad\times\prod_{i\in V_b^*}\left(M^{-1/2}N^{-(\deg_\Gamma(i)-1)/2}\phi^{1/2-\deg_\Gamma(i)/4}\right).\end{aligned}$$

By definition, τ_0 and ρ leave $\deg(i)$ invariant, and τ_1 increases $\deg(i)$ by 0 or 4. In particular, they all leave the parity of $\deg(i)$ invariant for $i\in V_b$. We conclude that $\deg_\Gamma(i)$ is odd for each $i\in V_b^*$. Since each block of ζ_i has size at least two, we find that

$$n_\zeta(i)\;\leqslant\;\begin{cases}\frac{\deg_\Gamma(i)}{2} & \text{if } \deg_\Gamma(i) \text{ is even}\\ \frac{\deg_\Gamma(i)-1}{2} & \text{if } \deg_\Gamma(i) \text{ is odd}\,.\end{cases}$$

We therefore conclude that

$$\left|W_{\mathbf{a}_b,\mathbf{a}_w}(\Gamma_\zeta)\right| \;\leqslant\; C_\Gamma\prod_{i\in V_b\setminus V_b^*}\left(N^{-n_\zeta(i)}\phi^{-\deg_\Gamma(i)/4}\right)\prod_{i\in V_b^*}\left(M^{-1/2}N^{-n_\zeta(i)}\phi^{1/2-\deg_\Gamma(i)/4}\right).$$

The proof is then completed by the following claim.

If $\deg_\Gamma(i)\geqslant 2$ for all $i\in V_b^*$ then

$$\prod_{i\in V_b}\phi^{-\deg_\Gamma(i)/4}\prod_{i\in V_b^*}\phi^{1/2}\;\leqslant\;\phi^{-p/2}\,. \tag{5.47}$$

What remains is to prove (5.47). Since Δ has p edges, we find

$$\prod_{i\in V_b}\phi^{-\deg_\Delta(i)/4}\;=\;\phi^{-p/2}\,.$$

As observed above, τ_0 and ρ leave $\deg(i)$ invariant, and τ_1 increases $\deg(i)$ by 0 or 4. Let $i\in V_b^*$. Since by assumption $\deg_\Gamma(i)\geqslant 2$, we find that in fact $\deg_\Gamma(i)\geqslant 5$. This yields

$$\begin{aligned}\prod_{i\in V_b}\phi^{-\deg_\Gamma(i)/4} &\leqslant \prod_{i\in V_b\setminus V_b^*}\phi^{-\deg_\Gamma(i)/4}\prod_{i\in V_b^*}\phi^{-5/4}\\ &\leqslant \prod_{i\in V_b\setminus V_b^*}\phi^{-\deg_\Delta(i)/4}\prod_{i\in V_b^*}\phi^{-\deg_\Delta(i)/4}\phi^{-1}\,,\end{aligned}$$

from which (5.47) follows. □

5.14 The nontrivial leaves III: summing over a and conclusion of the proof of Proposition 5.12

As above, fix a tree vertex $\sigma\in L_1(\mathcal{T})$, a graph $\Gamma\in\mathfrak{G}(\Theta_\sigma)$, and a partition ζ. In order to conclude the proof, we use Lemma 5.13 on each Γ_ζ to sum over $\mathbf{a}_w$ in (5.45).

Recall the quantity d from (5.31), defined as the number of off-diagonal G-edges plus the number of off-diagonal R-groups. By definition of Δ, $d(\Delta) = p$. Moreover, τ_0, τ_1, and ρ do not decrease d. Since by construction Γ has no G-entries, we conclude that Γ has at least p off-diagonal R-groups. We may therefore choose a set $E_o(\Gamma) \subset E_R(\Gamma)$ of size at least p, such that each $e \in E_o(\Gamma)$ is the centre of an off-diagonal R-group (see Section 5.6). The set E_o is naturally mapped into $E_R(\Gamma_\zeta)$, and is denoted by $E_o(\Gamma_\zeta)$. We denote by $\alpha(e)$ and $\beta(e)$ the end points of e in Γ_ζ. By (5.4), we have

$$\prod_{e \in E_R(\Gamma_\zeta)} \mathcal{A}_{\mathbf{a}_b, \mathbf{a}_w}(e, \Gamma_\zeta) \prec \prod_{e \in E_o(\Gamma_\zeta)} \Psi^{\mathbf{1}(\alpha(e) \neq \beta(e))} .$$

As in Section 5.11, it is easy to take the expectation using Lemma 3.2 to get

$$\mathbb{E} \prod_{e \in E_R(\Gamma_\zeta)} \mathcal{A}_{\mathbf{a}_b, \mathbf{a}_w}(e, \Gamma_\zeta) \prec \prod_{e \in E_o(\Gamma_\zeta)} \Psi^{\mathbf{1}(\alpha(e) \neq \beta(e))} .$$

We may now sum over $\mathbf{a}_w$ on the right-hand side of (5.45): from Lemma 5.13 we get

$$\begin{aligned}
&\sum_{\mathbf{a}_w} \chi_{\mathbf{a}_w}(\Gamma_\zeta) \left(\mathbb{E} \prod_{e \in E_R(\Gamma_\zeta)} \mathcal{A}_{\mathbf{a}_b, \mathbf{a}_w}(e, \Gamma_\zeta) \right) W_{\mathbf{a}_b, \mathbf{a}_w}(\Gamma_\zeta) \\
&\quad = \sum_{\mathbf{a}_w} \chi_{\mathbf{a}_w}(\Gamma_\zeta) \left(\mathbb{E} \prod_{e \in E_R(\Gamma_\zeta)} \mathcal{A}_{\mathbf{a}_b, \mathbf{a}_w}(e, \Gamma_\zeta) \right) W_{\mathbf{a}_b, \mathbf{a}_w}(\Gamma_\zeta) \\
&\qquad \times \prod_{e \in E_o(\Gamma_\zeta)} \Big(\mathbf{1}(\alpha(e) = \beta(e)) + \mathbf{1}(\alpha(e) \neq \beta(e)) \Big) \\
&\quad \prec C_\Gamma \, \phi^{-p/2} \prod_{i \in V_b} N^{-n_\zeta(i)} \prod_{i \in V_b^*} M^{-1/2} \sum_{k=0}^{p} \Psi^{p-k} N^{|V_w(\Gamma_\zeta)|-k} \\
&\quad \leqslant C_\Gamma \phi^{-p/2} \Psi^p \prod_{i \in V_b^*} M^{-1/2} .
\end{aligned}$$

In the second step we multiplied out the last p-fold product on the second line and classified all terms according to number, k, of factors $\mathbf{1}(\alpha(e) = \beta(e))$; we used that the total number of free summation variables is $|V_w(\Gamma_\zeta)| - k$. In the third step we used that $\sum_{i \in V_b} n_\zeta(i) = |V_w(\Gamma_\zeta)|$ and the bound $\Psi \geqslant N^{-1}$.

Returning to (5.45), we find

$$\sum_{\mathbf{a}_w} \mathbb{E} \mathcal{A}_{\mathbf{a}_b, \mathbf{a}_w}(\Gamma) \prec C_\Gamma \phi^{-p/2} \Psi^p \prod_{i \in V_b^*} M^{-1/2} .$$

We may now sum over $\mathbf{a}_b$ to prove (5.44). Using the bound (5.19), we therefore get

$$\begin{aligned}
\sum_{\mathbf{a}_b}^{*} w_{\mathbf{a}_b}(\Delta) \sum_{\mathbf{a}_w} \mathbb{E} \, \mathcal{A}_{\mathbf{a}_b, \mathbf{a}_w}(\Gamma) &\prec C_\Gamma \phi^{-p/2} \Psi^p \sum_{\mathbf{a}_b} \prod_{i \in V_b^*} M^{-1/2} \prod_{i \in V(\Delta)} |v_{a_i}|^{\deg_\Delta(i)} \\
&\prec C_\Gamma \phi^{-p/2} \Psi^p ,
\end{aligned}$$

where the last step follows from the fact that, by definition of Δ, $\deg_\Delta(i) \geqslant 1$ for all $i \in V_b$, as well as the estimate

$$\sum_a |v_a|^k \leqslant \begin{cases} M^{1/2} & \text{if } k = 1 \\ 1 & \text{if } k \geqslant 2 . \end{cases}$$

Summing over $\sigma \in L_1(\mathcal{T})$ concludes the proof of Proposition 5.12.

5.15 Conclusion of the proof of Theorem 3.11

Combining (5.38) and Proposition 5.12, and recalling (5.36) and (5.16), yieds

$$\mathbb{E}|\mathcal{Z}|^p \prec C_{p,\omega}(\phi^{-1}\Psi)^p .$$

Now (5.7), and hence (3.16), follows by a simple application of Chebyshev's inequality. Let $\varepsilon > 0$ and D be given. Then

$$\mathbb{P}(|\mathcal{Z}| > M^\varepsilon \phi^{-1}\Psi) \leqslant C_{p,\omega} M^\varepsilon M^{-\varepsilon p} \leqslant M^{-D}$$

for $p \geqslant \varepsilon^{-1}D + 2$. This, together with Remark 2.2 which allows us to interchange M and N in Definition 2.1, concludes the proof of (3.16).

Finally, we outline the proof of (3.17), which is very similar to that of (3.16). The expansion of Sections 5.4–5.12 may be taken over by swapping the roles of R and $\widetilde{G}$. In other words, we use Lemma 3.8 instead of Lemma 3.6. The arguments from Sections 5.13 and 5.14 carry over with straightforward adjustments in the power counting of Section 5.14. We leave the details to the interested reader. This concludes the proof of Theorem 3.11.

6 Proof of Theorems 3.12 and 3.13

Proof of Theorem 3.13. Note that, by the definition (2.17) of γ_α, we have $\gamma_\alpha \in [\gamma_-, \gamma_+]$ for all $\alpha = 1, \dots, N$. Hence, given $\varepsilon > 0$ and $c > 0$ as in Theorem 3.13, for small enough $\omega \in (0,1)$ we have $\gamma_\alpha \geqslant 2\omega$ provided that either $\alpha \leqslant (1-\varepsilon)N$ or $\phi \geqslant 1+c$. Set $\eta := N^{-1+\omega}$. We therefore conclude from Theorem 2.10 and the definition (2.11) of $\mathbf{S}$ that $\lambda_\alpha + \mathrm{i}\eta \in \mathbf{S}$ with high probability (see Definition 2.3), provided that either $\alpha \leqslant (1-\varepsilon)N$ or $\phi \geqslant 1+c$. Let α be such an index, and abbreviate $\Xi := \{\lambda_\alpha + \mathrm{i}\eta \in \mathbf{S}\}$. Then we get from (3.17), Remark 2.6, and (3.5) that $\mathbf{1}(\Xi)\operatorname{Im}\langle \mathbf{v}, R(\lambda_\alpha + \mathrm{i}\eta)\mathbf{v}\rangle \prec 1$. From

$$\operatorname{Im}\langle \mathbf{v}, R(\lambda_\alpha + \mathrm{i}\eta)\mathbf{v}\rangle = \sum_{\beta=1}^N \frac{\eta |\langle \mathbf{u}^{(\beta)}, \mathbf{v}\rangle|^2}{(\lambda_\alpha - \lambda_\beta)^2 + \eta^2} \geqslant \frac{|\langle \mathbf{u}^{(\alpha)}, \mathbf{v}\rangle|^2}{\eta}$$

we therefore get $\mathbf{1}(\Xi)|\langle \mathbf{u}^{(\alpha)}, \mathbf{v}\rangle|^2 \prec N^{-1+\omega}$. Since $\omega \in (0,1)$ can be made arbitrarily small and $1 - \mathbf{1}(\Xi) \prec 0$, the first estimate of Theorem 3.13 follows.

In order to prove the second estimate of Theorem 3.13, we use the same $\eta = N^{-1+\omega}$ as above and write $z = \lambda_\alpha + \mathrm{i}\eta$. Taking the imaginary part inside the absolute value on the left-hand side of (3.16), we get

$$\mathbf{1}(\Xi)\operatorname{Im}\langle \mathbf{w}, G(z)\mathbf{w}\rangle \prec \operatorname{Im} m_{\phi^{-1}}(z) + \frac{1}{\phi} \leqslant \frac{2}{\phi},$$

where in the second step we used (3.22), (3.5), and $z \in \mathbf{S}$ with high probability; this latter estimates follows from (2.5), the fact that $\gamma_\alpha \geqslant 2\omega$ by assumption, and Theorem 2.10. Repeating the above argument, we therefore find $\mathbf{1}(\Xi)|\langle \widetilde{\mathbf{u}}^{(\alpha)}, \mathbf{w}\rangle|^2 \prec \phi^{-1}N^{-1+\omega}$, and the second claim of Theorem 3.13 follows. □

Proof of Theorem 3.12. We only prove (3.19); the proof of (3.20) is the same, using (3.17) instead of (3.16). Moreover, to simplify notation, we assume that $E \geqslant \gamma_+ + N^{-2/3+\omega}$; the case $E \leqslant \gamma_- - N^{-2/3+\omega}$ is handled in exactly the same way.

Note first that if $\eta \geqslant \kappa$ then it is easy to see that (3.19) follows from (3.16), (3.6), and the lower bound $\eta \geqslant \kappa \geqslant N^{-2/3}$. For the following we therefore assume that $\eta \leqslant \kappa$. By Lemma 3.3, for $\eta \leqslant \kappa$ we have

$$\sqrt{\frac{\operatorname{Im} m_\phi(z)}{N\eta}} \asymp N^{-1/2}\kappa^{-1/4}.$$

By polarization and linearity, it therefore suffices to prove that

$$\big|\langle \mathbf{v}, G(z)\mathbf{v}\rangle - m_{\phi^{-1}}(z)\big| \;\prec\; \phi^{-1}N^{-1/2}\kappa^{-1/4}. \tag{6.1}$$

Define $\eta_0 := N^{-1/2}\kappa^{1/4}$. By definition of the domain $\widetilde{\mathbf{S}}$, we have $\eta_0 \leqslant \kappa$. Using (3.16), we find that (6.1) holds if $\eta \geqslant \eta_0$. For the following we therefore take $0 < \eta \leqslant \eta_0$. We proceed by comparison using the two spectral parameters

$$z \;:=\; E + \mathrm{i}\eta\,, \qquad z_0 \;:=\; E + \mathrm{i}\eta_0\,.$$

Since (6.1) holds at z_0 by (3.16), it is enough to prove the estimates

$$\big|m_\phi(z) - m_\phi(z_0)\big| \;\leqslant\; CN^{-1/2}\kappa^{-1/4} \tag{6.2}$$

and

$$\big|\langle \mathbf{v}, G(z)\mathbf{v}\rangle - \langle \mathbf{v}, G(z_0)\mathbf{v}\rangle\big| \;\prec\; \phi^{-1}N^{-1/2}\kappa^{-1/4}. \tag{6.3}$$

(The third required estimate, that of $\big|\frac{1-\phi}{\phi z} - \frac{1-\phi}{\phi z_0}\big|$, is trivial by $|z|^2 \asymp \phi$ for any $z \in \widetilde{\mathbf{S}}$.) We start with (6.2). From the definition $m_\phi(z) = \int \frac{\varrho_\phi(\mathrm{d}x)}{x-z}$ and the square root decay of the density of ϱ_ϕ near γ_+ from (2.4), it is not hard to derive the bound $m_\phi'(z) \leqslant C\kappa^{-1/2}$ for $z \in \widetilde{\mathbf{S}}$. Therefore we get

$$\big|m_\phi(z) - m_\phi(z_0)\big| \;\leqslant\; C\kappa^{-1/2}\eta_0 \;=\; CN^{-1/2}\kappa^{-1/4}\,,$$

which is (6.2).

What remains is to prove (6.3). By Theorem 2.10 we have $E \geqslant \lambda_1 + \eta_0$ with high probability since $\eta_0 \geqslant N^{-2/3+\omega/4}$. Therefore, since $\eta \leqslant \eta_0 \leqslant E - \lambda_1 \leqslant E - \lambda_\alpha$ for all $\alpha \geqslant 1$, we get

$$\begin{aligned}\mathrm{Im}\langle \mathbf{v}, G(z)\mathbf{v}\rangle \;&=\; \sum_\alpha \frac{|\langle \mathbf{v}, \mathbf{u}^{(\alpha)}\rangle|^2\eta}{(E-\lambda_\alpha)^2+\eta^2} \;\leqslant\; 2\sum_\alpha \frac{|\langle \mathbf{v}, \mathbf{u}^{(\alpha)}\rangle|^2\eta_0}{(E-\lambda_\alpha)^2+\eta_0^2} \\ &=\; 2\,\mathrm{Im}\langle \mathbf{v}, G(z_0)\mathbf{v}\rangle \;\prec\; \phi^{-1}N^{-1/2}\kappa^{-1/4}\end{aligned} \tag{6.4}$$

by (3.16) at z_0 and the estimate

$$\mathrm{Im}\bigg(\frac{1-\phi}{\phi z_0} + \frac{m_\phi(z_0)}{\phi}\bigg) \;\leqslant\; C\phi^{-1}N^{-1/2}\kappa^{-1/4}\,,$$

as follows from (3.6) and the estimate $|z_0|^2 \asymp \phi$.

Finally, we estimate the real part of the error in (6.3) using

$$\begin{aligned}\big|\mathrm{Re}\langle \mathbf{v}, G(z)\mathbf{v}\rangle - \mathrm{Re}\langle \mathbf{v}, G(z_0)\mathbf{v}\rangle\big| \;&=\; \sum_\alpha \frac{(E-\lambda_\alpha)(\eta_0^2-\eta^2)|\langle \mathbf{u}^{(\alpha)}, \mathbf{v}\rangle|^2}{\big((E-\lambda_\alpha)^2+\eta^2\big)\big((E-\lambda_\alpha)^2+\eta_0^2\big)} \\ &\leqslant\; \frac{\eta_0}{E-\lambda_1}\sum_\alpha \frac{\eta_0|\langle \mathbf{u}^{(\alpha)}, \mathbf{v}\rangle|^2}{(E-\lambda_\alpha)^2+\eta_0^2} \;\leqslant\; \mathrm{Im}\langle \mathbf{v}, G(z_0)\mathbf{v}\rangle\end{aligned} \tag{6.5}$$

with high probability, where in the last step we used that $\eta_0 \leqslant E - \lambda_1$ with high probability. Combining (6.4) and (6.5) completes the proof of (6.3), and hence of Theorem 3.12. □

7 Proofs for generalized Wigner matrices

In this section we explain how to modify the arguments of Sections 3–6 to the case of generalized Wigner matrices, and hence how to complete the proof of the results from Section 2.2. Since we are dealing with generalized Wigner matrices, in this section we consistently use the notations from Section 2.2 instead of Section 2.1.

7.1 Tools for generalized Wigner matrices

We begin by recalling some basic facts about the Stieltjes transform m from (2.24). In analogy to (2.9), for $E \in \mathbb{R}$ we define

$$\kappa \equiv \kappa_E := \big||E| - 2\big|, \tag{7.1}$$

the distance from E to the spectral edges ± 2.

Lemma 7.1. *For $|z| \leqslant 2\omega^{-1}$ we have*

$$|m(z)| \asymp 1, \qquad |1 - m(z)^2| \asymp \sqrt{\kappa + \eta} \tag{7.2}$$

and

$$\operatorname{Im} m(z) \asymp \begin{cases} \sqrt{\kappa+\eta} & \text{if } |E| \leqslant 2 \\ \frac{\eta}{\sqrt{\kappa+\eta}} & \text{if } |E| \geqslant 2\,. \end{cases}$$

(All implicit constants depend on ω.)

Proof. The proof is an elementary calculation; see Lemma 4.2 in [14]. □

The following definition is the analogue of Definitions 3.5 and 3.7. (Note that for generalized Wigner matrices we always simultaneously remove a row and the corresponding column.)

Definition 7.2 (Minors)**.** *For $T \subset \{1, \dots, N\}$ we define $H^{(T)}$ by*

$$(H^{(T)})_{ij} := \mathbf{1}(i \notin T)\mathbf{1}(j \notin T) H_{ij}\,.$$

Moreover, for $i, j \notin T$ we define the resolvent of the minor through

$$G^{(T)}_{ij}(z) := (H^{(T)} - z)^{-1}_{ij}\,.$$

We also set

$$\sum_i^{(T)} := \sum_{i : i \notin T}\,.$$

When $T = \{a\}$, we abbreviate $(\{a\})$ by (a) in the above definitions; similarly, we write (ab) instead of $(\{a, b\})$.

We shall also need the following resolvent identities, proved in Lemma 4.2 of [15] and Lemma 6.10 of [8].

Lemma 7.3 (Resolvent identities)**.** *For $i, j \neq k$ and $i, j, k \notin T$ the identity* (3.7) *holds. Moreover, for $i \neq j$ and $i, j \notin T$ we have*

$$G^{(T)}_{ij} = -G^{(T)}_{ii} \sum_k^{(Ti)} H_{ik} G^{(Ti)}_{kj} = -G^{(T)}_{jj} \sum_k^{(Tj)} G^{(Tj)}_{ik} H_{kj}\,. \tag{7.3}$$

Finally, for $i \notin T$ we have Schur's formula

$$\frac{1}{G^{(T)}_{ii}} = H_{ii} - z - \sum_{k,l}^{(Ti)} H_{ik} G^{(Ti)}_{kl} H_{li}\,. \tag{7.4}$$

From (7.3) we find for $i \neq j$ and $i, j \notin T$ that

$$G^{(T)}_{ij} = G^{(T)}_{ii} G^{(Ti)}_{jj} \bigg(-H_{ij} + \sum_{k,l}^{(Tij)} H_{ik} G^{(Tij)}_{kl} H_{lj} \bigg)\,, \tag{7.5}$$

7.2 The isotropic law: proof of Theorem 2.12

As in (5.7), and using (2.28) instead of (4.2), it suffices to prove that

$$|\mathcal{Z}| \prec \sqrt{\frac{\operatorname{Im} m(z)}{N\eta}} + \frac{1}{N\eta},$$

where, as in Section 5, $\mathcal{Z} = \sum_{a \neq b} \overline{v}_a G_{ab} v_b$. (Note that here there is no factor ϕ and hence no rescaled quantities bearing a tilde.) The estimate of $\mathbb{E}|\mathcal{Z}|^p$ follows the argument of Section 5; in particular, it consists of the eight steps sketched at the end of Section 5.3, which we follow closely. Throughout the argument we use the identities (7.4) and (7.5) instead of (3.8) and (3.9). In them, the two differences as compared to the argument of Section 5 are apparent.

1. The quadratic term in the expansion of $1/G_{ii}$ and G_{ij} contains an entry of G and not of R. (The matrix R is not even defined for generalized Wigner matrices.)
2. Both (7.4) and (7.5) contain an additional term, an entry of H.

Of these differences, the first one is minor. In order to mimic the bookkepping of Section 5, we still speak of black and white vertices. The simplest definition of our colouring is that the vertices of Δ are black and any other vertices that were added to Δ are white; see the explanation below (5.20). The second difference leads to a slightly larger class of graphs, but the new graphs will always be of subleading order. An alternative viewpoint is that the additional entry of H on the right-hand side of (7.4) and (7.5) should be regarded as a negligible error term. Almost all of the differences highlighted below stem from this additional term.

We now sketch the argument for each of the eight steps, by highlighting the changes as compared to Section 5.

Step 1. The reduced family of matrix indices is, as in Section 5, the set of indices $\mathbf{a}_b$ associated with the black vertices of Δ.

Step 2. We use (3.7) to make all entries of G maximally expanded of the black indices $\mathbf{a}_b = (a_i)_{i \in V_b}$. The graphical representation is the same as in Section 5.

Step 3. We use (7.5) to expand each maximally expanded off-diagonal entry of $G_{a_i a_j}$. As compared to the expansion based on (3.8) and (3.9) and used in Section 5, we get an additional term, $-H_{a_i a_j}$. See Figure 10 for a graphical representation of this expansion. Note that (7.5) is always invoked with $T = \mathbf{a}_b \setminus \{a_i, a_j\}$. Hence, for any graph Γ at any point of the argument, each white index is summed over the set $\{1, \dots, N\} \setminus \mathbf{a}_b$.

Step 4. Repeating Steps 2 and 3 in tandem yields a sum of monomials which consist only of maximally expanded diagonal entries of G with black indices, entries of $G^{(\mathbf{a}_b)}$ with white indices, and entries of H.

Step 5. We apply (7.4) to each maximally expanded diagonal entry of G. The graphical representation of this operation is similar to Figure 8, except that we also get a diagonal entry of H, depicted as a dotted loop. Note that there are no R-entries, but we still use the terminology of Section 5 and speak of R-groups; these refer to the to the same structure as in Figure 4, except that the edge e connecting the two white vertices encodes a G-edge and not an R-edge. We end up with entries of $G^{(\mathbf{a}_b)}$ and entries of H. Note that, by construction, each entry of H carries at least one black index, and that the G-edges are only incident to white vertices. In particular, all entries of H are independent of all entries of $G^{(\mathbf{a}_b)}$.

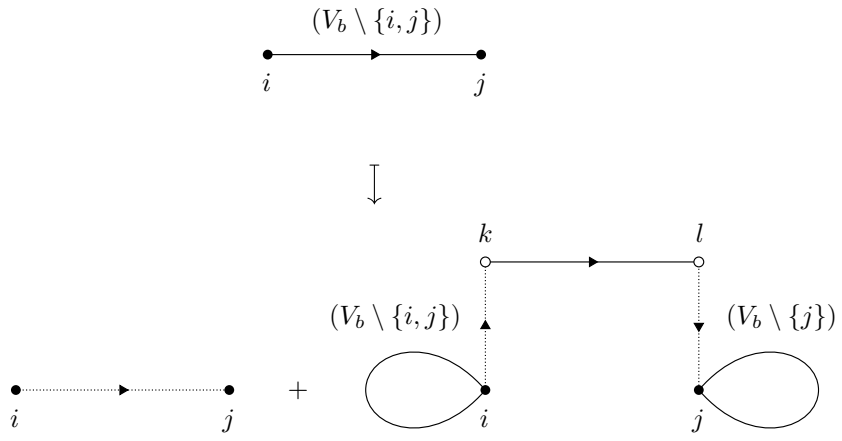

Figure 10: How Figure 6 is modified for generalized Wigner matrices. We use an oriented dotted line from i to j to draw the entry $-H_{a_i a_j}$.

Step 6. Using the independence of the entries of H and $G^{(\mathbf{a}_b)}$, we may take the partial expectation in the rows (or, equivalently, columns) indexed by $\mathbf{a}_b$. Note that now we have two classes of H-edges: white-black (incident to a black and a white vertex) and black-black (incident to two black vertices). Since the white indices are distinct from the black ones, the expectation factorizes over these two classes of H-edges. Exactly as in Section 5, taking the expectation in the white-black H-edges yields, for each $i \in V_b$, a partition of the white vertices adjacent to i, whereby each block of the partition must contain at least two vertices. The expectation over the black-black H-edges imposes an additional constraint among the loops incident to the white vertices, which are unimportant for the argument. Finally, for $i \neq j \in V_b$, we have the constraint that the number of edges joining i and j cannot be one.

Steps 7 and 8. The parity argument from the proof of Theorem 5.14 may be taken over with minor modifications, which arise from the additional black-black H-edges described in Step 6. Recall that the goal is to gain a factor $N^{-1/2}$ from each black vertex $i \in V_b$ that has an odd degree. If i is incident to a black-white H-edge, the counting from Section 5 applied unchanged and yields a power of $N^{-1/2}$. If i is not incident to a black-white H-edge, it must be incident to a black-black H-edge (recall that all graphs must be connected). By the constraints arising from the expectation in Step 6, i must then in fact be incident to at least two black-black H-edges which connect i to the same black vertex j. This yields a factor $\mathbb{E}|H_{a_i a_j}|^2 \leqslant C/N$, which is the desired small factor. (We may in general only allocate $N^{-1/2}$ from the factor N^{-1} to the vertex i, since j may also be a vertex that has degree one in Δ, in which case we have to allocate the other factor in $N = N^{-1/2}N^{-1/2}$ to j.)

This concludes sketch of how the argument of Section 5 is to be modified for the proof of Theorem 2.12. We omit further details. Finally, Theorems 2.15 and 2.16 follow from Theorem 2.12 by repeating the arguments of Section 6 almost to the letter.

References

[1] J. Baik, G. Ben Arous, and S. Péché, *Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices*, Ann. Prob. **33** (2005), 1643–1697. MR-2165575

[2] A. Bloemendal, A. Knowles, H.-T. Yau, and J. Yin, *On the principal components of sample covariance matrices*, In preparation.

[3] A. Bloemendal and B. Virág, *Limits of spiked random matrices II*, Preprint arXiv:1109.3704. MR-3078286

[4] _____, *Limits of spiked random matrices I*, Prob. Theor. Rel. Fields **156** (2013), 795–825. MR-3078286

[5] L. Erdős, *Universality for random matrices and log-gases*, Lecture Notes for Current Developments in Mathematics, Preprint arXiv:1212.0839 (2012).

[6] L. Erdős, A. Knowles, and H.-T. Yau, *Averaging fluctuations in resolvents of random band matrices*, to appear in Ann. H. Poincaré. Preprint arXiv:1205.5664. MR-3119922

[7] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Delocalization and diffusion profile for random band matrices*, Preprint arXiv:1205.5669. MR-3085669

[8] _____, *Spectral statistics of Erdős-Rényi graphs II: Eigenvalue spacing and the extreme eigenvalues*, Comm. Math. Phys. **314** (2012), 587–640. MR-2964770

[9] _____, *The local semicircle law for a general class of random matrices*, Electron. J. Probab **18** (2013), 1–58. MR-3068390

[10] _____, *Spectral statistics of Erdős-Rényi graphs I: Local semicircle law*, Ann. Prob. **41** (2013), 2279–2375. MR-3098073

[11] L. Erdős, B. Schlein, and H.-T. Yau, *Local semicircle law and complete delocalization for Wigner random matrices*, Comm. Math. Phys. **287** (2009), 641–655. MR-2481753

[12] _____, *Universality of random matrices and local relaxation flow*, Invent. Math. **185** (2011), no. 1, 75–119. MR-2810797

[13] L. Erdős, B. Schlein, H.-T. Yau, and J. Yin, *The local relaxation flow approach to universality of the local statistics of random matrices*, Ann. Inst. Henri Poincaré (B) **48** (2012), 1–46. MR-2919197

[14] L. Erdős, H.-T. Yau, and J. Yin, *Universality for generalized Wigner matrices with Bernoulli distribution*, J. Combinatorics **1** (2011), no. 2, 15–85. MR-2847916

[15] _____, *Bulk universality for generalized Wigner matrices*, Prob. Theor. Rel. Fields **154** (2012), 341–407. MR-2981427

[16] _____, *Rigidity of eigenvalues of generalized Wigner matrices*, Adv. Math **229** (2012), 1435–1515. MR-2871147

[17] I. M. Johnstone, *On the distribution of the largest eigenvalue in principal components analysis*, Ann. Stat. **29** (2001), 295–327. MR-1863961

[18] A. Knowles and J. Yin, *The isotropic semicircle law and deformation of Wigner matrices*, to appear in Comm. Pure Appl. Math. Preprint arXiv:1110.6449. MR-3103909

[19] _____, *The outliers of a deformed wigner matrix*, to appear in Ann. Prob. Preprint arXiv:1207.5619.

[20] V. A. Marchenko and L. A. Pastur, *Distribution of eigenvalues for some sets of random matrices*, Mat. Sbornik **72** (1967), 457–483.

[21] N. S. Pillai and J. Yin, *Universality of covariance matrices*, Preprint arXiv:1110.2501.

[22] E. P. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, Ann. Math. **62** (1955), 548–564. MR-0077805

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)

Economical model of EJP-ECP

- Low cost, based on free software (OJS[1])
- Non profit, sponsored by IMS[2], BS[3], PKP[4]
- Purely electronic and secure (LOCKSS[5])

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

[1]OJS: Open Journal Systems `http://pkp.sfu.ca/ojs/`
[2]IMS: Institute of Mathematical Statistics `http://www.imstat.org/`
[3]BS: Bernoulli Society `http://www.bernoulli-society.org/`
[4]PK: Public Knowledge Project `http://pkp.sfu.ca/`
[5]LOCKSS: Lots of Copies Keep Stuff Safe `http://www.lockss.org/`
[6]IMS Open Access Fund: `http://www.imstat.org/publications/open.htm`