Commun. Math. Phys. 333, 1365–1416 (2015)
Digital Object Identifier (DOI) 10.1007/s00220-014-2119-5

Communications in
Mathematical Physics

The Altshuler–Shklovskii Formulas for Random Band Matrices I: the Unimodular Case

László Erdős[1], **Antti Knowles**[2]

[1] IST Austria, Am Campus 1, Klosterneuburg 3400, Austria. E-mail: lerdos@ist.ac.at
[2] ETH Zürich, Zurich, Switzerland. E-mail: knowles@math.ethz.ch

Received: 25 September 2013 / Accepted: 3 March 2014
Published online: 17 July 2014 – © Springer-Verlag Berlin Heidelberg 2014

Abstract: We consider the spectral statistics of large random band matrices on mesoscopic energy scales. We show that the correlation function of the local eigenvalue density exhibits a universal power law behaviour that differs from the Wigner–Dyson–Mehta statistics. This law had been predicted in the physics literature by Altshuler and Shklovskii in (Zh Eksp Teor Fiz (Sov Phys JETP) 91(64):220(127), 1986); it describes the correlations of the eigenvalue density in general metallic samples with weak disorder. Our result rigorously establishes the Altshuler–Shklovskii formulas for band matrices. In two dimensions, where the leading term vanishes owing to an algebraic cancellation, we identify the first non-vanishing term and show that it differs substantially from the prediction of Kravtsov and Lerner in (Phys Rev Lett 74:2563–2566, 1995). The proof is given in the current paper and its companion (Ann. H. Poincaré. arXiv:1309.5107, 2014).

1. Introduction

The eigenvalue statistics of large random Hermitian matrices with independent entries are known to exhibit universal behaviour. Wigner proved [50] that the eigenvalue density converges (on the macroscopic scale) to the semicircle law as the dimension of the matrix tends to infinity. He also observed that the local statistics of individual eigenvalues (e.g., the gap statistics) are universal, in the sense that they depend only on the symmetry class of the matrix but are otherwise independent of the distribution of the matrix entries. In the Gaussian case, the local spectral statistics were identified by Gaudin, Mehta, and Dyson [36], who proved that they are governed by the celebrated sine kernel.

In this paper and its companion [11], we focus the universality of the eigenvalue density statistics on intermediate, so-called *mesoscopic*, scales, which lie between the

László Erdős on leave from Institute of Mathematics, University of Munich. Partially supported by SFB-TR 12 Grant of the German Research Council.

Antti Knowles partially supported by Swiss National Science Foundation Grant 144662.

macroscopic and the local scales. We study *random band matrices*, commonly used to model quantum transport in disordered media. Unlike the mean-field Wigner matrices, band matrices possess a nontrivial spatial structure. Apart from the obvious mathematical interest, an important motivation for this question arises from physics, namely from the theory of conductance fluctuations developed by Thouless [49]. In the next sections we explain the physical background of the problem. Thus, readers mainly interested in the mathematical aspects of our results may skip much of the introduction.

1.1. Metal-insulator transition. According to the Anderson metal-insulator transition [6], general disordered quantum systems are believed to fall into one of two very distinctive regimes. In the *localized regime* (also called the *insulator regime*), physical quantities depending on the position, such as eigenvectors and resolvent entries, decay on a length scale ℓ (called the *localization length*) that is independent of the system size. The unitary time evolution generated by the Hamiltonian remains localized for all times and the local spectral statistics are Poisson. In contrast, in the *delocalized regime* (also called the *metallic regime*), the localization length is comparable with the linear system size. The overlap of the eigenvectors induces strong correlations in the local eigenvalue statistics, which are believed to be universal and to coincide with those of a Gaussian matrix ensemble of the appropriate symmetry class. Moreover, the unitary time evolution generated by the Hamiltonian is diffusive for large times. Strongly disordered systems are in the localized regime. In the weak disorder regime, the localization properties depend on the dimension and on the energy.

Despite compelling theoretical arguments and numerical evidence, the Anderson metal-insulator transition has been rigorously proved only in a few very special cases. The basic model is the random Schrödinger operator, $-\Delta + V$, typically defined on $\mathbb{R}^d$ or on a graph (e.g., on a subset of $\mathbb{Z}^d$). Here V is a random potential with short-range spatial correlations; for instance, in the case of a graph, V is a family of independent random variables indexed by the vertices. The localized regime is relatively well understood since the pioneering work of Fröhlich and Spencer [28,29], followed by an alternative approach by Aizenman and Molchanov [1]. The Poissonian nature of the local spectral statistics was proved by Minami [37]. On the other hand, the delocalized regime has seen far less progress. With the exception of the Bethe lattice [2,26,32], only partial results are available. They indicate delocalization and quantum diffusion in certain limiting regimes [16–18,21], or in a somewhat different model where the static random potential is replaced with a dynamic phonon field in a thermal state at positive temperature [27,38].

Another much studied family of models describing disordered quantum systems is random matrices. Delocalization is well understood for random Wigner matrices [19,23], but, owing to their mean-field character, they are always in the delocalized regime, and hence no phase transition takes place. The local eigenvalue statistics are universal. This fundamental fact about random matrices, also known as the Wigner–Dyson–Mehta conjecture, has been recently proved [20,22,24] (see also [48] for a partially alternative argument in the Hermitian case).

1.2. Mesoscopic statistics. In a seminal paper [4], Altshuler and Shklovskii computed a new physical quantity: the variance of the number $\mathcal{N}_\eta$ of eigenvalues on a mesoscopic energy scale η in d-dimensional metallic samples with disorder for $d \leqslant 3$; here *mesoscopic* refers to scales η that are much larger than the typical eigenvalue spacing δ but much smaller than the total (macroscopic) energy scale of the system. Their motivation was to study fluctuations of the conductance in mesoscopic metallic samples; see also

[3,34]. The relationship between $\mathcal{N}_\eta$ and the conductance is given by a fundamental result of Thouless [49], asserting that the conductance of a sample of linear size L is determined by the (one-particle) energy levels in an energy band of a specific width η around the Fermi energy. In particular, the variance of $\mathcal{N}_\eta$ directly contributes to the conductance fluctuations. This specific value of η is given by $\eta = \max\{\eta_c, T\}$, where T is the temperature and η_c is the *Thouless energy* [49]. In diffusive models the Thouless energy is defined as $\eta_c := D/L^2$, where D is the diffusion coefficient. (In a conductor the dynamics of the particles, i.e., the itinerant electrons, is typically diffusive.) The Thouless energy may also be interpreted as the inverse diffusion time, i.e., the time needed for the particle to diffuse through the sample.

As it turns out, the mesoscopic linear statistics $\mathcal{N}_\eta$ undergo a sharp transition precisely at[1] $\eta \asymp \eta_c$. For small energy scales, $\eta \ll \eta_c$, Altshuler and Shklovskii found that the variance of $\mathcal{N}_\eta$ behaves according to[2]

$$\operatorname{Var} \mathcal{N}_\eta \asymp \log \mathcal{N}_\eta \asymp \log L, \tag{1.1}$$

as predicted by the Dyson–Mehta statistics [10]. The unusually small variance is due to the strong correlations among eigenvalues (arising from level repulsion). In the opposite regime, $\eta \gg \eta_c$, the variance is typically much larger, and behaves according to

$$\operatorname{Var} \mathcal{N}_\eta \asymp (\eta/\eta_c)^{d/2} = L^d (\eta/D)^{d/2} \qquad (d = 1, 2, 3). \tag{1.2}$$

The threshold η_c may be understood by introducing the concept of (an energy-dependent) *diffusion length* ℓ_η, which is the typical spatial scale on which the off-diagonal matrix entries of those observables decay that live on an energy scale η (e.g. resolvents whose spectral parameters have imaginary part η). Alternatively, ℓ_η is the linear scale of an initially localized state evolved up to time η^{-1}. The diffusion length is related to the localization length ℓ through $\ell = \lim_{\eta \to 0} \ell_\eta$. Assuming that the dynamics of the quantum particle can be described by a classical diffusion process, one can show that $\ell_\eta \asymp \sqrt{D/\eta}$ and the relation $\eta \ll \eta_c = D/L^2$ may be written as $L \ll \ell_\eta$. The physical interpretation is that the sample is so small that the system is essentially mean-field from the point of view of observables on the energy scale η, so that the spatial structure and dimensionality of the system are immaterial. The opposite regime $\eta \gg \eta_c$ corresponds to large samples, $L \gg \ell_\eta$, where the behaviour of the system can be approximated by a diffusion that has not reached the boundary of the sample. These two regimes are commonly referred to as *mean-field* and *diffusive* regimes, respectively.

A similar transition occurs if one considers the correlation of the number of eigenvalues $\mathcal{N}_\eta(E_1)$ and $\mathcal{N}_\eta(E_2)$ around two distinct energies $E_1 < E_2$ whose separation is much larger than the energy window η (i.e., $E_2 - E_1 \gg \eta$). For small samples, $\eta \ll \eta_c$, the correlation decays according to

$$\big\langle \mathcal{N}_\eta(E_1) \,;\, \mathcal{N}_\eta(E_2) \big\rangle \asymp (E_2 - E_1)^{-2}. \tag{1.3}$$

This decay holds for systems both with and without time reversal symmetry. The decay (1.3) is in agreement with the Dyson–Mehta statistics, which in the complex Hermitian case (corresponding to a system without time reversal symmetry) predict a correlation

$$\left(\frac{\sin\big((E_2 - E_1)/\delta\big)}{(E_2 - E_1)/\delta} \right)^2$$

[1] We use the notation $a \asymp b$ to indicate that a and b have comparable size. See the conventions at the end of Sect. 1.

[2] We use the notation $\langle \cdot \,;\, \cdot \rangle$ to denote the covariance and abbreviate $\operatorname{Var} X := \langle X \,;\, X \rangle$. See (2.10) below.

for highly localized observables on the scale $\eta \asymp \delta$. For mesoscopic scales, $\eta \gg \delta$, the oscillations in the numerator are averaged out and may be replaced with a positive constant to yield (1.3). A similar formula with the same decay holds for the real symmetric case (corresponding to a system with time reversal symmetry). On the other hand, for large samples, $\eta \gg \eta_c$, we have

$$\left\langle \mathcal{N}_\eta(E_1)\,;\mathcal{N}_\eta(E_2)\right\rangle \asymp (E_2-E_1)^{-2+d/2} \qquad (d=1,3)\,; \tag{1.4}$$

for $d = 2$ the correlation vanishes to leading order. The power laws in the energies η and $E_2 - E_1$ given in (1.2) and (1.4) respectively are called the *Altshuler–Shklovskii formulas*. They express the variance and the correlation of the density of states in the regime where the diffusion approximation is valid and the spatial extent of the diffusion, ℓ_η, is much less than the system size L. In contrast, the mean-field formulas (1.1) and (1.3) describe the situation where the diffusion has reached equilibrium. Note that the behaviours (1.2) and (1.4) as well as (1.1) and (1.3) are very different from the ones obtained if the distribution of the eigenvalues were governed by Poisson statistics; in that case, for instance, (1.3) and (1.4) would be zero.

From a mathematical point of view, the significance of these mesoscopic quantities is that their statistics are amenable to rigorous analysis even in the delocalized regime. In this paper we demonstrate this by proving the Altshuler–Shklovskii formulas for random band matrices.

1.3. Random band matrices. We consider d-dimensional random band matrices, which interpolate between random Schrödinger operators and mean-field Wigner matrices by varying the band width W; see [47] for an overview of this idea. These matrices represent quantum systems in a d-dimensional discrete box of side length L, where the quantum transition probabilities are random and their range is of order $W \ll L$. We scale the matrix so that its spectrum is bounded, i.e., the macroscopic energy scale is of order 1, and hence the eigenvalue spacing is of order $\delta \asymp L^{-d}$. Band matrices exhibit diffusion in all dimensions d, with a diffusion coefficient $D \asymp W^2$; see [30] for a physical argument in the general case and [12,13] for a proof up to certain large time scales. In [14] it was shown that the resolvent entries with spectral parameter $z = E + \mathrm{i}\eta$ decay exponentially on a scale $\ell_\eta \asymp W/\sqrt{\eta}$, as long as this scale is smaller than the system size, $W/\sqrt{\eta} \ll L$. (For technical reasons the proof is valid only if L is not too large, $L \ll W^{1+d/4}$.) The resolvent entries do not decay if $W/\sqrt{\eta} \gg L$, in which case the system is in the mean-field regime for observables living on energy scales of order $\eta \ll \eta_c$. Notice that the crossover at $W/\sqrt{\eta} \asymp L$ corresponds exactly to the crossover at $\eta \asymp \eta_c$ mentioned above.

1.4. Outline of results. Our main result is the proof of the formulas (1.2) (with $D = W^2$) and (1.4) for d-dimensional band matrices for $d \leqslant 3$; we also obtain similar results for $d = 4$, where the powers of η and $E_2 - E_1$ are replaced with a logarithm. This rigorously justifies the asymptotics of Silvestrov [42, Equation (40)], which in turn reproduced the earlier result of [4]. For technical reasons, we have to restrict ourselves to the regime $\eta \gg W^{-d/3}$. For convenience, we also assume that $L \gg W^{1+d/6}$, which guarantees that $L \gg \ell_\eta$ (or, equivalently, $\eta \gg \eta_c$). Hence we work in the diffusive regime. However, our method may be easily extended to the case $L \ll \ell_\eta$ as well (see Remark 2.8 and Sect. 2.3 below). We also show that for $d \geqslant 5$ the universality of the formulas (1.2) and (1.4) breaks down, and the variance and the correlation functions of $\mathcal{N}_\eta$ depend on the detailed structure of the band matrix. We also compute the leading correction to

the density–density correlation. In summary, we find that for $d = 1, 2$ the leading and subleading terms are universal, for $d = 3, 4$ only the leading terms are universal, and for $d \geqslant 5$ the density–density correlation is not universal.

The case $d = 2$ is special, since the coefficient of the leading term in (1.4) vanishes owing to an algebraic cancellation. The first non-vanishing term was predicted in [33]. We rigorously identify this term in the regime $E_2 - E_1 \gg \eta \gg W^{-2/3}$, and find a substantial discrepancy between it and the prediction of [33].

For an outline of our proof, and the relation between this paper and its companion [11], see Sect. 2.4.

1.5. Summary of previous related results. Our analysis is valid in the mesoscopic regime, i.e. when $\delta \ll \eta \ll 1$, and concerns only density fluctuations. For completeness, we mention what was previously known in this and other regimes.

Macroscopic statistics. In the macroscopic regime, $\eta \asymp 1$, the quantity $\mathcal{N}_\eta$ should fluctuate on the scale $(L/W)^{d/2}$ according to (1.2). For the Wigner case, $L = W$, it has been proved that a smoothed version of $\mathcal{N}_\eta$, the linear statistics of eigenvalues $\sum_i \phi(\lambda_i) = \operatorname{Tr} \phi(H)$, is asymptotically Gaussian. The first result in this direction for analytic ϕ was given in [43], and this was later extended by several authors to more general test functions; see [46] for the latest result. The first central limit theorem for matrices with a nontrivial spatial structure and for polynomial test functions was proved in [5]. Very recently, it was proved in [35] for one-dimensional band matrices that, provided that $\phi \in C^1(\mathbb{R})$, the quantity $\operatorname{Tr} \phi(H)$ is asymptotically Gaussian with variance of order $(L/W)^{d/2}$. For a complete list of references in this direction, see [35].

Mesoscopic statistics. The asymptotics (1.3) in the completely mean-field case, corresponding to Wigner matrices (i.e. $W = L$ so that $\eta_c \asymp 1$), was proved in [8,9]; see the remarks following Theorem 2.9 for more details about this work. We note that the formula (1.4) for random band matrices with $d = 1$ was derived in [7], using an unphysical double limit procedure, in which the limit $L \to \infty$ was first computed for a fixed η, and subsequently the limit of small η was taken. Note that the mesoscopic correlations cannot be recovered after the limit $L \to \infty$. Hence the result of [7] describes only the macroscopic, and not the mesoscopic, correlations.

Local spectral statistics. Much less is known about the local spectral statistics of random band matrices, even for $d = 1$. The Tracy–Widom law at the spectral edge was proved in [44]. Based on a computation of the localization length, the metal-insulator transition is predicted to occur at $W^2 \asymp L$; see [30] for a non-rigorous argument and [12–14,39] for the best currently known lower and upper bounds. Hence, the local spectral statistics are expected to be governed by the sine kernel from random matrix theory in the regime $W^2 \gg L$. Very recently, the sine kernel was proved [41] for a special Hermitian Gaussian random band matrix with band width W comparable with L. Universality for a more general class of band matrices but with an additional tiny mean-field component was proved in [15]. We also mention that the local correlations of determinants of a special Hermitian Gaussian random band matrix have been shown to follow the sine kernel [40], up to the expected threshold $L \lesssim W^2$.

1.6. Transition to Poisson statistics. The diagrammatic calculation of [4] uses the diffusion approximation, and formulas (1.1)–(1.4) are supposed to be valid in the delocalized

regime. Nevertheless, our results also hold in the localized regime, in particular even $d = 1$ and for $L \gg W^8$, in which case the eigenvectors are known to be localized [39]. In this regime, and for $\eta \gg W^{-1/3}$, we also prove (1.2) and (1.4). Both formulas show that Poisson statistics do not hold on large mesoscopic scales, despite the system being in the localized regime. Indeed, if $\mathcal{N}_\eta$ were Poisson-distributed, then we would have $\operatorname{Var}\mathcal{N}_\eta \asymp \mathcal{N}_\eta \asymp L\eta$. On the other hand, (1.2) gives $\operatorname{Var}\mathcal{N}_\eta \asymp L\sqrt{\eta}/W$. We conclude that the prediction of (1.2) for the magnitude of $\operatorname{Var}\mathcal{N}_\eta$ is much smaller than that predicted by Poisson statistics provided that $\eta \gg W^{-2}$.

The fact that the Poisson statistics breaks down on mesoscopic scales is not surprising. Indeed, the basic intuition behind the emergence of Poisson statistics is that eigenvectors belonging to different eigenvalues are exponentially localized on a scale $\ell \asymp W^2$, typically at different spatial locations. Hence the associated eigenvalues are independent. For larger η, however, the observables depend on many eigenvalues, which exhibit nontrivial correlations since the supports of their eigenvectors overlap. A simple counting argument shows that such overlaps become significant if $\eta \gg 1/\ell$, at which point correlations are expected to develop. In other words, we expect a transition to/from Poisson statistics at $\eta \asymp 1/\ell$. In the previous paragraph, we noted that (1.2) predicts a transition in the behaviour of $\mathcal{N}_\eta$ to/from Poisson statistics at $\eta \asymp W^{-2}$. Combining these observations, we therefore expect a transition to/from Poisson statistics for $\ell \asymp W^2$. This argument predicts the correct localization length $\ell \asymp W^2$ without resorting to Grassmann integration. It remains on a heuristic level, however, since our results do not cover the full range $\eta \gg W^{-2}$. We note that this argument may also be applied to $d \geqslant 2$, in which case it predicts the absence of a transition provided that $W \gg 1$.

The main conclusion of our results is that the local eigenvalue statistics, characterized by either Poisson or sine kernel statistics, do not in general extend to mesoscopic scales. On mesoscopic scales, a different kind of universality emerges, which is expressed by the Altshuler–Shklovskii formulas.

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. We use $a \asymp b$ to mean $ca \leqslant b \leqslant Ca$ for some constants $c, C > 0$. Also, for any finite set A we use $|A|$ to denote the cardinality of A. If the implicit constants in the usual notation $O(\cdot)$ depend on some parameters α, we sometimes indicate this explicitly by writing $O_\alpha(\cdot)$.

2. Setup and Results

2.1. Definitions and assumptions. Fix $d \in \mathbb{N}$, the physical dimension of the configuration space. For $L \in \mathbb{N}$ we define the discrete torus of size L

$$\mathbb{T} \equiv \mathbb{T}_L^d := \big([-L/2, L/2) \cap \mathbb{Z}\big)^d,$$

and abbreviate

$$N := |\mathbb{T}_L| = L^d. \tag{2.1}$$

Let $1 \ll W \leqslant L$ denote the band width, and define the deterministic matrix $S = (S_{xy})$ through

$$S_{xy} := \frac{\mathbf{1}(1 \leqslant |x - y| \leqslant W)}{M - 1}, \qquad M := \sum_{x \in \mathbb{T}} \mathbf{1}(1 \leqslant |x| \leqslant W), \tag{2.2}$$

where $|\cdot|$ denotes the periodic Euclidean norm on $\mathbb{T}$, i.e. $|x| := \min_{v\in\mathbb{Z}^d} |x+Lv|_{\mathbb{Z}^d}$. Note that

$$M \asymp W^d. \tag{2.3}$$

The fundamental parameters of our model are the linear dimension of the torus, L, and the band width, W. The quantities N and M are introduced for notational convenience, since most of our estimates depend naturally on N and M rather than L and W. We regard L as the independent parameter, and $W \equiv W_L$ as a function of L.

Next, let $A = A^* = (A_{xy})$ be a Hermitian random matrix whose upper-triangular[3] entries $(A_{xy} : x \leqslant y)$ are independent random variables with zero expectation. We consider two cases.

- *The real symmetric case ($\beta = 1$)*, where A_{xy} satisfies $\mathbb{P}(A_{xy} = 1) = \mathbb{P}(A_{xy} = -1) = 1/2$.
- *The complex Hermitian case ($\beta = 2$)*, where A_{xy} is uniformly distributed on the unit circle $\mathbb{S}^1 \subset \mathbb{C}$.

Here the index $\beta = 1, 2$ is the customary symmetry index of random matrix theory.

We define the *random band matrix* $H = (H_{xy})$ through

$$H_{xy} := \sqrt{S_{xy}}\, A_{xy}. \tag{2.4}$$

Note that H is Hermitian and $|H_{xy}|^2 = S_{xy}$, i.e. $|H_{xy}|$ is deterministic. Moreover, we have for all x

$$\sum_y S_{xy} = \frac{M}{M-1}. \tag{2.5}$$

With this normalization, as $N, W \to \infty$ the bulk of the spectrum of $H/2$ lies in $[-1, 1]$ and the eigenvalue density is given by the Wigner semicircle law with density

$$\nu(E) := \frac{2}{\pi}\sqrt{1-E^2} \qquad \text{for} \quad E \in [-1, 1]. \tag{2.6}$$

Let ϕ be a smooth, integrable, real-valued function on $\mathbb{R}$ satisfying $\int \phi(E)\,\mathrm{d}E \neq 0$. We call such functions ϕ *test functions*. We also require that our test functions ϕ satisfy one of the two following conditions.

(C1) ϕ is the Cauchy kernel

$$\phi(E) = \operatorname{Im} \frac{2}{E - \mathrm{i}} = \frac{2}{E^2+1}. \tag{2.7}$$

(C2) For every $q > 0$ there exists a constant C_q such that

$$|\phi(E)| \leqslant \frac{C_q}{1+|E|^q}. \tag{2.8}$$

[3] We introduce an arbitrary and immaterial total ordering $\leqslant$ on the torus $\mathbb{T}$.

A typical example of a test function ϕ satisfying **(C2)** is the Gaussian $\phi(E) = \sqrt{2\pi}\,\mathrm{e}^{-E^2/2}$. We introduce the rescaled test function $\phi^\eta(E) := \eta^{-1}\phi(\eta^{-1}E)$. We shall be interested in correlations of observables depending on $E \in (-1, 1)$ of the form

$$Y_\phi^\eta(E) := \frac{1}{N}\sum_i \phi^\eta(\lambda_i - E) = \frac{1}{N}\operatorname{Tr}\phi^\eta(H/2 - E),$$

where $\lambda_1, \ldots, \lambda_N$ denote the eigenvalues of $H/2$. (The factor $1/2$ is a mere convenience, chosen because, as noted above, the asymptotic spectrum of $H/2$ is the interval $[-1, 1]$.) The quantity $Y_\phi^\eta(E)$ is the smoothed local density of states around the energy E on the scale η. We always choose

$$\eta = M^{-\rho}$$

for some fixed $\rho \in (0, 1/3)$, and we frequently drop the index η from our notation. The strongest results are for large ρ, so that one should think of ρ being slightly less than $1/3$.

We are interested in the correlation function of the local densities of states, $Y_{\phi_1}^\eta(E_1)$ and $Y_{\phi_2}^\eta(E_2)$, around two energies $E_1 \leqslant E_2$. We shall investigate two regimes: $\eta \ll E_2 - E_1$ and $E_1 = E_2$. In the former regime, we prove that the correlation decay in the energy difference $E_2 - E_1$ is universal (in particular, independent of η, ϕ_1, and ϕ_2), and we compute the correlation function explicitly. In the latter regime, we prove that the variance has a universal dependence on η, and depends on ϕ_1 and ϕ_2 via their inner product in a homogeneous Sobolev space.

The case **(C2)** for our test functions is the more important one, since we are typically interested in the statistics of eigenvalues contained in an interval of size η. The Cauchy kernel from the case **(C1)** has a heavy tail, which introduces unwanted correlations arising from the overlap of the test functions and not from the long-distance correlations that we are interested in. Nevertheless, we give our results also for the special case **(C1)**. We do this for two reasons. First, the case **(C1)** is pedagogically useful, since it results in a considerably simpler computation of the main term (see [11, Section 3] for more details). Second, the case **(C1)** is often the only one considered in the physics literature (essentially because it corresponds to the imaginary part of the resolvent of H). Hence, our results in particular decouple the correlation effects arising from the heavy tails of the test functions from those arising from genuine mesoscopic correlations. As proved in Theorem 2.4 below, the effect of the heavy tail is only visible in the leading nonzero corrections for $d = 2$.

For simplicity, throughout the following we assume that both of our test functions satisfy **(C1)** or both satisfy **(C2)**. Since the covariance is bilinear, one may also consider more general test functions that are linear combinations of the cases **(C1)** and **(C2)**.

Definition 2.1. Throughout the following we use the quantities $E_1, E_2 \in (-1, 1)$ and

$$E := \frac{E_1 + E_2}{2}, \qquad \omega := E_2 - E_1$$

interchangeably. Without loss of generality we always assume that $\omega \geqslant 0$.

For the following we choose and fix a positive constant κ. We always assume that

$$E_1, E_2 \in [-1 + \kappa, 1 - \kappa], \qquad \omega \leqslant c_* \tag{2.9}$$

for some small enough positive constant c_* depending on κ. These restrictions are required since the nature of the correlations changes near the spectral edges ± 1. Throughout the following we regard the constants κ and c_* as fixed and do not track the dependence of our estimates on them.

We now state our results on the density–density correlation for band matrices in the diffusive regime (Sect. 2.2). The proofs are given in the current paper and its companion [11]. As a reference, we also state similar results for Wigner matrices, corresponding to the mean-field regime (Sect. 2.3).

2.2. Band matrices. Our first theorem gives the leading behaviour of the density–density correlation function in terms of a function $\Theta^{\eta}_{\phi_1,\phi_2}(E_1, E_2)$, which is explicit but has a complicated form. In the two subsequent theorems we determine the asymptotics of this function in two physically relevant regimes, where its form simplifies substantially. We use the abbreviations

$$\langle X \rangle := \mathbb{E} X, \qquad \langle X; Y \rangle := \mathbb{E}(XY) - \mathbb{E}X\,\mathbb{E}Y. \tag{2.10}$$

Theorem 2.2 (Density–density correlations). *Fix $\rho \in (0, 1/3)$ and $d \in \mathbb{N}$, and set $\eta := M^{-\rho}$. Suppose that the test functions ϕ_1 and ϕ_2 satisfy either both **(C1)** or both **(C2)**. Suppose moreover that*

$$W^{1+d/6} \leqslant L \leqslant W^{C} \tag{2.11}$$

for some constant C.

Then there exist a constant $c_0 > 0$ and a function $\Theta^{\eta}_{\phi_1,\phi_2}(E_1, E_2)$—which is given explicitly in (4.90) and (4.37) below, and whose asymptotic behaviour is derived in Theorems 2.3 and 2.4 below—such that, for any E_1, E_2 satisfying (2.9) for small enough $c_* > 0$, the local density–density correlation satisfies

$$\frac{\langle Y^{\eta}_{\phi_1}(E_1)\,; Y^{\eta}_{\phi_2}(E_2)\rangle}{\langle Y^{\eta}_{\phi_1}(E_1)\rangle\langle Y^{\eta}_{\phi_2}(E_2)\rangle} = \frac{1}{(LW)^d}\Big(\Theta^{\eta}_{\phi_1,\phi_2}(E_1, E_2) + O\big(M^{-c_0} R_2(\omega+\eta)\big)\Big), \tag{2.12}$$

where we defined

$$R_2(s) := 1 + \mathbf{1}(d = 1)s^{-1/2} + \mathbf{1}(d = 2)|\log s|. \tag{2.13}$$

*Moreover, if ϕ_1 and ϕ_2 are analytic in a strip containing the real axis (e.g. as in the case **(C1)**), we may replace the upper bound $L \leqslant W^C$ in (2.11) $L \leqslant \exp(W^c)$ for some small constant $c > 0$.*

We shall prove that the error term in (2.12) is smaller than the main term Θ for all $d \geqslant 1$. The main term Θ has a simple, and universal, explicit form only for $d \leqslant 4$. Why $d = 4$ is the critical dimension for the universality of the correlation decay is explained in Sect. 3.2 below. The two following theorems give the leading behaviour of the function Θ for $d \leqslant 4$ in the two regimes $\omega = 0$ and $\omega \gg \eta$. In fact, one may also compute the subleading corrections to Θ. These corrections turn out to be universal for $d \leqslant 2$ but not for $d \geqslant 3$; see Theorem 2.4 and the remarks following it.

In order to describe the leading behaviour of the variance, i.e. the case $\omega = 0$, we introduce the Fourier transform

$$\phi(E) = \int_{\mathbb{R}} \mathrm{d}t\, \mathrm{e}^{-\mathrm{i}Et}\, \widehat{\phi}(t), \qquad \widehat{\phi}(t) = \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{d}E\, \mathrm{e}^{\mathrm{i}Et}\, \phi(E).$$

For $d \leqslant 4$ we define the quadratic form V_d through

$$V_d(\phi_1, \phi_2) := \int_{\mathbb{R}} \mathrm{d}t\, |t|^{1-d/2}\, \overline{\widehat{\phi_1}(t)}\, \widehat{\phi_2}(t) \quad (d \leqslant 3), \qquad V_4(\phi_1, \phi_2) := 2\overline{\widehat{\phi_1}(0)}\, \widehat{\phi_2}(0). \tag{2.14}$$

Note that $V_d(\phi_1, \phi_2)$ is real since both ϕ_1 and ϕ_2 are. In the case **(C1)** we have the explicit values

$$V_0(\phi_1, \phi_2) = \frac{1}{2}, \quad V_1(\phi_1, \phi_2) = \frac{\sqrt{\pi}}{2\sqrt{2}}, \quad V_2(\phi_1, \phi_2) = 1,$$
$$V_3(\phi_1, \phi_2) = \sqrt{2\pi}, \quad V_4(\phi_1, \phi_2) = 2.$$

For the following statements of results, we recall the density $\nu(E)$ of the semicircle law from (2.6), and remind the reader of the index $\beta = 1, 2$ describing the symmetry class of H.

Theorem 2.3 (The leading term Θ for $\omega = 0$). *Suppose that the assumptions in the first paragraph of Theorem 2.2 hold, and let $\Theta^{\eta}_{\phi_1,\phi_2}(E_1, E_2)$ be the function from Theorem 2.2. Suppose in addition that $\omega = 0$. Then there exists a constant $c_1 > 0$ such that the following holds for $E = E_1 = E_2$ satisfying (2.9).*

(i) *For $d = 1, 2, 3$ we have*

$$\Theta^{\eta}_{\phi_1,\phi_2}(E, E) = \frac{(d+2)^{d/2}}{2\beta\pi^{2+d}\nu(E)^4}\left(\frac{\eta}{\nu(E)}\right)^{d/2-2}\big(V_d(\phi_1, \phi_2) + O(M^{-c_1})\big). \tag{2.15}$$

(ii) *For $d = 4$ we have*

$$\Theta^{\eta}_{\phi_1,\phi_2}(E, E) = \frac{36}{\beta\pi^6\nu(E)^4}\big(V_4(\phi_1, \phi_2)|\log\eta| + O(1)\big). \tag{2.16}$$

In order to describe the behaviour of Θ in the regime $\omega \gg \eta$, for $d = 1, 2, 3$ we introduce the constants

$$K_d := 2\,\mathrm{Re}\int_{\mathbb{R}^d} \frac{\mathrm{d}x}{(\mathrm{i} + |x|^2)^2}\,; \tag{2.17}$$

explicitly,

$$K_1 = -\frac{\pi}{\sqrt{2}}, \qquad K_2 = 0, \qquad K_3 = \sqrt{2}\pi^2.$$

Theorem 2.4 (The leading term Θ in the regime $\omega \gg \eta$). *Suppose that the assumptions in the first paragraph of Theorem 2.2 hold, and let $\Theta^{\eta}_{\phi_1,\phi_2}(E_1, E_2)$ be the function from Theorem 2.2. Suppose in addition that*

$$\eta \leqslant M^{-\tau}\omega \tag{2.18}$$

for some arbitrary but fixed $\tau > 0$. Then there exists a constant $c_1 > 0$ such that the following holds for E_1, E_2 satisfying (2.9) for small enough $c_* > 0$.

(i) *For $d = 1, 2, 3$ we have*

$$\Theta^{\eta}_{\phi_1,\phi_2}(E_1, E_2) = \frac{(d+2)^{d/2}}{2\beta\pi^{2+3d/2}\nu(E)^4}\left(\frac{\omega}{\nu(E)}\right)^{d/2-2}\Big(K_d + O\big(\sqrt{\omega} + M^{-c_1}\big)\Big). \tag{2.19}$$

(ii) *For $d = 2$ (2.19) does not identify the leading term since $K_2 = 0$. The leading nonzero correction to the vanishing leading term is*

$$\Theta^{\eta}_{\phi_1,\phi_2}(E_1, E_2) = \frac{8}{\beta\pi^5\nu(E)^4}\left(\pi\nu(E)\,\frac{\eta}{\omega^2+4\eta^2} - \frac{|\log\omega|}{3} + O(1)\right) \tag{2.20}$$

*in the case **(C1)** and*

$$\Theta^{\eta}_{\phi_1,\phi_2}(E_1, E_2) = \frac{8}{\beta\pi^5\nu(E)^4}\left(-\frac{|\log\omega|}{3} + O(1)\right) \tag{2.21}$$

*in the case **(C2)**.*

(iii) *For $d = 4$ we have*

$$\Theta^{\eta}_{\phi_1,\phi_2}(E_1, E_2) = \frac{36}{\beta\pi^6\nu(E)^4}\big(|\log\omega| + O(1)\big). \tag{2.22}$$

Note that the leading non-zero terms in the expressions (2.15), (2.16), (2.19)–(2.22) are much larger than the additive error term in (2.12). Hence, Theorems 2.2 and 2.3 give a proof of the first Altshuler–Shklovskii formula, (1.2). Similarly, Theorems 2.2 and 2.4 give a proof of the second Altshuler–Shklovskii formula, (1.4).

The additional term in (2.20) as compared to (2.21) originates from the heavy Cauchy tail in the test functions ϕ_1, ϕ_2 at large distances. In Theorem 2.4 (ii), we give the leading correction, of order $|\log\omega|$, to the vanishing main term for $d = 2$. Similarly, for $d = 1$ one can also derive the leading correction to the nonzero main term (which is of order $\omega^{-3/2}$). This correction turns out to be of order $\omega^{-1/2}$; we omit the details.

Remark 2.5. The leading term in (2.12) originates from the so-called one-loop diagrams in the terminology of physics. The next-order term after the vanishing leading term for $d = 2$ (recall that to $K_2 = 0$) was first computed by Kravtsov and Lerner [33, Equation (13)]. They found that for $\beta = 1$ it is of order $(LW)^{-2}W^{-2}\omega^{-1}$ and for $\beta = 2$ even smaller, of order $(LW)^{-2}W^{-4}\omega^{-1}$. Part (ii) of Theorem 2.4 shows that, at least in the regime $\omega \gg \eta \gg M^{-1/3} = W^{-2/3}$, the true behaviour is much larger. The origin of this term is a more precise computation of the one-loop diagrams, in contrast to [33] where the authors attribute the next-order term to the two-loop diagrams. (See [11, Section 3] for more details.)

Remark 2.6. If the distribution of the eigenvalues λ_i of $H/2$ were governed by Poisson statistics, the behaviour of the covariance (2.12) would be very different. Indeed, suppose that $\{\lambda_i\}$ is a stationary Poisson point process with intensity N. Then, setting $Y^{\eta}_{\phi}(E) := \frac{1}{N}\sum_{\alpha}\phi^{\eta}(\lambda_i - E)$ and supposing that $\int\phi_1 = \int\phi_2 = 1$, we find

$$\frac{\langle Y^{\eta}_{\phi_1}(E_1)\,;\,Y^{\eta}_{\phi_2}(E_2)\rangle}{\langle Y^{\eta}_{\phi_1}(E_1)\rangle\langle Y^{\eta}_{\phi_2}(E_2)\rangle} = \frac{1}{N\eta}\big(\phi_1 * \widetilde{\phi}_2\big)\left(\frac{E_2 - E_1}{\eta}\right) = \frac{1}{N}\big(\phi_1^{\eta} * \widetilde{\phi}_2^{\eta}\big)(\omega),$$

where $\widetilde{\phi}_2(x) := \phi_2(-x)$. This is in stark contrast to (2.15), (2.16), (2.19), and (2.22). In particular, in the case $\omega \gg \eta$ the behaviour of the covariance on ω depends on the tails of ϕ_1 and ϕ_2, unlike in (2.19) and (2.22). Hence, if ϕ_1 and ϕ_2 are compactly supported then the covariance for the Poisson process is zero, while for the eigenvalue process of a band matrix it has a power law decay in ω.

Remark 2.7. We emphasize that Theorem 2.2 is true under the sole restrictions (2.9) on ω and E_1, E_2. However, the leading term Θ only has a simple and universal form in the two (physically relevant) regimes $\omega = 0$ and $\eta \ll \omega \ll 1$ of Theorems 2.3 and 2.4. If neither of these conditions holds, the expression for Θ is still explicit but much more cumbersome and opaque. It is given by the sum of the values of eight (sixteen for $\beta = 1$) skeleton graphs, after a ladder resummation; these skeleton graphs are referred to as the "dumbbell skeletons" $D_1, \ldots, D_8$ in Sect. 4 below, and are depicted in Fig. 6 below. (They are the analogues of the *diffusion* and *cooperon* Feynman diagrams in the physics literature.)

Remark 2.8. The upper bound in the assumption (2.11) is technical and can be relaxed. The lower bound in (2.11), however, is a natural restriction, and is related to the *quantum diffusion* generated by the band matrix H. In [13], it was proved that the propagator $|(\mathrm{e}^{-\mathrm{i}tH/2})_{x0}|$ behaves diffusively for $1 \ll t \ll M^{1/3} \asymp W^{d/3}$, whereby the spatial extent of the diffusion is $x \asymp \sqrt{t}W \ll W^{1+d/6}$. Similarly, in [14], it was proved that the resolvent $\left|(H/2 - E - \mathrm{i}\eta)^{-1}_{x0}\right|^2$ has a nontrivial profile on the scale $x \asymp \eta^{-1/2}W$. (Note that η is the conjugate variable to t, i.e. the time evolution up to time t describes the same regime as the resolvent with a spectral parameter z whose imaginary part is $\eta \asymp 1/t$.) Since in Theorem 2.2 we assume that $\eta \gg M^{-1/3} \gg W^{-d/3}$, the condition (2.11) simply states that the diffusion profile associated with the spectral resolution η does not reach the edge of the torus $\mathbb{T}$. Thus, the lower bound in (2.11) imposes a regime in which boundary effects are irrelevant. Hence we are in the diffusive regime—a basic assumption of the Altshuler–Shklovskii formulas (1.2) and (1.4).

2.3. A remark on Wigner matrices. Our method can easily be applied to the case where the lower bound in (2.11) is not satisfied. In this case, however, the leading behaviour $\Theta^\eta(E_1, E_2)$ is modified by boundary effects. To illustrate this phenomenon, we state the analogue of Theorems 2.2–2.4 for the case of $W = L$. In this case, the physical dimension d in Sect. 2.1 is irrelevant. The off-diagonal entries of H are all identically distributed, i.e. H is a standard Wigner matrix (neglecting the irrelevant diagonal entries), and we have $M = N - 1$, and $S = N(N-2)^{-1}(\mathbf{e}\mathbf{e}^* - N^{-1})$ where $\mathbf{e} := N^{-1/2}(1, 1, \ldots, 1)^*$. In particular, H is a mean-field model in which the geometry of $\mathbb{T}$ plays no role; the effective dimension is $d = 0$. In this case (2.12) remains valid, and we get the following result.

Theorem 2.9 (Theorems 2.2–2.4 for Wigner matrices). *Suppose that $W = L = N$. Fix $\rho \in (0, 1/3)$ and set $\eta := N^{-\rho}$. Suppose that the test functions ϕ_1 and ϕ_2 satisfy either both **(C1)** or both **(C2)**. Then there exists a constant $c_0 > 0$ and a function $\widetilde{\Theta}^\eta_{\phi_1,\phi_2}(E_1, E_2)$ such that for any E_1, E_2 satisfying (2.9) for small enough $c_* > 0$ the following holds.*

(i) *The local density–density correlations satisfy*

$$\frac{\langle Y^\eta_{\phi_1}(E_1)\,;\, Y^\eta_{\phi_2}(E_2)\rangle}{\langle Y^\eta_{\phi_1}(E_1)\rangle\langle Y^\eta_{\phi_2}(E_2)\rangle} = \frac{1}{N^2}\Big(\widetilde{\Theta}^\eta_{\phi_1,\phi_2}(E_1, E_2) + O\big(N^{-c_0}(\omega+\eta)^{-1}\big)\Big). \tag{2.23}$$

(ii) *If* (2.18) *holds then*

$$\widetilde{\Theta}^{\eta}_{\phi_1,\phi_2}(E_1, E_2) = \frac{4}{\beta\pi^4\nu(E)^4}\,\frac{1}{\omega^2}\Big(-1 + O\big(\sqrt{\omega} + N^{-\tau/2}\big)\Big).$$

(iii) *If* $\omega = 0$ *then*

$$\widetilde{\Theta}^{\eta}_{\phi_1,\phi_2}(E, E) = \frac{2}{\beta\pi^4\nu(E)^4}\,\frac{1}{\eta^2}\,\big(V_0(\phi_1, \phi_2) + O(N^{-c_0})\big),$$

where V_0 *was defined in* (2.14).

The proof of Theorem 2.9 proceeds along the same lines as that of Theorems 2.2–2.4. In fact, the simple form of S results in a much easier proof; we omit the details. We remark that a result analogous to parts (i) and (ii) of Theorem 2.9, in the case where $\phi_1 = \phi_2$ are given by (2.7), was derived in [8,9]. More precisely, in [8], the authors assume that H is a GOE matrix and derive (i) and (ii) of Theorem 2.9 for any $0 < \rho < 1$; in [9], they extend these results to arbitrary Wigner matrices under the additional constraint that $0 < \rho < 1/8$. Moreover, results analogous to (iii) for the Gaussian Circular Ensembles were proved in [45]. More precisely, in [45] it is proved that in Gaussian Circular Ensembles the appropriately scaled mesoscopic linear statistics $Y^{\eta}_{\phi}(E)$ with $1/N \ll \eta \ll 1$ are asymptotically Gaussian with variance proportional to $V_0(\phi, \phi)$. We remark that for random band matrices the mesoscopic linear statistics also satisfy a Central Limit Theorem; see [11, Corollary 2.6].

2.4. Structure of the proof. The starting point of the proof is to use the Fourier transform to rewrite $\mathrm{Tr}\,\phi^{\eta}(H/2 - E)$, the spectral density on scale η, in terms of $\mathrm{e}^{\mathrm{i}tH}$ up to times $|t| \lesssim \eta^{-1}$. The large-t behaviour of this unitary group has been extensively analysed in [12,13] by developing a graphical expansion method which we also use in this paper. The main difficulty is to control highly oscillating sums. Without any resummation, the sum of the absolute values of the summands diverges exponentially in L, although their actual sum remains bounded. The leading divergence in this expansion is removed using a resummation that is implemented by expanding $\mathrm{e}^{\mathrm{i}tH}$ in terms of Chebyshev polynomials of H instead of powers of H. This step, motivated by [25], is algebraic and requires the deterministic condition $|H_{xy}| = 1$. (The removal of this condition is possible, but requires substantial technical efforts that mask the essence of the argument; see Sect. 2.5). In the jargon of diagrammatic perturbation theory, this resummation step corresponds to the self-energy renormalization.

The goal of [12,13] was to show that the unitary propagator $\mathrm{e}^{\mathrm{i}tH}$ can be described by a diffusive equation on large space and time scales. This analysis identified only the leading behaviour of $\mathrm{e}^{\mathrm{i}tH}$, which was sufficient to prove quantum diffusion emerging from the unitary time evolution. The quantity studied in the current paper—the local density–density correlation—is considerably more difficult to analyse because it arises from higher-order terms of $\mathrm{e}^{\mathrm{i}tH}$ than the quantum diffusion. Hence, not only does the leading term have to be computed more precisely, but the error estimates also require a much more delicate analysis. In fact, we have to perform a second algebraic resummation procedure, where oscillatory sums corresponding to families of specific subgraphs, the so-called ladder subdiagrams, are bundled together and computed with high precision. Estimating individual ladder graphs in absolute value is not affordable: a term-by-term estimate is possible only after this second renormalization step. Although

the expansion in nonbacktracking powers of H is the same as in [12,13], our proof in fact has little in common with that of [12,13]; the only similarity is the basic graphical language. In contrast to [12,13], almost all of the work in this paper involves controlling oscillatory sums, both in the error estimates and in the computation of the main term.

The complete proof is given in the current paper and its companion [11]. In order to highlight the key ideas, the current paper contains the proof assuming three important simplifications, given precisely in **(S1)**–**(S3)** in Sect. 4.1 below. They concern certain specific terms in the multiple summations arising from our diagrammatic expansion. Roughly, these simplifications amount to only dealing with typical summation label configurations (hence ignoring exceptional label coincidences) and restricting the summation over all partitions to a summation over pairings. As explained in [11], dealing with exceptional label configurations and non-pairings requires significant additional efforts, which are, however, largely unrelated to the essence of the argument presented in the current paper. How to remove these simplifications, and hence complete the proofs, is explained in [11]. In addition, the precise calculation of the leading term is also given in [11]; in the current paper we give a sketch of the calculation (see Sect. 3.2 below).

We close this subsection by noting that the restriction $\rho < 1/3$ for the exponent of $\eta = M^{-\rho}$ is technical and stems from a fundamental combinatorial fact that underlies our proof—the so-called *2/3-rule*. The 2/3-rule was introduced in [12,13] and is stated in the current context in Lemma 4.11 below. In [13, Section 11], it was shown that the 2/3-rule is sharp, and is in fact saturated for a large family of graphs. For more details on the 2/3-rule and how it leads the the restriction on ρ, we refer to the end Sect. 4.4 below.

2.5. Outlook and generalizations. We conclude this section by summarizing some extensions of our results from the companion paper [11]. First, our results easily extend from the two-point correlation functions of (2.12) to arbitrary k-point correlation functions of the form

$$\mathbb{E}\prod_{i=1}^{k}\left(\frac{Y_{\phi_i}^{\eta}(E_i)-\mathbb{E}Y_{\phi_i}^{\eta}(E_i)}{\mathbb{E}Y_{\phi_i}^{\eta}(E_i)}\right).$$

In [11, Theorem 2.5], we prove that the joint law of the smoothed densities $Y_{\phi_i}^{\eta}(E_i)$ is asymptotically Gaussian with covariance matrix $(\Theta_{\phi_i,\phi_j}^{\eta}(E_i,E_j))_{i,j}$, given by the Altshuler–Shklovskii formulas. This result may be regarded as a Wick theorem for the mesoscopic densities, i.e. a central limit theorem for the mesoscopic linear statistics of eigenvalues. In particular, if $E_1 = \cdots = E_k$, the finite-dimensional marginals of the process $(Y_{\phi}^{\eta}(E))_{\phi}$ converge (after an appropriate affine transformation) to those of a Gaussian process with covariance $V_d(\cdot,\cdot)$.

Second, in [11, Section 2.4] we introduce a general family of band matrices, where we allow the second moments $S_{xy} = \mathbb{E}|H_{xy}|^2$ and $T_{xy} = \mathbb{E}H_{xy}^2$ to be arbitrary translation-invariant matrices living on the scale W. In particular, we generalize the sharp step profile from (2.2) and relax the deterministic condition $|A_{xy}| = 1$. Note that we allow T_{xy} to be arbitrary up to the trivial constraint $|T_{xy}| \leqslant S_{xy}$, thus embedding the real symmetric matrices and the complex Hermitian matrices into a single large family of band matrices. In particular, this generalization allows us to probe the transition from $\beta = 1$ to $\beta = 2$ by rotating the entries of H or by scaling T_{xy}. Note that $S = T$ corresponds to the real symmetric case, while $T = 0$ corresponds to a complex Hermitian

case where the real and imaginary parts of the matrix elements are uncorrelated and have the same variance. We can combine this rotation and scaling into a two-parameter family of models; roughly, we consider $T_{xy} \approx (1-\varphi)\mathrm{e}^{\mathrm{i}\lambda}S_{xy}$ where $\varphi, \lambda \in [0,1]$ are real parameters. We show that the mesoscopic statistics described by Theorems 2.3 and 2.4 take on a more complicated form in the case of the general band matrix model; they depend on the additional parameter $\sigma = \lambda^2 + \varphi$, which also characterizes the transition from $\beta = 1$ (small σ) to $\beta = 2$ (large σ). We refer to [11, Section 2.4] for the details.

3. The Renormalized Path Expansion

Since the left-hand side of (2.12) is invariant under the scaling $\phi \mapsto \lambda\phi$ for $\lambda \neq 0$, we assume without loss of generality that $\int \mathrm{d}E\, \phi_i(E) = 2\pi$ for $i = 1, 2$. We shall make this assumption throughout the proof without further mention.

3.1. Expansion in nonbacktracking powers. We expand $\phi^\eta(H/2 - E)$ in nonbacktracking powers $H^{(n)}$ of H, defined through

$$H^{(n)}_{x_0x_n} := \sum_{x_1,\dots,x_{n-1}} H_{x_0x_1}\cdots H_{x_{n-1}x_n} \prod_{i=0}^{n-2} \mathbf{1}(x_i \neq x_{i+2}). \tag{3.1}$$

From [13], Section 5, we find that

$$H^{(n)} = U_n(H/2) - \frac{1}{M-1}U_{n-2}(H/2), \tag{3.2}$$

where U_n is the n-th Chebyshev polynomial of the second kind, defined through

$$U_n(\cos\theta) = \frac{\sin(n+1)\theta}{\sin\theta}. \tag{3.3}$$

The identity (3.2) first appeared in [25]. Note that it requires the deterministic condition $|A_{xy}| = 1$ on the entries of H. However, our basic approach still works even if this condition is not satisfied; in that case the proof is more complicated due to the presence of a variety of error terms in (3.2). See [11, Section 5.3] for more details.

From [13], Lemmas 5.3 and 7.9, we recall the expansion in nonbacktracking powers of H.

Lemma 3.1. *For $t \geqslant 0$ we have*

$$\mathrm{e}^{-\mathrm{i}tH/2} = \sum_{n\geqslant 0} a_n(t) H^{(n)}, \tag{3.4}$$

where

$$a_n(t) := \sum_{k\geqslant 0} \frac{\alpha_{n+2k}(t)}{(M-1)^k}, \qquad \alpha_k(t) := 2(-\mathrm{i})^k \frac{k+1}{t} J_{k+1}(t) \tag{3.5}$$

and J_ν denotes the ν-th Bessel function of the first kind.

Throughout the following we denote by arcsin the analytic branch of arcsin extended to the real axis by continuity from the upper half-plane. The following coefficients will play a key role in the expansion. For $n \in \mathbb{N}$ and $E \in \mathbb{R}$ define

$$\gamma_n(E) := \int_0^\infty \mathrm{d}t\, \mathrm{e}^{\mathrm{i}Et}\, a_n(t).$$

Lemma 3.2. *We have*

$$\gamma_n(E) = \frac{2(-\mathrm{i})^n \mathrm{e}^{\mathrm{i}(n+1)\arcsin E}}{1-(M-1)^{-1}\mathrm{e}^{2\mathrm{i}\arcsin E}}. \tag{3.6}$$

Proof. Using (3.5) we find

$$\gamma_n(E) = \sum_{k=0}^\infty \frac{\int_0^\infty \mathrm{d}t\, \mathrm{e}^{\mathrm{i}Et}\, \alpha_{n+2k}(t)}{(M-1)^k} = \sum_{k=0}^\infty \frac{2(-\mathrm{i})^n \mathrm{e}^{\mathrm{i}(n+2k+1)\arcsin E}}{(M-1)^k}, \tag{3.7}$$

where in the second step we used the identity

$$\int_0^\infty \mathrm{d}t\; t^{-1}\mathrm{e}^{\mathrm{i}Et} J_\nu(t) = \frac{1}{\nu}\,\mathrm{e}^{\mathrm{i}\nu\arcsin E}, \tag{3.8}$$

which is an easy consequence of [31, Formulas 6.693.1–6.693.2] and analytic continuation. This concludes the proof. □

Define

$$F^\eta_{\phi_1,\phi_2}(E_1,E_2) \equiv F^\eta(E_1,E_2) := \big\langle \mathrm{Tr}\,\phi_1^\eta(H/2-E_1)\,;\,\mathrm{Tr}\,\phi_2^\eta(H/2-E_2)\big\rangle, \tag{3.9}$$

where we used the notation (2.10). Note that the left-hand side of (2.12) may be written as

$$\frac{\langle Y^\eta_{\phi_1}(E_1)\,;\,Y^\eta_{\phi_2}(E_2)\rangle}{\langle Y^\eta_{\phi_1}(E_1)\rangle\langle Y^\eta_{\phi_2}(E_2)\rangle} = \frac{1}{N^2}\frac{F^\eta(E_1,E_2)}{\mathbb{E}Y^\eta_{\phi_1}(E_1)\,\mathbb{E}Y^\eta_{\phi_2}(E_2)}. \tag{3.10}$$

The expectations in the denominator are easy to compute using the local semicircle law for band matrices; see Lemma 4.24 below. Our main goal is to compute $F^\eta(E_1,E_2)$.

Throughout the following we use the abbreviation

$$\psi(E) := \phi(-E), \tag{3.11}$$

and define ψ^η, ψ_i, and ψ_i^η similarly similarly in terms of ϕ^η, ϕ_i, and ϕ_i^η. We also use the notation

$$(\varphi * \chi)(E) := \frac{1}{2\pi}\int \mathrm{d}E'\,\varphi(E-E')\,\chi(E') \tag{3.12}$$

to denote convolution. The normalizing factor $(2\pi)^{-1}$ is chosen so that $\widehat{\varphi * \chi} = \widehat{\varphi}\,\widehat{\chi}$. Observe that

$$(\psi^\eta * \gamma_n)(E) = \int_0^\infty \mathrm{d}t\, \mathrm{e}^{\mathrm{i}Et}\,\widehat{\phi}(\eta t)\, a_n(t). \tag{3.13}$$

We note that in the case where $\phi(E) = \frac{2}{E^2+1}$, we have $\widehat{\phi}(t) = \mathrm{e}^{-|t|}$. Hence (3.13) implies

$$(\psi^\eta * \gamma_n)(E) = \int_0^\infty \mathrm{d}t\, \mathrm{e}^{\mathrm{i}(E+\mathrm{i}\eta)t}\, a_n(t) = \gamma_n(E+\mathrm{i}\eta). \tag{3.14}$$

This may also be interpreted using the identity

$$\frac{1}{\pi}\int \mathrm{d}E'\, \mathrm{e}^{\mathrm{i}n \arcsin E'} \frac{\eta}{(E-E')^2+\eta^2} = \mathrm{e}^{\mathrm{i}n \arcsin(E+\mathrm{i}\eta)}.$$

We now return to the case of a general real ϕ. Since ϕ is real, we have $\overline{\widehat{\phi}(t)} = \widehat{\phi}(-t)$. We may therefore use Lemma 3.1 and Fourier transformation to get

$$\begin{aligned}\phi^\eta(H/2-E) &= \int_{-\infty}^\infty \mathrm{d}t\, \widehat{\phi}(\eta t)\, \mathrm{e}^{-\mathrm{i}t(H/2-E)} = 2\,\mathrm{Re}\int_0^\infty \mathrm{d}t\, \widehat{\phi}(\eta t)\, \mathrm{e}^{\mathrm{i}tE}\, \mathrm{e}^{-\mathrm{i}tH/2} \\ &= 2\,\mathrm{Re}\sum_{n=0}^\infty H^{(n)} \int_0^\infty \mathrm{d}t\, \widehat{\phi}(\eta t)\, \mathrm{e}^{\mathrm{i}tE} a_n(t) = \sum_{n=0}^\infty H^{(n)}\, 2\,\mathrm{Re}(\psi^\eta * \gamma_n)(E),\end{aligned} \tag{3.15}$$

where Re denotes the Hermitian part of a matrix, i.e. $\mathrm{Re}\, A := (A + A^*)/2$, and in the last step we used (3.13) and the fact that $H^{(n)}$ is Hermitian. We conclude that

$$F^\eta(E_1, E_2) = \sum_{n_1, n_2 \geqslant 0} 2\,\mathrm{Re}\big((\psi_1^\eta * \gamma_{n_1})(E_1)\big) 2\,\mathrm{Re}\big((\psi_2^\eta * \gamma_{n_2})(E_2)\big) \big\langle \mathrm{Tr}\, H^{(n_1)}; \mathrm{Tr}\, H^{(n_2)} \big\rangle. \tag{3.16}$$

Because the combinatorial estimates of Sect. 4 deteriorate rapidly for $n \gg \eta^{-1}$, it is essential to cut off the terms $n > M^\mu$ in the expansion (3.16), where $\rho < \mu < 1/3$. Thus, we choose a cutoff exponent μ satisfying $\rho < \mu < 1/3$. All of the estimates in this paper depend on ρ, μ, and ϕ; we do not track this dependence. The following result gives the truncated version of (3.16), whereby the truncation is done in n_i and in the support of $\widehat{\phi}_i$.

Proposition 3.3 (Path expansion with truncation). *Choose $\mu < 1/3$ and $\delta > 0$ satisfying $2\delta < \mu - \rho < 3\delta$. Define*

$$\widetilde{\gamma}_n(E, \phi) := \int_0^{M^{\rho+\delta}} \mathrm{d}t\, \mathrm{e}^{\mathrm{i}Et}\, \widehat{\phi}(\eta t)\, a_n(t) \tag{3.17}$$

and

$$\widetilde{F}^\eta(E_1, E_2) := \sum_{n_1+n_2 \leqslant M^\mu} 2\,\mathrm{Re}\big(\widetilde{\gamma}_{n_1}(E_1, \phi_1)\big)\, 2\,\mathrm{Re}\big(\widetilde{\gamma}_{n_2}(E_2, \phi_2)\big)\, \big\langle \mathrm{Tr}\, H^{(n_1)}; \mathrm{Tr}\, H^{(n_2)} \big\rangle. \tag{3.18}$$

Let $q > 0$ be arbitrary. Then for any $n \in \mathbb{N}$ and recalling (3.11) we have the estimates

$$|(\psi_i^\eta * \gamma_n)(E_i) - \widetilde{\gamma}_n(E_i, \phi_i)| \leqslant C_q M^{-q} \qquad (i = 1, 2) \tag{3.19}$$

and

$$\big|F^\eta(E_1, E_2) - \widetilde{F}^\eta(E_1, E_2)\big| \leqslant C_q N^2 M^{-q}. \tag{3.20}$$

Moreover, for all $q > 0$ we have

$$\big|\widetilde{\gamma}_n(E_i, \phi_i)\big| + \big|(\psi_i^\eta * \gamma_n)(E_i)\big| \leqslant \min\big\{C, C_q(\eta n)^{-q}\big\}. \tag{3.21}$$

If ϕ_1 and ϕ_2 are analytic in a strip containing the real axis, the factors $C_q M^{-q}$ on the right-hand sides of (3.19) *and* (3.20) *may be replaced with $\exp(-M^c)$ for some $c > 0$, and the factor $C_q(\eta n)^{-q}$ on the right-hand side of* (3.21) *by $\exp(-(\eta n)^c)$.*

The proof of Proposition 3.3 is given in [11, Appendix A].

3.2. Heuristic calculation of the leading term. At this point we make a short digression to outline how we compute the leading term of $F^\eta(E_1, E_2)$. The precise calculation is given in the companion paper [11]. In Sect. 4 below, we express the right-hand side of (3.16) as a sum of terms indexed by graphs, reminiscent of Feynman graphs in perturbation theory. We prove that the leading contribution is given by a certain set of relatively simple graphs, which we call the *dumbbell skeletons*. Their value $\mathcal{V}_{\mathrm{main}}$ may be explicitly computed and is essentially given by

$$\mathcal{V}_{\mathrm{main}} \approx \sum_{b_1,b_2,b_3,b_4=0}^{\infty} 2\,\mathrm{Re}\big(\gamma_{2b_1+b_3+b_4} * \psi_1^\eta\big)(E_1)\, 2\,\mathrm{Re}\big(\gamma_{2b_2+b_3+b_4} * \psi_2^\eta\big)(E_2)\,\mathrm{Tr}\, S^{b_3+b_4} \tag{3.22}$$

(see (4.37) below for the precise statement). The summations represent "ladder subdiagram resummations" in the terminology of graphs. Proving that the contribution of all other graphs is negligible, and hence that (3.22) gives the leading behaviour of (3.16), represents the main work, and is done in Sect. 4. Assuming that this approximation is valid, we compute (3.22) as follows. We use

$$(2\,\mathrm{Re}\, x_1)(2\,\mathrm{Re}\, x_2) = 2\,\mathrm{Re}(x_1\overline{x}_2 + x_1 x_2) \tag{3.23}$$

on the right-hand side of (3.22), and only consider the first resulting term; the second one will turn out to be subleading in the regime $\omega, \eta \ll \kappa$, owing to a phase cancellation. Recalling the definition of γ_n from (3.7), we find that the summations over $b_1, \ldots, b_4$ are simply geometric series, so that

$$\begin{aligned} F^\eta(E_1, E_2) \approx 2\,\mathrm{Re}\bigg(4\frac{\mathrm{e}^{\mathrm{i}A_1}}{1+\mathrm{e}^{2\mathrm{i}A_1}}\frac{\mathrm{e}^{-\mathrm{i}A_2}}{1+\mathrm{e}^{-2\mathrm{i}A_2}}\,\mathrm{Tr}\,\frac{\mathrm{e}^{\mathrm{i}(A_1-A_2)}S}{\big(1-\mathrm{e}^{\mathrm{i}(A_1-A_2)}S\big)^2}\bigg) \\ * \psi_1^\eta(E_1) * \psi_2^\eta(E_2), \end{aligned} \tag{3.24}$$

where we abbreviated $A_i := \arcsin E_i$, and wrote, by a slight abuse of notation, $(\varphi * \chi)(E) \equiv \varphi(E) * \chi(E)$.

In order to understand the behaviour of this expression, we make some basic observations about the spectrum of S. Since S is translation invariant, i.e. $S_{xy} = S_{x-y\,0}$, it may be diagonalized by Fourier transformation,

$$\widehat{S}_W(q) := \sum_{x\in\mathbb{T}} \mathrm{e}^{-\mathrm{i}q\cdot x/W} S_{x0} \approx \int \mathrm{e}^{-\mathrm{i}q\cdot x} f(x)\,\mathrm{d}x,$$

where f is the normalized indicator function of the unit ball in $\mathbb{R}^d$; in the last step we used the definition of S and a Riemann sum approximation. (Note that, since S lives on

the scale W, it is natural to rescale the argument q of the Fourier transform by W^{-1}.) From this representation it is not hard to see that $S \geqslant -1 + c$ for some constant $c > 0$. Moreover, for small q we may expand $\widehat{S}_W(q)$ to get $\widehat{S}_W(q) \approx 1 - q \cdot Dq$, where we defined the covariance matrix[4] $D_W \equiv D = (D_{ij})$ of S through

$$D_{ij} := \frac{1}{2} \sum_{x \in \mathbb{T}} \frac{x_i x_j}{W^2} S_{x0}. \tag{3.25}$$

We deduce that S has a simple eigenvalue at 1, with associated eigenvector $(1, 1, \ldots, 1)$, and all remaining eigenvalues lie in the interval $[-1 + c, 1 - c(W/L)^2]$ for some small constant $c > 0$. Therefore the resolvent on the right-hand side of (3.24) is near-singular (hence yielding a large contribution) for $\mathrm{e}^{\mathrm{i}(A_1 - A_2)} \approx 1$. This implies that the leading behaviour of (3.24) is governed by small values of q in Fourier space.

We now outline the computation of (3.24) in more detail. Let us first focus on the regime $\omega \gg \eta$, i.e. the regime from Theorem 2.4. Thus, the function ψ_i^η may be approximated by 2π times a delta function, so that the convolutions may be dropped. What therefore remains is the calculation of the trace. We write

$$\alpha := \mathrm{e}^{\mathrm{i}(A_1 - A_2)} \approx 1 - \mathrm{i}\omega(1 - E^2)^{-1/2} = 1 - \mathrm{i}\frac{2\omega}{\pi \nu}, \qquad \nu \equiv \nu(E),$$

in the regime $\omega \ll 1$. We use the Fourier representation of S and only consider the contribution of small values of q. After some elementary computations we get, for $d \leqslant 3$,

$$\begin{aligned} \operatorname{Tr} \frac{S}{(1 - \alpha S)^2} &\approx \frac{L^d}{W^d} \int_{\mathbb{R}^d} \mathrm{d}q \, \frac{\widehat{S}_W(q)}{(1 - \alpha \widehat{S}_W(q))^2} \approx \frac{L^d}{W^d} \int_{\mathbb{R}^d} \frac{\mathrm{d}q}{(1 - \alpha + q \cdot Dq)^2} \\ &\approx \frac{L^d}{W^d} \frac{1}{\sqrt{\det D}} \left(\frac{2\omega}{\pi \nu} \right)^{d/2 - 2} \int_{\mathbb{R}^d} \frac{\mathrm{d}q}{(\mathrm{i} + q^2)^2}. \end{aligned}$$

A similar calculation may be performed for $d = 4$, which results in a logarithmic behaviour in ω. This yields the right-hand sides of (2.19) and (2.22). For $d \leqslant 4$ the main contribution arises from the regime $q \approx 0$ and is therefore universal. If $d \geqslant 5$ the leading contribution to (3.22) arises from all values of q. While (3.22) may still be computed for $d \geqslant 5$, it loses its universal character and depends on the whole function $\widehat{S}_W(q)$.

In the regime $\omega = 0$, i.e. the regime from Theorem 2.3, we introduce $e := \psi_1 * \phi_2$ (recall (3.11)) and write (for simplicity setting $E_1 = E_2 = 0$ and $d \leqslant 3$)

$$\begin{aligned} \operatorname{Tr} \frac{\mathrm{e}^{\mathrm{i}(A_1 - A_2)} S}{\left(1 - \mathrm{e}^{\mathrm{i}(A_1 - A_2)} S\right)^2} * \psi_1^\eta(E_1) * \psi_2^\eta(E_2) &\approx \int_{\mathbb{R}} \mathrm{d}v \, e^\eta(v) \operatorname{Tr} \frac{S}{\left(1 - (1 - \mathrm{i}v)S\right)^2} \\ &\approx \frac{L^d}{W^d} \int_{\mathbb{R}} \mathrm{d}v \, e^\eta(v) \int_{\mathbb{R}^d} \mathrm{d}q \, \frac{1}{\left(\mathrm{i}v + q \cdot Dq\right)^2} \\ &= \frac{C L^d}{W^d} \int_{\mathbb{R}^d} \mathrm{d}q \int_0^\infty \mathrm{d}t \, \mathrm{e}^{-tq \cdot Dq} \, t \, \overline{\widehat{e}(\eta t)} \\ &= \frac{C L^d}{W^d \sqrt{\det D}} \int_0^\infty \mathrm{d}t \, t^{1 - d/2} \, \overline{\widehat{e}(\eta t)}, \end{aligned}$$

[4] To avoid confusion, we remark that this D differs from the D used in the introduction by a factor of order W^{-2}.

where in the third step we used an elementary identity of Fourier transforms. From this expression it will be easy to conclude (2.15), and an analogous calculation for $d = 4$ yields (2.16).

4. Proof of Theorems 2.2–2.4 for $\beta = 2$

In this section we prove Theorems 2.2–2.4 by computing the limiting behaviour of $\widetilde{F}^{\eta}(E_1, E_2)$. For simplicity, throughout this section we assume that we are in the complex Hermitian case, $\beta = 2$. The real symmetric case, $\beta = 1$, can be handled by a simple extension of the arguments of this section, and is presented in Sect. 5.

Due to the independence of the matrix entries (up to the Hermitian symmetry), the expectation of a product of matrix entries in (3.18) can be computed simply by counting how many times a matrix entry (or its conjugate) appears. We therefore group these factors according to the equivalence relation $H_{xy} \sim H_{uv}$ if $\{x, y\} = \{u, v\}$ as (unordered) sets. Since $\mathbb{E}H_{xy} = 0$, every block of the associated partition must contain at least two elements; otherwise the corresponding term is zero. If H were Gaussian, then by Wick's theorem only partitions with blocks of size exactly two (i.e. *pairings*) would contribute. Since H is not Gaussian, we have to do deal with blocks of arbitrary size; nevertheless, the pairings yield the main contribution.

In order to streamline the presentation and focus on the main ideas of the proof, in the current paper we do not deal with certain errors resulting from partitions that contain a block of size greater than two (i.e. that are not pairings), and from some exceptional coincidences among summation indices. Ignoring these issues results in three simplifications, denoted by **(S1)**–**(S3)** below, to the argument. Throughout the proof we use the letter $\mathcal{E}$ to denote any error term arising from these simplifications. In the companion paper [11], we show that the error terms $\mathcal{E}$ are indeed negligible; see Proposition 4.23 below. The proof of Proposition 4.23 is presented in a separate paper, as it requires a different argument from the one in the current paper.

In order clarify our main argument, it is actually helpful to generalize the assumptions on the matrix of variances S. This more general setup is also used in the generalized band matrix model analysed in [11] and outlined in Sect. 2.5. Instead of (2.2), we set

$$S_{xy} := \frac{1}{M-1} f\left(\frac{[x-y]_L}{W}\right), \qquad M := \sum_{x \in \mathbb{T}} f\left(\frac{x}{W}\right), \tag{4.1}$$

where $[x]_L$ denotes the representative of $x \in \mathbb{Z}^d$ in $\mathbb{T}$, and $f : \mathbb{R}^d \to \mathbb{R}$ is an even, bounded, nonnegative, piecewise[5] C^1 function, such that f and $|\nabla f|$ are integrable. We also assume that

$$\int \mathrm{d}x \, f(x) \, |x|^{4+c} \; < \; \infty \tag{4.2}$$

for some $c > 0$.

Note that M and W satisfy (2.3). We introduce the covariance matrices of S (see also (3.25)) and f, defined through

$$D_{ij} := \frac{1}{2} \sum_{x \in \mathbb{T}} \frac{x_i x_j}{W^2} S_{x0}, \qquad (D_0)_{ij} := \frac{1}{2} \int_{\mathbb{R}^d} x_i x_j f(x) \, \mathrm{d}x. \tag{4.3}$$

[5] We say that f is *piecewise* C^1 if there exists a finite collection of disjoint open sets $U_1, \ldots, U_n$ with piecewise C^1 boundaries, whose closures cover $\mathbb{R}^d$, such that f is C^1 on each U_i.

It is easy to see that $D = D_0 + O(W^{-1})$. We always assume that

$$c \leqslant D_0 \leqslant C \tag{4.4}$$

in the sense of quadratic forms, for some positive constants c and C. Note that, since (4.4) holds for D_0, it also holds for D for large enough W. In the case (2.2) we have the explicit diagonal form

$$D_0 = \frac{1}{2(d+2)} I_d. \tag{4.5}$$

In addition, for $d = 2$ we introduce the quantities

$$Q := \frac{1}{32} \sum_{x \in \mathbb{T}} S_{x0} \left| D^{-1/2} \frac{x}{W} \right|^4, \qquad Q_0 := \frac{1}{32} \int_{\mathbb{R}^2} \left| D_0^{-1/2} x \right|^4 f(x)\, \mathrm{d}x, \tag{4.6}$$

which also depend on the fourth moments of S and f respectively. (Here $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^2$.) As above, it is easy to see that $Q = Q_0 + O(W^{-1})$. In the case (2.2) we have the explicit form

$$Q_0 = \frac{2}{3}. \tag{4.7}$$

The main result of this section is summarized in the following Proposition 4.1, which establishes the leading asymptotics of $\widetilde{F}^{\eta}(E_1, E_2)$, defined in (3.18), for small $\omega = E_2 - E_1$. Once Proposition 4.1 is proved, our main results, Theorems 2.2–2.4 will follow easily (see Sect. 4.7). Recall the definition of R_2 from (2.13).

Proposition 4.1. *Suppose that the assumptions of the first paragraph of Theorem 2.2 as well as the assumptions on H made in (4.1)–(4.4) hold. Then there is a constant $c_0 > 0$ such that, for any E_1, E_2 satisfying (2.9) for small enough $c_* > 0$, we have*

$$\widetilde{F}^{\eta}(E_1, E_2) = \mathcal{V}_{\mathrm{main}} + \frac{N}{M} O\big(M^{-c_0} R_2(\omega + \eta)\big),$$

where the leading contribution $\mathcal{V}_{\mathrm{main}} \equiv (V_{\mathrm{main}})^{\eta}_{\phi_1,\phi_2}(E_1, E_2)$ satisfies the following estimates.

(i) *Suppose that (2.18) holds. Then for $d = 1, 2, 3$ we have*

$$\mathcal{V}_{\mathrm{main}} = \frac{(2/\pi)^{d/2}}{\nu(E)^2 \sqrt{\det D}} \left(\frac{L}{2\pi W}\right)^d \left(\frac{\omega}{\nu(E)}\right)^{d/2-2} \Big(K_d + O\big(\omega^{1/2} + M^{-\tau/2}\big)\Big) \tag{4.8}$$

where K_d was defined in (2.17). Moreover, for $d = 4$ we have

$$\mathcal{V}_{\mathrm{main}} = \frac{8}{\nu(E)^2 \sqrt{\det D}} \left(\frac{L}{2\pi W}\right)^d \big(|\log \omega| + O(1)\big). \tag{4.9}$$

(ii) *Suppose that* (2.18) *holds and that* $d = 2$. *If* ϕ_1 *and* ϕ_2 *satisfy* ***(C1)*** *then*

$$\mathcal{V}_{\mathrm{main}} = \frac{8}{\pi \nu(E)^2 \sqrt{\det D}} \left(\frac{L}{2\pi W} \right)^2 \left(\frac{\pi \eta \nu(E)}{\omega^2 + 4\eta^2} + (Q-1)|\log \omega| + O(1) \right), \tag{4.10}$$

and if ϕ_1 *and* ϕ_2 *satisfy* ***(C2)*** *then*

$$\mathcal{V}_{\mathrm{main}} = \frac{8}{\pi \nu(E)^2 \sqrt{\det D}} \left(\frac{L}{2\pi W} \right)^2 \big((Q-1)|\log \omega| + O(1) \big). \tag{4.11}$$

(iii) *Suppose that* $\omega = 0$. *Then the exponent* μ *from Proposition* 3.3 *may be chosen so that there exists an exponent* $c_1 > 0$ *such that for* $d = 1, 2, 3$ *we have*

$$\mathcal{V}_{\mathrm{main}} = \frac{2^{d/2}}{\nu(E)^2 \sqrt{\det D}} \left(\frac{L}{2\pi W} \right)^d \left(\frac{\eta}{\nu(E)} \right)^{d/2-2} \big(V_d(\phi_1, \phi_2) + O(M^{-c_1}) \big) \tag{4.12}$$

and for $d = 4$ *we have*

$$\mathcal{V}_{\mathrm{main}} = \frac{4}{\nu(E)^2 \sqrt{\det D}} \left(\frac{L}{2\pi W} \right)^4 \big(V_4(\phi_1, \phi_2) |\log \eta| + O(1) \big). \tag{4.13}$$

The rest of this section is devoted to the proof of Proposition 4.1.

4.1. Introduction of graphs. In order to express the nonbacktracking powers of H in terms of the entries of H, it is convenient to index the two multiple summations arising from (3.1) when plugged into (3.18) using a graph. We note that a similar graphical language was developed in [13], and many of basic definitions from Sects. 4.1 and 4.2 (such as bridges, ladders, and skeletons) are similar to those from [13]. We introduce a directed graph $\mathcal{C}(n_1, n_2) := \mathcal{C}_1(n_1) \sqcup \mathcal{C}_2(n_2)$ defined as the disjoint union of a directed chain $\mathcal{C}_1(n_1)$ with n_1 edges and a directed chain $\mathcal{C}_2(n_2)$ with n_2 edges. Throughout the following, to simplify notation we often omit the arguments n_1 and n_2 from the graphs $\mathcal{C}$, $\mathcal{C}_1$, and $\mathcal{C}_2$. For an edge $e \in E(\mathcal{C})$, we denote by $a(e)$ and $b(e)$ the initial and final vertices of e. Similarly, we denote by $a(\mathcal{C}_i)$ and $b(\mathcal{C}_i)$ the initial and final vertices of the chain $\mathcal{C}_i$. We call vertices of degree two *black* and vertices of degree one *white*. See Fig. 1 for an illustration of $\mathcal{C}$ and for the convention of the orientation.

We assign a label $x_i \in \mathbb{T}$ to each vertex $i \in V(\mathcal{C})$, and write $\mathbf{x} = (x_i)_{i \in V(\mathcal{C})}$. For an edge $e \in E(\mathcal{C})$ define the associated pairs of ordered and unordered labels

$$x_e := (x_{a(e)}, x_{b(e)}), \qquad [x_e] := \{x_{a(e)}, x_{b(e)}\}.$$

Using the graph $\mathcal{C} = \mathcal{C}(n_1, n_2)$ we may now write the covariance

$$\begin{aligned} \big\langle \operatorname{Tr} H^{(n_1)}; \operatorname{Tr} H^{(n_2)} \big\rangle &= \mathbb{E}\big[\big(\operatorname{Tr} H^{(n_1)} \big) \big(\operatorname{Tr} H^{(n_2)} \big) \big] - \mathbb{E}\big(\operatorname{Tr} H^{(n_1)} \big) \mathbb{E}\big(\operatorname{Tr} H^{(n_2)} \big) \\ &= \sum_{\mathbf{x} \in \mathbb{T}^{V(\mathcal{C})}} I(\mathbf{x}) A(\mathbf{x}), \end{aligned} \tag{4.14}$$

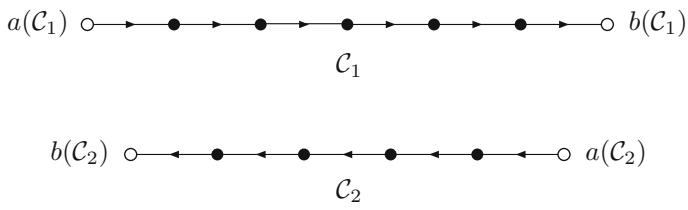

Fig. 1. The graph $\mathcal{C} = \mathcal{C}_1 \sqcup \mathcal{C}_2$. Here we chose $n_1 = 6$ and $n_2 = 5$. We indicate the orientation of the chains $\mathcal{C}_1$ and $\mathcal{C}_2$ using *arrows*. In subsequent pictures, we systematically drop the *arrows* to avoid clutter, but we consistently use this orientation when drawing graphs

where we introduced

$$A(\mathbf{x}) := \mathbb{E}\Big(\prod_{e\in E(\mathcal{C})} H_{x_e}\Big) - \mathbb{E}\Big(\prod_{e\in E(\mathcal{C}_1)} H_{x_e}\Big)\mathbb{E}\Big(\prod_{e\in E(\mathcal{C}_2)} H_{x_e}\Big) \tag{4.15}$$

and the indicator function

$$I(\mathbf{x}) := I_0(\mathbf{x}) \prod_{\substack{i,j\in V(\mathcal{C}):\\ \mathrm{dist}(i,j)=2}} \mathbf{1}(x_i \neq x_j), \quad I_0(\mathbf{x}) := \mathbf{1}(x_{a(\mathcal{C}_1)} = x_{b(\mathcal{C}_1)})\mathbf{1}(x_{a(\mathcal{C}_2)} = x_{b(\mathcal{C}_2)}). \tag{4.16}$$

The indicator function $I_0(\mathbf{x})$ implements the fact that the final and initial vertices of each chain have the same label, while $I(\mathbf{x})$ in addition implements the nonbacktracking condition. When drawing $\mathcal{C}$ as in Fig. 1, we draw vertices of $\mathcal{C}$ with degree two using black dots, and vertices of $\mathcal{C}$ with degree one using white dots. The use of two different colours also reminds us that each black vertex i gives rise to a nonbacktracking condition in $I(\mathbf{x})$, constraining the labels of the two neighbours of i to be distinct.

In order to compute the expectation in (4.15), we decompose the label configurations $\mathbf{x}$ according to partitions of $E(\mathcal{C})$.

Definition 4.2. We denote by $\mathfrak{P}(U)$ for the set of partitions of a set U and by $\mathfrak{M}(U) \subset \mathfrak{P}(U)$ the set of pairings (or matchings) of U. (In the applications below the set U will be either $E(\mathcal{C})$ or $V(\mathcal{C})$.) We call blocks of a pairing *bridges*. Moreover, for a label configuration $\mathbf{x} \in \mathbb{T}^{V(\mathcal{C})}$ we define the partition $P(\mathbf{x}) \in \mathfrak{P}(E(\mathcal{C}))$ as the partition of $E(\mathcal{C})$ generated by the equivalence relation $e \sim e'$ if and only if $[x_e] = [x_{e'}]$.

Hence we may write

$$\sum_{\mathbf{x}} I(\mathbf{x})A(\mathbf{x}) = \sum_{\Pi\in\mathfrak{P}(E(\mathcal{C}))} \sum_{\mathbf{x}} \mathbf{1}(P(\mathbf{x}) = \Pi) I(\mathbf{x})A(\mathbf{x}). \tag{4.17}$$

At this stage we introduce our first simplification.

(S1) We only keep the pairings $\Pi \in \mathfrak{M}(E(\mathcal{C}))$ in the summation (4.17).

Using Simplification **(S1)**, we write

$$\sum_{\mathbf{x}} I(\mathbf{x})A(\mathbf{x}) = \sum_{\Pi\in\mathfrak{M}(E(\mathcal{C}))} \sum_{\mathbf{x}} \mathbf{1}(P(\mathbf{x}) = \Pi) I(\mathbf{x})A(\mathbf{x}) + \mathcal{E}. \tag{4.18}$$

Here, as explained at the beginning of this section, we use the symbol $\mathcal{E}$ to denote an error term that arises from any simplification that we make. All such error terms are in fact negligible, as recorded in Proposition 4.23 below, and proved in the companion paper [11]. We use the symbol $\mathcal{E}$ without further comment throughout the following to denote such error terms arising from any of our simplifications **(S1)**–**(S3)**.

Fix $\Pi \in \mathfrak{M}(E(\mathcal{C}))$. In order to analyse the term resulting from the first term of (4.15), we write

$$\mathbf{1}(P(\mathbf{x}) = \Pi)\,\mathbb{E}\Big(\prod_{e\in E(\mathcal{C})} H_{x_e}\Big) = \mathbf{1}(P(\mathbf{x}) = \Pi)\Big(\prod_{\{e,e'\}\in\Pi} \mathbf{1}(x_e \neq x_{e'})S_{x_e}\Big), \tag{4.19}$$

where we used that H_{x_e} and $H_{x_{e'}}$ are independent if $[x_e] \neq [x_{e'}]$ as well as $\mathbb{E}H_{xy}^2 = 0$ and $\mathbb{E}H_{xy}H_{yx} = S_{xy}$. Note that the indicator function $\mathbf{1}(P(\mathbf{x}) = \Pi)$ imposes precisely two things: first, if e and e' belong to the same bridge of Π then $[x_e] = [x_{e'}]$ and, second, if e and e' belong to different bridges of Π then $[x_e] \neq [x_{e'}]$. The second simplification that we make neglects the second restriction, hence eliminating interactions between the labels associated with different bridges.

(S2) After taking the expectation, we replace the indicator function $\mathbf{1}(P(\mathbf{x}) = \Pi)$ with the larger indicator function $\prod_{\{e,e'\}\in\Pi} \mathbf{1}([x_e] = [x_{e'}])$.

Thus we have

$$\mathbf{1}(P(\mathbf{x}) = \Pi)\,\mathbb{E}\Big(\prod_{e\in E(\mathcal{C})} H_{x_e}\Big) = \Big(\prod_{\{e,e'\}\in\Pi} \mathbf{1}([x_e] = [x_{e'}])\mathbf{1}(x_e \neq x_{e'})S_{x_e}\Big) + \mathcal{E}. \tag{4.20}$$

A similar analysis may be used for the term resulting from the second term of (4.15) to get

$$\begin{aligned}\mathbf{1}(P(\mathbf{x}) = \Pi)\,\mathbb{E}\Big(\prod_{e\in E(\mathcal{C}_1)} H_{x_e}\Big)\mathbb{E}\Big(\prod_{e\in E(\mathcal{C}_2)} H_{x_e}\Big)\Big(\prod_{\pi\in\Pi} \mathbf{1}\big(|\pi \cap E(\mathcal{C}_1)| \neq 1\big)\Big)\\ = \Big(\prod_{\{e,e'\}\in\Pi} \mathbf{1}([x_e] = [x_{e'}])\mathbf{1}(x_e \neq x_{e'})S_{x_e}\Big) + \mathcal{E},\end{aligned} \tag{4.21}$$

where we used that if any bridge $\pi \in \Pi$ intersects both $E(\mathcal{C}_1)$ and $E(\mathcal{C}_2)$ then the left-hand side vanishes since $\mathbb{E}H_{xy} = 0$.

Next, we note that

$$\mathbf{1}([x_e] = [x_{e'}])\mathbf{1}(x_e \neq x_{e'}) = \mathbf{1}(x_{a(e)} = x_{b(e')})\mathbf{1}(x_{a(e')} = x_{b(e)}) \;=:\; J_{\{e,e'\}}(\mathbf{x}). \tag{4.22}$$

Plugging (4.20) and (4.21) back into (4.18) therefore yields

$$\big\langle \operatorname{Tr} H^{(n_1)}\,;\operatorname{Tr} H^{(n_2)}\big\rangle = \sum_{\Pi\in\mathfrak{M}_c(E(\mathcal{C}))}\sum_{\mathbf{x}} I(\mathbf{x})\Big(\prod_{\{e,e'\}\in\Pi} J_{\{e,e'\}}(\mathbf{x})\,S_{x_e}\Big) + \mathcal{E}, \tag{4.23}$$

where we introduced the *subset of connected pairings* of $E(\mathcal{C})$

$$\begin{aligned}\mathfrak{M}_c(E(\mathcal{C})) := \{\Pi \in \mathfrak{M}(E(\mathcal{C})) : \text{there is a } \pi \in \Pi\\ \text{such that } \pi \cap E(\mathcal{C}_1) \neq \emptyset \text{ and } \pi \cap E(\mathcal{C}_2) \neq \emptyset\}.\end{aligned} \tag{4.24}$$

The formula (4.23) provides the desired expansion in terms of pairings.

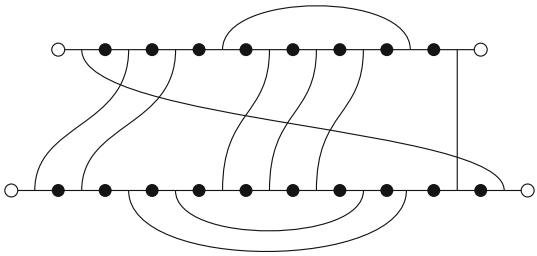

Fig. 2. A pairing of edges

A pairing Π may be conveniently represented graphically by drawing a line (or *bridge*) joining the edges e and e' whenever $\{e, e'\} \in \Pi$. See Fig. 2 for an example.

The following notations will prove helpful. We introduce the set of all connected pairings,

$$\mathfrak{M}_c := \bigsqcup_{\substack{n_1,n_2 \geqslant 0: \\ n_1+n_2 \text{ even}}} \mathfrak{M}_c\big(E(\mathcal{C}(n_1, n_2))\big).$$

Definition 4.3. With each pairing $\Pi \in \mathfrak{M}_c$ we associate its underlying graph $\mathcal{C}(\Pi)$, and regard n_1 and n_2 as functions on $\mathfrak{M}_c$ in self-explanatory notation. We also frequently abbreviate $V(\Pi) \equiv V(\mathcal{C}(\Pi))$, and refer to $V(\Pi)$ as the vertices of Π.

Next, we observe that the indicator function

$$\mathbf{1}(x_{a(\mathcal{C}_1)} = x_{b(\mathcal{C}_1)})\mathbf{1}(x_{a(\mathcal{C}_2)} = x_{b(\mathcal{C}_2)}) \prod_{\pi \in \Pi} J_\pi(\mathbf{x}) \tag{4.25}$$

in (4.23) constrains some labels of $\mathbf{x}$ to coincide. We introduce a corresponding partition $Q(\Pi) \in \mathfrak{P}(V(\Pi))$ of the vertices of Π, whereby i and j are in the same block of $Q(\Pi)$ if and only if x_i and x_j are constrained to be equal by (4.25). Equivalently, we define $Q(\Pi)$ as the finest partition of $V(\Pi)$ with the following properties.

(i) $a(e)$ and $b(e')$ belong to the same block of $Q(\Pi)$ whenever $\{e, e'\} \in \Pi$. (Note that, by symmetry, $a(e')$ and $b(e)$ also belong to the same block.)
(ii) $a(\mathcal{C}_1)$ and $b(\mathcal{C}_1)$ belong to the same block of $Q(\Pi)$.
(iii) $a(\mathcal{C}_2)$ and $b(\mathcal{C}_2)$ belong to the same block of $Q(\Pi)$.

Graphically, the first condition means that the two vertices on either side of a bridge are constrained to have the same label. See Fig. 3 for an illustration of $Q(\Pi)$. We emphasize that we constantly have to deal with two different partitions. Taking the expectation originally introduced a partition on the edges, which, after Simplification **(S1)**, is in fact a pairing. This pairing, in turn, induces constraints on the labels that are assigned to vertices; more precisely, it forces the labels of certain vertices to coincide. Together with the coincidence of the first and last labels on $\mathcal{C}_1$ and $\mathcal{C}_2$, imposed by taking the trace, this defines a partition on the vertices. Depending on $\mathbf{x}$ it may happen that more labels coincide than required by $Q(\Pi)$; the partition $Q(\Pi)$ encodes the minimal set of constraints. We therefore call $Q(\Pi)$ the *minimal vertex partition induced by* Π. Notice that, by construction, $Q(\Pi)$ does not depend on $\mathbf{x}$.

Next, suppose that there is a block $q \in Q(\Pi)$ that contains two vertices $i, j \in q$ such that $\mathrm{dist}(i, j) = 2$. We conclude that the contribution of Π to the right-hand side

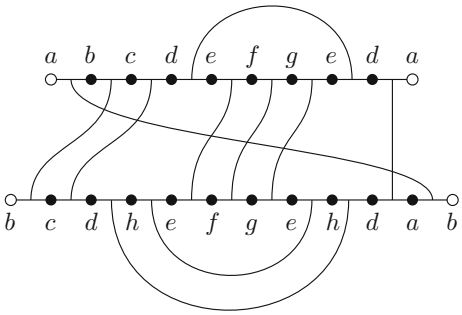

Fig. 3. The pairing Π from Fig. 2, where we in addition indicate the eight blocks of $Q(\Pi)$ by assigning a letter to each block

of (4.23) vanishes, since the indicator function $I(\mathbf{x})$ vanishes by the nonbacktracking condition $\mathbf{1}(x_i \neq x_j)$. Hence we may restrict the summation over Π in (4.23) to the subset of pairings

$$\mathfrak{R} := \big\{\Pi \in \mathfrak{M}_c : \text{if } \operatorname{dist}(i,j) = 2 \text{ then } i \text{ and } j \text{ belong to different blocks of } Q(\Pi)\big\}. \tag{4.26}$$

Lemma 4.4. *For any $\Pi \in \mathfrak{R}$, all blocks of $Q(\Pi)$ have size at least two.*

Proof. If $\{i\} \in Q(\Pi)$ then, by definition of $Q(\Pi)$, the degree of i is two and both edges incident to i belong to the same bridge. This implies that the two vertices adjacent to i belong to the same block of $Q(\Pi)$, which is impossible by definition of $\mathfrak{R}$. □

At this point we introduce our final simplification.

(S3) After restriction the summation over Π to the set $\mathfrak{R}$ in (4.23), we neglect the indicator function $I(\mathbf{x})$.

Note that the main purpose of $I(\mathbf{x})$ was to restrict the summation over pairings Π to the set $\mathfrak{R}$, which is still taken into account if one assumes **(S3)**. The presence of $I(\mathbf{x})$ in (4.23) simply results in some additional error terms $\mathcal{E}$ that are ultimately negligible. Note that $I(\mathbf{x})$ also restricts the summation to labels satisfying $x_{a(C_i)} = x_{b(C_i)}$; this condition is still imposed in the definition of $\mathfrak{R}$.

Hence we get

$$\begin{aligned}\big\langle \operatorname{Tr} H^{(n_1)} ; \operatorname{Tr} H^{(n_2)} \big\rangle = \sum_{\Pi \in \mathfrak{R}} \mathbf{1}(n_1(\Pi) = n_1)\mathbf{1}(n_2(\Pi) = n_2) \\ \times \sum_{\mathbf{y} \in \mathbb{T}^{Q(\Pi)}} \sum_{\mathbf{x} \in \mathbb{T}^{V(\Pi)}} \Big(\prod_{q \in Q(\Pi)} \prod_{i \in q} \mathbf{1}(x_i = y_q) \Big) \Big(\prod_{\{e,e'\} \in \Pi} S_{x_e} \Big) + \mathcal{E},\end{aligned} \tag{4.27}$$

where we introduced a set of independent summation labels $\mathbf{y}$, indexed by the blocks of $Q(\Pi)$.

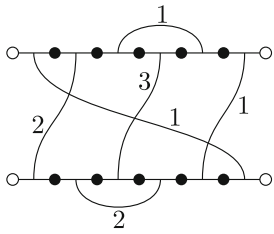

Fig. 4. The skeleton Σ of the pairing Π from Fig. 2. Next to each skeleton bridge $\sigma \in \Sigma$ we indicate the multiplicity b_σ describing how many bridges of Π were collapsed into σ

4.2. Skeletons. The summation in (3.18) is highly oscillatory, which requires a careful resummation of graphs of different order. We perform a local resummation procedure of the so-called *ladder* subdiagrams, which are subdiagrams with a pairing structure that consists only of parallel bridges. This is the second resummation procedure mentioned in Sect. 2.4. Concretely, we regroup pairings Π into families that have a similar structure, differing only in the number of parallel bridges per ladder subdiagram. Their common structure is represented by the simplest element of the family, the *skeleton*, whose ladders consist of a single bridge.

We now introduce these concepts precisely. The skeleton of a pairing $\Pi \in \mathfrak{M}_c$ is generated from Π by collapsing parallel bridges. By definition, the bridges $\{e_1, e_1'\}$ and $\{e_2, e_2'\}$ are *parallel* if $b(e_1) = a(e_2)$ and $b(e_2') = a(e_1')$. With each $\Pi \in \mathfrak{M}_c$ we associate a couple $\mathcal{S}(\Pi) = (\Sigma, \mathbf{b})$, where $\Sigma \in \mathfrak{M}_c$ has no parallel bridges, and $\mathbf{b} = (b_\sigma)_{\sigma \in \Sigma} \in \mathbb{N}^\Sigma$. The pairing Σ is obtained from Π by successively collapsing parallel bridges until no parallel bridges remain. The integer b_σ denotes the number of parallel bridges of Π that were collapsed into the bridge σ. Conversely, for any given couple $(\Sigma, \mathbf{b})$, where $\Sigma \in \mathfrak{M}_c$ has no parallel bridges and $\mathbf{b} \in \mathbb{N}^\Sigma$, we define $\Pi = \mathcal{G}(\Sigma, \mathbf{b})$ as the pairing obtained from Σ by replacing, for each $\sigma \in \Sigma$, the bridge σ with b_σ parallel bridges. Thus we have a one-to-one correspondence between pairings Π and couples $(\Sigma, \mathbf{b})$. The map $\mathcal{S}$ corresponds to the collapsing of parallel bridges of Π, and the map $\mathcal{G}$ to the "expanding" of bridges of Σ according to the multiplicities $\mathbf{b}$. Instead of burdening the reader with formal definitions of the operations $\mathcal{S}$ and $\mathcal{G}$, we refer to Figs. 2 and 4 for an illustration. When no confusion is possible, in order to streamline notation we shall omit $\mathcal{S}$ and $\mathcal{G}$ and identify Π with $(\Sigma, \mathbf{b})$. In particular, the minimal vertex partition $Q(\Pi)$ induced by $\Pi = \mathcal{G}(\Sigma, \mathbf{b})$ is denoted by $Q(\Sigma, \mathbf{b})$, and is not to be confused with $Q(\Sigma)$, the minimal vertex partition on the skeleton Σ.

Definition 4.5. Fix $\Sigma \in \mathfrak{M}_c$ and $\mathbf{b} \in \mathbb{N}^\Sigma$. As above, abbreviate $\Pi := \mathcal{G}(\Sigma, \mathbf{b})$.

(i) For $\sigma \in \Sigma$ we introduce the *ladder encoded by* σ, denoted by $L_\sigma(\Sigma, \mathbf{b}) \subset \Pi$ and defined as the set of bridges $\pi \in \Pi$ that are collapsed into the skeleton bridge σ by the operation $\mathcal{S}$. Note that $L_\sigma(\Sigma, \mathbf{b})$ consists of $|L_\sigma(\Sigma, \mathbf{b})| = b_\sigma$ parallel bridges.
(ii) We say that a vertex $i \in V(\Pi)$ *touches* the bridge $\{e, e'\} \in \Pi$ if i is incident to e or e'. We call a vertex i a *ladder vertex* of $L_\sigma(\Sigma, \mathbf{b})$ if it touches two bridges of $L_\sigma(\Sigma, \mathbf{b})$. Note that a ladder consisting of b parallel bridges gives rise to $2(b-1)$ ladder vertices.
(iii) We say that $i \in V(\Pi)$ is a *ladder vertex* of Π if it is a ladder vertex of $L_\sigma(\Sigma, \mathbf{b})$ for some $\sigma \in \Sigma$. We decompose the vertices $V(\Pi) = V_s(\Pi) \sqcup V_l(\Pi)$, where $V_l(\Pi)$ denotes the set of ladder vertices of Π.

See Fig. 5 for an illustration. Due to the nonbacktracking condition and the requirement that parallel bridges are collapsed, not every pairing can be a skeleton, and not every

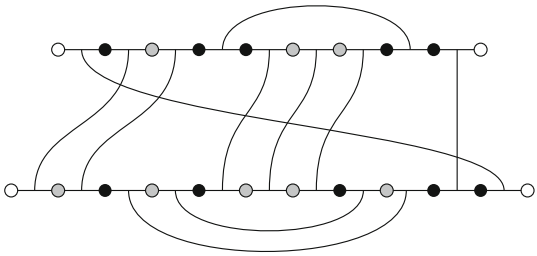

Fig. 5. The ladder vertices $V_l(\Pi)$, drawn in *grey* , of the pairing Π from Fig. 2. The vertices $V_s(\Pi)$ are drawn in *black or white*. In this example, there are $|\Sigma| = 6$ ladders

family of multiplicities is admissible; however, the few exceptions are easy to describe. The following lemma characterizes the explicit set $\mathfrak{S}$ of allowed skeletons Σ and the set of allowed multiplicities, $B(\Sigma)$, which may arise from some graph $\Pi \in \mathfrak{R} \subset \mathfrak{M}_c$.

Lemma 4.6. *For any* $\Pi \in \mathfrak{R}$ *with* $(\Sigma, \mathbf{b}) := \mathcal{S}(\Pi)$ *we have* $\Sigma \in \mathfrak{S}$*, where*

$$\mathfrak{S} := \big\{\Sigma \in \mathfrak{M}_c : \Sigma \text{ has no parallel bridges and no block of } Q(\Sigma) \text{ has size one}\big\}.$$

Moreover, defining

$$B(\Sigma) := \big\{\mathbf{b} \in \mathbb{N}^{\Sigma} : \mathcal{G}(\Sigma, \mathbf{b}) \in \mathfrak{R}\big\}, \tag{4.28}$$

for any $\Sigma \in \mathfrak{S}$*, we have that* $\mathbb{N}^{\Sigma} \setminus B(\Sigma)$ *is finite.*

Roughly, this lemma states two things. First, if a skeleton bridge $\sigma \in \Sigma$ touches two adjacent vertices of Σ that belong to the same block of $Q(\Sigma)$, then we have $b_\sigma \neq 2$. Second, if $Q(\Sigma)$ yields the label structure aba for three consecutive vertices of Σ, then $b_\sigma + b_{\sigma'} \geqslant 3$ where σ and σ' are the two bridges touching the innermost of these three vertices (in such a situation $\sigma = \sigma'$ is impossible by nonbacktracking condition implemented by $\mathfrak{R}$). See Fig. 7 below for an illustration of this latter restriction. Both of these restrictions are consequences of the nonbacktracking condition implemented in the definition of $\mathfrak{R}$.

For example, the skeleton D_4, defined in Fig. 6 below, may arise as a skeleton of some Π, so that $D_4 \in \mathfrak{S}$. Using b_1, b_2, and b_3 to denote the multiplicities of the top, bottom, and middle bridges respectively, we have $B(D_4) = \{\mathbf{b} = (b_1, b_2, b_3) : b_1, b_2, b_3 \geqslant 1,\ b_3 \neq 2\}$. Indeed, it is easy to check that the condition on the right-hand side of (4.26) is satisfied if and only if $b_2 \neq 2$.

Proof of Lemma 4.6. Let $\Pi \in \mathfrak{R}$ and $(\Sigma, \mathbf{b}) := \mathcal{S}(\Pi)$. Clearly, Σ has no parallel bridges. Moreover, if $Q(\Sigma)$ has a block of size one then Σ must have a bridge that connects two adjacent edges. Hence Π also has a bridge that connects two adjacent edges. By definition of $\mathfrak{R}$, this is impossible. This proves the first claim.

In order to prove the second claim, we simply observe that if $\Sigma \in \mathfrak{S}$ and $b_\sigma \geqslant 2$ for all $\sigma \in \Sigma$, then $\mathcal{G}(\Sigma, \mathbf{b}) \in \mathfrak{R}$. This follows easily from the definition of $\mathfrak{R}$ and the fact that the two vertices located between two parallel bridges of Π always form a block of size two in $Q(\Pi)$. □

Lemma 4.6 proves that there is a one-to-one correspondence, given by the maps $\mathcal{S}$ and $\mathcal{G}$, between pairings $\Pi \in \mathfrak{R}$ and couples $(\Sigma, \mathbf{b})$ with $\Sigma \in \mathfrak{S}$ and $\mathbf{b} \in B(\Sigma)$. Throughout

the following, we often make use of this correspondence and tacitly identify Π with $(\Sigma, \mathbf{b})$. We now use skeletons to rewrite $\widetilde{F}^{\eta}(E_1, E_2)$: from (4.27) we get

$$\begin{aligned}\widetilde{F}^{\eta}(E_1, E_2) &= \sum_{\Pi\in\mathfrak{R}} \mathbf{1}(2|\Pi| \leqslant M^{\mu}) 2\,\mathrm{Re}\big(\widetilde{\gamma}_{n_1(\Pi)}(E_1, \phi_1)\big)\, 2\,\mathrm{Re}\big(\widetilde{\gamma}_{n_2(\Pi)}(E_2, \phi_2)\big) \\ &\quad \times \sum_{\mathbf{y}\in\mathbb{T}^{Q(\Pi)}} \sum_{\mathbf{x}\in\mathbb{T}^{V(\Pi)}} \Big(\prod_{q\in Q(\Pi)} \prod_{i\in q} \mathbf{1}(x_i = y_q)\Big)\Big(\prod_{\{e,e'\}\in\Pi} S_{x_e}\Big) + \mathcal{E}, \\ &= \sum_{\Sigma\in\mathfrak{S}} \sum_{\mathbf{b}\in B(\Sigma)} \mathbf{1}\Big(2\sum_{\sigma\in\Sigma} b_{\sigma} \leqslant M^{\mu}\Big) 2\,\mathrm{Re}\big(\widetilde{\gamma}_{n_1(\Sigma,\mathbf{b})}(E_1, \phi_1)\big)\, 2\,\mathrm{Re}\big(\widetilde{\gamma}_{n_2(\Sigma,\mathbf{b})}(E_2, \phi_2)\big) \\ &\quad \times \sum_{\mathbf{y}\in\mathbb{T}^{Q(\Sigma,\mathbf{b})}} \sum_{\mathbf{x}\in\mathbb{T}^{V(\Sigma,\mathbf{b})}} \Big(\prod_{q\in Q(\Sigma,\mathbf{b})} \prod_{i\in q} \mathbf{1}(x_i = y_q)\Big)\Big(\prod_{\{e,e'\}\in\mathcal{G}(\Sigma,\mathbf{b})} S_{x_e}\Big) + \mathcal{E}. \end{aligned} \tag{4.29}$$

Next, we observe that we have a splitting

$$Q(\Pi) = Q(\Pi)|_{V_s(\Pi)} \sqcup Q(\Pi)|_{V_l(\Pi)},$$

so that the indicator function in (4.29) factors into an indicator function involving only labels y_q and x_i with $q \in Q(\Pi)|_{V_s(\Pi)}$ and $i \in V_s(\Pi)$, and another indicator function involving only labels y_q and x_i with $q \in Q(\Pi)|_{V_l(\Pi)}$ and $i \in V_l(\Pi)$. Summing over the latter ("ladder") labels yields

$$\widetilde{F}^{\eta}(E_1, E_2) = \sum_{\Sigma\in\mathfrak{S}} \mathcal{V}(\Sigma) + \mathcal{E}, \tag{4.30}$$

where we defined the *value* of the skeleton $\Sigma \in \mathfrak{S}$ as

$$\begin{aligned}\mathcal{V}(\Sigma) &:= \sum_{\mathbf{b}\in B(\Sigma)} \mathbf{1}\Big(2\sum_{\sigma\in\Sigma} b_{\sigma} \leqslant M^{\mu}\Big) 2\,\mathrm{Re}\big(\widetilde{\gamma}_{n_1(\Sigma,\mathbf{b})}(E_1, \phi_1)\big)\, 2\,\mathrm{Re}\big(\widetilde{\gamma}_{n_2(\Sigma,\mathbf{b})}(E_2, \phi_2)\big) \\ &\quad \times \sum_{\mathbf{y}\in\mathbb{T}^{Q(\Sigma)}} \sum_{\mathbf{x}\in\mathbb{T}^{V(\Sigma)}} \Big(\prod_{q\in Q(\Sigma)} \prod_{i\in q} \mathbf{1}(x_i = y_q)\Big)\Big(\prod_{\{e,e'\}\in\Sigma} (S^{b_{\{e,e'\}}})_{x_e}\Big). \end{aligned} \tag{4.31}$$

Here we recall Definition 4.3 for the meaning of the vertex set $V(\Sigma)$. The entry $(S^{b_{\{e,e'\}}})_{x_e}$ arises from summing out the $b_{\{e,e'\}} - 1$ independent labels associated with the ladder vertices of $L_{\{e,e'\}}(\Sigma, \mathbf{b})$, according to

$$\sum_{x_1,\ldots,x_{b-1}} S_{x_0x_1} S_{x_1x_2} \cdots S_{x_{b-1}x_b} = (S^b)_{x_0x_b}. \tag{4.32}$$

The labels $\mathbf{x} \in \mathbb{T}^{V(\Sigma)}$ in (4.31) are not free; we use them for notational convenience. They are a function $\mathbf{x} = \mathbf{x}(\mathbf{y})$ of the independent labels $\mathbf{y} \in \mathbb{T}^{Q(\Sigma)}$. The function $\mathbf{x}(\mathbf{y})$ is defined by the indicator function in the second parentheses on the second line of (4.31), i.e. $x_i(\mathbf{y}) := y_q$ where $q \ni i$.

In summary, we have proved that $\widetilde{F}^{\eta}(E_1, E_2)$ can be written as a sum of contributions of skeleton graphs (up to errors $\mathcal{E}$ that will prove to be negligible). The value of each skeleton is computed by assigning a positive power b_e of S to each bridge of Σ, and summing up all powers b_e and all labels that are compatible with Σ (in the sense that the vertices touching a bridge, on the same side of the bridge, must have identical labels).

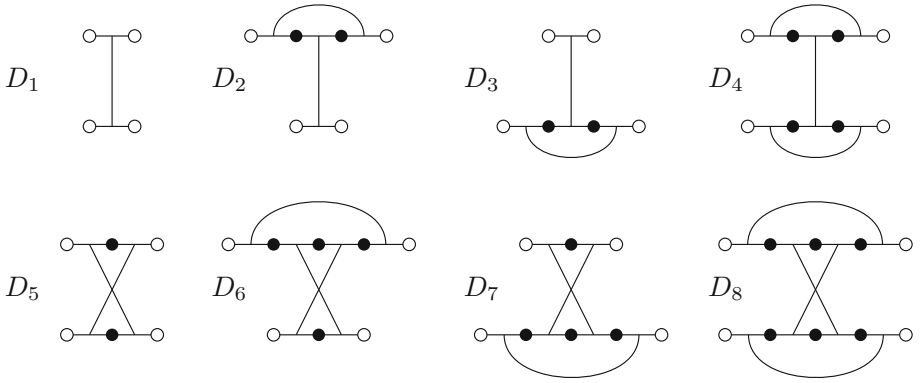

Fig. 6. The eight dumbbell skeletons $D_1, \dots, D_8$

4.3. The leading term. We now compute the leading contribution to (4.30). As it turns out, it arises from a family of eight skeleton pairings, which we call *dumbbell* skeletons. They are defined in Fig. 6. We denote by D_i the i-th dumbbell skeleton, where $i = 1, \dots, 8$. At this point in the argument, it is not apparent why precisely these eight skeletons yield the leading contribution. In fact, our analysis will reveal the graph-theoretic properties that single them out as the leading skeletons; see Sect. 4.5 below for the details.

We now *define* $\mathcal{V}_{\mathrm{main}}$ as the contribution of the dumbbell skeletons:

$$\mathcal{V}_{\mathrm{main}} := \sum_{i=1}^{8} \mathcal{V}(D_i). \tag{4.33}$$

Proposition 4.7 (Dumbbell skeletons). *Under the assumptions of Proposition 4.1, the contribution of the dumbbell skeletons defined in* (4.33) *satisfies (i), and (ii), and (iii) of Proposition 4.1.*

Proof. See [11, Propositions 3.4 and 3.7]. □

While the proof of Proposition 4.7 is given in the companion paper [11], here we explain how to obtain the (approximate) expression (3.22) from the definition (4.33). The main work, performed in [11, Sections 3.3 and 3.4], is the asymptotic analysis of the right-hand side of (3.22), which was outlined in Sect. 3.2.

We first focus on the most important skeleton, D_8. See Fig. 7 for our choice of labelling the vertex labels and the multiplicities of the bridges of D_8.

In particular, $Q(D_8)$ consists of four blocks, which are assigned the independent summation vertices $x_1, \dots, x_4$. From (4.31) we get

$$\begin{aligned}\mathcal{V}(D_8) = \sum_{b_1,b_2,b_3,b_4 \geqslant 1} & \mathbf{1}(b_3 + b_4 \geqslant 3)\mathbf{1}\big(b_1 + b_2 + b_3 + b_4 \leqslant M^{\mu}/2\big) \\ & \times 2\,\mathrm{Re}\big(\widetilde{\gamma}_{2b_1+b_3+b_4}(E_1, \phi_1)\big)\, 2\,\mathrm{Re}\big(\widetilde{\gamma}_{2b_2+b_3+b_4}(E_2, \phi_2)\big) \\ & \times \sum_{y_1,y_2,y_3,y_4 \in \mathbb{T}} (S^{b_1})_{y_1 y_3} (S^{b_2})_{y_2 y_4} (S^{b_3})_{y_3 y_4} (S^{b_4})_{y_3 y_4}.\end{aligned}$$

Here we used that $B(D_8) = \{(b_1, b_2, b_3, b_4) : b_1, b_2, b_3, b_4 \geqslant 1,\ b_3 + b_4 \geqslant 3\}$, as may be easily checked from the definition of $\mathfrak{R}$. Similarly, we may compute $\mathcal{V}(D_i)$ for

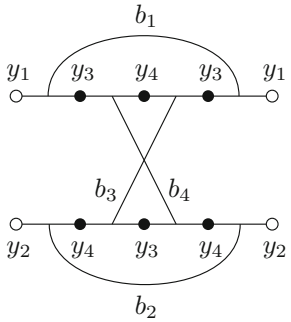

Fig. 7. The skeleton D_8. We indicate the independent labels $y_1, \dots, y_4$ next to their associated vertices, and the multiplicities $b_1, \dots, b_4$ next to their associated bridges of D_8

$i = 1, \dots, 7$; it is not hard to see that all of them arise from the expression for $\mathcal{V}(D_8)$ by setting b_1, b_2, or b_4 to be zero; setting a multiplicity b_i to be zero amounts to removing the corresponding bridge from the skeleton. Since the skeleton has to be connected, b_3 and b_4 cannot both be zero and we choose to assign b_3 to the bridge with nonzero multiplicity. The eight combinations generated by $b_1 = 0$ or $b_1 \neq 0, b_2 = 0$ or $b_2 \neq 0, b_4 = 0$ or $b_4 \neq 0$ correspond precisely to the eight graphs $D_1, \dots, D_8$ (the case $b_3 \geqslant 1$, $b_4 = 0$ corresponds to the first four graphs, $D_1, \dots, D_4$, while $b_3, b_4 \geqslant 1$ corresponds to $D_5, \dots, D_8$). Moreover, recalling (4.1), we can perform the sum over $y_1, \dots, y_4$:

$$\sum_{y_1, y_2, y_3, y_4 \in \mathbb{T}} (S^{b_1})_{y_1 y_3} (S^{b_2})_{y_2 y_4} (S^{b_3})_{y_3 y_4} (S^{b_4})_{y_3 y_4} = \mathcal{I}^{b_1+b_2} \operatorname{Tr} S^{b_3+b_4},$$

where we defined

$$\mathcal{I} \equiv \mathcal{I}_M := \frac{M}{M-1}. \tag{4.34}$$

The choice of the symbol $\mathcal{I}$ suggests that for most purposes $\mathcal{I}$ should be thought of as 1. Putting everything together, we find

$$\begin{aligned} \mathcal{V}_{\text{main}} = \sum_{b_1, b_2 = 0}^{\infty} \sum_{(b_3, b_4) \in \mathcal{A}} & \mathbf{1}\big(b_1 + b_2 + b_3 + b_4 \leqslant M^{\mu}/2\big) \\ & \times 2 \operatorname{Re}\big(\widetilde{\gamma}_{2b_1+b_3+b_4}(E_1, \phi_1)\big)\, 2 \operatorname{Re}\big(\widetilde{\gamma}_{2b_2+b_3+b_4}(E_2, \phi_2)\big)\, \mathcal{I}^{b_1+b_2} \operatorname{Tr} S^{b_3+b_4}, \end{aligned} \tag{4.35}$$

where

$$\mathcal{A} := \big(\{1, 2, \dots\} \times \{0, 1, \dots\}\big) \backslash \big\{(2, 0), (1, 1)\big\}. \tag{4.36}$$

Note that here we exclude the two cases where $b_3 + b_4 = 2$, since in those cases it may be easily checked that $Q(\mathcal{G}(\Sigma, \mathbf{b}))$ violates the defining condition of $\mathfrak{R}$. In all other cases, this condition is satisfied.

Next, we use (3.21) to decouple the upper bound in the summations over b_1, b_2, b_3, and b_4. Using (3.21) we easily find

$$\begin{aligned} \mathcal{V}_{\text{main}} = \sum_{b_1, b_2 = 0}^{\infty} \sum_{(b_3, b_4) \in \mathcal{A}} & 2 \operatorname{Re}\big(\widetilde{\gamma}_{2b_1+b_3+b_4}(E_1, \phi_1)\big)\, 2 \operatorname{Re}\big(\widetilde{\gamma}_{2b_2+b_3+b_4}(E_2, \phi_2)\big) \\ & \times \mathcal{I}^{b_1+b_2} \operatorname{Tr} S^{b_3+b_4} + O_q(NM^{-q}). \end{aligned}$$

Similarly, using (3.19) to replace $\widetilde{\gamma}$ by γ, we get

$$\mathcal{V}_{\mathrm{main}} = \sum_{b_1,b_2=0}^{\infty} \sum_{(b_3,b_4)\in\mathcal{A}} 2\,\mathrm{Re}\big(\gamma_{2b_1+b_3+b_4} * \psi_1^{\eta}\big)(E_1)\, 2\,\mathrm{Re}\big(\gamma_{2b_2+b_3+b_4} * \psi_2^{\eta}\big) \times (E_2)\, \mathcal{I}^{b_1+b_2}\,\mathrm{Tr}\, S^{b_3+b_4} + O_q(NM^{-q}). \tag{4.37}$$

This is the precise version of (3.22). For the asymptotic analysis of the right-hand side of (4.37), see [11, Sections 3.3 and 3.4].

4.4. The error terms: large skeletons. We now focus on the essence of the proof of Proposition 4.1: the estimate of the non-dumbbell skeletons. We have to estimate the contribution to the right-hand side of (4.30) of all skeletons Σ in the set

$$\mathfrak{S}^* := \mathfrak{S}\backslash\{D_1, \dots, D_8\}. \tag{4.38}$$

It turns out that when estimating $\mathcal{V}(\Sigma)$ we are faced with two independent difficulties. First, strong oscillations in the **b**-summations in the definition of $\mathcal{V}(\Sigma)$ (4.31) give rise to cancellations which have to be exploited carefully. Second, due to the combinatorial complexity of the skeletons, the size of $\mathfrak{S}^*$ grows exponentially with M, which means that we have to deal with combinatorial estimates. It turns out that these two difficulties may be effectively decoupled: if $|\Sigma|$ is small then only the first difficulty matters, and if $|\Sigma|$ is large then only the second one matters. The sets of small and large skeletons are defined as

$$\mathfrak{S}^{\leqslant} \equiv \mathfrak{S}^{\leqslant}_K := \big\{\Sigma \in \mathfrak{S}^* : |\Sigma| \leqslant K\big\}, \qquad \mathfrak{S}^{>} \equiv \mathfrak{S}^{>}_K := \big\{\Sigma \in \mathfrak{S}^* : |\Sigma| > K\big\}, \tag{4.39}$$

where $K \in \mathbb{N}$ is a cutoff, independent of N, to be fixed later.

In this subsection, we deal with large $|\Sigma|$, i.e. we estimate $\sum_{\Sigma\in\mathfrak{S}^>} \mathcal{V}(\Sigma)$. The only input on $\widetilde{\gamma}_n(E_i, \phi_i)$ that the argument of this subsection requires is the estimate (3.21). In particular, in this subsection we deal with both cases **(C1)** and **(C2)** simultaneously.

Proposition 4.8. *For large enough K, depending on μ, we have*

$$\sum_{\Sigma\in\mathfrak{S}^{>}_K} |\mathcal{V}(\Sigma)| \leqslant C_K NM^{-2}. \tag{4.40}$$

Recall that, according to Proposition 4.1, the value of the main terms (the dumbbell skeletons) is larger than NM^{-1}. The rest of this subsection is devoted to the proof of Proposition 4.8. We begin by introducing the following construction, which we shall make use of throughout the remainder of the paper. See Fig. 8 for an illustration.

Definition 4.9. Let $\Sigma \in \mathfrak{S}$ be a skeleton pairing. We define a graph $\mathcal{Y}(\Sigma)$ on the vertex set $V(\mathcal{Y}(\Sigma)) := Q(\Sigma)$ as follows. Each bridge $\{e, e'\} \in \Sigma$ gives rise to the edge $\{q, q'\}$ of $\mathcal{Y}(\Sigma)$, where q and q' are defined as the blocks of $Q(\Sigma)$ that contain $a(e)$ and $b(e)$ respectively. (Note that, by definition of Q, we also have $a(e') \in q$ and $b(e') \in q'$). We call $\mathcal{Y}(\Sigma)$ the graph associated with Σ.

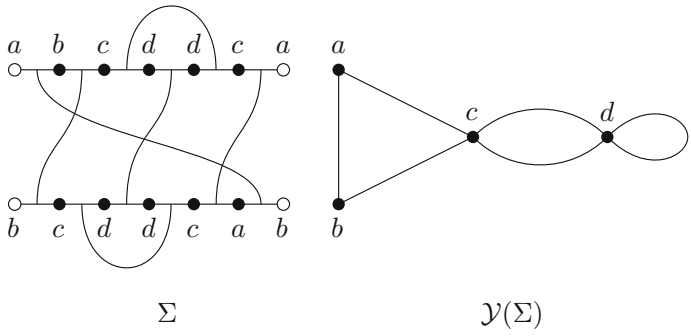

Fig. 8. A skeleton pairing Σ together with its associated graph $\mathcal{Y}(\Sigma)$. In Σ we use the letters a, b, c, d next to the vertices to indicate the four blocks of $Q(\Sigma)$. (We emphasize that the vertices of $\mathcal{Y}(\Sigma)$, unlike those of Σ, are not classified using colours; our use of *black dots* in the right-hand picture has no mathematical relevance)

Recall Definition 4.3 for the meaning of $\mathcal{C}(\Sigma)$. Then $\mathcal{Y}(\Sigma)$ is simply obtained as a minor of $\mathcal{C}(\Sigma)$ after contracting (identifying) vertices that belong to the same blocks of $Q(\Sigma)$ and replacing every pair of edges of $\mathcal{C}(\Sigma)$ forming a bridge with a single edge. In particular, the skeleton bridges of Σ become the edges of $\mathcal{Y}(\Sigma)$, i.e. Σ and $E(\mathcal{Y}(\Sigma))$ may be canonically identified. Similarly, $Q(\Sigma)$ is canonically identified with $V(\mathcal{Y}(\Sigma))$, the vertex set of the associated graph.

Lemma 4.10. *For any $\Sigma \in \mathfrak{S}$ the associated graph $\mathcal{Y}(\Sigma)$ is connected.*

Proof. This follows immediately from the definition of $\mathcal{Y}(\Sigma)$ and the fact that $\Sigma \in \mathfrak{M}_c$. □

Next, let $\Sigma \in \mathfrak{S}$ be fixed. Starting from the definition (4.31), we use (3.21) to get

$$|\mathcal{V}(\Sigma)| \leqslant C \sum_{\mathbf{b}\in\mathbb{N}^\Sigma} \mathbf{1}\Big(2\sum_{\sigma\in\Sigma} b_\sigma \leqslant M^\mu\Big) \sum_{\mathbf{y}\in\mathbb{T}^{Q(\Sigma)}} \sum_{\mathbf{x}\in\mathbb{T}^{V(\Sigma)}} \Big(\prod_{q\in Q(\Sigma)} \prod_{i\in q} \mathbf{1}(x_i = y_q)\Big) \times \Big(\prod_{\{e,e'\}\in\Sigma} (S^{b_{\{e,e'\}}})_{x_e}\Big). \tag{4.41}$$

For future reference we note that the right-hand side of (4.41) may also be written without the partition $Q(\Sigma)$ as

$$C \sum_{\mathbf{b}\in\mathbb{N}^\Sigma} \mathbf{1}\Big(2\sum_{\sigma\in\Sigma} b_\sigma \leqslant M^\mu\Big) \sum_{\mathbf{x}\in\mathbb{T}^{V(\Sigma)}} I_0(\mathbf{x}) \prod_{\{e,e'\}\in\Sigma} J_{\{e,e'\}}(\mathbf{x}) \left(S^{b_{\{e,e'\}}}\right)_{x_e}, \tag{4.42}$$

where I_0 was defined in (4.16) and $J_{\{e,e'\}}$ in (4.22). Recall that the free variables in (4.41) are $\mathbf{y}$. Using $\mathcal{Y}(\Sigma)$, we may rewrite (4.41) in the form

$$|\mathcal{V}(\Sigma)| \leqslant C \sum_{\mathbf{b}\in\mathbb{N}^{E(\mathcal{Y}(\Sigma))}} \mathbf{1}\Big(2\sum_{e\in E(\mathcal{Y}(\Sigma))} b_e \leqslant M^\mu\Big) \sum_{\mathbf{y}\in\mathbb{T}^{V(\mathcal{Y}(\Sigma))}} \Big(\prod_{e\in E(\mathcal{Y}(\Sigma))} (S^{b_e})_{y_e}\Big), \tag{4.43}$$

where we recall the convention $y_{\{q,q'\}} = (y_q, y_{q'})$.

Let

$$Q_b(\Sigma) := \big\{q \in Q(\Sigma) : q \text{ contains a black vertex of } V(\Sigma)\big\}. \tag{4.44}$$

It is easy to see that $|Q(\Sigma)\backslash Q_b(\Sigma)| \leqslant 2$. Next, we state the fundamental counting rule behind our estimates; its analogue in [13] was called the 2/3-rule. It says that each block of $Q(\Sigma)$ contains at least three vertices, with the possible exception of blocks consisting exclusively of white vertices.

Lemma 4.11 (2/3-rule). *Let $\Sigma \in \mathfrak{S}$. For all $q \in Q_b(\Sigma)$ we have $|q| \geqslant 3$. Moreover,*

$$|Q_b(\Sigma)| \leqslant \frac{2}{3}|\Sigma| + \frac{2}{3}. \tag{4.45}$$

Proof. By definition of $\mathfrak{S}$, we have that $|q| \geqslant 2$. Now suppose that $|q| = 2$. Let $i \in q$ be a black vertex of $V(\Sigma)$. Since $|q| = 2$, we conclude that the two bridges of Σ touching i (see Definition 4.5 (ii)) are parallel. This is in contradiction with the definition of $\mathfrak{S}$. Finally, (4.45) follows directly from $|q| \geqslant 3$, since

$$3|Q_b(\Sigma)| \leqslant \sum_{q \in Q_b(\Sigma)} |q| \leqslant |V(\Sigma)| = 2|\Sigma| + 2. \tag{4.46}$$

□

Since $|Q(\Sigma)| \leqslant |Q_b(\Sigma)| + 2$, we get from (4.45) that

$$|Q(\Sigma)| \leqslant \frac{2|\Sigma|}{3} + \frac{8}{3}. \tag{4.47}$$

Next, using Lemma 4.10 we choose some (immaterial) spanning tree $\mathcal{T}$ of $\mathcal{Y}(\Sigma)$. Clearly, $|E(\mathcal{T})| = |Q(\Sigma)| - 1$ and $|E(\mathcal{Y}(\Sigma))| = |\Sigma|$, so that (4.47) yields

$$\big|E(\mathcal{Y}(\Sigma))\backslash E(\mathcal{T})\big| \geqslant \frac{|\Sigma|}{3} - \frac{5}{3}. \tag{4.48}$$

We now sum over $\mathbf{y}$ in (4.43), using the estimates, valid for any $b \leqslant M^{\mu}$,

$$\sum_z (S^b)_{yz} \leqslant C, \qquad (S^b)_{yz} \leqslant \frac{C}{M}, \tag{4.49}$$

which are easy consequences of $S_{yz} \leqslant CM^{-1}$ and $\sum_z S_{yz} = \mathcal{I}$. In the product on the right-hand side of (4.43), we estimate each factor associated with $\{q, q'\} \notin E(\mathcal{T})$ by CM^{-1}, using the second estimate of (4.49). We then sum out all of the $\mathbf{y}$-labels, starting from the leaves of $\mathcal{T}$ (after some immaterial choice of root), at each summation using the first estimate of (4.49). This yields

$$\sum_{\mathbf{y} \in \mathbb{T}^{V(\mathcal{Y}(\Sigma))}} \Bigg(\prod_{e \in E(\mathcal{Y}(\Sigma))} (S^{b_e})_{y_e} \Bigg) \leqslant N \bigg(\frac{C}{M}\bigg)^{|E(\mathcal{Y}(\Sigma))\backslash E(\mathcal{T})|} \leqslant N \bigg(\frac{C}{M}\bigg)^{|\Sigma|/3 - 5/3},$$

where in the last step we used (4.48). The factor N results from the summation over the label associated with the root of $\mathcal{T}$. Thus we find from (4.43)

$$|\mathcal{V}(\Sigma)| \leqslant N\left(\frac{C}{M}\right)^{|\Sigma|/3-5/3} \sum_{\mathbf{b}\in N^{E(\mathcal{Y}(\Sigma))}} \mathbf{1}\Big(2\sum_{e\in E(\mathcal{Y}(\Sigma))} b_e \leqslant M^{\mu}\Big)$$
$$= N\left(\frac{C}{M}\right)^{|\Sigma|/3-5/3} \binom{[M^{\mu}/2]-1}{|\Sigma|-1} \leqslant N\left(\frac{C}{M}\right)^{|\Sigma|/3-5/3} \frac{M^{\mu|\Sigma|}}{(|\Sigma|-1)!}.$$

Next, for any $m \in \mathbb{N}$, a simple combinatorial argument shows that the number of skeleton pairings $\Sigma \in \mathfrak{S}$ satisfying $|\Sigma| = m$ is bounded by

$$(2m+1)\,\frac{(2m)!}{m!2^m} \leqslant C^m\, m!\,; \tag{4.50}$$

here the factor $\frac{(2m)!}{m!2^m}$ is the number of pairings of $2m$ edges, and the factor $2m+1$ is the number of graphs $\mathcal{C}$ with $2m$ edges. We therefore conclude that

$$\sum_{\Sigma\in\mathfrak{S}^{>}} |\mathcal{V}(\Sigma)| \leqslant N\sum_{m=K}^{\infty} C^m m!\left(\frac{C}{M}\right)^{m/3-5/3} \frac{M^{\mu m}}{(m-1)!} \leqslant NM^{5/3}\sum_{m=K}^{\infty}(CM^{\mu-1/3})^m$$
$$\leqslant C_K N M^{5/3+K(\mu-1/3)}.$$

Choosing K large enough completes the proof of Proposition 4.8.

We conclude this subsection by summarizing the origin of the restriction $\mu < 1/3$ (and hence $\rho < 1/3$), as it appears in the preceding proof of Proposition 4.8. The total contribution of a skeleton is determined by a competition between its *size* (given by the number of bridges) and its *entropy factor* (given by the number of independent summation labels $\mathbf{y}$). Each bridge yields, after resummation, a factor $(M\eta)^{-1}$, so that the size of the graph is $(M\eta)^{-s}$ where $s = |\Sigma|$ is the number of ladders. The entropy factor is M^{ℓ} where $\ell = |Q(\Sigma)|$ is the number independent summation labels. The 2/3-rule from Lemma 4.11 states roughly that $\ell \leqslant 2b/3$. The sum of the contributions of all skeletons is convergent if $(M\eta)^{-s}M^{\ell} \ll 1$, which, by the 2/3-rule, holds provided that $\eta \gg M^{-1/3}$.

4.5. The error terms: small skeletons. We now focus on the estimate of the small skeletons, i.e. we estimate $\mathcal{V}(\Sigma)$ for $\Sigma \in \mathfrak{S}^{\leqslant}$ (recall the splitting (4.39)). The details of the following estimates will be somewhat different for the two cases **(C1)** and **(C2)**; for definiteness, we focus on the (harder) case **(C2)**, i.e. we assume that ϕ_1 and ϕ_2 both satisfy (2.8). The analogue of the following result in the case **(C1)** is given in Proposition 4.21 at the end of this subsection.

Proposition 4.12. *Suppose that ϕ_1 and ϕ_2 satisfy (2.8). Suppose moreover that (2.9) holds for some small enough $c_* > 0$. Then for any fixed $K \in \mathbb{N}$ and small enough $\delta > 0$ in Proposition 3.3 there exists a constant $c_0 > 0$ such that*

$$\sum_{\Sigma\in\mathfrak{S}_K^{\leqslant}} |\mathcal{V}(\Sigma)| \leqslant \frac{C_K N}{M} R_2(\omega+\eta)M^{-c_0}, \tag{4.51}$$

where we recall the definition of R_2 from (2.13).

Note that, by Proposition 4.7, the size of the dumbbell skeletons is

$$|\mathcal{V}_{\mathrm{main}}| \asymp \frac{N}{M}\Big(1 + \mathbf{1}(d \leqslant 3)(\omega+\eta)^{d/2-2} + \mathbf{1}(d=4)\big|\log(\omega+\eta)\big|\Big), \tag{4.52}$$

unless $d = 2$ and $\omega \gg \eta$, in which case we have

$$|\mathcal{V}_{\mathrm{main}}| \asymp \frac{N}{M}(1 + |\log\omega|). \tag{4.53}$$

We conclude that the right-hand side of (4.51) is much smaller than the contribution of the dumbbell skeletons. In particular, the proof of Proposition 4.12 reveals why precisely the dumbbell skeletons provide the leading contributions.

In this section we give a sketch of the proof of Proposition 4.12, followed by the actual proof in the next section. As explained at the beginning of Sect. 4.4, the combinatorics of the summation over Σ are now trivial, since the cardinality of the set $\mathfrak{S}^{\leqslant} \equiv \mathfrak{S}_K^{\leqslant}$ depends only on K, which is fixed. However, the brutal estimate of (4.41), which neglects the oscillations present in the coefficients γ, is not good enough. For small skeletons, it is essential to exploit these oscillations.

First we undo the truncation in the definition of $\widetilde{\gamma}_{n_i}$ and use (3.19) to replace $\widetilde{\gamma}_{n_i}$ with $\psi_i^{\eta} * \gamma_{n_i}$ in the definition (4.31) of $\mathcal{V}(\Sigma)$. Then we rewrite the real parts in (4.31) using (3.23) this gives rise to two terms, and we focus on the first one, which we call $\mathcal{V}'(\Sigma)$. (The other one may be estimate in exactly the same way and is in fact smaller.) The summation over $\mathbf{b}$ in (4.31) can now be performed explicitly using geometric series. The result is that each skeleton bridge $\sigma \in \Sigma$ encodes an entry of the quantity $\mathcal{Z}(\sigma)$, which is roughly a resolvent of S multiplied by a phase α, i.e. $(1-\alpha S)^{-1}$. It turns out that these phases α depend strongly on the type of bridge they belong to. We split the set of skeleton bridges $\Sigma = \Sigma_d \sqcup \Sigma_c$ into the "domestic bridges" which join edges within the same component of $\mathcal{C}$ and "connecting bridges" which join edges in different components of $\mathcal{C}$; see Definition 4.13 below for more details. The critical regime is when $\alpha \approx 1$, which yields a singular resolvent $(1-\alpha S)^{-1}$ (see the discussion on the spectrum of S in Sect. 3.2). The phase α associated with a domestic bridge is separated away from 1, which yields a regular resolvent. (This may also be interpreted as strong oscillations in the geometric series of the resolvent expansion.) The phase α associated with a connecting bridge is close to 1 and the associated resolvent is therefore much more singular. More precisely (see Lemma 4.15 below), we find that these resolvents $\mathcal{Z}(\sigma)$ satisfy the bounds

$$\big|\mathcal{Z}(\sigma)_{yz}\big| \lesssim M^{-1}, \qquad \sum_z \big|\mathcal{Z}(\sigma)_{yz}\big| \lesssim 1 \tag{4.54}$$

for domestic bridges $\sigma \in \Sigma_d$ and

$$\big|\mathcal{Z}(\sigma)_{yz}\big| \lesssim M^{-1} R_2(\omega+\eta), \qquad \sum_z \big|\mathcal{Z}(\sigma)_{yz}\big| \lesssim M^{\mu}, \tag{4.55}$$

for connecting bridges $\sigma \in \Sigma_c$. (Recall the definition of R_2 from (2.13).)

Using the bounds (4.54) and (4.55) we get a simple bound on $\mathcal{V}'(\Sigma)$. The rest of the argument is purely combinatorics and power counting: we have to make sure that for any $\Sigma \in \mathfrak{S}^{\leqslant}$ this bound is small enough, i.e. $o(N/M)$. Without loss of generality we may assume that Σ does not contain a bridge that touches (see Definition 4.5) the two white vertices of the same component of $\mathcal{C}$. Indeed, if Σ contains such a bridge, we can sum up the (coinciding) labels of the two white vertices using the second bound of (4.54), which

effectively removes such a bridge, as depicted in Fig. 10 below. In particular, we have $Q_b(\Sigma) = Q(\Sigma)$. (Recall the definitions of $Q(\Sigma)$ after (4.25) and of $Q_b(\Sigma)$ from (4.44)).

We perform the summation over the labels $\mathbf{x}$ as in Sect. 4.4: by choosing a spanning tree on the graph $\mathcal{Y}(\Sigma)$. Recall that there is a canonical bijection between the edges of $\mathcal{Y}(\Sigma)$ and the bridges of Σ. Denote by Σ_t the bridges associated with the spanning tree of $\mathcal{Y}(\Sigma)$. The combinatorics rely on the following quantities:

$$\begin{aligned}
\ell &:= |Q(\Sigma)| = |V(\mathcal{Y}(\Sigma))| \;=\; \text{number of independent labels,}\\
s_d &:= |\Sigma_d| = \text{number of domestic bridges,}\\
s_t &:= |\Sigma_c \cap \Sigma_t| = \text{number of connecting tree bridges,}\\
s_l &:= |\Sigma_c \backslash \Sigma_t| = \text{number of connecting loop (i.e. non-tree) bridges.}
\end{aligned}$$

Note that the total number of bridges is $s := |\Sigma| = s_d + s_t + s_l$. Moreover, $s \geqslant \ell - 1$ since $\mathcal{Y}(\Sigma)$ is connected and $s_t \leqslant \ell - 1$ since s_t is part of a spanning tree. From the 2/3-rule in (4.45) we conclude that $|q| \geqslant 3$ for all $q \in Q(\Sigma)$ and

$$\ell \leqslant \frac{2(1+s)}{3}. \tag{4.56}$$

Using the bounds (4.54) and (4.55), we sum over the labels $\mathbf{x}$ associated with the vertices of Σ, and find the estimate

$$|\mathcal{V}'(\Sigma)| \lesssim N M^{\ell - s - 1} R_2^{s_l} M^{\mu s_t}. \tag{4.57}$$

Indeed, the root of the spanning tree gives rise to a factor N; each one of the $s - \ell + 1$ bridges not associated with the spanning tree gives rise to a factor M^{-1}; each one of the s_l connecting loop bridges gives rise to an additional factor R_2; and each one of the s_t connecting tree bridges gives rise to a factor M^{μ}.

It is instructive to compare the upper bound (4.57) for Σ being a dumbbell to the true size of the dumbbell skeletons from (4.52). Since we exclude pairings with bridges touching the two white vertices of the same component of $\mathcal{C}$, we may take Σ to be D_1 or D_5 (see Fig. 6). Of these two, D_5 saturates the 2/3-rule and is of leading order. For $\Sigma = D_5$ we have $\ell = 2$, $s = 2$, $s_l = 1$, $s_t = 1$. Hence the bound (4.57) reads

$$|\mathcal{V}'(D_5)| \lesssim \frac{N}{M} R_2 (\omega + \eta) M^{\mu}. \tag{4.58}$$

This is in general much larger than the true size (4.52); they become comparable for $\omega + \eta \asymp M^{-\mu}$ (i.e. on very small scales), which is ruled out by our assumptions on ω and η.

Now we explain how the estimate on $\mathcal{V}'(\Sigma)$ can be improved if Σ *is not a dumbbell skeleton*. We rely on two simple but fundamental observations. First, if Σ does not saturate the 2/3-rule then the right-hand side of (4.57) contains an extra power of $M^{-1/3}$ as compared to the leading term (4.58). Second, if Σ saturates the 2/3-rule and is not a dumbbell skeleton then Σ must contain a domestic bridge (joining edges within the same component of $\mathcal{C}$). Having a domestic bridge implies that $s_l + s_t \leqslant s - 1$ instead of the trivial bound $s_l + s_t \leqslant s$. This implies that the power of one of the large factors R_2 or M^{μ} on the right-hand side of (4.57) will be reduced by one; as it turns out, this is sufficient to make the right-hand side of (4.57) subleading. Note that the absence of such domestic bridges in Σ is the key feature that singles out the dumbbells among all skeletons that saturate the 2/3-rule. This explains why the leading contribution in (4.30) comes from the dumbbell skeletons.

We now explain these two scenarios more precisely. For the rest of this subsection we suppose that Σ is not a dumbbell skeleton. Hence

$$s \geqslant 3, \qquad \ell \geqslant 2, \qquad s - \ell \geqslant 1; \tag{4.59}$$

the first two estimates are immediate, and the last one follows from the first combined with (4.56) and the fact that $s - \ell \in \mathbb{N}$.

Suppose first that Σ saturates the 2/3-rule (4.56). Then $|q| = 3$ for all $q \in Q(\Sigma)$, and it is not too hard to see that Σ must contain a domestic bridge, i.e. $s_d \geqslant 1$. Roughly, this follows from the observation that in order to get a block of size three, the bridges touching the vertices of this block must be as in Fig. 11 below. Plugging (4.56) into (4.57) yields

$$\begin{aligned} |\mathcal{V}'(\Sigma)| &\lesssim \frac{N}{M} M^{2/3-s/3} R_2^{s_l} M^{\mu s_t} \leqslant \frac{N}{M} M^{2/3-s/3} R_2^{s_l+s_t-\ell+1} M^{\mu(\ell-1)} \\ &\leqslant \frac{N}{M} M^{2/3-s/3} R_2^{s-\ell} M^{\mu(\ell-1)}, \end{aligned}$$

where the second step follows from $R_2 \leqslant M^\mu$ and the third step from $s_l + s_t \leqslant s - 1$ (since $s_d \geqslant 1$). We conclude that

$$|\mathcal{V}'(\Sigma)| \lesssim \frac{N}{M} M^{1/3} (M^{-1/3} R_2)^{s-\ell} (M^{-1/3} M^\mu)^{\ell-1} \ll \frac{N}{M} R_2,$$

where we used (4.59) and $R_2 + M^\mu \leqslant M^{1/3}$.

Next, consider the case where Σ does not saturate the 2/3-rule (4.56). In this case it may well be that $s_d = 0$. However, if (4.56) is not saturated, then there must exist a $q \in Q(\Sigma)$ satisfying $|q| \geqslant 4$. Thus (4.56) improves to

$$\ell \leqslant \frac{1}{3} + \frac{2s}{3}.$$

Thus we find that

$$|\mathcal{V}'(\Sigma)| \lesssim \frac{N}{M} M^{1/3-s/3} R_2^{s_l} M^{\mu s_t}.$$

Note that we have $s_t + s_l \leqslant s$ and $s_t \leqslant \ell - 1$. Using $R_2 \leqslant M^\mu$ we therefore get

$$\begin{aligned} |\mathcal{V}'(\Sigma)| &\lesssim \frac{N}{M} M^{1/3-s/3} R_2^{s_l} M^{\mu s_t} \leqslant \frac{N}{M} M^{1/3-s/3} R_2^{s-\ell+1} M^{\mu(\ell-1)} \\ &= \frac{N}{M} M^{1/3} (M^{-1/3} R_2)^{s-\ell+1} (M^{-1/3} M^\mu)^{\ell-1}. \end{aligned}$$

From (4.59) we therefore get $|\mathcal{V}'(\Sigma)| \ll \frac{N}{M} R_2$. This concludes the sketch of the proof of Proposition 4.12.

4.6. Proof of Proposition 4.12. We begin the proof by rewriting (4.31) in a form where the oscillations in the summation over $\mathbf{b}$ may be effectively exploited. This consists of three steps, each of which results in negligible errors of order $C_q N M^{-q}$ for any $q > 0$. In the first step, we decouple the $\mathbf{b}$-summations by replacing the indicator function $\mathbf{1}\big(2\sum_{\sigma\in\Sigma} b_\sigma \leqslant M^\mu\big)$ with the product $\prod_{\sigma\in\Sigma}\mathbf{1}(b_\sigma \leqslant M^\mu)$, using the estimate (3.21). In the second step, we replace the factors $\widetilde{\gamma}_{n_i(\Sigma,\mathbf{b})}(E_1,\phi_i)$ with $(\gamma_{n_i(\Sigma,\mathbf{b})} * \psi_i^\eta)(E_i)$, using the estimate (3.19). These two steps are analogous to the steps from (4.35) to (4.37).

In the third step, we truncate in the tails of the functions ψ_i on the scale $M^{\delta/2}$, where $\delta > 0$ is the constant from Proposition 3.3. To that end, we choose a smooth, nonnegative, symmetric function χ satisfying $\chi(E) = 1$ for $|E| \leqslant 1$ and $\chi(E) = 0$ for $|E| \geqslant 2$. We split $\psi_i = \psi_i^{\leqslant} + \psi_i^{>}$, where

$$\psi_i^{\leqslant}(E) := \psi_i(E)\chi(M^{-\delta/2}E), \qquad \psi_i^{>}(E) := \psi_i(E)\big(1 - \chi(M^{-\delta/2}E)\big) \tag{4.60}$$

This yields the splitting $\psi_i^\eta = \psi_i^{\leqslant,\eta} + \psi_i^{>,\eta}$ of the rescaled test function $\psi^\eta(E) = \eta^{-1}\psi(\eta^{-1}E)$. This splitting is done on the scale $\eta M^{\delta/2}$, and we have

$$\operatorname{supp}\psi_i^{\leqslant,\eta} \subset [-2\eta M^{\delta/2}, 2\eta M^{\delta/2}]. \tag{4.61}$$

Moreover, recalling (2.8) and using the trivial bound $|\gamma_n(E)| \leqslant C$ we find

$$\big|(\psi_i^{>,\eta} * \gamma_n)(E_i)\big| \leqslant C_q M^{-q} \tag{4.62}$$

for any $q > 0$. The truncation of the third step is the replacement of $(\gamma_{n_i(\Sigma,\mathbf{b})} * \psi_i^\eta)(E_i)$ with $(\gamma_{n_i(\Sigma,\mathbf{b})} * \psi_i^{\leqslant,\eta})(E_i)$, using (4.62).

Applying these three steps to the definition (4.31) yields

$$\begin{aligned}\mathcal{V}(\Sigma) = \sum_{\mathbf{b}\in B(\Sigma)}\Big(\prod_{\sigma\in\Sigma}\mathbf{1}(b_\sigma \leqslant M^\mu)\Big)\, 2\operatorname{Re}\big(\gamma_{n_1(\Sigma,\mathbf{b})} * \psi_1^{\leqslant,\eta}\big)(E_1)\, 2\operatorname{Re}\big(\gamma_{n_2(\Sigma,\mathbf{b})} * \psi_2^{\leqslant,\eta}\big)(E_2)\\ \times \sum_{\mathbf{y}\in\mathbb{T}^{Q(\Sigma)}}\sum_{\mathbf{x}\in\mathbb{T}^{V(\Sigma)}}\Big(\prod_{q\in Q(\Sigma)}\prod_{i\in q}\mathbf{1}(x_i = y_q)\Big)\Big(\prod_{\{e,e'\}\in\Sigma}(S^{b_{\{e,e'\}}})_{x_e}\Big) + O_{q,\Sigma}(NM^{-q}).\end{aligned} \tag{4.63}$$

The errors arising from each of the three steps are estimated using (3.21), (3.19), and (4.62) respectively. The summations over $\mathbf{b}$ and $\mathbf{y}$ in the error terms are performed brutally, exactly as in the proof of Proposition 4.8 (in fact here we only need that $\mathcal{Y}(\Sigma)$ be connected); we omit the details.

Next, we use (3.23) to write $\mathcal{V}(\Sigma) = 2\operatorname{Re}(\mathcal{V}'(\Sigma) + \mathcal{V}''(\Sigma)) + O_{q,\Sigma}(NM^{-q})$, where

$$\begin{aligned}\mathcal{V}'(\Sigma) := \sum_{\mathbf{b}\in B(\Sigma)}\Big(\prod_{\sigma\in\Sigma}\mathbf{1}(b_\sigma \leqslant M^\mu)\Big)\big(\gamma_{n_1(\Sigma,\mathbf{b})} * \psi_1^{\leqslant,\eta}\big)(E_1)\,\big(\overline{\gamma_{n_2(\Sigma,\mathbf{b})}} * \psi_2^{\leqslant,\eta}\big)(E_2)\\ \times \sum_{\mathbf{y}\in\mathbb{T}^{Q(\Sigma)}}\sum_{\mathbf{x}\in\mathbb{T}^{V(\Sigma)}}\Big(\prod_{q\in Q(\Sigma)}\prod_{i\in q}\mathbf{1}(x_i = y_q)\Big)\Big(\prod_{\{e,e'\}\in\Sigma}(S^{b_{\{e,e'\}}})_{x_e}\Big),\end{aligned} \tag{4.64}$$

and $\mathcal{V}''(\Sigma)$ is defined similarly but without the complex conjugation on $\gamma_{n_2(\Sigma,\mathbf{b})}$. We shall gives the details of the estimate for the larger error term, $\mathcal{V}'(\Sigma)$. The term $\mathcal{V}''(\Sigma)$ may be estimated using an almost identical argument; we sketch the minor differences below.

In order to estimate the right-hand side of (4.63), we shall have to classify the bridges of Σ into three classes according to the following definition.

Definition 4.13. For $i = 1, 2$ we define

$$\Sigma_i := \big\{\sigma \in \Sigma : \sigma \subset E(\mathcal{C}_i)\big\},$$

the set of bridges consisting only of edges of $\mathcal{C}_i$. We abbreviate $\Sigma_d := \Sigma_1 \cup \Sigma_2$ (the set of "domestic bridges"). We also define $\Sigma_c := \Sigma \setminus \Sigma_d$, the set of bridges connecting the two components of $\mathcal{C}$. Moreover, for $\# = 1, 2, c, d$ we introduce the set $E_\#(\mathcal{Y}(\Sigma))$ defined as the subset of $E(\mathcal{Y}(\Sigma))$ encoded by $\Sigma_\#$ under the canonical identification $\Sigma \simeq E(\mathcal{Y}(\Sigma))$, according to Definition 4.9.

Since each $\sigma \in \Sigma_c$ contains one edge of $\mathcal{C}_1$ and one edge of $\mathcal{C}_2$, and each $\sigma \in \Sigma_i$ contains two edges of $\mathcal{C}_i$, we find that the number of edges in the i-th chain $\mathcal{C}_i(n_i)$ of the graph $\mathcal{C}(n_1, n_2)$ with pairing $\Pi = (\Sigma, \mathbf{b})$ is

$$n_i(\Sigma, \mathbf{b}) = \sum_{\sigma \in \Sigma_c} b_\sigma + 2 \sum_{\sigma \in \Sigma_i} b_\sigma.$$

Here we identify Π with $(\Sigma, \mathbf{b})$, as remarked after Lemma 4.6.

We may now plug into (4.64) the explicit expression for γ_n from (3.6), at which point it is convenient to introduce the abbreviations

$$T(E) := \frac{2}{1 - (M-1)^{-1} \mathrm{e}^{2\mathrm{i} \arcsin E}}, \qquad A_i := \arcsin E_i. \tag{4.65}$$

Thus we get from (4.64)

$$\begin{aligned}
\mathcal{V}'(\Sigma) = {} & \sum_{\mathbf{b} \in B(\Sigma)} \Bigg(\prod_{\sigma \in \Sigma} \mathbf{1}(b_\sigma \leqslant M^\mu) \Bigg) \\
& \times \Bigg(T(E_1) \overline{T(E_2)}\, \mathrm{e}^{\mathrm{i}(A_1 - A_2)} \prod_{\sigma \in \Sigma_1} (-\mathrm{e}^{2\mathrm{i}A_1})^{b_\sigma} \prod_{\sigma \in \Sigma_2} (-\mathrm{e}^{-2\mathrm{i}A_2})^{b_\sigma} \prod_{\sigma \in \Sigma_c} \mathrm{e}^{\mathrm{i}(A_1 - A_2) b_\sigma} \Bigg) \\
& * \psi_1^{\leqslant, \eta}(E_1) * \psi_2^{\leqslant, \eta}(E_2) \sum_{\mathbf{y} \in \mathbb{T}^{Q(\Sigma)}} \sum_{\mathbf{x} \in \mathbb{T}^{V(\Sigma)}} \Bigg(\prod_{q \in Q(\Sigma)} \prod_{i \in q} \mathbf{1}(x_i = y_q) \Bigg) \Bigg(\prod_{\{e, e'\} \in \Sigma} (S^{b_{\{e,e'\}}})_{x_e} \Bigg).
\end{aligned}$$

Here, by a slight abuse of notation, we write $(\varphi * \chi)(E) \equiv \varphi(E) * \chi(E)$. Using the associated graph $\mathcal{Y}(\Sigma)$ from Definition 4.9, we may rewrite this as

$$\begin{aligned}
\mathcal{V}'(\Sigma) = {} & \sum_{\mathbf{b} \in \{1, \dots, [M^\mu]\}^{E(\mathcal{Y}(\Sigma))}} \mathbf{1}(\mathbf{b} \in B(\Sigma)) \\
& \times \Bigg(T(E_1) \overline{T(E_2)}\, \mathrm{e}^{\mathrm{i}(A_1 - A_2)} \prod_{e \in E_1(\mathcal{Y}(\Sigma))} (-\mathrm{e}^{2\mathrm{i}A_1})^{b_e} \prod_{e \in E_2(\mathcal{Y}(\Sigma))} (-\mathrm{e}^{-2\mathrm{i}A_2})^{b_e} \prod_{e \in E_c(\mathcal{Y}(\Sigma))} \mathrm{e}^{\mathrm{i}(A_1 - A_2) b_e} \Bigg) \\
& * \psi_1^{\leqslant, \eta}(E_1) * \psi_2^{\leqslant, \eta}(E_2) \sum_{\mathbf{y} \in \mathbb{T}^{V(\mathcal{Y}(\Sigma))}} \Bigg(\prod_{e \in E(\mathcal{Y}(\Sigma))} (S^{b_e})_{y_e} \Bigg),
\end{aligned} \tag{4.66}$$

where we used the canonical identification between Σ and $E(\mathcal{Y}(\Sigma))$ to rewrite the set $B(\Sigma) \subset \mathbb{N}^\Sigma$ from Lemma 4.6 as a subset $B(\Sigma) \subset \mathbb{N}^{E(\mathcal{Y}(\Sigma))}$ (by a slight abuse of notation). Also, to avoid confusion, we emphasize that the expressions E_i and $E_i(\mathcal{Y}(\Sigma))$ have nothing to do with each other.

Next, we split $\mathcal{V}'(\Sigma) = \mathcal{V}'_0(\Sigma) - \mathcal{V}'_1(\Sigma)$ using the splitting $\mathbf{1}(\mathbf{b} \in B(\Sigma)) = 1 - \mathbf{1}(\mathbf{b} \notin B(\Sigma))$ in (4.66). We first focus on main term, $\mathcal{V}'_0(\Sigma)$. In the definition of $\mathcal{V}'_0(\Sigma)$, we may sum the geometric series associated with each summation variable b_e to get

$$\mathcal{V}'_0(\Sigma) = \sum_{\mathbf{y} \in \mathbb{T}^{V(\mathcal{Y}(\Sigma))}} \Bigg(T(E_1)\overline{T(E_2)}\, \mathrm{e}^{\mathrm{i}(A_1 - A_2)} \prod_{e \in E_1(\mathcal{Y}(\Sigma))} Z(-\mathrm{e}^{2\mathrm{i}A_1} S)_{y_e} \times \prod_{e \in E_2(\mathcal{Y}(\Sigma))} Z(-\mathrm{e}^{-2\mathrm{i}A_2} S)_{y_e} \prod_{e \in E_c(\mathcal{Y}(\Sigma))} Z(\mathrm{e}^{\mathrm{i}(A_1 - A_2)} S)_{y_e} \Bigg) * \psi_1^{\leqslant, \eta}(E_1) * \psi_2^{\leqslant, \eta}(E_2), \tag{4.67}$$

where we abbreviated

$$Z(x) := \sum_{b=1}^{[M^\mu]} x^b = \frac{x(1 - x^{[M^\mu]})}{1 - x} \tag{4.68}$$

for any quantity x, which may be a number or a matrix. The explicit summation over $\mathbf{b}$ exploits the cancellations associated with the highly oscillating summands. From now on, we shall freely estimate the summation over $\mathbf{y}$ by taking the absolute value inside the sum.

On the right-hand side of (4.67), each edge $e \in E(\mathcal{Y}(\Sigma))$ encodes a symmetric matrix of the form $Z(\alpha S)$, where $|\alpha| = 1$. In order to estimate the right-hand side of (4.67), we therefore need appropriate resolvent bounds on the entries of $Z(\alpha S)$. To that end, we improve the second bound of (4.49) using the following local decay bound. Recall the definition of $\mathcal{I}$ from (4.34).

Lemma 4.14. *For all $b \in \mathbb{N}$ we have*

$$(\mathcal{I}^{-1} S^b)_{yz} \leqslant \frac{C}{M b^{d/2}} + \frac{C}{N}$$

for some constant C depending only on f.

Proof. This follows from a standard local central limit theorem; see for instance the proof in [44, Section 3]. □

In particular, for $1 \leqslant b \leqslant (L/W)^2$ we have

$$(S^b)_{yz} \leqslant \frac{C}{M b^{d/2}}. \tag{4.69}$$

Recalling (2.11) and (2.13), we find from (4.69) that for $|\alpha| \leqslant 1$ we have

$$|Z(\alpha S)_{yz}| \leqslant \frac{C}{M} R_2(M^{-\mu}). \tag{4.70}$$

The bound (4.70) is sharp if $\alpha = 1$, i.e. if the sum in (4.68) is not oscillating. If oscillations are present, we get better bounds which we record in the following lemma. It is a special case of [11, Proposition 3.5].

Lemma 4.15. *Let S be as in (4.1) and $\alpha \in \mathbb{C}$ satisfy $|\alpha| \leqslant 1$ and $|1-\alpha| \geqslant 4/M + (W/L)^2$. There exists a constant $C > 0$, depending only on d and the profile function f, such that*

$$\left\| \frac{1}{1-\alpha S} \right\|_{\ell^\infty \to \ell^\infty} \leqslant \frac{C \log N}{2 - |1+\alpha|}. \tag{4.71}$$

Under the same assumptions we have

$$\sup_{x,y} \left| \left(\frac{S}{1-\alpha S} \right)_{xy} \right| \leqslant \frac{C}{M} R_2(|1-\alpha|), \tag{4.72}$$

where the constant C depends only on d and f.

From (4.49), (4.70), and Lemma 4.15 we get

$$\begin{aligned} |Z(\alpha S)_{yz}| &\leqslant \frac{C}{M} \min\bigl\{R_2(|1-\alpha|),\ R_2(M^{-\mu})\bigr\}, \qquad \sum_z |Z(\alpha S)_{yz}| \\ &\leqslant C \min\left\{ \frac{\log N}{2-|1+\alpha|},\ M^\mu \right\}, \end{aligned} \tag{4.73}$$

We apply (4.73) to estimating (4.67) via the following key estimate.

Lemma 4.16. *Let $v = (v_1, v_2)$ and denote by $A_{i,v} := \arcsin(E_i - v_i)$ the value of A_i in the convolution integral (4.67). For small enough $\delta > 0$ and $|v_1|, |v_2| \leqslant 2\eta M^{\delta/2}$ (i.e. v_1 and v_2 in the support of the convolution integral (4.67)) we have*

$$\left| Z(-\mathrm{e}^{\pm 2\mathrm{i}A_{i,v}} S)_{yz} \right| \leqslant \frac{C}{M}, \qquad \sum_z \left| Z(-\mathrm{e}^{\pm 2\mathrm{i}A_{i,v}} S)_{yz} \right| \leqslant C \log N \tag{4.74}$$

and

$$\left| Z(\mathrm{e}^{\mathrm{i}(A_{1,v} - A_{2,v})} S)_{yz} \right| \leqslant \frac{C}{M} M^{2\delta} R_2(\omega + \eta), \qquad \sum_z \left| Z(\mathrm{e}^{\mathrm{i}(A_{1,v} - A_{2,v})} S)_{yz} \right| \leqslant C M^\mu. \tag{4.75}$$

Proof. To prove (4.74), we set $\alpha_i = -\mathrm{e}^{\pm 2\mathrm{i}A_{i,v}}$, in which case an elementary estimate yields $2 - |1+\alpha_i| \geqslant c$. Similarly, we have $|1-\alpha_i| \geqslant c$, which yields $R_2(|1-\alpha_i|) \leqslant C$. Now (4.74) follows from (4.73) and (2.9).

To prove (4.75), we set $\alpha = \mathrm{e}^{\mathrm{i}(A_{1,v} - A_{2,v})}$. In order to estimate $Z(\alpha S)_{yz}$, we distinguish two cases according to whether $\eta \leqslant M^{-\delta}\omega$. Suppose first that $\eta \leqslant M^{-\delta}\omega$. Then we have $|1-\alpha| \asymp \omega(1 + O(\omega)) \geqslant c\omega$. We therefore find from the first inequality of (4.73) that

$$|Z(\alpha S)_{yz}| \leqslant \frac{C}{M} R_2(\omega) \leqslant \frac{C}{M} R_2(\omega + \eta),$$

where in the second step we used that $\omega + \eta \leqslant 2\omega$ and that R_2 is monotone decreasing for small enough arguments (see its definition in (2.13)). On the other hand, if $\omega < M^\delta \eta$ then we get from (4.73) that

$$\begin{aligned} |Z(\alpha S)_{yz}| &\leqslant \frac{C}{M} R_2(M^{-\mu}) \leqslant \frac{C}{M} \bigl((\omega+\eta) M^\mu\bigr)^{1/2} R_2(\omega+\eta) \\ &\leqslant \frac{C}{M} M^{(\delta+\mu-\rho)/2} R_2(\omega+\eta) \leqslant \frac{C}{M} M^{2\delta} R_2(\omega+\eta), \end{aligned}$$

where in the second step we used that $(\omega+\eta)M^{\mu} \geqslant 1$, and in the last step that $\mu-\rho < 3\delta$. Putting both cases together we get (4.75). □

Next, we plug the estimates (4.74) and (4.75) into (4.67) and sum over $\mathbf{y}$; we use (4.74) for $e \in E_i(\mathcal{Y}(\Sigma))$ with $i = 1, 2$, and (4.75) for $e \in E_c(\mathcal{Y}(\Sigma))$. We perform the summation over $\mathbf{y}$ as in Sect. 4.4: by choosing an arbitrary spanning tree $\mathcal{T}$ of $\mathcal{Y}(\Sigma)$ along with an arbitrary root of $\mathcal{T}$. In the summation over $\mathbf{y}$ on the right-hand side of (4.67), each edge $e \in E(\mathcal{Y}(\Sigma))$ encodes a matrix entry that we estimate as follows. For $e \in E_d(\mathcal{Y}(\Sigma))\setminus E(\mathcal{T})$ we use the first estimate of (4.74), for $e \in E_d(\mathcal{Y}(\Sigma)) \cap E(\mathcal{T})$ the second estimate of (4.74), for $e \in E_c(\mathcal{Y}(\Sigma))\setminus E(\mathcal{T})$ the first estimate of (4.75), and for $e \in E_c(\mathcal{Y}(\Sigma)) \cap E(\mathcal{T})$ the second estimate of (4.75). The result is

$$\begin{aligned} |\mathcal{V}_0'(\Sigma)| &\leqslant C^{|\Sigma|} N M^{-|E_d(\mathcal{Y}(\Sigma))\setminus E(\mathcal{T})|} (\log N)^{|E_d(\mathcal{Y}(\Sigma))\cap E(\mathcal{T})|} \\ &\quad \times \left(\frac{M^{2\delta} R_2(\omega+\eta)}{M}\right)^{|E_c(\mathcal{Y}(\Sigma))\setminus E(\mathcal{T})|} M^{\mu|E_c(\mathcal{Y}(\Sigma))\cap E(\mathcal{T})|} \\ &\leqslant \frac{N}{M} \frac{(\log N)^{|\Sigma|}}{M^{|\Sigma|-|Q(\Sigma)|}} \left(M^{2\delta} R_2(\omega+\eta)\right)^{|E_c(\mathcal{Y}(\Sigma))\setminus E(\mathcal{T})|} M^{\mu|E_c(\mathcal{Y}(\Sigma))\cap E(\mathcal{T})|}, \end{aligned}$$

where we used that $|E(\mathcal{Y}(\Sigma))| = |\Sigma|$ and $|E(\mathcal{T})| = |Q(\Sigma)| - 1$. As before, the factor N arises from the summation over the label of $\mathbf{y}$ associated with the root of $\mathcal{T}$.

Next, we remark that the above proof may be repeated verbatim for the other error term, $\mathcal{V}_1'(\Sigma)$. This case is in fact easier: since $\mathbb{N}^{E(\mathcal{Y}(\Sigma))}\setminus B(\Sigma)$ is a finite set (see Lemma 4.6), we do not have to exploit the cancellations from the summation over $\mathbf{b}$. Repeating the above argument for $\mathcal{V}_1'(\Sigma)$, with the right-hand sides of the corresponding estimates from (4.74) and (4.75) replaced with C/M, C, C/M, and C respectively, we find

$$|\mathcal{V}'(\Sigma)| \leqslant \mathcal{R}(\Sigma) := \frac{N}{M} \frac{M^{3\delta|\Sigma|}}{M^{|\Sigma|-|Q(\Sigma)|}} R_2(\omega+\eta)^{|E_c(\mathcal{Y}(\Sigma))\setminus E(\mathcal{T})|} M^{\mu|E_c(\mathcal{Y}(\Sigma))\cap E(\mathcal{T})|}. \tag{4.76}$$

In order to show that $\mathcal{R}(\Sigma)$ is small enough, we shall use a graph-theoretic argument to derive appropriate bounds on the exponents. It relies on the following further partition of the set Σ_d according to whether a bridge touches both endpoints (white vertices) of a chain.

Definition 4.17. We partition $\Sigma_d = \Sigma_d^0 \sqcup \Sigma_d^1$, where

$$\Sigma_d^1 := \big\{\sigma \in \Sigma_d : \sigma \text{ touches } a(\mathcal{C}_i) \text{ and } b(\mathcal{C}_i) \text{ for some } i = 1, 2\big\}.$$

We also use $E_d^0(\mathcal{Y}(\Sigma))$ and $E_d^1(\mathcal{Y}(\Sigma))$ to denote the corresponding disjoint subsets of $E_d(\mathcal{Y}(\Sigma))$.

Note that Σ_d^1 may contain at most two bridges: one only touching the white vertices of $\mathcal{C}_1$ and one only touching the white vertices of $\mathcal{C}_2$. See Fig. 9 for an illustration of these three types of bridges.

For the following counting arguments, for definiteness it will be convenient to assume that $\Sigma_d^1 = \emptyset$. Hence, we first show that skeleton pairings with $\Sigma_d^1 \neq \emptyset$ can be easily estimated by those with $\Sigma_d^1 = \emptyset$, at the expense of an unimportant factor. The following lemma states this fact precisely. Let

$$\overline{\mathfrak{S}}^{\leqslant} := \big\{\Sigma \in \mathfrak{S}^{\leqslant} : \Sigma_d^1 = \emptyset\big\}.$$

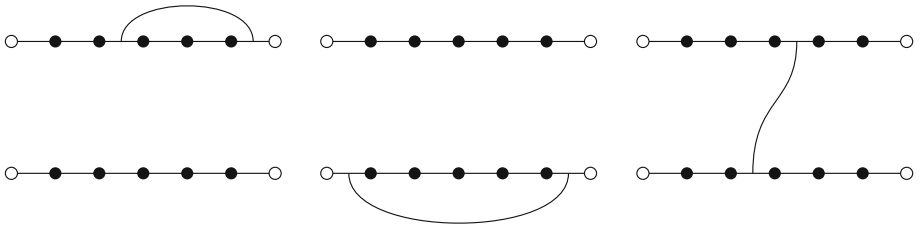

Fig. 9. A bridge in Σ_d^0 (*left*), Σ_d^1 (*centre*), and Σ_c (*right*)

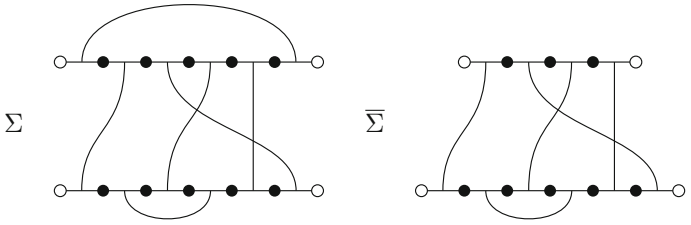

Fig. 10. The operation $\Sigma \mapsto \overline{\Sigma}$

Lemma 4.18. *For each* $\Sigma \in \mathfrak{S}^{\leqslant}$ *there exists a* $\overline{\Sigma} \in \overline{\mathfrak{S}}^{\leqslant}$ *such that and* $\mathcal{R}(\Sigma) \leqslant (\log N)^2 \mathcal{R}(\overline{\Sigma})$.

Proof. The operation $\Sigma \mapsto \overline{\Sigma}$ amounts to simply removing all bridges of Σ_d^1 from Σ. Instead of a formal definition, we refer to Fig. 10 for a graphical depiction of this operation.

By definition of $Q(\cdot)$, we find that the operation $\Sigma \mapsto \overline{\Sigma}$ amounts to removing any of the two vertices $\{a(\mathcal{C}_1), b(\mathcal{C}_1)\}$ and $\{a(\mathcal{C}_2), b(\mathcal{C}_2)\}$ that belongs to $Q(\Sigma)$. This results in a removal of the corresponding number of leaves from the spanning tree $\mathcal{T}$. (The removed bridges always correspond to leaves in $\mathcal{T}$. In particular, $|E_c(\mathcal{Y}(\Sigma)) \setminus E(\mathcal{T})|$ and $|E_c(\mathcal{Y}(\Sigma)) \cap E(\mathcal{T})|$ are remain unchanged by this removal.) Note that if $\Sigma \in \mathfrak{S}^{\leqslant}$ then $\overline{\Sigma} \in \overline{\mathfrak{S}}^{\leqslant}$, since by construction if $\Sigma \notin \{D_1, \ldots, D_8\}$ then $\overline{\Sigma} \notin \{D_1, \ldots, D_8\}$. The claim now follows easily from the bound (4.76) with argument Σ, as well as the observations that $|\Sigma| - |Q(\Sigma)| = |\overline{\Sigma}| - |Q(\overline{\Sigma})|$, that $|\Sigma_d| \leqslant |\overline{\Sigma}_d| + 2$, that $|\Sigma| \leqslant |\overline{\Sigma}| + 2$, and that the two last exponents on the right-hand side of (4.76) are the same for Σ and $\overline{\Sigma}$. □

By Lemma 4.18, it suffices to estimate $\mathcal{R}(\Sigma)$ for $\Sigma \in \overline{\mathfrak{S}}^{\leqslant}$. For $\Sigma \in \overline{\mathfrak{S}}^{\leqslant}$ we have $\Sigma_d^0 = \Sigma_d$. Moreover, if there is a bridge touching $a(\mathcal{C}_1)$ and $a(\mathcal{C}_2)$ as well as a bridge touching $b(\mathcal{C}_1)$ and $b(\mathcal{C}_2)$, we find that all four white vertices constitute a single block of $Q(\Sigma)$. Otherwise, since $\Sigma_d^1 = \emptyset$, every block of $Q(\Sigma)$ contains a black vertex, so that $Q_b(\Sigma) = Q(\Sigma)$, where $Q_b(\Sigma)$ was defined in (4.44). Either way, recalling Lemma 4.11, we conclude for $\Sigma \in \overline{\mathfrak{S}}^{\leqslant}$ and $q \in Q(\Sigma)$ that

$$|q| \geqslant 3. \tag{4.77}$$

In order to complete the estimate of (4.76), and hence the proof of Proposition 4.12, we shall have to distinguish between the case where $|q| = 3$ for all $q \in Q(\Sigma)$ and the case where there exists a $q \in Q(\Sigma)$ with $|q| > 3$.

Lemma 4.19. *Suppose that* $\Sigma \in \overline{\mathfrak{S}}^{\leqslant}$ *and* $|q| = 3$ *for all* $q \in Q(\Sigma)$. *Then*

$$\mathcal{R}(\Sigma) \leqslant \frac{N}{M} R_2(\omega + \eta) M^{3\mu - 1} M^{C\delta}. \tag{4.78}$$

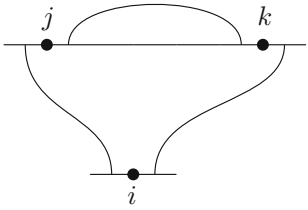

Fig. 11. A block $q = \{i, j, k\} \in Q(\Sigma)$ along with the three bridges of $\Sigma(q)$. We do not draw the other vertices or bridges

Lemma 4.20. *Suppose that* $\Sigma \in \overline{\mathfrak{S}}^{\leqslant}$ *and there exists a* $\overline{q} \in Q(\Sigma)$ *with* $|\overline{q}| > 3$. *Then*

$$\mathcal{R}(\Sigma) \leqslant \frac{N}{M} R_2(\omega+\eta) M^{\mu-1/3} M^{C\delta}. \tag{4.79}$$

Proof of Lemma 4.19. We first claim that there is at least one domestic bridge, i.e. that $\Sigma_d \neq \emptyset$.

Clearly, $|V(\Sigma)|$ is even. Recall that $\bigcup_{q\in Q(\Sigma)} q = V(\Sigma)$. Since each block of $Q(\Sigma)$ has size 3, we conclude that $|V(\Sigma)|$ is multiple of 3, and hence of 6. A simple exhaustion of all possible pairings $\Sigma \in \overline{\mathfrak{S}}^{\leqslant}$ that saturate the first inequality in (4.46) shows that there is no such Σ satisfying $\left|\bigcup_{q\in Q(\Sigma)} q\right| = 6$. (In fact, any connected pairing with at most six vertices is a dumbbell pairing, which are excluded by the definition (4.38) of $\mathfrak{S}^*$.) Hence we find that $\left|\bigcup_{q\in Q(\Sigma)} q\right| \geqslant 12$, so that $|Q(\Sigma)| \geqslant 4$.

Next, note that $Q(\Sigma)$ contains at most two blocks q that contain white vertices of Σ, since Σ contains four white vertices, and, for each $i \in \{1, 2\}$, those of $\mathcal{C}_i$ are in the same block of $Q(\Sigma)$. (Recall items (ii) and (iii) after (4.25)). Since $|Q(\Sigma)| \geqslant 4$, we find that there is a block $q \in Q(\Sigma)$ that contains only black vertices. Let $\Sigma(q)$ be the set of bridges of Σ touching a vertex of q; see Fig. 11. By definition of $Q(\Sigma)$, we have $|\Sigma(q)| = 3$. Now if all vertices of q belong to the same connected component of $\mathcal{C}$, then $\Sigma(q) \subset \Sigma_d$. Otherwise, let $q = \{i, j, k\}$ with j and k belonging to the same connected component of $\mathcal{C}$. Then both bridges touching i are in Σ_c; the remaining bridge of $\Sigma(q)$ must touch both j and k, and is therefore in Σ_d. Either way, we find that $\Sigma_d \neq \emptyset$, as claimed above.

For the following, abbreviate $s_l := |E_c(\mathcal{Y}(\Sigma)) \setminus E(\mathcal{T})|$ and $s_t := |E_c(\mathcal{Y}(\Sigma)) \cap E(\mathcal{T})|$. From the saturated inequality (4.46) we find

$$|Q(\Sigma)| = \frac{2|\Sigma|}{3} + \frac{2}{3}. \tag{4.80}$$

Plugging this into (4.76) yields

$$\mathcal{R}(\Sigma) = \frac{N}{M} M^{3\delta|\Sigma|} M^{2/3-|\Sigma|/3} R_2(\omega+\eta)^{s_l} M^{\mu s_t}. \tag{4.81}$$

Recall that $|\Sigma_d| \geqslant 1$ and $|\Sigma_d| + s_l + s_t = |\Sigma|$. Moreover, $s_t \leqslant |E(\mathcal{Y}(\Sigma)) \cap E(\mathcal{T})| = |Q(\Sigma)| - 1$. Since $R_2(\omega+\eta) \leqslant M^{\mu}$, we conclude

$$\begin{aligned} R_2(\omega+\eta)^{s_l} M^{\mu s_t} &\leqslant R_2(\omega+\eta)^{s_l+s_t-|Q(\Sigma)|+1} M^{\mu(|Q(\Sigma)|-1)} \\ &= R_2(\omega+\eta)^{|\Sigma|-|\Sigma_d|-|Q(\Sigma)|+1} M^{\mu(|Q(\Sigma)|-1)} \leqslant R_2(\omega+\eta)^{|\Sigma|-|Q(\Sigma)|} M^{\mu(|Q(\Sigma)|-1)}. \end{aligned}$$

Thus we get

$$\begin{aligned}\mathcal{R}(\Sigma) &\leqslant \frac{N}{M} M^{3\delta|\Sigma|} M^{2/3-|\Sigma|/3} R_2(\omega+\eta)^{|\Sigma|-|Q(\Sigma)|} M^{\mu(|Q(\Sigma)|-1)} \\ &= \frac{NM^{3\delta}M^{1/3}}{M} \left(M^{-1/3+3\delta} R_2(\omega+\eta)\right)^{|\Sigma|-|Q(\Sigma)|} (M^{\mu-1/3+3\delta})^{|Q(\Sigma)|-1} \\ &\leqslant \frac{N}{M} R_2(\omega+\eta) M^{3\mu-1} M^{15\delta},\end{aligned}$$

where in the last step we used (4.80) to get $|\Sigma| - |Q(\Sigma)| = |\Sigma|/3 - 2/3 \geqslant 1$, as well as $|Q(\Sigma)| \geqslant 4$ and $\mu < 1/3$. Here we chose $\delta > 0$ in Proposition 3.3 small enough that $\mu < 1/3 - 3\delta$. We also used that $M^{-1/3+3\delta} R_2(\omega+\eta) \leqslant 1$. This concludes the proof. □

Proof of Lemma 4.20. Since $|\overline{q}| \geqslant 4$ and all other blocks of $Q(\Sigma)$ have size at least 3 by (4.77), we find that

$$|Q(\Sigma)| \leqslant 1 + \frac{2|\Sigma| + 2 - |\overline{q}|}{3} \leqslant \frac{2|\Sigma|}{3} + \frac{1}{3}, \tag{4.82}$$

where $2|\Sigma| + 2 - |\overline{q}|$ is the number of vertices of Σ not in $\overline{q}$. Note the improvement of (4.82) over (4.80). Using the notation of the proof of Lemma 4.19, we get from (4.76), in analogy to (4.81),

$$\mathcal{R}(\Sigma) \leqslant \frac{N}{M} M^{3\delta|\Sigma|} M^{1/3-|\Sigma|/3} R_2(\omega+\eta)^{s_l} M^{\mu s_t}.$$

Now we proceed as in the proof of Lemma 4.19, using $|\Sigma_d| \geqslant 0$, $|\Sigma_d| + s_l + s_t = |\Sigma|$, and $s_t \leqslant |Q(\Sigma)| - 1$. We get

$$\begin{aligned}\mathcal{R}(\Sigma) &\leqslant \frac{NM^{1/3}}{M} \left(M^{-1/3+3\delta} R_2(\omega+\eta)\right)^{|\Sigma|-|Q(\Sigma)|+1} (M^{\mu-1/3+3\delta})^{|Q(\Sigma)|-1} \\ &\leqslant \frac{N}{M} R_2(\omega+\eta) M^{\mu-1/3} M^{6\delta}.\end{aligned}$$

In the last step we used that $|\Sigma| - |Q(\Sigma)| \geqslant 0$, which follows from (4.82) and from $|\Sigma| \geqslant 3$ for $\Sigma \in \overline{\mathfrak{S}}^{\leqslant}$, and that $|Q(\Sigma)| \geqslant 2$. (In fact, since $\Sigma \notin \{D_1, \ldots, D_8\}$ one may easily check that $|Q(\Sigma)| \geqslant 3$.) This concludes the proof. □

From Lemmas 4.19, 4.20, and 4.18, we conclude that for all $\Sigma \in \mathfrak{S}^{\leqslant}$ we have

$$|\mathcal{V}'(\Sigma)| \leqslant \mathcal{R}(\Sigma) \leqslant \frac{N}{M} R_2(\omega+\eta) M^{3\mu-1} M^{4\delta K}. \tag{4.83}$$

In order to conclude the proof of Proposition 4.12, we need an analogous estimate of $\mathcal{V}''(\Sigma)$. This may be obtained by repeating the above argument almost verbatim; the only nontrivial difference is that, on the right-hand side of (4.67), the factor $Z(\mathrm{e}^{\mathrm{i}(A_1-A_2)}S)_{y_e}$ associated with the edge $e \in E_c(\mathcal{Y}(\Sigma))$ is replaced with $Z(\mathrm{e}^{\mathrm{i}(A_1+A_2)}S)_{y_e}$. Since $|1 - \mathrm{e}^{\mathrm{i}(A_1+A_2)}| \geqslant c$ on the support of the convolution integral, we replace (4.75) with

$$\left|Z(-\mathrm{e}^{\mathrm{i}(A_1+A_2)}S)_{yz}\right| \leqslant \frac{C}{M} \leqslant \frac{C}{M} R_2(\omega+\eta), \qquad \sum_z \left|Z(\mathrm{e}^{\mathrm{i}(A_1-A_2)}S)_{yz}\right| \leqslant CM^{\mu}.$$

Thus we find, for any $\Sigma \in \mathfrak{S}^{\leqslant}$, that

$$|\mathcal{V}''(\Sigma)| \leqslant \mathcal{R}(\Sigma). \tag{4.84}$$

Hence Proposition 4.12 follows from (4.83), (4.84), and the observation that $\mathfrak{S}^{\leqslant}$ is a finite set that is independent of N. This concludes the proof of Proposition 4.12.

We conclude this subsection with an analogue of Proposition 4.12 in the case **(C1)**. Its proof follows along the same lines as that of Proposition 4.12, and is omitted.

Proposition 4.21. *Suppose that ϕ_1 and ϕ_2 satisfy (2.7). Suppose moreover that (2.9) holds for some small enough $c_* > 0$. Then for any fixed $K \in \mathbb{N}$ we have (4.51).*

For future reference we emphasize that the only information about the matrix entries of $Z(\cdot)$ that is required for the estimate (4.51) to hold is (4.74) and (4.75). Thus, the conclusion of the above argument may be formulated in the following more general form.

Lemma 4.22. *Let $\Sigma \notin \{D_1, \dots D_8\}$, and suppose that we have a family of matrices $\mathcal{Z}(\sigma, E_1, E_2, L) \equiv Z(\sigma)$ parametrized by $\sigma \in \Sigma$ satisfying*

$$|\mathcal{Z}_{xy}(\sigma)| \leqslant \frac{C}{M}, \qquad \sum_y |\mathcal{Z}_{xy}(\sigma)| \leqslant C \log N \tag{4.85}$$

for $\sigma \in \Sigma_1 \cup \Sigma_2$ and

$$|\mathcal{Z}_{xy}(\sigma)| \leqslant \frac{C}{M} M^{2\delta} R_2(\omega + \eta), \qquad \sum_y |\mathcal{Z}_{xy}(\sigma)| \leqslant C M^{\mu} \tag{4.86}$$

for $\sigma \in \Sigma_c$.

Then for small enough δ there exists a $c_0 > 0$ such that

$$\sum_{\mathbf{y} \in \mathbb{T}^{Q(\Sigma)}} \sum_{\mathbf{x} \in \mathbb{T}^{V(\Sigma)}} \Bigg(\prod_{q \in Q(\Sigma)} \prod_{i \in q} \mathbf{1}(x_i = y_q) \Bigg) \Bigg(\prod_{\{e,e'\} \in \Sigma} \mathcal{Z}_{x_e}(\{e,e'\}) \Bigg)$$
$$= \sum_{\mathbf{x} \in \mathbb{T}^{V(\Sigma)}} I_0(\mathbf{x}) \Bigg(\prod_{\{e,e'\} \in \Sigma} J_{\{e,e'\}} \mathcal{Z}_{x_e}(\{e,e'\}) \Bigg) \leqslant \frac{C_\Sigma N}{M} R_2(\omega + \eta) M^{-c_0}.$$

4.7. Conclusion of the proof of Proposition 4.1 and Theorems 2.2–2.4. We may now conclude the proof of Proposition 4.1. As indicated before, the error terms $\mathcal{E}$ resulting from the simplifications **(S1)–(S3)** are small; the precise statement is the following proposition that is proved in [11].

Proposition 4.23. *The error term $\mathcal{E}$ in (4.30) arising from the simplifications* **(S1)–(S3)** *satisfies*

$$|\mathcal{E}| \leqslant \frac{CN}{M} M^{-c_0} R_2(\omega + \eta), \tag{4.87}$$

for some constant $c_0 > 0$.

Proof. This is an immediate consequence of Propositions 4.5, 4.6, and 4.15 in [11]. □

Combining Propositions 4.8, 4.12, 4.21, and 4.23 yields

$$\widetilde{F}^{\eta}(E_1, E_2) = \mathcal{V}_{\mathrm{main}} + \frac{N}{M}\left(O\left(M^{-1} + M^{-c_0} R_2(\omega + \eta)\right) + O_q(NM^{-q})\right).$$

Together with Proposition 4.7, this concludes the proof of Proposition 4.1.

Using (3.20), (2.11), and Proposition 4.1 we therefore get, for H as in Sect. 2,

$$F^{\eta}(E_1, E_2) = \mathcal{V}_{\mathrm{main}} + \frac{N}{M}\left(O\left(M^{-1} + M^{-c_0} R_2(\omega + \eta)\right) + O_q(NM^{-q})\right). \quad (4.88)$$

In order to compute the left-hand side of (2.12), and hence conclude the proof of Theorems 2.2–2.4, we need to control the denominator of (2.12) using the following result.

Lemma 4.24. *For $E \in [-1 + \kappa, 1 - \kappa]$ we have*

$$\mathbb{E}\, Y_{\phi}^{\eta}(E) = 4\sqrt{1 - E^2} + O(\eta) \;=\; 2\pi\nu(E) + O(\eta). \quad (4.89)$$

Proof. In the case **(C1)** we have

$$\begin{aligned}\mathbb{E}\, Y_{\phi}^{\eta}(E) &= \mathbb{E}\,\frac{1}{N}\operatorname{Tr}\phi^{\eta}(H/2 - E) = \mathbb{E}\frac{1}{N}\operatorname{Im}\operatorname{Tr}\frac{4}{H - 2(E + \mathrm{i}\eta)} \\ &= 4\operatorname{Im} m(2E + 2\mathrm{i}\eta) + O(M^{-2/3+c})\end{aligned}$$

for any $c > 0$. Here in the last step we introduced the Stieltjes transform of the semicircle law, $m(z)$, and invoked [15, Theorem 2.3 and Equation (7.6)]. The claim then follows from the estimate $4\operatorname{Im} m(2E + 2\mathrm{i}\eta) = 4\sqrt{1 - E^2} + O(\eta)$, which itself follows from [14, Equations (3.3) and (3.5)].

In the case **(C2)**, we first split $\phi^{\eta} = \phi^{\leqslant,\eta} + \phi^{>,\eta}$ as in (4.60). The contribution of $\phi^{>,\eta}$ is small by the strong decay of ϕ. The error in the main term,

$$\mathbb{E}\,\frac{1}{N}\operatorname{Im}\operatorname{Tr}\phi^{\leqslant,\eta}(H/2 - E) - \frac{1}{2\pi}\int_{-2}^{2}\mathrm{d}x\,\sqrt{4 - x^2}\,\phi^{\leqslant,\eta}(x/2 - E),$$

may be estimated using [15, Theorem 2.3] and Helffer–Sjöstrand functional calculus, as in e.g. [15, Section 7.1]; we omit the details. Then the claim follows from $\frac{1}{2\pi}\int_{-2}^{2}\mathrm{d}x\,\sqrt{4 - x^2}\,\phi^{\leqslant,\eta}(x/2 - E) = 4\sqrt{1 - E^2} + O(\eta)$. □

Now we define

$$\Theta_{\phi_1,\phi_2}^{\eta}(E_1, E_2) := \frac{(LW)^d}{N^2}\,\frac{\mathcal{V}_{\mathrm{main}}}{\mathbb{E}Y_{\phi_1}^{\eta}(E_1)\mathbb{E}Y_{\phi_2}^{\eta}(E_2)}. \quad (4.90)$$

Then Theorems 2.3 and 2.4 follow from Lemma 4.24 and (4.88), recalling (4.5) and (4.7). Moreover, Theorem 2.2 follows from (3.10) and Lemma 4.24. This concludes the proof of Theorems 2.2–2.4 under the simplifications **(S1)**–**(S3)**.

5. The Real Symmetric Case ($\beta = 1$)

In this section we explain the changes needed to the arguments of Sect. 4 to prove Theorems 2.2, 2.3, and 2.4 for $\beta = 1$ instead of $\beta = 2$. The difference is that for $\beta = 1$ we have $\mathbb{E}H_{xy}^2 = S_{xy}$, while for $\beta = 2$ we have $\mathbb{E}H_{xy}^2 = 0$ (in addition to $\mathbb{E}H_{xy}H_{yx} = \mathbb{E}|H_{xy}|^2 = S_{xy}$, which is valid in both cases). This leads to additional terms for $\beta = 1$, which may be conveniently tracked in our graphical notation by introducing *twisted bridges*, in analogy to Section 9 of [13]. As it turns out, allowing twisted bridges results in eight new dumbbell skeletons, called $\widetilde{D}_1, \ldots, \widetilde{D}_8$ below, each of which has the same value $\mathcal{V}(\cdot)$ as its counterpart without a tilde. Hence, for $\beta = 1$ the leading term is simply twice the leading term of $\beta = 2$, which accounts for the trivial prefactor $2/\beta$ in the final formulas. Any other skeleton may be estimated by a trivial modification of the argument from Sects. 4.4–4.6. As in Sect. 4, we make the simplifications **(S1)**–**(S3)**, and do not deal with the errors terms $\mathcal{E}$ resulting from them. They are handled in [11].

We now give a more precise account of the proof for $\beta = 1$. We start from (4.18), which remains unchanged. Since $\mathbb{E}H_{xy}H_{xy} = \mathbb{E}H_{xy}H_{yx} = S_{xy}$, (4.19) holds for $\beta = 1$ without the indicator function $\mathbf{1}(x_e \neq x_{e'})$ that was present for $\beta = 2$. Hence (4.20) also holds without the indicator function $\mathbf{1}(x_e \neq x_{e'})$. We now write

$$\mathbf{1}([x_e] = [x_{e'}]) = \mathbf{1}([x_e] = [x_{e'}])\mathbf{1}(x_e \neq x_{e'}) + \mathbf{1}(x_e = x_{e'}) \;=:\; J_{\{e,e'\}}(\mathbf{x}) + \widetilde{J}_{\{e,e'\}}(\mathbf{x}),$$

in self-explanatory notation (recall that $J_{\{e,e'\}}(\mathbf{x})$ was already defined in (4.22)). Thus (4.23) becomes

$$\left\langle \operatorname{Tr} H^{(n_1)} ; \operatorname{Tr} H^{(n_2)} \right\rangle = \sum_{\Pi \in \mathfrak{M}_c(E(\mathcal{C}))} \sum_{\mathbf{x}} I(\mathbf{x}) \left(\prod_{\{e,e'\} \in \Pi} \left(J_{\{e,e'\}}(\mathbf{x}) + \widetilde{J}_{\{e,e'\}}(\mathbf{x}) \right) S_{x_e} \right) + \mathcal{E}. \tag{5.1}$$

Multiplying out the parentheses in (5.1) yields $2^{|\Pi|}$ terms, each of which is characterized by the set of bridges of Π associated with a factor J; the other bridges are associated with a factor $\widetilde{J}$. We call the former *straight bridges* and the latter *twisted bridges*. This terminology originates from the fact that a twisted bridge forces the labels of the adjacent vertices to coincide on opposite sides of the bridge; see Fig. 12 for an illustration.

More formally, we assign to each bridge of Π a binary tag, *straight* or *twisted*. We represent straight bridges (as before) by solid lines and twisted bridges by dashed lines.

Next, we extend the definition of skeletons from Sect. 4.2 to pairings containing twisted bridges. Recall that the key observation behind the definition of a skeleton was that parallel straight bridges yield a large contribution but a small combinatorial complexity. Now *antiparallel* twisted bridges behave analogously, whereby two bridges $\{e_1, e_1'\}$

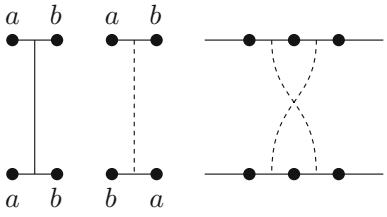

Fig. 12. *Left picture* a straight bridge (*left*) and a twisted bridge (*right*); *labels with the same name* are forced to coincide by the bridge. *Right picture* two antiparallel twisted bridges, which form an antiladder of size two

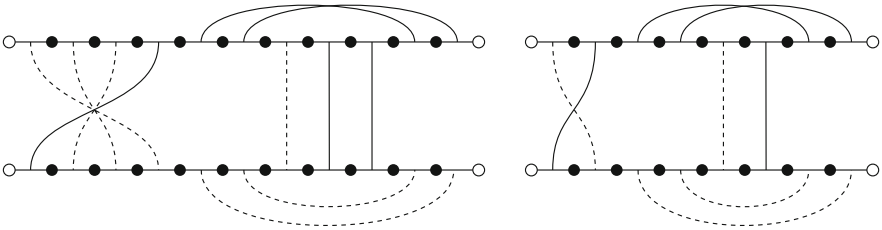

Fig. 13. A tagged pairing (*left*) and its tagged skeleton (*right*)

and $\{e_2, e_2'\}$ are *antiparallel* if $b(e_1) = a(e_2)$ and $b(e_1') = a(e_2')$. (Recall that they are parallel if $b(e_1) = a(e_2)$ and $b(e_2') = a(e_1')$.) See Fig. 12 for an illustration. An *antiladder* is a sequence of bridges such that two consecutive bridges are antiparallel. All of Sect. 4.1, in particular the partition $Q(\Pi)$, may now be taken over with trivial modifications.

As in Sect. 4.2, to each tagged pairing Π we assign a tagged skeleton Σ with associated multiplicities $\mathbf{b}$. The skeleton Σ is obtained from Π by successively collapsing *parallel straight bridges* and *antiparallel twisted bridges* until none remains. Parallel twisted bridges and antiparallel straight bridges remain unaltered. The skeleton Σ inherits the tagging of its bridges in the natural way: two parallel straight bridges are collapsed into a single straight bridge, and two antiparallel twisted bridges are collapsed into a single twisted bridge. See Fig. 13 for an illustration.

We take over all notions from Sect. 4.2, such as $\mathcal{V}(\cdot)$, with the appropriate straightforward modifications for tagged skeletons.

Allowing twisted bridges leads to a further eight skeleton graphs, which we denote by $\widetilde{D}_1, \dots, \widetilde{D}_8$, whose contribution is of leading order. They are the same graphs as $D_1, \dots, D_8$ from Fig. 6, except that the (one or two) vertical antiparallel straight bridges (depicted by solid lines) are replaced with the same number of vertical parallel twisted bridges (depicted by dashed lines). We use the notations

$$\mathcal{V}_{\mathrm{main}} := \sum_{i=1}^{8} \mathcal{V}(D_i), \qquad \widetilde{\mathcal{V}}_{\mathrm{main}} := \sum_{i=1}^{8} \mathcal{V}(\widetilde{D}_i).$$

We record the following simple result, whose proof is immediate.

Lemma 5.1. *If $\beta = 1$ then for $i = 1, \dots, 8$ we have $\mathcal{V}(D_i) = \mathcal{V}(\widetilde{D}_i)$.*

For $\beta = 1$ we may therefore write $\mathcal{V}_{\mathrm{main}} + \widetilde{\mathcal{V}}_{\mathrm{main}} = 2\mathcal{V}_{\mathrm{main}}$. Thus, the main term for $\beta = 1$ is simply twice the main term for $\beta = 2$.

What remains is the estimate of $\mathcal{V}(\Sigma)$ for $\Sigma \notin \{D_1, \dots, D_8, \widetilde{D}_1, \dots, \widetilde{D}_8\}$. We proceed exactly as in Sects. 4.4–4.6. The key observation is that the 2/3-rule from Lemma 4.11 remains true thanks to the definition of skeletons. When estimating the large skeletons (without making use of oscillations) in Sect. 4.4, we get an extra factor 2^m to the left-hand side of (4.50) arising from the sum over all possible taggings of a skeleton; this factor is clearly immaterial. Finally, the argument of Sects. 4.5 and 4.6 may be taken over with merely cosmetic changes. The set Σ_d^{I} from Definition 4.17 remains unchanged, and in particular only contains straight bridges. Note that the basic graph-theoretic argument from Lemmas 4.19 and 4.20 remains unchanged. In particular, exactly as in the proof of Lemma 4.19, if all blocks of $Q(\Sigma)$ have size three and Σ is not a dumbbell skeleton, then Σ contains a domestic bridge (which may be straight or twisted). This concludes the proof of Theorems 2.2–2.4 for the case $\beta = 1$ under the simplifications **(S1)**–**(S3)**.

Acknowledgements. We are very grateful to Alexander Altland for detailed discussions on the physics of the problem and for providing references.

References

1. Aizenman, M., Molchanov, S.: Localization at large disorder and at extreme energies: an elementary derivation. Commun. Math. Phys. **157**, 245–278 (1993)
2. Aizenman, M., Sims, R., Warzel, S.: Absolutely continuous spectra of quantum tree graphs with weak disorder. Commun. Math. Phys. **264**, 371–389 (2006)
3. Altschuler, B.L.: Fluctuations in the extrinsic conductivity of disordered conductors. JETP Lett. **41**, 648–651 (1985)
4. Altshuler, B.L., Shklovskii, B.I.: Repulsion of energy levels and the conductance of small metallic samples. Zh. Eksp. Teor. Fiz. (Sov. Phys. JETP) **91**(64), 220(127) (1986)
5. Anderson, G., Zeitouni, O.: A CLT for a band matrix model. Probab. Theory Relat. Fields **134**, 283–338 (2006)
6. Anderson, P.W.: Absence of diffusion in certain random lattices. Phys. Rev. **109**, 1492 (1958)
7. Ayadi, S.: Asymptotic properties of random matrices of long-range percolation model. Random Oper. Stoch. Equ. **17**, 295–341 (2009)
8. Boutetde Monvel, A., Khorunzhy, A.: Asymptotic distribution of smoothed eigenvalue density. I. Gaussian random matrices. Random Oper. Stoch. Equ. **7**, 1–22 (1999)
9. Boutetde Monvel, A., Khorunzhy, A.: Asymptotic distribution of smoothed eigenvalue density. II. Wigner random matrices. Random Oper. Stoch. Equ. **7**, 149–168 (1999)
10. Dyson, F.J., Mehta, M.L.: Statistical theory of the energy levels of complex systems. IV. J. Math. Phys. **4**, 701–713 (1963)
11. Erdős, L., Knowles, A.: The Altshuler–Shklovskii formulas for random band matrices II: the general case. Ann. H. Poincaré (2014, preprint). arXiv:1309.5107
12. Erdős, L., Knowles, A.: Quantum diffusion and delocalization for band matrices with general distribution. Ann. H. Poincaré **12**, 1227–1319 (2011)
13. Erdős, L., Knowles, A.: Quantum diffusion and eigenfunction delocalization in a random band matrix model. Commun. Math. Phys. **303**, 509–554 (2011)
14. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Delocalization and diffusion profile for random band matrices. Commun. Math. Phys. **323**(1), 367–416 (2013)
15. Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: The local semicircle law for a general class of random matrices. Electron. J. Probab. **18**, 1–58 (2013)
16. Erdős, L., Salmhofer, M., Yau, H.-T.: Quantum diffusion for the Anderson model in the scaling limit. Ann. H. Poincaré **8**, 621–685 (2007)
17. Erdős, L., Salmhofer, M., Yau, H.-T.: Quantum diffusion of the random Schrödinger evolution in the scaling limit II. the recollision diagrams. Commun. Math. Phys. **271**, 1–53 (2007)
18. Erdős, L., Salmhofer, M., Yau, H.-T.: Quantum diffusion of the random Schrödinger evolution in the scaling limit. Acta Math. **200**, 211–277 (2008)
19. Erdős, L., Schlein, B., Yau, H.-T.: Local semicircle law and complete delocalization for Wigner random matrices. Commun. Math. Phys. **287**, 641–655 (2009)
20. Erdős, L., Schlein, B., Yau, H.-T.: Universality of random matrices and local relaxation flow. Invent. Math. **185**(1), 75–119 (2011)
21. Erdős, L., Yau, H.-T.: Linear Boltzmann equation as the weak coupling limit of the random Schrödinger equation. Commun. Math. Phys. **53**, 667–735 (2000)
22. Erdős, L., Yau, H.-T.: Universality of local spectral statistics of random matrices. Bull. Am. Math. Soc **49**, 377–414 (2012)
23. Erdős, L., Yau, H.-T., Yin, J.: Bulk universality for generalized Wigner matrices. Probab. Theor. Relat. Fields **154**, 341–407 (2012)
24. Erdős, L., Yau, H.-T., Yin, J.: Rigidity of eigenvalues of generalized Wigner matrices. Adv. Math. **229**, 1435–1515 (2012)
25. Feldheim, O.N., Sodin, S.: A universality result for the smallest eigenvalues of certain sample covariance matrices. Geom. Funct. Anal. **20**, 88–123 (2010)
26. Froese, R., Hasler, D., Spitzer, W.: Transfer matrices, hyperbolic geometry and absolutely continuous spectrum for some discrete Schrödinger operators on graphs. J. Funct. Anal. **230**, 184–221 (2006)
27. Fröhlich, J., Roeck, W.de : Diffusion of a massive quantum particle coupled to a quasi-free thermal medium in dimension $d \geqslant 4$. Commun. Math. Phys. **303**, 613–707 (2011)
28. Fröhlich, J., Martinelli, F., Scoppola, E., Spencer, T.: Constructive proof of localization in the Anderson tight binding model. Commun. Math. Phys. **101**, 21–46 (1985)

29. Fröhlich, J., Spencer, T.: Absence of diffusion in the Anderson tight binding model for large disorder or low energy. Commun. Math. Phys. **88**, 151–184 (1983)
30. Fyodorov, Y.V., Mirlin, A.D.: Scaling properties of localization in random band matrices: a σ-model approach. Phys. Rev. Lett. **67**, 2405–2409 (1991)
31. Gradshteyn, I.S., Ryzhik, I.M.: Table of Integrals, Series, and Products, 7th edn. Academic Press, London (2007)
32. Klein, A.: Absolutely continuous spectrum in the anderson model on the Bethe lattice. Math. Res. Lett. **1**, 399–407 (1994)
33. Kravtsov, V.E., Lerner, I.V.: Level correlations driven by weak localization in 2d systems. Phys. Rev. Lett. **74**, 2563–2566 (1995)
34. Lee, P.A., Stone, A.D.: Universal conductance fluctuations in metals. Phys. Rev. Lett. **55**, 1622–1625 (1985)
35. Li, L., Soshnikov, A.: Central limit theorem for linear statistics of eigenvalues of band random matrices. Random Matrices Theory Appl. **02**, 1350009 (2013)
36. Mehta, M.L.: Random Matrices. Academic press, London (2004)
37. Minami, N.: Local fluctuation of the spectrum of a multidimensional Anderson tight binding model. Commun. Math. Phys. **177**, 709–725 (1996)
38. De Roeck, W., Kupiainen, A.: Diffusion for a quantum particle coupled to phonons in $d \geqslant 3$. Commun. Math. Phys. **323**(3), 889–973 (2013)
39. Schenker, J.: Eigenvector localization for random band matrices with power law band width. Commun. Math. Phys. **290**, 1065–1097 (2009)
40. Shcherbina, T.: On the second mixed moment of the characteristic polynomials of the 1D band matrices. Commun. Math. Phys. **328**(1), 45–82 (2014)
41. Shcherbina, T.: Universality of the local regime for the block band matrices with a finite number of blocks. J. Stat. Phys. **155**(3), 466–499 (2014)
42. Silvestrov, P.G.: Summing graphs for random band matrices. Phys. Rev. E **55**, 6419–6432 (1997)
43. Sinai, Y., Soshnikov, A.: Central limit theorem for traces of large random symmetric matrices with independent matrix elements. Bol. Soc. Bras. Mat. **29**, 1–24 (1998)
44. Sodin, S.: The spectral edge of some random band matrices. Ann. Math. **172**(3), 2223–2251 (2010)
45. Soshnikov, A.: The central limit theorem for local linear statistics in classical compact groups and related combinatorial identities. Ann. Probab. **28**, 1353–1370 (2000)
46. Sosoe, P., Wong, P.: Regularity conditions in the CLT for linear eigenvalue statistics of Wigner matrices. Adv. Math. **249**(20), 37–87 (2013)
47. Spencer, T.: Random banded and sparse matrices, Chapter 23. In: Akemann, G., Baik, J., Di Francesco, P. (eds.) Oxford Handbook of Random Matrix Theory (2011)
48. Tao, T., Vu, V.: Random matrices: universality of local eigenvalue statistics. Acta Math. **206**, 1–78 (2011)
49. Thouless, D.J.: Maximum metallic resistance in thin wires. Phys. Rev. Lett. **39**, 1167–1169 (1977)
50. Wigner, E.P.: Characteristic vectors of bordered matrices with infinite dimensions. Ann. Math. **62**, 548–564 (1955)

Communicated by H.-T. Yau