Averaging Fluctuations in Resolvents of Random Band Matrices

László Erdős[1*] Antti Knowles[2†] Horng-Tzer Yau[2‡]

Institute of Mathematics, University of Munich,
Theresienstrasse 39, D-80333 Munich, Germany
lerdos@math.lmu.de [1]

Department of Mathematics, Harvard University
Cambridge MA 02138, USA
knowles@math.harvard.edu, htyau@math.harvard.edu [2]

January 10, 2013

We consider a general class of random matrices whose entries are centred random variables, independent up to a symmetry constraint. We establish precise high-probability bounds on the averages of arbitrary monomials in the resolvent matrix entries. Our results generalize the previous results of [5, 16, 17] which constituted a key step in the proof of the local semicircle law with optimal error bound in mean-field random matrix models. Our bounds apply to random band matrices, and improve previous estimates from order 2 to order 4 in the cases relevant for applications. In particular, they lead to a proof of the diffusion approximation for the magnitude of the resolvent of random band matrices. This, in turn, implies new delocalization bounds on the eigenvectors. The applications are presented in a separate paper [3].

AMS Subject Classification: 15B52, 82B44, 82C44

Keywords: random band matrix, delocalization, sums of correlated random variables.

[*] Partially supported by SFB-TR 12 Grant of the German Research Council
[†] Partially supported by NSF grant DMS-0757425
[‡] Partially supported by NSF grants DMS-0757425, 0804279

1. Introduction

Let $H = (h_{ij})$ be a complex Hermitian or real symmetric $N \times N$ random matrix with centred matrix entries that are independent up to the symmetry constraint. We assume that the variances $s_{ij} := \mathbb{E}|h_{ij}|^2$ are normalized so that $\sum_j s_{ij} = 1$ for each i, and let $\|s\|_\infty := \max_{ij} s_{ij}$ denote the maximal variance. Let $G_{ab}(z) := (H-z)^{-1}_{ab}$ denote the resolvent matrix entries evaluated at a spectral parameter $z = E + i\eta$ whose imaginary part η is positive and small. It was established in [4, 15] that

$$\Lambda \;:=\; \max_{a \neq b} |G_{ab}| \;\lesssim\; \sqrt{\frac{\|s\|_\infty}{\eta}} \tag{1.1}$$

with high probability for large N, up to factors of N^ε.

The matrix entries $G_{ab} \equiv G_{ab}(z)$ depend strongly on the entries of the a-th and b-th columns of H, but weakly on the other columns. Focusing on the dependence on a only, this can be seen from the simple expansion formula

$$G_{ab} \;=\; -G_{aa} \sum_{i \neq a} h_{ai} G^{(a)}_{ib}\,, \tag{1.2}$$

where $G^{(a)}$ denotes the resolvent of the $(N-1)\times(N-1)$ minor of H obtained by removing the a-th row and column (see Lemma 3.7 below for the general statement). Since $G^{(a)}$ is independent of the family $(h_{ai})_{i=1}^N$, the formula (1.2) expresses G_{ab} as a sum of independent centred random variables (neglecting the prefactor G_{aa} which still depends on $(h_{ai})_{i=1}^N$). Therefore the size of G_{ab} is governed by a fluctuation averaging mechanism, similar to the central limit theorem. This is the main reason why the bound (1.1) is substantially better than the naive estimate $|G_{ab}| \leqslant \eta^{-1}$.

In this paper we investigate a more subtle phenomenon. To take a simple example, we are interested in averages of resolvent matrix entries of the form

$$\frac{1}{N} \sum_a G_{ab} \tag{1.3}$$

or, more generally, its weighted version

$$\sum_a s_{\mu a} G_{ab}\,, \tag{1.4}$$

where μ and b are fixed. We aim to show that, with high probability, these averages are of order Λ^2 – much smaller than the naive bound Λ which results from an application of (1.1) to each summand (we shall always work in the regime where $\Lambda \ll 1$). The mechanism behind this improved bound is that for $a \neq a'$ the matrix entries G_{ab} and $G_{a'b}$ are only weakly correlated. To see this, note that, since h_{ai} in (1.2) and $h_{a'i}$ in the analogous formula

$$G_{a'b} \;=\; -G_{a'a'} \sum_{i' \neq a'} h_{a'i'} G^{(a')}_{i'b}\,,$$

are independent, the correlation between G_{ab} and $G_{a'b}$ primarily comes from correlations between h_{ai} and $G^{(a')}_{i'b}$ and between $h_{a'i'}$ and $G^{(a)}_{ib}$. (As above, here we neglect the less important prefactors G_{aa} and $G_{a'a'}$.) Now $G^{(a')}_{i'b}$ depends only weakly on h_{ai} unless some lower indices coincide: $i = i'$ or $i = b$ or $i' = a$. Such coincidences are atypical, however, and consequently give rise to lower-order terms. Once the smallness of the correlation between G_{ab} and $G_{a'b}$ is established, the variance of the averages (1.3) or (1.4) can be estimated. The smallness of the higher-order correlations between different resolvent matrix entries allows one to compute high moments and turn the variance bound into a high-probability bound. However, keeping track of all weak correlations among a large product of expressions of the form (1.3) with different a's is rather involved, and we shall need to develop a graphical representation to do this effectively.

This idea of exploiting the weak dependence among different resolvent entries of random matrices first appeared in [17] and was subsequently used in [5, 16, 18]. Such estimates provide optimal error bounds in the *local semicircle law* – a basic ingredient in establishing the universality of local statistics of for Wigner matrices.

Our main result in this paper estimates with high probability (weighted) averages of general monomials in the resolvent matrix entries and their complex conjugates, where the averaging is performed on a subset of the indices. A more complicated example is

$$\sum_{a,b} s_{\mu a} s_{\nu b} \Big(|G_{ab}|^2 |G_{a\rho}|^2 - \mathbb{E}_{ab} |G_{ab}|^2 |G_{a\rho}|^2 \Big) \,, \tag{1.5}$$

where μ, ν, and ρ are fixed. Here we subtract from each summand its partial expectation $\mathbb{E}_{ab}$ with respect to the random variables in the a-th and b-th columns of H. (Note that we could also have subtracted $\mathbb{E}_a G_{ab}$ in (1.3) and (1.4) as well, but this expectation turns out to be negligible, unlike the expectations of the manifestly positive quantity $|G_{ab}|^2 |G_{a\rho}|^2$ in (1.5)).

The expression (1.5) can trivially be estimated by Λ^4 with high probability using the estimate (1.1) on each summand (neglecting that diagonal resolvent matrix entries G_{aa} require a different estimate). However, we can in fact do better: the averaging over two indices gives rise to a cancellation of fluctuations, due to the weak correlations among the summands. Since each averaging independently yields an extra factor Λ as in (1.3) and (1.4), it seems plausible that the naive estimate of order Λ^4 on (1.5) can be improved to Λ^6. This in fact turns out to be correct in the example (1.5), but in general the principle that each averaging yields one extra Λ factor is not optimal. Depending on the structure of the monomial, the gain may be more than a single factor Λ per averaged index. For example, averaging in the index a in the quantities

$$(\mathrm{I}) \;:=\; \sum_a s_{\mu a} \big(G_{ba} G^*_{ab} - \mathbb{E}_a G_{ba} G^*_{ab} \big) \qquad \text{and} \qquad (\mathrm{II}) \;:=\; \sum_a s_{\mu a} \big(G_{ba} G_{ab} - \mathbb{E}_a G_{ba} G_{ab} \big) \tag{1.6}$$

has different effects. The naive estimate using (1.1) yields Λ^2 for both quantities, but (I) is in fact of order Λ^4 while the (II) is only of order Λ^3 (all estimates are understood with high probability).

The reason behind the gain of a factor Λ^2 over the naive size in case of (I) is quite subtle. We already mentioned that the dependence of G_{ab} on the random variables in the c-th column is weak if $c \neq a, b$. This is manifested in the identity

$$G_{ab} \;=\; G^{(c)}_{ab} + \frac{G_{ac} G_{cb}}{G_{cc}} \,. \tag{1.7}$$

(This identity first appeared in [15]; see Lemma 3.7 below for a precise statement and related formulas.) Since $G^{(c)}_{ab}$ is independent of the c-th column, the c-dependence of G_{ab} is contained in the second term of (1.7). This term is naively of order Λ^2, i.e. smaller than the main term (accepting that G_{cc} in the denominator is harmless; in fact it turns out to be bounded from above and below by universal positive constants). Computing the variance of (I) results in a double sum $\sum_a \sum_c$. We shall see that, since the first term of (1.7) is independent of c, the leading order contribution to the variance in fact comes from the second term. This yields an improvement of one Λ over the naive bound Λ^2. These ideas lead to a bound of order Λ^3 for both

(I) and (II). The idea of using averaging to improve a trivial bound on resolvent entries by an extra factor Λ was central in [17]. In that paper this idea was applied to a specific quantity analogous to

$$\frac{1}{N}\sum_a (1-\mathbb{E}_a)\frac{1}{G_{aa}}\,. \tag{1.8}$$

When we compute a high moment of the quantities in (1.6), we successively use formulas (1.7) and (1.2) and take partial expectation in the expanded indices. The result is the average of a high-order monomial of resolvent matrix entries. Whether this averaging reduces the naive size depends on the precise structure of the monomial. For example,

$$\sum_c s_{\mu c} G_{bc} G^*_{cb} \;=\; \sum_c s_{\mu c}|G_{bc}|^2 = O(\Lambda^2) \tag{1.9}$$

and this estimate is optimal, while

$$\sum_c s_{\mu c} G_{bc} G_{cb} \;=\; O(\Lambda^3). \tag{1.10}$$

It turns out that average of the high-order monomial obtained from computing a high moment of (I) in (1.6) contains several summations of the type (1.10), while the analogous formula for (II) contains only summations of the type (1.9) (at least to leading order). Whether the additional gain is present or not depends on the precise structure of the original monomial, in particular on how many times the averaging index appears in an entry of G or G^*. In this regard the expressions (I) and (II) differ, which is the reason why their sizes differ. Our main result (Theorem 4.8) expresses the precise relation between the maximal gain and the structure of the monomial. As it turns out, this dependence is quite subtle. The main purpose of this paper is to give a systematic rule, applicable to arbitrary monomials in the resolvent entries, which determines the gain from all indices over which an average is taken. In particular, averaging over certain indices yields an improvement of order Λ^2; this is a novel phenomenon. This observation is crucial in the application of our results to the problem of quantum diffusion in random band matrices [3].

Finally, we shortly explain the improvement from the naive size Λ^2 to Λ^3 for the left-hand side of (1.10). It follows from the estimate of order Λ^3 on (II) in (1.6) and from the fact that $\mathbb{E}_c G_{ac} G_{cb} = O(\Lambda^3)$ for any a, b. That the expectation $\mathbb{E}_c G_{ac} G_{cb}$ itself is smaller than its naive size Λ^2 may be seen by expanding $G_{ac} G_{cb}$ in the index c using formulas of the type (1.2). It turns out that $\mathbb{E}_c G_{ac} G_{cb}$, viewed as a vector indexed by c and keeping a and b fixed, satisfies a stable self-consistent vector equation (see (7.16)). The analysis of this equation leads to the improved bound on $\mathbb{E}_c G_{ac} G_{cb}$ of order Λ^3.

Bounds on averages of resolvents of random matrices have played an essential role in establishing the local semicircle law with an optimal error bound. We recall that in the simplest case of Wigner matrices, where $s_{ij} = N^{-1}$, the trace of the resolvent

$$m_N(z) \;:=\; \frac{1}{N}\operatorname{Tr} G(z) \;=\; \frac{1}{N}\sum_a G_{aa}(z)$$

is well approximated by the Stieltjes transform of the celebrated Wigner semicircle law

$$m(z) \;:=\; \frac{1}{2\pi}\int_{-2}^{2}\frac{\sqrt{4-x^2}}{x-z}\,\mathrm{d}x\,.$$

The optimal bound is

$$|m(z) - m_N(z)| \;\lesssim\; \frac{1}{N\eta} \tag{1.11}$$

with high probability (see [16] for the precise statement and the history of this result). One of the main steps in proving this optimal bound is to exploit that G_{aa} and $G_{a'a'}$ are only weakly correlated for $a \neq a'$. Hence the average of G_{aa} in a in the definition of $m_N(z)$ fluctuates on a smaller scale than the fluctuations of G_{aa}. Various forms of this fluctuation averaging were formulated in [5, 16, 17]. They were the key inputs to prove (1.11) and its analogue for the sample covariance matrices in [18]. In Proposition 6.1, we present a simple special case of our main result, Theorem 4.8. This proposition yields generalizations of estimates analogous to the previous fluctuation averaging bounds with a more streamlined proof. A somewhat different simplification was given in [18].

On the one hand, Theorem 4.8 is more general than its predecessors since it is applicable to arbitrary monomials in G and G^*, and also holds for universal Wigner matrices with nonconstant variances. On the other hand, and more importantly, Theorem 4.8 gives a stronger bound because it exploits the additional cancellation effect explained in connection with the different bounds on the two quantities in (1.6). This extra cancellation mechanism was not present in [5, 16–18].

In a separate paper [3] we apply the stronger bound

$$\sum_a s_{\mu a}\big(|G_{ab}|^2 - \mathbb{E}_a|G_{ab}|^2\big) \;=\; O(\Lambda^4) \tag{1.12}$$

to derive a lower bound on the localization length of random band matrices. Extensions of the methods of [5, 16–18] would have yielded only

$$\sum_a s_{\mu a}\big(|G_{ab}|^2 - \mathbb{E}_a|G_{ab}|^2\big) \;=\; O(\Lambda^3)\,. \tag{1.13}$$

Had we had only (1.13) available in [3], the resulting estimate on the localization length would not have improved the previously known results [1, 2] on eigenvector delocalization.

We conclude this section with a roadmap of the paper. In Section 2 we define our main objects and introduce notation used throughout the paper. Our main result is Theorem 4.8 in Section 4. Before stating it in full generality, we first present a special case, Proposition 3.3, in Section 3. In order to motivate the concepts underlying Theorem 4.8, we not only state this special case but also give a sketch of its proof, in Section 3.2. This is done before the main theorem is stated. A reader who prefers an inductive presentation should follow our sections in sequential order. A reader who wants to jump quickly to the main result may skip Section 3.2. However, some concepts introduced in Section 3.2 are needed later in the proof (but not in the statement) of Theorem 4.8. The full proof of Theorem 4.8 is presented in Sections 6–9, following Section 5 where we give an outline of the proof and explain how Sections 6–9 are related.

2. Setup

Let $(h_{ij} : i \leqslant j)$ be a family of independent, complex-valued random variables $h_{ij} \equiv h_{ij}^{(N)}$ satisfying $\mathbb{E}h_{ij} = 0$ and $h_{ii} \in \mathbb{R}$ for all i. For $i > j$ we define $h_{ij} := \overline{h}_{ji}$, and denote by $H \equiv H_N = (h_{ij})_{i,j=1}^N$ the $N \times N$ matrix with entries h_{ij}. By definition, H is Hermitian: $H = H^*$. We abbreviate

$$s_{ij} \;:=\; \mathbb{E}|h_{ij}|^2\,, \qquad M \;\equiv\; M_N \;:=\; \frac{1}{\max_{i,j} s_{ij}}\,. \tag{2.1}$$

In particular, we have the bound

$$s_{ij} \;\leqslant\; M^{-1} \tag{2.2}$$

for all i and j. We introduce the $N \times N$ symmetric matrix $S \equiv S_N = (s_{ij})_{i,j=1}^N$. We assume that S is (doubly) stochastic:

$$\sum_j s_{ij} \;=\; 1 \tag{2.3}$$

for all i. We shall always assume the bounds

$$N^\delta \;\leqslant\; M \;\leqslant\; N \tag{2.4}$$

for some fixed $\delta > 0$.

Example 2.1 (Band matrix). Fix $d \in \mathbb{N}$. Let f be a bounded and symmetric (i.e. $f(x) = f(-x)$) probability density on $\mathbb{R}^d$. Let L and W be integers satisfying

$$L^{\delta'} \;\leqslant\; W \;\leqslant\; L$$

for some fixed $\delta' > 0$. Define the d-dimensional discrete torus

$$\mathbb{T}_L^d \;=\; [-L/2, L/2)^d \cap \mathbb{Z}^d\,.$$

Thus, $\mathbb{T}_L^d$ has $N = L^d$ lattice points, and we may identify $\mathbb{T}_L^d$ with $\{1, \dots, N\}$. We define the canonical representative of $i \in \mathbb{Z}^d$ through

$$[i]_L \;:=\; (i + L\mathbb{Z}^d) \cap \mathbb{T}_L^d\,.$$

Then H is a *d-dimensional band matrix* with band width W and profile function f if

$$s_{ij} \;=\; \frac{1}{Z_L}\, f\biggl(\frac{[i-j]_L}{W}\biggr)\,.$$

It is not hard to see that $M = \bigl(W^d + O(W^{d-1})\bigr)/\|f\|_\infty$ as $L \to \infty$. The rows and columns of H are thus indexed by the lattice points in $\mathbb{T}_L^d$, i.e. they are equipped with the geometry of $\mathbb{Z}^d$. For $d = 1$, assuming that f is compactly supported, the matrix entry h_{ij} vanishes if $|i-j|$ is larger than CW, i.e. H is a band matrix in the traditional sense.

It is often convenient to use the normalized entries

$$\zeta_{ij} \;:=\; (s_{ij})^{-1/2} h_{ij}\,,$$

which satisfy $\mathbb{E}\zeta_{ij} = 0$ and $\mathbb{E}|\zeta_{ij}|^2 = 1$. (If $s_{ij} = 0$ we set for convenience ζ_{ij} to be a normalized Gaussian, so that these relations remain valid. Of course in this case the law of ζ_{ij} is immaterial.) We assume that the random variables ζ_{ij} have finite moments, uniformly in N, i, and j, in the sense that for all $p \in \mathbb{N}$ there is a constant μ_p such that

$$\mathbb{E}|\zeta_{ij}|^p \;\leqslant\; \mu_p \tag{2.5}$$

for all N, i, and j. We make this assumption to streamline notation in the statements of results such as Theorem 4.8 and the proofs. In fact, our results hold, with the same proof, provided (2.5) is valid for some large but fixed p. See Remark 4.11 below for a more precise statement.

Throughout the following we use a spectral parameter $z \in \mathbb{C}$ satisfying $\operatorname{Im} z > 0$. We shall use the notation

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

without further comment. The Stieltjes transform of Wigner's semicircle law is defined by

$$m \;\equiv\; m(z) \;:=\; \frac{1}{2\pi} \int_{-2}^{2} \frac{\sqrt{4-\xi^2}}{\xi - z}\, \mathrm{d}\xi\,. \tag{2.6}$$

To avoid confusion, we remark that the Stieltjes transform m was denoted by m_{sc} in the papers [5–17], in which m had a different meaning from (2.6). It is well known that the Stieltjes transform m satisfies the identity

$$m(z) + \frac{1}{m(z)} + z \;=\; 0\,. \tag{2.7}$$

We define the *resolvent* of H through

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

and denote its entries by $G_{ij}(z)$. We also write $G^*(z) := (G(z))^* = (H - \bar{z})^{-1}$. We often drop the argument z and write $G \equiv G(z)$ as well as $G^* \equiv G^*(z)$.

Definition 2.2 (Minors). For $T \subset \{1, \dots, N\}$ we define $H^{(T)}$ by

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

Moreover, we define the resolvent of $H^{(T)}$ through

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

We also set

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

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

Definition 2.3 (Partial expectation and independence). Let $X \equiv X(H)$ be a random variable. For $i \in \{1, \dots, N\}$ define the operations P_i and Q_i through

$$P_i X \;:=\; \mathbb{E}(X | H^{(i)})\,, \qquad Q_i X \;:=\; X - P_i X\,.$$

We call P_i *partial expectation* in the index i. Moreover, we say that X is *independent of* $T \subset \{1, \dots, N\}$ if $X = P_i X$ for all $i \in T$.

The following definition introduces a notion of a high-probability bound that is suited for our purposes.

Definition 2.4 (Stochastic domination). Let $X = \big(X^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}\big)$ be a family of random variables, where $U^{(N)}$ is a possibly N-dependent parameter set. Let $\Psi = \big(\Psi^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}\big)$ be a deterministic family satisfying $\Psi^{(N)}(u) \geqslant 0$. We say that X is *stochastically dominated by Ψ, uniformly in u,* if for all (small) $\varepsilon > 0$ and (large) $D > 0$ we have

$$\sup_{u \in U^{(N)}} \mathbb{P}\Big[\big|X^{(N)}(u)\big| > N^{\varepsilon} \Psi^{(N)}(u)\Big] \;\leqslant\; N^{-D}$$

for large enough $N \geqslant N_0(\varepsilon, D)$. Unless stated otherwise, throughout this paper the stochastic domination will always be uniform in all parameters apart from the parameter δ in (2.4) and the sequence of constants μ_p in (2.5); thus, $N_0(\varepsilon, D)$ also depends on δ and μ_p. If X is stochastically dominated by Ψ, uniformly in u, we use the equivalent notations

$$X \;\prec\; \Psi \qquad \text{and} \qquad X \;=\; O_{\prec}(\Psi)\,.$$

For example, using Chebyshev's inequality and (2.5) one easily finds that

$$h_{ij} \;\prec\; (s_{ij})^{1/2} \;\prec\; M^{-1/2}\,, \tag{2.8}$$

so that we may also write $h_{ij} = O_{\prec}((s_{ij})^{1/2})$. The relation $\prec$ satisfies the familiar algebraic rules of order relations. For instance if $A_1 \prec \Psi_1$ and $A_2 \prec \Psi_2$ then $A_1 + A_2 \prec \Psi_1 + \Psi_2$ and $A_1 A_2 \prec \Psi_1 \Psi_2$. Moreover, if $A \prec \Psi$ and there is a constant $C > 0$ such that $\Psi \geqslant N^{-C}$ and $|A| \leqslant N^{C}$ almost surely, then $P_i A \prec \Psi$ and $Q_i A \prec \Psi$. More general statements in this spirit are given in Lemma 3.6 below.

Let $\gamma > 0$ be a fixed small positive constant and let $(\mathbf{S}^{(N)})$ be a sequence of domains satisfying

$$\mathbf{S}^{(N)} \;\subset\; \big\{z \in \mathbb{C} : -10 \leqslant E \leqslant 10\,,\; M^{-1+\gamma} \leqslant \eta \leqslant 10\big\}\,.$$

As usual, we shall systematically omit the index N on $\mathbf{S}$.

Definition 2.5. A positive N-dependent deterministic function $\Psi \equiv \Psi^{(N)}$ on $\mathbf{S}$ is called a *control parameter*. The control parameter Ψ is *admissible* if there is a constant $c > 0$ such that

$$M^{-1/2} \;\leqslant\; \Psi(z) \;\leqslant\; M^{-c} \tag{2.9}$$

for all N and $z \in \mathbf{S}$.

In this paper we always consider families $X^{(N)}(u) = X_i^{(N)}(z)$ indexed by $u = (z, i)$, where $z \in \mathbf{S}$ and i takes on values in some finite (possibly N-dependent or empty) index set.

We slightly modify the definition (1.1) to include a control on the diagonal entries of G in addition to the off-diagonal entries. For the rest of the paper, we define the (z-dependent) random variable

$$\Lambda(z) \;:=\; \max_{x,y} \big|G_{xy}(z) - \delta_{xy} m(z)\big|\,.$$

The variable Λ will play the role of a *random* control parameter. If Ψ is an admissible control parameter, the lower bound on Ψ in (2.9) together with (2.8) implies that

$$h_{ij} \;\prec\; \Psi\,. \tag{2.10}$$

3. Simple examples and ingredients of the proof

In this section we give an informal overview of fluctuation averaging, by stating and sketching the proofs of a few simple, yet representative, cases. Our starting point will always be an admissible control parameter Ψ that controls Λ, i.e. $\Lambda \prec \Psi$. In addition to Ψ, we introduce the secondary control parameter

$$\Phi \;\equiv\; \Phi_\Psi \;:=\; \min\bigl\{\varrho\bigl(\Psi + M^{-1/2}\Psi^{-1}\bigr)\,,\,1\bigr\}\,, \tag{3.1}$$

where we defined the coefficient[1]

$$\varrho \;:=\; \bigl\|(1-m^2S)^{-1}\bigr\|_{\ell^\infty\to\ell^\infty}\,. \tag{3.2}$$

Thus, Φ is defined in terms of the primary control parameter Ψ, although we usually do not indicate this explicitly.

Remark 3.1. We use the somewhat complicated definitions (3.1) and (3.2) because they emerge naturally from our argument, and do not require us to impose any further conditions on the matrix H or the spectral parameter z. The parameter Φ will describe the gain associated with a charged vertex or a chain vertex; see Definitions 4.7 and 5.1 below.

In the motivating example of band matrices (Example 2.1), the parameter Φ may be considerably simplified. Indeed, in that case there is a positive constant C such that

$$\varrho \;\leqslant\; \frac{C\log N}{(\operatorname{Im} m)^2}\,, \tag{3.3}$$

as proved in Proposition B.2 below. For most applications, we are interested in the bulk spectrum of the band matrix, i.e. $E\in[-2+\kappa,2-\kappa]$ for some fixed $\kappa>0$. In that case the relation $\operatorname{Im} m(z)\asymp\sqrt{\eta+2-|E|}$ (proved e.g. in [17, Lemma 4.2]) yields $\operatorname{Im} m \geqslant c$ for some positive constant c depending on κ. We conclude that $1\leqslant\varrho\leqslant C\log N$; the logarithmic factor in the upper bound is irrelevant, since Φ will always be used as a deterministic control parameter in Definition 2.4. In summary: for the bulk spectrum of a band matrix, we may replace Φ with $\Psi+M^{-1/2}\Psi^{-1}$.

Moreover, in typical applications the imaginary part η of the spectral parameter z is small enough that $\Psi\geqslant M^{-1/4}$. In this case Φ and Ψ are comparable (in the bulk spectrum), and hence interchangeable as control parameters in Definition 2.4.

Remark 3.2. We have the lower bound

$$1/2 \;\leqslant\; |1-m^2|^{-1} \;\leqslant\; \varrho\,, \tag{3.4}$$

where the first inequality follows from (3.11) below, and the second from the identity $(1-m^2S)^{-1}\mathbf{e} = (1-m^2)^{-1}\mathbf{e}$ with the vector $\mathbf{e}:=(1,\dots,1)$. We therefore have the bounds $\Psi\leqslant 2\Phi\leqslant 2$.

In this section we sketch the proof of the following result.

Proposition 3.3 (Simple examples). *Suppose that $\Lambda\prec\Psi$ for some admissible control parameter Ψ. Then we have*

$$\frac{1}{N}\sum_a^{(\mu)} G_{\mu a}G_{a\mu} \;\prec\; \Psi^2\Phi\,,\qquad \frac{1}{N}\sum_a^{(\mu)} G_{\mu a}G^*_{a\mu} \;\prec\; \Psi^2 \tag{3.5}$$

[1] Here we use the notation $\|A\|_{\ell^\infty\to\ell^\infty} = \max_i\sum_j|A_{ij}|$ for the operator norm on $\ell^\infty(\mathbb{C}^N)$.

as well as

$$\frac{1}{N}\sum_{a}^{(\mu)} Q_a(G_{\mu a}G_{a\mu}) \prec \Psi^3\,, \qquad \frac{1}{N}\sum_{a}^{(\mu)} Q_a(G_{\mu a}G^*_{a\mu}) \prec \Psi^3\Phi\,. \tag{3.6}$$

In addition, we have the bounds

$$\frac{1}{N}\sum_{a}(G_{aa}-m) \prec \Psi\Phi\,, \qquad \frac{1}{N}\sum_{a} Q_a G_{aa} \prec \Psi^2\,. \tag{3.7}$$

Remark 3.4. As explained after (3.1), typically Φ and Ψ are comparable. In this case the right-hand sides of the estimates in (3.5) can be replaced with Ψ^3 and Ψ^2, those of (3.6) with Ψ^3 and Ψ^4, and those of (3.7) with Ψ^3 and Ψ^3. Thus we may keep track of the improving effect of the average using a simple power counting in the single parameter Ψ, replacing each Φ with a Ψ.

The significance of Proposition 3.3 is the following. The trivial bound $G_{a\mu} \prec \Psi$ (which follows immediately from $\Lambda \prec \Psi$) implies, for example, that $\frac{1}{N}\sum_a^{(\mu)} G_{\mu a}G_{a\mu} \prec \Psi^2$. The first estimate in (3.5) represents an improvement from Ψ^2 to $\Psi^2\Phi$. This improvement is due to the averaging over the index a of fluctuating quantities with almost vanishing expectation. We shall refer to such vertices as *charged*; see Definition 4.7 below. In contrast, there is no such improvement in the second estimate of (3.5), since $G_{\mu a}G^*_{a\mu} = |G_{\mu a}|^2$ is always positive. If we subtract the expectation (for technical reasons, we subtract only the partial expectation, i.e. take $Q_a = 1 - P_a$), then the averaging becomes effective and it improves the average of $G_{\mu a}G^*_{a\mu}$ by two orders, from Ψ^2 to $\Psi^3\Phi$. Interestingly, subtracting the expectation in the average of $G_{\mu a}G_{a\mu}$ does not improve the estimate further; compare the first bounds in (3.5) and (3.6). (In fact, we get the only slightly stronger bound Ψ^3 instead of $\Psi^2\Phi$.) These examples indicate that the improving effect of the averaging heavily depends on the structure of the resolvent monomials.

We shall be concerned with averages of more general expressions. Roughly, we consider arbitrary monomials in the resolvent entries (G_{ij}). Some of the indices are summed. The summation is always performed with respect to a *weight*, a nonnegative quantity which sums to one. In the examples of Proposition 3.3, the weight was N^{-1}. Generally, we want to allow weights consisting of factors N^{-1} as well as s_{ij}; recall that $\sum_j s_{ij} = \sum_j N^{-1} = 1$. Thus, in addition to (3.5), (3.6), and (3.7) we have for example the bounds

$$\sum_{a}^{(\mu)} s_{\nu a}G_{\mu a}G_{a\mu} \prec \Psi^2\Phi\,, \qquad \sum_{a}^{(\mu)} s_{\nu a}Q_a(G_{\mu a}G^*_{a\mu}) \prec \Psi^3\Phi\,, \qquad \sum_{a} s_{\nu a}(G_{aa}-m) \prec \Psi\Phi\,. \tag{3.8}$$

A slightly more involved average is

$$\sum_{a,b} s_{\mu a}s_{\rho b}\, Q_b\big(G_{\mu a}G_{ab}G^*_{b\nu}G^*_{ab}G_{\nu a}\big) \tag{3.9}$$

where μ, ν, and ρ are fixed. In Theorem 4.8 we shall see that (3.9) is stochastically dominated by $\Psi^6\Phi^2$. This means that the double averaging and the effect of one Q-operation amounts to an improvement of a power three, from the trivial bound Ψ^5 to $\Psi^6\Phi^2$. It may be tempting to think that each average and each factor Q improves the trivial bound by one power of Ψ or Φ, but this naive rule already fails in the some of the simplest examples in (3.5) and (3.6). The relation between the averaging structure and the improved power of Ψ and Φ is more intricate. Our final goal (see Theorem 4.8) is to establish an optimal result for general monomials, which takes into account the precise effect of all averages.

More generally, we shall be interested in averaging arbitrary monomials $\mathcal{Z}_{\mathbf{a}}$ in the resolvent entries. Each such monomial contains a family of *summation indices* $\mathbf{a}$ and *external indices* $\boldsymbol{\mu}$. In the example (3.9), we have

$$\mathcal{Z}_{\mathbf{a}} \;=\; G_{\mu a} G_{ab} G^*_{b\nu} G^*_{ab} G_{\nu a}\,, \qquad \mathbf{a} = (a,b)\,, \qquad \boldsymbol{\mu} = (\mu,\nu)\,. \tag{3.10}$$

The most convenient way to define such a monomial $\mathcal{Z}_{\mathbf{a}}$ is using a graph. The vertices are associated with the summation and external indices, and a resolvent entry G_{xy} is represented as a directed edge from vertex x to vertex y. We draw an edge associated with a resolvent entry G_{xy} with a solid line, and an edge associated with a resolvent entry G^*_{xy} with a dashed line. See Figure 3.1. As it turns out, the gain in powers of Ψ resulting from the averaging has a simple expression in terms of such graphs. Moreover, this graphical representation is a key tool in our proofs.

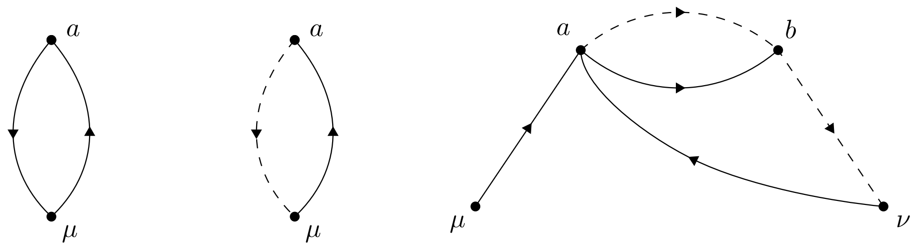

FIGURE 3.1. Graphs associated with the monomials (from left to right) $G_{\mu a}G_{a\mu}$, $G_{\mu a}G^*_{a\mu}$, and $G_{\mu a}G_{ab}G^*_{b\nu}G^*_{ab}G_{\nu a}$ (from (3.10)).

Note that neither the Q-factors nor the averaging weights are encoded in the graphical structure. Later we shall give a more precise definition of the class of weights we consider, but as an orientation to the reader, we emphasize that they play a secondary role. As long as the weights ensure an effective averaging over at least M values of each summation index, their final role is simply accounted for in the additional factor $M^{-1/2}\Psi^{-1}$ in the definition of Φ. The key improvement on the power of Ψ in the final estimate is solely determined by the structure of $\mathcal{Z}_{\mathbf{a}}$ and by the locations of the Q-factors.

3.1. Preliminaries. In this subsection we collect some basic facts that will be used throughout the paper. 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. For two positive quantities A_N and B_N we use the notation $A_N \asymp B_N$ to mean $C^{-1}A_N \leqslant B_N \leqslant C A_N$.

LEMMA 3.5. *There is a constant $c > 0$ such that for $E \in [-10, 10]$ and $\eta \in (0, 10]$*

$$c \;\leqslant\; |m(z)| \;\leqslant\; 1\,. \tag{3.11}$$

PROOF. See Lemma 4.2 in [17]. □

The following lemma collects basic algebraic properties of stochastic domination $\prec$.

LEMMA 3.6. *(i) Suppose that $X(u,v) \prec \Psi(u,v)$ uniformly in $u \in U$ and $v \in V$. If $|V| \leqslant N^C$ for some constant C then*

$$\sum_{v\in V} X(u,v) \;\prec\; \sum_{v \in V} \Psi(u,v)$$

uniformly in u.

(ii) *Suppose that $X_1(u) \prec \Psi_1(u)$ uniformly in u and $X_2(u) \prec \Psi_2(u)$ uniformly in u. Then*

$$X_1(u)X_2(u) \;\prec\; \Psi_1(u)\Psi_2(u)$$

uniformly in u.

(iii) *Suppose that $\Psi(u) \geqslant N^{-C}$ for all u and that for all p there is a constant C_p such that $\mathbb{E}|X(u)|^p \leqslant N^{C_p}$ for all u. Then, provided that $X(u) \prec \Psi(u)$ uniformly in u, we have*

$$P_a X(u) \;\prec\; \Psi(u) \qquad \text{and} \qquad Q_a X(u) \;\prec\; \Psi(u)$$

uniformly in u and a.

Proof. The claims (i) and (ii) follow from a simple union bound. The claim (iii) follows from Chebyshev's inequality, using a high-moment estimate combined with Jensen's inequality for partial expectation. We omit the details. □

We shall frequently make use of Schur's well-known complement formula, which we write as

$$\frac{1}{G_{ii}^{(T)}} \;=\; h_{ii} - z - \sum_{k,l}^{(Ti)} h_{ik} G_{kl}^{(Ti)} h_{li}\,, \tag{3.12}$$

where $i \notin T \subset \{1, \ldots, N\}$.

The following resolvent identities form the backbone of all of our calculations. The idea behind them is that a resolvent matrix entry G_{ij} depends strongly on the i-th and j-th columns of H, but weakly on all other columns. The first set of identities (called Family A) determines how to make a resolvent matrix entry G_{ij} independent of an additional index $k \neq i, j$. The second set of identities (Family B) expresses the dependence of a resolvent matrix entry G_{ij} on the matrix entries in the i-th or in the j-th column of H.

Lemma 3.7 (Resolvent identities). *For any Hermitian matrix H and $T \subset \{1, \ldots, N\}$ the following identities hold.*

Family A. *For $i, j, k \notin T$ and $k \neq i, j$, we have*

$$G_{ij}^{(T)} \;=\; G_{ij}^{(Tk)} + \frac{G_{ik}^{(T)} G_{kj}^{(T)}}{G_{kk}^{(T)}}\,, \qquad \frac{1}{G_{ii}^{(T)}} \;=\; \frac{1}{G_{ii}^{(Tk)}} - \frac{G_{ik}^{(T)} G_{ki}^{(T)}}{G_{ii}^{(T)} G_{ii}^{(Tk)} G_{kk}^{(T)}}\,. \tag{3.13}$$

Family B. *For $i, j \notin T$ satisfying $i \neq j$ we have*

$$G_{ij}^{(T)} \;=\; -G_{ii}^{(T)} \sum_{k}^{(Ti)} h_{ik} G_{kj}^{(Ti)} \;=\; -G_{jj}^{(T)} \sum_{k}^{(Tj)} G_{ik}^{(Tj)} h_{kj}\,, \tag{3.14a}$$

$$G_{ij}^{(T)} \;=\; G_{ii}^{(T)} G_{jj}^{(Ti)} \biggl(-h_{ij} + \sum_{k,l}^{(Tij)} h_{ik} G_{kl}^{(Tij)} h_{lj}\biggr)\,, \tag{3.14b}$$

$$\frac{1}{G_{ii}^{(T)}} \;=\; \frac{1}{m} - \bigl(-h_{ii} + Z_i^{(T)} + U_i^{(Ti)}\bigr)\,, \tag{3.14c}$$

where we defined

$$Z_i^{(T)} \;:=\; Q_i \sum_{k,l}^{(Ti)} h_{ik} G_{kl}^{(Ti)} h_{li}\,, \qquad U_i^{(S)} \;:=\; \sum_{k}^{(S)} s_{ik} G_{kk}^{(S)} - m\,. \tag{3.15}$$

Proof. The first identity of (3.13) was proved in Lemma 4.2 of [15]. The second identity of (3.13) is an immediate consequence of the first. The identities (3.14a) were proved in Lemma 6.10 of [6], and (3.14b) follows by iterating (3.14a) twice. Finally, (3.14c) (together with (3.15)) follows easily from (3.12), (2.7), the partition $1 = Q_i + P_i$, and the definition (2.1). □

Next, we record a simple estimate on resolvent entries of minors. For $T \subset \{1, \dots, N\}$ define the random variable

$$\Lambda^{(T)}(z) \;:=\; \max_{i,j \notin T} \Big| G_{ij}^{(T)}(z) - \delta_{ij} m(z) \Big|\,.$$

Lemma 3.8 (Bound on $\Lambda^{(T)}$). *Suppose that $\Lambda \prec \Psi$ for some admissible control parameter Ψ. Then for any fixed $\ell \in \mathbb{N}$ we have*

$$\Lambda^{(T)} \;\prec\; \Psi \tag{3.16}$$

provided that $|T| \leqslant \ell$. (The threshold $N_0(\varepsilon, D)$ in Definition 2.4 may also depend on ℓ).

Proof. See Appendix A. □

In particular, if $\Lambda \prec \Psi$ for some admissible Ψ, then Lemmas 3.8 and 3.5 imply that for any fixed $\ell \in \mathbb{N}$ we have

$$\frac{1}{G_{ii}^{(T)}} \;\prec\; 1 \tag{3.17}$$

provided that $|T| \leqslant \ell$. We conclude this section with rough bounds on the entries of G, which will be used to deal with exceptional, low-probability events.

Lemma 3.9 (Rough bounds on G). *Suppose that $\Lambda \prec \Psi$ for some admissible control parameter Ψ.*

(i) We have

$$\big| G_{ij}^{(T)}(z) \big| \;\leqslant\; \eta^{-1} \;\leqslant\; M \tag{3.18}$$

for all $z \in \mathbf{S}$, $T \subset \{1, \dots, N\}$, and $i, j \notin T$.

(ii) For every $p \in \mathbb{N}$ and $\ell \in \mathbb{N}$ there is a constant $C_{p,\ell}$ such that

$$\mathbb{E} \Big| 1/G_{ii}^{(T)}(z) \Big|^p \;\leqslant\; C_{p,\ell} \tag{3.19}$$

for all $T \subset \{1, \dots, N\}$ satisfying $|T| \leqslant \ell$, all $z \in \mathbf{S}$, and all $i \notin T$.

Proof. See Appendix A. □

3.2. Some ingredients of the proof of Proposition 3.3. A reader interested only in our main theorem (Theorem 4.8) may skip this section and proceed to Section 4 directly. Here we sketch the proof of Proposition 3.3. Our goal is to motivate some concepts underlying our main theorem, and to give an impressionistic overview of some ideas in its proof. The actual proof of Proposition 3.3 will not be needed, since Theorem 4.8 implies Proposition 3.3 as a special case.

To avoid needless complications in our proof, we additionally assume that we are dealing with one of the two classical symmetry classes of random matrices: real symmetric and complex Hermitian. For *real symmetric band matrices* we assume

$$\zeta_{ij} \in \mathbb{R} \quad \text{for all} \quad i \leqslant j \,. \tag{3.20}$$

For *complex Hermitian band matrices* we assume

$$\mathbb{E}\zeta_{ij}^2 = 0 \quad \text{for all} \quad i < j \,. \tag{3.21}$$

A common way to satisfy (3.21) is to choose the real and imaginary parts of ζ_{ij} to be independent with identical variance. In Remark 4.13 below we explain how to remove the assumption that (3.20) or (3.21) holds, i.e. how to remove the assumption $\mathbb{E}\zeta_{ij}^2 = 0$ in the case (3.21).

The second estimate of (3.5) follows trivially from $|G_{\mu a}| \leqslant \Lambda \prec \Psi$. We shall sketch the proofs of the remaining inequalities in the following order:

(A) first estimate of (3.6) and second estimate of (3.7);

(B) first estimate of (3.5) and first estimate of (3.7);

(C) second estimate of (3.6).

This order corresponds to an increasing degree of complication of the proofs. These three steps thus serve as simple examples in which to introduce four basic concepts underlying our proof. More specifically, in the language of the full proof (Sections 5 – 9), (A) requires only the simple high-moment estimate from Section 6, (B) requires in addition the inversion of a stable self-consistent equation (Section 7.2), and (C) requires in addition a priori bounds on chains (Sections 7.2 and 7.1) as well as the procedure of vertex resolution (Section 8).

3.2.1. Proof of (A). We focus first on the first estimate of (3.6). We derive the stochastic bound from high-moment bounds and Chebyshev's inequality. To simplify the presentation, we only estimate the variance

$$\mathbb{E}\biggl|\frac{1}{N}\sum_a^{(\mu)} Q_a(G_{\mu a}G_{a\mu})\biggr|^2 \;=\; \frac{1}{N^2}\sum_{a,b}^{(\mu)} \mathbb{E}\, Q_a\Bigl(G_{\mu a}G_{a\mu}\Bigr)\overline{Q_b\Bigl(G_{\mu b}G_{b\mu}\Bigr)}\,. \tag{3.22}$$

Our goal is to prove that (3.22) is bounded by $C\Psi^6$. We partition the summation into the cases $a = b$ and $a \neq b$. For the case $a = b$, we easily get from Lemmas 3.6 and 3.9 the bound $CN^{-1}\Psi^4 \leqslant C\Psi^6$, where we used (2.4) and the fact that Ψ satisfies Definition 2.9.

Let us therefore focus on the case $a \neq b$. We use (3.13) to get

$$\begin{aligned}
&\mathbb{E}\, Q_a(G_{\mu a}G_{a\mu})Q_b\overline{(G_{\mu b}G_{b\mu})} \\
&\quad = \mathbb{E}\, Q_a\biggl[\biggl(G^{(b)}_{\mu a} + \frac{G_{\mu b}G_{ba}}{G_{bb}}\biggr)\biggl(G^{(b)}_{a\mu} + \frac{G_{ab}G_{b\mu}}{G_{bb}}\biggr)\biggr] Q_b \overline{\biggl[\biggl(G^{(a)}_{\mu b} + \frac{G_{\mu a}G_{ab}}{G_{aa}}\biggr)\biggl(G^{(a)}_{b\mu} + \frac{G_{ba}G_{a\mu}}{G_{aa}}\biggr)\biggr]} \\
&\quad = \mathbb{E}\, Q_a\biggl[\biggl(G^{(b)}_{\mu a} + \frac{G^{(a)}_{\mu b}G_{ba}}{G^{(a)}_{bb}}\biggr)\biggl(G^{(b)}_{a\mu} + \frac{G_{ab}G^{(a)}_{b\mu}}{G^{(a)}_{bb}}\biggr)\biggr] Q_b \overline{\biggl[\biggl(G^{(a)}_{\mu b} + \frac{G^{(b)}_{\mu a}G_{ab}}{G^{(b)}_{aa}}\biggr)\biggl(G^{(a)}_{b\mu} + \frac{G_{ba}G^{(b)}_{a\mu}}{G^{(b)}_{aa}}\biggr)\biggr]} + \cdots\,,
\end{aligned} \tag{3.23}$$

where we dropped the higher order terms of the expansion. The philosophy behind this expansion is to make each resolvent entry independent of as many indices in (a, b) as possible by using (3.13) iteratively. We call such terms *maximally expanded* in (a, b), i.e. a maximally expanded resolvent entry cannot be made independent of a or b using the identity (3.13); the reason is that either it already has a and b as upper indices or an index from (a, b) appears as a lower index. See Definition 6.4 below for a precise statement. The iteration is stopped if either (3.13) cannot be applied to any resolvent entry or if a sufficient number of resolvent entries (in our case a total of six) have been generated. (In the proof of Proposition 6.3 we give a precise definition of this stopping rule.)

We now multiply everything out on the right-hand side of (3.23) to get terms of the form $\mathbb{E}Q_a(A)Q_b(B)$. The key observation is that if B is independent of a then the expectation vanishes (in fact, already the partial expectation P_a renders the whole term zero). Similarly, if A is independent of b then the expectation vanishes. An example of a leading-order term from (3.23) that does not vanish is

$$\mathbb{E}\, Q_a\biggl[\frac{G^{(a)}_{\mu b}G_{ba}}{G^{(a)}_{bb}}G^{(b)}_{a\mu}\biggr] Q_b \overline{\biggl[\frac{G^{(b)}_{\mu a}G_{ab}}{G^{(b)}_{aa}}G^{(a)}_{b\mu}\biggr]}. \tag{3.24}$$

(Note that all resolvent entries are maximally expanded in (a, b).) In this fashion each Q imposes the presence of at least one additional off-diagonal entry. Since every off-diagonal resolvent entry contributes a factor Ψ (see Lemma 3.8), we find that (3.23) is of order Ψ^6 instead of the naive Ψ^4. This concludes the sketch of the proof of the first estimate of (3.6).

The sketch of the proof of the second estimate of (3.7) is almost identical, and therefore omitted.

3.2.2. Introduction of graphs. Before moving on to (B) and (C), we take this opportunity to introduce a graphical language which is useful for keeping track of terms such as (3.24). Although not needed here, since the example in (A) is very simple, this language will prove essential when defining more complicated expressions, as well as for the actual proof of Theorem 4.8. Recall from Figure 3.1 that we can represent the expression $G_{\mu a}G_{a\mu}$ graphically by regarding μ and a as vertices, and by drawing two directed edges associated with $G_{\mu a}$ and $G_{a\mu}$. We adopt the convention given after (3.10). Thus, an off-diagonal resolvent entry of G_{ab}, $a \neq b$, is represented with a directed solid line from a to b, and the analogous entry G^*_{ab} with a directed dashed line from a to b.

Convention. We sometimes identify a vertex with its associated summation index, and hence use the letter a to denote two different things: a vertex of a graph and the value of the associated index. This allows us to avoid a proliferation of double subscripts in expressions like $G_{a_i a_j}$. When depicting graphs, we always label a vertex using the name of the associated index.

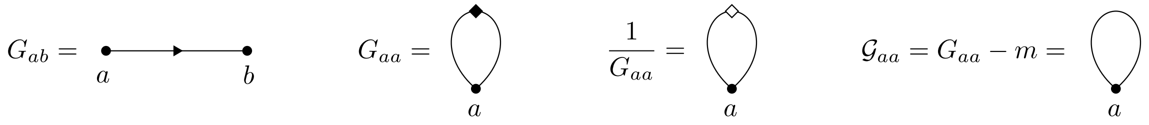

FIGURE 3.2. The graphical representations of resolvent entries. The versions associated with G^* are the same with a dashed line.

We shall also have to deal with diagonal resolvent entries; in fact we introduce separate notations the three most common functions of them. Our graphical conventions are summarized in Figure 3.2. We may thus represent the expression on the left-hand side of (3.23), i.e. $Q_a(G_{\mu a}G_{a\mu})Q_b(G^*_{\mu b}G^*_{b\mu})$, See Figure 3.3; note that our graphical notation does not keep track of the factors Q. Having drawn the graph in Figure 3.3, we start making all resolvent entries (corresponding to edges) independent of the indices a and b, using the identities (3.13). As explained above, this gives rise to a sum of terms, each one of which is a fraction of resolvent entries that are maximally expanded in (a, b). The denominator of each term contains diagonal resolvent entries, while its numerator is a product of off-diagonal resolvent entries; this follows from the structure of (3.13). A simple such example was given in (3.24). The associated monomial,

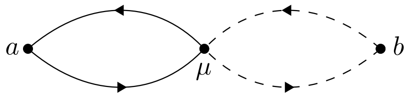

FIGURE 3.3. Graph associated with the monomial $G_{\mu a}G_{a\mu}G^*_{\mu b}G^*_{b\mu}$. Here we draw the case $a \neq b$.

$$\frac{G^{(a)}_{\mu b}G_{ba}}{G^{(a)}_{bb}}G^{(b)}_{a\mu}\,\frac{G^{(b)*}_{a\mu}G^*_{ba}}{G^{(b)*}_{aa}}G^{(a)*}_{\mu b}, \tag{3.25}$$

may be represented graphically as in Figure 3.4.

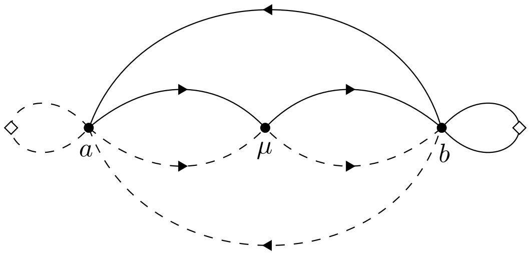

FIGURE 3.4. Graph associated with (3.25). Here we draw the case $a \neq b$.

We remark that the graphs depicted in Figures 3.3 and 3.4 are fundamentally different in the following sense. In Figure 3.3, each edge of the graph represents a resolvent entry with no upper indices; in Figure 3.4, each edge of the graph represents a resolvent entry that is maximally expanded in (a, b). In the language of Section 6, the former graph will be called $\gamma^2(\Delta)$ while the latter will be called Γ. It is the latter graphs that play a major role in our proofs. The former type is simply a trivial concatenation of basic graphs, and serves as an intermediate step in the construction of graphs of the latter type (i.e. whose edges represent maximally expanded resolvent entries). If one wanted to be more precise, one could keep track of the upper indices associated with each edge in the graphs. By definition, the edges of the graph in Figure 3.3 have no upper indices, and the edges of the graph in Figure 3.4 have upper indices as given in (3.25). However, these upper indices are unambiguously determined by the condition that each resolvent entry be maximally expanded in (a, b). This means that a appears as upper index of any edge that is not incident to a (and similarly with b). In practice, however, we do not indicate the upper indices, as they are uniquely determined by the condition that all edges are maximally expanded in (a, b).

It is possible, and indeed important for our proof, to introduce a graphical rule that generates graphs like the one depicted in Figure 3.4 from graphs like the one depicted in Figure 3.3 through a sequence of graphs whose edges are not yet maximally expanded. Before the maximal expansion is achieved, we shall temporarily indicate the upper indices on the graph edges in parenthesis. Recall that the underlying algebra was simply governed by the identities (3.13). Figure 3.5 depicts the identity

$$G_{ij} \;=\; G_{ij}^{(k)} + \frac{G_{ik}G_{kj}}{G_{kk}}\,. \tag{3.26}$$

Similarly, the corresponding identities for the diagonal entries,

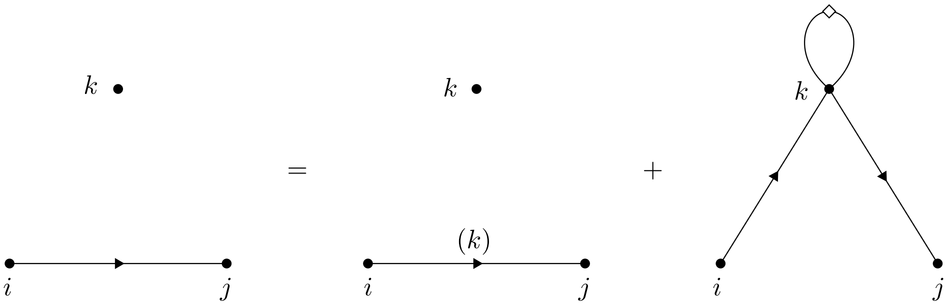

FIGURE 3.5. The graphical representation of the formula (3.26).

$$\frac{1}{G_{ii}} \;=\; \frac{1}{G_{ii}^{(j)}} - \frac{G_{ij}G_{ji}}{G_{ii}G_{ii}^{(j)}G_{jj}}\,, \qquad G_{ii} \;=\; G_{ii}^{(j)} + \frac{G_{ij}G_{ji}}{G_{jj}}\,, \tag{3.27}$$

are depicted in Figure 3.6. Applying the graphical rules of Figures 3.5 and 3.6 to Figure 3.3 results e.g. in Figure 3.4 (and many others). To be precise, we should keep track of the upper indices associated with each edge at each step, as is done in Figure 3.6. When all edges are maximally expanded, we stop the application of the rules of Figures 3.5 and 3.6. However, as explained above, we usually omit the explicit indication of upper indices in graphs after the maximal expansion is achieved. For future use, we record the following definition associated with the operations depicted in Figures 3.5 and 3.6.

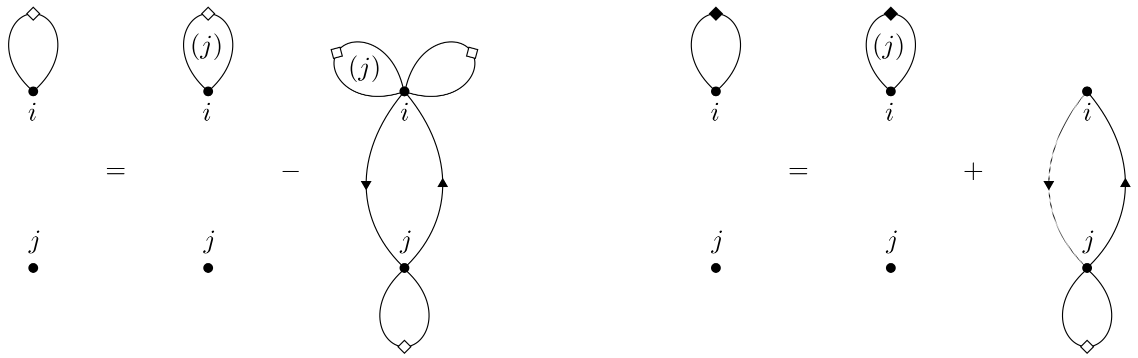

FIGURE 3.6. Adding an upper index j to the diagonal entries $1/G_{ii}$ and G_{ii}. These pictures correspond to (3.27). We exceptionally also mark the upper indices associated with each edge.

DEFINITION 3.10. We refer to the second graph on the right-hand side of Figure 3.5 as arising from *linking* the edge (ij) with the vertex k. We also say that the vertex k was *linked to* by the edge (ij). Similarly, in both connected graphs in Figure 3.6, the vertex j was linked to by the edge (ii).

The argument underlying (3.23) may now be formulated graphically as follows. We start from Figure 3.3 and apply the identities from Figures 3.5 and 3.6 until all resolvent entries associated with the edges are maximally expanded in (a, b). Since these identities can be applied in various orders, this procedure is not unique. This lack of uniqueness does not concern us, however: we need only a maximally expanded representation. By the argument given after (3.23), we know that only those graphs in which both a and b have been linked to by an edge survive. Such graphs (as the one from Figure 3.4) have (at least) two additional edges as compared to the one from Figure 3.3. This results in a size $O_\prec(\Psi^6)$.

3.2.3. Sketch of the proof of (B). We focus first on the first estimate of (3.5). The idea is to derive a stable self-consistent equation for the quantity

$$\frac{1}{N}\sum_a^{(\mu)} G_{\mu a}G_{a\mu}\,. \tag{3.28}$$

We do this by introducing the partition $1 = P_a + Q_a$ inside the summation. The second resulting term was estimated in (A). The first resulting term may be written as

$$\begin{aligned}
\frac{1}{N}\sum_a^{(\mu)} P_a\big(G_{\mu a}G_{a\mu}\big) &= \frac{1}{N}\sum_a^{(\mu)} P_a\bigg(\frac{m^2}{G_{aa}^2}G_{\mu a}G_{a\mu}\bigg) + O_\prec(\Psi^3) \\
&= m^2\,\frac{1}{N}\sum_a^{(\mu)} P_a\bigg(\sum_{x,y}^{(a)} G_{\mu x}^{(a)} h_{xa} h_{ay} G_{y\mu}^{(a)}\bigg) + O_\prec(\Psi^3) \\
&= m^2\,\frac{1}{N}\sum_a^{(\mu)}\sum_x^{(a)} s_{ax} G_{\mu x}^{(a)} G_{x\mu}^{(a)} + O_\prec(\Psi^3) \\
&= m^2\,\frac{1}{N}\sum_a^{(\mu)}\sum_x^{(a)} s_{ax} G_{\mu x} G_{x\mu} + O_\prec(\Psi^3) \\
&= m^2\,\frac{1}{N}\sum_x^{(\mu)} G_{\mu x} G_{x\mu} + O_\prec(\Psi^3 + N^{-1})\,.
\end{aligned}$$

In the first step we used (3.17). In the second step we used the identity (3.14a) (note the usefulness of smuggling in G_{aa} in the previous step). In the third step we used the identity $P_a h_{xa} h_{ay} = \mathbb{E} h_{xa} h_{ay} = s_{ax}\delta_{xy}$, as follows from the definition of H and the fact that $G^{(a)}$ is independent of a. In the fourth step we used the identity (3.13) to remove the upper indices. In the fifth step we used a simple analysis of coinciding indices together with the estimates (2.2) and (3.11). Together with the bound from (A), we therefore get for the quantity (3.28) the self-consistent equation

$$\begin{aligned}
\frac{1}{N}\sum_a^{(\mu)} G_{\mu a}G_{a\mu} &= \frac{1}{N}\sum_a^{(\mu)} P_a\big(G_{\mu a}G_{a\mu}\big) + O_\prec(\Psi^3) \\
&= m^2\frac{1}{N}\sum_x^{(\mu)} G_{\mu x}G_{x\mu} + O_\prec\big(\Psi^3 + M^{-1}\big) \\
&= m^2\frac{1}{N}\sum_x^{(\mu)} G_{\mu x}G_{x\mu} + O_\prec\Big(\Psi^2\big(\Psi + M^{-1/2}\Psi^{-1}\big)\Big)\,,
\end{aligned}$$

where in the last step we used (2.9). Using (3.4) and the trivial bound $\frac{1}{N}\sum_a^{(\mu)} G_{\mu a}G_{a\mu} \prec \Psi^2$ we therefore get

$$\frac{1}{N}\sum_a^{(\mu)} G_{\mu a}G_{a\mu} \;\prec\; \min\bigg\{\Psi^2\,,\,\frac{\Psi^2\big(\Psi + M^{-1/2}\Psi^{-1}\big)}{|1-m^2|}\bigg\} \;\leqslant\; \min\big\{\Psi^2\,,\,\varrho\Psi^2\big(\Psi + M^{-1/2}\Psi^{-1}\big)\big\} \;=\; \Psi^2\Phi\,,$$

which is the first estimate of (3.5).

The proof of the first estimate of (3.7) is similar, except that we derive the self-consistent equation using (3.14c) instead of (3.14a). Using the second estimate of (3.7) we find

$$\frac{1}{N}\sum_a (G_{aa} - m) \;=\; \frac{1}{N}\sum_a P_a(G_{aa} - m) + \frac{1}{N}\sum_a Q_a G_{aa} \;=\; \frac{1}{N}\sum_a P_a(G_{aa} - m) + O_\prec(\Psi^2)\,. \tag{3.29}$$

Next, from a simple large deviation estimate (see the first paragraph in the proof of Lemma 9.1 in Appendix A) we find $Z_a \prec \Psi$. Moreover, Lemma 3.8, (2.3), (2.2), and (2.9) readily imply that $U_a^{(a)} \prec \Psi$. Recalling the estimate (2.10), we may therefore expand the identity (3.14c) using (2.7) to get

$$G_{aa} \;=\; m + m^2(-h_{aa} + Z_a + U_a^{(a)}) + O_\prec(\Psi^2)\,.$$

Using $P_a h_{aa} = 0$ and $P_a Z_a = 0$ we therefore find

$$\begin{aligned}\frac{1}{N}\sum_a P_a(G_{aa} - m) \;&=\; m^2 \frac{1}{N}\sum_a P_a U_a^{(a)} + O_\prec(\Psi^2)\\ &=\; m^2 \frac{1}{N}\sum_a \sum_x^{(a)} s_{ax}(G_{xx}^{(a)} - m) + O_\prec(\Psi^2)\\ &=\; m^2 \frac{1}{N}\sum_x (G_{xx} - m) + O_\prec(\Psi^2)\,,\end{aligned}$$

where in the second step we recalled the definition (3.15) and used (2.2) as well as (2.3) to write $m = \sum_x^{(a)} s_{ax} m + O(M^{-1})$ with $M^{-1} = O(\Psi^2)$ by (2.9), and in the third (3.13) to get rid of the upper index a as well as (2.3). Thus, together with (3.29), we get the self-consistent equation

$$\frac{1}{N}\sum_a (G_{aa} - m) \;=\; m^2 \frac{1}{N}\sum_x (G_{xx} - m) + O_\prec(\Psi^2)\,, \tag{3.30}$$

from which we easily conclude the first estimate of (3.7) as before.

In both of the above examples the averaging was performed with respect to the uniform weight $w_a = N^{-1}$. We conclude by sketching the differences in the case of a nontrivial weight, e.g. $w_a = s_{\nu a}$. Consider for example the average $\sum_a s_{\nu a}(G_{aa} - m)$ from (3.8). Repeating the above derivation of (3.30), we find the self-consistent *system of equations*

$$\sum_a s_{\nu a}(G_{aa} - m) \;=\; m^2 \sum_a s_{\nu a} \sum_x s_{ax}(G_{xx} - m) + E_\nu$$

for each ν. Here the error satisfies $E_\nu = O_\prec(\Psi^2)$. Introducing the vectors $\mathbf{v} = (v_a)_{a=1}^N$ defined by $v_a := \sum_x s_{ax}(G_{xx} - m)$ and $\mathbf{E} = (E_\nu)_{\nu=1}^N$, we have

$$\mathbf{v} \;=\; m^2 S \mathbf{v} + \mathbf{E}\,.$$

Thus we find

$$\mathbf{v} \;=\; (1 - m^2 S)^{-1}\mathbf{E}\,,$$

from which we conclude that $v_a \prec \varrho\Psi^2 \leqslant \Psi\Phi$.

3.2.4. Sketch of the proof of (C). As in (A), the proof is based on a high-moment estimate. We again restrict our attention to the variance

$$\mathbb{E}\biggl|\frac{1}{N}\sum_a^{(\mu)} Q_a(G_{\mu a}G^*_{a\mu})\biggr|^2 \;=\; \frac{1}{N^2}\sum_{a,b}^{(\mu)} \mathbb{E}\, Q_a(G_{\mu a}G^*_{a\mu})\, Q_b(G_{\mu b}G^*_{b\mu})\,. \tag{3.31}$$

Our goal is to derive the stochastic bound $\Psi^6\Phi^2$ for (3.31). The case $a = b$ yields the bound

$$\frac{1}{N^2}\sum_a^{(\mu)}|G_{\mu a}|^4 \;\prec\; \Psi^4 N^{-1} \;\leqslant\; \Psi^6\Phi^2\,. \tag{3.32}$$

Let us therefore assume for the following that $a \neq b$. The first part of the argument follows precisely the proof of (A) above. We expand all resolvent entries of

$$\mathbb{E}\, Q_a(G_{\mu a}G^*_{a\mu})\, Q_b(G_{\mu b}G^*_{b\mu}) \tag{3.33}$$

using (3.13) and obtain a sum of monomials whose resolvent entries are maximally expanded in (a, b). A typical example of a nonvanishing term arising from the expansion of (3.31) is

$$\mathbb{E}\, Q_a\left(\frac{G^{(a)}_{\mu b}G_{ba}}{G^{(a)}_{bb}}G^{(b)*}_{a\mu}\right) Q_b\left(G^{(a)}_{\mu b}\frac{G^*_{ba}G^{(b)*}_{a\mu}}{G^{(a)*}_{aa}}\right).$$

As in the proof of (A), this immediately gives the stochastic bound Ψ^6. See Figure 3.7 for a graphical summary of the argument in this context.

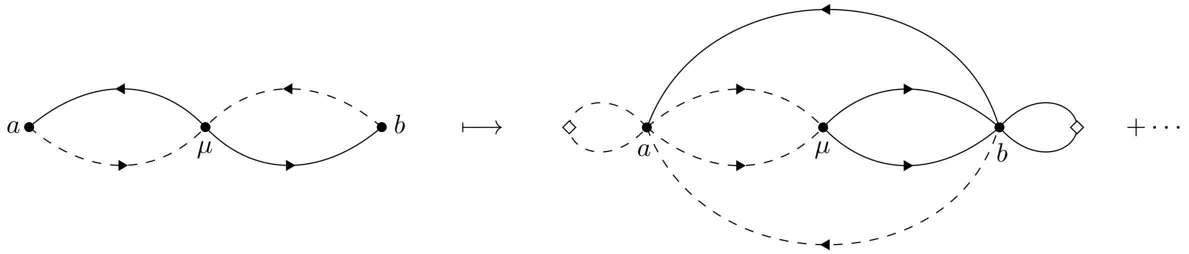

FIGURE 3.7. The process of making all edges of the graph associated with (3.33) maximally expanded in (a, b).

The bound Ψ^6 is not enough, however. In order to improve this to $\Psi^6\Phi^2$, we introduce a new operation which we call *vertex resolution*. In order to simplify the presentation, in the following we systematically replace any diagonal entry $G^{(T)}_{aa}$ by m. The resulting error terms are small by definition of Λ (of course, they have to be dealt with, which is done Section 9.1 of the full proof below). Thus, we have to estimate the expression

$$\mathbb{E}\, Q_a\big(G^{(a)}_{\mu b}G_{ba}G^{(b)*}_{a\mu}\big)\, Q_b\big(G^{(a)}_{\mu b}G^*_{ba}G^{(b)*}_{a\mu}\big) \tag{3.34}$$

for $a \neq b$. We begin by expanding all resolvent entries using the Family B identity (3.14a), again neglecting the diagonal prefactors in (3.14a). This gives

$$\sum_{x,y,z,w}^{(ab)}\;\sum_{x',y',z',w'}^{(ab)} \mathbb{E}\, Q_a\big(G^{(ab)}_{\mu x}h_{xb}h_{by}G^{(ab)}_{yz}h_{za}h_{aw}G^{(ab)*}_{w\mu}\big)\, Q_b\big(G^{(ab)}_{\mu x'}h_{x'b}h_{by'}G^{(ab)*}_{y'z'}h_{z'a}h_{aw'}G^{(ab)*}_{w'\mu}\big)\,. \tag{3.35}$$

(Here we also ignored a few special cases of coinciding indices when expanding both a and b in G_{ab} using (3.14a). As usual, the resulting terms are subleading and unimportant for this sketchy discussion.) The idea behind (3.35) is to expand all of the randomness that depends on a and b explicitly (i.e. in entries of H), so that partial expectations may be explicitly taken. Note that all resolvent entries in (3.35) are independent of a and b. We may now take the expectation $\mathbb{E}$ in (3.35); more precisely, we reorganize (3.35) as

$$\sum_{x,y,z,w}^{(ab)} \sum_{x',y',z',w'}^{(ab)} \mathbb{E}\, G_{\mu x}^{(ab)} G_{yz}^{(ab)} G_{w\mu}^{(ab)*} G_{\mu x'}^{(ab)} G_{y'z'}^{(ab)*} G_{w'\mu}^{(ab)*}\, P_a\Big[Q_a(h_{za}h_{aw})h_{z'a}h_{aw'}\Big] P_b\Big[h_{xb}h_{by}Q_b(h_{x'b}h_{by'})\Big]\,. \tag{3.36}$$

The two square brackets in (3.35) may be computed explicitly. Since $\mathbb{E}h_{uv} = 0$, each matrix entry h_{uv} must (at least) be paired with another copy of the same factor or its conjugate $\bar{h}_{uv} = h_{vu}$. Assume first that we are dealing with a complex Hermitian band matrix (condition (3.21)). In that case, each h_{uv} must be paired with its conjugate h_{vu} since $\mathbb{E}h_{uv}^2 = 0$. Of course, it may happen that more than two entries have coinciding indices, but this leads to a term that is subleading by a factor $M^{-1/2}$, and which we neglect here. Thus, h_{za} in (3.36) may be paired with h_{aw} (resulting in $z = w$) or with $h_{aw'}$ (resulting in $z = w'$). However, the pairing h_{za} with h_{aw} gives a vanishing contribution owing to the presence of Q_a, since

$$\mathbb{E}_{za} Q_a(h_{za}h_{az}) \;=\; 0$$

where $\mathbb{E}_{za}$ denotes partial expectation with respect to h_{za}. In other words, Q_a forbids the pairing of h_{za} with h_{aw}, and similarly Q_b the pairing of $h_{x'b}$ with $h_{by'}$. Thus the leading order term resulting from the square brackets in (3.36), on which we focus here, is the pairing

$$s_{az}s_{aw}s_{bx}s_{by}\,\delta_{wz'}\delta_{zw'}\delta_{x'y}\delta_{xy'}\,. \tag{3.37}$$

In the real symmetric case (condition (3.20)), where $\mathbb{E}h_{uv}^2$ does not vanish, h_{za} can also be paired with $h_{z'a}$. (Note that Q_a still forbids the pairing of h_{za} with h_{aw}.) This yields the three further allowed pairings

$$s_{az}s_{aw}s_{bx}s_{by}\delta_{zz'}\delta_{ww'}\delta_{x'y}\delta_{xy'}\,, \qquad s_{az}s_{aw}s_{bx}s_{by}\delta_{wz'}\delta_{zw'}\delta_{xx'}\delta_{yy'}\,, \qquad s_{az}s_{aw}s_{bx}s_{by}\delta_{zz'}\delta_{ww'}\delta_{xx'}\delta_{yy'}\,. \tag{3.38}$$

Assuming again condition (3.21), only (3.37) contributes, and we get the expression (up to lower order error terms in $M^{-1/2}$)

$$\begin{aligned}\sum_{x,y,z,w}^{(ab)} & s_{bx}s_{by}s_{az}s_{aw}\,\mathbb{E}\,G_{\mu x}^{(ab)}G_{yz}^{(ab)}G_{w\mu}^{(ab)*}G_{\mu y}^{(ab)}G_{xw}^{(ab)*}G_{z\mu}^{(ab)*} \\ &= \mathbb{E}\Bigg(\sum_{x,w} s_{aw}s_{bx}G_{\mu x}^{(ab)}G_{xw}^{(ab)*}G_{w\mu}^{(ab)*}\Bigg)\Bigg(\sum_{y,z} s_{by}s_{az}G_{\mu y}^{(ab)}G_{yz}^{(ab)}G_{z\mu}^{(ab)*}\Bigg)\,.\end{aligned} \tag{3.39}$$

Now each of the expressions in the parentheses is stochastically bounded by $\Psi^3\Phi$. Indeed, an argument very similar to the proof of (B) above yields

$$\sum_y s_{by}G_{\mu y}^{(ab)}G_{yz}^{(ab)} \;\prec\; \Psi^2\Phi\,. \tag{3.40}$$

(The additional upper indices (ab) are unimportant.) Thus, from the summation over y in (3.39) we gain an additional factor Φ (and, similarly, from the summation over w). We therefore find that (3.33) is stochastically bounded by $\Psi^6\Phi^2$, which was the claim of (C).

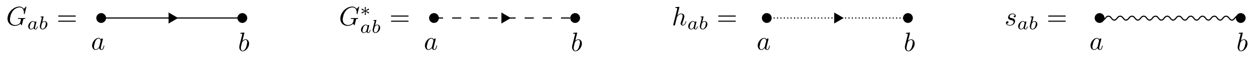

FIGURE 3.8. The graphical notation for entries of G, G^*, H, and S. Since S is symmetric, the edge associated with s_{ab} is undirected.

The use of graphs greatly clarifies the mechanism underlying the above sketch of the proof of (C). In order to depict vertex resolution, we need a graphical notation for edges associated with matrix entries h_{uv} and s_{ab}; we represent the former using dotted lines and the latter using wiggly lines. See Figure 3.8. As seen above, the starting point for the operation of vertex resolution is (3.34). The expanded expression (3.35) may be graphically represented as in Figure 3.9. Thus, the vertex a is "resolved" into four (i.e. the degree of a) new vertices, which are drawn in white and are connected to their parent vertex a by dotted lines (corresponding to a matrix entry of H). White vertices are either *incoming* or *outgoing*, depending on the orientation of the dotted edge that joins them to their parent vertex a. Similarly, the vertex b is resolved into four new vertices. Note that each solid or dashed edge in the right-hand graph of Figure 3.9 represents a resolvent matrix entry that is independent of a and b. The expression (3.39) was obtained from (3.35) by computing the partial expectations P_a and P_b of the associated entries of H. Graphically, this amounts to a pairing of the white vertices surrounding each black parent vertex. (Note that the factors Q, which yielded constraints on the allowed pairings, are not visible in the graphs. This is not a problem, however, as the ensuing bounds will hold for all pairings, even if these restrictions are relaxed.) The pairing of two dotted lines gives rise to a wiggly line, in accordance with the identity $\mathbb{E}_a|h_{ax}|^2 = s_{ax}$. See Figure 3.10 for a graphical representation of the pairing in (3.39). In Figure 3.10 we represented the pairing (3.37), which is the only one in the complex Hermitian case (3.21). In this case, the orientation of the edges must be matched when pairing white vertices, i.e. an incoming white vertex can only be paired with an outgoing one. This is an immediate consequence of the condition $\mathbb{E}h_{ax}^2 = 0$, as explained after (3.37). In the real symmetric case (3.20), where $\mathbb{E}_a|h_{ax}|^2 = \mathbb{E}h_{ax}^2 = s_{ax}$, the other pairings (3.38) are also possible. Graphically, this means that, when pairing white vertices, there are no constraints on the orientation of the incident edges. In other

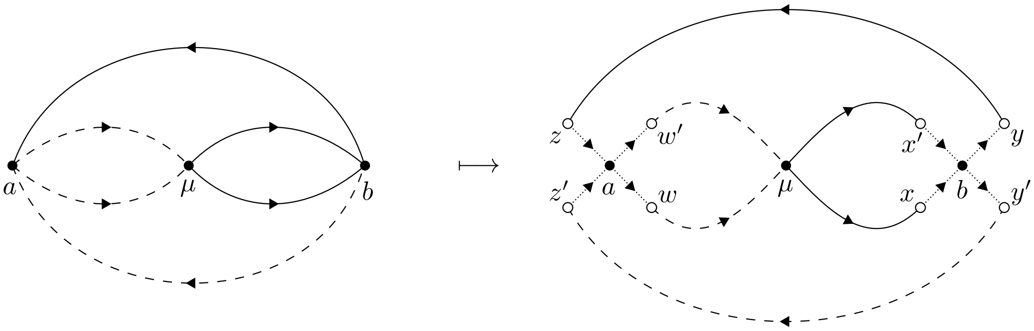

FIGURE 3.9. Resolving the vertices a and b. In accordance with (3.35), we ignore the loops associated with diagonal terms. (These lead to corrections that are higher order in Ψ.)

words, the arrows on the dotted edges may be ignored.

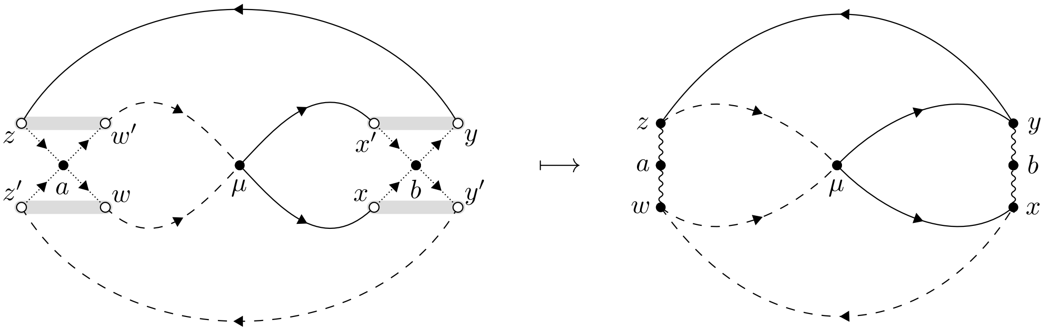

FIGURE 3.10. Taking a pairing of the white vertices to get the completed resolution.

The result of the vertex resolution is graphically evident when comparing the first graph in Figure 3.9 and second graph of Figure 3.10: the vertex a, of degree four, has been split (or "resolved") into two vertices of degree two. (The same happened for b). This *resolution* also entails the creation of new summation indices, z and w. Each one of them is connected to the original vertex a by a factor s_{az} (respectively s_{aw}), which implies that the summation over the larger family of summation indices is still performed with respect to a normalized weight. Generally, vertex resolution splits vertices of high degree into several vertices of degree two. The reason why this helps is that we can gain an extra factor Φ from any summation vertex of degree two whose incident edges are of the same "colour" (solid or dashed). We shall use the name *marked vertex* (see Definition 8.1 below) to denote a vertex whose resolution yields at least one new summation vertex whose (two) incident edges are of the same colour. The mechanism behind the gain of a factor Φ from a newly created (via resolution) index is roughly the content of (B), and was used in (3.40). In our case, we gain from the summations over y and w (but not z or x). Generally, the process of vertex resolution yields long "chains" (i.e. subgraphs whose vertices have degree two), each vertex of which yields an extra factor Φ provided both incident edges have the same colour. In fact, establishing such estimates for chains is an important step in our proof (see Proposition 5.3 below). This concludes our overview of the proof of Proposition 3.3.

4. General monomials and main result

In this section we state the fluctuation averaging theorem in full generality. To that end, we introduce a general class of monomials which we shall average. We consider monomials in the variables

$$\mathcal{G}_{ij}(z) \;:=\; G_{ij}(z) - \delta_{ij} m(z)\,,$$

which yield a more consistent power counting for diagonal resolvent entries. Indeed, by definition $|\mathcal{G}_{ij}| \leqslant \Lambda$ for all i and j. As we saw in Section 3, monomials in the resolvent entries are best described using graphs; see (3.10) and Figure 3.1.

We may now define the graphs Δ used to describe monomials.

Definition 4.1 (Admissible graphs). (i) Let V_s and V_e be finite disjoint sets. Let $V := V_s \sqcup V_e$ be their disjoint union[2] and E be a subset of the ordered pairs $V \times V$. The quadruple

$$\Delta \;=\; (V_s, V_e, E, \xi)$$

is an *admissible graph* if it is a directed, edge-coloured, multigraph with set of vertices V. The edges are ordered pairs of vertices with multiplicity, i.e. we allow loops and multiple edges. We shall also use the notation $V_s \equiv V_s(\Delta)$, $V_e \equiv V_e(\Delta)$, and $E \equiv E(\Delta)$.

More formally, we can view $E(\Delta)$ as an arbitrary finite set equipped with maps $\alpha, \beta : E(\Delta) \to V(\Delta)$. Here $\alpha(e)$ and $\beta(e)$ represent the source and target vertices of the edge $e \in E(\Delta)$. The colouring $\xi : E(\Delta) \to \{1, *\}$ is a mapping that assigns one of two "colours", 1 or $*$, to each edge. If no confusion is possible with the multiplicity of an edge $e \in E(\Delta)$, we shall identify it with the ordered pair $(\alpha(e), \beta(e))$.

(ii) We denote by $\mathfrak{Z}$ the set of admissible graphs Δ on arbitrary V_s and V_e.

(iii) The *degree* of $\Delta \in \mathfrak{Z}$ is

$$\deg(\Delta) \;:=\; |E(\Delta)|\,.$$

The set $V_s(\Delta)$ will label the family of summation indices ($(a,b) = (a_i)_{i \in V_s(\Delta)}$ in the example (3.9)), and $V_e(\Delta)$ the set of external indices ($(\mu,\nu) = (\mu_i)_{i \in V_e(\Delta)}$ in the example (3.9)). We use the notation

$$\mathbf{u} \;=\; (\mathbf{a}, \boldsymbol{\mu})\,, \qquad \mathbf{a} \;=\; (a_i)_{i \in V_s(\Delta)}\,, \qquad \boldsymbol{\mu} \;=\; (\mu_i)_{i \in V_e(\Delta)} \tag{4.1}$$

for the matrix indices. Generally, we try to use Latin letters $a, b, c, d, x, y, z, \dots$ for summation indices and Greek letters $\mu, \nu \dots$ for external indices.

Although our statements and proofs hold for any admissible graph Δ, in order to avoid trivial cases in our applications we shall always consider graphs without isolated vertices and with the property that each edge is incident to at least one vertex $V_s(\Delta)$, i.e. every resolvent entry contains at least one summation index.

Next, we introduce the monomials in $(\mathcal{G}_{xy})$ whose average we shall estimate.

Definition 4.2 (Monomials). Let $\Delta \in \mathfrak{Z}$ be an admissible graph and let $\boldsymbol{\mu} \in \{1, \dots, N\}^{V_e(\Delta)}$ be a collection of external indices. We define the monomial

$$\mathcal{Z}_{\mathbf{a}} \;\equiv\; \mathcal{Z}_{\mathbf{a}}^{\boldsymbol{\mu}}(\Delta) \;:=\; \prod_{e \in E(\Delta)} \mathcal{G}^{\xi_e}_{u_{\alpha(e)} u_{\beta(e)}} \tag{4.2}$$

which is regarded as a function of the summation indices $\mathbf{a}$, recalling the splitting of the indices (4.1). We also denote by

$$\mathcal{Z} \;=\; \big(\mathcal{Z}_{\mathbf{a}}^{\boldsymbol{\mu}}(\Delta) : \mathbf{a} \in \{1, \dots, N\}^{V_s(\Delta)}\big)$$

the family of monomials associated with $(\Delta, \boldsymbol{\mu})$ by (4.2), and say that Δ *encodes* the monomial $\mathcal{Z}_{\mathbf{a}}$.

Note that $\deg(\Delta)$ is the degree of the monomial $\mathcal{Z}_{\mathbf{a}}$ encoded by Δ. Throughout the following we shall frequently drop the explicit dependence of $\mathcal{Z}_{\mathbf{a}}$ on $\boldsymbol{\mu}$ and Δ.

[2] Here, and throughout the following, we use the symbol $\sqcup$ to denote disjoint union.

The averaging over $\mathbf{a}$ will be performed with respect to a weight $w(\mathbf{a})$. In the example (3.9), this weight was $w(a,b) = s_{\mu a} s_{\rho b}$. A typical example of a weight is

$$w(a,b,c) \;=\; \frac{1}{N} \sum_d s_{\mu d} s_{db} s_{bc} \qquad \text{for} \quad \mathbf{a} = (a,b,c)\,. \tag{4.3}$$

In order to define a general class of weights, the following notion of partitioning of summation indices is helpful.

Definition 4.3 (Partition of indices). Let I be a finite index set. For $\mathbf{a} = (a_i)_{i \in I} \in \{1, \dots, N\}^I$ we denote by $\mathcal{P}(\mathbf{a})$ the partition of I defined by the equivalence relation $k \sim l$ if and only if $a_k = a_l$.

Generally, we consider weights satisfying the following definition; when reading it, it is good to keep examples of the type (4.3) in mind.

Definition 4.4 (Weights). A map $w : \{1, \dots, N\}^{V_s(\Delta)} \to [0,1]$ is a *weight adapted to* $\Delta \in \mathfrak{Z}$ if it satisfies the following condition. Let $V_s(\Delta) = I \sqcup J$ be a (possibly trivial) partition of $V_s(\Delta)$ into two disjoint subsets, inducing a splitting $\mathbf{a} = (\mathbf{a}_I, \mathbf{a}_J)$ of the summation indices. Then we require that, for any partition P of J, we have

$$\max_{\mathbf{a}_I} \sum_{\mathbf{a}_J} \mathbf{1}\big(\mathcal{P}(\mathbf{a}_J) = P\big)\, w(\mathbf{a}_I, \mathbf{a}_J) \;\leqslant\; M^{|P| - |V_s(\Delta)|}\,, \tag{4.4}$$

where $|P|$ denotes the number of blocks in P.

The interpretation of (4.4) is that the left-hand side of (4.4) has $|P|$ free summation indices; the remaining summation indices have been either frozen (i.e. they belong to $\mathbf{a}_I$) or merged with others (i.e. they belong to a nontrivial block of P). Then (4.4) simply states that each suppressed summation yields a factor M^{-1}. In particular, with $J = V_s(\Delta)$ and the trivial atomic partition P we have

$$\sum_{\mathbf{a}} w(\mathbf{a}) \;\leqslant\; 1\,,$$

i.e. the total sum of all weights is always bounded by one.

When estimating averages such as (3.9), we shall always impose that all indices that have distinct names also have distinct values. In the case that two indices have the same value, we give them the same name. Thus, for example we write

$$\begin{aligned}
&\frac{1}{N^2} \sum_{a,b} G_{\mu a} G_{ab} G_{b \mu} \\
&\;=\; \frac{1}{N^2} \sum_{a,b}^{(\mu)*} G_{\mu a} G_{ab} G_{b\mu} + \frac{1}{N^2} \sum_{a}^{(\mu)} G_{\mu a} G_{aa} G_{a\mu} + \frac{1}{N^2} \sum_{a}^{(\mu)} G_{\mu a} G_{a\mu} G_{\mu\mu} + \frac{1}{N^2} \sum_{b}^{(\mu)} G_{\mu\mu} G_{\mu b} G_{b\mu} + \frac{1}{N^2} G_{\mu\mu}^3\,,
\end{aligned}$$

where a star on top of a summation means that all summation indices are constrained to be distinct. (Recall also the notation $\sum^{(S)}$ for $S \subset \{1, \dots, N\}$ from Definition 2.2.)

We may now define our central quantity. Let $\Delta \in \mathfrak{Z}$ and $\boldsymbol{\mu} \in \{1, \dots, N\}^{V_e(\Delta)}$ be a collection of external indices. Let $F \subset V_s(\Delta)$ and w be a weight adapted to Δ. We define

$$X_F^w(\Delta) \;\equiv\; X_F^{w, \boldsymbol{\mu}}(\Delta) \;:=\; \sum_{\mathbf{a}}^{(\boldsymbol{\mu})*} w(\mathbf{a}) \left[\prod_{i \in F} Q_{a_i}\right] \mathcal{Z}_{\mathbf{a}}^{\boldsymbol{\mu}}(\Delta)\,. \tag{4.5}$$

Thus, F denotes the set of summation indices that come with an operator Q. As explained above, the symbol $(\boldsymbol{\mu})$ on top of them sum means that $a_i \neq \mu_j$ for all $i \in V_s(\Delta)$ and $j \in V_e(\Delta)$, and the star means that $a_i \neq a_j$ for all distinct $i, j \in V_s(\Delta)$. Throughout the following, we shall frequently drop the explicit dependence of $X_F^{w,\boldsymbol{\mu}}(\Delta)$ on $\boldsymbol{\mu}$.

Remark 4.5. In (4.5) each operator Q acts on all resolvent entries in $\mathcal{Z}_{\mathbf{a}}$. We make this choice to simplify the presentation; also, this is sufficient for all of our current applications. However, our results may be easily extended to more complicated quantities, in which each Q acts only on a subset of the resolvent entries in $\mathcal{Z}_{\mathbf{a}}$. Thus, in general, there a resolvent entry is either *outside* or *inside* Q_{a_i}, for each $i \in F$. We require that each resolvent entry outside Q_{a_i} have no index a_i, and at least one resolvent entry inside Q_{a_i} have an index a_i. Then our proof carries over with merely cosmetic changes. For example, expressions such as

$$\sum_{a,b,c,d}^{(\mu\nu)*} s_{\mu a} s_{\rho b} s_{bc}\, Q_a \Big(G_{\mu a} Q_b (G_{ab} G^*_{b\nu}) Q_c (G^*_{ac} G_{cd}) G_{d\mu} \Big)$$

may be estimated in this fashion.

From Lemmas 3.6 and 3.9, we find the trivial bound

$$X_F^w(\Delta) \;\prec\; \Psi^{\deg(\Delta)} \tag{4.6}$$

for any adapted weight w, provided that $\Lambda \prec \Psi$. We call (4.6) trivial because we also have the bound

$$\mathcal{Z}_{\mathbf{a}}^{\boldsymbol{\mu}}(\Delta) \;\prec\; \Psi^{\deg(\Delta)}\,.$$

Hence the estimate (4.6) has not been improved by the averaging over $\mathbf{a}$.

Next, we define indices which count the gain in the size of $X_F^w(\Delta)$ resulting from the averaging over $\mathbf{a}$ and from the factors Q.

Definition 4.6. Let Δ be an edge-coloured graph as in Definition 4.1. For $i \in V(\Delta)$ we set

$$\begin{aligned}
\nu_i(\Delta) \;&:=\; \sum_{(j,k)\in E(\Delta)} \mathbf{1}(\xi_{(j,k)} = 1) \Big[\mathbf{1}(i = j) + \mathbf{1}(i = k) \Big]\,, \\
\nu_i^*(\Delta) \;&:=\; \sum_{(j,k)\in E(\Delta)} \mathbf{1}(\xi_{(j,k)} = *) \Big[\mathbf{1}(i = j) + \mathbf{1}(i = k) \Big]\,.
\end{aligned}$$

Informally, $\nu_i(\Delta)$ is the number of legs of colour 1 incident to i, and $\nu_i^*(\Delta)$ the number of legs of colour $*$ incident to i.

We shall use $\deg(i) \equiv \deg_\Delta(i)$ to denote the degree of the vertex $i \in V(\Delta)$. It is sometimes important to emphasize that this degree is computed with respect to the graph Δ, which we indicate using the subscript[3] Δ. By definition, $\deg_\Delta(i)$ is the number of legs incident to i, i.e. a loop at i counts twice. In particular, $\deg_\Delta(i) = \nu_i(\Delta) + \nu_i^*(\Delta)$.

In terms of the monomials $\mathcal{Z}$ encoded by Δ, the index $\nu_i(\Delta)$ (respectively $\nu_i^*(\Delta)$) is the number of resolvent entries of $\mathcal{G}$ (respectively of $\mathcal{G}^*$) in which the index a_i appears. (Note that if the index a_i appears twice in a resolvent entry, this entry is counted twice.)

[3] Of course, $\deg_\Delta$ is not the same as $\deg(\Delta)$. In fact, we have $\deg(\Delta) = \frac{1}{2}\sum_{i\in V(\Delta)} \deg_\Delta(i)$.

Definition 4.7 (Charged vertex). We call a summation vertex $i \in V_s(\Delta)$ *charged* if either

(i) $i \notin F$ and $\nu_i \neq \nu_i^*$, or

(ii) $i \in F$ and $|\nu_i - \nu_i^*| \neq 2$.

We denote by $V_c(\Delta) \subset V_s(\Delta)$ the set of charged vertices.

We may now state our main result.

Theorem 4.8 (Averaging theorem). *Suppose that $\Lambda \prec \Psi$ for some admissible control parameter Ψ. Let $\Delta \in \mathfrak{Z}$ (recall Definitions 4.1 and 4.2) and $F \subset V_s(\Delta)$. Then*

$$X_F^{w,\boldsymbol{\mu}}(\Delta) \;\prec\; \Psi^{\deg(\Delta)+|F|}\,\Phi^{|V_c(\Delta)|} \tag{4.7}$$

for any $\boldsymbol{\mu}$ and weight w adapted to Δ (recall Definition 4.4).

Thus, Theorem 4.8 states that we gain a factor Ψ from each Q and a factor Φ from each charged vertex. The rationale behind the name "charged" is that, in the vertex resolution process from the proof of Theorem 4.8, a charged vertex gives rise, in leading order, to a collection vertices of degree two, at least one of which will be a *chain vertex* (see Definition 5.1) and hence yield a factor Φ using the a priori bounds of Section 7.

Remark 4.9. The right-hand side of (4.7) can be estimated from above by

$$(\Psi + M^{-1/4})^{\deg(\Delta)+|F|+|V_c(\Delta)|}\,,$$

which gives a simple power counting in terms of the quantity $\Psi + M^{-1/4}$. From each summation index a_i without an associated Q_{a_i} we gain a factor $\Psi + M^{-1/4}$ if $\nu_i \neq \nu_i^*$. If there is a Q_{a_i} then we gain at least a factor $\Psi + M^{-1/4}$, and, provided that $|\nu_i - \nu_i^*| \neq 2$, one additional factor $\Psi + M^{-1/4}$. Note that we gain at most two additional factors $\Psi + M^{-1/4}$ from each summation index.

Remark 4.10. As explained after (3.6), the additional term $M^{-1/2}\Psi^{-1}$ in the definition of Φ is a (necessary) technical nuisance and should be thought of as a lower order term in typical applications. In general, however, it cannot be eliminated, and Theorem 4.8 cannot be formulated in terms of powers of Ψ alone. This may be seen for instance from the variance calculation of the quantity $\frac{1}{N}\sum_a^{(\mu)} Q_a(G_{\mu a}G_{a\mu}^*)$. Indeed, as is apparent from (3.32), the term arising from $a = b$ is of order $N^{-1}\Psi^4$, which is in general not bounded by Ψ^8.

Remark 4.11. The requirement that (2.5) hold for all p can be easily relaxed. Indeed, Theorem 4.8 has the following variant. Fix $\varepsilon > 0$ and $D > 0$. Then there exists a $p(\varepsilon, D) \in \mathbb{N}$ such that the following holds. Suppose that the hypotheses of Theorem 4.8 hold, and that (2.5) holds for $p(\varepsilon, D)$. Then

$$\mathbb{P}\Big[|X_F^{w,\boldsymbol{\mu}}(\Delta)| > N^{\varepsilon}\,\Psi^{\deg(\Delta)+|F|}\,\Phi^{|V_c(\Delta)|}\Big] \;\leqslant\; N^{-D}\,,$$

for all $z \in \mathbf{S}$, all $\boldsymbol{\mu}$, and all weights w adapted to Δ.

This variant is an immediate consequence of the proof of Theorem 4.8, using the observation that, for any fixed ε and D, the estimate on $X_F^w(\Delta)$ consists of a finite number of steps s, each of them using a bound on $\mathbb{E}|\zeta_{ij}|^{p_s}$ for some finite p_s. As $\varepsilon \to 0$ or $D \to \infty$, the number of these steps tends to infinity. Moreover, as the step index s tends to infinity, the exponent p_s in $\mathbb{E}|\zeta_{ij}|^{p_s}$ also tends to infinity.

Remark 4.12. Our result applies verbatim if (some or all) diagonal entries of the form $\mathcal{G}_{ii} = G_{ii} - m$ in the monomial (4.2) are replaced by $1/G_{ii} - 1/m$. (This would be a mere notational complication in the statement of Theorem 4.8). After a little algebra (multiplying out a product of terms of the form $1/G_{ii} - 1/m$), we consequently find that our result applies to monomials divided by diagonal entries G_{ii}, i.e. expressions of the form

$$\frac{\mathcal{G}_{xy}\mathcal{G}_{uv}\mathcal{G}_{wz}\cdots}{G_{aa}G_{bb}G_{cc}\cdots}\,,$$

where the indices can be either summation or external indices. This extension may be proved in two ways.

The first way is to observe that if we replace the identity

$$G_{ii} - m \;=\; \frac{1}{1/m - \bigl(-h_{ii} + Z_i + U_i^{(i)}\bigr)} - m \;=\; m^2\bigl(-h_{ii} + Z_i + U_i^{(i)}\bigr) + m^3\bigl(-h_{ii} + Z_i + U_i^{(i)}\bigr)^2 + \cdots\,,$$

used in our proof by (3.14c) for the quantity $1/G_{ii} - 1/m$, the proof of Theorem 4.8 carries over unchanged.

The second way is to write

$$\frac{1}{G_{ii}} - \frac{1}{m} \;=\; \frac{m - G_{ii}}{m^2} + \frac{(m - G_{ii})^2}{m^3} + \frac{(m - G_{ii})^3}{m^3 G_{ii}}\,.$$

This induces a splitting of $\mathcal{Z}$ into three parts, which are treated separately. It is a simple matter to check that Theorem 4.8 may be applied to the first two parts. The third part is treated trivially, by freezing the index i; in this case we already get a factor Ψ^3 from the index i, and hence the averaging effect of the summation over i is not needed, since we already gained the maximal two additional factors of Ψ from i.

Remark 4.13. As in Section 3, in our proofs we shall assume that either (3.20) or (3.21) holds (see Section 3.2). We impose these conditions in order to simplify the derivation and analysis of self-consistent equations such as the ones in Sections 3.2.3 and 7.1. Without them, however, our core argument remains unchanged. For instance, when estimating $\sum_a s_{ba} G_{\mu a} G_{\mu a}$, we instead consider the quantity $V_a := P_a G_{\mu a} G_{\mu a}$. Using (3.14a), we may do a calculation similar to the one following (3.28), and get a self-consistent equation for V_a. Solving the self-consistent equation entails the analysis of the Hermitian operator $R = (r_{ij})$ where $r_{ij} := \mathbb{E} h_{ij}^2$. Using $|r_{ij}| \leqslant s_{ij}$, the spectral analysis from the end of Section 7.2 and Appendix A carries over with minor modifications. We omit the extraneous details of this generalization.

Remark 4.14. In [17, Lemma 5.2], a fluctuation averaging theorem of the form

$$\frac{1}{N}\sum_i Q_i \sum_{k,l}^{(i)} h_{ik} G_{kl}^{(i)} h_{li} \;\prec\; \Psi^2 \tag{4.8}$$

was proved. This result was further generalized in [5, 16, 18]. The estimate (4.8) also follows from Theorem 4.8. To see this, we use Schur's formula (3.12) to get

$$\frac{1}{N}\sum_i Q_i \frac{1}{G_{ii}} \;=\; \frac{1}{N}\sum_i h_{ii} - \frac{1}{N}\sum_i Q_i \sum_{k,l}^{(i)} h_{ik} G_{kl}^{(i)} h_{li}\,. \tag{4.9}$$

The first term on the right-hand side of (4.9) is easily proved to be stochastically bounded by $N^{-1} \leqslant \Psi^2$. The second term on the right-hand side of (4.9) is the left-hand side of (4.8). Moreover, the left-hand side of (4.9) is stochastically bounded by Ψ^2, as follows from Theorem 4.8; see Remark 4.12. In fact, the left-hand side of (4.9) may be estimated using the much simpler Proposition 6.1 (whose proof trivially holds for expressions like the one the left-hand side of (4.9)). In particular, Proposition 6.1 and this remark provide a simpler proof than [5, 16–18] of the previously known estimate (4.8).

Theorem 4.8 has the following, simpler, variant in which the averaging with respect to a weight w is replaced with partial expectation.

Theorem 4.15 (Averaging using partial expectation). *Suppose that $\Lambda \prec \Psi$ for some admissible control parameter Ψ. Let $\Delta \in \mathfrak{Z}$ and $F = \emptyset$. Then*

$$\prod_{a \in \mathbf{a}} P_a \, \mathcal{Z}^{\boldsymbol{\mu}}_{\mathbf{a}}(\Delta) \;\prec\; \Psi^{\deg(\Delta)} \, \Phi^{|V_c(\Delta)|} \tag{4.10}$$

for all $\mathbf{a}$ and $\boldsymbol{\mu}$ such that all indices of the collection $(\mathbf{a}, \boldsymbol{\mu})$ are distinct.

Thus in Theorem 4.15 we set $F = \emptyset$, i.e. there are no factors Q, whose presence would be nonsensical because the identity $P_a Q_a = 0$ implies that the partial expectation of any monomial preceded by a factor Q vanishes. The condition $F = \emptyset$ is still used indirectly in the theorem since the definition of $V_c(\Delta)$ depends on F.

Remark 4.16. It is possible to combine Theorems 4.8 and 4.15 by splitting $\mathbf{a} = (\mathbf{a}', \mathbf{a}'')$, and averaging over $\mathbf{a}'$ with respect to a weights $w(\mathbf{a}')$ and taking the partial expectation $\prod_{a \in \mathbf{a}''} P_a$ over $\mathbf{a}''$. We omit the details.

Remark 4.17. Remarks 4.9 – 4.13 also apply to Theorem 4.15 with the obvious modifications.

5. Outline of proof

We now outline the strategy behind the proof of Theorem 4.8. The first part of the proof relies on an inductive argument to prove the claim of Theorem 4.8 for a special class of Δ's (the *chains*) that encode monomials containing only factors $\mathcal{G}$ and not $\mathcal{G}^*$ (or the other way around). These Δ's act as building blocks which are used to estimate the error terms arising in the estimate of arbitrary Δ's, in the second part of the proof. The need to have a priori bounds on chains was already hinted at in Section 3.2. Indeed, the estimate (3.40) is the simplest prototype of a chain estimate, and was used to estimate quantities arising from the process of vertex resolution. This is in fact a general phenomenon: a priori bounds on chains will be used used in combination with vertex resolution.

Definition 5.1 (Chains). Let $\Delta \in \mathfrak{Z}$.

(i) We call a vertex $i \in V_s(\Delta)$ a *chain vertex* if i is not adjacent to itself, i has degree two, and both incident edges have the same colour. We denote by $c(\Delta)$ the number of chain vertices in Δ.

(ii) We call Δ an *open (undirected) chain* if all vertices $i \in V_s(\Delta)$ are chain vertices, $|V_e(\Delta)| = 2$, and $\deg(i) = 1$ for both $i \in V_e(\Delta)$.

(iii) We call Δ a *closed (undirected) chain* if all vertices $i \in V_s(\Delta)$ are chain vertices, $|V_e(\Delta)| \leqslant 1$, and $\deg(i) = 2$ for $i \in V_e(\Delta)$.

(iv) A chain vertex $i \in V_s(\Delta)$ is *directed* if one incident edge is incoming and the other outgoing. A chain is *directed* if every $i \in V_s(\Delta)$ is directed.

Figure 5.1 gives a few examples of chains. The notion of a directed chain will be used in the complex Hermitian case (3.21), in which all chains that arise in our proof will be directed. In the real symmetric case (3.20), there is no such restriction.

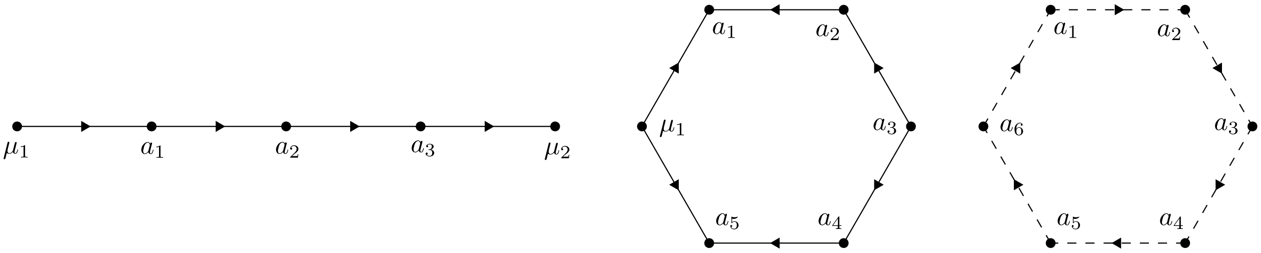

FIGURE 5.1. From left to right: an open directed chain, a closed undirected chain with one external vertex, a closed directed chain with no external vertices.

If Δ is a chain then by definition $X^w_F(\Delta)$ contains no diagonal entries $\mathcal{G}_{ii}$. Since $\mathcal{G}_{ij} = G_{ij}$ for $i \neq j$, we may (and shall) therefore replace all entries of $\mathcal{G}$ with entries of G when Δ is a chain.

Chains are useful in combination with the following family of special weights.

DEFINITION 5.2 (CHAIN WEIGHTS). Let $n \in \mathbb{N}$. For any fixed $\mathbf{b} = (b_1, \dots, b_n)$ define the weight

$$w_{\mathbf{b}}(\mathbf{a}) \;\equiv\; w(\mathbf{a}) \;:=\; s_{a_1 b_1} \cdots s_{a_n b_n}\,. \tag{5.1}$$

We call weights of the form (5.1) *chain weights*.

Using (2.2), it is easy to check that a chain weight from Definition 5.2 is a weight in the sense of Definition 4.4. The role of chains is highlighted by the two following facts.

- If Δ is a chain and $w_{\mathbf{b}}$ is an adapted chain weight, then the family $\big(\sum_{\mathbf{a}} w_{\mathbf{b}}(\mathbf{a}) \mathcal{Z}_{\mathbf{a}}(\Delta)\big)_{b_1}$ for fixed $(b_2, \dots, b_{n-1})$ satisfies a stable self-consistent equation; See Step I_2 below.
- Proving Theorem 4.8 (in fact, a weaker version given in Proposition 5.3 below) for a chain Δ and an adapted chain weight is a key tool for proving Theorem 4.8 for arbitrary Δ.

PROPOSITION 5.3. *Suppose that $\Lambda \prec \Psi$ for some admissible control parameter Ψ, and recall the definition* (3.1) *of Φ. Let Δ be a chain, w an adapted chain weight, and $F \subset V_s(\Delta)$. Then we have*

$$X^w_F(\Delta) \;\prec\; \Psi^{\deg(\Delta) + |F|} \Phi^{c(\Delta) - |F|} \tag{5.2}$$

for any $\boldsymbol{\mu}$ and adapted chain weight w.

As an a priori bound in Sections 8 and 9, we shall always use Proposition 5.3 with $F = \emptyset$. The statement of Proposition 5.3 for $F = \emptyset$ may be summarized by saying that from each chain vertex we gain a factor Φ (as compared to the trivial bound (4.6)).

Next, we outline the proof Theorem 4.8. The argument consists of two main steps: establishing a priori bounds on chains (i.e. proving Proposition 5.3) and proving Theorem 4.8 using Proposition 5.3 as input.

Proposition 5.3 is proved first for open chains, using a two-step induction. The induction parameter is the length of the chain $\ell := \deg(\Delta)$. The induction is started at $\ell = 1$, and consists of two steps, I_1 and I_2. It may be summarized in the form

$$(\ell = 1\,,\, F = \emptyset) \xrightarrow{\mathrm{I}_1} (\ell = 2\,,\, F \neq \emptyset) \xrightarrow{\mathrm{I}_2} (\ell = 2\,,\, F = \emptyset) \xrightarrow{\mathrm{I}_1} (\ell = 3\,,\, F \neq \emptyset) \xrightarrow{\mathrm{I}_2} (\ell = 3\,,\, F = \emptyset) \xrightarrow{\mathrm{I}_1} \cdots\,.$$

What follows is a sketch of steps I_1 and I_2.

Step I_1**.** The input for Step I_2 is the claim of Proposition 5.3 with $F = \emptyset$, for all open chains Δ' satisfying $\deg(\Delta') < \ell$. Using a high moment expansion, we estimate $X_F^w(\Delta)$, where Δ is an open chain, $\deg(\Delta) = \ell$, and $F \neq \emptyset$. The details are carried out in Section 7.1.

Step I_2**.** We fix an open chain Δ and prove the claim of Proposition 5.3 for $F = \emptyset$, under the assumption that the claim of Proposition 5.3 has been established for

(i) Δ with $F \neq \emptyset$;

(ii) all open chains Δ' satisfying $\deg(\Delta') < \deg(\Delta)$ with $F = \emptyset$.

The proof is based on a self-consistent equation for the family $\big(\sum_{\mathbf{a}} w_{\mathbf{b}}(\mathbf{a}) \mathcal{Z}_{\mathbf{a}}\big)_{b_1}$ for fixed $b_2, \dots, b_n$. This self-consistent equation will be stable provided $E = \operatorname{Re} z$ lies away from the spectral edges ± 2. This stability is ensured by the fact that $\mathcal{Z}$ only contains factors G and not G^*. The details are carried out in Section 7.2.

The induction is started by noting that Proposition 5.3 holds trivially for the open chain of length 1 (which has no chain vertex), encoding the monomial $G_{\mu\nu} \prec \Psi$. After Steps I_1 and I_2 are complete, the induction argument outlined above completes the proof of Proposition 5.3 for open chains. The proof for closed chains is almost identical, except that no induction is needed; the only required assumption is that Proposition 5.3 hold for *open* chains of arbitrary degree.

Once Proposition 5.3 has been proved, we use it as input to prove Theorem 4.8 for a general $\Delta \in \mathfrak{Z}$. Similarly to Step I_2, we use a high-moment expansion. The estimates are considerably more involved than in Step I_2, however. (In the language of Sections 3.2.4 and 8, we use vertex resolution to gain extra powers of Ψ from the charged vertices.) The details are carried out in Sections 8 − 9.

We record the following guiding principle for the entire proof of Theorem 4.8. It is a *basic power counting* that can be summarized as follows. The size of $X_F^w(\Delta)$ is given by a product of three main ingredients:

(a) The naive size $\Psi^{\deg(\Delta)}$, which is simply the number of entries of $\mathcal{G}$ in $X_F^w(\Delta)$ (obtained by a trivial power counting and $\Lambda \prec \Psi$).

(b) The smallness arising from F, i.e. $\Psi^{|F|}$ (obtained from the linking imposed by the factors Q).

(c) The smallness arising from the charged vertices, i.e. $\Phi^{|V_c(\Delta)|}$ (obtained from vertex resolution and the a priori bounds of Proposition 5.3 applied to chain vertices).

We shall frequently refer to the factors $\Psi^{|F|}$ and $\Phi^{|V_c(\Delta)|}$ from (b) and (c) as *gain* over the naive size $\Psi^{\deg(\Delta)}$. It is very important for the whole proof that the mechanism of this gain is *local* in the graph, i.e. operates on the level of individual vertices. Each factor gained in the case (c) can be associated with a charged vertex. In the case (b), a linking results in an additional edge adjacent to the vertex on which a linking was performed. There will be some technical complications which somewhat obscure this picture, such as occasionally coinciding indices. We shall always analyse these exceptional situations by comparing them to the basic power counting dictated by the generic situation. We remark that these "exceptional" situations sometimes in fact lead to leading-order error terms, which is for instance the reason why the parameter Φ cannot in general be replaced with Ψ in (4.7).

Figure 5.2 contains a diagram summarizing all key steps of the proof.

We conclude this section with an outline of Sections 6 – 9. In Section 6 we present a simple high-moment estimate that only uses the process of *linking* (see Definition 3.10); more algebraically, the argument of Section 6 only uses Family A identities (and not Family B). The result is Proposition 6.1, which obtains a gain of a factor Ψ from each Q but no gain from charged vertices (see Definition 4.7). The goal of Section 6 is twofold, the first goal being pedagogical. It provides a complete but vastly simplified proof of a special case of Theorem 4.8, thereby illustrating the process of linking. In addition, it lays the ground for Step I_1 used to derive a priori bounds on chains, as well as for the more complicated high-moment estimates used in the full proof of Theorem 4.8.

Section 6 is devoted to chains; its goal is to prove Proposition 5.3. Step I_1 is proved in Section 7.1 and Step I_2 in Section 7.2. The induction, and hence the proof of Proposition 5.3, is completed in Section 7.3. In Section 8 we prove Theorem 4.8 under four simplifying assumptions, **(S1)** – **(S4)** listed in Sections 6 and 8. These simplifications allow us to ignore some additional complications, and give a streamlined argument in which the fundamental mechanism is evident. The starting point for the argument in Section 8 is the high-moment expansion using vertex linking, already introduced in Section 6. In addition, we make use of Family B identities, which leads us to the process of vertex resolution (sketched in Section 3.2.4). In Section 9 we present the additional arguments needed to drop Simplifications **(S1)** – **(S4)**, and hence prove Theorem 4.8 in full generality. Finally, in Section 10 we prove Theorem 4.15 as a relatively easy consequence of Theorem 4.8.

6. Warmup: simple high-moment estimates

We now move on to the high-moment estimates which underlie our proofs. The idea is to derive high-probability bounds on $X_F^w(\Delta)$ by controlling its high moments using a graphical expansion scheme.

For pedagogical reasons, we shall throughout the following selectively ignore some complications so as to make the core strategy clearer. We shall eventually put back the complications one by one. In this section we consistently assume the following simplification.

(S1) All summation indices in the expanded summation $\mathbb{E}|X_F^w(\Delta)|^p$ (see (6.5) below) are distinct. (I.e. we ignore repeated indices which give rise to a smaller combinatorics of the summation.)

In this section we present a simple argument which proves the following weaker estimate.

Proposition 6.1. *Suppose that $\Lambda \prec \Psi$ for some admissible control parameter Ψ, $\Delta \in \mathfrak{Z}$, and w is an adapted weight. Then for all $F \subset V_s(\Delta)$ and $\boldsymbol{\mu}$ we have*

$$X_F^w(\Delta) \;\prec\; \Psi^{\deg(\Delta)+|F|}\,. \tag{6.1}$$

The estimate (6.1) expresses that from each Q in $X_F^w(\Delta)$ one gains an additional factor Ψ.

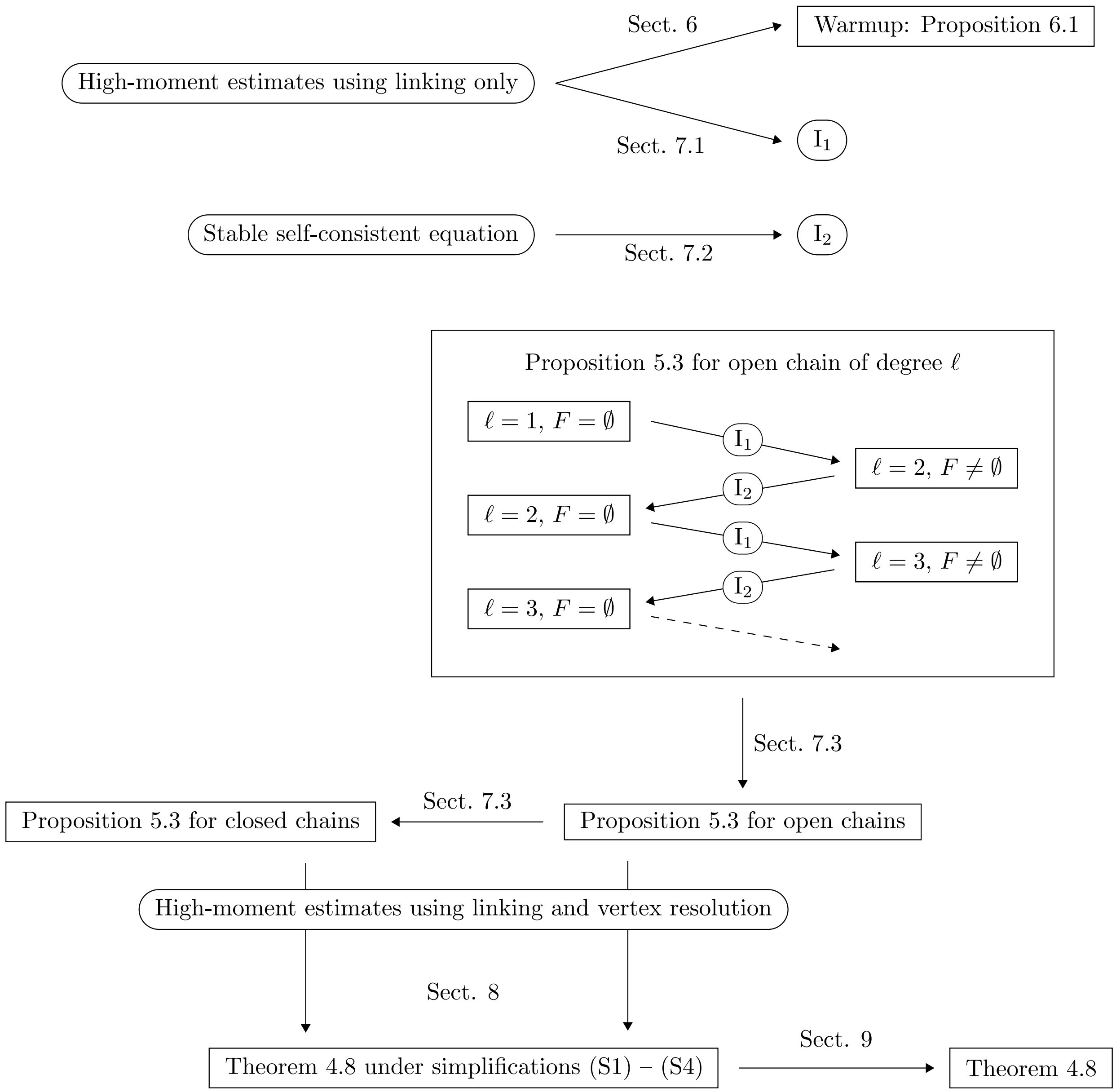
Sect. 6
Warmup: Proposition 6.1
High-moment estimates using linking only
Sect. 7.1
I_1
Stable self-consistent equation
Sect. 7.2
I_2
Proposition 5.3 for open chain of degree ℓ
$\ell = 1$, $F = \emptyset$
I_1
$\ell = 2$, $F \neq \emptyset$
I_2
$\ell = 2$, $F = \emptyset$
I_1
$\ell = 3$, $F \neq \emptyset$
I_2
$\ell = 3$, $F = \emptyset$
Sect. 7.3
Sect. 7.3
Proposition 5.3 for closed chains
Proposition 5.3 for open chains
High-moment estimates using linking and vertex resolution
Sect. 8
Sect. 9
Theorem 4.8 under simplifications (S1) – (S4)
Theorem 4.8

FIGURE 5.2. The structure of the proof of Theorem 4.8. Concepts and arguments are displayed in rounded boxes, statements and results in rectangular boxes.

REMARK 6.2. As in Remark 4.12, the statement of Proposition 6.1 remains true if some (or all) diagonal entries of the form $\mathcal{G}_{aa} = G_{aa} - m$ are replaced by $1/G_{aa} - 1/m$. The proof is exactly the same.

The simplified argument behind the proof of Proposition 6.1 uses only the Family A identities, i.e. (3.13). It relies on a high-moment estimate of the following form. The precise statement is somewhat complicated by the need to keep track of low-probability exceptional events. The sum over $\Gamma \in \mathfrak{G}$ in Lemma 6.3 will arise as a summation over graphs.

LEMMA 6.3. *Suppose that $\Lambda \prec \Psi$ for some admissible control parameter Ψ, and let $p \in 2\mathbb{N}$ be even. Then we have*

$$\mathbb{E}|X_F^w(\Delta)|^p \;\leqslant\; \sum_{\Gamma \in \mathfrak{G}} \mathbb{E} X_\Gamma\,, \tag{6.2}$$

where $\mathfrak{G}$ is a finite set (depending on Δ, F, and p) and X_Γ is a random variable satisfying

$$X_\Gamma \;\prec\; \Psi^{p(\deg(\Delta)+|F|)} \tag{6.3}$$

as well as the rough bound

$$\mathbb{E}|X_\Gamma|^2 \;\leqslant\; N^{C_p} \tag{6.4}$$

for some constant C_p.

Before proving Lemma 6.3, we show how it implies Proposition 6.1.

PROOF OF PROPOSITION 6.1. Let $\varepsilon > 0$ and $D > 0$ be given. Define p as the smallest even number greater than $4D/\varepsilon$, and abbreviate $q := \deg(\Delta) + |F|$. Then by Lemma 6.3, for each $\Gamma \in \mathfrak{G}$ there exists an event Ξ_Γ such that

$$|X_\Gamma| \mathbf{1}(\Xi_\Gamma) \;\leqslant\; N^{\varepsilon p/2} \Psi^{pq}\,, \qquad \mathbb{P}(\Xi_\Gamma^c) \;\leqslant\; N^{-C_p - pq}$$

for all w, $\boldsymbol{\mu}$, and $z \in \mathbf{S}$. Then we find, using Lemma 6.3 again,

$$\begin{aligned}
\mathbb{E}|X_F^w(\Delta)|^p &\leqslant \sum_{\Gamma \in \mathfrak{G}} \Big(\mathbb{E}\big(X_\Gamma \mathbf{1}(\Xi_\Gamma)\big) + \mathbb{E}\big(X_\Gamma \mathbf{1}(\Xi_\Gamma^c)\big) \Big) \\
&\leqslant \sum_{\Gamma \in \mathfrak{G}} \Big(N^{\varepsilon p/2} \Psi^{pq} + \big(\mathbb{E}|X_\Gamma|^2\big)^{1/2} \mathbb{P}(\Xi_\Gamma^c)^{1/2} \Big) \\
&\leqslant \big|\mathfrak{G}\big| \Big(N^{\varepsilon p/2} \Psi^{pq} + N^{-pq/2} \Big) \\
&\leqslant 2\big|\mathfrak{G}\big| N^{\varepsilon p/2} \Psi^{pq}\,,
\end{aligned}$$

for all w, $\boldsymbol{\mu}$, and $z \in \mathbf{S}$. Therefore Chebyshev's inequality gives

$$\mathbb{P}\Big(|X_F^w(\Delta)| > N^\varepsilon \Psi^q \Big) \;\leqslant\; 2\big|\mathfrak{G}\big| N^{-\varepsilon p/2} \;\leqslant\; N^{-D}$$

for all w, $\boldsymbol{\mu}$, and $z \in \mathbf{S}$. □

The rest of this section is devoted to the proof of Lemma 6.3. All of our estimates will be uniform in w and $\boldsymbol{\mu}$, and we shall henceforth no longer mention this explicitly. Throughout this section we assume Simplification **(S1)**.

Proof of Lemma 6.3. The idea of the proof was already outlined in Section 3.2.1. Let $\Delta \in \mathfrak{Z}$ have n summation indices, denoted by $a_1, \dots, a_n$, and k external indices, denoted by $\mu_1, \dots, \mu_k$. Let $F \subset \{1, \dots, n\}$. Let $p \in 2\mathbb{N}$ be even and write

$$\mathbb{E}|X_F^w(\Delta)|^p \;=\; \sum_{\mathbf{a}^1}^{(\boldsymbol{\mu})*} w(\mathbf{a}^1) \cdots \sum_{\mathbf{a}^p}^{(\boldsymbol{\mu})*} w(\mathbf{a}^p) \mathbb{E} \prod_{j=1}^{p/2} \left[\left(\prod_{i \in F} Q_{a_i^j} \right) \mathcal{Z}_{\mathbf{a}^j} \right] \prod_{j=p/2+1}^{p} \overline{\left[\left(\prod_{i \in F} Q_{a_i^j} \right) \mathcal{Z}_{\mathbf{a}^j} \right]}\,, \tag{6.5}$$

where we abbreviated

$$\mathbf{a}^j \;:=\; (a_i^j : 1 \leqslant i \leqslant n)\,, \qquad \mathbf{a} \;:=\; \big(a_i^j : 1 \leqslant i \leqslant n\,,\, 1 \leqslant j \leqslant p\big)\,.$$

We now make the crucial observation that $\widehat{w}(\mathbf{a}) := w(\mathbf{a}^1) \dots w(\mathbf{a}^j)$ is a weight on the set of indices (i,j); this is an elementary consequence of the Definition (4.4). In particular, $\sum_{\mathbf{a}} \widehat{w}(\mathbf{a}) \leqslant 1$.

By Simplification **(S1)**, we assume that all indices $\mathbf{a}$ are distinct: in addition to the constraint $a_i^j \notin \{\mu_1, \dots, \mu_k\}$, we introduce into (6.5) an indicator function that imposes $a_i^j \neq a_{i'}^{j'}$ if $(i,j) \neq (i',j')$.

We now make each $\mathcal{G}_{xy}$ independent of as many summation indices as possible using Family A identities. To that end, we define

$$\mathcal{G}_{ij}^{(T)} \;:=\; G_{ij}^{(T)} - \delta_{ij}\, m\,.$$

Using (3.13) iteratively, we expand every factor $\mathcal{G}_{xy}$ appearing in (6.5) in all the indices

$$\mathbf{a}_F \;:=\; \big(a_i^j : i \in F\,,\, 1 \leqslant j \leqslant p\big)$$

associated with a factor Q. Let $\mathcal{G}_{xy}$ be a fixed entry in (6.5). The idea is to successively add to $\mathcal{G}_{xy}$ as many upper indices from the collection $\mathbf{a}_F$ as possible. The goal is to obtain a quantity satisfying the following definition.

Definition 6.4. An entry $G_{xy}^{(T)}$ or $\mathcal{G}_{xy}^{(T)}$ is *maximally expanded in* S if $S \subset T \sqcup \{x,y\}$. In other words, a maximally expanded resolvent entry cannot be expanded any further in the indices S using (3.13).

Along the expansion of each $\mathcal{G}_{xy}$ using (3.13), new terms in (6.5) appear; each such term is a monomial of entries of $\mathcal{G}$ divided by diagonal entries of G. We stop expanding a term if either

(a) all its factors are maximally expanded in $\mathbf{a}_F$, or

(b) it contains $\deg(\Delta) + 2pn$ entries of $\mathcal{G}$ in the numerator.

The precise recursive procedure is as follows. We start by setting $A := \mathcal{G}_{xy}$, where $\mathcal{G}_{xy}$ is an entry on the right-hand side of (6.5).

1. Let $\mathcal{G}_{uv}^{(T)}$ denote an entry in A and d an index in $\mathbf{a}_F$ such that $d \notin T \cup \{u,v\}$. (This choice is arbitrary and unimportant.) If (a) no such pair exists, or (b) A contains $\deg(\Delta) + 2pn$ factors $\mathcal{G}$ in the numerator, stop the recursion of the term A.

2. Using Family A identities, write

$$\mathcal{G}_{uv}^{(T)} \;=\; \mathcal{G}_{uv}^{(Td)} + \frac{\mathcal{G}_{ud}^{(T)} \mathcal{G}_{dv}^{(T)}}{G_{dd}^{(T)}} \tag{6.6}$$

if $\mathcal{G}_{uv}^{(T)}$ is a resolvent entry in the numerator and

$$\frac{1}{G_{uu}^{(T)}} \;=\; \frac{1}{G_{uu}^{(Td)}} - \frac{\mathcal{G}_{ud}^{(T)}\mathcal{G}_{du}^{(T)}}{G_{uu}^{(T)}G_{uu}^{(Td)}G_{dd}^{(T)}} \tag{6.7}$$

if $G_{uv}^{(T)} = G_{uu}^{(T)}$ is a diagonal resolvent entry in the denominator. This yields the splitting $A = A' + A''$, where both terms have the form of a product of entries in the numerator and diagonal entries in the denominator. Repeat step 1 for both A' and A'' (playing the role of A in step 1).

It is not hard to see that the stopping rule defined by the conditions (a) or (b) ensures that the recursion terminates after a finite number of steps. Indeed, the quantity "number of entries of $\mathcal{G}$" + "number of upper indices" must remain bounded by the stopping rules (a) and (b).

The result is of the form

$$\mathcal{G}_{xy} \;=\; \sum_{\alpha} H_\alpha + R\,,$$

where each summand H_α is a fraction with entries of $\mathcal{G}$ in the numerator and diagonal entries of G in the denominator, all of them maximally expanded in $\mathbf{a}_F$. Here the rest term R satisfies

$$R \;\prec\; \Psi^{\deg(\Delta)+2pn}\,, \qquad \mathbb{E}|R|^2 \;\leqslant\; N^{C_{p,\Delta}} \tag{6.8}$$

for some constant $C_{p,\Delta}$. The first estimate of (6.8) follows from (3.16) combined with (3.11) and Lemma 3.6, and the second estimate of (6.8) from (3.18) and (3.19).

We then multiply the resulting sums on the right-hand side of (4.2) out to get

$$\mathcal{Z}_{\mathbf{a}^j} \;=\; \sum_{\alpha} Y_{\mathbf{a}}^{j,\alpha}\,, \tag{6.9}$$

where $Y_{\mathbf{a}}^{j,\alpha}$ is a monomial and α a counting index ranging over some finite set. Each term $Y_{\mathbf{a}}^{j,\alpha}$ is a fraction with entries of $\mathcal{G}$ in the numerator and diagonal entries of F in the denominator. Moreover, either (i) all entries of $Y_{\mathbf{a}}^{j,\alpha}$ are maximally expanded in $\mathbf{a}_F$ or (ii) $Y_{\mathbf{a}}^{j,\alpha} \prec \Psi^{\deg(\Delta)+2pn}$ (the latter arises if $Y_{\mathbf{a}}^{j,\alpha}$ contains one or more rest terms R). We now multiply out the expectation in (6.5) as

$$\mathbb{E}\Biggl[\biggl(\prod_{i\in F} Q_{a_i^1}\biggr)\mathcal{Z}_{\mathbf{a}^1}\Biggr]\cdots\Biggl[\biggl(\prod_{i\in F} Q_{a_i^p}\biggr)\overline{\mathcal{Z}_{\mathbf{a}^p}}\Biggr] \;=\; \sum_{\alpha_1,\dots,\alpha_p}\mathbb{E}\Biggl[\biggl(\prod_{i\in F} Q_{a_i^1}\biggr)Y_{\mathbf{a}}^{1,\alpha_1}\Biggr]\cdots\Biggl[\biggl(\prod_{i\in F} Q_{a_i^p}\biggr)\overline{Y_{\mathbf{a}}^{p,\alpha_p}}\Biggr]. \tag{6.10}$$

We plug this into (6.5) and pull out the summation over $\alpha_1,\dots,\alpha_p$. This gives rise to the summation in (6.2), indexed by the set $\mathfrak{G} = \{(\alpha_1,\dots,\alpha_p)\}$. If, for some $(\alpha_1,\dots,\alpha_p)$, one or more of $Y_{\mathbf{a}}^{1,\alpha_1},\dots,Y_{\mathbf{a}}^{p,\alpha_p}$ is not maximally expanded in $\mathbf{a}_F$, it is easy to see that

$$X_{\alpha_1\cdots\alpha_p} \;:=\; \sum_{\mathbf{a}^1}^{(\boldsymbol{\mu})*} w(\mathbf{a}^1)\cdots\sum_{\mathbf{a}^p}^{(\boldsymbol{\mu})*} w(\mathbf{a}^p)\Biggl[\biggl(\prod_{i\in F} Q_{a_i^1}\biggr)Y_{\mathbf{a}}^{1,\alpha_1}\Biggr]\cdots\Biggl[\biggl(\prod_{i\in F} Q_{a_i^p}\biggr)\overline{Y_{\mathbf{a}}^{p,\alpha_p}}\Biggr] \tag{6.11}$$

satisfies (6.3) and (6.4). Indeed, each term $Y_{\mathbf{a}}^{j,\alpha_j}$ contains at least $\deg(\Delta)$ entries of $\mathcal{G}$; thus the trivial bound $Y_{\mathbf{a}}^{j,\alpha_j} \prec \Psi^{\deg(\Delta)}$ always holds by Lemma 3.9. Using Lemma 3.6 we can multiply these estimates. Recalling (6.8), (3.18), and (3.19), we find (6.3) and (6.4).

It therefore suffices to consider products of $Y_{\mathbf{a}}^{j,\alpha_j}$'s in (6.10) which are all maximally expanded in $\mathbf{a}_F$ (i.e. terms which are products of H_α's only and not R's). The presence of Q's leads to the following crucial restriction on terms yielding a nonzero contribution to (6.10). For each $i \in F$, we claim that at least one of $Y_{\mathbf{a}}^{2,\alpha_2}, \dots, Y_{\mathbf{a}}^{p,\alpha_p}$ is not independent of a_i^1. This follows from the observation that generally $\mathbb{E}[Q_a(X)Y] = 0$ if Y is independent of a. More generally, we require that, for any $i \in F$ and $j = 1, \dots, p$, at least of one of

$$Y_{\mathbf{a}}^{1,\alpha_1}, \dots, \widehat{Y_{\mathbf{a}}^{j,\alpha_j}}, \dots, Y_{\mathbf{a}}^{p,\alpha_p}$$

is not independent of a_i^j (hat indicates omission from the list). This imposes a constraint on the terms that survive the expansion.

Moreover, the term that is not independent of a_i^j contains at least one additional entry of $\mathcal{G}$, since at some point the formula (6.6) or (6.7) had to be applied with $d = a_i^j$ and the second term of (6.6) or (6.7) contains at least one additional entry of $\mathcal{G}$. Since we assumed Simplification **(S1)**, i.e. all a_i^j's are different, it is a general fact that each Q gives rise to an additional off-diagonal entry of $\mathcal{G}$ and contributes a factor $\Lambda \prec \Psi$ to (6.5). In other words, any $X_{\alpha_1 \cdots \alpha_p}$ yielding a nonzero contribution to (6.5) has at least $p(\deg(\Delta) + |F|)$ entries of $\mathcal{G}$ in the numerator. Recalling Lemma 3.6 and Lemma 3.9, we find that any term $X_{\alpha_1 \cdots \alpha_p}$ yielding a nonzero contribution to (6.5) satisfies (6.3) and (6.4). This concludes the proof of Lemma 6.3. □

6.1. Graphical representation. The phenomenon behind the proof of Lemma 6.3 in fact has a simple graphical representation, which will prove essential for later, more intricate, estimates. We illustrate its usefulness by applying it to the proof of Lemma 6.3. We recall the basic graphical notation introduced in Section 3.2.2.

The quantity whose expectation we are estimating, $|X_F^w(\Delta)|^p = \big[X_F^w(\Delta)\big]^{p/2}\big[\overline{X_F^w(\Delta)}\big]^{p/2}$, has a natural representation in terms of a multigraph, which we call $\gamma^p(\Delta)$ and which is essentially a p-fold copy of the graph Δ encoding $\mathcal{Z}$. The graph $\gamma^p(\Delta)$ is obtained as follows.

(i) Take $p/2$ copies of Δ and $p/2$ copies of Δ whose edges have inverted direction and colour arising from the relation $\overline{\mathcal{G}}_{ab} = \mathcal{G}_{ba}^*$. More precisely, this inversion means that each edge $e \in E(\Delta)$ gives rise to an inverted edge e' satisfying

$$\xi_{e'} \;=\; \begin{cases} 1 & \text{if } \xi_e = * \\ * & \text{if } \xi_e = 1\,, \end{cases} \qquad \alpha(e') \;=\; \beta(e)\,, \qquad \beta(e') \;=\; \alpha(e)\,.$$

(ii) For each external vertex $i \in V_e(\Delta)$ merge all p copies of i to form a single vertex.

Note that the set F is not depicted in $\gamma^p(\Delta)$. The vertex set of $\gamma^p(\Delta)$ consists of summation vertices and external vertices (this classification is inherited from the vertices of Δ in the obvious way), so that we may write $V(\gamma^p(\Delta)) = V_s(\gamma^p(\Delta)) \sqcup V_e(\gamma^p(\Delta))$.

Definition 6.5 (Projection π). We introduce the p-to-one canonical projection $\pi : V(\gamma^p(\Delta)) \to V(\Delta)$, defined as $\pi(i) = j$ if i is a copy of j in the construction of $\gamma^p(\Delta)$.

We start with the graph $\gamma^p(\Delta)$ (see Figure 6.1). We shall construct a set $\widetilde{\mathfrak{G}}_F^p(\Delta)$ of graphs, denoted by Γ, on the same vertex set $V(\gamma^p(\Delta))$. The algorithm that generates $\widetilde{\mathfrak{G}}_F^p(\Delta)$ is precisely the one given after Definition 6.4. On the level of graphs, this algorithm consists of a repeated application of the graphical rules in Figures 3.5 and 3.6. (Note that the second identity of Figure 3.6 is also valid for $\mathcal{G}$ instead of G, i.e. without the black diamonds.) As indicated in Figures 3.5 and 3.6, we keep track of the upper indices

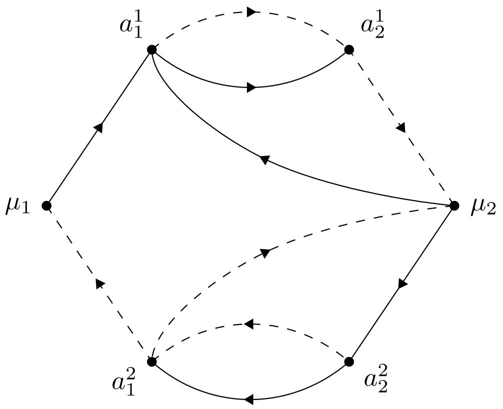

FIGURE 6.1. The graph $\gamma^2(\Delta)$ that encodes $\mathbb{E}|X_F^w(\Delta)|^2$, where $\mathcal{Z}_{a_1a_2} = \mathcal{G}_{\mu_1a_1}\mathcal{G}_{a_1a_2}\mathcal{G}^*_{a_2\mu_2}\mathcal{G}^*_{a_1a_2}\mathcal{G}_{\mu_2a_1}$.

associated with an edge by attaching a list of upper indices to each edge. The algorithm terminates when either all edges are maximally expanded in $\mathbf{a}_F$ or there are $\deg(\Delta) + pn$ edges that do not bear a diamond (i.e. that contribute a factor Ψ). Indicating these upper indices may be more precisely implemented using decorated edges, but we shall not need such formal constructions.

Recall from Definition 3.10 that choosing the second graph on the right-hand side of any identity in Figures 3.5 and 3.6 is called *linking* (an edge with a vertex). As shown above, any graph $\Gamma \in \widetilde{\mathfrak{G}}_F^p(\Delta)$ whose edges are not maximally expanded yields a small enough contribution by a trivial power counting. In the following we shall therefore only consider the remaining graphs, i.e. we shall assume that all edges of $\Gamma \in \widetilde{\mathfrak{G}}_F^p(\Delta)$ are maximally expanded in $\mathbf{a}_F$. Moreover, the upper indices of an edge are uniquely determined by the constraint that the edge be maximally expanded: the entry encoded by the edge (x, y) is $\mathcal{G}_{xy}^{(\mathbf{a}_F \setminus \{x,y\})}$. Thus, we shall consistently drop the upper indices associated with edges from our graphs Γ.

We introduce some convenient notions when dealing with graphs in $\widetilde{\mathfrak{G}}_F^p(\Delta)$.

DEFINITION 6.6. Let $\Gamma \in \widetilde{\mathfrak{G}}_F^p(\Delta)$.

(i) We denote the vertex set of Γ by $V(\Gamma) = V_s(\Gamma) \sqcup V_e(\Gamma)$, where $V_s(\Gamma)$ denotes the summation vertices and $V_e(\Gamma)$ external (fixed) vertices. By definition, all three sets are the same as those of $\gamma^p(\Delta)$.

(ii) We denote the set of edges of Γ by $E(\Gamma)$; the set $E(\Gamma)$ has a colouring $\xi : E(\Gamma) \to \{1, *\}$.

(iii) The p-to-one canonical projection $\pi : V(\Gamma) \to V(\Delta)$ is taken over from Definition 6.5.

Thus, $\mathbf{a}_F = \big(a_i : i \in \pi^{-1}(F)\big)$. In this manner we write the sum of maximally expanded terms on the right-hand side of (6.10) as a sum of graphs $\Gamma \in \widetilde{\mathfrak{G}}_F^p(\Delta)$. By definition, each vertex in $\pi^{-1}(F)$ has been linked with at least one edge, possibly more. Each such linking adds an edge to the graph, and hence contributes a factor $\Lambda \prec \Psi$ to its size. This concludes the graphical discussion behind the proof of (6.1). Figure 6.2 shows two sample graphs from $\widetilde{\mathfrak{G}}_F^2(\Delta)$ for the graph Δ from Figure 6.1, where F consists of a single vertex associated with the summation variable a_1.

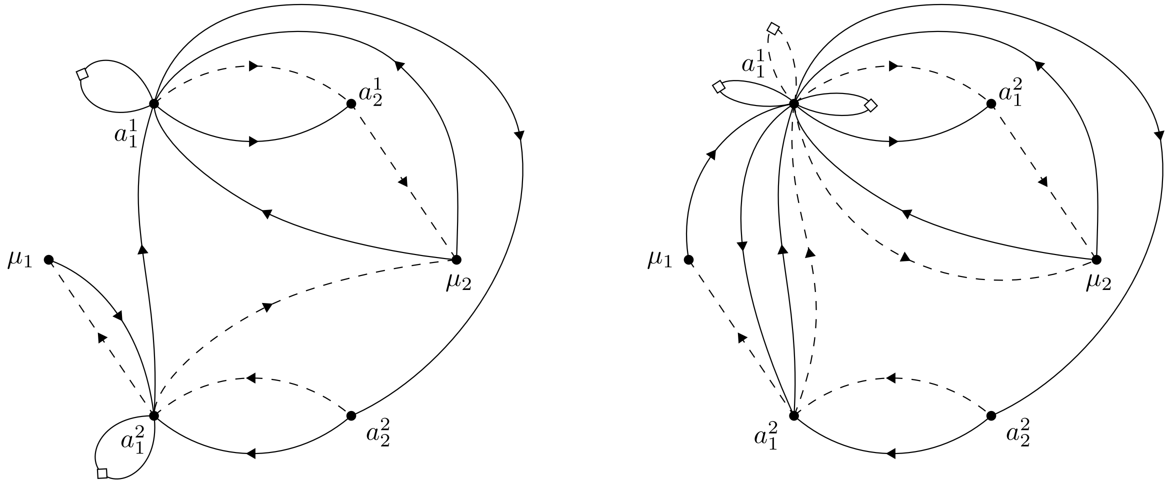

Figure 6.2. Left: a graph obtained from the one in Figure 6.1 with $|F| = 1$ by linking both vertices in $\mathbf{a}_F = (a_1^1, a_1^2)$ with an edge; a_1^1 was linked with the edge (μ_2, a_2^2) and a_1^2 was linked with (μ_1, a_1^1). This graph contains the minimal number of edges to yield a nonzero contribution after taking the expectation. Hence it is of leading order. Right: a graph obtained by linking two further edges with a_1^1, namely the edges (a_1^2, μ_2) and (μ_1, a_1^2). Its size is subleading.

7. Chains

In this section we derive the a priori estimate on chains, Proposition 5.3.

7.1. Step I$_1$: chains with $F \neq \emptyset$. Step I$_1$ is an application of the simple high-moment expansion method from Section 6. It is formulated in the following proposition. In Section 7.3, it will be used in conjunction with Proposition 7.2 below to complete the induction and hence the proof of Proposition 5.3.

Proposition 7.1 (Induction Step I$_1$). *Suppose that $\Lambda \prec \Psi$ for some admissible control parameter Ψ, and let $\ell \geqslant 2$. Suppose that*

$$X_F^w(\Delta) \;\prec\; \Psi^{\deg(\Delta)+|F|}\Phi^{c(\Delta)-|F|} \tag{7.1}$$

holds for any open chain Δ of degree strictly less than ℓ, $F = \emptyset$, and any adapted chain weight w. Then (7.1) holds for any open chain Δ of degree ℓ, $F \neq \emptyset$, and any adapted chain weight w.

In this section we continue to assume Simplification **(S1)** (see the beginning of Section 6).

Proof of Proposition 7.1. For simplicity of notation, we focus on the case where Δ is a directed chain of degree ℓ; the undirected case is proved in the same way. The argument is best understood in a representative example,

$$\mathcal{Z}_{\mathbf{a}} \;=\; G_{\mu_1 a_1} G_{a_1 a_2} G_{a_2 a_3} G_{a_3 a_4} G_{a_4 \mu_2}\,, \tag{7.2}$$

which is encoded by the graph Δ depicted in Figure 7.1. Let us take $F = \{1\}$ and compute the variance of $X_F^w(\Delta)$. In the following we use the terminology and notation of Section 6 without further comment. In Section 6 is was shown that the only graphs $\Gamma \in \widetilde{\mathfrak{G}}_F^2(\Delta)$ that contribute are those in which the vertices a_1^1 and a_1^2 have both been linked to some edge. Figure 7.2 shows such a graph Γ of leading order. Since the

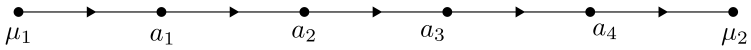

FIGURE 7.1. The graph Δ that encodes $\mathcal{Z}$ defined in (7.2).

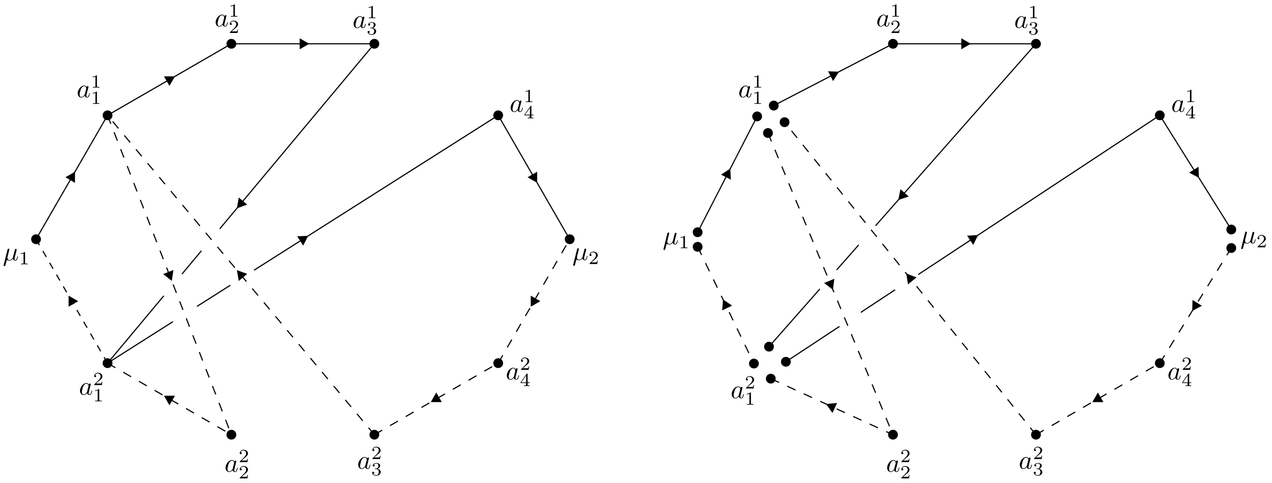

FIGURE 7.2. Left: a graph Γ of leading order in $\widetilde{\mathfrak{G}}^2_F(\Delta)$ with Δ defined through (7.2), and $F = \{1\}$. We do not draw the loops that encode diagonal resolvent entries in the denominator. Right: the same graph broken down to chains.

two vertices a_1^1 and a_1^2 have been linked, they each contribute a factor $\Lambda \prec \Psi$ (two edges were added by the linking process). Now we break the graph in Figure 7.2 down to its chains, i.e. we freeze all those summation vertices, a_1^1 and a_1^2, that were linked to. What remains is a collection of chains, each shorter than the original chain Δ. In this example there are four nontrivial subchains:

$$a_1^1 \to a_2^1 \to a_3^1 \to a_1^2\,, \qquad a_1^2 \to a_4^1 \to \mu_2\,, \qquad \mu_2 \to a_4^2 \to a_3^2 \to a_1^1\,, \qquad a_1^1 \to a_2^2 \to a_1^2\,.$$

Moreover, the monomial encoded by each subchain lies either inside $Q_{a_1^1}(\cdot)$, inside $Q_{a_1^2}(\cdot)$, or inside neither. Thus the monomials encoded by the first two subchains lie inside $Q_{a_1^1}(\cdot)$, and the monomials encoded by the two last subchains inside $Q_{a_1^2}(\cdot)$.

Now we may invoke the induction assumption (i.e. Proposition 5.3 for $F = \emptyset$) on each of the four subchains. We use that they all have degree strictly less than $\deg(\Delta)$. To be precise, before invoking Proposition 5.3, we have to get rid of the upper indices using (3.13); see below for details.

Moreover, we ignore some minor technicalities associated with coinciding indices. By Simplification **(S1)**, we assumed that all summation indices of Γ were distinct. In particular, the indices associated with different subchains of Γ are distinct, which implies that the subchains of Γ are coupled. This coupling is manifested in the fact that summation indices within a subchain are subject to additional restrictions that are unrelated to that subchain: these summation indices cannot take on values of indices in other subchains. This means that summations cannot be performed independently within each subchain. Hence we may not strictly speaking invoke Proposition 5.3 for each subchain; in order to do so, we first have to *decouple* the subchains so as to get a product of terms associated with the subchains. In order to achieve this decoupling we have to allow indices associated with different subchains to coincide. This decoupling is a simple inclusion-exclusion argument whose details are postponed to Lemma 9.5 in Section 9.3.

Summarizing this example, we obtain an estimate of order $\Psi^{12}\Phi^6$ for the graph depicted in Figure 7.2. Since, by Proposition 5.3, the contribution of an open subchain of degree d is $\Psi^d\Phi^{d-1}$, the four non-trivial subchains yield a contribution $\Psi^3\Phi^2\,\Psi^2\Phi\,\Psi^3\Phi^2\,\Psi^3\Phi = \Psi^{10}\Phi^6$. There are also two trivial subchains, thus resulting in a total contribution $\Psi^{12}\Phi^6$. Another way to think about such estimates is to count the additional factors of Ψ and Φ gained along the proof. The naive size of the original graph, before linking, was $\Psi^{2\deg(\Delta)} = \Psi^{10}$ since $\mathcal{Z}$ in (7.2) contains five factors and we consider its second moment (i.e. set $p = 2$). Since $|F| = 1$, we gain an additional $\Psi^{2|F|} = \Psi^2$ from the linking; this step increases the number of edges from 10 to 12 in the graph on the left-hand side of Figure 7.2. Moreover, we gain an additional Φ factor from each internal summation vertex in the subchains, in this example we gain a factor Φ from each of the six vertices a_2^1, a_3^1, a_4^1, a_4^2, a_3^2, and a_2^2. Thus we recover the bound $\Psi^{12}\Phi^6$.

Let us now give the general argument, which is in fact a trivial generalization of the above example. We start with a graph $\Gamma \in \widetilde{\mathfrak{G}}_F^p(\Delta)$, as constructed in Section 6. We split the summation indices $\mathbf{a} = (\mathbf{a}', \mathbf{a}'')$, where $\mathbf{a}'$ consists of the chain vertices of Γ. Thus, $\mathbf{a}''$ contains in particular the indices associated with vertices which have been linked to an edge. By the argument of Section 6, $\mathbf{a}''$ contains all indices of $\mathbf{a}_F$, so that $|\mathbf{a}''| \geqslant |\mathbf{a}_F| = p|F|$. Since each linked vertex is incident to an additional edge resulting from linking, the graph Γ contains at least $p\deg(\Delta) + |\mathbf{a}_F| \geqslant p(\deg(\Delta) + |F|)$ edges. So far we have simply repeated the argument of Section 6 and reproved the bound (6.1).

In order to gain an additional factor Φ from each of the summation indices in $\mathbf{a}'$, we use the induction assumption. The assumption is used on open chains of vertices, i.e. subgraphs of Γ which are open chains. We fix $\mathbf{a}''$ and regard $\mathbf{a}'$ as the summation indices. Then Γ becomes a collection of open (sub)chains, and the vertices associated with $\mathbf{a}'$ are the chain vertices of these subchains. If we can ensure that each subchain has degree strictly less than Δ, we can apply the induction assumption to get an additional factor Φ from each chain vertex in represented in $\mathbf{a}'$. This will give us a bound of size

$$\Psi^{p\deg(\Delta)+|\mathbf{a}''|}\Phi^{|\mathbf{a}'|} \;\leqslant\; \Psi^{p(\deg(\Delta)+|F|)}\Phi^{p(c(\Delta)-|F|)}\,,$$

where we used that $|\mathbf{a}''| \geqslant p|F|$, $|\mathbf{a}''| + |\mathbf{a}'| = p\,c(\Delta)$, and $\Psi \leqslant \Phi$.

In order to carry out this argument, we make the following observations.

(i) All subchains of Γ have degree strictly less than $\deg(\Delta)$. This property is crucial for the induction. It is a consequence of the two following facts. First, the linking of vertices never produces new subchains nor lengthens pre-existing subchains. Note that vertices in $\mathbf{a}''$ are fixed, and subchains terminate at them. Second, since $F \neq \emptyset$, at least one vertex of every subchain of degree $\deg(\Delta)$ in $\gamma^p(\Delta)$ will be linked to an edge, hence cutting the subchain of degree $\deg(\Delta)$ into smaller subchains.

(ii) The expression $\mathcal{Z}'_{\mathbf{b}}$ encoded by any subchain Γ' of Γ always appears in conjunction with a chain weight $w'(\mathbf{b})$. This is an immediate consequence of the fact that the weight $w(\mathbf{a}^1)\cdots w(\mathbf{a}^p)$ is a chain weight by assumption.

(iii) Let $\mathcal{Z}'_{\mathbf{b}}$ denote the monomial encoded by a subchain Γ' of Γ. Then any Q_a has an index a in $\mathbf{a}''$ (i.e. is fixed), and acts either on all resolvent entries of $\mathcal{Z}'_{\mathbf{b}}$ or none of them.

In order to invoke the induction assumption, we still have to get rid of the upper indices in the maximally expanded resolvent entries. The procedure is almost identical to the one following Definition 6.4, but in the opposite direction. In particular, the key formula (3.13) should be viewed in the form

$$G_{ij}^{(Tk)} \;=\; G_{ij}^{(T)} - \frac{G_{ik}^{(T)}G_{kj}^{(T)}}{G_{kk}^{(T)}}\,, \qquad \frac{1}{G_{ii}^{(Tk)}} \;=\; \frac{1}{G_{ii}^{(T)}} + \frac{G_{ik}^{(T)}G_{ki}^{(T)}}{G_{ii}^{(T)}G_{ii}^{(Tk)}G_{kk}^{(T)}}\,. \tag{7.3}$$

We start removing the upper indices one by one using (7.3), and stop if either all upper indices have been removed or if the number of off-diagonal resolvent entries exceeds $\deg(\Delta) + 2p\ell$. The size of the latter terms is already sufficiently small by the trivial bound $\Lambda \prec \Psi$. As for the former terms, they are represented by a new (but still finite) set of graphs in which every vertex is either a chain vertex or has been linked with an edge.

Now the induction assumption is applicable to each subchain, and the proof is completed by invoking Lemma 3.6. (Note that as before we ignored issues related to coinciding indices according to Simplification **(S1)**; these are dealt with using the inclusion-exclusion argument of Lemma 9.5.) □

7.2. Step I$_2$: chains with $F = \emptyset$. Step I$_2$ is completed in the following proposition.

PROPOSITION 7.2 (INDUCTION STEP I$_2$). *Suppose that $\Lambda \prec \Psi$ for some admissible control parameter Ψ, and let $\ell \geqslant 2$. Suppose that*

$$X_F^w(\Delta) \;\prec\; \Psi^{\deg(\Delta)+|F|}\Phi^{c(\Delta)-|F|} \tag{7.4}$$

holds for any open chain Δ of degree ℓ, any $F \neq \emptyset$, and any adapted chain weight w. If $\ell \geqslant 3$, suppose in addition that (7.4) *holds for any open chain Δ of degree strictly less than ℓ, $F = \emptyset$, and any adapted chain weight w.*

Then (7.4) *holds for any open chain Δ of degree ℓ, $F = \emptyset$, and any adapted chain weight w.*

PROOF. As before, we focus on the case where Δ is a directed open chain; the proof in in the undirected case is the same. Thus,

$$\mathcal{Z}_{a_1\cdots a_n} \;=\; G_{\mu_1 a_1}G_{a_1 a_2}\cdots G_{a_n \mu_2}\,, \qquad w_{\mathbf{b}}(\mathbf{a}) \;\equiv\; w(\mathbf{a}) \;=\; s_{a_1 b_1}\cdots s_{a_n b_n}$$

for some $\mathbf{b} = \{b_1, \dots, b_n\}$. (Recall that $G_{ij} = \mathcal{G}_{ij}$ for $i \neq j$.) Note that $n = \deg(\Delta) - 1$. We have to prove that

$$X_\emptyset^w(\Delta) \;\prec\; \Psi^{n+1}\Phi^n\,. \tag{7.5}$$

(Here we used that $c(\Delta) = n$.) The main idea of the proof was given in Section 3.2.3: derive a stable self-consistent equation whose error terms may be estimated using the induction assumption. We subdivide the proof into six steps.

To simplify notation, throughout this proof we use $\mathcal{E} \equiv \mathcal{E}(\mathbf{b}, \mu_1, \mu_2)$ to denote a random error term satisfying $\mathcal{E} \prec \Psi^{n+2}\Phi^{n-1}$. Like generic constants C, these error terms may change from line to line without changing name.

Moreover, in order to keep the presentation more concise, we shall sometimes ignore unimportant subtleties arising from coinciding summation indices. These complications are harmless and will be dealt with precisely using the inclusion-exclusion argument of Lemma 9.5. The general philosophy is the following: if we constrain a pair of indices to coincide instead of being distinct, we lose at most two factors of Ψ. Indeed, we lose at most one chain vertex (resulting in a loss of $\Phi \geqslant \Psi$), and at most one off-diagonal entry of G may become diagonal (resulting in a loss of Ψ). Note that, since Δ is an open chain, at most one off-diagonal entry may become diagonal when setting two summation indices to be equal. (This is not true for closed chains; see Section 7.3.) This loss of Ψ^2 is compensated by the factor $M^{-1} \leqslant \Psi^2$ we gain from the reduction in the number of summation variables.

Step (i). We introduce a factor P_{a_1} into the summation in $X^w_\emptyset(\Delta)$. We find

$$\begin{aligned} X^w_\emptyset(\Delta) \;&=\; \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1} \mathcal{Z}_{a_1\cdots a_n} + X^w_{\{1\}}(\Delta) \\ &=\; \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1} \mathcal{Z}_{a_1\cdots a_n} + \mathcal{E}\,. \end{aligned}$$

where the second equality follows from the induction assumption.

Step (ii). We introduce a factor $m^2/(G_{a_1a_1})^2$ in front of $X^w_\emptyset(\Delta)$; this prefactor will be important in the fourth step below, as the factor $1/(G_{a_1a_1})^2$ will be used to cancel diagonal resolvent entries arising from two applications of the identity (3.14a). We find

$$\begin{aligned} \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1} \mathcal{Z}_{a_1\cdots a_n} \;&=\; \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a}) P_{a_1} \biggl[\frac{(G_{a_1a_1}-m)^2 + 2m(G_{a_1a_1}-m) + m^2}{(G_{a_1a_1})^2} \mathcal{Z}_{a_1\cdots a_n}\biggr] \\ &=\; \sum_{a_1}^{(\mu_1\mu_2)} s_{a_1b_1} P_{a_1} \biggl[G_{\mu_1a_1} \frac{(G_{a_1a_1}-m)^2 + 2m(G_{a_1a_1}-m) + m^2}{(G_{a_1a_1})^2} \\ &\qquad \times \sum_{a_2,\dots,a_n}^{(a_1\mu_1\mu_2)*} s_{a_2b_2}\cdots s_{a_nb_n} G_{a_1a_2}\cdots G_{a_n\mu_2}\biggr]\,. \end{aligned}$$

For $n=1$ (i.e. $\deg(\Delta)=2$) the last line is understood to be $G_{a_1\mu_2}$.

We now show that the only the term m^2 in the numerator is relevant. By induction assumption and Lemma 3.6, we find that

$$\sum_{a_2,\dots,a_n}^{(a_1\mu_1\mu_2)*} s_{a_2b_2}\cdots s_{a_nb_n} G_{a_1a_2}\cdots G_{a_n\mu_2} \;\prec\; \Psi^n \Phi^{n-1}\,. \tag{7.6}$$

(Note that the induction assumption is only used if $n \geqslant 2$; for the initial value $n=1$ (7.6) is trivial.) Using $\sum_{a_1} s_{a_1b_1} \leqslant 1$, (3.17), and Lemma 3.6 again, we get

$$X^w_\emptyset(\Delta) \;=\; \widetilde{X}^w_\emptyset(\Delta) + \mathcal{E}\,, \qquad \widetilde{X}^w_\emptyset(\Delta) \;:=\; \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1} \biggl[\frac{m^2}{(G_{a_1a_1})^2} \mathcal{Z}_{a_1\cdots a_n}\biggr]\,. \tag{7.7}$$

Step (iii). We make all resolvent entries which do not contain the index a_1 independent of a_1 using (3.13). Thus, we assume that $n \geqslant 2$; if $n=1$ there is nothing to be done and this step is trivial. Using (3.13) we find

$$
\begin{aligned}
\widetilde{X}^w_{\emptyset}(\Delta) &= \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1}\biggl[\frac{m^2}{(G_{a_1a_1})^2} G_{\mu_1a_1}G_{a_1a_2}G_{a_2a_3}G_{a_3a_4}\cdots G_{a_n\mu_2}\biggr] \\
&= \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1}\biggl[\frac{m^2}{(G_{a_1a_1})^2} G_{\mu_1a_1}G_{a_1a_2}\biggl(G^{(a_1)}_{a_2a_3} + \frac{G_{a_2a_1}G_{a_1a_3}}{G_{a_1a_1}}\biggr)G_{a_3a_4}\cdots G_{a_n\mu_2}\biggr] \\
&= \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1}\biggl[\frac{m^2}{(G_{a_1a_1})^2} G_{\mu_1a_1}G_{a_1a_2}G^{(a_1)}_{a_2a_3}G_{a_3a_4}\cdots G_{a_n\mu_2}\biggr] + \mathcal{E}\,.
\end{aligned}
$$

Here the bound on the error term

$$
\sum_{a_1}^{(\mu_1\mu_2)} s_{a_1b_1}P_{a_1}\Biggl[\frac{m^2}{(G_{a_1a_1})^3}G_{\mu_1a_1}\sum_{a_2,\dots,a_n}^{(a_1\mu_1\mu_2)*} s_{a_2b_2}\cdots s_{a_nb_n}\, G_{a_1a_2}G_{a_2a_1}G_{a_1a_3}G_{a_3a_4}\cdots G_{a_n\mu_2}\Biggr]
$$

follows by first fixing the summation index a_1 and using the induction assumption combined with Lemma 3.6, similarly to Step (ii) above. The induction assumption is used on two chains: one of degree 2 (corresponding to $G_{a_1a_2}G_{a_2a_1}$) and one of degree $n-1$ (corresponding to $G_{a_1a_3}G_{a_3a_4}\cdots G_{a_n\mu_2}$); here a_1 is regarded as an external index. (Here we swept under the rug a minor technicality. Strictly speaking, the expressions encoded by different subchains do not factor, since their summations are still coupled by the constraint $a_2 \notin \{a_1, a_3, a_4, \dots, a_n\}$. As outlined above, we ignore such complications here; they are dealt with using the inclusion-exclusion argument from Lemma 9.5 in Section 9.3 by introducing a partitioning on the values of a_2, which results in a decoupled expression plus a series of small error terms.)

Next, we write

$$
\begin{aligned}
&\sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1}\biggl[\frac{m^2}{(G_{a_1a_1})^2} G_{\mu_1a_1}G_{a_1a_2}G^{(a_1)}_{a_2a_3}G_{a_3a_4}G_{a_4a_5}\cdots G_{a_n\mu_2}\biggr] \\
&\qquad = \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1}\biggl[\frac{m^2}{(G_{a_1a_1})^2} G_{\mu_1a_1}G_{a_1a_2}G^{(a_1)}_{a_2a_3}G^{(a_1)}_{a_3a_4}G_{a_4a_5}\cdots G_{a_n\mu_2}\biggr] + \mathcal{R}\,, \qquad (7.8)
\end{aligned}
$$

where the error term is

$$
\begin{aligned}
\mathcal{R} &:= \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1}\biggl[\frac{m^2}{(G_{a_1a_1})^2} G_{\mu_1a_1}G_{a_1a_2}G^{(a_1)}_{a_2a_3}\frac{G_{a_3a_1}G_{a_1a_4}}{G_{a_1a_1}}G_{a_4a_5}\cdots G_{a_n\mu_2}\biggr] \\
&= \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1}\biggl[\frac{m^2}{(G_{a_1a_1})^2} G_{\mu_1a_1}G_{a_1a_2}\biggl(G_{a_2a_3} - \frac{G_{a_2a_1}G_{a_1a_3}}{G_{a_1a_1}}\biggr)\frac{G_{a_3a_1}G_{a_1a_4}}{G_{a_1a_1}}G_{a_4a_5}\cdots G_{a_n\mu_2}\biggr]\,.
\end{aligned}
$$

We may estimate this exactly as above, by first fixing a_1 and regarding it as an external index. The induction assumption allows us to estimate the two resulting terms

$$
\sum_{a_2,\dots,a_n}^{(a_1\mu_1\mu_2)*} s_{a_2b_2}\cdots s_{a_nb_n}\, G_{a_1a_2}G_{a_2a_3}G_{a_3a_1}G_{a_1a_4}G_{a_4a_5}\cdots G_{a_n\mu_2}
$$

(a product of two subchains) and

$$\sum_{a_2,\dots,a_n}^{(a_1\mu_1\mu_2)*} s_{a_2b_2}\cdots s_{a_nb_n}\, G_{a_1a_2}G_{a_2a_1}G_{a_1a_3}G_{a_3a_1}G_{a_1a_4}G_{a_4a_5}\cdots G_{a_n\mu_2}\,,$$

(a product of three subchains). Note that the induction assumption is always used on subchains of degree strictly less than $\deg(\Delta) = n+1$. Using Lemma 3.6, one therefore finds that $\mathcal{R}$ in (7.8) can be replaced with an $\mathcal{E}$. Continuing in this manner, we eventually get

$$\widetilde{X}^w_{\emptyset}(\Delta) \;=\; \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1}\biggl[\frac{m^2}{(G_{a_1a_1})^2}G_{\mu_1a_1}G_{a_1a_2}G^{(a_1)}_{a_2a_3}G^{(a_1)}_{a_3a_4}\cdots G^{(a_1)}_{a_n\mu_2}\biggr] + \mathcal{E}\,. \tag{7.9}$$

Step (iv). We apply the identity (3.14a) to both resolvent entries with lower index a_1. This yields

$$\begin{aligned}\widetilde{X}^w_{\emptyset}(\Delta) \;&=\; m^2\sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w(\mathbf{a})\, P_{a_1}\Biggl[\sum_{d,d'}^{(a_1)} G^{(a_1)}_{\mu_1d}h_{da_1}h_{a_1d'}G^{(a_1)}_{d'a_2}G^{(a_1)}_{a_2a_3}G^{(a_1)}_{a_3a_4}\cdots G^{(a_1)}_{a_n\mu_2}\Biggr] + \mathcal{E}\\ &=\; m^2\sum_{\mathbf{a}}^{(\mu_1\mu_2)*}\sum_{d}^{(a_1)} w(\mathbf{a})s_{a_1d}\, G^{(a_1)}_{\mu_1d}G^{(a_1)}_{da_2}G^{(a_1)}_{a_2a_3}G^{(a_1)}_{a_3a_4}\cdots G^{(a_1)}_{a_n\mu_2} + \mathcal{E}\,,\end{aligned} \tag{7.10}$$

where in the second step we used that $P_{a_1}(h_{da_1}h_{a_1d'}) = \delta_{dd'}s_{a_1d}$, and that all resolvent entries are independent of a_1. Renaming $(a_1,d)\mapsto(d,a_1)$ and interchanging the order of summation, we find from (7.7) and (7.10)

$$X^w_{\emptyset}(\Delta) \;=\; m^2\sum_{d}^{(\mu_1\mu_2)} s_{b_1d}\sum_{a_1}^{(d)}\sum_{a_2,\dots,a_n}^{(d\mu_1\mu_2)*} s_{da_1}s_{b_2a_2}\cdots s_{b_na_n}G^{(d)}_{\mu_1a_1}G^{(d)}_{a_1a_2}\cdots G^{(d)}_{a_n\mu_2} + \mathcal{E}. \tag{7.11}$$

Step (v). Having performed the expectation, we now get rid of the upper indices d in all of the resolvent entries of (7.11) to go back to the original resolvent entries. To that end, we write

$$G^{(d)}_{\mu_1a_1}G^{(d)}_{a_1a_2}\cdots G^{(d)}_{a_n\mu_2} \;=\; \biggl(G_{\mu_1a_1} - \frac{G_{\mu_1d}G_{da_1}}{G_{dd}}\biggr)\biggl(G_{a_1a_2} - \frac{G_{a_1d}G_{da_2}}{G_{dd}}\biggr)\cdots\biggl(G_{a_n\mu_2} - \frac{G_{a_nd}G_{d\mu_2}}{G_{dd}}\biggr) \tag{7.12}$$

in the summand of (7.11), and multiply everything out. As before, each term results in a collection of subchains of degree strictly less than $\deg(\Delta) = n+1$, to which the induction assumption may be applied. More precisely, suppose that we have chosen the second term in $k \leqslant n+1$ of the factors in (7.12). Then we get k additional resolvent entries of G (yielding a total of $n+k+1$), as well as a collection of subchains whose total number of chain vertices is $n-k$. Notice that the index structure of every terms after multiplying (7.12) out is chain-like. The result is

$$X^w_{\emptyset}(\Delta) \;=\; m^2\sum_{d}^{(\mu_1\mu_2)} s_{b_1d}\sum_{a_1,\dots,a_n}^{(d\mu_1\mu_2)*} s_{da_1}s_{b_2a_2}\cdots s_{b_na_n}G_{\mu_1a_1}G_{a_1a_2}\cdots G_{a_n\mu_2} + \mathcal{E}\,. \tag{7.13}$$

(As before, we ignore the issues related to the cases $a_1 \in \{\mu_1, \mu_2, a_2, \dots, a_n\}$; see the inclusion-exclusion argument of Lemma 9.5.) We have the rough bound

$$\sum_{a_1,\dots,a_n}^{(d\mu_1\mu_2)*} s_{da_1} s_{b_2a_2} \cdots s_{b_na_n} G_{\mu_1a_1} G_{a_1a_2} \cdots G_{a_n\mu_2} \;\prec\; \Psi^{n+1}\Phi^{n-1} \,. \tag{7.14}$$

Indeed, by fixing a_1 and using the induction assumption on the subchain of degree n that encodes the expression $G_{a_1a_2} \cdots G_{a_n\mu_2}$, (7.14) follows by Lemma 3.6. It follows using $s_{b_1d} \leqslant M^{-1}$ that we may replace the summation $\sum_d^{(\mu_1\mu_2)}$ in (7.13) by $\sum_d$ up to an error of type $\mathcal{E}$. Finally, we may replace the sum $\sum_{a_1,\dots,a_n}^{(d\mu_1\mu_2)*}$ in (7.13) by $\sum_{a_1,\dots,a_n}^{(\mu_1\mu_2)*}$ up to an error type $\mathcal{E}$, by the inclusion-exclusion argument of Lemma 9.5. The result is

$$X_\emptyset^w(\Delta) \;=\; m^2 \sum_d s_{b_1d} \sum_{a_1,\dots,a_n}^{(\mu_1\mu_2)*} s_{da_1} s_{b_2a_2} \cdots s_{b_na_n} G_{\mu_1a_1} G_{a_1a_2} \cdots G_{a_n\mu_2} + \mathcal{E} \,. \tag{7.15}$$

Step (vi). We fix $b_2, \dots, b_n$ and regard b_1 as a free index. Define

$$X_\emptyset^w(\Delta) \;\equiv\; v_{b_1} \;:=\; \sum_{\mathbf{a}}^{(\mu_1\mu_2)*} w_{b_1b_2\dots b_n}(\mathbf{a})\, \mathcal{Z}_{a_1\dots a_n} \,.$$

Then (7.15) reads

$$v_{b_1} \;=\; (m^2 S v)_{b_1} + \mathcal{E}_{b_1} \,, \tag{7.16}$$

where $\mathcal{E}_{b_1} \prec \Psi^{n+2}\Phi^{n-1}$. (Here we use the notation $\mathcal{E}_{b_1}(b_2, \dots, b_n, \mu_1, \mu_2) \equiv \mathcal{E}(b_1, \dots, b_n, \mu_1, \mu_2)$, indicating that b_1 is the variable index and all other indices are fixed for this argument.) Inverting the self-consistent equation yields

$$v_{b_1} \;=\; \big((1 - m^2 S)^{-1} \mathcal{E}\big)_{b_1} \,.$$

In order to complete the proof, we observe that if $X_j^k \prec \Psi$ uniformly in (j, k), then for any matrix $A = (A_{ij})$ we have $\sum_j A_{ij} X_j^k \prec \|A\|_{\ell^\infty \to \ell^\infty} \Psi$ uniformly in (i, k). Recalling the definition (3.2), we therefore get

$$X_\emptyset^w(\Delta) \;=\; \big((1 - m^2 S)^{-1} \mathcal{E}\big)_{b_1} \;\prec\; \varrho\, \Psi^{n+2}\Phi^{n-1} \;\leqslant\; \Psi^{n+1}\Phi^n \,.$$

The last inequality is valid only if $\Phi < 1$, but the final bound is still correct even if $\Phi = 1$ by using the trivial bound $X_\emptyset^w(\Delta) \prec \Psi^{n+1}$. This concludes the proof. □

7.3. Completion of the induction and the proof of Proposition 5.3. We may now complete the proof of Proposition 5.3. We begin with open chains. As outlined in Section 5, the proof is by induction on $\deg(\Delta)$. The induction is started with the trivial open chain Δ corresponding to $\mathcal{Z} = G_{\mu\nu}$, for which we have the trivial bound $G_{\mu\nu} \prec \Psi$. Then (5.2) for an arbitrary open chain Δ follows by induction, using Propositions 7.1 and 7.2.

In order to prove (5.2) for an arbitrary closed chain Δ, we follow almost to the letter the arguments from Sections 7.1 and 7.2. The proof consists of two steps, each repeated twice.

(a) Prove (5.2) for $F \neq \emptyset$.

(b) Prove (5.2) for $F = \emptyset$.

The order of the argument is as follows. First we do step (a) for closed chains with one external vertex, then step (b) for closed chains with one external vertex, then step (a) again but now for closed chains with no external vertex, and finally step (b) for closed chains with no external vertex. Here no induction is required; the necessary input is (5.2) for arbitrary open chains. Each one of the four above steps uses the previous ones as input. The proof of either step (a) is almost identical to that of Proposition 7.1, and the proof of either step (b) almost identical to that of Proposition 7.2. The only nontrivial difference is associated with coinciding indices, where we may lose a factor $\Psi^2\Phi^2$ if two indices coincide. Here the worst case is the closed chain of degree two with no external vertices: $\sum_{a,b} s_{\mu a} s_{\nu b} G_{ab} G_{ba}$. The associated monomial $G_{ab}G_{ba}$ is of degree two and has two chain vertices. However, setting $a = b$ yields a contribution of order M^{-1}, i.e. we lost a factor Ψ^2 (from the two off-diagonal entries that became diagonal) and Φ^2 (from the two chain vertices). However, this loss is compensated by the gain M^{-1}: we get the bound $\Psi^2\Phi^2 + M^{-1} \leqslant 2\Psi^2\Phi^2$.

Let us sketch the general cases. Consider a closed chain Δ with no external vertices. Thus, Δ has $c(\Delta) = \deg(\Delta)$ chain vertices. If we ignore coinciding indices, we get the bound $\Psi^{\deg(\Delta)}\Phi^{\deg(\Delta)}$ on its size. On the other hand, if all of the $\deg(\Delta)$ indices coincide, we get the bound $M^{-\deg(\Delta)+1}$. Indeed, all but one of the entries of S in the chain weight can be estimated using their maximum M^{-1}; moreover, all resolvent entries are diagonal and hence of size 1. This yields the combined bound

$$\Psi^{\deg(\Delta)}\Phi^{\deg(\Delta)} + M^{-\deg(\Delta)+1} \asymp \Psi^{\deg(\Delta)}\Phi^{\deg(\Delta)}, \tag{7.17}$$

where we used that $\deg(\Delta)/2 \geqslant \deg(\Delta) - 1$. This is (5.2).

The case of a closed chain Δ with one external vertex is similar. In this case we have $c(\Delta) = \deg(\Delta) - 1$. Ignoring coinciding indices, we get the bound $\Psi^{\deg(\Delta)}\Phi^{\deg(\Delta)-1}$. On the other hand, if all indices coincide we get the bound $M^{-\deg(\Delta)+1}$ exactly as before. This yields the combined bound

$$\Psi^{\deg(\Delta)}\Phi^{\deg(\Delta)-1} + M^{-\deg(\Delta)+1} \asymp \Psi^{\deg(\Delta)}\Phi^{\deg(\Delta)-1}, \tag{7.18}$$

where we used (2.9). This is (5.2).

Note that in the bounds (7.17) and (7.18) we only considered the two extreme cases: when all summation indices are distinct, and when they all coincide. In Lemma 9.4, we prove that these bounds in fact cover all possible index configurations. The full details on coinciding indices are given in Section 9.3. This concludes the proof of (5.2), and hence of Proposition 5.3.

8. General monomials and vertex resolution

In this section we conclude the proof of Theorem 4.8 for general Δ under certain simplifying assumptions. A sketch of the argument presented in this section was given in Section 3.2.4; the key new concept is that of *vertex resolution*, which relies on Family B identities.

Our starting point is the graphical expansion from Section 6 as well as the chain estimates from Proposition 5.3. Throughout this section we assume Simplification **(S1)** from Section 6. Moreover, we shall tacitly make use of Lemma 3.6 as well as the notations and definitions from Section 6.

In order to perform the vertex resolution, it will prove necessary to expand all resolvent entries in *all* of the summation indices $\mathbf{a}$ instead of the smaller set $\mathbf{a}_F$ (as was done in Section 6). Thus, as first step, we repeat the construction of Section 6: we start with the graph $\gamma^p(\Delta)$ that encodes the p-th moment of $X^w_F(\Delta)$, and perform the expansion given after Definition 6.4, except we now expand in the full set $\mathbf{a}$ of summation indices instead of $\mathbf{a}_F$. This gives rise to a family of graphs which we denote by $\mathfrak{G}^p_F(\Delta)$. Note that $\mathfrak{G}_F^p(\Delta) \supset \widetilde{\mathfrak{G}}_F^p(\Delta)$, where, we recall, the set $\widetilde{\mathfrak{G}}_F^p(\Delta)$ is the set generated in Section 6 by expanding in the indices $\mathbf{a}_F$ only. Thus, each graph $\Gamma \in \mathfrak{G}_F^p(\Delta)$ encodes a monomial of entries of $\mathcal{G}$, the edge (x,y) giving rise to the maximally expanded entry $\mathcal{G}_{xy}^{(\mathbf{a}\setminus\{x,y\})}$ or $\mathcal{G}_{xy}^{(\mathbf{a}\setminus\{x,y\})*}$ depending on its colour. Here, and throughout the following, we use the phrase *maximally expanded* to mean maximally expanded in $\mathbf{a}$ (see Definition 6.4). As in Section 6, we do not keep track of the Q's in our notation. (Indeed, this information will turn out to be unimportant for our proof.)

Note that Definition 6.6 carries over verbatim for $\Gamma \in \mathfrak{G}_F^p(\Delta)$. For the following we pick and fix a $\Gamma \in \mathfrak{G}_F^p(\Delta)$. Thus, Γ encodes a monomial, whereby each edge of Γ gives rise to a maximally expanded entry of $\mathcal{G}$. As explained in Section 6, the linking procedure used to make all entries maximally expanded ensures, thanks to the presence of the Q's, that $|E(\Gamma)| \geqslant p(\deg(\Delta) + |F|)$.

The main idea of vertex resolution already appeared in Section 3.2.4. Roughly, we *resolve* each summation vertex using the Family B identities (3.14a) and (3.14b), which results in a new set of summation vertices which we call fresh and draw using white dots. We call the resulting graph Θ. (More precisely, from each graph Γ we get a finite family of new graphs $\{\Theta_\alpha\}$.) Next, we take the expectation, which results in a summation over all pairings (in fact, more generally, over all lumpings) of the white vertices adjacent to the original summation vertex. Each pairing gives rise to a new graph, which we call Υ. (As above, from each graph Θ we get a finite family of new graphs $\{\Upsilon_\alpha\}$.) Although each steps results in an increase in the number of graphs, it is easy to see that this combinatorial factor is bounded by a constant depending only on $|V(\Gamma)|$, the number of vertices in Γ. In other words, the above families $\{\Theta_\alpha\}$ and $\{\Upsilon_\alpha\}$ are finite and do not depend on N.

The step $\Gamma \mapsto \Theta$ is performed in Section 8.2, and the step $\Theta \mapsto \Upsilon$ in Section 8.3. Figure 8.1 contains a summary of this process, on the level of a single vertex, which is helpful to keep in mind while reading the following. The idea is that, provided the vertex being resolved arose as a copy of a charged vertex (see Definition 4.7), the resolution process will (in leading order) generate at least one fresh chain vertex. From this chain vertex we shall gain a factor Φ by invoking Proposition 5.3, and this will conclude the proof.

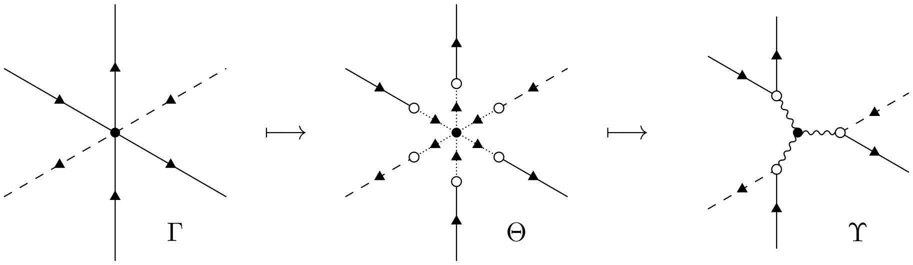

FIGURE 8.1. A graphical overview of vertex resolution on the level of a single vertex. (We do not draw the other vertices.) In the second step we draw only one of the possible six pairings (in the Hermitian case).

Note that the notion of charged vertex can be lifted from Δ to Γ using the projection π (see Definition 8.1 below). The following class of vertices, called *marked vertices*, plays the central role in this section. Informally, a marked vertex $i \in V_s(\Gamma)$ is a charged vertex that, in the construction of Γ from $\gamma^p(\Delta)$, was linked to by the smallest allowed number of edges (zero if $\pi(i) \notin F$ and one if $\pi(i) \in F$).

Recall that we intend to gain an additional factor Φ from any charged vertex of Γ. However, the possibility of doing this may be destroyed if a charged vertex was linked to in the construction of Γ from $\gamma^p(\Delta)$. In this case we gain from the linking, as always, but we may not additionally gain from the vertex's being charged. If the vertex was excessively linked (i.e. more than minimally required), then we have the additional gain from this extra linking. Marked vertices are exactly those charged vertices which have been minimally linked. Thus, in order to gain from a marked vertex we cannot use the simple power counting that underlies linking, but need the more refined mechanism of vertex resolution.

Definition 8.1 (Charged and Marked vertices in Γ). The set of *charged vertices* of Γ is by definition $V_c(\Gamma) := \pi^{-1}(V_c(\Delta))$ (see Definition 4.7).

The vertex $i \in V_c(\Gamma)$ is called *marked* if either

(i) $\pi(i) \notin F$ and $\deg_\Gamma(i) = \deg_\Delta(\pi(i))$, or

(ii) $\pi(i) \in F$ and $\deg_\Gamma(i) = \deg_\Delta(\pi(i)) + 2$.

We denote the set of marked vertices by $V_m(\Gamma) \subset V_c(\Gamma)$.

Note that (i) corresponds to the case where i was not linked to at all, and (ii) to the case where i was linked to exactly once. The following lemma gives a lower bound on the number of edges of Γ. Informally, it states that if i is not marked but $\pi(i)$ is charged then i was linked to at least once more than the minimum required amount (zero if $\pi(i) \notin F$ and one if $\pi(i) \in F$).

Lemma 8.2. *We have the bound*

$$|E(\Gamma)| \;\geqslant\; p(\deg(\Delta) + |F|) + |V_c(\Gamma)| - |V_m(\Gamma)|\,. \tag{8.1}$$

Proof. By definition of $V_c(\Gamma)$ and $V_m(\Gamma)$, we find that $i \in V_s(\Gamma)$ was linked to at least once if

$$i \;\in\; \pi^{-1}(F^c) \cap V_c(\Gamma) \cap V_m(\Gamma)^c \qquad \text{or} \qquad i \;\in\; \pi^{-1}(F) \cap \bigl(V_c(\Gamma) \cap V_m(\Gamma)^c\bigr)^c\,,$$

and i was linked to at least twice if

$$i \;\in\; \pi^{-1}(F) \cap V_c(\Gamma) \cap V_m(\Gamma)^c\,.$$

Since each linking adds an edge to the $p\deg(\Delta)$ edges of $\gamma^p(\Delta)$, we find

$$\begin{aligned} |E(\Gamma)| \;&\geqslant\; p\deg(\Delta) + \Bigl|\pi^{-1}(F^c) \cap V_c(\Gamma) \cap V_m(\Gamma)^c\Bigr| + \Bigl|\pi^{-1}(F) \cap \bigl(V_c(\Gamma) \cap V_m(\Gamma)^c\bigr)^c\Bigr| \\ &\qquad + 2\Bigl|\pi^{-1}(F) \cap V_c(\Gamma) \cap V_m(\Gamma)^c\Bigr| \\ &=\; p\deg(\Delta) + p|F| + p|V_c(\Delta)| - |V_m(\Gamma)|\,, \end{aligned}$$

where in the last step we used that π is p-to-one and $V_m(\Gamma) \subset V_c(\Gamma)$. □

The goal of this section is to gain an extra factor Φ from each marked vertex of Γ using vertex resolution. Provided we can do this, the proof of Theorem 4.8 will be complete. This can be informally understood as follows. In order to get the estimate (4.7), we need a bound of size $\Psi^{p(\deg(\Delta)+|F|)}\Phi^{p|V_c(\Delta)|}$. Each vertex $i \in V_s(\Gamma)$ contributes factors Ψ and Φ (in addition to the trivial $p\deg(\Delta)$) to the estimate as follows.

(i) $\pi(i) \notin F$ and $i \notin V_c(\Gamma)$. In this case i yields no factor Ψ or Φ.

(ii) $\pi(i) \notin F$ and $i \in V_c(\Gamma)$. If $\deg_\Gamma(i) = \deg_\Delta(\pi(i))$ then i is marked and will yield a factor Φ by vertex resolution. If $\deg_\Gamma(i) > \deg_\Delta(\pi(i))$ then i is not marked but carries an extra factor Ψ since it has been linked to more times than needed. (Thus, Γ has at least one extra edge, corresponding to an off-diagonal entry $G_{uv} \prec \Psi$, incident to i).

(iii) $\pi(i) \in F$ and $i \notin V_c(\Gamma)$. In this case i has been linked to at least once and is consequently incident to at least one extra edge. This yields a factor Ψ.

(iv) $\pi(i) \in F$ and $i \in V_c(\Gamma)$. As in (iii), i has been linked to at least once and hence yields a factor Ψ. In addition, i yields a second factor Φ as follows. If $\deg_\Gamma(i) = \deg_\Delta(\pi(i)) + 2$ then i is marked and will yield an extra factor Φ by vertex resolution. If $\deg_\Gamma(i) > \deg_\Delta(\pi(i)) + 2$ then i has been linked to at least twice, hence yielding a second factor Ψ. In either case the vertex i generates a factor $\Psi\Phi$.

Thus, from each case (i) – (iv) we gain ℓ_Ψ factors Ψ and ℓ_Φ factors Φ in addition to the trivially available $p\deg(\Delta)$ factors of Ψ, where the values of ℓ_Ψ and ℓ_Φ is as follows: (i) $\ell_\Psi = \ell_\Phi = 0$, (ii) $\ell_\Psi = 0$, $\ell_\Phi = 1$, (iii) $\ell_\Psi = 1$, $\ell_\Phi = 0$, (iv) $\ell_\Psi = \ell_\Phi = 1$. From this the bound $\Psi^{p(\deg(\Delta)+|F|)}\Phi^{p|V_c(\Delta)|}$ follows immediately.

Before moving on to the main argument of this section, we outline how a marked vertex yields an additional factor Φ. We claim that Definitions 4.7 and 8.1 imply that

$$i \in V_s(\Gamma) \text{ is marked} \qquad \Longrightarrow \qquad \nu_i(\Gamma) \;\neq\; \nu_i^*(\Gamma) \tag{8.2}$$

(see Definition 4.6). To see this, let i be an arbitrary marked vertex. If $\pi(i) \notin F$ then by Definition 8.1 i has not been linked to, and hence $\nu_i^\xi(\Gamma) = \nu_{\pi(i)}^\xi(\Delta)$ for $\xi = 1, *$. Therefore (8.2) follows from Definition 4.7. On the other hand, if $\pi(i) \in F$ then by Definition 8.1 i has been linked to once, and either (a) $\nu_i(\Gamma) = \nu_{\pi(i)}(\Delta)$ and $\nu_i^*(\Gamma) = \nu_{\pi(i)}^*(\Delta) + 2$ or (b) $\nu_i(\Gamma) = \nu_{\pi(i)}(\Delta) + 2$ and $\nu_i^*(\Gamma) = \nu_{\pi(i)}^*(\Delta)$. Either way, Definition 4.7 yields (8.2).

Roughly, vertex resolution splits each summation vertex of degree $2d$ (we assume for simplicity that the vertex is of even degree) into d fresh summation vertices of degree two. If the right-hand side of (8.2) holds (as it does if we are resolving a marked vertex), at least one of the fresh summation vertices will be a chain vertex (i.e. both of its incident edges will have the same colour). The desired gain of a factor Φ will then come from an application of Proposition 5.3.

8.1. General graphical representation.

Throughout Sections 8 and 9, we shall fix $\mathbf{a}$ and apply three algebraic operations to the monomial encoded by Γ:

(i) Family A identities,

(ii) Family B identities,

(iii) partial expectation in $\mathbf{a}$.

In order to keep track of the structure of the ensuing terms, we make heavy use of graphs. For pedagogical reasons, we shall develop the algebraic and graphical languages in parallel. Each algebraic expression (a monomial in the matrix entries of G, $\mathcal{G}$, H, and S) is represented by a graph. Application of one of the three elementary algebraic identities listed above can, as before, be described by a elementary transformation on graphs. Although the entire argument could be stated in terms of graphs alone, this would rather obscure the underlying mechanism, which always corresponds to applying one of the three algebraic operations listed above. Instead, we introduce each graph operation when it naturally arises in our argument. In order to obtain a set of graphs that is closed under all of the operations we shall need, we extend our set of graphs according to the following definition.

Definition 8.3 (General graph). By a *graph* we mean a quintuple (V_f, V_s, V_e, E, ξ) with the following properties. The set E is a set of edges on the vertex set $V := V_f \sqcup V_s \sqcup V_e$. Multiple edges as well as loops are allowed. The colouring ξ is a map

$$\xi : E \longrightarrow \big\{\text{solid, dashed, dotted, wiggly}\big\}\,.$$

As in Definition 4.1, we sometimes use the alternative notations solid $\equiv 1$ and dashed $\equiv *$.

An edge that is solid or dashed is called a *resolvent edge*. Dotted and resolvent edges are directed, while wiggly edges are undirected. The vertices in V_s are called the *original summation vertices*, in V_f the *fresh summation vertices*, and in V_e the *external vertices*. Vertices in V_f are drawn using white dots and vertices in $V_s \sqcup V_e$ using black dots.

Figure 3.8 contains the dictionary of the colour-code: a solid edge encodes an entry of G, a dashed edge an entry of G^*, a dotted edge an entry of H, and a wiggly edge an entry of S. More precisely, each resolvent edge encodes a *maximally expanded* (in $\mathbf{a}$) resolvent entry (see Definition 6.4), and each dotted edge an $\mathbf{a}$-*admissible entry of* H, which is the subject of the following definition.

Definition 8.4. The entry h_{uv} is an $\mathbf{a}$-*admissible entry of* H if $u \in \mathbf{a}$ or $v \in \mathbf{a}$.

The arguments below consist of a series of operations on the set of graphs from Definition 8.3. To be completely precise, below we shall in fact adorn the graphs from Definition 8.3 with *decorations*: resolvent loops may be decorated with a black or white diamond (see Figure 3.3), wiggly edges with an arbitrary number of crossing strokes, and original summation vertices with an arbitrary number of rings. (The latter two concepts are defined above Definition 8.5 and in the beginning of the proof of Lemma 9.3 respectively.)

To each vertex $i \in V(\Gamma)$ of a graph Γ we assign an index u_i. We shall consistently use the splitting

$$\mathbf{u} \;=\; (u_i)_{i \in V(\Gamma)} \;=\; (\mathbf{x}, \mathbf{a}, \boldsymbol{\mu})\,, \qquad \mathbf{x} \;=\; (x_i)_{i \in V_f(\Gamma)}\,, \qquad \mathbf{a} \;=\; (a_i)_{i \in V_s(\Gamma)}\,, \qquad \boldsymbol{\mu} \;=\; (\mu_i)_{i \in V_e(\Gamma)}\,.$$

We say that the index u_i is *associated with* the vertex i, and vice versa. We introduce the notation

$$\mathcal{A}(\Gamma) \;\equiv\; \mathcal{A}_{\mathbf{a},\mathbf{x}}(\Gamma) \tag{8.3}$$

for the monomial (in the entries of $G^{(T)}$, $\mathcal{G}^{(T)}$, H, and S where $T \subset \mathbf{a}$) encoded by the graph Γ. Note that $\mathcal{A}(\Gamma)$ has an explicit formula, analogous to (4.2) except needing much heavier notation. As we shall not need it, we shall not give it. (Note also that our graphs do not keep track of any factors of Q, as we shall not need this information.)

8.2. Generation of the fresh summation vertices. Next, we define the vertex resolution operation precisely. Our starting point is a fixed $\Gamma \in \mathfrak{G}_F^p(\Delta)$. In order to streamline the argument, we at first make the following simplifying assumption on Δ, which is removed in Section 9.

(S2) There are no diagonal entries $\mathcal{G}_{aa} = G_{aa} - m$ in $\mathcal{Z}(\Delta)$. (I.e. Δ has no loops.)

The operation of vertex resolution consists of two main steps: the *generation* and *lumping* of fresh summation vertices. The idea behind the first step – the generation of fresh summation vertices – is to resolve, using the Family B identities (3.14a) and (3.14b), each (already maximally expanded) off-diagonal resolvent entry $G_{uv}^{(\mathbf{a}\setminus\{u,v\})}$, with $u \neq v$, in any summation index from the set $\{u, v\}$. (The word "resolve" here refers to explicitly identifying the dependence on all matrix entries h_{ij} with $i, j \in \mathbf{a}$ so that partial expectation in these variables can be taken. After taking the partial expectation, we shall get expressions that can again be represented by admissible graphs.) More precisely, we write

$$G_{uv}^{(\mathbf{a}\setminus\{u,v\})} = \begin{cases} -G_{uu}^{(\mathbf{a}\setminus\{u\})} \sum_x^{(\mathbf{a})} h_{ux} G_{xv}^{(\mathbf{a})} & (u \text{ summation and } v \text{ external})\,, \\ -G_{vv}^{(\mathbf{a}\setminus\{v\})} \sum_x^{(\mathbf{a})} G_{ux}^{(\mathbf{a})} h_{xv} & (u \text{ external and } v \text{ summation})\,, \\ G_{uu}^{(\mathbf{a}\setminus\{u,v\})} G_{vv}^{(\mathbf{a}\setminus\{v\})} \big(-h_{uv} + \sum_{x,y}^{(\mathbf{a})} h_{ux} G_{xy}^{(\mathbf{a})} h_{xv}\big) & (u \text{ and } v \text{ summation})\,. \end{cases} \tag{8.4}$$

(As stated after Definition 4.1, we exclude the trivial case where both u and v are external indices.) The proof of (8.4) is a straightforward consequence of the identities (3.14a) and (3.14b), and the fact that $u \neq v$ by definition of $X_F^w(\Delta)$. For example, if a and b are summation indices and μ and ν are external indices, we write

$$\begin{aligned} & G_{\mu a}^{(b)} G_{ab} G_{b\mu}^{(a)*} G_{a\nu}^{(b)} G_{\nu a}^{(b)*} \\ &\quad = G_{aa}^{(b)} G_{aa} G_{bb}^{(a)} G_{bb}^{(a)*} G_{aa}^{(b)} G_{aa}^{(b)*} \sum_{x,y,z,u,v,w}^{(ab)} G_{\mu x}^{(ab)} h_{xa} h_{ay} G_{yz}^{(ab)} h_{zb} h_{bu} G_{u\mu}^{(ab)*} h_{av} G_{v\nu}^{(ab)} G_{\nu w}^{(ab)*} h_{wa} \\ &\qquad - G_{aa}^{(b)} G_{aa} G_{bb}^{(a)} G_{bb}^{(a)*} G_{aa}^{(b)} G_{aa}^{(b)*} \sum_{x,u,v,w}^{(ab)} G_{\mu x}^{(ab)} h_{xa} h_{ab} h_{bu} G_{u\mu}^{(ab)*} h_{av} G_{v\nu}^{(ab)} G_{\nu w}^{(ab)*} h_{wa}\,. \end{aligned} \tag{8.5}$$

Notice that along this procedure we may generate non-maximally expanded diagonal terms, (e.g. G_{aa} above), but off-diagonal terms always have upper indices $\mathbf{a}$. Moreover, all entries of H on the right-hand side of (8.5) are $\mathbf{a}$-admissible.

At this point we make the following further simplification, which leaves the essence of the argument unchanged but removes some technicalities.

(S3) We replace any diagonal term $G_{aa}^{(T)}$ with m and any diagonal term $G_{aa}^{(T)*}$ with $\bar{m}$. (Recall that $G_{aa}^{(T)} \approx m$ in the sense that $G_{aa}^{(T)} - m \prec \Psi$ by definition of Λ.) This replacement is done in two places: in $\mathcal{A}(\Gamma)$ and in the identities (3.14a) and (3.14b) which underlie the algebra of vertex resolution.

Again, Simplification **(S3)** is removed in Section 9. Thus, under Simplification **(S3)**, we neglect all diagonal terms in (8.5). (More precisely, we replace each one of them with m or $\bar{m}$; the resulting powers of m and $\bar{m}$ are irrelevant for estimates by (3.11).)

The use of graphs greatly simplifies the analysis of complicated expressions like (8.5). The identities (8.4) all have obvious graphical representations. In Figure 8.2 we give a graphical depiction of (8.5). (Recall the conventions introduced in Figure 3.8.) In the typical case, the summation vertex of degree four associated with a gives rise to four fresh summation vertices, associated with x, y, v, w; likewise the summation vertex of degree two associated with b creates two fresh vertices associated with u and z (first term in the right side of (8.5)). Due to presence of the term h_{uv} in (8.4), sometimes two summation indices are directly connected with a dotted line at the expense of one fewer fresh summation vertex adjacent to each of these two indices. This results in the second term on the right-hand side of (8.5), which contains a factor h_{ab} (note that G_{ab} plays the role of $G_{uv}^{\mathbf{a}\setminus\{u,v\}}$ from (8.4)).

The generation of fresh summation vertices for a general $\Gamma \in \mathfrak{G}_F^p(\Delta)$ is no different from the above example. Applying the graphical rules associated with (8.4) to each edge of Γ, we get a finite family of graphs which we denote by $\widetilde{\mathfrak{R}}(\Gamma)$. In accordance with Simplifications **(S2)** and **(S3)**, in this section we drop

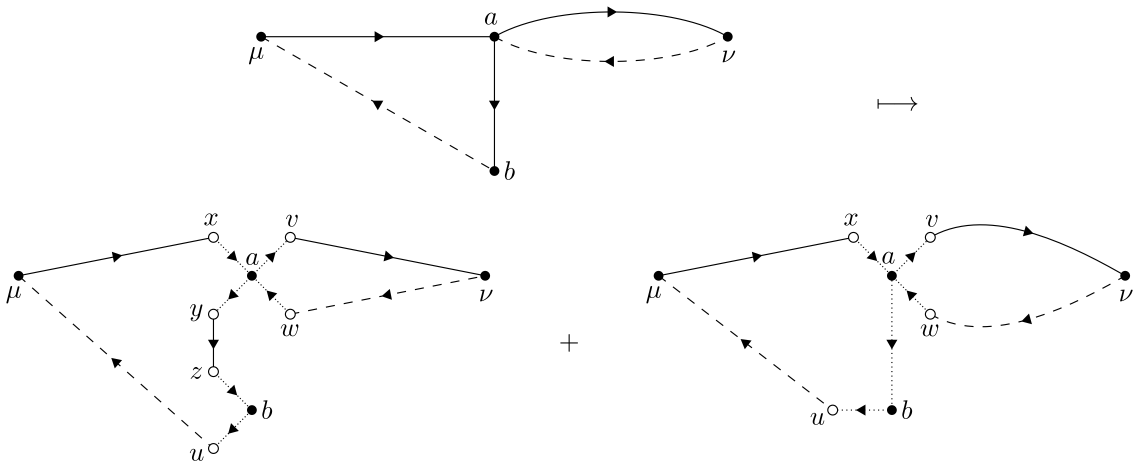

Figure 8.2. The vertex resolution from (8.5). The graph Γ is represented on the top and both graphs of $\widetilde{\mathfrak{R}}(\Gamma)$ are represented on the bottom. In accordance with Simplification **(S3)** we do not draw the diagonal entries of G.

all diagonal terms, and hence all loops from the graphs in $\widetilde{\mathfrak{R}}(\Gamma)$. (In Section 9 below we keep track of the loops, which will lead to the larger set $\mathfrak{R}$.)

Any graph $\Theta \in \widetilde{\mathfrak{R}}(\Gamma)$ has the following properties.

(i) Θ has only straight, dashed or dotted edges but no wiggly edge (see Figure 3.8).

(ii) Θ has no loops or multiple edges.

(iii) The sets $V_s(\Theta) = V_s(\Gamma)$ and $V_e(\Theta) = V_e(\Gamma)$ remain unchanged, as do the associated indices $\mathbf{a}$ and $\boldsymbol{\mu}$. In addition, we now have a new set of vertices, $V_f(\Theta) \neq \emptyset$, which indexes the fresh summation vertices $\mathbf{x}$.

(iv) Each resolvent edge of Θ encodes an entry of $G^{(\mathbf{a})}$ or $G^{(\mathbf{a})*}$ in $\mathcal{A}(\Theta)$ (in particular, no resolvent edge of Θ is incident to $V_s(\Theta)$). Each dotted edge of Θ encodes an $\mathbf{a}$-admissible entry of H in $\mathcal{A}(\Theta)$.

(v) For $i, j \in V_s(\Theta)$ let the symmetric function $\sigma(i,j)$ denote the number of dotted edges joining i and j. The number of edges in Γ and Θ is conserved in the sense that

$$|E(\Gamma)| \;=\; \sum_{e \in E(\Theta)} \big(\mathbf{1}(\xi_e = 1) + \mathbf{1}(\xi_e = *)\big) + \frac{1}{2} \sum_{i,j \in V_s(\Theta)} \sigma(i,j)\,. \tag{8.6}$$

Informally: in the process $\Gamma \mapsto \Theta$ that generates fresh summation vertices, each resolvent entry either remains a resolvent entry or is replaced with an entry of H with original summation vertices. This simply corresponds to the two terms in the last line of the right-hand side of (8.4).

(vi) Each original summation vertex is incident only to dotted edges.

(vii) Each fresh summation vertex has degree two and is incident to precisely one dotted edge.

We remark that, by construction, the sets $\mathbf{x}$ and $\mathbf{a}$ are disjoint, as are the sets $\mathbf{a}$ and $\boldsymbol{\mu}$. However, $\mathbf{x}$ and $\boldsymbol{\mu}$ are in general not disjoint, and indices of $\mathbf{x}$ may coincide.

8.3. Lumping of the fresh summation vertices. We now take the expectation of $\mathcal{A}(\Theta)$. In fact, all that we shall need is the partial expectation in $\mathbf{a}$. The key observation is that, by Property (iv) in Section 8.2, each resolvent entry of $\mathcal{A}(\Theta)$ is independent of $\mathbf{a}$ and each entry of H is $\mathbf{a}$-admissible. In particular, the partial expectation $\prod_{a \in \mathbf{a}} P_a$ acting on $\mathcal{A}(\Theta)$ acts on the product of the entries of H alone. Since this product is an explicit monomial, we can evaluate its expectation directly. If the random variables h_{uv} were Gaussian, this would correspond to a simple Wick-pairing of the dotted edges. Pairing two dotted edges, each of them incident to a fresh summation vertex and a common original summation vertex, results in a pairing of two fresh summation vertices. Since $\mathbf{x}$ and $\mathbf{a}$ are distinct, a dotted edge incident to a fresh summation vertex cannot be paired with a dotted edge incident to two original summation vertices. In the non-Gaussian case, higher-order moments are also present, but they are suppressed by a combinatorial factor (i.e. a positive power of M^{-1}). Graphically, we represent the procedure of taking expectation by pairing (or in general lumping) fresh summation indices, and replace the corresponding doubled dotted line by a wiggly line. What follows is a more precise description.

We define the second step of vertex resolution – the lumping of the fresh summation vertices. Before giving the general procedure, we complete the analysis of the example (8.5). From (8.5), assuming Simplification **(S3)**, we get

$$\begin{aligned}
&\mathbb{E} G^{(b)}_{\mu a} G_{ab} G^{(a)*}_{b\mu} G^{(b)}_{a\nu} G^{(b)*}_{\nu a} \\
&\overset{\textbf{(S3)}}{=} \; m^4 \bar{m}^2 \sum_{x,y,z,u,v,w}^{(ab)} \mathbb{E} G^{(ab)}_{\mu x} G^{(ab)}_{yz} G^{(ab)*}_{u\mu} G^{(ab)}_{v\nu} G^{(ab)*}_{\nu w} P_a\big(h_{xa} h_{ay} h_{av} h_{wa}\big) P_b(h_{zb} h_{bu}) \\
&\qquad - m^4 \bar{m}^2 \sum_{x,u,v,w}^{(ab)} \mathbb{E} G^{(ab)}_{\mu x} G^{(ab)*}_{u\mu} G^{(ab)}_{v\nu} G^{(ab)*}_{\nu w} h_{av} h_{wa} h_{xa} P_b\big(h_{ab} h_{bu}\big) \,, \qquad (8.7)
\end{aligned}$$

where we used the trivial identity $\mathbb{E} X = \mathbb{E} P_a P_b X$ and the fact that $G^{(ab)}$ is independent of a and b (see Definition 2.3). Since $a \neq u$, the partial expectation P_b on the second line of (8.7) vanishes. The partial expectations on the first line of (8.7) may be computed explicitly, similarly to the computation (3.38):

$$\begin{aligned}
&P_a\big(h_{xa} h_{ay} h_{av} h_{wa}\big) P_b(h_{zb} h_{bu}) \\
&\qquad = \; s_{ax} s_{av} s_{bz} \delta_{xy} \delta_{vw} \delta_{zu} + s_{ax} s_{ay} s_{bz} \delta_{xv} \delta_{yw} \delta_{zu} + \mathbf{1}(x = y = v = w) s_{ax}^2 s_{bz} \delta_{zu} \mathbb{E} |\zeta_{ax}|^4 \,.
\end{aligned}$$

Here we assumed the condition (3.21). In the case (3.20), we get the additional term

$$s_{ax} s_{ay} \delta_{xw} s_{bz} \delta_{yv} \delta_{zu} \,.$$

Thus we may write (in the case (3.21) for simplicity)

$$\begin{aligned}\mathbb{E}G^{(b)}_{\mu a}G_{ab}G^{(a)*}_{b\mu}G^{(b)}_{a\nu}G^{(b)*}_{\nu a} \overset{\textbf{(S3)}}{=} {} & m^4\bar{m}^2\mathbb{E}\sum_{x,v,z}^{(ab)} s_{ax}s_{av}s_{bz}\, G^{(ab)}_{\mu x}G^{(ab)}_{xz}G^{(ab)*}_{z\mu}G^{(ab)}_{v\nu}G^{(ab)*}_{\nu v} \\ & + m^4\bar{m}^2\mathbb{E}\sum_{x,y,z}^{(ab)} s_{ax}s_{ay}s_{bz}\, G^{(ab)}_{\mu x}G^{(ab)}_{yz}G^{(ab)*}_{z\mu}G^{(ab)}_{x\nu}G^{(ab)*}_{\nu y} \\ & + m^4\bar{m}^2\mathbb{E}\sum_{x,z}^{(ab)} s_{ax}^2 s_{bz}\big(\mathbb{E}|\zeta_{ax}|^4\big)\, G^{(ab)}_{\mu x}G^{(ab)}_{xz}G^{(ab)*}_{z\mu}G^{(ab)}_{z\nu}G^{(ab)*}_{\nu x}\,. \end{aligned} \tag{8.8}$$

Note that the summations on the right-hand side are performed with respect to weights (see Definition 4.4). Now it is apparent how better estimates are available for each term on the right-hand side than the term on the left-hand side. Indeed, all terms contain five off-diagonal entries of G. In addition, however, the first two terms on the right-hand side contain a summation index associated with a chain vertex x (see Definition 5.1) which is summed over with respect to the chain weight s_{ax} (see Definition 5.2). The last term is suppressed by an additional factor $s_{ax} \leqslant M^{-1}$. Invoking Proposition 5.3 (and neglecting the upper indices (ab) which are dealt with easily in the full proof below), we find that the arguments of $\mathbb{E}$ on the right-hand side of (8.8) are all $O_\prec(\Psi^5\Phi)$. Here the extra factor Φ was extracted from the resolution of the vertex i associated with the summation variable a, and arose from the fact that $\nu_i(\Gamma) \neq \nu_i^*(\Gamma)$.

Again, the lumping of fresh summation vertices is best represented graphically. In order to represent terms like the last term on the right-hand side of (8.8) graphically, we represent the expression $\mathbb{E}|h_{ij}|^{2+k}$ using a wiggly line crossed by k strokes. Thus, a wiggly edge may be either *uncrossed* or *crossed.* We shall always use the bound

$$P_i|h_{ij}|^{2+k} \;=\; s_{ij}^{1+k/2}\mathbb{E}|\zeta_{ij}|^{2+k} \;\leqslant\; s_{ij}M^{-k/2}$$

in combination with crossed wiggly lines. See Figure 8.3 for a graphical depiction of (8.8).

The general case is similar to the example (8.8).

Definition 8.5 (Lumping). A *lumping* is a partition whose blocks, called *lumps*, have size greater than or equal to two.

Let $\Gamma \in \mathfrak{G}^p_F(\Delta)$ and $\Theta \in \widetilde{\mathfrak{R}}(\Gamma)$. From Θ we generate a finite family of graphs Υ by taking all lumpings of the white vertices (fresh summation vertices) of Θ. (This lumping, or identification of vertices, arises from taking the partial expectation $\prod_{a\in\mathbf{a}} P_a$ of all entries of H in $\mathcal{A}(\Theta)$. Note that each fresh summation vertex of Θ must be lumped with at least another one because h_{ij} and h_{kl} are independent if $\{i,j\} \neq \{k,l\}$, and $P_i h_{ij} = 0$.) Thus, the result of the lumping is to merge some white vertices into lumps, where each lump consists of at least two vertices, and is again represented as a single white vertex. We denote by $\mathfrak{L}(\Theta)$ the set of such graphs Υ obtained by lumping the fresh summation vertices of Θ. Recall that all resolvent matrix entries encoded by the resolvent edges of Θ have upper indices $\mathbf{a}$. Hence any factors Q_a with $a \in \mathbf{a}$ act trivially on them according to the identity $Q_a\big(G^{(\mathbf{a})}_{ij}X\big) = G^{(\mathbf{a})}_{ij}Q_aX$ for all $a \in \mathbf{a}$. Therefore the factors Q only act on entries of H. As in the example (3.36), they simply forbid some pairings; this restriction is no importance for us, and we shall estimate the contribution of arbitrary pairings. Note that after the partial expectation in $\mathbf{a}$ has been taken, no factors Q remain. In particular, $\mathcal{A}(\Upsilon)$ has no factors Q.

Any graph $\Upsilon \in \mathfrak{L}(\Theta)$ has the following properties.

(i) Υ has no loops or multiple edges.

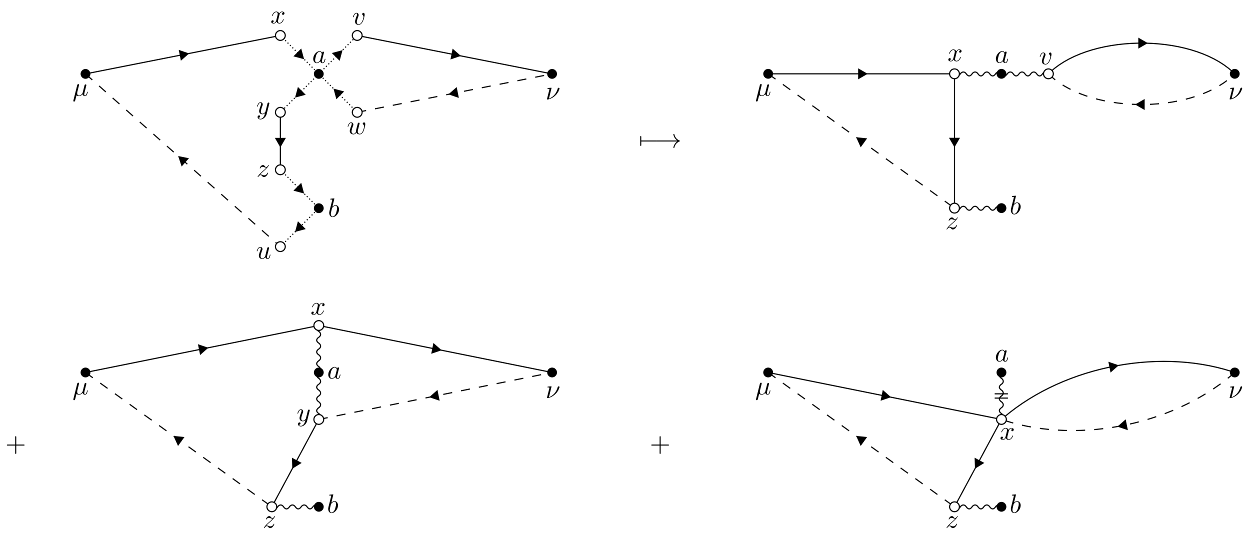

FIGURE 8.3. The lumping of the fresh summation vertices in the example (8.5). The graph on the left-hand side is Θ (representing the first term on the right-hand side of (8.7)), and the three graphs on the right-hand side are the elements of $\mathfrak{L}(\Theta)$ (representing the three terms on the right-hand side of (8.8)).

(ii) The vertex set of Υ is a disjoint union $V(\Upsilon) = V_f(\Upsilon) \sqcup V_s(\Upsilon) \sqcup V_e(\Upsilon)$, where $V_s(\Upsilon) = V_s(\Theta) = V_s(\Gamma)$ is the set of original summation vertices, and $V_e(\Upsilon) = V_e(\Theta) = V_e(\Gamma)$ is the set of external summation vertices. (The set of fresh summation vertices $V_f(\Upsilon)$ is strictly smaller than $V_f(\Theta)$.)

(iii) A directed edge (i,j) of Θ has one of two colours: solid (encoding $G_{ij}^{(\mathbf{a})}$) or dashed (encoding $G_{ij}^{(\mathbf{a})*}$). An undirected edge is always wiggly. An uncrossed wiggly edge $\{i,j\}$ encodes $\mathbb{E}|h_{ij}|^2 = s_{ij}$, and a wiggly edge $\{i,j\}$ crossed by k strokes encodes $\mathbb{E}|h_{ij}|^{2+k} = s_{ij}^{1+k/2}\mathbb{E}|\zeta_{ij}|^{2+k}$.

(iv) Each $i \in V_f(\Upsilon)$ is adjacent (via a wiggly line) to a unique vertex $p(i) \in V_s(\Upsilon)$. Thus, the map $p : V_f(\Upsilon) \to V_s(\Upsilon)$ is a projection which associates with each fresh summation vertex i its "parent" original summation vertex $p(i)$. (E.g. in the first graph on the right-hand side of Figure 8.3 we have $p(x) = p(v) = a$ and $p(z) = b$.)

(v) For each $i, j \in V_s(\Upsilon)$, we have $\sigma(i,j) \in \{0, 2, 3, 4, \dots\}$. (Recall that $\sigma(i,j)$ is the number of dotted lines in the graph Θ between vertices $i, j \in V_s(\Upsilon) = V_s(\Theta)$.) If $\sigma(i,j) \geqslant 2$ then i and j are connected by a wiggly line crossed by $\sigma(i,j) - 2$ strokes. Moreover, the number of edges is conserved in the sense that

$$|E(\Gamma)| \;=\; \sum_{e \in E(\Upsilon)} \big(\mathbf{1}(\xi_e = 1) + \mathbf{1}(\xi_e = *)\big) + \frac{1}{2} \sum_{i,j \in V_s(\Upsilon)} \sigma(i,j)\,, \tag{8.9}$$

as follows from (8.6).

See Figure 8.4 for an illustration of (v).

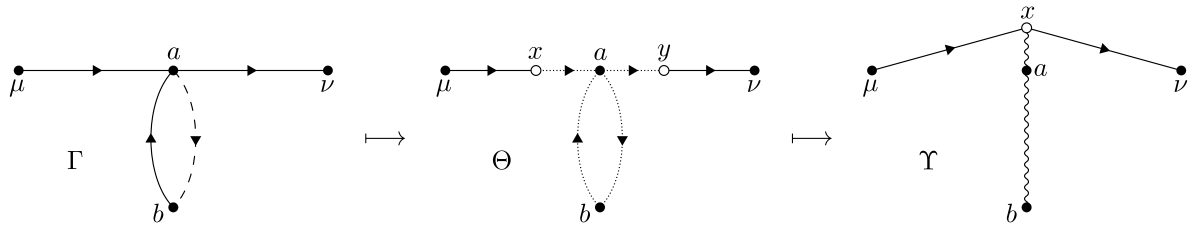

FIGURE 8.4. The resolution process $\Gamma \mapsto \Theta \mapsto \Upsilon$, where in the first step we chose a Θ satisfying $\sigma(i,j) = 2$ where i and j are the original summation vertices associated with a and b respectively. This figure illustrates the property (v) above, as well as the conservation of the number of edges from (8.6) and (8.9).

8.4. Summing over the fresh summation indices and completion of the proof. The complete process of vertex resolution may be summarized as

$$\Gamma \in \mathfrak{G}_F^p(\Delta) \quad \longmapsto \quad \Theta \in \widetilde{\mathfrak{R}}(\Gamma) \quad \longmapsto \quad \Upsilon \in \mathfrak{L}(\Theta)\,.$$

Here the first step represents explicitly resolving all maximally expanded entries of G (encoded by resolvent edges of Γ) in $\mathbf{a}$-admissible entries of H. The second step represents taking partial expectation in all these h-variables.

At this point, we introduce a further, and final, simplification that allows us to postpone some needless technicalities to Section 9.

(S4) The sets $\mathbf{x} = (x_i)_{i \in V_f(\Upsilon)}$ and $\boldsymbol{\mu} = (\mu_i)_{i \in V_e(\Upsilon)}$ are disjoint, and all indices of $\mathbf{x}$ are distinct.

Thus, Simplification **(S4)** is in the same spirit as Simplification **(S1)**. We assume Simplification **(S4)** throughout Section 8.4.

Let $\Gamma \in \mathfrak{G}_F^p(\Delta)$, $\Theta \in \widetilde{\mathfrak{R}}(\Gamma)$, and $\Upsilon \in \mathfrak{L}(\Theta)$. Choose and fix a marked vertex $i \in V_m(\Gamma)$ (see Definition 8.1). In order to gain an extra factor Φ from i, we consider three cases, (a), (b), and (c).

First we explain them informally. Case (a) is the typical situation. Since i is marked, it is a charged vertex in Δ, i.e. after having been minimally linked to, the number of solid and dashed edges adjacent to it are different. In case (a) we show that this property is inherited by at least one of the fresh summation vertices whose parent is i. This fact is fairly clear since neither vertex generation nor lumping alters the number of solid or dashed edges. This will imply that one fresh summation vertex generated by i is a chain vertex in Υ; this in turn will give the extra factor Φ associated with $i \in V_m(\Gamma)$. The other two cases represent exceptional cases. Case (b) deals with a higher-order lumping (i.e. a lumping which has a lump of size greater than two). As indicated above, this results in a combinatorial gain expressed in terms of powers of M^{-1}. Such a factor is depicted graphically using a crossed wiggly line. Finally, case (c) deals with the consequence of the factor h_{uv} on the last line of (8.4), i.e. when a dotted edge joins two original summation vertices. We note that choosing the term h_{uv} is the only way to change the number of solid or dashed edges (off-diagonal entries of G) in the process of vertex resolution. Since dotted edges must be paired, we find that at least two parallel solid or dashed edges must be replaced with a dotted edge; this corresponds to replacing two off-diagonal factors G_{uv} with $|h_{uv}|^2$. After expectation, this means trading in a factor Ψ^2 for M^{-1}. Since $\Psi\Phi \geqslant M^{-1/2}$, the factor M^{-1} may be estimated by $\Psi^2\Phi^2$. Out of this, Ψ^2 is used to compensate for the loss of Ψ^2 mentioned above (losing two off-diagonal entries of G), and the remaining Φ^2 provides us with the

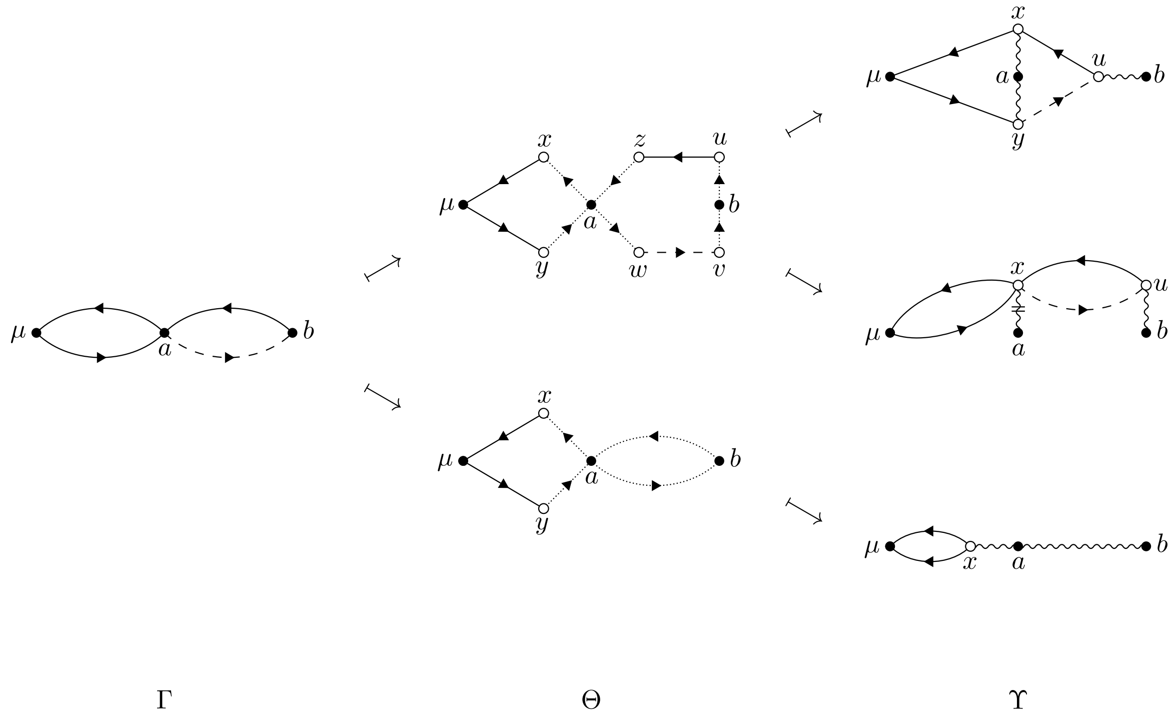

FIGURE 8.5. A simple example of vertex resolution giving rise to the three cases (a), (b), and (c) when resolving the vertex associated with a. From top to bottom in the right-hand column: cases (a), (b), and (c). In the top graph, corresponding to case (a), x is a chain vertex (within the chain $u \to x \to \mu$) which is a remnant of the path $b \to a \to \mu$ consisting of solid edges in the graph Γ. In the middle graph, corresponding to case (b), the crossed wiggly line expresses that all four fresh summation indices arising from the resolution of a coincide, and the four associated dotted edges of Θ were collapsed into one of Υ. Finally, in the bottom figure, corresponding to case (c), the solid and dashed edges between the vertices associated with a and b in Γ give rise to a single wiggly line according to $s_{ab} = \mathbb{E}|h_{ab}|^2$.

gains of Φ associated with each of the two vertices incident to the dotted edge. See Figure 8.5 for a graphical depiction of the cases (a), (b), and (c). Below we formalize these ideas.

Recall the definitions of ν_i and ν_i^* for an arbitrary edge-coloured, directed multigraph (Definition 4.6). Informally, ν_i^ξ gives the number of legs of colour ξ incident to i. The following definition gives a natural extension of chain weights to graphs with wiggly edges.

DEFINITION 8.6. A vertex $i \in V_f(\Upsilon)$ is a *chain vertex* if it has degree three, such that: (i) i is incident to exactly one wiggly edge, and (ii) i is incident to exactly two resolvent edges, which are of the same colour.

Thus a chain vertex $i \in V_f(\Upsilon)$ corresponds to a chain vertex in the sense of Definition 4.6 (i.e. $\nu_i(\Upsilon) + \nu_i^*(\Upsilon) = 2$ and one of the terms vanishes), with the added condition that the summation over the index of i is done with respect to a chain weight (encoded in $\mathcal{A}(\Upsilon)$ by the wiggly edge of Υ incident to i).

Now we present the details of the cases (a), (b), and (c). Let $\Gamma \in \mathfrak{G}_F^p(\Delta)$, $\Theta \in \widetilde{\mathfrak{R}}(\Gamma)$, and $\Upsilon \in \mathcal{L}(\Theta)$. Let

$i \in V_m(\Gamma)$ be marked.

(a) Suppose that $\sigma(i,j) = 0$ for all $j \in V_s(\Upsilon)$ (i.e. there are no wiggly edges in Υ that join two original summation vertices). Suppose moreover that the lumping of the fresh summation vertices of $p^{-1}(i)$ that generates Υ is a pairing (hence $\deg_\Gamma(i)$ must be even).

Then $\nu_j(\Upsilon) + \nu_j^*(\Upsilon) = 2$ for any $j \in p^{-1}(i)$. More precisely, each $j \in p^{-1}(i)$ has degree three in Υ; two of the incident edges are resolvent edges, and the third edge is a wiggly uncrossed edge connecting j with i. Moreover,

$$\nu_i^\xi(\Gamma) \;=\; \sum_{j \in p^{-1}(i)} \nu_j^\xi(\Upsilon) \tag{8.10}$$

for $\xi = 1, *$, since, under the assumption that there are no wiggly edges in Υ that join two original summation vertices, the total number of resolvent edges incident to i does not change by vertex resolution and taking expectation. Hence we find from (8.2) and from $\nu_j(\Upsilon) + \nu_j^*(\Upsilon) = 2$ that there must exist a $j \in p^{-1}(i)$ such that either $\nu_j(\Upsilon) = 0$ or $\nu_j^*(\Upsilon) = 0$; this implies that j is a chain vertex of Υ. We conclude:

At least one $j \in p^{-1}(i)$ is a chain vertex of Υ.

(b) Suppose that $\sigma(i,j) = 0$ for all $j \in V_s(\Upsilon)$, and the lumping of the fresh summation vertices of $p^{-1}(i)$ that generates Υ is a not pairing.

In this case one fresh summation vertex $j \in p^{-1}(i)$ was obtained by lumping together three or more fresh summation vertices of Θ. We conclude:

At least one $j \in p^{-1}(i)$ is connected in Υ to i by a crossed wiggly edge.

(c) Suppose that $\sigma(i,j) > 0$ for some $j \in V_s(\Gamma)$. By property (v) of Section 8.3, $\sigma(i,j) \geqslant 2$. By construction of Υ, if $\sigma(i,j) \geqslant 2$ then i is joined to j in Υ by a wiggly edge that is crossed by $\sigma(i,j) - 2$ strokes. We conclude:

The vertex i is connected in Υ to some $j \in V_s(\Upsilon)$ by a wiggly edge.

Partition $V_m(\Gamma) = V_m^{(a)} \sqcup V_m^{(b)} \sqcup V_m^{(c)}$ into three subsets according to the three case (a), (b), and (c). Now we may estimate the contribution of Υ. For the whole estimate we freeze the original summation vertices $\mathbf{a}$. We shall use Proposition 5.3 on the $|V_m^{(a)}|$ chain vertices of Υ. In order to do so, we still have to get rid of the upper indices $(\mathbf{a})$ from each resolvent entry. This is done exactly as in the end of Section 7.1, using the identity (7.3). We omit further details. Thus Proposition 5.3 is applicable to each chain vertex of Υ. The remainder of the proof is a simple counting of different types of vertices.

Let

$$\sigma \;:=\; \frac{1}{2} \sum_{i,j \in V_s(\Theta)} \sigma(i,j)$$

denote the number of dotted edges of Θ that join two vertices in $V_s(\Theta)$. From (8.9) we find that Υ has $|E(\Gamma)| - \sigma$ resolvent edges. Moreover, Υ has $|V_m^{(a)}|$ chain vertices. The gain from cases (b) and (c) is as follows. From the vertices of type (b) we gain $M^{-|V_m^{(b)}|/2}$. (Each such vertex is incident to a crossed wiggly line, and dropping the strokes crossing such a lines yields a factor $M^{-1/2}$.) From the vertices of type (c) we gain $M^{-\sigma/2}$. (Each dotted edge, encoding $h_{uv} = (s_{uv})^{1/2}\zeta_{uv}$, yields a contribution of size $(s_{uv})^{1/2} \leqslant CM^{-1/2}$ after taking the expectation in the h-variables.)

Now we sum over $\mathbf{x}$ (while still keeping $\mathbf{a}$ frozen). Invoking Proposition 5.3 to estimate the chain vertices of Υ, we therefore find that the contribution of Υ is bounded by

$$\mathcal{X} \;:=\; \Psi^{|E(\Gamma)|-\sigma}\, \Phi^{|V_m^{(a)}|}\, M^{-|V_m^{(b)}|/2}\, M^{-\sigma/2} \;\leqslant\; \Psi^{|E(\Gamma)|-\sigma}\, \Phi^{|V_m^{(a)}|+|V_m^{(b)}|}\, M^{-\sigma/2}\,,$$

where we used $M^{-1/2} \leqslant \Phi$. Since $\sigma(i,j) \geqslant 2$ in case (c), it is easy to see that $|V_m^{(c)}| \leqslant \sigma$. Thus we have $|V_m^{(a)}| + |V_m^{(b)}| + \sigma \geqslant |V_m(\Gamma)|$. This yields the bound

$$\mathcal{X} \;\leqslant\; \Psi^{|E(\Gamma)|}\Phi^{|V_m^{(a)}|+|V_m^{(b)}|+\sigma}\left(\Psi^{-1}\Phi^{-1}M^{-1/2}\right)^{\sigma} \;\leqslant\; \Psi^{|E(\Gamma)|}\Phi^{|V_m(\Gamma)|}\,,$$

where we also used that $\Psi\Phi \geqslant M^{-1/2}$. Recalling Lemma 8.2, we find

$$\mathcal{X} \;\leqslant\; \Psi^{p(\deg(\Delta)+|F|)+|V_c(\Gamma)|-|V_m(\Gamma)|}\, \Phi^{|V_m(\Gamma)|} \;\leqslant\; \Psi^{p(\deg(\Delta)+|F|)}\, \Phi^{p|V_c(\Delta)|}\,, \tag{8.11}$$

where we used $|V_m(\Gamma)| \leqslant |V_c(\Gamma)| = p|V_c(\Delta)|$ and $\Psi \leqslant \Phi$. This estimate was obtained for the sum over $\mathbf{x}$ with fixed $\mathbf{a}$. Finally, we sum over $\mathbf{a}$ trivially, using the fact the $\mathbf{a}$-summation is performed with respect to a weight. This concludes the proof of Theorem 4.8 under Simplifications **(S1)** – **(S4)**.

9. Removing Simplifications (S1) – (S4)

In this section we go back to the proof of Theorem 4.8 of Section 8, and give the additional arguments required to remove the Simplifications **(S1)** – **(S4)** which were assumed there. For ease of reference, we recall them here.

(S1) All summation indices in the expanded summation $\mathbb{E}|X_F^w(\Delta)|^p$ (see (6.5) below) are distinct. (I.e. we ignore repeated indices which give rise to a smaller combinatorics of the summation.)

(S2) There are no diagonal entries $\mathcal{G}_{aa} = G_{aa} - m$ in $\mathcal{Z}(\Delta)$. (I.e. Δ has no loops.)

(S3) We replace any diagonal term $G_{aa}^{(T)}$ with m and any diagonal term $G_{aa}^{(T)*}$ with $\bar{m}$. (Recall that $G_{aa}^{(T)} \approx m$ in the sense that $G_{aa}^{(T)} - m \prec \Psi$ by definition of Λ.) This replacement is done in two places: in $\mathcal{A}(\Gamma)$ and in the identities (3.14a) and (3.14b) which underlie the algebra of vertex resolution.

(S4) The families $\mathbf{x} = (x_i)_{i \in V_f(\Upsilon)}$ and $\boldsymbol{\mu} = (\mu_i)_{i \in V_e(\Upsilon)}$ are disjoint, and all indices of $\mathbf{x}$ are distinct.

9.1. Removing Simplification (S3). We start by removing Simplification **(S3)**, while still assuming Simplifications **(S1)**, **(S2)**, and **(S4)**. In order to remove Simplification **(S3)**, we have deal with the error terms made in the replacement $G_{aa}^{(T)} \mapsto m$. The key formula for dealing with the diagonal terms is (3.14c) with the error terms h_{aa}, $Z_a^{(T)}$, and $U_a^{(Ta)}$ (see (3.15)). Recall that by (2.8) we have $h_{aa} \prec M^{-1/2}$. The other error terms are estimated in the following lemma.

Lemma 9.1. *Suppose that $\Lambda \prec \Psi$ for some admissible control parameter Ψ. Fix $\ell \in \mathbb{N}$. Then we have*

$$Z_a^{(T)} \;\prec\; \Psi\,, \qquad U_a^{(S)} \;\prec\; \min\{\varrho\Psi^2, \Psi\} \;\leqslant\; \Psi\Phi\,, \tag{9.1}$$

for $|T|, |S| \leqslant \ell$ and $a \in \{1, \dots, N\} \setminus T$. Moreover, $Z_a^{(T)}$ is independent of T and $U_a^{(S)}$ is independent of S.

Proof. See Appendix A. □

Using Lemma 9.1 and (3.14c) we write, for any fixed $K \in \mathbb{N}$,

$$\frac{1}{G_{aa}^{(\mathbf{a}\setminus\{a\})}} \;=\; \frac{1}{m} + h_{aa} - Z_a^{(\mathbf{a}\setminus\{a\})} - U_a^{(\mathbf{a})}\,, \qquad G_{aa}^{(\mathbf{a}\setminus\{a\})} \;=\; \sum_{k=0}^{K-1} m^{k+1}\big(-h_{aa} + Z_a^{(\mathbf{a}\setminus\{a\})} + U_a^{(\mathbf{a})}\big)^k + O_\prec(\Psi^K)\,. \tag{9.2}$$

(The error term $O_\prec(\Psi^K)$ is uniform in the same sense as the estimates of (9.1).)

In this section we revisit the argument from Section 8, and explain the differences resulting from the added diagonal terms.

9.1.1. Generation of the fresh summation vertices (revisited), Step I: from Γ *to* Π. As in Section 8.2, we start with $\Gamma \in \mathfrak{G}_F^p(\Delta)$. The goal in Section 8.2 was to decompose Γ into a finite union of graphs (called $\widetilde{\mathfrak{R}}(\Gamma)$) whereby a resolvent edge of $\Theta \in \widetilde{\mathfrak{R}}(\Gamma)$ encoded in $\mathcal{A}(\Theta)$ an entry of $G^{(\mathbf{a})}$ or $G^{(\mathbf{a})*}$, and a dotted edge an $\mathbf{a}$-admissible entry of H. Thus $\mathcal{A}(\Theta)$ was well-suited for taking the partial expectation in $\mathbf{a}$. In this section we keep track of the diagonal entries (represented graphically by loops) that arise both in Family A and B identities and were previously freely replaced by powers of m and $\bar{m}$, according to Simplification **(S3)**. The main difficulty here is that resolving diagonal entries requires the more complicated formulas (9.2) instead of (8.4) (which immediately yielded resolvents with upper indices $\mathbf{a}$).

The ultimate goal of this section and of Section 9.1.2 is the same as that of Section 8.2: to obtain graphs Θ whose resolvent edges encode resolvent entries with upper indices $\mathbf{a}$ (up to a negligible error term that can be estimated brutally). We shall reach this in two steps. In the first step, which is the content of this section (Section 9.1.1), we express $\mathcal{A}(\Gamma)$ as a sum of monomials whose off-diagonal resolvent entries have upper indices $\mathbf{a}$ and whose diagonal resolvent entries are maximally expanded. We denote by $\mathfrak{D}(\Gamma)$ the family of graphs encoding these new monomials, and we shall use the letter Π for a generic element of $\mathfrak{D}(\Gamma)$. Graphically, therefore, this step corresponds to the mapping $\Gamma \mapsto \{\Pi_\alpha\} = \mathfrak{D}(\Gamma)$. Sometimes we shall refer to it informally as $\Gamma \mapsto \Pi$.

In the second step, which is the content of Section 9.1.2, we use (9.2) to replace all maximally expanded diagonal resolvent entries (in $\mathcal{A}(\Pi)$ for any $\Pi \in \mathfrak{D}(\Gamma)$) with resolvent entries having upper indices $\mathbf{a}$ (again up to a negligible error term that can be estimated brutally). We generically call the resulting graphs Θ, and let $\mathfrak{R}(\Pi) = \{\Theta_\alpha\}$ be the collection of such graphs obtained from a fixed $\Pi \in \mathfrak{D}(\Gamma)$. Graphically, this step corresponds to the mapping $\Pi \mapsto \{\Theta_\alpha\} = \mathfrak{R}(\Pi)$. The graphs Θ play the same role as the graphs Θ in Section 8. Indeed, each resolvent edge of Θ encodes in $\mathcal{A}(\Theta)$ a resolvent entry that has upper indices $\mathbf{a}$, and each dotted edge an $\mathbf{a}$-admissible entry of H. Hence $\mathcal{A}(\Theta)$ is amenable to taking the partial expectation in $\mathbf{a}$. (This will be done in Section 9.1.3.)

Lemma 9.2. *For any $K \in \mathbb{N}$ we have the decomposition*

$$\mathcal{A}(\Gamma) \;=\; \sum_\alpha A_\alpha + O_\prec(\Psi^K)\,,$$

where the summation is over a finite N-independent set, and A_α is a monomial in the entries of $G^{(T)}$, the entries of $G^{(T)*}$, and the $\mathbf{a}$-admissible entries of H. Moreover, each entry $G_{uv}^{(T)}$ of A_α satisfies the condition

$(*)$ $G_{uv}^{(T)}$ either has upper indices $T = \mathbf{a}$ or is a maximally expanded diagonal entry ($u = v$ and $T = \mathbf{a}\setminus\{u\}$).

The same condition also applies to each entry $G_{uv}^{(T)*}$.

We shall apply this lemma below with the choice $K := p(\deg(\Delta) + 2|V_s(\Delta)|)$, which will ensure that the error term $O_\prec(\Psi^K)$ is negligible.

Proof of Lemma 9.2. We apply (8.4) to each off-diagonal (maximally expanded) resolvent entry of $\mathcal{A}(\Gamma)$. After an application of (8.4), the resulting expression does not in general satisfy $(*)$, due to the factor $G_{uu}^{(\mathbf{a}\setminus\{u,v\})}$ on the last line of (8.4), which is not maximally expanded. (Note that all other factors on the right-hand side of (8.4) satisfy $(*)$.) As always, we use (3.13) to make $G_{uu}^{(\mathbf{a}\setminus\{u,v\})}$ maximally expanded:

$$G_{uu}^{(\mathbf{a}\setminus\{u,v\})} = G_{uu}^{(\mathbf{a}\setminus\{u\})} + \frac{G_{uv}^{(\mathbf{a}\setminus\{u,v\})}G_{vu}^{(\mathbf{a}\setminus\{u,v\})}}{G_{vv}^{(\mathbf{a}\setminus\{u,v\})}}\,, \qquad \frac{1}{G_{uu}^{(\mathbf{a}\setminus\{u,v\})}} = \frac{1}{G_{uu}^{(\mathbf{a}\setminus\{u\})}} - \frac{G_{uv}^{(\mathbf{a}\setminus\{u,v\})}G_{vu}^{(\mathbf{a}\setminus\{u,v\})}}{G_{uu}^{(\mathbf{a}\setminus\{u,v\})}G_{uu}^{(\mathbf{a}\setminus\{u\})}G_{vv}^{(\mathbf{a}\setminus\{u,v\})}}\,. \tag{9.3}$$

Here we use the first identity of (9.3). The first term is maximally expanded and good as it is. The second consists of two maximally expanded off-diagonal terms in the numerator and one diagonal term in the denominator which is not maximally expanded. We now apply (8.4) to each of the terms in the numerator. The result is an expression with entries of G that either have upper indices $\mathbf{a}$ or are diagonal. The diagonal entries are not maximally expanded, and hence we must apply (3.13) to each of them. Moreover, the diagonal entry in the denominator is not maximally expanded, and must be further expanded using the second identity of (9.3). We continue in this manner, successively using (8.4) on maximally expanded off-diagonal entries and (3.13) on diagonal entries that are not maximally expanded. This procedure is reminiscent of the one introduced after Definition 6.4. As in Section 6, although this procedure in general does not terminate, it does increase the number of off-diagonal terms, which allows us to stop brutally once a sufficient number of off-diagonal terms have been generated. Note that, unlike the one-step iteration of Section 6 (which only used (3.13)), we now have a two-step iteration, which repeatedly uses (3.13) and (8.4) in tandem.

More formally, the algorithm may be described as follows. In order to define the brutal stopping rule precisely, we set

$$\ell(A) := (\text{number of entries of } G^{(\mathbf{a})} \text{ in } A) + \sum_{a,b \in V_s(\Gamma)} (\text{number of entries } h_{ab} \text{ in } A)\,,$$

where A is a monomial in the entries of $G^{(T)}$ and H. Now set $A := \mathcal{A}(\Gamma)$; A will denote the running monomial in the algorithm, and $\mathcal{A}(\Gamma)$ is its initial value.

Step 1. Pick an off-diagonal term $G_{uv}^{(\mathbf{a}\setminus\{u,v\})}$ in A which does not have upper indices $\mathbf{a}$. If no such term exists, go to Step 2. Otherwise apply (8.4) to $G_{uv}^{(\mathbf{a}\setminus\{u,v\})}$. This yields the splitting $A = A' + A''$ (where A' is the main term that contains a factor $G^{(\mathbf{a})}$ and $A'' = 0$ unless both u and v are summation vertices, in which case A'' contains the special factor h_{uv} from the third line of (8.4)). Repeat step 1 for A' and A'' (provided $A'' \neq 0$). (Notice that at each repetition of Step 1 the number of off-diagonal terms with no upper index $\mathbf{a}$ decreases by one, so after finitely many steps the algorithm exits to Step 2.)

Step 2. If $\ell(A) \geqslant K$, stop. Otherwise go to Step 3.

Step 3. Pick a diagonal term $G_{uu}^{(\mathbf{a}\setminus\{u,v\})}$ in A that is not maximally expanded. If no such term exists, stop. Otherwise apply (9.3) to $G_{uu}^{(\mathbf{a}\setminus\{u,v\})}$. This induces a splitting $A = A' + A''$ according to the two summands in either identity of (9.3). Repeat Step 1 for both A' and A''.

Since Step 1 increases ℓ by exactly one, it follows by Step 2 that the algorithm must terminate after a finite, N-independent, number of steps. The result is a finite sum of terms $\{A_\alpha\}$ whose number does not depend on N. Pick one such $A_\alpha \equiv A$. We consider two cases depending on whether the algorithm, in generating A, stopped at Step 2 or Step 3. In other words, we differentiate based on whether it is stopped because there are sufficiently many small factors (stopping at Step 2) or because each resolvent entry satisfies $(*)$ (stopping at Step 3).

Consider first the case where the algorithm stopped at Step 2. This corresponds to a brutal stopping, where we may estimate A by a simple power counting using the lower bound on $\ell(A)$. We claim that we have the trivial bound

$$A \;=\; O_\prec(\Psi^K). \tag{9.4}$$

In order to see this, we note that (8.4) (reading these formulas from right to left) and (2.10) imply[4]

$$\sum_x^{(\mathbf{a})} h_{ux} G^{(\mathbf{a})}_{xv} \;=\; O_\prec(\Psi), \qquad \sum_x^{(\mathbf{a})} G^{(\mathbf{a})}_{ux} h_{xv} \;=\; O_\prec(\Psi)\,, \qquad \sum_{x,y}^{(\mathbf{a})} h_{ux} G^{(\mathbf{a})}_{xy} h_{xv} \;=\; O_\prec(\Psi)\,. \tag{9.5}$$

Moreover, by definition of Step 1 and the explicit expressions in (8.4) each entry of $G^{(\mathbf{a})}$ in A comes in one of the three forms in (9.5). Hence the lower bound $\ell(A) \geqslant K$, the definition of ℓ, and (2.10) yield (9.4).

Apart from this error term, all other terms resulted in stopping the algorithm at Step 3. In this case it is immediate that each entry $G^{(T)}_{uv}$ of A satisfies $(*)$. This proves Lemma 9.2. □

The algorithm from the proof of Lemma 9.2 (Steps 1 – 3) has a trivial reformulation on the level of graphs. This also yields a convenient graphical representation of the monomials $\{A_\alpha\}$. For future use, we give more details on the graphical version of Step 3. Let $\mathfrak{D}(\Gamma)$ denote the set of graphs that encode the monomials $\{A_\alpha\}$ obtained through Steps 1–3 starting from Γ and stopping at Step 3. The elements of $\mathfrak{D}(\Gamma)$ will generically be denoted by Π.

We use the graphical notations from Figures 3.2 and 3.8. In Figure 9.1 we summarize the rules (8.4) graphically.

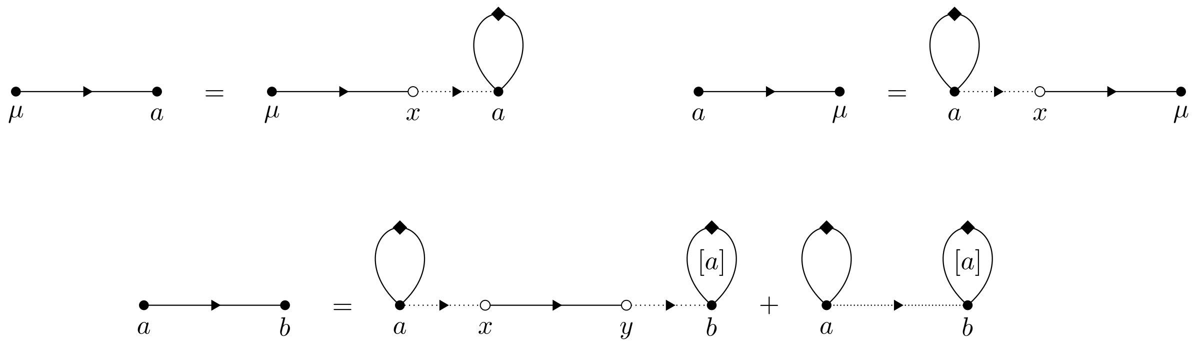

FIGURE 9.1. The graphical representation of the rules (8.4) (using a and b for summation indices and μ for external index instead of the generic indices u and v). Two of the loops on the last line correspond to resolvent entries $G^{(\mathbf{a}\setminus\{a,b\})}_{bb}$ which are not maximally expanded: they still depend on a, which is indicated by the label $[a]$ inside the loop.

[4] Note that these estimates also follow directly from basic large deviation results such as Lemmas B.1 and B.2 in [15].

The term A'' from Step 3 arises from taking the second term on the right-hand sides of (9.3), which in the graphical language translates to creating two (non-loop) resolvent edges connecting the vertices i and j associated with the indices u and v respectively, i.e. linking a loop at i with j. We call this process *linking i with j*. See Figure 9.2.

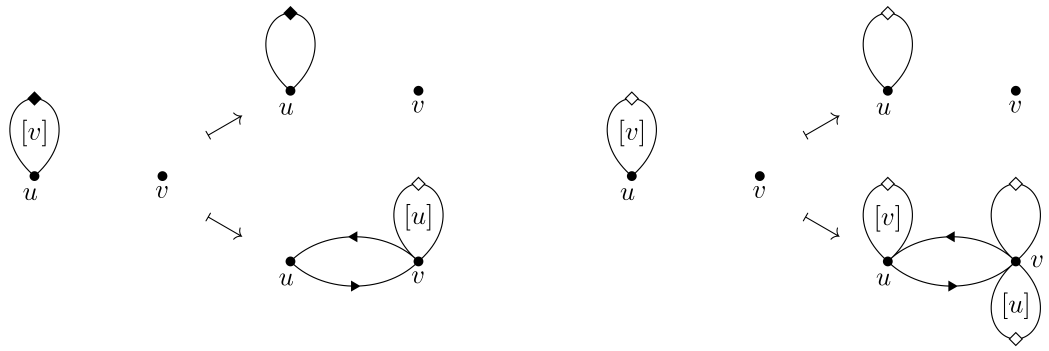

FIGURE 9.2. Linking the vertex i (depicted in the picture with its associated index u) with the vertex j (depicted with index v). These two diagrams correspond to the two identities in (9.3). As in Figure 9.1, if a diagonal resolvent entry encoded by a loop is not maximally expanded, we indicate the index on which it depends in angular brackets inside the loop.

This graphical algorithm provides an alternative, graphical, construction of $\mathfrak{D}(\Gamma)$ starting from Γ. Each $\Pi \in \mathfrak{D}(\Gamma)$ encodes a monomial $\mathcal{A}(\Pi)$ whose resolvent entries satisfy $(*)$. As in Section 8, the vertex set of Π may be written as $V(\Pi) = V_f(\Pi) \sqcup V_s(\Pi) \sqcup V_e(\Pi)$, corresponding to the fresh summation vertices, the original summation vertices, and the external vertices, respectively. Note that Π now contains loops, which bear either a black or white diamond (encoding diagonal entries of G in the numerator or denominator respectively). Moreover, each diagonal entry encoded by a loop of Π is maximally expanded, and each off-diagonal entry encoded by a non-loop edge of Π has upper indices $\mathbf{a}$. See Figure 9.3 for a simple example of the process $\Gamma \mapsto \Pi \in \mathfrak{D}(\Gamma)$.

9.1.2. Generation of the fresh summation vertices (revisited), Step II: from Π *to* Θ. In this section we complete the second part of the generation of the fresh summation vertices, by constructing a family of graphs $\Theta \in \mathfrak{R}(\Pi)$ from Π. The underlying algebraic identity is (9.2).

LEMMA 9.3. *For any $K \in \mathbb{N}$ and any $\Pi \in \mathfrak{D}(\Gamma)$ there is a decomposition*

$$\mathcal{A}(\Pi) \;=\; \sum_{\alpha} B_\alpha + O_\prec(\Psi^K)\,, \tag{9.6}$$

such that each B_α is a monomial in the entries of $G^{(\mathbf{a})}$ and the $\mathbf{a}$-admissible entries of H. The sum over α ranges over a finite set that is independent of N.

We shall apply Lemma 9.3 with the choice $K := p(\deg(\Delta) + 2|V_s(\Delta)|)$, which will ensure that the error term $O_\prec(\Psi^K)$ is negligible.

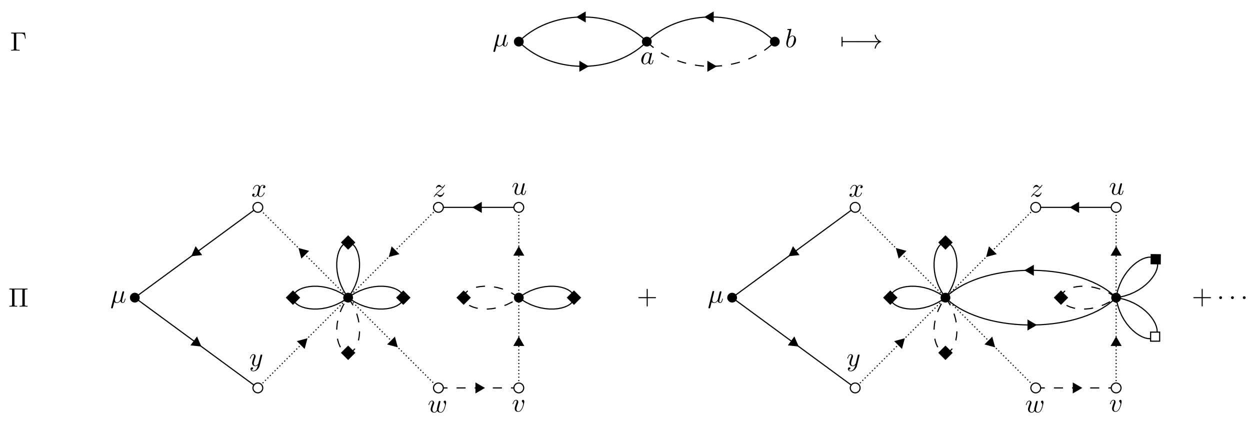

FIGURE 9.3. The process $\Gamma \mapsto \Pi \in \mathfrak{D}(\Gamma)$, where we draw the two simplest elements of $\mathfrak{D}(\Gamma)$ on the bottom line. (For reasons of space, we omit the labels a and b on the bottom line.) The first graph is just the top-middle graph of Figure 8.5 but with added loops. In the second graph one loop at a was linked with b.

PROOF OF LEMMA 9.3. We simply apply (9.2) to each diagonal resolvent entry of $\mathcal{A}(\Pi)$. Recall that each diagonal resolvent entry of $\mathcal{A}(\Pi)$ is maximally expanded, which implies that all resolvent entries explicitly appearing in the definition (3.15) for $Z_a^{(\mathbf{a}\setminus\{a\})}$ and $U_a^{(\mathbf{a})}$ have upper indices $\mathbf{a}$. Then, as above, it immediately follows that if we pick the rest term $O_\prec(\Psi^K)$ in (9.2) from any diagonal entry, the resulting monomial is $O_\prec(K)$ and may be absorbed into the error term on the right-hand side of (9.6).

For the following we therefore assume that there are no rest terms $O_\prec(\Psi^K)$ in the expansion (9.2) of the diagonal resolvent entries of $\mathcal{A}(\Pi)$. The result is a finite family of monomials whose number does not depend on N (but does of course depend on K), and which may again be represented graphically. In such graphs we represent a term $U_a^{(\mathbf{a})}$ with a solid ring around the vertex associated with a (these terms $U_a^{(\mathbf{a})}$ will not be expanded further, so their precise structure does not matter; the number of rings simply encode their size). See Figure 9.4 for a depiction of the three nontrivial terms arising from the expansion of $G_{aa}^{(\mathbf{a}\setminus\{a\})}$. Thus,

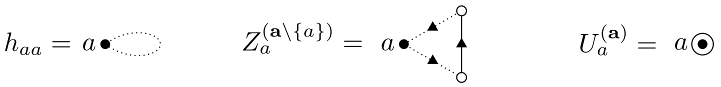

FIGURE 9.4. The graphical representation of the three error terms resulting from the expansion of $G_{aa}^{(\mathbf{a}\setminus\{a\})}$ using (9.2). Apart from the entries of H encoded by the dotted edges, all terms are independent of $\mathbf{a}$.

when expanding a loop at a with a white diamond (encoding $1/G_{aa}^{(\mathbf{a}\setminus\{a\})}$), we replace the loop with either nothing (corresponding to the term $1/m$ used in the argument of Section 8) or one of the three pieces in Figure 9.4. Similarly, when expanding a loop at a with a black diamond (encoding $G_{aa}^{(\mathbf{a}\setminus\{a\})}$), we replace the loop with either nothing (corresponding to a factor m coming from the zeroth order term in the summation in (9.2)) or an agglomeration of pieces from Figure 9.4 at the vertex associated with a. (We use concentric rings around a to depict several factors $U_a^{(\mathbf{a})}$). See Figure 9.5.

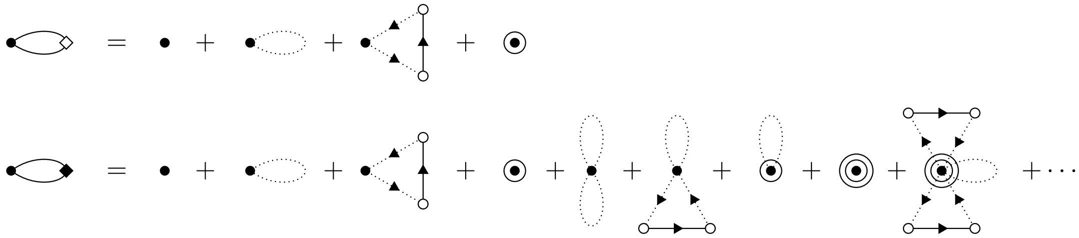

FIGURE 9.5. Expanding a loop. The two lines correspond to the two identities of (9.2) respectively.

The application of (9.2) to the diagonal entries encoded by Π yields a new family of graphs, which we call $\mathfrak{R}(\Pi)$ and whose elements we denote by Θ. Each resolvent edge of $\Theta \in \mathfrak{R}(\Pi)$ now encodes an entry of $G^{(\mathbf{a})}$ or $G^{(\mathbf{a})*}$. We also have the usual self-explanatory splitting $V(\Theta) = V_f(\Theta) \sqcup V_s(\Theta) \sqcup V_e(\Theta)$. See Figure 9.6 for an example of the process $\Pi \mapsto \mathfrak{R}(\Pi)$. □

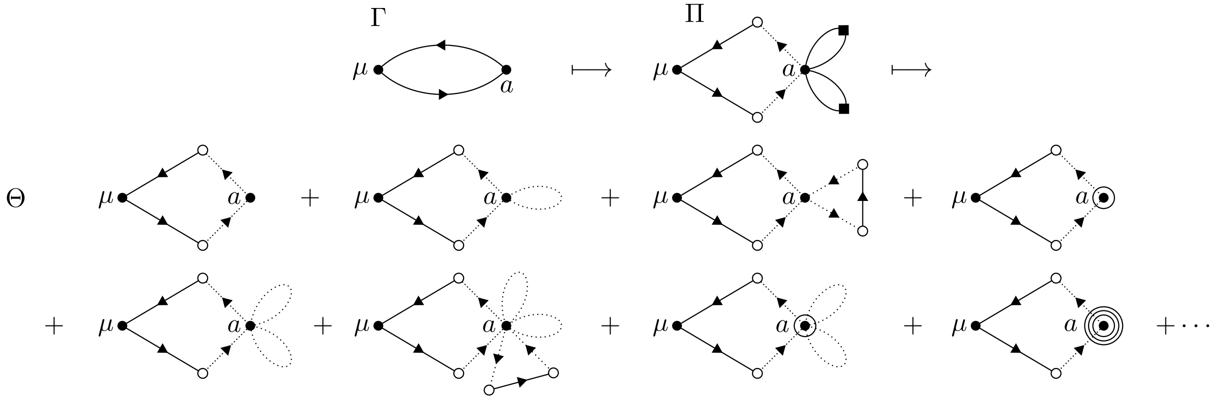

FIGURE 9.6. The process $\Gamma \mapsto \Pi \mapsto \Theta$. On the second line we draw four graphs Θ corresponding to choosing one of the four terms on the right-hand side of the first equation of (9.2). On the third line we draw some more complicated graphs Θ.

Summarizing the results of this Sections 9.1.1 and 9.1.2, for a given $\Gamma \in \mathfrak{G}_F^p(\Delta)$, we have constructed an N-independent set of graphs,

$$\mathfrak{R}(\mathfrak{D}(\Gamma)) \;=\; \bigcup \bigl\{ \mathfrak{R}(\Pi) : \Pi \in \mathfrak{D}(\Gamma) \bigr\} \,.$$

If $\Theta \in \mathfrak{R}(\mathfrak{D}(\Gamma))$ then each resolvent entry of $\mathcal{A}(\Theta)$ has upper indices $\mathbf{a}$, and the fresh summation indices $\mathbf{x} = (x_i)_{i \in V_f(\Theta)}$ and the original summation indices $\mathbf{a} = (a_i)_{i \in V_s(\Theta)}$ are disjoint. Moreover, we have the splitting

$$\mathcal{A}_{\mathbf{a}}(\Gamma) \;=\; \sum_{\Theta \in \mathfrak{R}(\mathfrak{D}(\Gamma))} \sum_{\mathbf{x}}^{(\mathbf{a})} \mathcal{A}_{\mathbf{a},\mathbf{x}}(\Theta) + O_{\prec}\big(\Psi^{p(\deg(\Delta)+2|V_s(\Delta)|)}\big)$$

where we explicitly indicated the set of summation indices in the subscript of $\mathcal{A}$, see (8.3). Note that the elements of the family $\mathfrak{R}(\mathfrak{D}(\Gamma))$ have the same properties as the elements of the smaller set $\widetilde{\mathfrak{R}}(\Gamma)$ from Section 8.

9.1.3. Lumping of the fresh summation vertices (revisited) and conclusion of the estimate. Fix a $\Theta \in \mathfrak{R}(\mathfrak{D}(\Gamma))$. Now we may proceed as in Section 8.3 and take the lumping of the entries of H in $\mathcal{A}(\Theta)$ by computing their partial expectation $\prod_{a \in \mathbf{a}} P_a$. Since all resolvent entries of $\mathcal{A}(\Theta)$ are independent of $\mathbf{a}$, this partial expectation acts only on the entries of H, and leads to lumpings exactly as in Section 8.3. This gives rise to a family of graphs $\Upsilon \in \mathfrak{L}(\Theta)$. As before, we seek to gain a factor Φ from each marked vertex $i \in V_m(\Gamma)$.

Thus, let us fix a sequence $\Gamma \mapsto \Pi \mapsto \Theta \mapsto \Upsilon$. It is convenient to extend the definition of the degree of a vertex as follows. By definition, the *degree* of $i \in V(\Theta)$, written $\deg_\Theta(i)$, is equal to the number of legs incident to i plus two times the number of rings around i. This convention is chosen so that each error term in Figure 9.4 increases the degree of i by two.

Now take a marked vertex $i \in V_m(\Gamma)$. Note that, by construction of Π and Θ, we have $\deg_\Theta(i) \geqslant \deg_\Gamma(i)$. We consider two cases.

(i) Suppose that $\deg_\Theta(i) = \deg_\Gamma(i)$. This means that in the process $\Gamma \mapsto \Pi$ the original summation vertex i was not linked with another original summation vertex (see Section 9.1.1), and that in the process $\Pi \mapsto \Theta$ (see Section 9.1.2) we always chose the main term ($1/m$ or m) on the right-hand sides of (9.2) when applying (9.2) to any diagonal entries with lower indices $a_i a_i$. In particular, (8.10) holds. We may therefore proceed exactly as in Section 8: any pairing in $\Theta \mapsto \Upsilon$ of the white vertices adjacent to i gives rise to at least one chain vertex of Υ (see Definition 8.6). A higher-order lumping (i.e. one that is not a pairing) gives rise to a positive power of $M^{-1/2} \leqslant \Psi$. Either way, we shall gain a factor Φ from i after summing over $\mathbf{x}$ and invoking Proposition 5.3.

(ii) Suppose that $\deg_\Theta(i) > \deg_\Gamma(i)$. In this case we have that either

(ii.1) in the process $\Gamma \mapsto \Pi$ the vertex i was linked to another original summation vertex, or

(ii.2) in the process $\Pi \mapsto \Theta$ we chose at least one error term (represented graphically by one of the graphs in Figure 9.4) on the right-hand sides of (9.2) when applying (9.2) to the diagonal entries with lower indices i.

We claim that either case, (ii.1) or (ii.2), results in an extra error factor in Υ of order Ψ.

In order to see this, consider first the case (ii.1). Here the linking means that there is a $j \in V_s(\Upsilon)$ such that in Υ we have two extra resolvent edges (as compared to case (i)), each connecting a vertex in $p^{-1}(i)$ to a vertex in $p^{-1}(j)$. This yields a factor Ψ^2. Thus we gain a factor Ψ that we ascribe to i (the other factor Ψ is in general not available, as it may be needed for exactly the same reason at the vertex j). Next, consider the case (ii.2). If there is a term h_{aa} or $U_a^{(\mathbf{a})}$ in $\mathcal{A}(\Theta)$, then we immediately get a factor $\Psi\Phi$. (For $U_a^{(\mathbf{a})}$ this is trivial by (9.1) and for h_{aa} taking P_a implies that there must be at least another factor h_{aa}, in which case we get a factor $M^{-1} \leqslant \Psi\Phi$.) Finally, if we have a factor

$Z_a^{(\mathbf{a}\setminus\{a\})}$ observe that in the expression

$$Z_a^{(\mathbf{a}\setminus\{a\})} \;=\; Q_a\left(\sum_{x,y}^{(\mathbf{a})} h_{ax}G_{xy}^{(\mathbf{a})}h_{ya}\right) \tag{9.7}$$

we cannot pair h_{ax} with h_{ya} when computing the partial expectation P_a (since $P_aQ_a = 0$). (Of course in a higher-order lumping, they could be in the same lump provided this lump contains at least three elements.) This implies that, in the leading-order pairing, the fresh summation vertices of Θ associated with x and y will be paired into different vertices of Υ. In particular, we gain an additional off-diagonal resolvent entry $G_{xy}^{(\mathbf{a})} \prec \Psi$. See Figure 9.7 for a graphical depiction of this lumping.

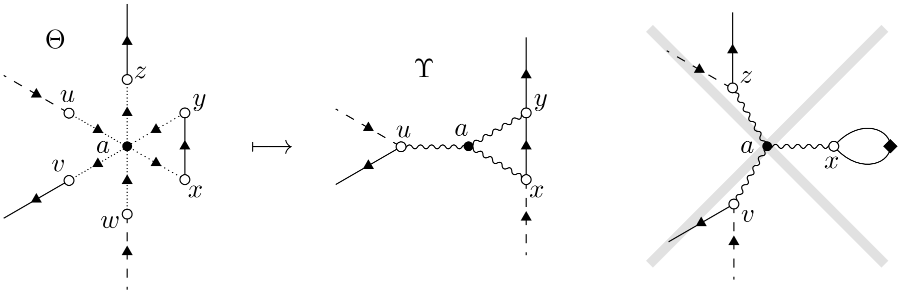

FIGURE 9.7. The lumping of fresh summation vertices in the presence of a factor $Z_a^{(\mathbf{a}\setminus\{a\})}$ (represented by a triangle in Θ.). Due to the Q_a in the definition of $Z_a^{(\mathbf{a}\setminus\{a\})}$, the last graph (crossed out in grey) does not contribute.

In summary, each marked vertex i therefore yields a gain of Φ upon summation over $\mathbf{x}$. This concludes the proof of Theorem 4.8 without Simplification **(S3)**.

9.2. Removing Simplification (S2). In this section we revisit the arguments of Sections 8 and 9.1, and explain the modifications required if we relax Simplification **(S2)**, i.e. allow diagonal entries $\mathcal{G}_{aa} = G_{aa} - m$ in the definition of $\mathcal{Z}$. (On the level of Δ, this amounts to allowing loops.) The construction of $\Gamma \in \mathfrak{G}_F^p(\Delta)$ remains unchanged. Each diagonal entry of $\mathcal{A}(\Gamma)$ is maximally expanded, i.e. of the form $\mathcal{G}_{aa}^{(\mathbf{a}\setminus\{a\})}$. Hence the construction of $\Pi \in \mathfrak{D}(\Gamma)$ from Γ carries over unchanged from Section 9.1.1. Now Π has loops of three kinds: with a black diamond (encoding $G_{aa}^{(\mathbf{a}\setminus\{a\})}$), with a white diamond (encoding $1/G_{aa}^{(\mathbf{a}\setminus\{a\})}$), and plain (encoding $\mathcal{G}_{aa}^{(\mathbf{a}\setminus\{a\})}$). Note that a decorated loop encodes a factor of size $O_\prec(1)$ while a plain loop encodes a factor of size $O_\prec(\Psi)$. The additional difficulty in this section as compared to Section 8 is that the naive size of a plain loop is smaller than the size of the decorated loops dealt with in Section 8. Thus we have to establish bounds which, in addition to the gain extracted in Section 8, also contain the smallness associated with the naive size of a plain loop.

The process $\Pi \mapsto \Theta \in \mathfrak{R}(\Pi)$, in which all (maximally expanded) diagonal entries are expanded using (9.2) is again the same as that of Section 9.1.2. For $1/G_{aa}^{(\mathbf{a}\setminus\{a\})}$ and $G_{aa}^{(\mathbf{a}\setminus\{a\})}$ we use (9.2), and, in addition, for $\mathcal{G}^{(\mathbf{a}\setminus\{a\})}_{aa}$ we use

$$\mathcal{G}^{(\mathbf{a}\setminus\{a\})}_{aa} \;=\; \sum_{k=1}^{K-1} m^{k+1}\big(-h_{aa} + Z^{(\mathbf{a}\setminus\{a\})}_a + U^{(\mathbf{a})}_a\big)^k + O_{\prec}(\Psi^K) \tag{9.8}$$

(note that the sum starts with $k = 1$). Finally, lumping the white summation vertices yields the graph Υ. As in Section 9.1.3, the important observation is that the two white vertices associated with a $Z^{(\mathbf{a}\setminus\{a\})}_a$ (see Figure 9.4) cannot be paired. In summary, the resolution process $\Gamma \mapsto \Pi \mapsto \Theta \mapsto \Upsilon$ is almost identical to that in Sections 8 and 9.1. There is only one new ingredient: the expansion (9.8) which starts with $k = 1$.

To illustrate this procedure, let us consider the simple example with $\mathbf{a} = \{a\}$

$$\begin{aligned}
\mathbb{E} G^*_{\mu a} G_{a\mu} \mathcal{G}_{aa} &= \mathbb{E} \sum_{x,y}^{(a)} G_{aa} G^*_{aa} h_{ax} G^{(a)}_{x\mu} G^{(a)*}_{\mu y} h_{ya} \mathcal{G}_{aa} \\
&= m^2 \bar{m}\, \mathbb{E} \sum_{x,y}^{(a)} h_{ax} G^{(a)}_{x\mu} G^{(a)*}_{\mu y} h_{ya} \bigg(Q_a \sum_{z,w}^{(a)} h_{az} G^{(a)}_{zw} h_{wa} - h_{aa} + U^{(a)}_a \bigg) + \cdots \\
&= m^2 \bar{m}\, \mathbb{E} \sum_{x,y}^{(a)} s_{ax} s_{ay} G^{(a)*}_{\mu y} G^{(a)}_{yx} G^{(a)}_{x\mu} + 0 + m^2 \bar{m}\, \mathbb{E} \sum_{x} s_{ax} G^{(a)*}_{\mu x} G^{(a)}_{x\mu} U^{(a)}_a + \cdots ,
\end{aligned} \tag{9.9}$$

where $+\cdots$ denotes higher-order terms in the expansion (9.2) and (9.8). The expectation of the middle term in the parentheses vanishes because $\mathbb{E} h_{aa} = 0$. See Figure 9.8 for a graphical version of (9.9).

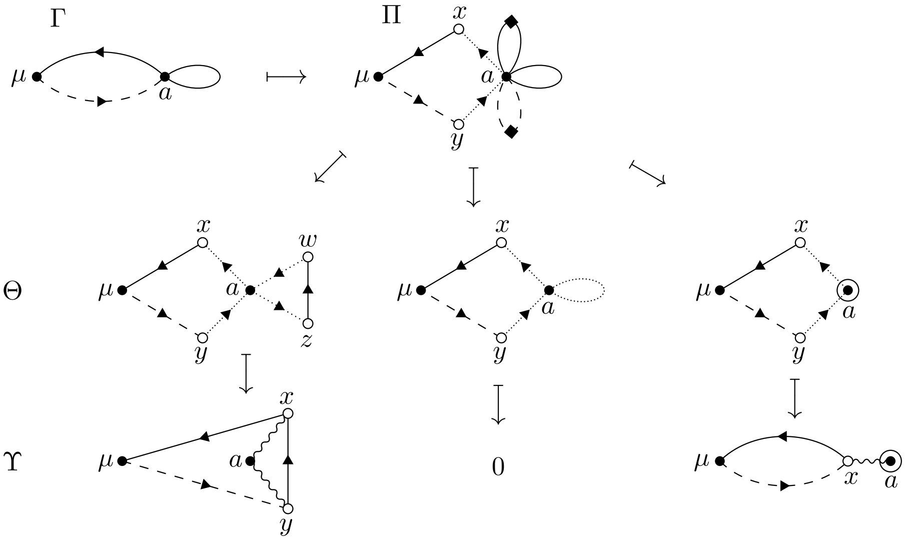

FIGURE 9.8. The complete resolution process $\Gamma \mapsto \Pi \mapsto \Theta \mapsto \Upsilon$ for the example (9.9). At each step we only draw the leading-order graphs. The first, second, and third lines of the figure correspond to the first, second, and third lines of (9.9) respectively.

Note that each unmarked loop (encoding a diagonal entry of $\mathcal{G}$) contributes a factor $O_\prec(\Psi)$ to $\mathcal{A}(\Gamma)$. When performing the vertex resolution $\Gamma \mapsto \Upsilon$, we therefore have to ensure that this gain of Ψ is not lost (i.e. that $\mathcal{A}(\Upsilon)$ has an associated factor of size $O_\prec(\Psi)$). In addition, we have to gain a factor Φ from each marked vertex of Υ.

In the example (9.9), the vertex a is marked and each term on the bottom line of (9.9) is of order $\Psi^3\Phi$. This bound should be read as $\Psi^2\Psi\Phi$, where Ψ^2 is the trivial bound on the off-diagonal entries, Ψ is the bound on the diagonal entry of $\mathcal{G}$, and Φ is the additional gain arising from the fact that a is marked. Indeed, the first term on the bottom line of (9.9) is of order $\Psi^3\Phi$ by Proposition 5.3, and the last term of order Ψ^4 by Lemma 9.1.

This is in fact a general phenomenon. Let $i \in V_m(\Gamma)$ be marked, with associated summation index a. We shall give the details only for a leading-order graph Υ, i.e. a graph Υ that satisfies:

(i) In the process $\Gamma \mapsto \Pi$ the original summation vertex i was not linked with another original summation vertex.

(ii) In the process $\Pi \mapsto \Theta$ we always chose the main term ($1/m$ or m) on the right-hand sides (9.2), and the term $mZ_a^{(\mathbf{a}\setminus\{a\})}$ on the right-hand side of (9.8).

(iii) In the process $\Theta \mapsto \Upsilon$, we chose a pairing of the white vertices incident to i. (I.e. no higher-order lumping is allowed.)

If Υ does not satisfy (i) – (iii), an argument almost identical to that of Sections 8.4 and 9.1.3 yields an extra factor Φ, in addition to Ψ^ℓ where ℓ is the number of plain loops incident to i in Γ. (This is a simple power counting that uses the fact that each $U_a^{(\mathbf{a})}$ yields a factor $\Psi\Phi$, and each h_{aa} and $Z_a^{(\mathbf{a}\setminus\{a\})}$ yield a factor Ψ each. That $Z_a^{(\mathbf{a}\setminus\{a\})}$ yields a factor Ψ follows from the observation that after resolution it yields an off-diagonal resolvent entry in $\mathcal{A}(\Upsilon)$, as explained after (9.7). Note that, unlike in Section 9.1.3 where it was enough to gain a factor Ψ from $U_a^{(\mathbf{a})}$, here it is crucial that $U_a^{(\mathbf{a})} \prec \Psi\Phi$.)

Let us therefore assume that Υ satisfies (i) – (iii). By (i) and (ii) we have (8.10). Recall the definition of the projection p from (iv) in Section 8.3. Each $j \in p^{-1}(i)$ is incident to precisely two resolvent edges and one wiggly edge that is also incident to i. Moreover, no vertex of $p^{-1}(i)$ is incident to a loop; this follows from the above observation that the two white vertices associated with $mZ_a^{(\mathbf{a}\setminus\{a\})}$ cannot be paired. From (8.2) and (8.10), we therefore find that at least one vertex in $p^{-1}(i)$ is a chain vertex. Consequently summation over $\mathbf{x}$ results in an extra factor Φ by Proposition 5.3, and hence completes the argument.

9.3. Removing Simplification (S4). In this section we remove Simplification **(S4)**, by allowing the fresh summation indices $\mathbf{x}$ to coincide with each other and with external indices $\boldsymbol{\mu}$. This entails proving Proposition 5.3 without the simplifying assumption **(S4)** that was assumed in its proof. Roughly, there are two kinds of problems arising from such coincidences: an off-diagonal resolvent entry G_{xy} may become diagonal (hence leading to a loss of a factor Ψ), and a chain vertex may cease to be one (hence leading to a loss of a factor Φ). However, these losses are compensated by powers of M^{-1} resulting from a reduction in the number of independent summation variables. The main point is to prove each coincidence of summation variables results in a loss of at most two factors of Ψ and at most two factors of Φ. Since $M^{-1} \leqslant \Psi^2\Phi^2$, the gain of M^{-1} will be enough to compensate this loss.

Throughout Sections 8, 9.1, and 9.2, we invoked Proposition 5.3 in order to gain from chain vertices. To that end, we had to assume Simplification **(S4)** (since the indices $(\mathbf{x}, \boldsymbol{\mu})$ are assumed to be disjoint in Proposition 5.3). The main result of this section is the following extension of Proposition 5.3. It states that the stochastic bound of Proposition 5.3 is valid even if the summation over $\mathbf{x}$ has no restriction. (As in Proposition 5.3, we use $\mathbf{a}$ to denote the summation indices; in our applications of Proposition 9.4 $\mathbf{a}$ always consists of fresh summation vertices which we denoted by $\mathbf{x}$ in Sections 8, 9.1, and 9.2.)

PROPOSITION 9.4. *Suppose that $\Lambda \prec \Psi$ for some admissible control parameter Ψ. Let Δ be a chain encoding $\mathcal{Z}_{\mathbf{a}}$. Then*

$$\sum_{\mathbf{a}} w(\mathbf{a}) \mathcal{Z}_{\mathbf{a}} \;\prec\; \Psi^{\deg(\Delta)} \Phi^{c(\Delta)} \tag{9.10}$$

for any $\boldsymbol{\mu}$ and chain weight w.

PROOF. The basic idea is to split the summation into partitions

$$\sum_{\mathbf{a}} w(\mathbf{a}) \mathcal{Z}_{\mathbf{a}} \;=\; \sum_P \sum_{\mathbf{a}} \mathbf{1}\big(\mathcal{P}(\boldsymbol{\mu}, \mathbf{a}) = P\big) w(\mathbf{a}) \mathcal{Z}_{\mathbf{a}}\,,$$

where P ranges over all partitions of $V(\Delta)$, and $\mathcal{P}$ was introduced in Definition 4.3. Note that, since $\boldsymbol{\mu}$ are constrained to be distinct, if P yields a nonzero contribution each of its blocks may contain at most one vertex in $V_e(\Delta)$. For the following we fix a partition P and prove that

$$\mathcal{X}_P \;:=\; \sum_{\mathbf{a}} \mathbf{1}\big(\mathcal{P}(\boldsymbol{\mu}, \mathbf{a}) = P\big) w(\mathbf{a}) \mathcal{Z}_{\mathbf{a}}$$

is stochastically bounded by the right-hand side of (9.10). On the level of the graph Δ, a nontrivial partition P of $V(\Delta)$ results in a merging of vertices $V(\Delta)$. By merging vertices of Δ we therefore get a new graph which we denote by $P(\Delta)$. The vertex set of $P(\Delta)$ has the usual decomposition $V(P(\Delta)) = V_e(P(\Delta)) \sqcup V_s(P(\Delta))$, where $V_e(P(\Delta)) = V_e(\Delta)$ and $V_s(P(\Delta))$ is given by the set of blocks of P that do not contain a vertex from $V_e(\Delta)$. A vertex $i \in V(P(\Delta))$ is *unmerged* if the corresponding block has size one, and *merged* otherwise. See Figure 9.9 for an example of the merging $\Delta \mapsto P(\Delta)$.

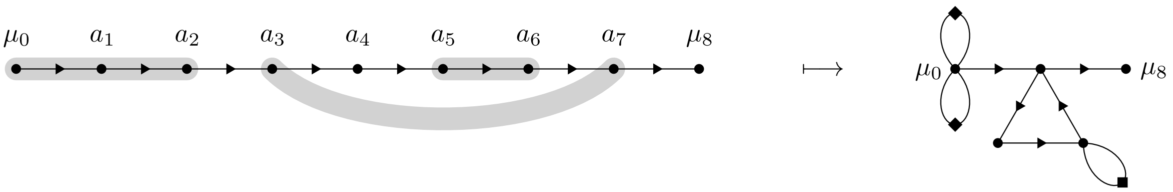

FIGURE 9.9. The merging $\Delta \mapsto P(\Delta)$ of summation vertices of a chain. The vertices of Δ are $V_s(\Delta) = \{1, \dots, 7\}$ and $V_e(\Delta) = \{0, 8\}$. We chose the partition $P = \{\{0,1,2\}, \{3,7\}, \{4\}, \{5,6\}, \{8\}\}$.

For concreteness assume first that Δ is an open chain with $V(\Delta) = \{0, \dots, n\}$, where $V_e(\Delta) = \{0, n\}$. Thus, $\deg(\Delta) = n$. For any graph Δ' define the set of vertices

$$V_g(\Delta') \;:=\; \big\{ i \in V_s(\Delta') : i \text{ has degree two without counting loops} \big\}\,.$$

Thus, the set $V_g(P(\Delta))$ includes not only the chain vertices of $P(\Delta)$ but also chain vertices to which one more loops are attached. Let $r(\Delta')$ denote the number of edges of Δ' that are not loops, and set $c(\Delta') := |V_g(\Delta')|$.

In particular, if Δ' is a chain then this definition agrees with that from Definition 5.1. (For example, in Figure 9.9 we have $c(P(\Delta)) = 2$ and $r(P(\Delta)) = 5$.)

Define $k := n-1-|V_s(P(\Delta))|$. Informally, k is the number of summation vertices of $V_s(\Delta)$ that have been merged into some other vertex. As we shall see, k is the exponent of M^{-1} which describes the reduction in the combinatorics of the summation. (In Figure 9.9 we have $n = 8$ and $|V_s(P(\Delta))| = 3$, which gives $k = 4$.) We claim that

$$n - k \;\leqslant\; r(P(\Delta)) \;\leqslant\; n\,, \qquad r(P(\Delta)) + c(P(\Delta)) \;\geqslant\; 2n - 1 - 2k\,. \tag{9.11}$$

The easiest way to prove (9.11) is by the following inductive argument. We construct $P(\Delta)$ from Δ by successively merging one vertex at a time, and follow the change of the functions $r(\cdot)$ and $r(\cdot) + c(\cdot)$ at each step. The formal procedure is the following. We construct a sequence of graphs $\Delta_0 = \Delta, \Delta_1, \dots, \Delta_k = P(\Delta)$ as follows. We start from $\Delta_0 := \Delta$. Recall that the vertices of Δ are naturally ordered by $\leqslant$. Let $i_1 \in V_s(\Delta)$ be the smallest vertex of Δ that is in a nontrivial block (i.e. of size greater than one) of P. Set Δ_1 to be the graph obtained from Δ_0 by merging i_1 with the (unique) vertex $j \in V(\Delta_0)$ satisfying $j < i_1$. The vertices of Δ_1 remain ordered after we assign the newly created merged vertex the index j. Similarly, Δ_{l+1} is obtained from Δ_l by choosing the smallest unmerged vertex $i_l \in V_s(\Delta_l)$ that is in a nontrivial block of P, and merging it with the unique $j \in V(\Delta_l)$ satisfying $j < i_l$. After k steps of this procedure, we obtain $\Delta_k = P(\Delta)$. Moreover, it is easy to see for $0 \leqslant l \leqslant k-1$ that

$$r(\Delta_l) - 1 \;\leqslant\; r(\Delta_{l+1}) \;\leqslant\; r(\Delta_l)\,, \qquad r(\Delta_{l+1}) + c(\Delta_{l+1}) \;\geqslant\; r(\Delta_l) + c(\Delta_l) - 2\,. \tag{9.12}$$

Indeed, either i_l is merged with a vertex adjacent to itself, in which case we have $r(\Delta_{l+1}) = r(\Delta_l) - 1$ and $c(\Delta_{l+1}) \geqslant c(\Delta_l) - 1$, or i_l is merged with a vertex not adjacent to itself, in which case we have $r(\Delta_{l+1}) = r(\Delta_l)$ and $c(\Delta_{l+1}) \geqslant c(\Delta_l) - 2$. Since $r(\Delta) + c(\Delta) = 2n - 1$, (9.11) follows from (9.12).

We may now sum over $(a_i)_{i \in V_s(P(\Delta))}$. To that end, if $i \in V_g(P(\Delta))$ and there is a loop (or several loops) at i, then we expand each corresponding diagonal term $G_{a_i a_i}$ as $G_{a_i a_i} = m + (G_{a_i a_i} - m)$. If we pick a factor m from each loop, i becomes a chain vertex. If we pick at least one factor $G_{a_i a_i} - m$, i is not a chain vertex but carries a factor of order Ψ. Either way, summing over a_i yields a factor Φ by Proposition 5.3. (Note that Proposition 5.3 is applicable to the graph $P(\Delta)$ because all summation indices are constrained to be distinct.) Thus we get the bound

$$\mathcal{X}_P \;\prec\; M^{-k} \Psi^{r(P(\Delta))} \Phi^{c(P(\Delta))} \;\leqslant\; M^{-k} \Psi^{n-k} \Phi^{n-1-2k}\,,$$

where in the last step we used (9.11). Since $M^{-1}\Psi^{-1}\Phi^{-2} \leqslant \Psi \leqslant 1$ we find $\mathcal{X}_P \prec \Psi^n \Phi^{n-1}$, which is (9.10).

The case of a closed chain Δ of degree n is handled similarly. For definiteness assume that Δ has no external vertex. Now we have $k := n - |V_s(P(\Delta))| \leqslant n-1$ and we let l range from 0 to k. Then (9.12) holds for $l = 0, \dots, n-3$. If $l = n-2$ then (9.12) is in general false (as can be seen e.g. on the open chain of degree two with $V_s(\Delta) = \{1, 2\}$ and $P = \{\{1, 2\}\}$). In that case we replace it with the trivial bounds $r(\Delta_{n-1}) \geqslant 0$ and $c(\Delta_{n-1}) \geqslant 0$. Thus if $k \leqslant n-2$ then we find (9.10) exactly as above, and if $k = n-1$ we get using $n \geqslant 2$

$$\mathcal{X}_P \;\prec\; M^{-k} \Psi^{r(P(\Delta))} \Phi^{c(P(\Delta))} \;\leqslant\; M^{-n+1} \;\leqslant\; \Psi^n \Phi^n\,,$$

which is (9.10). □

To conclude this section, we address an issue concerning coinciding indices that was repeatedly swept under the rug in Sections 7, 8, 9.1, and 9.2. Essentially, we do an inclusion-exclusion argument on the values of the summation indices of a union of chains so as to decouple the summations associated with different subchains. Recall the definition of $X_F^w(\Delta)$ from (4.5).

LEMMA 9.5. *If $\Delta = \Delta_1 \cup \cdots \cup \Delta_k$ is a union*[5] *of chains then*

$$X^w_\emptyset(\Delta) \;\prec\; \prod_{l=1}^k \Psi^{\deg(\Delta_l)}\Phi^{c(\Delta_l)}\,. \tag{9.13}$$

In words: if Δ is a union of chains, then in the summation over $\mathbf{a}$ in $X^w_\emptyset(\Delta)$ we can decouple the summations associated with different subchains of Δ. (In $X^w_\emptyset(\Delta)$ these summations are coupled by the constraint that indices associated with different subchains are distinct.)

PROOF OF LEMMA 9.5. This is a simple decoupling of the summation indices. Let $\mathbf{a}^l$ and $\boldsymbol{\mu}^l$ denote the summation and external indices of Δ_l. Abbreviate $\mathcal{Z}^{\boldsymbol{\mu}^l}_{\mathbf{a}^l}(\Delta_l) \equiv \mathcal{Z}^l_{\mathbf{a}^l}$. Thus we have

$$X_\emptyset(\Delta) \;=\; \sum_{\mathbf{a}^1\cdots\mathbf{a}^k}^{(\boldsymbol{\mu}^1\cdots\boldsymbol{\mu}^k)*} w(\mathbf{a}^1,\dots,\mathbf{a}^k)\mathcal{Z}^1_{\mathbf{a}^1}\cdots\mathcal{Z}^k_{\mathbf{a}^k} \;=\; \sum_{\mathbf{a}^1\cdots\mathbf{a}^k} I(\mathbf{a}^1,\dots,\mathbf{a}^k)\, w(\mathbf{a}^1,\dots,\mathbf{a}^k)\mathcal{Z}^1_{\mathbf{a}^1}\cdots\mathcal{Z}^k_{\mathbf{a}^k}\,,$$

where the indicator function I explicitly enforces that all a^l_i's are distinct from all μ^l_i's and they are district among themselves. Explicitly,

$$\begin{aligned} I(\mathbf{a}^1,\dots,\mathbf{a}^k) \;:=\; &\Bigg[\prod_{l=1}^k \prod_{i,j\in V_s(\Delta_l)}^{*} \big(1-\mathbf{1}(a^l_i = a^l_j)\big)\Bigg]\Bigg[\prod_{l,m\leqslant k}\prod_{i\in V_s(\Delta_l)}\prod_{j\in V_e(\Delta_m)} \big(1-\mathbf{1}(a^l_i = \mu^m_j)\big)\Bigg] \\ &\times\Bigg[\prod_{l,m\leqslant k}^{*}\prod_{i\in V_s(\Delta_l)}\prod_{j\in V_s(\Delta_m)} \big(1-\mathbf{1}(a^l_i = a^m_j)\big)\Bigg]\,. \end{aligned}$$

Multiplying out each parenthesis in the definition of I, we get a splitting of the form $I = \sum_\alpha I_\alpha$ (the sum ranges over a finite set which depends only on Δ). For each α, we may now estimate

$$\sum_{\mathbf{a}^1\cdots\mathbf{a}^k} I_\alpha(\mathbf{a}^1,\dots,\mathbf{a}^k)\, w(\mathbf{a}^1,\dots,\mathbf{a}^k)\mathcal{Z}^1_{\mathbf{a}^1}\cdots\mathcal{Z}^k_{\mathbf{a}^k} \;\prec\; \prod_{l=1}^k \Psi^{\deg(\Delta_l)}\Phi^{c(\Delta_l)}\,. \tag{9.14}$$

To see this, we note that picking the term $\mathbf{1}(\cdots)$ from the parenthesis $(1-\mathbf{1}(\cdots))$ results in the merging of two vertices. Thus, the left-hand side of (9.14) is encoded by a graph $\Delta^{(\alpha)}$ obtained from Δ by merging vertices according to I_α. Whenever two vertices are merged, we may lose two chain vertices, but gain a power M^{-1} from the chain weight (since if indices a and a' coincide, then one of the factors s_{ab} and $s_{a'b'}$ in the chain weight (see (5.1)) can be dropped from the weight and estimated by M^{-1}). The associated loss of Φ^2 is therefore compensated by $M^{-1} \leqslant \Phi^2$. In order to gain from the chain vertices in the merged graph, we invoke Lemma 9.4 to get

$$\sum_{\mathbf{a}'} w'(\mathbf{a}')\mathcal{Z}_{\mathbf{a}'}(\Delta') \;\prec\; \Phi^{\deg(\Delta')}\Phi^{c(\Delta')} \tag{9.15}$$

for each subchain Δ' of $\Delta^{(\alpha)}$. Here (9.15) is applicable because the left-hand side of (9.14) factors into a product of expressions encoded by the subchains of $\Delta^{(\alpha)}$ (i.e. there are no summation constraints that involve two different subchains of $\Delta^{(\alpha)}$). This completes the proof of (9.14), and hence of (9.13). □

[5] By union we mean that the chains $\Delta_1,\dots,\Delta_k$ may share external vertices but not summation vertices.

9.4. Removing Simplification (S1) and completion of the proof of Theorem 4.8. In this section we remove Simplification **(S1)** and put the arguments from Sections 6, 8, 9.1, 9.2, and 9.3 together to complete the proof of Theorem 4.8 in full generality.

The following tensorization property of weights plays a crucial role in this section.

Lemma 9.6. *If $w'(\mathbf{a}')$ and $w''(\mathbf{a}'')$ are weights then so is $w(\mathbf{a}', \mathbf{a}'') := w'(\mathbf{a}')w''(\mathbf{a}'')$.*

Proof. The claim easily follows from Definition 4.4. □

Recall that Simplification **(S1)** states that no index coincidences occur among the indices $\mathbf{a}$ when we compute the p-th power of $X_F(\Delta)$, i.e. in going from Δ to $\gamma^p(\Delta)$. In order to relax Simplification **(S1)**, we go back to Section 6. In this section we add a tilde to the original summation indices in (6.5): $\tilde{\mathbf{a}} = (\tilde{a}_i)_{i \in V_s(\gamma^p(\Delta))}$ (we shall use $\mathbf{a}$ to denote the merged summation indices; see below). Let $\tilde{w}(\tilde{\mathbf{a}})$ denote the product weight (see Lemma 9.6) in the p-fold copy of $X_F(\Delta)$. In general, if we do not assume Simplification **(S1)** then in (6.5) the original summation vertices $\tilde{\mathbf{a}}$ associated with different copies of Δ may coincide. As in the proof of Proposition 9.4, we split the summation using partitions by introducing the factor

$$1 \;=\; \sum_P \mathbf{1}(\mathcal{P}(\tilde{\mathbf{a}}) = P)$$

into the right-hand side of (6.5). Here the summation ranges over partitions of $V_s(\gamma^p(\Delta))$. Thus we get a finite collection of terms indexed by partitions P, which we estimate individually (The combinatorics stemming from the number of partitions is independent of N and will be included in the irrelevant constant prefactors in the final estimate).

Thus, for the sequel we choose and fix a partition P of $V_s(\gamma^p(\Delta))$. If two vertices of $V_s(\gamma^p(\Delta))$ are in the same block of P, we merge them and get a single vertex. Thus we get a new graph which we denote by $\gamma^p_P(\Delta)$. As before we have the splitting $V(\gamma^p_P(\Delta)) = V_e(\gamma^p_P(\Delta)) \sqcup V_s(\gamma^p_P(\Delta))$, where $V_e(\gamma^p_P(\Delta)) = V_e(\gamma^p(\Delta)) = V_e(\Delta)$ and $V_s(\gamma^p_P(\Delta))$ is given by the blocks of P. We use $\mathbf{a} = (a_i)_{i \in V_s(\gamma^p_P(\Delta))}$ to denote the summation indices of the graph $\gamma^p_P(\Delta)$. Each summation vertex of $\gamma^p_P(\Delta)$ is either *unmerged* or *merged*, depending on whether the associated block of P is of size one or greater than one. We have the trivial lift $\tilde{\mathbf{a}} = L_P(\mathbf{a})$ defined by $\tilde{a}_l = a_i$ if l belongs to the block i of P. In merging two vertices i and j in $V_s(\gamma^p(\Delta))$, we lose in general all mechanisms that extract smallness (ingredients (b) and (c) in the list of the guiding principle of Section 5) from them, including the linking associated with the possible factors Q_{a_i} or Q_{a_j}. On the other hand, we gain a factor M^{-1} from the reduction of the combinatorics of the summation. Generally, the reduced summation yields a factor $M^{|V_s(\gamma^p_P(\Delta))| - |V_s(\gamma^p(\Delta))|}$. More precisely,

$$\sum_{\mathbf{a}} w_P(\mathbf{a}) \;\leqslant\; 1\,, \qquad w_P(\mathbf{a}) \;:=\; \tilde{w}(L_P(\mathbf{a})) M^{|V_s(\gamma^p(\Delta))| - |V_s(\gamma^p_P(\Delta))|}\,. \tag{9.16}$$

This follows from (4.4) and the fact that $\tilde{w}$ is a weight by Lemma 9.6. We stress that this is the only point where the assumption (4.4) is needed in our proof.

Having fixed the merging of the vertices, we may now construct all graphs $\Gamma \in \mathfrak{G}^p_{F,P}(\Delta)$; note that this set now depends on P. Here $\mathfrak{G}^p_{F,P}(\Delta)$ is constructed using the same algorithm as $\mathfrak{G}^p_F(\Delta)$ in Section 6. In this case, however, each graph $\Gamma \in \mathfrak{G}^p_{F,P}(\Delta)$ has the property that *unmerged* summation vertices of $\gamma^p_P(\Delta)$ which come with a Q have have been linked with an edge of Γ. There is no similar constraint for merged vertices. (The proof is the same as that for $\mathfrak{G}^p_F(\Delta)$ in Section 6.)

Now we may repeat the arguments of Sections 8, 9.1, 9.2, and 9.3 almost verbatim. The only difference is that we only gain from the *unmerged* vertices of Γ. For example, if $i \in V_s(\Gamma)$ is unmerged and satisfies

$i \in \pi^{-1}(F)$, then it must have been linked with an edge. Similarly, if $i \in V_s(\Gamma)$ is unmerged and marked, it will give rise to a chain vertex after vertex resolution, and hence a factor Φ.

In order to account for the gain from the merged original summation vertices of Γ, we interpret the estimate (9.16) as stating that each summation vertex i of $\gamma^p(\Delta)$ carries a factor $M^{-1/2}$. This means if i is merged then we gain a factor $M^{-1/2}$ over the unmerged scenario. (It is easy to see that this counting corresponds to the worst-case scenario where vertices of $\gamma^p(\Delta)$ were paired to get $\gamma^p_P(\Delta)$. For example, if we have a weight $w(a,b,c,d)$ with $\sum_{abcd} w(a,b,c,d) = 1$ and we merge a with b and c with d, then the new weight $w_P(a,c) = w(a,a,c,c)$ will sum up to

$$\sum_{a,c} w_P(a,c) \;=\; \sum_{a,c} w(a,a,c,c) \;\leqslant\; M^{-2}$$

by (4.4). The gain of order M^{-2} can then be distributed among the four vertices involved in the merging, each receiving a factor $M^{-1/2}$.) This gain of $M^{-1/2}$ compensates any possible gain associated with i, which is at best $\Psi\Phi$ (in the case where $\pi(i)$ is marked and belongs to F). See the guiding principle in Section 5.

The proof is then completed by the simple observation that $M^{-1/2} \leqslant \Psi\Phi$.

10. Proof of Theorem 4.15

In this section we prove Theorem 4.15. The proof relies on some ideas from the proof of Theorem 4.8, but is considerably easier. The strategy is to resolve (using the Family B identities) the summation vertices (associated with indices $\mathbf{a}$) using the partial expectation $\prod_{a \in \mathbf{a}} P_a$, and to estimate the resulting averaging using Theorem 4.8. Thus, unlike in the proof of Theorem 4.8, there is no need to estimate high moments.

Before giving the general proof, let us consider the simple example

$$\begin{aligned}
P_a G_{\mu a} G_{a\mu} \;&=\; P_a \frac{m^2}{G_{aa}^2} G_{\mu a} G_{a\mu} + P_a \biggl(1 - \frac{m^2}{G_{aa}^2}\biggr) G_{\mu a} G_{a\mu} \\
&=\; m^2 P_a \sum_{x,y}^{(a)} G^{(a)}_{\mu x} h_{xa} h_{ay} G^{(a)}_{y\mu} + O_\prec(\Psi^3) \\
&=\; m^2 \sum_{x}^{(a)} s_{ax} G^{(a)}_{\mu x} G^{(a)}_{x\mu} + O_\prec(\Psi^3) \\
&=\; m^2 \sum_{x}^{(a)} s_{ax} G_{\mu x} G_{x\mu} + O_\prec(\Psi^3) \\
&=\; O_\prec(\Psi^2 \Phi)\,,
\end{aligned}$$

where in the second step we used (3.14a) and the bound $\Lambda \prec \Psi$, in the third step (2.1), in the fourth step (3.13), and in the last step Theorem 4.8 (or Proposition 5.3).

The argument for a general graph Δ is similar. We have to gain a factor Φ from each vertex $i \in V_c(\Delta)$ (in addition to the trivial $\deg(\Delta)$ factors Ψ). We use the terminology of Sections 6 – 9 without further comment. The proof consists of the following steps, which we merely sketch as they are almost identical to those of Sections 8 and 9.

(i) Make all entries of $\mathcal{Z}_{\mathbf{a}}(\Delta)$ maximally expanded in $\mathbf{a}$ using the algorithm from the proof of Lemma 6.3. The resulting linking yields a set of graphs $\mathfrak{G}(\Delta)$ satisfying (recall the notation (8.3))

$$\mathcal{Z}_{\mathbf{a}}(\Delta) \;=\; \sum_{\Gamma \in \mathfrak{G}(\Delta)} \mathcal{A}_{\mathbf{a}}(\Gamma) + O_{\prec}(\Psi^{\deg(\Delta)+|V_s(\Delta)|})\,,$$

where all resolvent entries of $\mathcal{A}_{\mathbf{a}}(\Gamma)$ are maximally expanded in $\mathbf{a}$. Each graph $\Gamma \in \mathfrak{G}(\Delta)$ resulted from Δ by a finite number (possibly zero) of linking operations. In particular, $V_s(\Gamma) = V_s(\Delta)$.

(ii) Let $V_m(\Gamma) \subset V_c(\Delta)$ denote those vertices of $V_c(\Delta)$ that were not linked to in the process $\Delta \mapsto \Gamma$. (In other words, $i \in V_m(\Gamma)$ if and only if $\deg_\Delta(i) = \deg_\Gamma(i)$.) We have to gain a factor Φ from each vertex $i \in V_m(\Gamma)$; note that each $i \in V_c(\Delta) \setminus V_m(\Gamma)$ yields a factor Ψ due to the additional edge incident to i produced by the linking to i.

Now we follow the vertex resolution of Sections 8, 9.1, and 9.2 to the letter. The only difference is that the $\mathcal{A}_{\mathbf{a}}(\Gamma)$ is not contained within a full expectation $\mathbb{E}$ but a partial expectation $\prod_{a\in\mathbf{a}} P_a$ instead. We resolve all vertices in $V_s(\Gamma)$, which yields the splitting

$$\mathcal{Z}_{\mathbf{a}}(\Delta) \;=\; \sum_{\Gamma \in \mathfrak{G}(\Delta)} \sum_{\Upsilon \in \mathfrak{L}(\mathfrak{R}(\Gamma))} \sum_{\mathbf{x}}^{(\mathbf{a})} \mathcal{A}_{\mathbf{a},\mathbf{x}}(\Upsilon) + O_{\prec}(\Psi^{\deg(\Delta)+|V_s(\Delta)|})\,,$$

where $\mathbf{x} \in \{1, \dots N\}^{V_f(\Upsilon)}$ denotes the fresh summation indices of Υ.

(iii) Exactly as in Sections 8.4 and 9.1, each vertex $i \in V_m(\Gamma)$ either carries an extra factor Φ (if an error term of subleading order was chosen in the resolution of i) or gives rise to a fresh summation vertex $j \in p^{-1}(i)$ that is a chain vertex of Υ. Hence we may invoke Theorem 4.8, for each fixed $\Upsilon \in \mathfrak{L}(\mathfrak{R}(\Gamma))$, to get

$$\sum_{\mathbf{x}}^{(\mathbf{a})} \mathcal{A}_{\mathbf{a},\mathbf{x}}(\Upsilon) \;\prec\; \Psi^{\deg(\Delta)}\Phi^{|V_c(\Delta)|}\,.$$

This concludes the proof of Theorem 4.15.

A. Basic resolvent bounds

In this appendix we collect some useful tools about resolvents, and in particular prove Lemmas 3.8, 3.9, and 9.1.

Proof of Lemma 3.8. Let $\varepsilon > 0$ and $D > 0$ be arbitrary. From (2.4) and (2.9) we find that there exists $c_0, c_1 \in (0, \varepsilon/2)$ and an event Ξ such that

$$\Lambda(z)\mathbf{1}(\Xi) \;\leqslant\; N^{c_0}\Psi(z) \;\leqslant\; N^{-c_1}$$

for all $z \in \mathbf{S}$ and large enough N, and $\mathbb{P}(\Xi^c) \;\leqslant\; N^{-D}$. Thus we conclude using (3.11) that

$$\sup_{z\in\mathbf{S}} \max_i \Big(\Big| 1/G_{ii}(z) \Big| \mathbf{1}(\Xi) \Big) \;\leqslant\; C$$

for large enough N. Using the first identity of (3.13) and (3.11) again, we find

$$\max_{|T|=\ell}\max_{i,j\notin T}\Big(\big|G_{ij}^{(T)}(z)-\delta_{ij}m(z)\big|\mathbf{1}(\Xi)\Big)\;\leqslant\;CN^{c_0}\Psi(z)\,,\qquad \sup_{z\in\mathbf{S}}\max_{|T|=\ell}\max_{i\notin T}\Big(\Big|1/G_{ii}^{(T)}(z)\Big|\mathbf{1}(\Xi)\Big)\;\leqslant\;C\,. \tag{A.1}$$

for $\ell=1$. Using the first identity of (3.13) and (3.11), we may now proceed inductively on $\ell=1,2,\dots$, at each step proving (A.1) for ℓ assuming it holds for $\ell-1$. The result is

$$\sup_{|T|\leqslant\ell}\max_{i,j\notin T}\Big(\big|G_{ij}^{(T)}(z)-\delta_{ij}m(z)\big|\mathbf{1}(\Xi)\Big)\;\leqslant\;C_\ell N^{c_0}\Psi(z)\;\leqslant\;N^{\varepsilon}\Psi(z) \tag{A.2}$$

for all $z\in\mathbf{S}$. This concludes the proof. □

Proof of Lemma 3.9. The estimate (3.18) follows immediately from $\big|G_{ij}^{(T)}(E+\mathrm{i}\eta)\big|\leqslant\eta^{-1}$ and the definition of $\mathbf{S}$.

In order to prove (3.19), we choose $D:=10p$ and let Ξ denote the event from the proof of Lemma 3.8 above. First we deal with the high-probability event Ξ. From (A.2) and (3.11) we immediately get

$$\sup_{z\in\mathbf{S}}\sup_{|T|\leqslant\ell}\max_{i\notin T}\Big(\Big|1/G_{ii}^{(T)}(z)\Big|\mathbf{1}(\Xi)\Big)\;\leqslant\;C\,. \tag{A.3}$$

In order to handle the exceptional event Ξ^c, we use Schur's formula (3.12). Then by Cauchy-Schwarz, (3.18), and (2.5), we find

$$\mathbb{E}\bigg(\Big|1/G_{ii}^{(T)}(z)\Big|^p\mathbf{1}(\Xi^c)\bigg)\;\leqslant\;\bigg[\mathbb{E}\bigg(\Big|1/G_{ii}^{(T)}(z)\Big|^{2p}\mathbf{1}(\Xi^c)\bigg)\bigg]^{1/2}\mathbb{P}(\Xi^c)^{1/2}\;\leqslant\;(C+N^3)^pN^{-5p}\,. \tag{A.4}$$

Combining (A.3) and (A.4) yields (3.19). □

Proof of Lemma 9.1. To simplify notation, we set $T=\emptyset$ (the proof for nonempty T is the same). A simple large deviation estimate (see e.g. Lemmas B.1 and B.2 in [15]) applied to

$$Z_i\;=\;\sum_k^{(i)}\big(|h_{ik}|^2-s_{ik}\big)G_{kk}^{(i)}+\sum_{k\neq l}^{(i)}h_{ik}G_{kl}^{(i)}h_{li}$$

implies $Z_i\prec\Psi$.

As above, for the estimate of $U_i^{(S)}$ we set $S=\emptyset$ to simplify notation. Using (2.3) we write

$$U_i\;=\;\sum_k s_{ik}(G_{kk}-m)\;=\;\sum_k s_{ik}P_k(G_{kk}-m)+\sum_k s_{ik}Q_k(G_{kk}-m)\;=\;\sum_k s_{ik}P_k(G_{kk}-m)+O_\prec(\Psi^2)\,,$$

where the last step follows from Proposition 6.1. Now we expand the inverse of (3.14c) using (2.7) to get

$$G_{kk}-m\;=\;m^2\big(-h_{kk}+Z_k+U_k^{(k)}\big)+O_\prec(\Psi^2)\,,$$

where we estimated the higher-order terms using (3.11) and the trivial bounds $h_{ii} \prec \Psi$, $U_i^{(i)} \prec \Psi$, and $Z_i \prec \Psi$ (as proved in the previous paragraph). Using $P_k h_{kk} = 0$ and $P_k Z_k = 0$ we therefore get

$$\begin{aligned} U_i \;&=\; m^2 \sum_k s_{ik} P_k U_k^{(k)} + O_\prec(\Psi^2) \\ &=\; m^2 \sum_k s_{ik} \Biggl(\sum_l^{(k)} s_{kl} P_k G_{ll}^{(k)} - m \Biggr) + O_\prec(\Psi^2) \\ &=\; m^2 \sum_{k,l} s_{ik} s_{kl} (G_{ll} - m) + O_\prec(\Psi^2) \\ &=\; m^2 \sum_k s_{ik} U_k + O_\prec(\Psi^2)\,, \end{aligned}$$

where in the third step we used (2.2), (2.3), (2.9), and (3.13). Inverting the operator $1 - m^2 S$ therefore yields $U_i \prec \varrho \Psi^2$. On the other hand, the estimate $U_i \prec \Psi$ is trivial. This concludes the proof. □

B. The coefficient ϱ for band matrices

In this section we prove an explicit bound for the coefficient ϱ defined in (3.2), in the case that S is the variance matrix of a band matrix H, as defined in Example 2.1. In fact, we need only that the spectrum $\sigma(S)$ of S is separated away from -1; that is this always true for band matrices is the content of the following lemma.

Lemma B.1. *Suppose that H is a d-dimensional band matrix from Example 2.1. Then there is a constant $\delta_- > 0$, depending only on the profile function f, such that $\sigma(S) \subset [-1 + \delta_-, 1]$.*

Proof. See [15, Lemma A.1]. □

Proposition B.2. *Let S be a doubly stochastic matrix satisfying $\sigma(S) \subset [-1+\delta_-, 1]$ for some $\delta_- > 0$. Then there is a universal constant C such that*

$$\varrho \;\leqslant\; \frac{C \log N}{\min\{\delta_-, (\operatorname{Im} m)^2\}}\,. \tag{B.1}$$

In particular, using Lemma B.1 we find that (B.1) *holds for a d-dimensional band matrix from Example 2.1, with a constant C depending only on the profile function f.*

The rest of this appendix is devoted to the proof of Proposition B.2. A similar argument was given in the proof of [15, Lemma 3.5]. The main difference is that here we do not assume the existence of a spectral gap near $+1$ in the spectrum of S.

Proof of Proposition B.2. Abbreviate $\zeta := m^2$ and write

$$\frac{1}{1 - \zeta S} \;=\; \frac{1/2}{1 - (1 + \zeta S)/2}\,.$$

We have the bound

$$\left\|\frac{1+\zeta S}{2}\right\|_{\ell^\infty\to\ell^\infty} \leqslant \max_i \sum_j \left|\left(\frac{1+\zeta S}{2}\right)_{ij}\right| \leqslant 1\,,$$

where we used that $|\zeta| \leqslant 1$ as follows from (3.11). By the condition on the spectrum, we have

$$\left\|\frac{1+\zeta S}{2}\right\|_{\ell^2\to\ell^2} \leqslant \max_{x\in[-1+\delta_-,1]} \frac{|1+\zeta x|}{2} \leqslant \max\left\{1-\frac{\delta_-}{2}, \frac{|1+\zeta|}{2}\right\}.$$

An elementary calculation yields $1 - |1+\zeta|/2 \geqslant c(\operatorname{Im} m)^2$ for some constant $c > 0$, from which we conclude

$$\left\|\frac{1+\zeta S}{2}\right\|_{\ell^2\to\ell^2} \leqslant 1 - c\min\{\delta_-, (\operatorname{Im} m)^2\} \tag{B.2}$$

for some small universal constant $c > 0$. For $n_0 \in \mathbb{N}$ we therefore have

$$\begin{aligned}\left\|\frac{1}{1-\zeta S}\right\|_{\ell^\infty\to\ell^\infty} &\leqslant \sum_{n=0}^{n_0-1}\left\|\frac{1+\zeta S}{2}\right\|^n_{\ell^\infty\to\ell^\infty} + \sqrt{N}\sum_{n=n_0}^{\infty}\left\|\frac{1+\zeta S}{2}\right\|^n_{\ell^2\to\ell^2}\\ &\leqslant n_0 + \sqrt{N}\big(1-c\min\{\delta_-,(\operatorname{Im} m)^2\}\big)^{n_0}\frac{C}{\min\{\delta_-,(\operatorname{Im} m)^2\}}\\ &\leqslant \frac{C\log N}{\min\{\delta_-,(\operatorname{Im} m)^2\}}\,,\end{aligned}$$

where the last step follows by taking $n_0 = C_0 \log N/\min\{\delta_-,(\operatorname{Im} m)^2\}$ for large enough $C_0 > 0$. □

References

[1] L. Erdős and A. Knowles, *Quantum diffusion and delocalization for band matrices with general distribution*, Ann. H. Poincaré **12** (2011), 1227–1319.

[2] ———, *Quantum diffusion and eigenfunction delocalization in a random band matrix model*, Comm. Math. Phys. **303** (2011), 509–554.

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

[4] ———, *The local semicircle law for a general class of random matrices*, Preprint arXiv:1212.0164.

[5] ———, *Spectral statistics of Erdős-Rényi graphs I: Local semicircle law*, to appear in Ann. Prob. Preprint arXiv:1103.1919.

[6] ———, *Spectral statistics of Erdős-Rényi graphs II: Eigenvalue spacing and the extreme eigenvalues*, to appear in Comm. Math. Phys. Preprint arXiv:1103.3869.

[7] L. Erdős, S. Péché, J.A. Ramirez, B. Schlein, and H.T. Yau, *Bulk universality for Wigner matrices*, Comm. Pure Appl. Math. **63** (2010), 895–925.

[8] L. Erdős, J. Ramirez, B. Schlein, T. Tao, V. Vu, and H.T. Yau, *Bulk universality for Wigner hermitian matrices with subexponential decay*, Math. Res. Lett. **17** (2010), 667–674.

[9] L. Erdős, J. Ramirez, B. Schlein, and H.T. Yau, *Universality of sine-kernel for Wigner matrices with a small Gaussian perturbation*, Electr. J. Prob. **15** (2010), 526–604.

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

[11] ______, *Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices*, Ann. Prob. **37** (2009), 815–852.

[12] ______, *Wegner estimate and level repulsion for Wigner random matrices*, Int. Math. Res. Not. **2010** (2009), 436–479.

[13] ______, *Universality of random matrices and local relaxation flow*, Invent. Math. **185** (2011), no. 1, 75–119.

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

[15] L. Erdős, H.T. Yau, and J. Yin, *Bulk universality for generalized Wigner matrices*, Preprint arXiv:1001.3453.

[16] ______, *Rigidity of eigenvalues of generalized Wigner matrices*, to appear in Adv. Math. Preprint arXiv:1007.4652.

[17] ______, *Universality for generalized Wigner matrices with Bernoulli distribution*, J. Combinatorics **1** (2011), no. 2, 15–85.

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