Delocalization and quantum diffusion of random band matrices in high dimensions I: Self-energy renormalization

Fan Yang
University of Pennsylvania
fyang75@wharton.upenn.edu

Horng-Tzer Yau
Harvard University
htyau@math.harvard.edu

Jun Yin
University of California, Los Angeles
jyin@math.ucla.edu

We consider Hermitian random band matrices $H = (h_{xy})$ on the d-dimensional lattice $(\mathbb{Z}/L\mathbb{Z})^d$. The entries h_{xy} are independent (up to Hermitian conditions) centered complex Gaussian random variables with variances $s_{xy} = \mathbb{E}|h_{xy}|^2$. The variance matrix $S = (s_{xy})$ has a banded structure so that s_{xy} is negligible if $|x-y|$ exceeds the band width W. In dimensions $d \geqslant 8$, we prove that, as long as $W \geqslant L^{\varepsilon}$ for a small constant $\varepsilon > 0$, with high probability most bulk eigenvectors of H are delocalized in the sense that their localization lengths are comparable to L. Denote by $G(z) = (H - z)^{-1}$ the Green's function of H. For $\operatorname{Im} z \gg W^2/L^2$, we also prove a widely used criterion in physics for quantum diffusion of this model, namely, the leading term in the Fourier transform of $\mathbb{E}|G_{xy}(z)|^2$ with respect to $x - y$ is of the form $(\operatorname{Im} z + a(p))^{-1}$ for some $a(p)$ quadratic in p, where p is the Fourier variable. Our method is based on an expansion of $T_{xy} = |m|^2 \sum_\alpha s_{x\alpha}|G_{\alpha y}|^2$ and it requires a self-energy renormalization up to error W^{-K} for any large constant K independent of W and L. We expect that this method can be extended to non-Gaussian band matrices.

The work of F.Y. is partially supported by the Wharton Dean's Fund for Postdoctoral Research. The work of H.-T. Y. is partially supported by the NSF grant DMS-1855509 and a Simons Investigator award. The work of J.Y. is partially supported by the NSF grant DMS-1802861.

1 Introduction

1.1 Random band matrices

Wigner envisioned [62] that spectral properties of quantum systems of high complexity can be modeled by Gaussian random matrices such as GOE (Gaussian orthogonal ensemble) or GUE (Gaussian unitary ensemble). Although many non-rigorous arguments and numerical simulations support his thesis, rigorous works have been mostly restricted to mean-field models such as Wigner matrices or the adjacency matrices of random graphs of various types. For non-mean-field models, the understanding of their spectral properties is much more limited. One important non-mean-field model is the random Schrödinger operator or more specifically, the Anderson model [5]. More precisely, the d-dimensional Anderson model is defined by a Hamiltonian $H = -\Delta + \lambda V$, where Δ is the graph Laplacian on $\mathbb{Z}^d$, V is a random potential with i.i.d. entries, and λ is a small coupling strength. This model is highly non-mean-field because the off-diagonal elements consist of only $2d$ entries of constant value in each row or column, while all the randomness is in the diagonal elements. In the strong disorder regime, i.e., when λ is large, the eigenvectors of the Anderson model are expected to be localized and the local eigenvalue statistics converge to a Poisson process; in the weak disorder regime, the eigenvectors are expected to be delocalized and the local eigenvalue statistics coincide with those of a GOE or GUE. The localization was first proved rigorously by Fröhlich and Spencer [38] using a multi-scale analysis; an alternative proof was given years later by Aizenman and Molchanov [2] using a fractional moment method. Many spectacular results have been proved regarding the localization of the Anderson model (see, e.g., [14, 15, 19, 20, 37, 40, 46]). The existence of the delocalized regime for the Anderson model has only been proved for the Bethe lattice [3, 4], but not for any finite-dimensional integer lattice $\mathbb{Z}^d$.

A model that is more tractable than the Anderson model but still preserves its key non-mean-field property is the following random band matrix ensemble. Let $\mathbb{Z}_L^d := \{1, 2, \cdots, L\}^d$ be a lattice of linear size L, and $N \equiv L^d$ be the total number of lattice sites. A d-dimensional random band matrix ensemble consists of $N \times N$ random Hermitian matrices $H = (h_{xy})_{x,y\in\mathbb{Z}_L^d}$, whose entries h_{xy} are centered random variables that are independent up to the Hermitian condition $h_{xy} = \overline{h}_{yx}$. In this paper, we require that s_{xy} be negligible when $|x - y| \gg W$ for some length scale $1 \ll W \ll L$ and satisfy the normalization condition

$$\sum_x s_{xy} = \sum_y s_{xy} = 1. \tag{1.1}$$

It is well-known that under the condition (1.1), the global eigenvalue distribution of H converges weakly to the Wigner's semicircle law supported in $[-2, 2]$. As W varies, the random band matrices naturally interpolate between the random Schrödinger operator [5] and the mean-field Wigner ensemble [62].

A key physical quantity for both the Anderson model and random band matrices is the *localization length* ℓ, which, roughly speaking, is the length scale of the region in which most weight of an eigenvector resides. There are different ways to define the localization length depending on how the eigenvector decays outside the localized region (e.g. polynomial decay, exponential decay, etc.). For the Anderson model in infinite volume, an eigenvector is localized if its localization length is finite, and delocalized otherwise. For random band matrices, one can define an eigenvector to be *delocalized* if its localization length ℓ is comparable with the linear size L of the system, and *localized* otherwise. It should be remarked that localization and delocalization in general depend on the energy levels. In this paper, we will restrict ourselves to the bulk eigenvectors, that is, eigenvectors with eigenvalues in $(-2 + \kappa, 2 - \kappa)$ for some small constant $\kappa > 0$ independent of L.

We assume for the moment that the majority of the bulk eigenvectors have similar localization lengths so that we can refer to the localization length of a random band matrix. The localization length $\ell \equiv \ell(d, W)$ is expected to increase with W and an (almost) sharp localization-delocalization transition occurs at some critical band width $W_c \equiv W_c(d, L)$ when the localization length ℓ becomes comparable to the system size L, i.e.,

- for $W \gg W_c$, the bulk eigenvectors are delocalized, i.e. ℓ is of order similar to L;
- for $W \ll W_c$, the bulk eigenvectors are localized, i.e. ℓ is much smaller than L.

Heuristically, the random band matrices and the Anderson model are expected to have same qualitative properties with $\lambda \sim W^{-1}$. In dimension $d = 1$, the localization length of the Anderson model is known to be of order $\ell \sim \lambda^{-2}$. By simulations [17, 18, 36, 63] and non-rigorous supersymmetric arguments [39], the localization length of one-dimensional random band matrices is conjectured to be of order $\ell \sim W^2$, leading to the critical band width $W_c \sim \sqrt{L}$. The localization length of the two-dimensional Anderson model is conjectured to be exponentially large in λ^{-2} [1] (although this conjecture is not universally accepted). Correspondingly, it is conjectured that the localization length of two-dimensional random band matrices also grows exponentially fast in W^2, leading to the critical band width $W_c \sim \sqrt{\log L}$. In dimensions $d \geqslant 3$, it is conjectured that there is a threshold energy in the Anderson model, the mobility edge, that separates the delocalized and localized states. For random band matrices with $d \geqslant 3$, the bulk eigenvectors are conjectured to be delocalized. More precisely, the localization length of the bulk eigenvectors is expected to be of macroscopic scale $\ell \sim L$ independently of the band width W, and the critical band width W_c is a large number independent of L. The previous summary on the localization-delocalization conjecture is mainly focused on the random band matrices, and we refer the reader to [10, 57, 59, 60] for more details. There are extensive works concerning this problem for random Schrödinger operators in the past several decades; they are beyond the scope of this paper and we refer the reader to [16, 45, 58] for extensive reviews.

There have been many partial results concerning these conjectures for random band matrices in dimension $d = 1$ [6, 11–13, 24, 25, 27, 28, 42, 48–56, 65]. Key results include that $\ell < W^7$ if the entries of H are Gaussian [48], and $\ell > W^{4/3}$ for general random band matrices without Gaussian assumption [12, 13, 65]. For a class of complex Hermitian Gaussian random band matrices with certain special variance profiles, supersymmetry techniques can be used [6, 21, 23, 50–55], and a transition in the two-point correlation function for the bulk eigenvalues at $W_c \sim L^{1/2}$ was proved in [52]. It is still not clear if the supersymmetry method can be adapted to prove localization or delocalization of the random band matrices treated in [52].

The understanding of the delocalization of random band matrices in dimensions $d \geqslant 2$, however, is much more limited. Based on studying the unitary operator $e^{\mathrm{i}tH}$, it was shown [24, 25] that the localization length for d-dimensional random band matrices satisfies $\ell > W^{1+d/6}$. The delocalization used in these papers is defined in a weak sense which we will explain later on. With a Green's function method, it was proved [65] that $\ell > W^{1+d/2}$, improving the earlier results obtained in [27, 42].

In this paper, we prove that with high probability in dimensions $d \geqslant 8$, the bulk eigenvectors of random band matrices are (weakly) delocalized in the sense defined in [24, 25] provided that $W \geqslant L^{\varepsilon}$ for a small constant $\varepsilon > 0$. Recall that the delocalization conjecture asserts that random band matrices are delocalized in dimensions $d \geqslant 3$ as long as $W \geqslant C$ for a large enough constant $C > 0$. Our result gives a positive answer to this conjecture for $d \geqslant 8$ in the weak delocalization sense under the slightly stronger assumption $W \geqslant L^{\varepsilon}$ (vs. $W \geqslant C$). The definition of the delocalization used in this paper, following [24, 25], is still far from the strong delocalization used for Wigner matrices [31, 32]. Major works remain to be done to prove the strong delocalization even under the conditions $d \geqslant 8$ and $W \geqslant L^{\varepsilon}$. We will discuss some of these problems after stating the main results.

1.2 Delocalization and local law

In this subsection, we define our model and state the first two main results, Theorem 1.3 and Theorem 1.4, of this paper. We will consider d-dimensional random band matrices indexed by a cube of linear size L in $\mathbb{Z}^d$, i.e.,

$$\mathbb{Z}_L^d := \left(\mathbb{Z} \cap (-L/2, L/2]\right)^d . \tag{1.2}$$

We will view $\mathbb{Z}_L^d$ as a torus and denote by $[x-y]_L$ the representative of $x-y$ in $\mathbb{Z}_L^d$, i.e.,

$$[x-y]_L := \left[(x-y) + L\mathbb{Z}^d\right] \cap \mathbb{Z}_L^d. \tag{1.3}$$

Clearly, $\|x-y\|_L := \|[x-y]_L\|$ is the periodic distance on $\mathbb{Z}_L^d$ for any norm $\|\cdot\|$ on $\mathbb{Z}^d$. For definiteness, we use ℓ^∞-norm in this paper, i.e. $\|x-y\|_L := \|[x-y]_L\|_\infty$. In this paper, we consider the following class of d-dimensional random band matrices.

Assumption 1.1 (Random band matrix $H \equiv H_{d,f,W,L}$)**.** *Fix any $d \in \mathbb{N}$. For $L \gg W \gg 1$ and $N := L^d$, we assume that $H \equiv H_{d,f,W,L}$ is an $N \times N$ complex Hermitian random matrix whose entries $(\operatorname{Re} h_{xy}, \operatorname{Im} h_{xy} : x, y \in \mathbb{Z}_L^d)$ are independent Gaussian random variables (up to the Hermitian condition $h_{xy} = \overline{h}_{yx}$) such that*

$$\mathbb{E}h_{xy} = 0, \quad \mathbb{E}(\operatorname{Re} h_{xy})^2 = \mathbb{E}(\operatorname{Im} h_{xy})^2 = s_{xy}/2, \quad x, y \in \mathbb{Z}_L^d, \tag{1.4}$$

where the variances s_{xy} satisfy that

$$s_{xy} = f_{W,L}\left([x-y]_L\right) \tag{1.5}$$

for some positive symmetric function $f_{W,L}$ satisfying Assumption 1.2 below. Then we say that H is a d-dimensional random band matrix with the linear size L, band width W and variance profile $f_{W,L}$. Denote the variance matrix by $S := (s_{xy})_{x,y\in\mathbb{Z}_L^d}$, which is a doubly stochastic symmetric $N \times N$ matrix.

Assumption 1.2 (Variance profile)**.** *We assume that $f_{W,L} : \mathbb{Z}_L^d \to \mathbb{R}_+$ is a positive symmetric function on $\mathbb{Z}_L^d$ that can be expressed by the Fourier transform*

$$f_{W,L}(x) := \frac{1}{(2\pi)^d Z_{W,L}} \int \psi(Wp) e^{\mathrm{i}p\cdot x} \mathrm{d}p. \tag{1.6}$$

Here $\mathbb{Z}_{W,L}$ is the normalization constant so that $\sum_{x\in\mathbb{Z}_L^d} f_{W,L}(x) = 1$, and $\psi \in C^\infty(\mathbb{R}^d)$ is a symmetric smooth function independent of W and L and satisfies the following properties:

(i) $\psi(0) = 1$ *and* $\|\psi\|_\infty \leqslant 1$;

(ii) $\psi(p) \leqslant \max\{1 - c_\psi|p|^2, 1 - c_\psi\}$ *for a constant* $c_\psi > 0$;

(iii) ψ *is in the Schwartz space, i.e.,*

$$\lim_{|p|\to\infty} (1 + |p|)^k |\psi^{(l)}(p)| = 0, \quad \text{for any } k, l \in \mathbb{N}. \tag{1.7}$$

Clearly, $f_{W,L}$ is of order $\mathrm{O}(W^{-d})$ and decays faster than any polynomial, that is, for any fixed $k \in \mathbb{N}$, there exists a constant $C_k > 0$ so that

$$|f_{W,L}(x)| \leqslant C_k W^{-d} \left(\|x\|_L/W\right)^{-k}. \tag{1.8}$$

Hence the variance profile S defined in (1.5) has a banded structure, namely, for any constants $\tau, D > 0$,

$$\mathbf{1}_{|x-y|\geqslant W^{1+\tau}} |s_{xy}| \leqslant W^{-D}. \tag{1.9}$$

Combining (1.7) and (1.8) with the Poisson summation formula, we obtain that

$$Z_{W,L} = \psi(0) + \mathrm{O}(W^{-D}) = 1 + \mathrm{O}(W^{-D}), \tag{1.10}$$

for any large constant $D > 0$ as long as $L \geqslant W^{1+\varepsilon}$ for a constant $\varepsilon > 0$. Note that Assumption 1.2 does not cover non-smooth profile functions. For example, it does not include the indicator function $f_{W,L}(x) = W^{-d}\mathbf{1}_{x\in(-W/2,W/2]^d}$. While we believe that Assumption 1.2 is not essential, we will not get into this technical issue in this paper.

Denote the eigenvalues and normalized eigenvectors of H by $\{\lambda_\alpha\}$ and $\{\mathbf{u}_\alpha\}$. According to [59], an eigenvector $\mathbf{u}_\alpha$ is localized with a localization length ℓ if for some $x_0 \in \mathbb{Z}_L^d$ its entries satisfy that

$$|u_\alpha(x)| \leqslant C e^{-c\|x-x_0\|_L/\ell}, \tag{1.11}$$

for some constants $c, C > 0$. Inspired by this definition, for any fixed constants $K > 1$ and $0 < \gamma \leqslant 1$, we define a random subset of indices as

$$\mathcal{B}_{\gamma,K,\ell} := \left\{ \alpha : \lambda_\alpha \in (-2+\kappa, 2-\kappa) \ \text{ so that } \min_{x_0\in\mathbb{Z}_L^d} \sum_x |u_\alpha(x)|^2 \exp\left[\left(\frac{\|x-x_0\|_L}{\ell}\right)^\gamma\right] \leqslant K \right\},$$

which contains all indices associated with bulk eigenvectors that have localization lengths bounded by $\mathrm{O}(\ell)$. Here we have relaxed the exponential function in (1.11) to a more general family of sub-exponential functions. Then we have the following theorem for random band matrices in dimensions $d \geqslant 8$.

Theorem 1.3 (Weak delocalization of bulk eigenvectors in high dimesnions)**.** *Fix $d \geqslant 8$, small constants $c_0, c_1, \gamma, \kappa > 0$ and a large constant $K > 1$. Suppose that $W \leqslant \ell \leqslant L^{1-c_0}$, $L^{c_1} \leqslant W \leqslant L$ and H is a d-dimensional random band matrix satisfying Assumptions 1.1 and 1.2. Then we have that for any constants $\tau, D > 0$,*

$$\mathbb{P}\left[\frac{|\mathcal{B}_{\gamma,K,\ell}|}{N} \leqslant W^\tau \left(\frac{\ell^2}{L^2} + W^{-d/2}\right)\right] \geqslant 1 - L^{-D}, \tag{1.12}$$

provided that L is sufficiently large depending on these constants. Moreover, for any eigenvalue λ_α of H satisfying $\lambda_\alpha \in (-2+\kappa, 2-\kappa)$, its eigenvector $\mathbf{u}_\alpha$ satisfies that

$$\mathbb{P}\left(\|\mathbf{u}_\alpha\|_\infty \leqslant W^{1+\tau}/L\right) \geqslant 1 - L^{-D}, \tag{1.13}$$

for any constants $\tau, D > 0$ and sufficiently large L.

The estimate (1.12) asserts that, for random band matrices with band width essentially of order one (L^{c_1} for any small constant $c_1 > 0$), the majority of bulk eigenvectors have localization lengths essentially of the size of the system (in the sense that they are larger than L^{1-c_0} for any small constant $c_0 > 0$). The bound (1.13) implies that $\|\mathbf{u}_\alpha\|_4^4 \leqslant W^{2+2\tau}/L^2$ with high probability, which converges to 0 as $L \to \infty$, another commonly used weak notion of delocalization in physics (see, e.g., [59]). We also remark that all the results in this paper hold only for large enough W and L, and, for simplicity, we do not repeat it again in all our statements.

To prove Theorem 1.3, we study the resolvent (or Green's function) of H defined by

$$G(z) = (H - z)^{-1}, \quad z \in \mathbb{C}_+ := \{x \in \mathbb{C} : \operatorname{Im} z > 0\}.$$

In [28, 33], it has been shown that for any small constant $\varepsilon > 0$,

$$\max_{x,y} |G_{xy}(z) - m(z)\delta_{xy}| \leqslant \frac{W^\varepsilon}{\sqrt{W^d \eta}}, \quad z = E + \mathrm{i}\eta, \tag{1.14}$$

with high probability for all $E \in (-2+\kappa, 2-\kappa)$ and $\eta \geqslant W^{-d+\varepsilon}$, where $m(z)$ is the Stieltjes transform of Wigner's semicircle law,

$$m(z) := \frac{-z + \sqrt{z^2 - 4}}{2} = \frac{1}{2\pi} \int_{-2}^{2} \frac{\sqrt{4 - \xi^2}}{\xi - z} \mathrm{d}\xi, \quad z \in \mathbb{C}_+. \tag{1.15}$$

The bound (1.14) implies a lower bound on the localization length of order W, which is far shorter than L^{1-c_0} stated in Theorem 1.3 when $W = L^{c_1}$. For our purpose, we need to decrease η from $\eta \geqslant W^{-d+\varepsilon}$ to a much smaller scale $\eta \geqslant W^2/L^{2-\varepsilon}$ and improve the error bound in (1.14) significantly. While the diagonal resolvent entry G_{xx} is expected to be given by the semicircle law for a large range of η, we will show that the off-diagonal entries can be approximated by a *diffusive kernel* Θ defined by

$$\Theta(z) := \frac{|m(z)|^2 S}{1 - |m(z)|^2 S}. \tag{1.16}$$

It is well-known that for $z = E + \mathrm{i}\eta$ with $E \in (-2+\kappa, 2-\kappa)$ and $\eta \geqslant W^2/L^{2-\varepsilon}$ for a constant $\varepsilon > 0$,

$$\Theta_{xy}(z) \leqslant \frac{W^\tau \mathbf{1}_{|x-y| \leqslant \eta^{-1/2} W^{1+\tau}}}{W^2 \left(\|x-y\|_L + W\right)^{d-2}} + \frac{1}{\left(\|x-y\|_L + W\right)^D} \leqslant W^\tau B_{xy}, \tag{1.17}$$

for any constants $\tau, D > 0$, where we have abbreviated that

$$B_{xy} := W^{-2} \left(\|x-y\|_L + W\right)^{-d+2}. \tag{1.18}$$

The reader can refer to [27, 65] for a proof of (1.17). The following theorem provides an essentially sharp local law on the resolvent entries under the assumptions of Theorem 1.3.

Theorem 1.4 (Local law)**.** *Under the assumptions of Theorem 1.3, for any small constants $\varepsilon, \tau > 0$ and large constant $D > 0$, we have the following estimate on $G(z)$ for $z = E + \mathrm{i}\eta$ and all $x, y \in \mathbb{Z}_L^d$:*

$$\mathbb{P}\bigg(\sup_{E \in (-2+\kappa, 2-\kappa)} \sup_{W^2/L^{2-\varepsilon} \leqslant \eta \leqslant 1} |G_{xy}(z) - m(z)\delta_{xy}|^2 \leqslant W^\tau B_{xy} \bigg) \geqslant 1 - L^{-D}. \tag{1.19}$$

The proof of Theorem 1.4 is based on an expansion method and can be readily adapted to non-Gaussian random band matrices after some technical modifications (cf. Remark 3.23 below for more details). We choose to present it for Gaussian cases to avoid technical complexities associated with non-Gaussian distributions (which will increase the number of terms in expansions). The proof for the real symmetric Gaussian case is very similar to the complex case except that, as usual, the number of terms will double in every expansion step. (This is due to the fact that $\mathbb{E} h_{xy}^2 = 0$ in the complex case but not in the real case.) The condition $d \geqslant 8$ can also be improved; it remains to be seen whether this method can reach the physical dimension $d = 3$. We will deal with these improvements in forthcoming papers.

A very strong notion of delocalization is to require that

$$\mathbb{P}\left(\|\mathbf{u}_\alpha\|_\infty \leqslant L^{-d/2+\tau} \right) \geqslant 1 - L^{-D}$$

for any constants $\tau, D > 0$. This was first proved for Wigner matrices in [31–34] and later extended to many other classes of mean-field type random matrices (see e.g. [7–9, 13, 29, 41, 43]). This estimate was proved in [33] as a consequence of the following bound on the diagonal resolvent entries, i.e., for some constant $C > 0$,

$$\max_{x\in\mathbb{Z}_L^d} |G_{xx}(z)| \leqslant C \quad \text{for all } \eta \gg L^{-d}.$$

We believe that the resolvents of random band matrices satisfy the following stronger estimate with high probability:

$$|G_{xy}(z) - m(z)\delta_{xy}|^2 \leqslant W^\tau B_{xy} + W^\tau (L^d\eta)^{-1} \quad \text{for all } \eta \gg L^{-d}. \tag{1.20}$$

The restriction $\eta \gg W^2/L^2$ in this paper is not intrinsic and can be substantially improved to, say, $\eta \gg L^{-d/4}$. However, it seems to be a difficult problem to reach the optimal threshold $\eta \gg L^{-d}$.

1.3 Quantum diffusion

A key quantity in the analysis of random band matrices is the T-matrix introduced in [27]:

$$T_{xy}(z) := |m|^2 \sum_\alpha s_{x\alpha} |G_{\alpha y}(z)|^2, \quad x, y \in \mathbb{Z}_L^d. \tag{1.21}$$

Note that T_{xy} is very similar to $|G_{xy}|^2$, and it is known that the T-matrix controls the asymptotic behaviors of the resolvent (see, e.g., Lemma 5.1 below). Moreover, the T-variables are slightly easier to use in our proof, because the diagonal T-variables T_{xx} can be dealt with in the same way as the off-diagonal T-variables T_{xy} with $x \neq y$, while this is not the case for the $|G_{xy}|^2$ variables. In the following theorem, we show that $\mathbb{E}T_{xy}$ is governed by a diffusion profile.

Theorem 1.5 (Quantum diffusion of the T-matrix)**.** *Suppose the assumptions of Theorem 1.3 hold. Fix any small constant $\varepsilon > 0$ and large constant $M \in \mathbb{N}$. Then for all $x, y \in \mathbb{Z}_L^d$ and $z = E + \mathrm{i}\eta$ with $E \in (-2+\kappa, 2-\kappa)$ and $W^2/L^{2-\varepsilon} \leqslant \eta \leqslant 1$, we have that*

$$\mathbb{E}T_{xy} = \left[\Theta^{(M)}\left(|m|^2 + \mathcal{G}^{(M)}\right)\right]_{xy} + \mathrm{O}(W^{-Md/2}). \tag{1.22}$$

Here $\Theta^{(M)}$ is the M-th order renormalized diffusive matrix

$$\Theta^{(M)} := \frac{1}{1 - |m|^2 S\left(1 + \Sigma^{(M)}\right)} |m|^2 S, \tag{1.23}$$

and it satisfies the bound

$$\left|\Theta^{(M)}_{xy}\right| \leqslant L^\tau B_{xy}, \tag{1.24}$$

for any constant $\tau > 0$. Furthermore, the self-energy correction *$\Sigma^{(M)}$ is given by $\Sigma^{(M)}(z) := \sum_{l=4}^M \mathcal{E}_l(z)$ where $\{\mathcal{E}_l\}_{l=4}^M$ is a sequence of deterministic matrices satisfying the following properties:*

$$\mathcal{E}_l(x, x+a) = \mathcal{E}_l(0, a), \quad \mathcal{E}_l(0, a) = \mathcal{E}_l(0, -a), \quad \forall\ x, a \in \mathbb{Z}_L^d, \tag{1.25}$$

and for any constant $\tau > 0$,

$$|(\mathcal{E}_l)_{0x}(z)| \leqslant L^\tau W^{-(l-4)d/2} B_{0x}^2, \quad \forall\ x \in \mathbb{Z}_L^d,\ \eta \in [W^2/L^{2-\varepsilon}, 1], \tag{1.26}$$

$$\Big|\sum_{x\in\mathbb{Z}_L^d} (\mathcal{E}_l)_{0x}(z)\Big| \leqslant L^\tau \eta W^{-(l-2)d/2}, \quad \forall\ \eta \in [W^2/L^{2-\varepsilon}, 1]. \tag{1.27}$$

Here in (1.25) and throughout the rest of this paper, we use $\mathcal{A}_{xy}$ and $\mathcal{A}(x, y)$ interchangeably for any matrix $\mathcal{A}$. The M-th order local correction *$\mathcal{G}^{(M)}$ satisfies that*

$$\left|\mathcal{G}^{(M)}_{xy}\right| \leqslant L^\tau B_{xy}^{3/2}, \tag{1.28}$$

for any constant $\tau > 0$.

We believe that the bound (1.28) can be improved to $|\mathcal{G}^{(M)}_{xy}| \leqslant L^\tau B^2_{xy}$ with some extra work, but we do not pursue this improvement in this paper for simplicity. From Theorem 1.5, we can readily obtain the following quantum diffusion of the resolvent entries.

Corollary 1.6 (Quantum diffusion)**.** *Under the assumptions of Theorem 1.5, we have that*

$$\mathbb{E}|G_{xy}|^2 = \left[\frac{1}{1-\left(1+\Sigma^{(M)}\right)|m|^2 S}\left(|m|^2+\mathcal{G}^{(M)}\right)\right]_{xy} + \mathrm{O}(W^{-Md/2}), \tag{1.29}$$

for all $x, y \in \mathbb{Z}_L^d$ and $z = E + \mathrm{i}\eta$ with $E \in (-2+\kappa, 2-\kappa)$ and $W^2/L^{2-\varepsilon} \leqslant \eta \leqslant 1$.

Taking $M = 0$, we get the (0-th order) diffusive matrix $\Theta^{(0)} \equiv \Theta$ in (1.16). We can expand Θ into a geometric series

$$\Theta_{xy} = \sum_{k=1}^{\infty} |m|^{2k}(S^k)_{xy}. \tag{1.30}$$

By (1.1), S is the transition matrix of a random walk on $\mathbb{Z}_L^d$ with step size $\mathrm{O}(W)$. With direct calculations, we can check that $|m|^2 = 1-\eta/r(E) + \mathrm{O}(\eta^2)$ where $r(E) := \sqrt{4-E^2}/2$ is proportional to the semicircle density. Hence (1.30) shows that Θ_{xy} is a superposition of random walks up to the time η^{-1}, which is the main reason why we call Θ the diffusive matrix. Due to the form of s_{xy} in (1.5), Θ_{xy} is translationally invariant on $\mathbb{Z}_L^d$. Moreover, the Fourier transform of Θ_{xy} with respect to $x-y$ is given by

$$\widehat{\Theta}(p) := \sum_x \Theta_{0x} e^{\mathrm{i}p\cdot x} = \frac{|m|^2\widehat{S}_{W,L}(p)}{1-|m|^2\widehat{S}_{W,L}(p)}, \quad \text{with} \quad p \in \mathbb{T}_L^d := \left(\frac{2\pi}{L}\mathbb{Z}_L\right)^d, \quad \widehat{S}_{W,L}(p) := \sum_x s_{0x} e^{\mathrm{i}p\cdot x}.$$

Note that by (1.6), $\widehat{S}_{W,L}(p)$ is equal to $\psi(Wp)$ up to a small error when L is large. In the regime $|p| \ll W^{-1}$ and $\eta \ll 1$, this equation gives the following diffusion approximation:

$$\widehat{\Theta}(p) = \frac{|m|^2\widehat{S}_{W,L}(p)}{(1-|m|^2)+|m|^2[1-\widehat{S}_{W,L}(p)]} = \frac{r(E)\left[1+\mathrm{O}(\eta+W^2|p|^2)\right]}{\eta + W^2 p\cdot \mathcal{D}(E)p + \mathrm{O}(\eta^2+W^3|p|^3)}, \tag{1.31}$$

with an effective diffusion coefficient (matrix) $\mathcal{D}(E)$ defined by

$$\mathcal{D}_{ij} := \frac{r(E)}{2}\sum_{i,j}\frac{x_i x_j}{W^2}s_{0x}, \quad 1 \leqslant i,j \leqslant d.$$

The matrix $\Theta^{(M)}$ can be viewed as a diffusion propagator with an M-th order *self-energy renormalization* to the diffusion constant. For any $4 \leqslant l \leqslant M$, the property (1.25) shows that $(\mathcal{E}_l)_{xy}$ is translationally invariant, and $(\mathcal{E}_l)_{0x}$ is symmetric in x. Thus its Fourier transform in x, $\widehat{\mathcal{E}}_l(p)$, is a symmetric function in p. Using the properties (1.26) and (1.27), it is easy to check that for $|p| \ll W^{-1}$ and any constant $\tau > 0$,

$$\widehat{\mathcal{E}}_l(p) = W^{-(l-2)d/2}W^2 p\cdot\mathcal{D}_l(z)p + \mathrm{O}\left[W^{-(l-2)d/2+\tau}\left(\eta+W^3|p|^3\right)\right], \tag{1.32}$$

where $\mathcal{D}_l(z)$ is defined by

$$(\mathcal{D}_l)_{ij}(z) := W^{(l-2)d/2}\cdot\frac{1}{2}\sum_{i,j}\frac{x_i x_j}{W^2}(\mathcal{E}_l)_{0x}(z), \quad 1 \leqslant i,j \leqslant d.$$

Note that the main error in (1.32) comes from the $l = 4$ case. Using (1.31) and (1.32), we can write the Fourier transform of $\Theta^{(M)}$ as

$$\widehat{\Theta}^{(M)}(p) = \frac{r(E)\left[1+\mathrm{O}(\eta+W^2|p|^2)\right]}{\eta+W^2 p\cdot\mathcal{D}^{(M)}_{eff}(z)p + \mathrm{O}\left(\eta^2+W^{-d+\tau}\eta+W^3|p|^3\right)}, \tag{1.33}$$

for $|p| \ll W^{-1}$, where the renormalized effective diffusion coefficient is defined as

$$\mathcal{D}^{(M)}_{eff}(z) := \mathcal{D}(E) + r(E)\sum_{l=4}^{M} W^{-(l-2)d/2}\mathcal{D}_l(z). \tag{1.34}$$

Therefore, $\Theta^{(M)}$ is a diffusion propagator with $\mathcal{E}_l$ being the l-th order self-energy.

The matrix $\mathcal{G}^{(M)}$ in (1.28) represents the collective effects of local recollisions. Notice that each row of $\mathcal{G}^{(M)}$ has a summable decay and its ℓ^1 norm is small in the sense that $\sum_y |\mathcal{G}^{(M)}_{xy}| = \mathrm{O}(W^{-d/2+\tau})$. In particular, this shows that $|\widehat{\mathcal{G}}^{(M)}(p)| = \mathrm{O}(W^{-d/2+\tau})$. Thus the Fourier transform of (1.22) is given by

$$\sum_x \mathbb{E}T_{0x}(z)e^{\mathrm{i}p\cdot x} = \widehat{\Theta}^{(M)}(p)\left[|m|^2 + \mathrm{O}(W^{-d/2+\tau})\right] + \mathrm{O}(W^{-M}), \tag{1.35}$$

as long as M is sufficiently large. By (1.29), the Fourier transform of $\mathbb{E}|G_{xy}|^2$ has a similar behavior for $|p| \ll W^{-1}$. To summarize, we have obtained the following corollary from Theorem 1.5 and Corollary 1.6.

Corollary 1.7. *Under the assumptions of Theorem 1.5, let M be a large constant satisfying $M \geqslant 4\log_W L$. For $p \in \mathbb{T}_L^d$ with $|p| \ll W^{-1}$ and $W^2/L^{2-\varepsilon} \leqslant \eta \ll 1$, (1.35) and the following estimate hold for any small constant $\tau > 0$:*

$$\sum_x \mathbb{E}|G_{0x}(z)|^2 e^{\mathrm{i}p\cdot x} = \frac{r(E)\left[|m|^2 + \mathrm{O}(W^{-d/2+\tau})\right]}{\eta + W^2 p\cdot \mathcal{D}^{(M)}_{eff}(z)p + \mathrm{O}\left(\eta^2 + W^{-d+\tau}\eta + W^3|p|^3\right)} + \mathrm{O}(W^{-M}). \tag{1.36}$$

It is commonly believed in physics literature (see, e.g., [59,60]) that

$$\sum_x \mathbb{E}|G_{0x}(z)|^2 e^{\mathrm{i}p\cdot x} \sim \frac{1}{\eta + a(p)} \tag{1.37}$$

with $a(p)$ being a quadratic form of p for small $|p|$ is a signature of quantum diffusion. Hence Corollary 1.7 shows that the resolvent is diffusive for $\eta \geqslant W^2/L^{2-\varepsilon}$.The quantum diffusion for the Anderson's model was proved in [30] for time scale $t \sim \lambda^{-2-c}$ for some small constant $c > 0$. If we take the correspondence $t \sim \eta^{-1}$ and $\lambda \sim W^{-d/2}$, the result in [30] amounts to establishing the quantum diffusion for $\eta \sim W^{-d-c}$ in the current language. The quantum diffusion in [30] was established for the unitary evolution $e^{\mathrm{i}tH}$ instead of (1.37) in terms of the resolvent. While the two formulations of the quantum diffusion are generally believed to be roughly equivalent, lots of works are still required to prove the quantum diffusion for the unitary evolution $e^{\mathrm{i}tH}$ of random band matrices. However, we believe that there are no intrinsic difficulties for such results.

The *Thouless time* [22,59,61] for random band matrices is defined to be the time for a particle to reach the boundary of the system, which is roughly $t_{Th} = L^2/W^2$ if we assume that the particle evolves as a diffusion. It is generally believed, at least heuristically, that the localization/delocalization and quantum diffusion properties of a disordered system can be determined by the behavior of the resolvent up to the Thouless time. Since η and the time t are dual variables, the assumption $\eta \gg W/L^2$ in Corollary 1.7 exactly corresponds to that the evolution time is less than the Thouless time. In other words, Corollary 1.7 establishes the quantum diffusion in resolvent sense up to the Thouless time.

1.4 T-expansion

The main tool to prove Theorems 1.3, 1.4 and 1.5 is an expansion of the T-matrix up to arbitrarily high order. In [27], the T-matrix was shown to satisfy a T-equation to the leading order, which gives a T-expansion up to second order in $W^{-d/2}$ (i.e., up to order W^{-d}) as follows. From (1.21), it is trivial to derive the following equation

$$T_{xy} = \Theta_{xy}(|G_{yy}|^2 - T_{yy}) + \sum_{\alpha\neq y}\Theta_{x\alpha}(|G_{\alpha y}|^2 - T_{\alpha y}). \tag{1.38}$$

Since we have

$$T_{yy} = |m|^2 s_{yy}|G_{yy}(z)|^2 + |m|^2\sum_{\alpha\neq y} s_{y\alpha}|G_{\alpha y}(z)|^2 \leqslant CW^{-d}|G_{yy}(z)|^2 + \sup_{\alpha\neq y}|G_{\alpha y}(z)|^2,$$

we expect that $T_{yy} = \mathrm{O}(W^{-d})$ with high probability and thus is an error term. We now show that $|G_{\alpha y}|^2 - T_{\alpha y}$ also gives a higher order term. Using the equation $(z+m)m = -1$ for $m(z)$, we get that

$$G = -\frac{1}{z+m} + \frac{1}{z+m}(H+m)G \quad \Rightarrow \quad G - m = -m(H+m)G. \tag{1.39}$$

Here the expansion in terms of $H+m$, instead of H, can be viewed as a naive renormalized expansion. Define $\mathbb{E}_x$ as the partial expectation with respect to the x-th row and column of H, i.e., $\mathbb{E}_x(\cdot) := \mathbb{E}(\cdot|H^{(x)})$, where $H^{(x)}$ denotes the $(N-1)\times(N-1)$ minor of H obtained by removing the x-th row and column. For simplicity, in this paper we will use the notations

$$P_x := \mathbb{E}_x, \quad Q_x := 1-\mathbb{E}_x.$$

Using (1.39), we get that for $x \neq y$,

$$|G_{xy}|^2 = P_x\left(G_{xy}\overline{G}_{xy}\right) + Q_x|G_{xy}|^2 = -P_x\Big[\Big(m^2 G_{xy} + m\sum_\alpha h_{x\alpha}G_{\alpha y}\Big)\overline{G}_{xy}\Big] + Q_x|G_{xy}|^2.$$

Using Gaussian integration by parts with respect to $h_{x\alpha}$, we obtain that

$$\begin{aligned}
|G_{xy}|^2 &= Q_x|G_{xy}|^2 - m^2 P_x|G_{xy}|^2 - mP_x\Big[\sum_\alpha s_{x\alpha}\partial_{h_{\alpha x}}\left(G_{\alpha y}\overline{G}_{xy}\right)\Big] \\
&= Q_x|G_{xy}|^2 + mP_x\Big[\sum_\alpha s_{x\alpha}\left(G_{\alpha\alpha}-m\right)|G_{xy}|^2\Big] + mP_x\Big(\overline{G}_{xx}\sum_\alpha s_{x\alpha}|G_{\alpha y}|^2\Big) \\
&= |m|^2\sum_\alpha s_{x\alpha}|G_{\alpha y}|^2 + \Omega_{xy} = T_{xy} + \Omega_{xy},
\end{aligned} \tag{1.40}$$

where Ω_{xy} consists of diagonal error terms (i.e., terms depending on $G_{\alpha\alpha}-m$ and $\overline{G}_{xx}-\overline{m}$) and fluctuations (i.e., terms of the form $Q_x[\cdot]$). Inserting (1.40) into (1.38), we obtain that

$$T_{xy} = \Big[|m|^2 + \mathrm{O}(W^{-d/2})\Big]\Theta_{xy} + \sum_{\alpha\neq y}\Theta_{x\alpha}\Omega_{\alpha y}, \tag{1.41}$$

if we have a diagonal estimate $G_{yy} = m + \mathrm{O}(W^{-d/2})$ with high probability.

We expect the second term in (1.41) to be an error term. But we have that

$$\sum_y \Theta_{xy}(z) = \frac{|m(z)|^2}{1-|m(z)|^2} \sim \eta^{-1}, \tag{1.42}$$

which makes the error $\sum_{\alpha\neq y}\Theta_{x\alpha}\Omega_{\alpha y}$ bigger than the order of $\max_{\alpha,y}|\Omega_{\alpha y}|$ by a huge factor η^{-1} if we bound the sum naively. Thus this error is very difficult to bound when $\eta \ll 1$ (in particular, when $\eta = W^2/L^{2-\varepsilon}$). The estimate of $\sum_{\alpha\neq y}\Theta_{x\alpha}\Omega_{\alpha y}$ can be improved by a fluctuation averaging lemma, which was first discovered in [35] and later extended to random band matrices in [26]. This leads to, roughly speaking, the following bound in [27]: for any small constant $\tau > 0$,

$$\Big|\sum_{\alpha\neq y}\Theta_{x\alpha}\Omega_{\alpha y}\Big| \leqslant \eta^{-1}W^{-3d/2+\tau} \quad \text{with high probability for} \quad \eta \gg W^{-d}. \tag{1.43}$$

It was noticed later [65] that one can take advantage of the decay of $\Theta_{x\alpha}$ and $\Omega_{\alpha y}$ with respect to α to improve the error estimate. In order to achieve the regime $L \geqslant W^C$ for an arbitrarily large constant $C>0$, the previous methods will require that $|\Omega_{\alpha y}| \ll \eta$, which is almost impossible to establish and very likely to be incorrect.

While the T-equation has drawbacks, it is already a big step towards the understanding of the T-matrix. Recall that Θ is a random walk expansion up to the time η^{-1}. Hence to prove that $T_{xy} \sim \Theta_{xy}$ for $\eta \sim W^{-d}$, it amounts to expanding the resolvent $(H-z)^{-1}$ at least W^d times. This will generate a huge combinatorial factor $(W^d \times W^d)!$ in calculating $\mathbb{E}|G_{xy}|^2$ using Gaussian contractions. This combinatorial factor makes it infeasible to use the naive expansion method even taking into account various renormalization simplifications in the calculations. The T-equation method bypasses the problem of analyzing the $(W^d\times W^d)!$ many error terms at the expense of showing that the error is bounded up to the accuracy $|\Omega_{\alpha y}| \ll \eta$. Returning to the current case with $W = L^\varepsilon$ and $\eta \sim W^2/L^2$, the naive expansion will generate $(L^2/W^2 \times L^2/W^2)!$ many terms and it is again hopeless to analyze them. Thus we have to study the T-equation more deeply and seek for a crucial replacement of the bound $|\Omega_{\alpha y}| \ll \eta$.

One key observation of this work is that main contributions to the term $\sum_{\alpha\neq y}\Theta_{x\alpha}\Omega_{\alpha y}$ come from self-energy related terms such as $(\Theta\Sigma^{(M)})^k\Theta$ in the Taylor expansion of $\Theta^{(M)}$. Suppose for now we replace the property (1.27) by a stronger sum zero property

$$\sum_x (\mathcal{E}_l)_{xy} = 0. \tag{1.44}$$

Together with the fact that $(\mathcal{E}_l)_{xy}$ is symmetric in x and y, we can sum by parts twice in the expression $\sum_\alpha \Theta_{x\alpha}\,(\mathcal{E}_l)_{\alpha y}$ to get

$$|(\Theta\mathcal{E}_l)_{xy}| \leqslant W^{-(l-4)d/2+\tau}\sum_\alpha \left|\partial_\alpha^2\Theta_{x\alpha}\right| B_{\alpha y}^2 \leqslant \frac{W^{-(l-2)d/2+2\tau}}{(\|x-y\|_L+W)^d}, \tag{1.45}$$

where we also used the bound (1.26) for $\mathcal{E}_l$ and $|\partial_\alpha^2\Theta_{x\alpha}| \lesssim (\|x-\alpha\|_L+W)^{-d+\tau}$ for any constant $\tau>0$. (Strictly speaking, $\partial_\alpha^2\Theta_{x\alpha}$ should be replaced by the second order difference of $\Theta_{x\alpha}$ in α.) Using this estimate, it is easy to get that for any small constant $\tau>0$,

$$\left|\left[(\Theta\Sigma^{(M)})^k\Theta\right]_{xy}\right| \leqslant W^\tau B_{xy}. \tag{1.46}$$

Although the row sums of $\mathcal{E}_l$ are not exactly equal to zero by (1.27), the η factor in the error term will be small enough to cancel the factor from $\sum_\alpha \Theta_{x\alpha}\sim\eta^{-1}$. To summarize, the self-energies in the T-expansion need to either satisfy a sum zero property or contain effectively an η factor. If we take $\eta\to 0$ as $L\to\infty$, then an exact sum zero property will hold for the *infinite space limit* of $\mathcal{E}_l$ (see equation (2.16) for a more precise statement). Hence we will call (1.27) a *sum zero property.*

Our main task is thus to design an expansion method to derive a T-equation with the leading term $\Theta^{(n)}$ and an error of order $\mathrm{O}(W^{-(n+1)d/2})$ for any fixed $n\in\mathbb{N}$. But there will also be many other types of terms. Roughly speaking, we will derive an expression of the form

$$T = \Theta^{(n)} + (\text{recollision term}) + (\text{higher order term}) + (\text{fluctuation term}) + (\text{error term}), \tag{1.47}$$

where the recollision term consists of expressions with coincidences in summation indices, the higher order term consists of expressions that are of order smaller than $W^{-nd/2}$, the fluctuation term consists of expressions that can be written into the form $\sum_x Q_x(\cdot)$ (which can be analyzed via the fluctuation averaging mechanism), and the error term can be neglected for all of our proofs. In the expansion process, we will need to give a precise construction of $\Theta^{(n)}$. Furthermore, the recollision, higher order and fluctuation terms will also need to be tracked relatively explicitly and some key structures (which we call the *doubly connected structures*) need to be maintained in order to derive the final estimates on these terms. The expansion (1.47) is constructed inductively in n. Roughly speaking, with the T-expansion (1.47) for a given n, we insert itself into a suitable subset of expressions in (1.47) to derive the $(n+1)$-th order T-expansion. The main technical difficulties are to verify the sum zero properties for the self-energies $\mathcal{E}_l$ order by order, and to maintain the doubly connected structures for all the other expressions so that we can estimate them. We want to point out that the typical sizes of Θ_{xy} and G_{xy} are of order B_{xy} and $B_{xy}^{1/2}$, respectively. Moreover, the row sums of B_{xy} are bounded by L^2/W^2, while the row sums of $B_{xy}^{3/2}$ are bounded by $\mathrm{O}(W^{-d/2})$. The doubly connected structures defined in Definition 6.5 below ensure that in each sum, we have at least a product of a Θ factor and a G factor, so that the sum can be bounded independently of L.

The proof of the main results in this paper and [64] can be roughly divided into the following three parts: (i) construction of the T-expansion, (ii) proof of the sum zero properties for the self-energies, (iii) proof of Theorems 1.3, 1.4 and 1.5 using the T-expansion. In this paper, we will complete (ii) and (iii), while (i) and some estimates used in (iii) will be proved in the second paper of this series [64]. We stress that our strategy to construct the T-expansion is not a straightforward extension of the one used in [26, 65]. In terms of the terminology to be introduced in this paper, the expansions in [26, 65] are *local expansions.* The construction of the full T-expansion will require the much more sophisticated *global expansions*, which will be explained in Section 3.5 and Section 9.

The rest of this paper is organized as follows. In Section 2, we introduce the graphical tools and use them to define the core concepts of this paper—the T-expansion and self-energies. In Section 3, we introduce the basic graph operations that are used to construct the T-expansion. In Section 4, we give some examples of how to use the basic graph operations to obtain some lower order T-expansions. In Section 5, we give the proof of Theorem 2.1, a slightly weaker version of Theorem 1.4, based on some lemmas that will be proved in Sections 6–7 and the second paper of this series [64]. We will also discuss the restriction $d \geqslant 8$ after the continuity estimate, Lemma 5.3. In Section 6, we introduce the doubly connected structures of the graphs. In Section 7, we study the infinite space limits of self-energies. Finally, the proofs of Theorem 1.3, Theorem 1.4, Theorem 1.5 and Corollary 1.6 will be presented in Section 8. In Section 9, we discuss some key new ideas in [64] that are used to prove the relevant lemmas in Section 5.

Acknowledgements. J.Y. would like to thank Xinyi Cui and Xuyang Tian for helpful discussions.

2 T-expansion and self-energies

The major part of this paper is devoted to proving Theorem 1.4. We will mainly focus on proving the following slightly weaker form of Theorem 1.4, which assumes a stronger compactly supported condition on ψ. The reason is that in the setting of Theorem 2.1, the sum zero property (2.16) can be stated in a cleaner form. In Section 8, we will show how to adapt the proof for Theorem 2.1 to the proof of Theorem 1.4.

Theorem 2.1. *Under the assumptions of Theorem 1.4, we assume in addition that ψ in Definition 1.2 is a compactly supported smooth function. Then for any small constants $\varepsilon, \tau > 0$ and large constant $D > 0$, the estimate* (1.19) *holds.*

In this section, we introduce the following three key tools for the proof of Theorem 2.1: self-energies in Definition 2.13, the T-expansion in Definition 2.15, and the T-equation in Definition 2.17. With these tools, we will give an outline of the proof of Theorem 2.1 in Section 5. Before stating the T-expansion, we introduce two deterministic matrices

$$S^+(z) := \frac{m^2(z)S}{1-m^2(z)S}, \quad S^-(z) := \overline{S}^+(z), \tag{2.1}$$

which satisfy the following estimate (2.3). For simplicity, throughout the rest of this paper, we abbreviate

$$|x-y| \equiv \|x-y\|_L, \quad \langle x-y\rangle \equiv \|x-y\|_L + W. \tag{2.2}$$

Lemma 2.2. *Suppose Assumptions 1.1 and 1.2 hold, and $z = E + \mathrm{i}\eta$ with $E \in (-2+\kappa, 2-\kappa)$ for a constant $\kappa > 0$. Then for any constants $\tau, D > 0$, we have that*

$$|S^{\pm}_{xy}(z)| \lesssim W^{-d}\mathbf{1}_{|x-y|\leqslant W^{1+\tau}} + \langle x-y\rangle^{-D}. \tag{2.3}$$

Proof. The estimate (2.3) is a folklore result. A formal proof for the $d = 1$ case is given in equation (4.21) of [12]. This proof can be extended directly to the general d case. □

In this paper, we adopt the following convention of stochastic domination [26].

Definition 2.3 (Stochastic domination and high probability event)**.** *(i) Let*

$$\xi = \left(\xi^{(W)}(u) : W \in \mathbb{N}, u \in U^{(W)}\right), \quad \zeta = \left(\zeta^{(W)}(u) : W \in \mathbb{N}, u \in U^{(W)}\right),$$

be two families of non-negative random variables, where $U^{(W)}$ is a possibly W-dependent parameter set. We say ξ is stochastically dominated by ζ, uniformly in u, if for any fixed (small) $\tau > 0$ and (large) $D > 0$,

$$\mathbb{P}\Bigg[\bigcup_{u\in U^{(W)}} \left\{\xi^{(W)}(u) > W^\tau \zeta^{(W)}(u)\right\}\Bigg] \leqslant W^{-D}$$

for large enough $W \geqslant W_0(\tau, D)$, and we will use the notation $\xi \prec \zeta$. If for some complex family ξ we have $|\xi| \prec \zeta$, then we will also write $\xi \prec \zeta$ or $\xi = \mathrm{O}_\prec(\zeta)$.

(ii) As a convention, for two deterministic non-negative quantities ξ and ζ, we will write $\xi \prec \zeta$ if and only if $\xi \leqslant W^\tau \zeta$ for any constant $\tau > 0$.

(iii) We say that an event Ξ holds with high probability (w.h.p.) if for any constant $D > 0$, $\mathbb{P}(\Xi) \geqslant 1 - W^{-D}$ for large enough W. More generally, we say that an event Ω holds w.h.p. in Ξ if for any constant $D > 0$, $\mathbb{P}(\Xi \setminus \Omega) \leqslant W^{-D}$ for large enough W.

2.1 *Second order T-expansion*

We generalize the T-variable in (1.21) to the following T-variables with three subscripts:

$$T_{x,yy'} := |m|^2 \sum_\alpha s_{x\alpha} G_{\alpha y} \overline{G}_{\alpha y'}, \quad \text{and} \quad T_{yy',x} := |m|^2 \sum_\alpha G_{y\alpha} \overline{G}_{y'\alpha} s_{\alpha x}. \tag{2.4}$$

By definition, the T-variable in (1.21) can be written as $T_{xy} \equiv T_{x,yy}$. Our T-expansion will be formulated in terms of these generalized T-variables. In this subsection, we define the second order T-expansion of $T_{x,yy'}$ using the following Θ expansion, which is derived from Gaussian integration by parts. The expansion of $T_{yy',x}$ can be obtained by considering the transposition of $T_{x,yy}$.

Lemma 2.4 (Θ-expansion)**.** *In the setting of Theorem 1.3, consider the expression $|m|^2 \sum_\alpha s_{x\alpha} G_{\alpha y} \overline{G}_{\alpha y'} f(G)$, where f is a differentiable function of G. Then we have the identity*

$$\begin{aligned} |m|^2 \sum_\alpha s_{x\alpha} G_{\alpha y} \overline{G}_{\alpha y'} f(G) &= m\Theta_{xy} \overline{G}_{yy'} f(G) + m \sum_{\alpha,\beta} \Theta_{x\alpha} s_{\alpha\beta} (G_{\beta\beta} - m) G_{\alpha y} \overline{G}_{\alpha y'} f(G) \\ &+ m \sum_{\alpha,\beta} \Theta_{x\alpha} s_{\alpha\beta} (\overline{G}_{\alpha\alpha} - \overline{m}) G_{\beta y} \overline{G}_{\beta y'} f(G) - m \sum_{\alpha,\beta} \Theta_{x\alpha} s_{\alpha\beta} G_{\beta y} \overline{G}_{\alpha y'} \partial_{h_{\beta\alpha}} f(G) + \mathcal{Q}_\Theta, \end{aligned} \tag{2.5}$$

where

$$\begin{aligned} \mathcal{Q}_\Theta := &\sum_\alpha \Theta_{x\alpha} Q_\alpha \left[G_{\alpha y} \overline{G}_{\alpha y'} f(G) \right] - m\Theta_{xy} Q_y \left[\overline{G}_{yy'} f(G) \right] - \sum_{\alpha,\beta} m\Theta_{x\alpha} s_{\alpha\beta} Q_\alpha \left[(G_{\beta\beta} - m) G_{\alpha y} \overline{G}_{\alpha y'} f(G) \right] \\ &- \sum_{\alpha,\beta} m\Theta_{x\alpha} s_{\alpha\beta} Q_\alpha \left[\overline{G}_{\alpha\alpha} G_{\beta y} \overline{G}_{\beta y'} f(G) \right] + \sum_{\alpha,\beta} m\Theta_{x\alpha} s_{\alpha\beta} Q_\alpha \left[G_{\beta y} \overline{G}_{\alpha y'} \partial_{h_{\beta\alpha}} f(G) \right]. \end{aligned}$$

Proof. With (1.39) and the identity $|m|^2 S = \Theta - |m|^2 \Theta S$, we can write that

$$\begin{aligned} |m|^2 \sum_\alpha s_{x\alpha} P_\alpha \left[G_{\alpha y} \overline{G}_{\alpha y'} f(G) \right] &= \sum_\alpha \left[\Theta_{x\alpha} - |m|^2 (\Theta S)_{x\alpha} \right] P_\alpha \left[G_{\alpha y} \overline{G}_{\alpha y'} f(G) \right] \\ &= - \sum_\alpha |m|^2 (\Theta S)_{x\alpha} P_\alpha \left[G_{\alpha y} \overline{G}_{\alpha y'} f(G) \right] + \sum_\alpha \Theta_{x\alpha} P_\alpha \left[(m\delta_{\alpha y} - m^2 G_{\alpha y} - m(HG)_{\alpha y}) \overline{G}_{\alpha y'} f(G) \right]. \end{aligned} \tag{2.6}$$

For the HG term, using Gaussian integration by parts we obtain that

$$\begin{aligned} P_\alpha \left[-m \sum_\beta h_{\alpha\beta} G_{\beta y} \overline{G}_{\alpha y'} f(G) \right] &= P_\alpha \left[m \sum_\beta s_{\alpha\beta} G_{\beta\beta} G_{\alpha y} \overline{G}_{\alpha y'} f(G) + m \sum_\beta s_{\alpha\beta} G_{\beta y} \overline{G}_{\alpha\alpha} \overline{G}_{\beta y'} f(G) \right] \\ &+ P_\alpha \left[-m \sum_\beta s_{\alpha\beta} G_{\beta y} \overline{G}_{\alpha y'} \partial_{h_{\beta\alpha}} f(G) \right]. \end{aligned}$$

Using this equation, we get that

$$\begin{aligned} &\sum_\alpha \Theta_{x\alpha} P_\alpha \left[(m\delta_{\alpha y} - m^2 G_{\alpha y} - m(HG)_{\alpha y}) \overline{G}_{\alpha y'} f(G) \right] \\ &= m\Theta_{xy} P_y \left[\overline{G}_{yy'} f(G) \right] + \sum_\alpha \Theta_{x\alpha} P_\alpha \left[m \sum_\beta s_{\alpha\beta} (G_{\beta\beta} - m) G_{\alpha y} \overline{G}_{\alpha y'} f(G) \right] \\ &+ \sum_\alpha \Theta_{x\alpha} P_\alpha \left[m(\overline{G}_{\alpha\alpha} - \overline{m}) \sum_\beta s_{\alpha\beta} G_{\beta y} \overline{G}_{\beta y'} f(G) \right] + \sum_\alpha \Theta_{x\alpha} P_\alpha \left[|m|^2 \sum_\beta s_{\alpha\beta} G_{\beta y} \overline{G}_{\beta y'} f(G) \right] \\ &- \sum_\alpha \Theta_{x\alpha} P_\alpha \left[m \sum_\beta s_{\alpha\beta} G_{\beta y} \overline{G}_{\alpha y'} \partial_{h_{\beta\alpha}} f(G) \right]. \end{aligned}$$

Plugging it into (2.6), writing $P_\alpha = 1 - Q_\alpha$, and using the identity

$$\sum_{\alpha,\beta} |m|^2 \Theta_{x\alpha} s_{\alpha\beta} P_\alpha \left[G_{\beta y} \overline{G}_{\beta y'} f(G) \right] - \sum_\alpha |m|^2 (\Theta S)_{x\alpha} P_\alpha \left[G_{\alpha y} \overline{G}_{\alpha y'} f(G) \right]$$

$$= \sum_{\alpha} |m|^2 (\Theta S)_{x\alpha} Q_\alpha \left[G_{\alpha y} \overline{G}_{\alpha y'} f(G) \right] - \sum_{\alpha,\beta} |m|^2 \Theta_{x\alpha} s_{\alpha\beta} Q_\alpha \left[G_{\beta y} \overline{G}_{\beta y'} f(G) \right],$$

we can obtain (2.5) after some simple calculations. □

Using the Θ-expansion (2.5), we obtain the following second order T-expansion.

Lemma 2.5. *Under the assumptions of Theorem 1.3, we have that for any* $\mathfrak{a}, \mathfrak{b}_1, \mathfrak{b}_2 \in \mathbb{Z}_L^d$,

$$T_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = m\Theta_{\mathfrak{a}\mathfrak{b}_1} \overline{G}_{\mathfrak{b}_1\mathfrak{b}_2} + (\mathcal{A}_T^{(>2)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + (\mathcal{Q}_T^{(2)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}, \tag{2.7}$$

where

$$(\mathcal{A}_T^{(>2)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} := m \sum_{x,y} \Theta_{\mathfrak{a}x} s_{xy} (G_{yy} - m) G_{x\mathfrak{b}_1} \overline{G}_{x\mathfrak{b}_2} + m \sum_{x,y} \Theta_{\mathfrak{a}x} s_{xy} (\overline{G}_{xx} - \overline{m}) G_{y\mathfrak{b}_1} \overline{G}_{y\mathfrak{b}_2}, \tag{2.8}$$

$$\begin{aligned}(\mathcal{Q}_T^{(2)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} &:= \sum_x Q_x \left(\Theta_{\mathfrak{a}x} G_{x\mathfrak{b}_1} \overline{G}_{x\mathfrak{b}_2} \right) - m Q_{\mathfrak{b}_1} \left(\Theta_{\mathfrak{a}\mathfrak{b}_1} \overline{G}_{\mathfrak{b}_1\mathfrak{b}_2} \right) \\ &\quad - m \sum_{x,y} Q_x \left[\Theta_{\mathfrak{a}x} s_{xy} (G_{yy} - m) G_{x\mathfrak{b}_1} \overline{G}_{x\mathfrak{b}_2} \right] - m \sum_{x,y} Q_x \left[\Theta_{\mathfrak{a}x} s_{xy} \overline{G}_{xx} G_{y\mathfrak{b}_1} \overline{G}_{y\mathfrak{b}_2} \right]. \end{aligned} \tag{2.9}$$

Proof. Taking $f(G) \equiv 1$ in Lemma 2.4 and replacing x, y, y' by $\mathfrak{a}, \mathfrak{b}_1, \mathfrak{b}_2$, we immediately conclude (2.7). □

We will see in Section 4 that the third and fourth order T-expansions are already rather lengthy. For even higher order T-expansions, the number of terms will grow exponentially (actually there are about n^{Cn} many terms in the n-th order T-expansion). These terms have complicated structures and we will use graphical notations to represent them.

2.2 Graphical notations

Our goal is to expand the generalized T-variable $T_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ for $\mathfrak{a}, \mathfrak{b}_1, \mathfrak{b}_2 \in \mathbb{Z}_L^d$. We represent these three indices by special vertices

$$\mathfrak{a} \equiv \otimes, \quad \mathfrak{b}_1 \equiv \oplus, \quad \mathfrak{b}_2 \equiv \ominus, \tag{2.10}$$

in the graphs. In other words, we use $\mathfrak{a}, \mathfrak{b}_1, \mathfrak{b}_2$ in expressions, and draw them as $\otimes, \oplus, \ominus$ in the graphs. Now we first introduce the atomic graphs, and the concept of subgraphs.

Definition 2.6 (Atomic graphs)**.** *Given a standard oriented graph with vertices and edges, we assign the following structures and call the resulting graph an atomic graph.*

- **Atoms:** *We will call the vertices atoms (vs. molecules in Definition 3.4 below). Each graph has some external atoms and internal atoms. The external atoms represent external indices whose values are fixed, while internal atoms represent summation indices that will be summed over. In particular, each graph in the expansions of* $T_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *has the following external atoms: one* $\otimes$ *atom representing the* $\mathfrak{a}$ *index, one* $\oplus$ *atom representing the* $\mathfrak{b}_1$ *index, and one* $\ominus$ *atom representing the* $\mathfrak{b}_2$ *index (where some of them can be the same atom). By fixing the value of an internal atom, it will become an external atom; by summing over an external atom, it will become an internal atom.*

- **Regular weights:** *A regular weight on the atom* x *represents a* G_{xx} *or* $\overline{G}_{xx}$ *factor. Each regular weight has a charge, where "+" charge indicates that the weight is a* G *factor, represented by a blue solid* Δ*, and "−" charge indicates that the weight is a* $\overline{G}$ *factor, represented by a red solid* Δ*.*

- **Light weights**: *Corresponding to the regular weights defined above, we define the light weights representing* $G_{xx} - m$ *and* $\overline{G}_{xx} - \overline{m}$*. They are drawn as blue or red hollow* Δ *in graphs depending on their charges.*

- **Edges:** *The edges are divided into the following types.*

  (i) **Solid edges:** *A solid edge represents a* G *factor. More precisely,*

  – *each oriented edge from atom* α *to atom* β *with + charge represents a* $G_{\alpha\beta}$ *factor;*
  – *each oriented edge from atom* α *to atom* β *with − charge represents a* $\overline{G}_{\alpha\beta}$ *factor.*

The plus G edges will be drawn as blue solid edges, while minus G edges will be drawn as red solid edges. In this paper, whenever we say "G edges", we mean both the plus and minus G edges.

(ii) **Waved edges:** *We have neutral black, positive blue and negative red waved edges:*

- *a neutral waved edge between atoms x and y represents an s_{xy} factor;*
- *a blue waved edge of positive charge between atoms x and y represents a S^+_{xy} factor;*
- *a red waved edge of negative charge between atoms x and y represents a S^-_{xy} factor.*

(iii) **Diffusive edges:** *A diffusive edge connecting atoms x and y represents a Θ_{xy} factor; we draw it as a double-line edge between atoms x and y.*

(iv) **Dotted edges:** *A dotted line connecting atoms α and β represents the factor $\mathbf{1}_{\alpha=\beta} \equiv \delta_{\alpha\beta}$; a dotted line with a cross ($\times$) represents the factor $\mathbf{1}_{\alpha\neq\beta} \equiv 1-\delta_{\alpha\beta}$. There is at most one dotted or $\times$-dotted edge between each pair of atoms. By definition, a $\times$-dotted edge between the two ending atoms of a G edge indicates that this G edge is off-diagonal. We also allow for dotted edges between external atoms.*

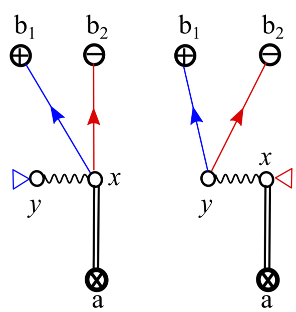
The orientations of non-solid edges do not matter. The edges between internal atoms are called internal edges*; the edges with at least one end at an external atom are called* external edges.

- P **and** Q **labels:** *Some solid edges and weights may have a label P_x or Q_x, where x is an atom in the graph. Moreover, each edge or weight can have at most one P or Q label.*

- **Coefficients:** *There is a coefficient (which is a polynomial of m, m^{-1}, $(1-m^2)^{-1}$ and their complex conjugates) associated with each graph.*

Definition 2.7 (Sugraphs)**.** *A graph $\mathcal{G}_1$ is said to be a subgraph of $\mathcal{G}_2$, denoted by $\mathcal{G}_1 \subset \mathcal{G}_2$, if every graphical component of $\mathcal{G}_1$ is also in $\mathcal{G}_2$. Moreover, $\mathcal{G}_1$ is a proper subgraph of $\mathcal{G}_2$ if $\mathcal{G}_1 \subsetneq \mathcal{G}_2$. Given a subset $\mathcal{S}$ of atoms in a graph $\mathcal{G}$, the subgraph induced on $\mathcal{S}$ refers to the subgraph of $\mathcal{G}$ with atoms in $\mathcal{S}$ as vertices, the edges between these atoms, and the weights on these atoms.*

Example 2.8. As an example, we draw the graphs for $\mathcal{A}_T^{(>2)}$ in (2.8):



(2.11)

For conciseness, we do not draw the coefficients of these graphs.

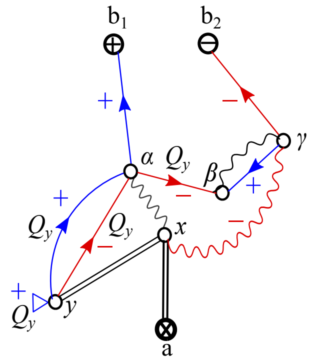
To each graph, we assign a value as follows.

Definition 2.9 (Values of graphs)**.** *For an atomic graph $\mathcal{G}$, we define its value, denoted by $[\![\mathcal{G}]\!]$, as an expression obtained as follows. We first take the product of all the edges, all the weights and the coefficient of $\mathcal{G}$. Then for the edges and weights with the same P_x or Q_x label, we group them together and apply P_x or Q_x to them. Finally, we sum over all the internal indices represented by the internal atoms. The values of the external indices are fixed by their given values. For a linear combination of graphs $\sum_i c_i\mathcal{G}_i$, where $\{c_i\}$ is a sequence of coefficients and $\{\mathcal{G}_i\}$ is a sequence of graphs, we define its value by*

$$\Big[\!\!\Big[\sum_i c_i \mathcal{G}_i \Big]\!\!\Big] = \sum_i c_i \, [\![\mathcal{G}_i]\!] \, .$$

For simplicity, we will abuse the notation by identifying a graph (which is a geometric object) with its value (which is an analytic expression).

Example 2.10. As an example of Definition 2.9, we write down the value for the following graph:



$$= \sum_{x,y,\alpha,\beta,\gamma} \Theta_{\mathfrak{a}x} s_{x\alpha} S^-_{x\gamma} s_{\beta\gamma} G_{\alpha\mathfrak{b}_1} G_{\gamma\beta} \overline{G}_{\gamma\mathfrak{b}_2} \Theta_{xy} Q_y \left[\overline{G}_{\alpha\beta} |G_{y\alpha}|^2 (G_{yy} - m) \right].$$

Next, we introduce the concept of *regular graphs*, which include (almost) all the graphs appearing in this paper, and a stronger concept of *normal regular graphs.*

Definition 2.11 (Normal regular graphs)**.** *We say an atomic graph $\mathcal{G}$ is* **regular** *if it satisfies the following properties:*

(i) it is a connected graph that contains at most $\mathrm{O}(1)$ *many atoms and edges;*

(ii) all the internal atoms are connected together through paths of waved and diffusive edges;

(iii) there are no dotted edges between internal atoms.

Moreover, we say a regular graph is **normal** *if it satisfies the following additional property:*

(iv) any pair of atoms α and β in the graph are connected by a $\times$-dotted edge if and only if *they are connected by a G edge.*

By this definition, every G edge in a normal regular graph is off-diagonal, while all the diagonal G factors will be represented by weights. There are two reasons for introducing the property (iv): (1) in Definition 2.12, we need to distinguish between the diagonal and off-diagonal G entries; (2) the weight expansion (cf. Definition 3.6 below) of diagonal G entries and the edge expansions (cf. Definitions 3.11, 3.15, 3.18 below) of off-diagonal G entries are very different in nature.

Definition 2.12 (Scaling order)**.** *Given a normal regular graph $\mathcal{G}$, we define its scaling order as*

$$\begin{aligned} ord(\mathcal{G}) := &\ \#\{\textit{off-diagonal } G \textit{ edges}\} + \#\{\textit{light weights}\} + 2\#\{\textit{waved edges}\} + 2\#\{\textit{diffusive edges}\} \\ & - 2\left[\#\{\textit{internal atoms}\} - \#\{\textit{dotted edges}\}\right]. \end{aligned} \tag{2.12}$$

Here each dotted edge in a normal regular graph means that an internal atom is equal to an external atom, so we lose one free summation index. The concept of scaling order can be also defined for subgraphs.

The motivation behind this definition is as follows. Consider the Wigner ensemble with $W = L$. By (1.8), (1.17) and (2.3), each waved edge is of order $\mathrm{O}(W^{-d})$ and each diffusive edge is of order $\mathrm{O}_\prec(W^{-d})$. Moreover, if we know that $|G_{xy} - m\delta_{xy}| \prec W^{-d/2}$, then each off-diagonal G edge or light weight is bounded by $\mathrm{O}_\prec(W^{-d/2})$. Finally, each summation leads to a factor W^d. Hence it is easy to obtain the bound

$$[\![\mathcal{G}]\!] \prec W^{-\mathrm{ord}(\mathcal{G})\cdot d/2}.$$

Later in Lemma 6.10, we will show that this bound holds even if $L \gg W$ as long as the graph satisfies the doubly connected property to be introduced in Section 6.2.

In the following proof whenever we say the order of a graph, we are referring to its scaling order. We emphasize that in general the scaling order does not imply the "order of the graph value" directly.

2.3 Self-energies

The T-expansion is defined using a collection of special sums of deterministic graphs, which satisfy some important properties given by Definition 2.13 below. Following the notations in Feynman diagrams, we call them "self-energies". As we explained before, the sum zero property (1.27) of the self-energies is one of the key reasons why we can define the T-expansion up to any order. In previous works [27, 65], the T-expansion can only be performed to third order without using the concept of self-energies.

Definition 2.13 (Self-energies)**.** *Under the assumptions of Theorem 2.1, for a fixed $l \in \mathbb{N}$, let $\mathcal{E}_l(z) \equiv \mathcal{E}_l^L(z)$ be a deterministic matrix depending on $m(z)$, S, $S^{\pm}(z)$ and $\Theta(z)$ only, and satisfying the following properties. (In this paper, we will often omit the dependence on L in $\mathcal{E}_l^L$.)*

(i) For any $x, y \in \mathbb{Z}_L^d$, $(\mathcal{E}_l)_{xy}$ is a sum of at most C_l many deterministic graphs of scaling order l and with external atoms x and y. Here C_l is a large constant depending on l. Some graphs, say $\mathcal{G}$, in $\mathcal{E}_l$ can be diagonal matrices satisfying $\mathcal{G}_{xy} = \mathcal{G}_{xx}\delta_{xy}$, i.e. there is a dotted edge between the atoms x and y.

(ii) $\mathcal{E}_l(z)$ satisfies the properties (1.25)–(1.27).

(iii) For any $x, y \in \mathbb{Z}^d$ and $z_L := E + \mathrm{i}\eta_L$ with $\eta_L \in [W^2/L^{2-\tau}, L^{-\tau}]$ for a small constant $\tau > 0$, we denote the infinite space limit (with W being fixed) of $(\mathcal{E}_l^L)_{xy}(z_L)$ by $(\mathcal{E}_l^\infty)_{xy}(E) := \lim_{L\to\infty}(\mathcal{E}_l^L)_{xy}(z_L)$, which is independent of L and with $\eta_\infty = 0$.

We call $\mathcal{E}_l$ the l-th order self-energy ($\mathcal{E}_l$ will be unique from our construction) and graphically we will use a square, $\square$, between atoms x and y with a label l to represent $(\mathcal{E}_l)_{xy}$.

We will show that the infinite space limits of the self-energies satisfy the following properties:

$$\mathcal{E}_l^\infty(x, x+a) = \mathcal{E}_l^\infty(0, a), \quad \mathcal{E}_l^\infty(0, a) = \mathcal{E}_l^\infty(0, -a), \quad \forall\, x, a \in \mathbb{Z}^d, \tag{2.13}$$

$$|(\mathcal{E}_l^\infty)_{0x}(E)| \leqslant W^{-ld/2}\frac{W^{2d-4}}{\langle x\rangle^{2d-4-\tau}}, \quad \forall\, x \in \mathbb{Z}^d, \tag{2.14}$$

$$\left|(\mathcal{E}_l^L)_{0x}(z) - (\mathcal{E}_l^\infty)_{0x}(E)\right| \leqslant W^{-ld/2}\frac{\eta W^{2d-6}}{\langle x\rangle^{2d-6-\tau}}, \quad \forall\, x \in \mathbb{Z}_L^d \subset \mathbb{Z}^d,\ \eta \in [W^2/L^{2-\tau}, L^{-\tau}], \tag{2.15}$$

$$\sum_{x\in\mathbb{Z}^d}(\mathcal{E}_l^\infty)_{0x}(E) = 0. \tag{2.16}$$

In (2.15), with slight abuse of notation, we identify the torus $\mathbb{Z}_L^d$ in (1.2) as a subset of $\mathbb{Z}^d$. The properties (2.13) and (2.14) take the same forms as the properties (1.25) and (1.26). The property (2.16) is an exact *sum zero property* and is thus stronger than (1.27). These properties will be proved in Lemma 5.8.

By Definition 2.12, the scaling order of a deterministic graph can only be even. Moreover, every nontrivial self-energy $\mathcal{E}_l$ used in this paper has scaling order $\geqslant 4$. Hence we always have

$$\mathcal{E}_1 = \mathcal{E}_2 = \mathcal{E}_3 = 0, \quad \text{and} \quad \mathcal{E}_{2l+1} := 0, \quad l \in \mathbb{N}. \tag{2.17}$$

By property (2.13), $\mathcal{E}_l^\infty$ is translationally invariant and symmetric (so are all the deterministic graphs in this paper by Lemma A.1). The properties (1.26) and (2.14) show that the rows of $\mathcal{E}_l$ or $\mathcal{E}_l^\infty$ are absolutely summable, i.e., for any constant $\tau > 0$,

$$\sum_x |(\mathcal{E}_l)_{0x}| \leqslant W^{-(l-2)d/2+\tau}, \quad \sum_x |(\mathcal{E}_l^\infty)_{0x}| \leqslant W^{-(l-2)d/2+\tau}. \tag{2.18}$$

The bound (1.27) is stronger than the first estimate in (2.18) by an extra η factor, which, as discussed in Section 1.4, is crucial for our proof.

The property (2.15) controls the difference between $(\mathcal{E}_l)_{0x}(z)$ and $(\mathcal{E}_l^\infty)_{0x}(E)$. The property (1.27) actually can be derived from (2.14), (2.15), and the sum zero property (2.16) for $\mathcal{E}_l^\infty$. More precisely,

$$\Big|\sum_{x\in\mathbb{Z}_L^d}(\mathcal{E}_l)_{0x}(z)\Big| = \Big|\sum_{x\in\mathbb{Z}_L^d}(\mathcal{E}_l)_{0x}(z) - \sum_{x\in\mathbb{Z}^d}(\mathcal{E}_l^\infty)_{0x}(E)\Big| \leqslant \sum_{|x|\leqslant L/2}|(\mathcal{E}_l)_{0x}(z) - (\mathcal{E}_l^\infty)_{0x}(E)| + \sum_{|x|>L/2}|(\mathcal{E}_l)_{0x}(z)|$$

$$\leqslant \sum_{|x|\leqslant L/2} W^{-ld/2}\frac{\eta W^{2d-6}}{\langle x\rangle^{2d-6-\tau}} + \sum_{|x|>L/2} W^{-ld/2}\frac{W^{2d-4}}{\langle x\rangle^{2d-4-\tau}}$$
$$\lesssim L^\tau\left(\eta + \frac{W^2}{L^2}\right)W^{-(l-2)d/2} \leqslant 2L^\tau \eta W^{-(l-2)d/2},$$

where in the first step we used (2.16), in the third step we used (2.14) and (2.15), and in the last step we used $\eta \gg W^2/L^2$.

The l-th order self-energy $\mathcal{E}_l$ in this paper is constructed through a specific expansion procedure of the T-variables. In general, if a different expansion procedure is used, a different l-th order self-energy may be obtained. Although we expect the self-energies constructed in different procedures to be the same up to negligible errors, this property is not needed in this paper and we will not pursue it.

2.4 Definition of the T-expansion

Given $n \in \mathbb{N}$, we will define the n-th order T-expansion in Definition 2.15, which is an extension of the second order T-expansion in (2.7). To this end, we first introduce the following two types of graphs.

Definition 2.14 (Recollision graphs and Q-graphs). *(i) We say a graph is a $\oplus/\ominus$-recollision graph, if there is at least one dotted edge connecting $\oplus$ or $\ominus$ to an internal atom. In other words, a recollision graph represents an expression where we set at least one summation index to be equal to $\mathfrak{b}_1$ or $\mathfrak{b}_2$.*

(ii) We say a graph is a Q-graph if all G edges and G weights in the graph have the same Q label with a specific atom x, i.e., all Q operators are given by the same Q_x.

We now define a general n-th order T-expansion for any fixed $n \in \mathbb{N}$. Besides the properties in Definition 2.15, the graphs in the definition satisfy several additional properties to be stated in Definition 6.6.

Definition 2.15 (n-th order T-expansion). *Fix any $n \in \mathbb{N}$ and let $D > n$ be an arbitrary large constant. For $\mathfrak{a}, \mathfrak{b}_1, \mathfrak{b}_2 \in \mathbb{Z}_L^d$, an n-th order T-expansion of $T_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ with D-th order error is an expression of the following form:*

$$\begin{aligned} T_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} &= m\Theta_{\mathfrak{a}\mathfrak{b}_1}\overline{G}_{\mathfrak{b}_1\mathfrak{b}_2} + m(\Theta\Sigma_T^{(n)}\Theta)_{\mathfrak{a}\mathfrak{b}_1}\overline{G}_{\mathfrak{b}_1\mathfrak{b}_2} \\ &\quad + (\mathcal{R}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + (\mathcal{A}_T^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + (\mathcal{Q}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + (\mathcal{E}rr_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}. \end{aligned} \tag{2.19}$$

The graphs on the right side depend only on n, D, $m(z)$, S, $S^{\pm}(z)$, $\Theta(z)$ and $G(z)$, but do not depend on W, L and d explicitly. Moreover, they satisfy the following properties with C_n and C_D denoting large constants depending on n and D, respectively.

(i) The graphs on the right side are normal regular graphs (recall Definition 2.11) with external atoms $\otimes \equiv \mathfrak{a}$, $\oplus \equiv \mathfrak{b}_1$ and $\ominus \equiv \mathfrak{b}_2$, and with at most C_D many atoms.

(ii) $\Sigma_T^{(n)}$ is a sum of at most C_n many deterministic normal regular graphs. We decompose it according to the scaling order as

$$\Sigma_T^{(n)} = \sum_{k\leqslant n} \Sigma_{T,k}. \tag{2.20}$$

Moreover, we have a sequence of self-energies $\mathcal{E}_k$ satisfying Definition 2.13 and properties (2.13)–(2.16) *for $4 \leqslant k \leqslant n$ such that $\Sigma_{T,k}$ can be written into the following form*

$$\Sigma_{T,k} = \mathcal{E}_k + \sum_{l=2}^{k} \sum_{\mathbf{k}=(k_1,\cdots,k_l)\in\Omega_k^{(l)}} \mathcal{E}_{k_1}\Theta\mathcal{E}_{k_2}\Theta\cdots\Theta\mathcal{E}_{k_l}. \tag{2.21}$$

Here all the deterministic graphs with $l = 1$ are included into $\mathcal{E}_k$ so that the summation starts with $l = 2$. Moreover, $\Omega_k^{(l)} \subset \mathbb{N}^l$ is the subset of vectors $\mathbf{k}$ satisfying that

$$4 \leqslant k_i \leqslant k-1, \quad \text{and} \quad \sum_{i=1}^{l} k_i - 2(l-1) = k. \tag{2.22}$$

The second condition in (2.22) *guarantees that the subgraph $(\mathcal{E}_{k_1}\Theta\mathcal{E}_{k_2}\Theta\cdots\Theta\mathcal{E}_{k_l})_{xy}$ has scaling order k.*

(iii) $(\mathcal{R}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *is a sum of at most* C_n *many* $\oplus/\ominus$*-recollision graphs of scaling order* $\leqslant n$ *and without any* P/Q *labels. Moreover, it can be decomposed as*

$$(\mathcal{R}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = \sum_{k=3}^{n} (\mathcal{R}_{T,k})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}, \tag{2.23}$$

where each $(\mathcal{R}_{T,k})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *is a sum of the* $\oplus/\ominus$*-recollision graphs of scaling order* k *in* $(\mathcal{R}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$.

(iv) $(\mathcal{A}_T^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *is a sum of at most* C_D *many graphs of scaling order* $> n$ *and without any* P/Q *labels.*

(v) $(\mathcal{Q}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *is a sum of at most* C_D *many* Q*-graphs. Moreover, it can be decomposed as*

$$(\mathcal{Q}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = \sum_{k=2}^{n} (\mathcal{Q}_{T,k})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + (\mathcal{Q}_T^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}, \tag{2.24}$$

where $(\mathcal{Q}_{T,k})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *is a sum of the scaling order* k Q*-graphs in* $(\mathcal{Q}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *and* $(\mathcal{Q}_T^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *is a sum of all the scaling order* $> n$ Q*-graphs in* $(\mathcal{Q}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$.

(vi) $\Sigma_{T,k}$, $\mathcal{R}_{T,k}$ *and* $\mathcal{Q}_{T,k}$ *are independent of* n.

(vii) $(\mathcal{E}rr_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *is a sum of at most* C_D *many graphs, each of which has scaling order* $> D$ *and may contain some* P/Q *labels in it.*

(viii) In each graph of $(\mathcal{R}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{A}_T^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{Q}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *and* $(\mathcal{E}rr_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$*, there is a unique diffusive edge connected to* $\otimes$*. Furthermore, there is at least an edge, which is either plus solid* G *or diffusive or dotted, connected to* $\oplus$*, and there is at least an edge, which is either minus solid* G *or diffusive or dotted, connected to* $\ominus$*.*

The graphs on the right-hand side of (2.19) *satisfy some additional properties, which will be given in Definition 6.6 below.*

In accordance with (2.17), we have that

$$\Sigma_{T,1} = \Sigma_{T,2} = \Sigma_{T,3} = \Sigma_{T,2l+1} = 0, \quad l \in \mathbb{N}.$$

With (1.45), we can bound $(\Theta\Sigma_{T,k}\Theta)_{\mathfrak{a}\mathfrak{b}_1}$ by

$$(\Theta\Sigma_{T,k}\Theta)_{\mathfrak{a}\mathfrak{b}_1} \prec W^{-(k-2)d/2} B_{\mathfrak{a}\mathfrak{b}_1}. \tag{2.25}$$

This bound shows that when $\mathfrak{b}_1 = \mathfrak{b}_2 = \mathfrak{b}$, the second term on the right-hand side of (2.19) can be bounded by $m\overline{G}_{\mathfrak{b}\mathfrak{b}}(\Theta\Sigma_T^{(n)}\Theta)_{\mathfrak{a}\mathfrak{b}} \prec B_{\mathfrak{a}\mathfrak{b}}$, which is necessary for (1.19) to hold. The rigorous proof of (2.25) will be given in Lemma 6.2. When $\mathfrak{b}_1 = \mathfrak{b}_2 = \mathfrak{b}$, by (2.8) the two graphs in $(\mathcal{R}_{T,3})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}$ are

$$m\sum_{x,y} \delta_{x\mathfrak{b}}\Theta_{\mathfrak{a}x}s_{xy}(G_{yy}-m)|G_{x\mathfrak{b}}|^2 + m\sum_{x,y}\delta_{y\mathfrak{b}}\Theta_{\mathfrak{a}x}s_{xy}(\overline{G}_{xx}-\overline{m})|G_{y\mathfrak{b}}|^2,$$

which can be easily bounded by $\Theta_{\mathfrak{a}\mathfrak{b}} \prec B_{\mathfrak{a}\mathfrak{b}}$. In general, there are many more complicated graphs in $(\mathcal{R}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, but they all satisfy good enough bounds for our prupose. If $D > 0$ is sufficiently large, the term $(\mathcal{E}rr_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ will be negligible for all proofs. If a graph in $(\mathcal{E}rr_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ does not contain any P/Q label, then it can be also included into $(\mathcal{A}_T^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$.

In Section 3, we will describe the basic graph operations that are used to obtain the T-expansion, and more details will be given in [64]. Assuming the n-th order T-expansion, we can prove Theorem 2.1.

Theorem 2.16. *Fix any* $n \in \mathbb{N}$*. Suppose the assumptions of Theorem 2.1 hold, and we have an* n*-th order* T*-expansion given in Definition 2.15 (together with the additional properties in Definition 6.6). Assume that* L *satisfies*

$$L^2/W^2 \leqslant W^{(n-1)d/2-c_0} \tag{2.26}$$

for some constant $c_0 > 0$*. Then for any constant* $\varepsilon > 0$*, the local law*

$$|G_{xy}(z) - m(z)\delta_{xy}|^2 \prec B_{xy} \tag{2.27}$$

holds uniformly in all $z = E + \mathrm{i}\eta$ *with* $E \in (-2+\kappa, 2-\kappa)$ *and* $\eta \in [W^2/L^{2-\varepsilon}, 1]$.

If we have obtained the n-th order T-expansion for $n = n_{W,L} := \left\lceil \frac{4}{d}\left(\log_W L - 1 + \frac{c_0}{2}\right)\right\rceil + 1$, then we can conclude Theorem 2.1 by using Theorem 2.16. The proof of Theorem 2.16 will be given Section 5.

2.5 Definition of the T-equation

In this subsection, we define the concept of T*-equation*.

Definition 2.17 (n-th order T-equation)**.** *Fix any $n \in \mathbb{N}$ and let $D > 0$ be an arbitrary large constant. For $\mathfrak{a}, \mathfrak{b}_1, \mathfrak{b}_2 \in \mathbb{Z}_L^d$, an n-th order T-equation of $T_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ with D-th order error is an expression of the following form:*

$$\begin{aligned} T_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = m\Theta_{\mathfrak{a}\mathfrak{b}_1}\overline{G}_{\mathfrak{b}_1\mathfrak{b}_2} + \sum_x (\Theta\Sigma^{(n)})_{\mathfrak{a}x} T_{x,\mathfrak{b}_1\mathfrak{b}_2} \\ + (\mathcal{R}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + (\mathcal{A}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + (\mathcal{Q}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + (\mathcal{E}rr'_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}, \end{aligned} \tag{2.28}$$

where the graphs on the right-hand side depend only on n, D, $m(z)$, S, $S^{\pm}(z)$, $\Theta(z)$ and $G(z)$, but do not depend on W, L and d explicitly. Moreover, they satisfy the following properties.

(i) $(\mathcal{R}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{A}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{Q}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *and* $(\mathcal{E}rr'_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *respectively satisfy the same properties as* $(\mathcal{R}_{T}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{A}_{T}^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{Q}_{T}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *and* $(\mathcal{E}rr_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *in Definition 2.15. Furthermore,* $(\mathcal{R}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *and* $(\mathcal{Q}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *can be decomposed as*

$$(\mathcal{R}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = \sum_{k=3}^{n} (\mathcal{R}_{IT,k})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}, \tag{2.29}$$

and

$$(\mathcal{Q}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = \sum_{k=2}^{n} (\mathcal{Q}_{IT,k})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + (\mathcal{Q}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}, \tag{2.30}$$

where $(\mathcal{R}_{IT,k})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *is a sum of the scaling order k $\oplus/\ominus$-recollision graphs in* $(\mathcal{R}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{Q}_{IT,k})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *is a sum of the scaling order k Q-graphs in* $(\mathcal{Q}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$*, and* $(\mathcal{Q}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *is a sum of the scaling order $> n$ Q-graphs in* $(\mathcal{Q}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$*. Moreover, $\mathcal{R}_{IT,k}$ and $\mathcal{Q}_{IT,k}$ are independent of n.*

(ii) $\Sigma^{(n)}$ *can be decomposed according to the scaling order as*

$$\Sigma^{(n)} = \mathcal{E}_n + \sum_{l=4}^{n-1} \mathcal{E}_l, \tag{2.31}$$

where $\mathcal{E}_l$, $1 \leqslant l \leqslant n-1$, is a sequence of self-energies satisfying Definition 2.13 and properties (2.13)–(2.16)*. For any $x, y \in \mathbb{Z}_L^d$, $(\mathcal{E}_n)_{xy}$ is a sum of at most C_n many deterministic graphs of scaling order n and with external atoms x and y.*

(iii) Each graph of $(\mathcal{R}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{A}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{Q}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *and* $(\mathcal{E}rr'_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ *can be written into*

$$\sum_x \Theta_{\mathfrak{a}x} \mathcal{G}_{x,\mathfrak{b}_1\mathfrak{b}_2},$$

where $\mathcal{G}_{x,\mathfrak{b}_1\mathfrak{b}_2}$ is a normal regular graph with external atoms x, $\mathfrak{b}_1$ and $\mathfrak{b}_2$. Moreover, $\mathcal{G}_{x,\mathfrak{b}_1\mathfrak{b}_2}$ has at least an edge, which is either plus solid G or diffusive or dotted, connected to $\oplus$, and at least an edge, which is either minus solid G or diffusive or dotted, connected to $\ominus$.

The graphs on the right side of (2.28) *satisfy some additional properties, which will be given in Definition 6.7 below.*

The form of (2.28) is different from (2.19) only in the second term on the right-hand side. We can regard (2.28) as a linear equation of the T-variable $T_{x,\mathfrak{b}_1\mathfrak{b}_2}$. In fact, taking $\mathfrak{b}_1 = \mathfrak{b}_2$ and $n = M$ in (2.28), if we move the second term on the right-hand side to the left-hand side, multiply both sides by $(1 - \Theta\Sigma^{(n)})^{-1}$ and take expectation, then we will get (1.22). More details of the proof will be given in Section 8.

The $\mathcal{E}_l$'s in (2.31) are the same self-energies as in Definition 2.15. We remark that the sequence of T-equations is constructed inductively. In particular, before constructing the n-th order T-equation, we have obtained the k-th order T-equation and proved the properties (1.25)–(1.27) and (2.13)–(2.16) for $\mathcal{E}_k$ for all $4 \leqslant k \leqslant n-1$. On the other hand, $\mathcal{E}_n$ is a new sum of deterministic graphs obtained in the n-th order T-equation, whose properties (1.25)–(1.27) and (2.13)–(2.16) are yet to be shown.

3 Basic graph operations

A graph operation $\mathcal{O}[\mathcal{G}]$ on a graph $\mathcal{G}$ is a linear combination of new graphs such that the graph value of $\mathcal{G}$ is unchanged, i.e. $[\![\mathcal{O}[\mathcal{G}]]\!] = [\![\mathcal{G}]\!]$. All graph operations are linear, that is,

$$\mathcal{O}\Big[\sum_i c_i \mathcal{G}_i\Big] = \sum_i c_i \mathcal{O}\left[\mathcal{G}_i\right]. \tag{3.1}$$

3.1 Dotted edge operations

Recall that a dotted edge between atoms α and β represents a $\delta_{\alpha\beta}$ factor. We will identify internal atoms connected by dotted edges, but we will *not* identify an external and an internal atom due to their different roles in graphs. Dotted edges between internal atoms may appear in intermediate steps, so we define the following merging operation.

Definition 3.1 (Merging operation)**.** *Given a graph $\mathcal{G}$ that contains dotted edges between different atoms, we define an operator $\mathcal{O}_{merge}$ in the following way: $\mathcal{O}_{merge}[\mathcal{G}]$ is a graph obtained by merging every pair of internal atoms, say α and β, that are connected by a path of dotted edges into a single internal atom, say γ. Moreover, the weights and edges attached to α and β are now attached to the atom γ in $\mathcal{O}_{merge}[\mathcal{G}]$. In particular, the G edges between α and β become weights on γ, and the waved and diffusive edges between α and β become self-loops on γ.*

It is easy to see that the graph operator $\mathcal{O}_{merge}$ is an identity in the sense of graph values: $[\![\mathcal{O}_{merge}[\mathcal{G}]]\!] = [\![\mathcal{G}]\!]$. Given any regular graph, we can rewrite it as a linear combination of normal regular graphs using the following *dotted edge partition* operation.

Definition 3.2 (Dotted edge partition)**.** *Given a regular graph $\mathcal{G}$, for any pair of atoms α and β, if there is at least one G edge but no $\times$-dotted edge between them, then we write*

$$1 = \mathbf{1}_{\alpha=\beta} + \mathbf{1}_{\alpha\neq\beta};$$

if there is a $\times$-dotted line $\mathbf{1}_{\alpha\neq\beta}$ but no G edge between them, then we write

$$\mathbf{1}_{\alpha\neq\beta} = 1 - \mathbf{1}_{\alpha=\beta}.$$

Expanding the product of all these sums on the right-hand sides, we can expand $\mathcal{G}$ as

$$\mathcal{O}_{dot}[\mathcal{G}] := \sum \mathcal{O}_{merge}[\mathbf{Dot} \cdot \mathcal{G}], \tag{3.2}$$

where each **Dot** *is a product of dotted and $\times$-dotted edges together with a $+$ or $-$ sign. In $\mathbf{Dot} \cdot \mathcal{G}$, if there is a $\times$-dotted edge between α and β, then the G edges between them are off-diagonal; otherwise, the G edges between them become weights after the merging operation. If* **Dot** *is "inconsistent" (i.e., two atoms are connected by a $\times$-dotted edge and a path of dotted edges), then we trivially have $[\![\mathbf{Dot} \cdot \mathcal{G}]\!] = 0$. Thus we will drop all inconsistent graphs. Finally, if the graph $\mathcal{G}$ is already normal, then $\mathcal{O}_{dot}$ acting on $\mathcal{G}$ is a null operation and we let $\mathcal{O}_{dot}[\mathcal{G}] := \mathcal{G}$.*

Lemma 3.3. *Given any regular graph $\mathcal{G}$, $\mathcal{O}_{dot}[\mathcal{G}]$ is a sum of normal regular graphs and $[\![\mathcal{O}_{dot}[\mathcal{G}]]\!] = [\![\mathcal{G}]\!]$.*

Lemma 3.3 trivially follows from Definition 3.2. We now introduce the concept of molecules and local expansions.

Definition 3.4 (Molecules)**.** *We partition the set of all atoms into a union of disjoint sets $\{$all atoms$\} = \cup_j \mathcal{M}_j$, where each $\mathcal{M}_j$ is called a molecule. Two internal atoms belong to the same molecule if and only if they are connected by a path of neutral/plus/minus waved edges and dotted edges (note there may be dotted edges between internal atoms if the graph is not regular). Each external atom will be called an external molecule (such as $\otimes$, $\oplus$ and $\ominus$ molecules) by definition. An edge is said to be inside a molecule if its ending atoms belong to this molecule.*

By (1.8) and (2.3), if two atoms x and y are in the same molecule, then we essentially have $|x-y| \leqslant W^{1+\tau}$ up to a negligible error $O(W^{-D})$. Given an atomic graph, we will call the subgraph inside a molecule (i.e. the subgraph induced on the atoms inside this molecule) the *local structure* of the molecule. The *molecular graph* (cf. Definition 6.4 below) is the quotient graph with each molecule regarded as a vertex. Then the *global structure* of a graph refers to its molecular graph. Note that the local structures can only vary on scales of order $O(W^{1+\tau})$, while the global structure varies on scales up to L. The *two-level structure* of an atomic graph—a global structure plus several local structures—has been explored in [65] already.

We will call an expansion *local* if it does not create new molecules, that is, every molecule in the new graphs is obtained by adding new atoms to existing molecules or merging some molecules in the original graph. It is easy to see that $\mathcal{O}_{dot}$ is a local expansion. In Sections 3.2 and 3.3, we will introduce more local expansions. We point out that local expansions can change the global structure. However, as we will explain in [64], they will maintain the doubly connected properties of the graphs (cf. Definition 6.5).

3.2 Weight expansion

Lemma 3.5. *In the setting of Theorem 1.3, suppose that f is a differentiable function of G. Then we have the following identity:*

$$(G_{xx}-m)f(G) = m\sum_{\alpha,\beta} b_{x\alpha}s_{\alpha\beta}(G_{\alpha\alpha}-m)(G_{\beta\beta}-m)f(G) - m\sum_{\alpha,\beta} b_{x\alpha}s_{\alpha\beta}P_\alpha\left[G_{\beta\alpha}\partial_{h_{\beta\alpha}}f(G)\right]$$
$$+\sum_{\alpha} b_{x\alpha}Q_\alpha\left[(G_{\alpha\alpha}-m)f(G)\right] - m\sum_{\alpha,\beta} b_{x\alpha}s_{\alpha\beta}Q_\alpha\left[(G_{\beta\beta}-m)G_{\alpha\alpha}f(G)\right], \tag{3.3}$$

where for simplicity we introduced the matrix

$$b := (1-m^2S)^{-1} = 1 + S^+. \tag{3.4}$$

Proof. Using (1.39) and Gaussian integration by parts, we obtain that

$$\begin{aligned}
&(G_{xx}-m)f(G) = Q_x\left[(G_{xx}-m)f(G)\right] + P_x\left[(G_{xx}-m)f(G)\right]\\
&= Q_x\left[(G_{xx}-m)f(G)\right] + P_x\Big[\Big(-m^2G_{xx} - m\sum_\alpha h_{x\alpha}G_{\alpha x}\Big)f(G)\Big]\\
&= Q_x\left[(G_{xx}-m)f(G)\right] + m\sum_\alpha s_{x\alpha}P_x\left[(G_{\alpha\alpha}-m)G_{xx}f(G)\right] - mP_x\sum_\alpha s_{x\alpha}\left[G_{\alpha x}\partial_{h_{\alpha x}}f(G)\right]\\
&= Q_x\left[(G_{xx}-m)f(G)\right] - m\sum_\alpha s_{x\alpha}Q_x\left[(G_{\alpha\alpha}-m)G_{xx}f(G)\right] + m^2\sum_\alpha s_{x\alpha}(G_{\alpha\alpha}-m)f(G)\\
&+ m\sum_\alpha s_{x\alpha}(G_{\alpha\alpha}-m)(G_{xx}-m)f(G) - m\sum_\alpha s_{x\alpha}P_x\left[G_{\alpha x}\partial_{h_{\alpha x}}f(G)\right],
\end{aligned}$$

which gives the equation

$$\begin{aligned}
\sum_\alpha(1-m^2S)_{x\alpha}(G_{\alpha\alpha}-m)f(G) &= Q_x\left[(G_{xx}-m)f(G)\right] - m\sum_\alpha s_{x\alpha}Q_x\left[(G_{\alpha\alpha}-m)G_{xx}f(G)\right]\\
&+m\sum_\alpha s_{x\alpha}(G_{\alpha\alpha}-m)(G_{xx}-m)f(G) - m\sum_\alpha s_{x\alpha}P_x\left[G_{\alpha x}\partial_{h_{\alpha x}}f(G)\right].
\end{aligned}$$

Multiplying both sides with $(1-m^2S)^{-1}$, we obtain (3.3). ☐

Expanding $b_{x\alpha}$ as $\delta_{x\alpha}+S^+_{x\alpha}$ and P_α as $1-Q_\alpha$ in (3.3), we obtain the following weight expansion operator.

Definition 3.6 (Weight expansion operator)**.** *Given a normal regular graph $\mathcal{G}$ which contains an atom x, if there is no weight on x, then we trivially define $\mathcal{O}^{(x)}_{weight}[\mathcal{G}] := \mathcal{G}$. Otherwise, we define $\mathcal{O}^{(x)}_{weight}[\mathcal{G}]$ in the following way.*

(i) Removing regular weights: *Suppose there are regular G_{xx} or $\overline{G}_{xx}$ weights on x. Then we rewrite*

$$G_{xx} = m + (G_{xx}-m), \quad \textit{and} \quad \overline{G}_{xx} = \overline{m} + (\overline{G}_{xx}-\overline{m}).$$

Expanding the product of all these sums, we can write $\mathcal{G}$ into a linear combination of normal regular graphs containing only light weights on x. We denote this graph operator as $\mathcal{O}^{(x),1}_{weight}$.

(ii) Expanding light weight: *If $\mathcal{G}$ has a light weight $G_{xx} - m$ of positive charge on x and is of the form $\mathcal{G} = (G_{xx} - m)f(G)$, then we define the light weight expansion on x by*

$$\begin{aligned}\mathcal{O}^{(x),2}_{weight}[\mathcal{G}] := m\sum_{\alpha} s_{x\alpha}(G_{xx}-m)(G_{\alpha\alpha}-m)f(G) + m\sum_{\alpha,\beta} S^+_{x\alpha}s_{\alpha\beta}(G_{\alpha\alpha}-m)(G_{\beta\beta}-m)f(G)\\ - m\sum_{\alpha} s_{x\alpha}G_{\alpha x}\partial_{h_{\alpha x}}f(G) - m\sum_{\alpha,\beta} S^+_{x\alpha}s_{\alpha\beta}G_{\beta\alpha}\partial_{h_{\beta\alpha}}f(G) + \mathcal{Q}_w,\end{aligned} \tag{3.5}$$

where $\mathcal{Q}_w$ is a sum of Q-graphs,

$$\begin{aligned}\mathcal{Q}_w := Q_x\left[(G_{xx}-m)f(G)\right] + \sum_{\alpha} Q_\alpha\left[S^+_{x\alpha}(G_{\alpha\alpha}-m)f(G)\right]\\ - mQ_x\Big[\sum_{\alpha} s_{x\alpha}(G_{\alpha\alpha}-m)G_{xx}f(G)\Big] - m\sum_{\alpha} Q_\alpha\Big[\sum_{\beta} S^+_{x\alpha}s_{\alpha\beta}(G_{\beta\beta}-m)G_{\alpha\alpha}f(G)\Big]\\ + mQ_x\Big[\sum_{\alpha} s_{x\alpha}G_{\alpha x}\partial_{h_{\alpha x}}f(G)\Big] + m\sum_{\alpha} Q_\alpha\Big[\sum_{\beta} S^+_{x\alpha}s_{\alpha\beta}G_{\beta\alpha}\partial_{h_{\beta\alpha}}f(G)\Big].\end{aligned}$$

If $\mathcal{G} = (\overline{G}_{xx} - \overline{m})f(G)$, then we define

$$\mathcal{O}^{(x),2}_{weight}[\mathcal{G}] := \overline{\mathcal{O}^{(x),2}_{weight}\left[(G_{xx}-m)\overline{f(G)}\right]}, \tag{3.6}$$

where the right-hand side can be defined using (3.5). When there are more than one light weights on x, we pick any positive light weight and apply (3.5); if there is no positive light weight, then we pick any negative light weight and apply (3.6).

Combining the above two graph operators, given any normal regular graph $\mathcal{G}$, we define

$$\mathcal{O}^{(x)}_{weight}[\mathcal{G}] := \mathcal{O}^{(x),1}_{weight} \circ \mathcal{O}_{dot} \circ \mathcal{O}^{(x),2}_{weight} \circ \mathcal{O}^{(x),1}_{weight}[\mathcal{G}], \tag{3.7}$$

where the operator $\mathcal{O}_{dot}$ is applied to make sure that the resulting graphs after applying $\mathcal{O}_{dot}$ are normal regular. The reason for the last $\mathcal{O}^{(1)}_{weight}(x)$ operator will be explained in Remark 3.7 below.

Remark 3.7. Consider the third term on the right-hand side of (3.5) as an example. First, when applying $\mathcal{O}_{dot}$, we will have a graph with $\alpha = x$, in which case $G_{\alpha x}$ becomes a weight G_{xx}. Second, we consider the partial derivative $\partial_{h_{\alpha x}} f(G)$. Suppose $f(G)$ is of the form

$$f(G) = \sum_{\{y_i\},\{y'_i\},\{w_i\},\{w'_i\}} \prod_{i=1}^{k_1} G_{xy_i} \cdot \prod_{i=1}^{k_2} \overline{G}_{xy'_i} \cdot \prod_{i=1}^{k_3} G_{w_i x} \cdot \prod_{i=1}^{k_4} \overline{G}_{w'_i x} \cdot G^{l_1}_{xx}\overline{G}^{l_2}_{xx}(G_{xx}-m)^{l_3}(\overline{G}_{xx}-\overline{m})^{l_4} g(G),$$

where $g(G) \equiv g(G, \{y_i\}, \{y'_i\}, \{w_i\}, \{w'_i\})$ does not contain any weight or solid edge attached to atom x. Then we take the partial derivative of the weights and solid edges in $f(G)$ using the identities

$$\partial_{h_{\alpha x}} G_{ab} = -G_{a\alpha}G_{xb}, \quad \partial_{h_{\alpha x}} \overline{G}_{ba} = -\overline{G}_{bx}\overline{G}_{\alpha a}, \quad a, b \in \mathbb{Z}^d_L. \tag{3.8}$$

Note that it is possible to have $b = x$ (e.g. when we take the partial derivative $\partial_{h_{\alpha x}}$ of $\overline{G}_{xy'_i}$, $G_{w_i x}$ or a weight on x), which will lead to a weight G_{xx} or $\overline{G}_{xx}$ on atom x. Hence we can have regular weights in the graphs in $\mathcal{O}_{dot} \circ \mathcal{O}^{(x),2}_{weight} \circ \mathcal{O}^{(x),1}_{weight}[\mathcal{G}]$. These regular weights are removed by applying another $\mathcal{O}^{(x),1}_{weight}$.

Definition 3.8 (Canonical local expansions)**.** *A local expansion $\mathcal{O}^{(x)}$ of a normal regular graph $\mathcal{G}$ at an atom x is said to be* canonical *if it satisfies the following properties.*

(i) The graph value is unchanged after the expansion, i.e., $[\![\mathcal{O}^{(x)}[\mathcal{G}]]\!] = [\![\mathcal{G}]\!]$.

(ii) $\mathcal{O}^{(x)}[\mathcal{G}]$ is a linear combination of normal regular graphs.

(iii) *Every graph in* $\mathcal{O}^{(x)}[\mathcal{G}]$ *has scaling order* $\geqslant \mathrm{ord}(\mathcal{G})$.

(iv) *If there is a new atom in a graph after the expansion, then it is connected to* x *through a path of waved edges.*

The property (iv) shows that all the new atoms created in the expansions are included in the existing molecule containing atom x and hence is consistent with the local property of $\mathcal{O}^{(x)}$.

Lemma 3.9. *Given a normal regular graph* $\mathcal{G}$ *with an atom* x, $\mathcal{O}^{(x)}_{weight}[\mathcal{G}]$ *is a canonical local expansion. If* $\mathcal{G}$ *contains at least one weight at* x, *then every graph without* Q*-labels, say* $\mathcal{G}_1$, *in* $\mathcal{O}^{(x)}_{weight}[\mathcal{G}]$ *satisfies one of the following two properties:*

(a) *its scaling order is strictly higher than* $\mathrm{ord}(\mathcal{G})$, *i.e.,* $\mathrm{ord}(\mathcal{G}_1) \geqslant \mathrm{ord}(\mathcal{G}) + 1$;

(b) $\mathrm{ord}(\mathcal{G}_1) = \mathrm{ord}(\mathcal{G})$, *and* $\mathcal{G}_1$ *has strictly fewer weights than* $\mathcal{G}$ *(more precisely, it contains at least one fewer weight on* x*, no weights on the new atoms, and the same number of weights on any other atom).*

The proof of Lemma 3.9 follows straightforwardly by using Definition 3.6 and we postpone it to Appendix C. The properties (a) and (b) in Lemma 3.9 show that, by applying the weight expansion repeatedly, we can get either new graphs without weights, or Q-graphs and graphs of sufficiently high scaling orders.

3.3 Edge expansions

In this subsection, we introduce three basic edge expansion operators. First, we define a multi-edge expansion, which aims to remove atoms that have degrees larger than 2. Here we use the following notion of degrees of solid edges (i.e. plus and minus G edges):

$$\deg(x) := \#\{\text{solid edges connected with } x\}. \tag{3.9}$$

Lemma 3.10. *In the setting of Theorem 1.3, suppose that* f *is a differentiable function of* G. *Consider a graph*

$$\mathcal{G} := \prod_{i=1}^{k_1} G_{xy_i} \cdot \prod_{i=1}^{k_2} \overline{G}_{xy_i'} \cdot \prod_{i=1}^{k_3} G_{w_i x} \cdot \prod_{i=1}^{k_4} \overline{G}_{w_i' x} \cdot f(G), \tag{3.10}$$

where the atoms y_i, y_i', w_i *and* w_i' *are all not equal to* x. *If* $k_1 \geqslant 1$, *then we have the following identity:*

$$\begin{aligned}\mathcal{G} = &\sum_{i=1}^{k_2} mP_x\left[\overline{G}_{xx}\left(\sum_\alpha s_{x\alpha}G_{\alpha y_1}\overline{G}_{\alpha y_i'}\right)\frac{\mathcal{G}}{G_{xy_1}\overline{G}_{xy_i'}}\right] + \sum_{i=1}^{k_3} mP_x\left[G_{xx}\left(\sum_\alpha s_{x\alpha}G_{\alpha y_1}G_{w_i\alpha}\right)\frac{\mathcal{G}}{G_{xy_1}G_{w_i x}}\right]\\ &+ mP_x\left[\sum_\alpha s_{x\alpha}\left(G_{\alpha\alpha}-m\right)\mathcal{G}\right] + (k_1-1)mP_x\left[\sum_\alpha s_{x\alpha}G_{\alpha y_1}G_{x\alpha}\frac{\mathcal{G}}{G_{xy_1}}\right] + k_4 mP_x\left[\sum_\alpha s_{x\alpha}G_{\alpha y_1}\overline{G}_{\alpha x}\frac{\mathcal{G}}{G_{xy_1}}\right]\\ &- mP_x\left[\sum_\alpha s_{x\alpha}\frac{\mathcal{G}}{G_{xy_1}f(G)}G_{\alpha y_1}\partial_{h_{\alpha x}}f(G)\right] + Q_x\left(\mathcal{G}\right).\end{aligned} \tag{3.11}$$

Here the fractions are used to simplify the expression. For example, the fraction $\mathcal{G}/(G_{xy_1}\overline{G}_{xy_i'})$ *is the graph obtained by removing the factor* $G_{xy_1}\overline{G}_{xy_i'}$ *from the product in* (3.10).

Proof. Using (1.39) and $x \neq y_1$, we can write that

$$P_x\left(\mathcal{G}\right) = P_x\left[\left(-m^2 G_{xy_1} - m\sum_\alpha h_{x\alpha}G_{\alpha y_1}\right)\prod_{i=2}^{k_1} G_{xy_i} \cdot \prod_{i=1}^{k_2} \overline{G}_{xy_i'} \cdot \prod_{i=1}^{k_3} G_{w_i x} \cdot \prod_{i=1}^{k_4} \overline{G}_{w_i' x} \cdot f(G)\right]. \tag{3.12}$$

We apply Gaussian integration by parts to the HG term to get that

$$-mP_x\left[\sum_\alpha h_{x\alpha}G_{\alpha y_1} \cdot \prod_{i=2}^{k_1} G_{xy_i} \cdot \prod_{i=1}^{k_2} \overline{G}_{xy_i'} \cdot \prod_{i=1}^{k_3} G_{w_i x} \cdot \prod_{i=1}^{k_4} \overline{G}_{w_i' x} \cdot f(G)\right]$$

$$
\begin{aligned}
&= mP_x\left[\left(\sum_\alpha s_{x\alpha}G_{\alpha\alpha}\right)\mathcal{G}\right] + (k_1-1)mP_x\left[\sum_\alpha s_{x\alpha}G_{\alpha y_1}G_{x\alpha}\frac{\mathcal{G}}{G_{xy_1}}\right] + k_4 mP_x\left[\sum_\alpha s_{x\alpha}G_{\alpha y_1}\overline{G}_{\alpha x}\frac{\mathcal{G}}{G_{xy_1}}\right]\\
&+\sum_{i=1}^{k_2} mP_x\left[\overline{G}_{xx}\left(\sum_\alpha s_{x\alpha}G_{\alpha y_1}\overline{G}_{\alpha y_i'}\right)\frac{\mathcal{G}}{G_{xy_1}\overline{G}_{xy_i'}}\right] + \sum_{i=1}^{k_3} mP_x\left[G_{xx}\left(\sum_\alpha s_{x\alpha}G_{\alpha y_1}G_{w_i\alpha}\right)\frac{\mathcal{G}}{G_{xy_1}G_{w_ix}}\right]\\
&- mP_x\left[\sum_\alpha s_{x\alpha}G_{\alpha y_1}\cdot\prod_{i=2}^{k_1}G_{xy_i}\cdot\prod_{i=1}^{k_2}\overline{G}_{xy_i'}\cdot\prod_{i=1}^{k_3}G_{w_ix}\cdot\prod_{i=1}^{k_4}\overline{G}_{w_i'x}\cdot\partial_{h_{\alpha x}}f(G)\right].
\end{aligned}
$$

Plugging it into (3.12) and using $\mathcal{G} = P_x(\mathcal{G}) + Q_x(\mathcal{G})$, we conclude (3.11). □

Applying $P_x = 1 - Q_x$ to (3.11), we can define the following multi-edge expansion operator.

Definition 3.11 (Multi-edge expansion operator)**.** *Given a normal regular graph $\mathcal{G}$, if there are no solid edges connected with an atom x, then we trivially define $\mathcal{O}^{(x)}_{multi-e}[\mathcal{G}] := \mathcal{G}$. Otherwise, we define $\mathcal{O}^{(x)}_{multi-e}$ in the following way. Suppose $\mathcal{G}$ takes the form (3.10), where the atoms y_i, y_i', w_i and w_i' are all not equal to x.*

(i) If $k_1 \geqslant 1$, then we define the multi-edge expansion on x as

$$
\begin{aligned}
\widehat{\mathcal{O}}^{(x)}_{multi-e}[\mathcal{G}] &:= \sum_{i=1}^{k_2}|m|^2\left(\sum_\alpha s_{x\alpha}G_{\alpha y_1}\overline{G}_{\alpha y_i'}\right)\frac{\mathcal{G}}{G_{xy_1}\overline{G}_{xy_i'}} + \sum_{i=1}^{k_3}m^2\left(\sum_\alpha s_{x\alpha}G_{\alpha y_1}G_{w_i\alpha}\right)\frac{\mathcal{G}}{G_{xy_1}G_{w_ix}}\\
&+\sum_{i=1}^{k_2}m(\overline{G}_{xx}-\overline{m})\left(\sum_\alpha s_{x\alpha}G_{\alpha y_1}\overline{G}_{\alpha y_i'}\right)\frac{\mathcal{G}}{G_{xy_1}\overline{G}_{xy_i'}} + \sum_{i=1}^{k_3}m(G_{xx}-m)\left(\sum_\alpha s_{x\alpha}G_{\alpha y_1}G_{w_i\alpha}\right)\frac{\mathcal{G}}{G_{xy_1}G_{w_ix}}\\
&+ m\sum_\alpha s_{x\alpha}\left(G_{\alpha\alpha}-m\right)\mathcal{G} + (k_1-1)m\sum_\alpha s_{x\alpha}G_{x\alpha}G_{\alpha y_1}\frac{\mathcal{G}}{G_{xy_1}} + k_4 m\sum_\alpha s_{x\alpha}\overline{G}_{\alpha x}G_{\alpha y_1}\frac{\mathcal{G}}{G_{xy_1}}\\
&- m\sum_\alpha s_{x\alpha}\frac{\mathcal{G}}{G_{xy_1}f(G)}G_{\alpha y_1}\partial_{h_{\alpha x}}f(G) + \mathcal{Q}_{multi-e}.
\end{aligned}\tag{3.13}
$$

On the right-hand side of (3.13)*, the first two terms are main terms with the same scaling order as $\mathcal{G}$, but the degree of atom x is reduced by 2 and a new atom α with degree 2 is created; the third to fifth terms contain one more light weight and hence are of strictly higher scaling orders than $\mathcal{G}$; the sixth to eighth terms contain at least one more off-diagonal G edge and hence are of strictly higher scaling orders than $\mathcal{G}$. The last term $\mathcal{Q}_{multi-e}$ is a sum of Q-graphs defined by*

$$
\begin{aligned}
\mathcal{Q}_{multi-e} := Q_x(\mathcal{G}) &- \sum_{i=1}^{k_2} mQ_x\left[\overline{G}_{xx}\left(\sum_\alpha s_{x\alpha}G_{\alpha y_1}\overline{G}_{\alpha y_i'}\right)\frac{\mathcal{G}}{G_{xy_1}\overline{G}_{xy_i'}}\right]\\
&- \sum_{i=1}^{k_3} mQ_x\left[G_{xx}\left(\sum_\alpha s_{x\alpha}G_{\alpha y_1}G_{w_i\alpha}\right)\frac{\mathcal{G}}{G_{xy_1}G_{w_ix}}\right] - mQ_x\left[\sum_\alpha s_{x\alpha}\left(G_{\alpha\alpha}-m\right)\mathcal{G}\right]\\
&- (k_1-1)mQ_x\left[\sum_\alpha s_{x\alpha}G_{x\alpha}G_{\alpha y_1}\frac{\mathcal{G}}{G_{xy_1}}\right] - k_4 mQ_x\left[\sum_\alpha s_{x\alpha}\overline{G}_{\alpha x}G_{\alpha y_1}\frac{\mathcal{G}}{G_{xy_1}}\right]\\
&+ mQ_x\left[\sum_\alpha s_{x\alpha}\frac{\mathcal{G}}{G_{xy_1}f(G)}G_{\alpha y_1}\partial_{h_{\alpha x}}f(G)\right].
\end{aligned}
$$

(ii) If $k_1 = 0$ and $k_2 \geqslant 1$, then we define

$$
\widehat{\mathcal{O}}^{(x)}_{multi-e}[\mathcal{G}] := \overline{\widehat{\mathcal{O}}^{(x)}_{multi-e}\left[\prod_{i=1}^{k_2}G_{xy_i'}\cdot\prod_{i=1}^{k_3}\overline{G}_{w_ix}\cdot\prod_{i=1}^{k_4}G_{w_i'x}\cdot\overline{f(G)}\right]},
$$

where the right-hand side can be defined using (i).

(iii) If $k_1 = k_2 = 0$ and $k_3 \geqslant 1$, then we define $\widehat{\mathcal{O}}^{(x)}_{multi-e}[\mathcal{G}]$ by exchanging the order of matrix indices in (i). More precisely, we define

$$\begin{aligned}\widehat{\mathcal{O}}^{(x)}_{multi-e}[\mathcal{G}] := &\sum_{i=1}^{k_4} |m|^2 \left(\sum_\alpha s_{x\alpha} G_{w_1\alpha}\overline{G}_{w_i'\alpha}\right) \frac{\mathcal{G}}{G_{w_1x}\overline{G}_{w_i'x}} + \sum_{i=1}^{k_4} m(\overline{G}_{xx} - \overline{m}) \left(\sum_\alpha s_{x\alpha} G_{w_1\alpha}\overline{G}_{w_i'\alpha}\right) \frac{\mathcal{G}}{G_{w_1x}\overline{G}_{w_i'x}} \\ &+ m\sum_\alpha s_{x\alpha}\left(G_{\alpha\alpha} - m\right)\mathcal{G} + (k_3-1)m\sum_\alpha s_{x\alpha} G_{w_1\alpha}G_{\alpha x}\frac{\mathcal{G}}{G_{w_1x}} \\ &- m\sum_\alpha s_{x\alpha}\frac{\mathcal{G}}{G_{w_1x}f(G)} G_{w_1\alpha}\partial_{h_{x\alpha}} f(G) + \mathcal{Q}_{multi-e},\end{aligned} \tag{3.14}$$

where

$$\begin{aligned}\mathcal{Q}_{multi-e} := &\, Q_x(\mathcal{G}) - \sum_{i=1}^{k_4} mQ_x\left[\overline{G}_{xx}\left(\sum_\alpha s_{x\alpha}G_{w_1\alpha}\overline{G}_{w_i'\alpha}\right)\frac{\mathcal{G}}{G_{w_1x}\overline{G}_{w_i'x}}\right] - mQ_x\left[\sum_\alpha s_{x\alpha}\left(G_{\alpha\alpha}-m\right)\mathcal{G}\right] \\ &- (k_3-1)mQ_x\left[\sum_\alpha s_{x\alpha}G_{w_1\alpha}G_{\alpha x}\frac{\mathcal{G}}{G_{w_1x}}\right] + mQ_x\left[\sum_\alpha s_{x\alpha}\frac{\mathcal{G}}{G_{w_1x}f(G)}G_{w_1\alpha}\partial_{h_{x\alpha}}f(G)\right].\end{aligned}$$

(iv) If $k_1 = k_2 = k_3 = 0$ and $k_4 \geqslant 1$, then we define

$$\widehat{\mathcal{O}}^{(x)}_{multi-e}[\mathcal{G}] := \overline{\widehat{\mathcal{O}}^{(x)}_{multi-e}\left[\prod_{i=1}^{k_4} G_{w_i'x}\cdot\overline{f(G)}\right]},$$

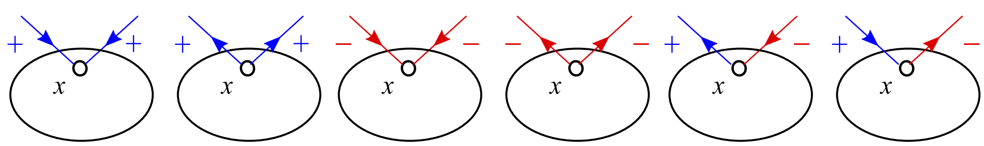
where the right-hand side can be defined using (iii).

Finally, applying the $\mathcal{O}_{dot}$ in Definition 3.2, we define

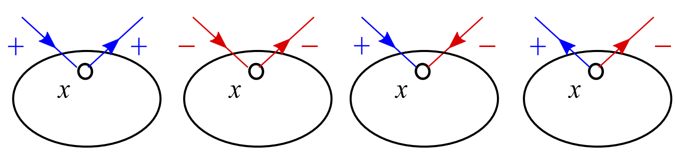
$$\mathcal{O}^{(x)}_{multi-e}[\mathcal{G}] := \mathcal{O}_{dot} \circ \widehat{\mathcal{O}}^{(x)}_{multi-e}[\mathcal{G}].$$

The multi-edge expansion motivates the following definition of *matched* and *mismatched* solid edges.

Definition 3.12 (Matched and mismatched edges). *Consider an internal atom x of degree 2 in a graph. We say the two edges connected with x are* **mismatched** *if they are of the following forms:*



Otherwise the two edges are **matched** *and of the following forms:*



Alternatively, a degree 2 atom x is said to be connected with two matched edges if and only if its charge is 0, where the charge of an atom is defined by

$$\#\{incoming + \text{ and } outgoing - \text{ solid edges}\} - \#\{outgoing + \text{ and } incoming - \text{ solid edges}\}.$$

By Definition 3.11, we can see that if x is connected with two mismatched edges, then $\mathcal{O}_{multi-e}(x)[\mathcal{G}]$ is a sum of graphs that are all of strictly higher scaling orders than $\mathcal{G}$. For example, we take $\mathcal{G} = G_{xy}G_{xy'}f(G)$, i.e., $k_1 = 2$ and $k_2 = k_3 = k_4 = 0$ in (3.10). Then the first two main terms on the right-hand side of (3.13) are both zero.

The following lemma describes the basic properties of multi-edge expansions. Its proof is a straightforward application of Definition 3.11, and we postpone it to Appendix C.

Lemma 3.13. *Consider a normal regular graph $\mathcal{G}$ taking the form* (3.10)*, where $f(G)$ does not contain any G edges or weights attached to x, and the atoms y_i, y_i', w_i and w_i' are all not equal to x. Then $\mathcal{O}^{(x)}_{multi-e}[\mathcal{G}]$ is a canonical local expansion satisfying the following properties.*

(a) Suppose that $\deg(x) \geqslant 4$ in $\mathcal{G}$. Then every graph without Q-labels, say $\mathcal{G}_1$, in $\mathcal{O}^{(x)}_{multi-e}[\mathcal{G}]$ either has a strictly higher scaling order than $\mathcal{G}$ or satisfies one of the following two properties:

(a.1) $\mathrm{ord}(\mathcal{G}_1) = \mathrm{ord}(\mathcal{G})$; $\mathcal{G}_1$ has one new atom with degree 2; $\deg(x)$ in $\mathcal{G}_1$ is smaller than $\deg(x)$ in $\mathcal{G}$ by 2, and the degree of any other atom stays the same as in $\mathcal{G}$;

(a.2) $\mathrm{ord}(\mathcal{G}_1) = \mathrm{ord}(\mathcal{G})$; $\mathcal{G}_1$ has no new atom; $\deg(x)$ in $\mathcal{G}_1$ is smaller than $\deg(x)$ in $\mathcal{G}$ by 2, and the degree of any other atom either stays the same or decreases by 2.

(b) Suppose that $\deg(x) = 1$, or x is connected with exactly two mismatched solid edges in $\mathcal{G}$. Then every graph without Q-labels has a strictly higher scaling order than $\mathcal{G}$.

Lemma 3.13 shows that, by applying the multi-edge expansion repeatedly, we can either make all atoms in the resulting graphs to be connected with *exactly two matched solid edges*, or get Q-graphs and graphs of sufficiently high scaling orders.

If an atom is connected with exactly two matched solid edges, then applying the multi-edge expansion cannot improve the graph anymore. Instead, we will apply the GG expansion given by the following lemma if these two edges are of the same charge.

Lemma 3.14. *In the setting of Theorem 1.3, consider a graph $\mathcal{G} = G_{xy}G_{y'x}f(G)$ where f is a differentiable function of G and $y, y' \neq x$. Then we have that*

$$\begin{aligned}\mathcal{G} = & \, m b_{xy} P_y \left[G_{y'y} f(G)\right] + m \sum_{\alpha,\beta} b_{x\alpha} s_{\alpha\beta} P_\alpha \left[(G_{\beta\beta} - m) G_{\alpha y} G_{y'\alpha} f(G)\right] \\ & + m \sum_{\alpha,\beta} b_{x\alpha} s_{\alpha\beta} P_\alpha \left[(G_{\alpha\alpha} - m) G_{\beta y} G_{y'\beta} f(G)\right] - m \sum_{\alpha,\beta} b_{x\alpha} s_{\alpha\beta} P_\alpha \left[G_{\beta y} G_{y'\alpha} \partial_{h_{\beta\alpha}} f(G)\right] \\ & + \sum_\alpha b_{x\alpha} Q_\alpha \left[G_{\alpha y} G_{y'\alpha} f(G)\right] - m^2 \sum_{\alpha,\beta} b_{x\alpha} s_{\alpha\beta} Q_\alpha \left[G_{\beta y} G_{y'\beta} f(G)\right],\end{aligned} \tag{3.15}$$

where b is defined in (3.4).

Proof. Using (1.39) and $I_N = b - m^2 bS$, we get that

$$\begin{aligned}P_x(\mathcal{G}) &= \sum_\alpha \left[b_{x\alpha} - m^2 (bS)_{x\alpha}\right] P_\alpha \left[G_{\alpha y} G_{y'\alpha} f(G)\right] \\ &= -\sum_\alpha m^2 (bS)_{x\alpha} P_\alpha \left[G_{\alpha y} G_{y'\alpha} f(G)\right] + \sum_\alpha b_{x\alpha} P_\alpha \left[\left(m\delta_{\alpha y} - m^2 G_{\alpha y} - m(HG)_{\alpha y}\right) G_{y'\alpha} f(G)\right].\end{aligned} \tag{3.16}$$

Applying Gaussian integration by parts to the HG term, we get that

$$\begin{aligned}& P_\alpha \Big[-m \sum_\beta h_{\alpha\beta} G_{\beta y} G_{y'\alpha} f(G)\Big] \\ =& P_\alpha \Big[m \sum_\beta s_{\alpha\beta} G_{\beta\beta} G_{\alpha y} G_{y'\alpha} f(G) + m \sum_\beta s_{\alpha\beta} G_{\alpha\alpha} G_{\beta y} G_{y'\beta} f(G) - m \sum_\beta s_{\alpha\beta} G_{\beta y} G_{y'\alpha} \partial_{h_{\beta\alpha}} f(G)\Big].\end{aligned}$$

Plugging it into (3.16), using $1 = P_x + Q_x$, $\delta_{x\alpha} + m^2(bS)_{x\alpha} = b_{x\alpha}$ and

$$\begin{aligned}& m \sum_{\alpha,\beta} b_{x\alpha} s_{\alpha\beta} P_\alpha \left[G_{\alpha\alpha} G_{\beta y} G_{y'\beta} f(G)\right] - \sum_\alpha m^2 (bS)_{x\alpha} P_\alpha \left[G_{\alpha y} G_{y'\alpha} f(G)\right] \\ =& \, m \sum_{\alpha,\beta} b_{x\alpha} s_{\alpha\beta} P_\alpha \left[(G_{\alpha\alpha} - m) G_{\beta y} G_{y'\beta} f(G)\right] + \sum_\alpha m^2 (bS)_{x\alpha} Q_\alpha \left[G_{\alpha y} G_{y'\alpha} f(G)\right] \\ & - m^2 \sum_{\alpha,\beta} b_{x\alpha} s_{\alpha\beta} Q_\alpha \left[G_{\beta y} G_{y'\beta} f(G)\right],\end{aligned}$$

we obtain equation (3.15). ∎

Using (3.15), $b = 1 + S^+$ and $P_\alpha = 1 - Q_\alpha$, we can define the following GG expansion operator.

Definition 3.15 (GG expansion operator)**.** *Given a normal regular graph $\mathcal{G}$, suppose an atom x is connected with exactly two matched G edges of the same charge. Suppose $\mathcal{G}$ takes the form $\mathcal{G} = G_{xy}G_{y'x}f(G)$ with $y, y' \neq x$. Then we define*

$$\begin{aligned}\widehat{\mathcal{O}}^{(x)}_{GG}[\mathcal{G}] := {} & mS^+_{xy}G_{y'y}f(G) + m\sum_\alpha s_{x\alpha}(G_{\alpha\alpha} - m)\mathcal{G} + m\sum_{\alpha,\beta} S^+_{x\alpha}s_{\alpha\beta}(G_{\beta\beta} - m)G_{\alpha y}G_{y'\alpha}f(G) \\ & + m(G_{xx} - m)\sum_\alpha s_{x\alpha}G_{\alpha y}G_{y'\alpha}f(G) + m\sum_{\alpha,\beta} S^+_{x\alpha}s_{\alpha\beta}(G_{\alpha\alpha} - m)G_{\beta y}G_{y'\beta}f(G) \\ & - m\sum_\alpha s_{x\alpha}G_{\alpha y}G_{y'x}\partial_{h_{\alpha x}}f(G) - m\sum_{\alpha,\beta} S^+_{x\alpha}s_{\alpha\beta}G_{\beta y}G_{y'\alpha}\partial_{h_{\beta\alpha}}f(G) + \mathcal{Q}_{GG}. \end{aligned} \tag{3.17}$$

On the right-hand side of (3.17)*, the first term is the main term which is either of the same scaling order as $\mathcal{G}$ if $y = y'$ or has a strictly higher scaling order if $y \neq y'$; the second to fifth terms contain one more light weight and hence are of strictly higher scaling orders than $\mathcal{G}$; the sixth and seventh terms contain at least one more off-diagonal G edge and hence are of strictly higher scaling orders than $\mathcal{G}$. The last term $\mathcal{Q}_{GG}$ is a sum of Q-graphs defined by*

$$\begin{aligned}\mathcal{Q}_{GG} := {} & Q_x(\mathcal{G}) + \sum_\alpha Q_\alpha\Big[S^+_{x\alpha}G_{\alpha y}G_{y'\alpha}f(G)\Big] - mQ_y\Big[S^+_{xy}G_{y'y}f(G)\Big] - mQ_x\Big[\sum_\alpha s_{x\alpha}(G_{\alpha\alpha} - m)\mathcal{G}\Big] \\ & - m\sum_\alpha Q_\alpha\Big[\sum_\beta S^+_{x\alpha}s_{\alpha\beta}(G_{\beta\beta} - m)G_{\alpha y}G_{y'\alpha}f(G)\Big] - mQ_x\Big[G_{xx}\sum_\alpha s_{x\alpha}G_{\alpha y}G_{y'\alpha}f(G)\Big] \\ & - m\sum_\alpha Q_\alpha\Big[\sum_\beta S^+_{x\alpha}s_{\alpha\beta}G_{\alpha\alpha}G_{\beta y}G_{y'\beta}f(G)\Big] + mQ_x\Big[\sum_\alpha s_{x\alpha}G_{\alpha y}G_{y'x}\partial_{h_{\alpha x}}f(G)\Big] \\ & + m\sum_\alpha Q_\alpha\Big[\sum_\beta S^+_{x\alpha}s_{\alpha\beta}G_{\beta y}G_{y'\alpha}\partial_{h_{\beta\alpha}}f(G)\Big]. \end{aligned} \tag{3.18}$$

On the other hand, if $\mathcal{G} = \overline{G}_{xy}\overline{G}_{y'x}f(G)$, then we define

$$\widehat{\mathcal{O}}^{(x)}_{GG}\left[\mathcal{G}\right] := \overline{\widehat{\mathcal{O}}^{(x)}_{GG}\left[G_{xy}G_{y'x}\overline{f(G)}\right]},$$

where the right-hand side can be defined using (3.17)*. Finally, we define*

$$\mathcal{O}^{(x)}_{GG}[\mathcal{G}] := \mathcal{O}_{dot} \circ \widehat{\mathcal{O}}^{(x)}_{GG}[\mathcal{G}].$$

We describe the basic properties of the GG expansions in the following lemma. Its proof is straightforward by using Definition 3.15, and we postpone it to Appendix C.

Lemma 3.16. *Given a normal regular graph $\mathcal{G} = G_{xy}G_{y'x}f(G)$, where $f(G)$ contains no weights or solid edges attached to x and $y, y' \neq x$. Then $\mathcal{O}^{(x)}_{GG}[\mathcal{G}]$ is a canonical local expansion. Moreover, every graph without Q-labels, say $\mathcal{G}_1$, in $\mathcal{O}^{(x)}_{GG}[\mathcal{G}]$ satisfies one of the following properties.*

(a) If $y \neq y'$, then $\mathcal{G}_1$ has a strictly higher scaling order than $\mathcal{G}$.

(b) If $y = y'$, then either $\mathcal{G}_1$ has a strictly higher scaling order than $\mathcal{G}$, or $\mathcal{G}_1$ is obtained by replacing $G_{xy}G_{yx}$ in $\mathcal{G}$ with $mS^+_{xy}G_{yy}$.

Similar statements hold if $\mathcal{G} = \overline{G}_{xy}\overline{G}_{y'x}f(G)$.

Lemma 3.16 shows that, by applying the GG expansion repeatedly, we can either get rid of atoms that are connected with a pair of edges of the same charge, or obtain Q-graphs and graphs of sufficiently high scaling orders.

Now we define the following concept of *standard neutral atoms.* Roughly speaking, the edges connected with a standard neutral atom almost form a T-variable (but not an exact T-variable because of the $\times$-dotted edges; see Section 3.5 for more details).

Definition 3.17 (Standard neutral atoms)**.** *An atom is said to be* standard neutral *if it is only connected with three edges besides the $\times$-dotted edges: two matched G edges of opposite charges and one waved S edge.*

Given a graph with a non-standard neutral atom x (for example, the atom x in graph (f) of (4.3) below) that is connected with two matched G edges of opposite charges, we can apply the following $G\overline{G}$ expansion. The $G\overline{G}$ expansion (3.19) is a special case of the multi-edge expansion in Definition 3.11 with $k_1 = k_2 = 1$, $k_3 = k_4 = 0$ or $k_1 = k_2 = 0$, $k_3 = k_4 = 1$.

Definition 3.18 ($G\overline{G}$ expansion operator)**.** *Given a normal regular graph $\mathcal{G}$, suppose the atom x is connected with exactly two matched G edges of opposite charges, and $\mathcal{G}$ takes the form $\mathcal{G} = G_{xy}\overline{G}_{xy'}f(G)$ with $y, y' \neq x$. Then we define*

$$\begin{aligned}\widehat{\mathcal{O}}^{(x)}_{G\overline{G}}[\mathcal{G}] &:= |m|^2 \sum_\alpha s_{x\alpha} G_{\alpha y}\overline{G}_{\alpha y'} f(G) + m \sum_\alpha s_{x\alpha}\left(G_{\alpha\alpha} - m\right)\mathcal{G} \\ &\quad + m(\overline{G}_{xx} - \overline{m}) \sum_\alpha s_{x\alpha} G_{\alpha y}\overline{G}_{\alpha y'} f(G) - m \sum_\alpha s_{x\alpha} G_{\alpha y}\overline{G}_{xy'} \partial_{h_{\alpha x}} f(G) + \mathcal{Q}_{G\overline{G}}, \qquad (3.19)\end{aligned}$$

where on the right-hand side, the first term is of the same scaling order as $\mathcal{G}$, and the new atom α is standard neutral; the second and third terms contain one more light weight and hence are of strictly higher scaling orders than $\mathcal{G}$; the fourth term contains at least one more off-diagonal G edge and hence is of strictly higher scaling order than $\mathcal{G}$. The last term $\mathcal{Q}_{G\overline{G}}$ is a sum of Q-graphs defined by

$$\begin{aligned}\mathcal{Q}_{G\overline{G}} &:= Q_x\left(\mathcal{G}\right) - mQ_x\Big[\sum_\alpha s_{x\alpha}\overline{G}_{xx}G_{\alpha y}\overline{G}_{\alpha y'}f(G)\Big] - mQ_x\Big[\sum_\alpha s_{x\alpha}\left(G_{\alpha\alpha} - m\right)\mathcal{G}\Big] \\ &\quad + mQ_x\Big[\sum_\alpha s_{x\alpha}G_{\alpha y}\overline{G}_{xy'}\partial_{h_{\alpha x}} f(G)\Big].\end{aligned}$$

On the other hand, if $\mathcal{G} = G_{yx}\overline{G}_{y'x}f(G)$, then we define $\mathcal{O}_{G\overline{G}}(x)[\mathcal{G}]$ by taking $k_1 = k_2 = 0$ and $k_3 = k_4 = 1$ in Definition 3.11, and we omit the explicit expression for simplicity. Finally, we define

$$\mathcal{O}^{(x)}_{G\overline{G}}[\mathcal{G}] := \mathcal{O}_{dot} \circ \widehat{\mathcal{O}}^{(x)}_{G\overline{G}}[\mathcal{G}].$$

The purpose of the $G\overline{G}$ expansion is to turn the non-standard neutral atom x into a new standard neutral atom α in the first term. The following lemma describes the basic properties of the $G\overline{G}$ expansion. Its proof is straightforward by using Definition 3.18, and we postpone it to Appendix C.

Lemma 3.19. *Given a normal regular graph $\mathcal{G} = G_{xy}\overline{G}_{xy'}f(G)$, where $f(G)$ contains no weights or solid edges attached to x and $y, y' \neq x$. Then $\mathcal{O}^{(x)}_{G\overline{G}}[\mathcal{G}]$ is a canonical local expansion. Moreover, every graph without Q-labels, say $\mathcal{G}_1$, in $\mathcal{O}^{(x)}_{G\overline{G}}[\mathcal{G}]$ either has a strictly higher scaling order than $\mathcal{G}$, or satisfies one of the following properties:*

(a) $\mathrm{ord}(\mathcal{G}_1) = \mathrm{ord}(\mathcal{G})$, $\deg(x) = 0$ *in* $\mathcal{G}_1$, *and* $\mathcal{G}_1$ *contains one more standard neutral atom;*

(b) $\mathrm{ord}(\mathcal{G}_1) = \mathrm{ord}(\mathcal{G})$, *and* $\mathcal{G}_1$ *is obtained by replacing* $G_{xy}\overline{G}_{xy}$ *with* $|m|^2 s_{xy}|G_{yy}|^2$ *in the* $y = y'$ *case.*

Similar statements hold if $\mathcal{G} = G_{yx}\overline{G}_{y'x}f(G)$.

Lemma 3.19 shows that by applying the $G\overline{G}$ expansions repeatedly, we can get either new graphs containing only standard neutral atoms and degree 0 atoms, or Q-graphs and graphs of sufficiently high scaling orders.

3.4 Local expansion strategy

We define the concept of *locally standard graphs.*

Definition 3.20 (Locally standard graphs)**.** *A graph $\mathcal{G}$ is* locally standard *if*

(i) it is a normal regular graph without P/Q labels;

(ii) it has no weights or light weights;

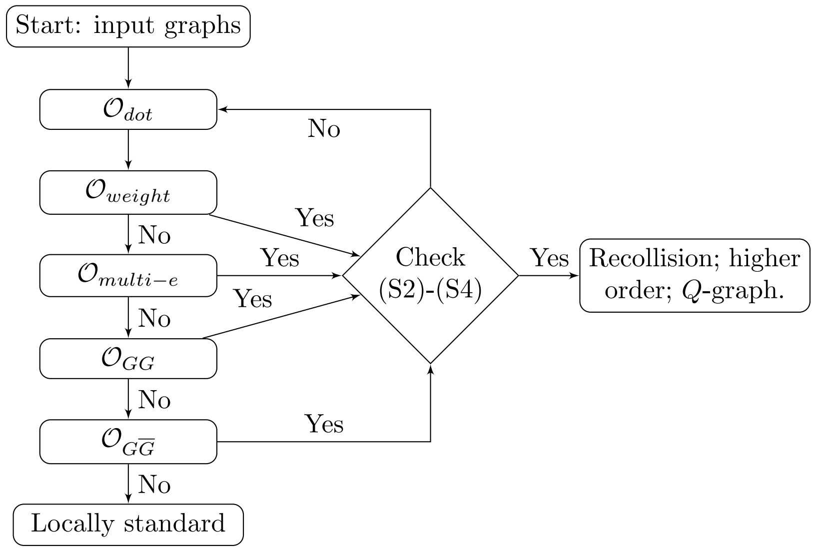


Figure 1: The flow chart for local expansions. If the weight, multi-edge, GG, or $G\overline{G}$ expansion does not do anything to an input graph (in which case we call it a *null operation*), then we have "No" and send it to the next operation. In particular, if all graph operations are null for a graph, then it is locally standard and will be sent to the output. On the other hand, if a non-trivial graph operation is acted on an input graph, then we have "Yes" and we will check whether the resulting graphs satisfy the stopping rules (S2)–(S4). If a graph indeed satisfies the stopping rules, then we send it to the output. Otherwise, we send it back to the first step $\mathcal{O}_{dot}$.

(iii) the degree of any internal atom is 0 *or* 2*;*

(iv) all degree 2 internal atoms are standard neutral atoms*.*

Applying local expansions in Definitions 3.2, 3.6, 3.11, 3.15 and 3.18 repeatedly, we can expand any regular graph into a linear combination of locally standard, recollision, higher order and Q graphs. The expansions will be performed according to the flow chart in Figure 1. More precisely, given a regular graph $\mathcal{G}$, we first apply $\mathcal{O}_{dot}$ to expand it into a sum of normal regular graphs, then apply the weight expansion to remove the weights, and then apply the multi-edge, GG and $G\overline{G}$ expansions one by one to remove all atoms that are not standard neutral. After an expansion, we may need to perform earlier expansions to the resulting graphs. For example, after a multi-edge expansion, we may get graphs that contain weights. Then before performing another multi-edge expansion, we first need to perform weight expansions to these graphs. This explains why we have loops in Figure 1.

To describe precisely the local expansion process in Figure 1, we define the following stopping rules. Given a cut-off order n, we stop the expansion of a graph if it is a normal regular graph and satisfies at least one of the following properties:

(S1) it is locally standard;

(S2) it is a $\oplus/\ominus$-recollision graph;

(S3) its scaling order is at least $n+1$;

(S4) it is a Q-graph.

Strategy 3.21 (Local expansion strategy)*.* We apply the following local expansion strategy.

(1) We first assign dotted edge partitions of the input graph using $\mathcal{O}_{dot}$ such that all resulting graphs are normal regular.

(2) For any input graph, pick an atom x and apply $\mathcal{O}^{(x)}_{weight}$ to expand the weights on x. For the resulting graphs from this expansion, we send the ones satisfying the stopping rules (S2)–(S4) to the outputs, and the remaining graphs back to the first operation $\mathcal{O}_{dot}$. If the input graph has no weight, then $\mathcal{O}_{weight}$ is a null operation and we send the graph to the next operation.

(3) For any input graph, if it contains atoms of degrees $\notin \{0,2\}$ or atoms connected with two mismatched edges, then we pick one of them, say x, and apply $\mathcal{O}^{(x)}_{multi-e}$ to expand the graph. For the resulting graphs, we send the ones satisfying the stopping rules (S2)–(S4) to the outputs, and the remaining graphs back to the first operation $\mathcal{O}_{dot}$. If every internal atom in the input graph either has degree 0 or is connected with exactly two matched solid edges, then $\mathcal{O}_{multi-e}$ is a null operation and we send the graph to the next operation.

(4) For any input graph, if it contains atoms connected with exactly two matched solid edges of the same charge, then we pick one of them, say x, and apply $\mathcal{O}^{(x)}_{GG}$ to expand the graph. For the resulting graphs, we send the ones satisfying the stopping rules (S2)–(S4) to the outputs, and the remaining graphs back to the first operation $\mathcal{O}_{dot}$. If every internal atom in an input graph is connected with exactly two matched edges of opposite charges, then $\mathcal{O}_{GG}$ is a null operation and we send the graph to the next operation.

(5) For any input graph, if it contains non-standard neutral atoms, then we pick one of them, say x, and apply $\mathcal{O}^{(x)}_{G\overline{G}}$ to expand the graph. For the resulting graphs, we send the ones satisfying the stopping rules (S2)–(S4) to the outputs, and the remaining graphs back to the first operation $\mathcal{O}_{dot}$.

(6) Finally, if all the above operations are null, then the input graph is locally standard, and we send it to the output.

Finally, we collect all the output graphs of Strategy 3.21 and obtain the following lemma. The proof of Lemma 3.22 is based on Lemmas 3.9, 3.13, 3.16 and 3.19, and is postponed to Appendix C.

Lemma 3.22. *Let $\mathcal{G}_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ be a normal regular graph without solid edges connected with $\mathfrak{a}$. Then for any fixed $n \in \mathbb{N}$, we can expand it into a sum of* $\mathrm{O}(1)$ *many graphs:*

$$\mathcal{G}_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = (\mathcal{G}_{local})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + \mathcal{R}^{(n)}_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + (\mathcal{A}^{(>n)}_{ho})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + \mathcal{Q}^{(n)}_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}, \tag{3.20}$$

where $(\mathcal{G}_{local})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ is a sum of locally standard graphs, $\mathcal{R}^{(n)}_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ is a sum of $\oplus/\ominus$-recollision graphs, $(\mathcal{A}^{(>n)}_{ho})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ is a sum of graphs of scaling order $> n$, and $(\mathcal{Q}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ is a sum of Q-graphs. Every molecule in the graphs on the right side is obtained by merging some molecules in the original graph $\mathcal{G}_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$.

We have noted that local expansions will not create new molecules. Hence if there are no dotted or waved edges added between different molecules, then the molecules in the new graphs are the same as those in $\mathcal{G}_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$. In general, there may be newly added dotted edges (due to the dotted edge partition $\mathcal{O}_{dot}$) or waved edges (due to the first term on the right-hand side of (3.17)) to the graphs, so the molecules in the new graphs are obtained from merging the molecules connected by dotted or waved edges.

3.5 Global expansions

In this section, we introduce the global expansions. Suppose that we have the $(n-1)$-th order T-expansion by induction. Given a locally standard graph, say $\mathcal{G}$, a global expansion consists of the following three steps:

(i) choosing a standard neutral atom in $\mathcal{G}$;

(ii) replacing the T-variable containing the atom in (i) by the $(n-1)$-th order T-expansion;

(iii) applying Q-expansions to the resulting graphs with Q-labels from (ii).

This procedure is called "global" because it may create new molecules in the resulting graphs. For example, if we replace T_{x,y_1y_2} with the right-hand side of (2.5) (with $f(G)\equiv 1$), then the new atoms α and β are in a different molecule from x. Unlike the local expansions, a global expansion may break the doubly connected properties of our graphs (cf. Definition 6.5). To avoid this issue, we need to follow a delicate procedure to choose the standard neutral atom in (i). This will be done fully in [64] and a brief discussion will be given in Section 9.

We now explain briefly the items (ii) and (iii) in the above procedure. Picking a standard neutral atom, say α, in a locally standard graph, the edges connected to it take one of the following forms:

$$t_{x,y_1y_2} := |m|^2 \sum_\alpha s_{x\alpha} G_{\alpha y_1}\overline{G}_{\alpha y_2}\mathbf{1}_{\alpha\ne y_1}\mathbf{1}_{\alpha\ne y_2}, \quad \text{or} \quad t_{y_1y_2,x} := |m|^2 \sum_\alpha G_{y_1\alpha}\overline{G}_{y_2\alpha} s_{\alpha x}\mathbf{1}_{\alpha\ne y_1}\mathbf{1}_{\alpha\ne y_2}. \tag{3.21}$$

Then we apply the $(n-1)$-th order T-expansion in (2.19) to these variables in the following way:

$$t_{x,y_1y_2} = m\Theta_{xy_1}\overline{G}_{y_1y_2} + m(\Theta\Sigma_T^{(n-1)}\Theta)_{xy_1}\overline{G}_{y_1y_2} + (\mathcal{R}_T^{(n-1)})_{x,y_1y_2} + (\mathcal{A}_T^{(>n-1)})_{x,y_1y_2} + (\mathcal{Q}_T^{(n-1)})_{x,y_1y_2}$$
$$+ (\mathcal{E}rr_{n-1,D})_{x,y_1y_2} - |m|^2\sum_\alpha s_{x\alpha}G_{\alpha y_1}\overline{G}_{\alpha y_2}\left(\mathbf{1}_{\alpha\neq y_1}\mathbf{1}_{\alpha=y_2} + \mathbf{1}_{\alpha=y_1}\mathbf{1}_{\alpha\neq y_2} + \mathbf{1}_{\alpha=y_1}\mathbf{1}_{\alpha=y_2}\right).$$

The last term on the right-hand side gives one (if $y_1 = y_2$) or two (if $y_1 \neq y_2$) recollision graphs, so we combine it with $(\mathcal{R}_T^{(n-1)})_{x,y_1y_2}$ and denote the resulting expression by $(\widetilde{\mathcal{R}}_T^{(n-1)})_{x,y_1y_2}$. Hence we have the final expansion formula

$$\begin{aligned} t_{x,y_1y_2} &= m\Theta_{xy_1}\overline{G}_{y_1y_2} + m(\Theta\Sigma_T^{(n-1)}\Theta)_{xy_1}\overline{G}_{y_1y_2} \\ &\quad + (\widetilde{\mathcal{R}}_T^{(n-1)})_{x,y_1y_2} + (\mathcal{A}_T^{(>n-1)})_{x,y_1y_2} + (\mathcal{Q}_T^{(n-1)})_{x,y_1y_2} + (\mathcal{E}rr_{n-1,D})_{x,y_1y_2}. \end{aligned} \tag{3.22}$$

The expansion of $t_{y_1y_2,x}$ can be obtained by exchanging the order of matrix indices in the above equation.

In a global expansion, if we replace t_{x,y_1y_2} in a graph, say $\mathcal{G}_0$, with a graph in $(\mathcal{Q}_T^{(n-1)})_{x,y_1y_2}$, we will get a graph of the form

$$\mathcal{G} = \sum_y \Gamma Q_y(\mathcal{G}_1), \tag{3.23}$$

where both Γ and $\mathcal{G}_1$ are graphs without P/Q labels (more precisely, Γ is the subgraph obtained by removing t_{x,y_1y_2} from $\mathcal{G}_0$, and $Q_y(\mathcal{G}_1)$ is a Q-graph in $(\mathcal{Q}_T^{(n-1)})_{x,y_1y_2}$). Applying the so-called Q-expansions, we can expand the above graph into a sum of Q-graphs and some graphs without P/Q labels. We will give the precise definition of Q-expansions in [64]. Here we only describe briefly the basic ideas. For any $y \in \mathbb{Z}_L^d$, let $H^{(y)}$ be the $(N-1)\times(N-1)$ minor of H obtained by removing the y-th row and column of H, and define the resolvent minor $G^{(y)}(z) := (H^{(y)} - z)^{-1}$. Using Schur complement formula, we can obtain the following resolvent identity:

$$G_{x_1x_2} = G^{(y)}_{x_1x_2} + \frac{G_{x_1y}G_{yx_2}}{G_{yy}}, \quad x_1, x_2 \in \mathbb{Z}_L^d.$$

Applying this identity to expand the resolvent entries in Γ one by one, we can write it as

$$\Gamma = \Gamma^{(y)} + \sum_\omega \Gamma_\omega. \tag{3.24}$$

Here $\Gamma^{(y)}$ is a graph whose weights and solid edges are $G^{(y)}$ entries, so it is independent of the y-th row and column of H. The other term is a sum of $\mathrm{O}(1)$ many graphs, where each Γ_ω has a strictly higher scaling order than Γ, at least two new solid edges connected with atom y, and a factor of the form $(G_{yy})^{-k}(\overline{G}_{yy})^{-l}$ for some $k, l \in \mathbb{N}$. The entry $1/G_{yy}$ can be expanded using Taylor expansion

$$\frac{1}{G_{yy}} = \frac{1}{m} + \sum_{k=1}^D \frac{1}{m}\left(-\frac{G_{yy}-m}{m}\right)^k + \mathcal{W}_{err}, \quad \mathcal{W}_{err} := \sum_{k>D}\left(-\frac{G_{yy}-m}{m}\right)^k.$$

We will regard $\mathcal{W}_{err}$ as a weight of scaling order $> D$ and collect all graphs containing it into $\mathcal{E}rr_{n,D}$ in (2.19). Using (3.24), we can expand (3.23) as

$$\mathcal{G} = \sum_\omega\sum_y \Gamma_\omega Q_y(\mathcal{G}_1) + \sum_y Q_y(\Gamma\mathcal{G}_1) - \sum_\omega\sum_y Q_y(\Gamma_\omega\mathcal{G}_1), \tag{3.25}$$

where the second and third terms are sums of Q-graphs. For the first term, we will remove Q_y using some operations that will be introduced in [64]. The above Q-expansion is an expansion of the commutator $[\Gamma, Q_y]$. It has the following important properties: (i) the scaling order of any graph $\sum_y \Gamma_\omega Q_y(\mathcal{G}_0)$ is strictly higher than $\mathcal{G}$; (ii) for any ω, at least one weight or solid edge in Γ is replaced by two solid edges connected with y in Γ_ω.

Remark 3.23. The local and global expansions can be readily extended to non-Gaussian band matrices. The Gaussian integration by parts will be replaced by the following cumulant expansion in [47, Proposition 3.1] and [44, Section II]. Fix an integer $l \in \mathbb{N}$ and let h be a real-valued random variable with finite moments up to order $l+2$. Then for any $f \in \mathcal{C}^{l+1}(\mathbb{R})$, we have that

$$\mathbb{E}[f(h)h] = \sum_{k=0}^{l} \frac{1}{k!}\kappa_{k+1}(h)\mathbb{E}f^{(k)}(h) + R_{l+1},$$

where $\kappa_k(h)$ is the k-th cumulant of h and R_{l+1} satisfies that for any $K>0$,

$$R_{l+1} \lesssim \mathbb{E}\left|h^{l+2}\mathbf{1}_{|h|>K}\right| \cdot \|f^{(l+1)}\|_\infty + \mathbb{E}\,|h|^{l+2} \cdot \sup_{|x|\leqslant K} |f^{(l+1)}(x)|.$$

Using the cumulant expansions, we can extend the expansions in Lemma 2.4 and Definitions 3.6, 3.11, 3.15 and 3.18 to general cases. These general expansions will make the Θ-expansion and the local expansions more complicated, but there are no new "essential" difficulties. Moreover, the global expansions defined in this subsection can be used without any change regardless of the distributions of the matrix entries. With these remarks, we can prove our main results for random band matrices with entries satisfying only certain moment assumptions. Due to the length constraint of the current paper, we will postpone the details of this generalization to a future work.

4 Examples of low order T-expansions

To help the reader to understand how operations in Section 3 are applied, in this section we give some examples of low order T-expansions. We remark that these examples will not be used in the proof of Theorem 2.1, so the reader can skip this section and go to Section 5 directly for the main proof.

4.1 Third order T-expansion

We can derive the third order T-expansion by further expanding (2.7). Applying the weight expansion (3.5) to the two terms in $(\mathcal{A}_T^{(>2)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, we can obtain that

$$\begin{aligned}(\mathcal{A}_T^{(>2)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} &= m^2 \sum_{x,y,\alpha,\beta} \Theta_{\mathfrak{a}x} s_{xy}\left(\delta_{y\alpha} + S^+_{y\alpha}\right) s_{\alpha\beta}(G_{\alpha\alpha}-m)(G_{\beta\beta}-m)G_{x\mathfrak{b}_1}\overline{G}_{x\mathfrak{b}_2} \\ &- m^2 \sum_{x,y,\alpha,\beta} \Theta_{\mathfrak{a}x} s_{xy}\left(\delta_{y\alpha} + S^+_{y\alpha}\right) s_{\alpha\beta}G_{\beta\alpha}\partial_{h_{\beta\alpha}}(G_{x\mathfrak{b}_1}\overline{G}_{x\mathfrak{b}_2}) \\ &+ |m|^2 \sum_{x,y,\alpha,\beta} \Theta_{\mathfrak{a}x} s_{xy}\left(\delta_{x\alpha} + S^-_{x\alpha}\right) s_{\alpha\beta}(\overline{G}_{\alpha\alpha}-\overline{m})(\overline{G}_{\beta\beta}-\overline{m})G_{y\mathfrak{b}}\overline{G}_{y\mathfrak{b}_2} \\ &- |m|^2 \sum_{x,y,\alpha,\beta} \Theta_{\mathfrak{a}x} s_{xy}\left(\delta_{x\alpha} + S^-_{x\alpha}\right) s_{\alpha\beta}\overline{G}_{\beta\alpha}\partial_{h_{\alpha\beta}}(G_{y\mathfrak{b}_1}\overline{G}_{y\mathfrak{b}_2}) + \mathcal{Q}_{T,3}.\end{aligned}$$

Here $\mathcal{Q}_{T,3}$ is a sum of Q-graphs that can be derived from (3.5), but we do not write down its expression for simplicity. If we expand the partial derivatives using (3.8), and use the identity

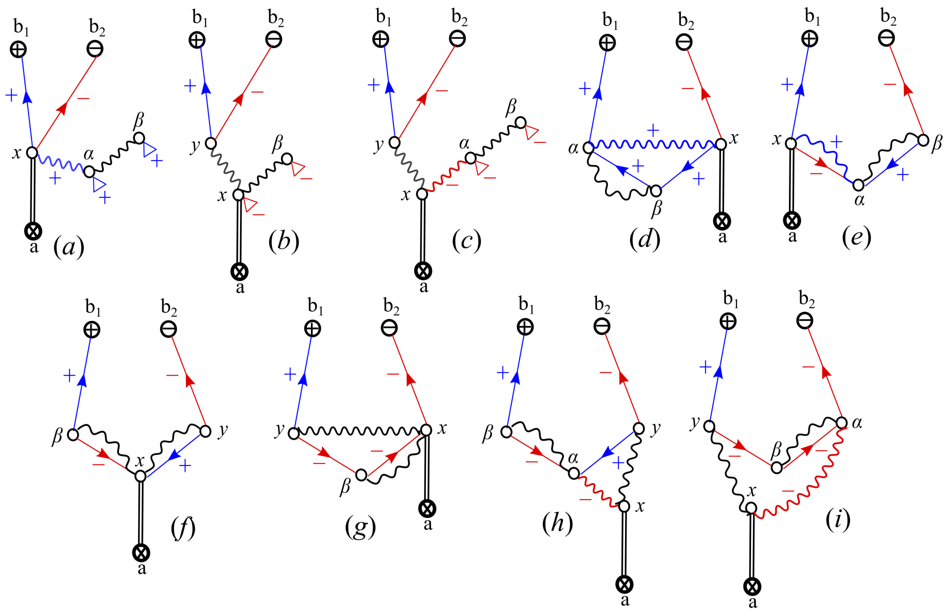
$$\sum_y m^2 s_{xy}\left(\delta_{y\alpha} + S^+_{y\alpha}\right) = S^+_{x\alpha}, \tag{4.1}$$

we can reduce the above expansion to $(\mathcal{A}_T^{(>2)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = (\mathcal{A}_T^{(>3)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} + (\mathcal{Q}_{T,3})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, where

$$\begin{aligned}(\mathcal{A}_T^{(>3)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} := &\sum_{x,\alpha,\beta} \Theta_{\mathfrak{a}x} S^+_{x\alpha} s_{\alpha\beta}(G_{\alpha\alpha}-m)(G_{\beta\beta}-m)G_{x\mathfrak{b}_1}\overline{G}_{x\mathfrak{b}_2} \\ &+ |m|^2 \sum_{x,y,\beta} \Theta_{\mathfrak{a}x} s_{xy} s_{x\beta}(\overline{G}_{xx}-\overline{m})(\overline{G}_{\beta\beta}-\overline{m})G_{y\mathfrak{b}_1}\overline{G}_{y\mathfrak{b}_2} \\ &+ |m|^2 \sum_{x,y,\alpha,\beta} \Theta_{\mathfrak{a}x} s_{xy} S^-_{x\alpha} s_{\alpha\beta}(\overline{G}_{\alpha\alpha}-\overline{m})(\overline{G}_{\beta\beta}-\overline{m})G_{y\mathfrak{b}_1}\overline{G}_{y\mathfrak{b}_2}\end{aligned} \tag{4.2}$$

$$+\sum_{x,\alpha,\beta}\Theta_{\mathfrak{a}x}S^+_{x\alpha}s_{\alpha\beta}G_{\beta\alpha}G_{x\beta}G_{\alpha\mathfrak{b}_1}\overline{G}_{x\mathfrak{b}_2}+\sum_{x,\alpha,\beta}\Theta_{\mathfrak{a}x}S^+_{x\alpha}s_{\alpha\beta}G_{\beta\alpha}G_{x\mathfrak{b}_1}\overline{G}_{x\alpha}\overline{G}_{\beta\mathfrak{b}_2}$$
$$+|m|^2\sum_{x,y,\beta}\Theta_{\mathfrak{a}x}s_{xy}s_{x\beta}\overline{G}_{\beta x}G_{yx}G_{\beta\mathfrak{b}_1}\overline{G}_{y\mathfrak{b}_2}+|m|^2\sum_{x,y,\beta}\Theta_{\mathfrak{a}x}s_{xy}s_{x\beta}\overline{G}_{\beta x}G_{y\mathfrak{b}_1}\overline{G}_{y\beta}\overline{G}_{x\mathfrak{b}_2}$$
$$+|m|^2\sum_{x,y,\alpha,\beta}\Theta_{\mathfrak{a}x}s_{xy}S^-_{x\alpha}s_{\alpha\beta}\overline{G}_{\beta\alpha}G_{y\alpha}G_{\beta\mathfrak{b}_1}\overline{G}_{y\mathfrak{b}_2}+|m|^2\sum_{x,y,\alpha,\beta}\Theta_{\mathfrak{a}x}s_{xy}S^-_{x\alpha}s_{\alpha\beta}\overline{G}_{\beta\alpha}G_{y\mathfrak{b}_1}\overline{G}_{y\beta}\overline{G}_{\alpha\mathfrak{b}_2}.$$

Now we draw the 9 graphs of (4.2) in the following figure:



(4.3)

For conciseness, we do not draw the coefficients of these graphs. The graphs in (4.3) are not yet normal regular, but it is easy to see that after applying $\mathcal{O}_{dot}$ to them, all the resulting graphs are of scaling order $\geqslant 4$. Thus we have obtained the following third order T-expansion

$$T_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}=m\Theta_{\mathfrak{a}\mathfrak{b}_1}\overline{G}_{\mathfrak{b}_1\mathfrak{b}_2}+(\mathcal{A}_T^{(>3)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}+(\mathcal{Q}_T^{(3)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}, \tag{4.4}$$

where $\mathcal{Q}_T^{(3)}:=\mathcal{Q}_{T,3}+\mathcal{Q}_T^{(2)}$.

4.2 Fourth order T-expansion

Next we can perform local and global expansions to the graphs in $\mathcal{A}_T^{(>3)}$ to construct the fourth order T-expansion. Since the expression of the fourth order T-expansion is rather lengthy and does not help our proof, we will not give its explicit form in this paper. Instead, we will describe the expansions of several typical graphs to show that we actually have $\Sigma_T^{(4)}=0$ in the current setting where H has complex Gaussian entries.

First, the graphs (a), (b), (c) in (4.3) all have two light weights in them. Taking graph (a) as an example, we apply the weight expansion in Definition 3.5 to the weight $G_{\beta\beta}-m$ and get that

$$\sum_{x,\alpha,\beta}\Theta_{\mathfrak{a}x}S^+_{x\alpha}s_{\alpha\beta}(G_{\alpha\alpha}-m)(G_{\beta\beta}-m)G_{x\mathfrak{b}_1}\overline{G}_{x\mathfrak{b}_2}=\mathcal{Q}_a$$

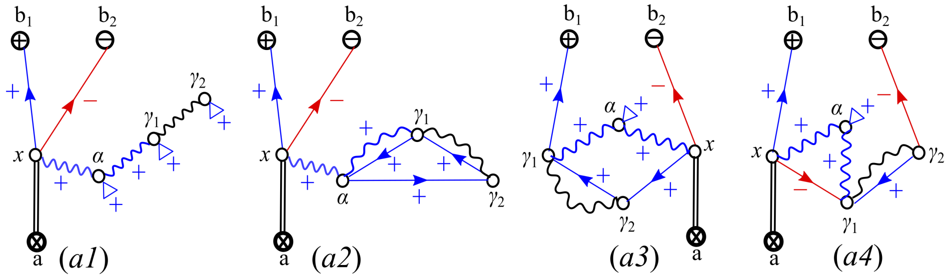
$$+m^{-1}\sum_{x,\alpha,\gamma_1,\gamma_2}\Theta_{\mathfrak{a}x}S^+_{x\alpha}S^+_{\alpha\gamma_1}s_{\gamma_1\gamma_2}(G_{\alpha\alpha}-m)(G_{\gamma_1\gamma_1}-m)(G_{\gamma_2\gamma_2}-m)G_{x\mathfrak{b}_1}\overline{G}_{x\mathfrak{b}_2} \tag{a1}$$

$$+m^{-1}\sum_{x,\alpha,\gamma_1,\gamma_2}\Theta_{\mathfrak{a}x}S^+_{x\alpha}S^+_{\alpha\gamma_1}s_{\gamma_1\gamma_2}G_{\gamma_2\gamma_1}G_{\alpha\gamma_2}G_{\gamma_1\alpha}G_{x\mathfrak{b}_1}\overline{G}_{x\mathfrak{b}_2} \tag{a2}$$

$$+m^{-1}\sum_{x,\alpha,\gamma_1,\gamma_2}\Theta_{\mathfrak{a}x}S^+_{x\alpha}S^+_{\alpha\gamma_1}s_{\gamma_1\gamma_2}(G_{\alpha\alpha}-m)G_{\gamma_2\gamma_1}G_{x\gamma_2}G_{\gamma_1\mathfrak{b}_1}\overline{G}_{x\mathfrak{b}_2} \tag{a3}$$

$$+\,m^{-1}\sum_{x,\alpha,\gamma_1,\gamma_2}\Theta_{\mathfrak{a}x}S^{+}_{x\alpha}S^{+}_{\alpha\gamma_1}s_{\gamma_1\gamma_2}(G_{\alpha\alpha}-m)G_{\gamma_2\gamma_1}G_{x\mathfrak{b}_1}\overline{G}_{x\gamma_1}\overline{G}_{\gamma_2\mathfrak{b}_2}, \tag{a4}$$

where we used (4.1) in the derivation, and $\mathcal{Q}_a$ is a sum of Q-graphs. In (4.5), we draw the four graphs (a1)–(a4), where for conciseness we do not draw the coefficients of them.



(4.5)

The graphs in (4.5) are not yet normal regular, but it is easy to see that after applying $\mathcal{O}_{dot}$ to them, all the resulting graphs are of scaling order $\geqslant 5$. Similarly, we can check that applying the weight expansion to the light weights in graphs (b) and (c) of (4.3) will give graphs of scaling order $\geqslant 5$.

Second, the graphs (d), (g) and (i) in (4.3) all have an atom β connected with two matched edges of the same charge. Taking graph (i) as an example, we apply the GG expansion in Definition 3.15 to the two edges connected with β, and get that

$$|m|^2\sum_{x,y,\alpha,\beta}\Theta_{\mathfrak{a}x}s_{xy}S^{-}_{x\alpha}s_{\alpha\beta}\overline{G}_{\beta\alpha}\overline{G}_{y\beta}G_{y\mathfrak{b}_1}\overline{G}_{\alpha\mathfrak{b}_2}=\mathcal{Q}_i$$

$$+\,|m|^2\overline{m}\sum_{x,y,\alpha,\beta}\Theta_{\mathfrak{a}x}s_{xy}S^{-}_{x\alpha}s_{\alpha\beta}S^{-}_{\alpha\beta}\overline{G}_{y\alpha}G_{y\mathfrak{b}_1}\overline{G}_{\alpha\mathfrak{b}_2} \tag{i1}$$

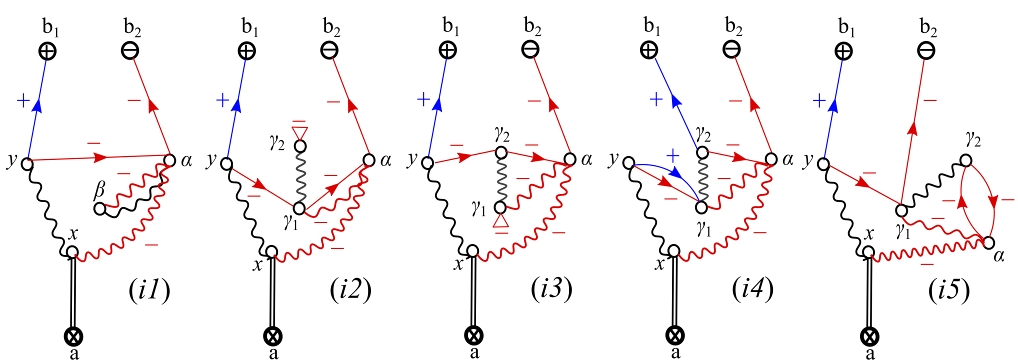
$$+\,m\sum_{x,y,\alpha,\gamma_1,\gamma_2}\Theta_{\mathfrak{a}x}s_{xy}S^{-}_{x\alpha}S^{-}_{\alpha\gamma_1}s_{\gamma_1\gamma_2}(\overline{G}_{\gamma_2\gamma_2}-\overline{m})\overline{G}_{\gamma_1\alpha}\overline{G}_{y\gamma_1}G_{y\mathfrak{b}_1}\overline{G}_{\alpha\mathfrak{b}_2} \tag{i2}$$

$$+\,m\sum_{x,y,\alpha,\gamma_1,\gamma_2}\Theta_{\mathfrak{a}x}s_{xy}S^{-}_{x\alpha}S^{-}_{\alpha\gamma_1}s_{\gamma_1\gamma_2}(\overline{G}_{\gamma_1\gamma_1}-\overline{m})\overline{G}_{\gamma_2\alpha}\overline{G}_{y\gamma_2}G_{y\mathfrak{b}_1}\overline{G}_{\alpha\mathfrak{b}_2} \tag{i3}$$

$$+\,m\sum_{x,y,\alpha,\gamma_1,\gamma_2}\Theta_{\mathfrak{a}x}s_{xy}S^{-}_{x\alpha}S^{-}_{\alpha\gamma_1}s_{\gamma_1\gamma_2}|G_{y\gamma_1}|^2\overline{G}_{\gamma_2\alpha}G_{\gamma_2\mathfrak{b}_1}\overline{G}_{\alpha\mathfrak{b}_2} \tag{i4}$$

$$+\,m\sum_{x,y,\alpha,\gamma_1,\gamma_2}\Theta_{\mathfrak{a}x}s_{xy}S^{-}_{x\alpha}S^{-}_{\alpha\gamma_1}s_{\gamma_1\gamma_2}\overline{G}_{y\gamma_1}\overline{G}_{\alpha\gamma_2}\overline{G}_{\gamma_2\alpha}G_{y\mathfrak{b}_1}\overline{G}_{\gamma_1\mathfrak{b}_2}, \tag{i5}$$

where we used the complex conjugate of (4.1) in the derivation, and $\mathcal{Q}_i$ is a sum of Q-graphs. In (4.6), we draw the five graphs (i1)–(i5), where for conciseness we do not draw the coefficients of the graphs.



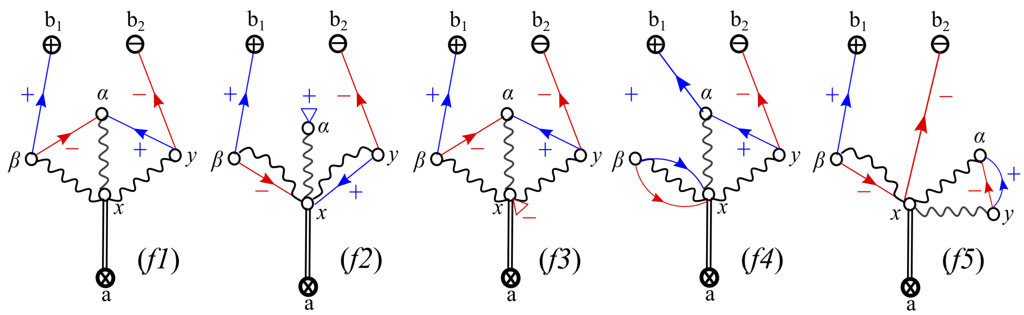
(4.6)

The graphs in (4.6) are not yet normal regular, but it is easy to see that after applying $\mathcal{O}_{dot}$ to them, all the resulting graphs are of scaling order $\geqslant 5$. Similarly, we can check that applying the GG expansion to the two G edges connected with atom β in the graphs (d) and (g) of (4.3) will give graphs of scaling order $\geqslant 5$.

Finally, the graphs (e), (f) and (h) of (4.3) only contain degree 2 atoms connected with two matched edges of opposite charges, but not all atoms in them are standard neutral, such as the atom x in (f) and the atom α in (e) and (h). Taking graph (f) as an example, we apply the $G\overline{G}$ expansion in Definition 3.18 to the two edges connected with atom x, and get that

$$
\begin{aligned}
&|m|^2 \sum_{x,y,\beta} \Theta_{\mathfrak{a}x} s_{xy} s_{x\beta} \overline{G}_{\beta x} G_{yx} G_{\beta \mathfrak{b}_1} \overline{G}_{y\mathfrak{b}_2} = \mathcal{Q}_f \\
&+ |m|^4 \sum_{x,y,\alpha,\beta} \Theta_{\mathfrak{a}x} s_{xy} s_{x\beta} s_{x\alpha} \overline{G}_{\beta\alpha} G_{y\alpha} G_{\beta\mathfrak{b}_1} \overline{G}_{y\mathfrak{b}_2} && \text{(f1)} \\
&+ |m|^2 m \sum_{x,y,\alpha,\beta} \Theta_{\mathfrak{a}x} s_{xy} s_{x\beta} s_{x\alpha} (G_{\alpha\alpha} - m) \overline{G}_{\beta x} G_{yx} G_{\beta\mathfrak{b}_1} \overline{G}_{y\mathfrak{b}_2} && \text{(f2)} \\
&+ |m|^2 m \sum_{x,y,\alpha,\beta} \Theta_{\mathfrak{a}x} s_{xy} s_{x\beta} s_{x\alpha} (\overline{G}_{xx} - \overline{m}) \overline{G}_{\beta\alpha} G_{y\alpha} G_{\beta\mathfrak{b}_1} \overline{G}_{y\mathfrak{b}_2} && \text{(f3)} \\
&+ |m|^2 m \sum_{x,y,\alpha,\beta} \Theta_{\mathfrak{a}x} s_{xy} s_{x\beta} s_{x\alpha} G_{y\alpha} |G_{\beta x}|^2 G_{\alpha\mathfrak{b}_1} \overline{G}_{y\mathfrak{b}_2} && \text{(f4)} \\
&+ |m|^2 m \sum_{x,y,\alpha,\beta} \Theta_{\mathfrak{a}x} s_{xy} s_{x\beta} s_{x\alpha} \overline{G}_{\beta x} |G_{y\alpha}|^2 G_{\beta\mathfrak{b}_1} \overline{G}_{x\mathfrak{b}_2}, && \text{(f5)}
\end{aligned}
$$

where $\mathcal{Q}_f$ is a sum of Q-graphs. In (4.7), we draw the five graphs (f1)–(f5), where for conciseness we do not draw the coefficients of them.

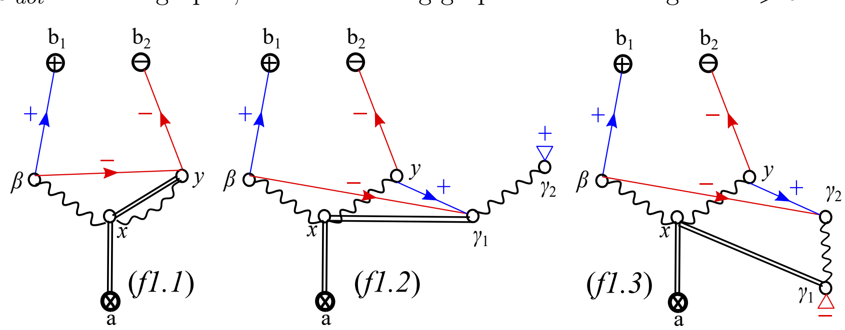


(4.7)

Here the graph (f1) is the main term, while the graphs (f2)–(f5) all give graphs of scaling order $\geqslant 5$ after a dotted edge partition $\mathcal{O}_{dot}$. Next we apply a global expansion to (f1), that is, we replace $|m|^2 \sum_\alpha s_{x\alpha} G_{y\alpha} \overline{G}_{\beta\alpha}$ with the second order T-expansion in (2.7):

$$
\begin{aligned}
&|m|^4 \sum_{x,y,\alpha,\beta} \Theta_{\mathfrak{a}x} s_{xy} s_{x\beta} s_{x\alpha} \overline{G}_{\beta\alpha} G_{y\alpha} G_{\beta\mathfrak{b}_1} \overline{G}_{y\mathfrak{b}_2} = |m|^2 \sum_{x,y,\beta} \Theta_{\mathfrak{a}x} s_{xy} s_{x\beta} (\mathcal{Q}_T^{(2)})_{y\beta,x} G_{\beta\mathfrak{b}_1} \overline{G}_{y\mathfrak{b}_2} \\
&+ |m|^2 m \sum_{x,y,\beta} \Theta_{\mathfrak{a}x} s_{xy} s_{x\beta} \Theta_{xy} \overline{G}_{\beta y} G_{\beta\mathfrak{b}_1} \overline{G}_{y\mathfrak{b}_2} && \text{(f1.1)} \\
&+ |m|^2 m \sum_{x,y,\beta,\gamma_1,\gamma_2} \Theta_{\mathfrak{a}x} s_{xy} s_{x\beta} \Theta_{x\gamma_1} s_{\gamma_1\gamma_2} (G_{\gamma_2\gamma_2} - m) \overline{G}_{\beta\gamma_1} G_{y\gamma_1} G_{\beta\mathfrak{b}_1} \overline{G}_{y\mathfrak{b}_2} && \text{(f1.2)} \\
&+ |m|^2 m \sum_{x,y,\beta,\gamma_1,\gamma_2} \Theta_{\mathfrak{a}x} s_{xy} s_{x\beta} \Theta_{x\gamma_1} s_{\gamma_1\gamma_2} (\overline{G}_{\gamma_1\gamma_1} - \overline{m}) \overline{G}_{\beta\gamma_2} G_{y\gamma_2} G_{\beta\mathfrak{b}_1} \overline{G}_{y\mathfrak{b}_2}. && \text{(f1.3)}
\end{aligned}
$$

We need to further apply a Q-expansion to the first term on the right-hand side. In (4.8), we draw the three graphs (f1.1)–(f1.3), where for conciseness we do not draw the coefficients of them. It is easy to see that after applying $\mathcal{O}_{dot}$ to these graphs, all the resulting graphs are of scaling order $\geqslant 5$.



(4.8)

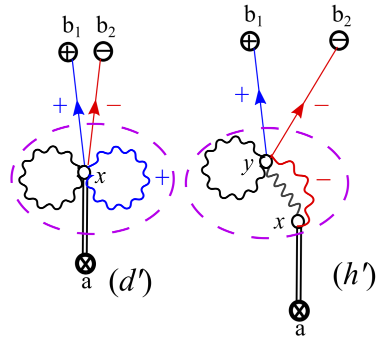
Similarly, we can check that after applying the $G\overline{G}$ expansion to the two G edges connected with atom α in graphs (e) and (h) of (4.3) and then applying a global expansion, we will get graphs of scaling order $\geqslant 5$.

To summarize, we have found that in the fourth order T-expansion, the fourth order self-energy $\mathcal{E}_4$ vanishes.

4.3 Examples of graphs in self-energy $\mathcal{E}_6$

In this subsection, we use some examples to show that the 6th order self-energy actually contains non-trivial graphs. Hence, unlike $\mathcal{E}_4$, its sum zero property (2.16) is not trivial anymore. We remark there are hundreds of ways to get graphs in $\mathcal{E}_6$, and we are not trying to exhaust all of them.

First, if we assign the dotted edge partition such that the two internal G edges in (d)–(i) of (4.3) are diagonal, then we will get sixth order graphs. For example, for the graph (d) of (4.3), we assign dotted edges $\delta_{x\beta}$ and $\delta_{\alpha\beta}$, and then replace the two weights G_{xx}^2 with m^2; for the graph (h), we assign dotted edges $\delta_{y\alpha}$ and $\delta_{\alpha\beta}$, and then replace the two weights $|G_{yy}|^2$ with $|m|^2$. Then we get the following two graphs:

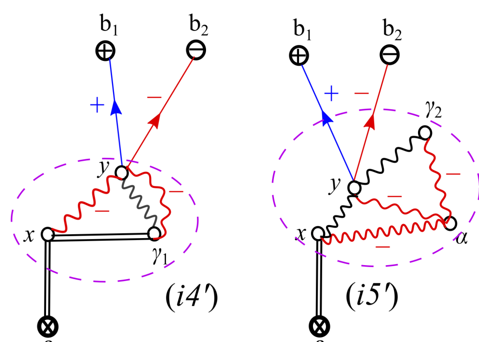


(4.9)

Inside the purple dashed circles are two deterministic graphs in $(\mathcal{E}_6)_{xy}$ (except for the coefficients):

$$(d'):\ m^2\delta_{xy}s_{xx}S^+_{xx},\quad (h'):\ |m|^4 s_{yy}s_{xy}S^-_{xy}.$$

As the second example, in the graph (i4) of (4.6), if we assign a dotted edge $\delta_{\alpha\gamma_2}$, replace the weight $\overline{G}_{\alpha\alpha}$ with $\overline{m}$, replace the T-variable $|m|^2\sum_y s_{xy}|G_{y\gamma_1}|^2$ with $|m|^2\Theta_{x\gamma_1}$ in a global expansion, and rename α as y, we then get the graph (i4′) in (4.10). In the graph (i5) of (4.6), if we assign a dotted edge $\delta_{y\gamma_1}$, replace the weight $\overline{G}_{yy}$ with $\overline{m}$, and replace the two edges $\overline{G}_{\alpha\gamma_2}\overline{G}_{\gamma_2\alpha}$ with $\overline{m}^2 S^-_{\alpha\gamma_2}$ in the GG expansion, then we get the graph (i5′) in (4.10).

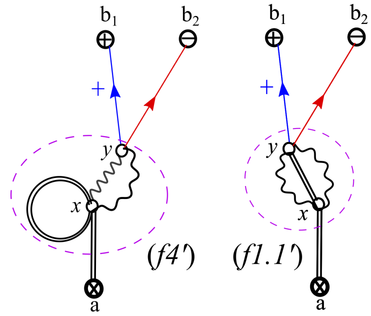


(4.10)

Inside the purple dashed circles are two deterministic graphs in $(\mathcal{E}_6)_{xy}$ (except for the coefficients):

$$(i4'):\ |m|^2 S^-_{xy}\sum_{\gamma_1} s_{y\gamma_1}S^-_{y\gamma_1}\Theta_{x\gamma_1},\quad (i5'):\ |m|^2\overline{m}^2 s_{xy}\sum_{\alpha,\gamma_2} S^-_{x\alpha}S^-_{y\alpha}S^-_{\alpha\gamma_2}s_{y\gamma_2}.$$

As the last example, in the graph (f4) of (4.7), if we assign a dotted edge $\delta_{y\alpha}$, replace the weight G_{yy} with m, and replace the T-variable $|m|^2\sum_\beta s_{x\beta}|G_{\beta x}|^2$ with a $|m|^2\Theta_{xx}$ edge in a global expansion, then we get the graph (f4′) in (4.11). In the graph (f1.1) of (4.8), if we assign a dotted edge $\delta_{y\beta}$ and replace the weight $\overline{G}_{yy}$ with $\overline{m}$, then we get the graph (f1.1′) in (4.11).



$$(4.11)$$

Inside the purple dashed circles are two deterministic graphs in $(\mathcal{E}_6)_{xy}$ (except for the coefficients):

$$(f4') : \ |m|^2 m^2 \Theta_{xx}(s_{xy})^2, \quad (f1.1') : \ |m|^4 (s_{xy})^2 \Theta_{xy}.$$

5 Proof of Theorem 2.1

In this section, we give an outline of the proof of Theorem 2.1. Some lemmas used in the proof will be proved in subsequent sections and [64]. We first recall the following large deviation estimates in Lemma 5.1, which show that the resolvent entries can be bounded using the T-variables in (1.21). The bound (5.2) was proved in equation (3.20) of [65], while (5.3) was proved in Lemma 5.3 of [27]. Given a matrix M, we will use $\|M\|_{\max} = \max_{i,j} |M_{ij}|$ to denote its maximum norm.

Lemma 5.1. *Suppose for a constant $\delta_0 > 0$ and deterministic parameter $W^{-d/2} \leqslant \Phi \leqslant W^{-\delta_0}$ we have that*

$$\|G(z) - m(z)\|_{\max} \prec W^{-\delta_0}, \quad \|T\|_{\max} \prec \Phi^2, \tag{5.1}$$

uniformly in $z \in \mathbf{D}$ for a subset $\mathbf{D} \subset \mathbb{C}_+$. Then

$$\mathbf{1}_{x\neq y} |G_{xy}(z)|^2 \prec T_{xy}(z) \tag{5.2}$$

uniformly in $x \neq y \in \mathbb{Z}_L^d$ and $z \in \mathbf{D}$, and

$$|G_{xx}(z) - m(z)| \prec \Phi \tag{5.3}$$

uniformly in $x \in \mathbb{Z}_L^d$ and $z \in \mathbf{D}$.

5.1 Main structure of the proof

The proof of Theorem 2.1 will proceed by induction on n, the scaling order of the T-expansion.

Step 1: Second order T-expansion. The second order T-expansion has been given by Lemma 2.5.

Step 2: Local law. Assume by induction that we have obtained the k-th order T-expansion for $2 \leqslant k \leqslant n-1$. Then we will prove in Theorem 2.16 that the local law (2.27) holds when L satisfies the condition (with n in (2.26) replaced by $n-1$)

$$L^2/W^2 \leqslant W^{(n-2)d/2 - c_0}. \tag{5.4}$$

Step 3: n-th order T-equation. Given the k-th order T-expansion for $2 \leqslant k \leqslant n-1$, we will construct an n-th order T-equation in Lemma 5.7.

Step 4: Sum zero property. With the n-th order T-equation in Step 3, using the local law proved in Step 2 we will show in Lemma 5.8 that the n-th order self-energy $\mathcal{E}_n$ satisfies the sum zero property.

Step 5: n-th order T-expansion. With the n-th order T-equation in Step 3 and the sum zero property for $\mathcal{E}_n$, we will construct an n-th order T-expansion in Lemma 5.13.

Combining these steps, by induction on n we obtain an n-th order T-expansion for any fixed $n \in \mathbb{N}$. Theorem 2.16 then implies that the local law (2.27) holds for L satisfying (2.26). This concludes Theorem 2.1 since n is arbitrary. In Figure 2, we illustrate the structure of the whole proof of Theorem 2.1 with a flow chart.

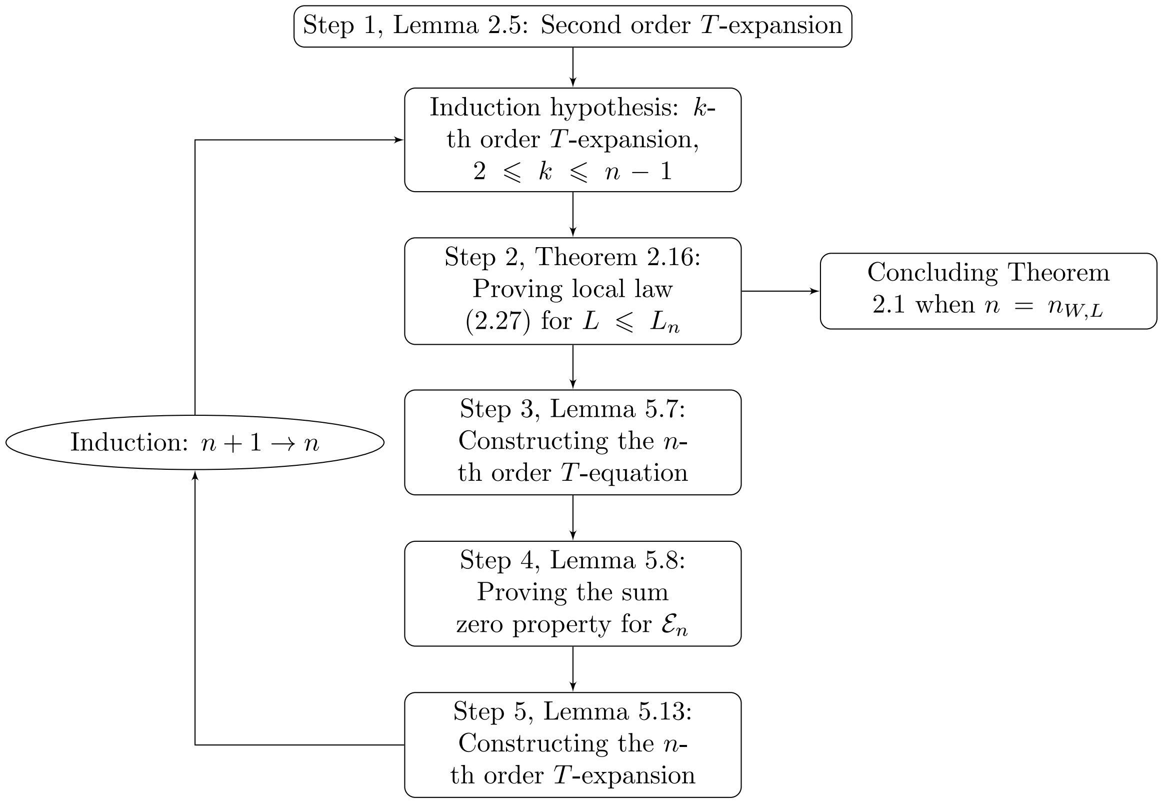


Figure 2: The main structure of the proof of Theorem 2.1. Corresponding to (5.4), we denote $L_n := W^{1+(n-2)d/4-c_0/2}$ for a constant $c_0 > 0$. Moreover, given the L in Theorem 2.1, it is enough to perform the induction up to $n_{W,L} := \lceil \frac{4}{d}\left(\log_W L - 1 + \frac{c_0}{2}\right)\rceil + 2$.

5.2 Step 2: Proof of Theorem 2.16

In this subsection, we prove Theorem 2.16, which is based on three main ingredients, Lemmas 5.2, 5.3 and 5.4. The first step is an initial estimate when $\eta = 1$. The following lemma is a folklore result, and has been proved in e.g. [27] in a different setting. We will give a formal proof in [64].

Lemma 5.2 (Initial estimate, Lemma 7.2 of [64])**.** *Under the assumptions of Theorem 2.16, for any $z = E + \mathrm{i}\eta$ with $E \in (-2+\kappa, 2-\kappa)$ and $\eta = 1$, we have that*

$$|G_{xy}(z) - m(z)\delta_{xy}|^2 \prec B_{xy}, \quad \forall\, x, y \in \mathbb{Z}_L^d. \tag{5.5}$$

The second step is the following continuity estimate, Lemma 5.3, whose proof will be given at the end of this subsection. It allows us to get some a priori estimates on $G(z)$ from the local law (2.27) on $G(\widetilde{z})$ for $\widetilde{z}$ with a larger imaginary part $\operatorname{Im} \widetilde{z} = W^{\varepsilon_0} \operatorname{Im} z$ for a small constant $\varepsilon_0 > 0$.

Lemma 5.3 (Continuity estimate)**.** *Under the assumptions of Theorem 2.16, suppose that*

$$|G_{xy}(\widetilde{z}) - m(\widetilde{z})\delta_{xy}|^2 \prec B_{xy}(\widetilde{z}), \quad \forall\, x, y \in \mathbb{Z}_L^d, \tag{5.6}$$

with $\widetilde{z} = E + \mathrm{i}\widetilde{\eta}$ for some $E \in (-2+\kappa, 2-\kappa)$ and $\widetilde{\eta} \in [W^2/L^{2-\varepsilon}, 1]$. Then we have that

$$\max_{x,x_0} \frac{1}{K^d} \sum_{y:|y-x_0|\leqslant K} \left(|G_{xy}(z)|^2 + |G_{yx}(z)|^2\right) \prec \left(\frac{\widetilde{\eta}}{\eta}\right)^2 \frac{1}{W^4 K^{d-4}}, \tag{5.7}$$

uniformly in $K \in [W, L/2]$ and $z = E + \mathrm{i}\eta$ with $W^2/L^{2-\varepsilon} \leqslant \eta \leqslant \widetilde{\eta}$. Moreover, for any constant $\varepsilon_0 \in (0, d/20)$, we have that

$$\|G(z) - m(z)\|_{\max} \prec W^{-d/2+\varepsilon_0}, \tag{5.8}$$

uniformly in $z = E + \mathrm{i}\eta$ with $\max\{W^{-\varepsilon_0}\widetilde{\eta}, W^2/L^{2-\varepsilon}\} \leqslant \eta \leqslant \widetilde{\eta}$.

Compared with (2.27), the ℓ^∞ bound (5.8) is sharp up to a factor W^{ε_0}. The estimate (5.7) is an averaged bound instead of an entrywise bound and the right-hand side of (5.7) loses an W^2/K^2 factor when compared with the sharp averaged bound $W^{-2}K^{-(d-2)}$. In our proof, we will need to bound terms of the form $\sum_x \Theta_{xy_1}|G_{xy_2}|$. Using (1.17) and (5.7), it is not hard to get the bound $\sum_x \Theta_{xy_1}|G_{xy_2}| \prec W^{-d/2}\widetilde{\eta}/\eta$ when $d \geqslant 8$ (cf. Claim 6.9). This is one key reason why we require $d \geqslant 8$ in Theorem 1.4 and Theorem 2.1.

In order to improve the weaker estimates (5.7) and (5.8) to the stronger local law (2.27), we use the following lemma, whose proof will be given in [64]. Note that (5.7) verifies the assumption (5.9) as long as we have $W^{-\varepsilon_0}\widetilde{\eta} \leqslant \eta \leqslant \widetilde{\eta}$.

Lemma 5.4 (Entrywise bound on T-variables, Lemma 7.4 of [64]). *Suppose the assumptions of Theorem 2.16 hold. Fix any $z = E + \mathrm{i}\eta$ with $E \in (-2+\kappa, 2-\kappa)$ and $\widetilde{\eta} \in [W^2/L^{2-\varepsilon}, 1]$. Suppose (5.8) and the following estimate hold:*

$$\max_{x,x_0} \frac{1}{K^d} \sum_{y:|y-x_0|\leqslant K} \left(|G_{xy}(z)|^2 + |G_{yx}(z)|^2\right) \prec \frac{W^{2\varepsilon_0}}{W^4 K^{d-4}}, \tag{5.9}$$

for all $K \in [W, L/2]$. As long as ε_0 is a sufficiently small constant (depending on n and c_0 in (5.4)), we have that

$$T_{xy}(z) \prec B_{xy}, \quad \forall\, x, y \in \mathbb{Z}_L^d. \tag{5.10}$$

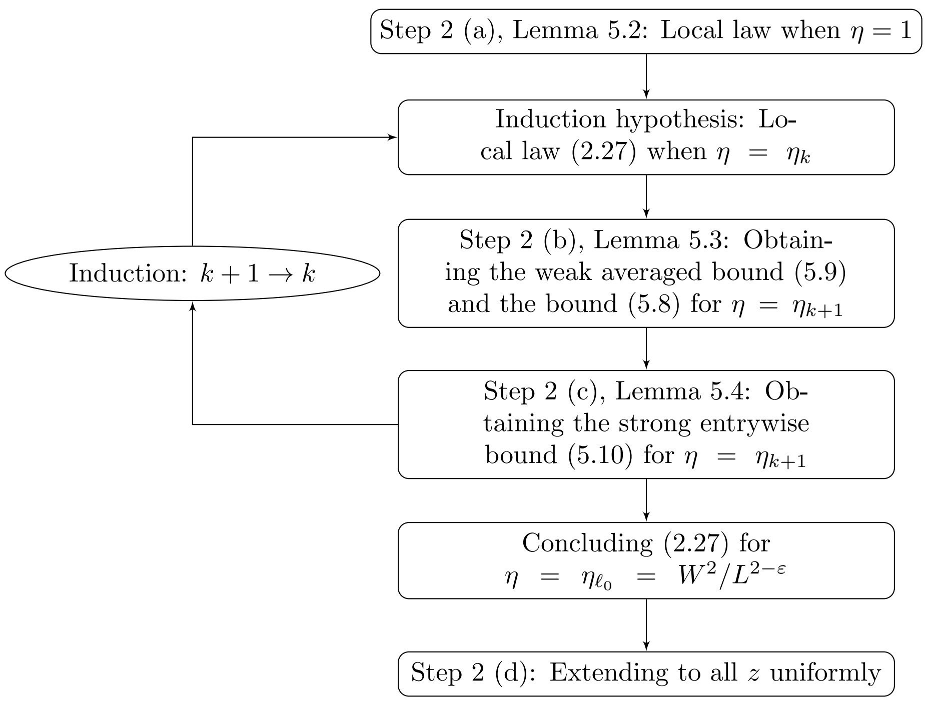


Figure 3: The structure of the proof of Theorem 2.16, where η_k is defined in (5.11).

Combining Lemmas 5.2–5.4, we can complete the proof of Theorem 2.16 using a bootstrapping argument on a sequence of multiplicatively decreasing η given below.

Proof of Theorem 2.16. Given a small constant $\varepsilon_0 > 0$ and a fixed $E \in (-2+\kappa, 2-\kappa)$, we define the following sequence of decreasing imaginary parts:

$$\eta_k := \max\left\{W^{-k\varepsilon_0},\ W^2/L^{2-\varepsilon}\right\}, \quad 0 \leqslant k \leqslant \ell_0, \tag{5.11}$$

where ℓ_0 is the smallest integer such that $W^{-\ell_0\varepsilon_0} \leqslant W^2/L^{2-\varepsilon}$. Note that by definition $\eta_{k+1} = W^{-\varepsilon_0}\eta_k$ for $k \leqslant \ell_0 - 2$ and we always have $\eta_{\ell_0} = W^2/L^{2-\varepsilon} \geqslant W^{-\varepsilon_0}\eta_{\ell_0-1}$. Then we prove Theorem 2.16 through an induction on k as illustrated in Figure 3. More precisely, we have the following procedure.

Step 2 (a): By Lemma 5.2, (2.27) holds for $z_0 = E + \mathrm{i}\eta_0$.

Step 2 (b): For any $0 \leqslant k \leqslant \ell_0 - 1$, suppose (2.27) holds for $z_k = E + \mathrm{i}\eta_k$. Then by Lemma 5.3, (5.8) and (5.9) hold for all $z = E + \mathrm{i}\eta$ with $\eta_{k+1} \leqslant \eta \leqslant \eta_k$.

Step 2 (c): Applying Lemma 5.4, we obtain that (5.10) holds for $z = z_{k+1}$. Using Lemma 5.1, we conclude (2.27) for $z = z_{k+1}$.

Repeating the above Steps 2 (b) and 2 (c) for ℓ_0 steps, we obtain that

(i) (2.27) holds for all z_k with $0 \leqslant k \leqslant \ell_0$;

(ii) (5.8) and (5.9) hold for all $z = E + \mathrm{i}\eta$ with $\eta_{\ell_0} \leqslant \eta \leqslant 1$.

To conclude Theorem 2.16, we still need to extend (2.27) uniformly to all $z = E+\mathrm{i}\eta$ with $E \in (-2+\kappa, 2-\kappa)$ and $\eta \in [\eta_{\ell_0}, 1]$.

Step 2 (d): For a fixed $E \in (-2+\kappa, 2-\kappa)$ and $0 \leqslant k \leqslant \ell_0 - 1$, we consider the following interpolations between z_k and z_{k+1}:

$$z_{k,j} = E + \mathrm{i}\eta_k - \mathrm{i}(\eta_k - \eta_{k+1}) \cdot jL^{-10d}, \quad j \in [\![0, L^{10d}]\!]. \tag{5.12}$$

By the above item (ii), (5.8) and (5.9) hold for all $z = z_{k,j}$, $j \in [\![0, L^{10d}]\!]$. Now applying Lemma 5.4, we obtain that (5.10) holds for all $z_{k,j}$:

$$T_{xy}(z_{k,j}) \prec B_{xy}(z_{k,j}), \quad \forall\, x, y \in \mathbb{Z}_L^d,\ j \in [\![0, L^{10d}]\!].$$

Using Lemma 5.1 and taking a union bound, we conclude that (2.27) holds uniformly for all $z_{k,j}$. Then using the simple resolvent identity

$$G_{xy}(E+\mathrm{i}\eta) = G_{xy}(E+\mathrm{i}\eta') + \mathrm{i}(\eta-\eta')\sum_\alpha G_{x\alpha}(E+\mathrm{i}\eta)G_{\alpha y}(E+\mathrm{i}\eta'), \tag{5.13}$$

and the trivial bound

$$\|G(E+\mathrm{i}\eta)\|_{\max} \leqslant \eta^{-1} \leqslant L^2/W^2, \quad \forall\, \eta \geqslant W^2/L^2,$$

we can easily obtain the perturbation estimate

$$\|G(z) - G(z_{k,j})\| \leqslant L^{-d}, \quad \forall\, z = E + \mathrm{i}\eta,\ \eta \in [\operatorname{Im} z_{k,j}, \operatorname{Im} z_{k,j-1}]. \tag{5.14}$$

Together with the local law (2.27) at $z_{k,j}$, (5.14) implies that (2.27) holds uniformly for all $z = E + \mathrm{i}\eta$ with $\eta \in [\eta_k, \eta_{k+1}]$. This concludes (2.27) for a fixed $E \in (-2+\kappa, 2-\kappa)$ and uniformly for all $\eta \in [\eta_{\ell_0}, 1]$. Finally, to extend (2.27) uniformly to all E, we choose an L^{-10d}-net of $(-2+\kappa, 2-\kappa)$ and use a similar perturbation argument as above. We omit the details. □

Now we give the proof of Lemma 5.3. We first recall the following classical Ward's identity. Its proof is a simple application of the spectral decomposition of $G(z)$.

Lemma 5.5 (Ward's identity)**.** *For any $y, y' \in \mathbb{Z}_L^d$ and $z = E + \mathrm{i}\eta$, we have*

$$\sum_x \overline{G}_{xy'}(z)G_{xy}(z) = \frac{G_{y'y}(z) - \overline{G_{yy'}(z)}}{2\mathrm{i}\eta}, \quad \sum_x \overline{G}_{y'x}(z)G_{yx}(z) = \frac{G_{yy'}(z) - \overline{G_{y'y}(z)}}{2\mathrm{i}\eta}. \tag{5.15}$$

As a special case, if $y = y'$, we have

$$\sum_x |G_{xy}(z)|^2 = \sum_x |G_{yx}(z)|^2 = \frac{\operatorname{Im} G_{yy}(z)}{\eta}. \tag{5.16}$$

Proof of Lemma 5.3. We first prove (5.7). The proof of (5.8) will be based on (5.7).

Proof of (5.7)**:** By (5.6), the following event is a high probability event:

$$\Xi := \left\{ \max_x |G_{xx}(\widetilde{z}) - m(\widetilde{z})| \leqslant W^{-d/4} \right\}.$$

With Ward's identity (5.16), we obtain that on Ξ,

$$\sum_x |G_{xy}(\widetilde{z})|^2 = \sum_x |G_{yx}(\widetilde{z})|^2 = \frac{\operatorname{Im} G_{yy}(\widetilde{z})}{\widetilde{\eta}} \sim \widetilde{\eta}^{-1}. \tag{5.17}$$

Moreover, using the inequality $\eta \operatorname{Im} G_{yy}(z) \leqslant \widetilde{\eta} \operatorname{Im} G_{yy}(\widetilde{z})$, we obtain that on Ξ,

$$\sum_x |G_{xy}(z)|^2 = \sum_x |G_{yx}(z)|^2 = \frac{\operatorname{Im} G_{yy}(z)}{\eta} \leqslant \frac{\widetilde{\eta} \operatorname{Im} G_{yy}(\widetilde{z})}{\eta^2} \lesssim \frac{\widetilde{\eta}}{\eta^2}. \tag{5.18}$$

Now we define the family of vectors $\mathbf{v}_x$, $\mathbf{w}_x$, $\widehat{\mathbf{v}}_x$ and $\widehat{\mathbf{w}}_x$ as

$$v_x(y) := G_{xy}(z), \quad \widehat{\mathbf{v}}_x := \frac{\mathbf{v}_x}{\|\mathbf{v}_x\|_2}, \quad w_x(y) := \overline{G}_{yx}(\widetilde{z}), \quad \widehat{\mathbf{w}}_x := \frac{\mathbf{w}_x}{\|\mathbf{w}_x\|_2}.$$

By (5.17) and (5.18), we have that on Ξ,

$$\|\mathbf{v}_x\|_2^2 \lesssim \widetilde{\eta}/\eta^2, \quad \|\mathbf{w}_x\|_2^2 \sim \widetilde{\eta}^{-1}, \quad \forall\, x \in \mathbb{Z}_L^d. \tag{5.19}$$

Now let $\mathcal{I}$ be any subset of indices. Suppose the orthogonal projection of $\widehat{\mathbf{v}}_x$ onto the subspace spanned by $\{\widehat{\mathbf{w}}_y : y \in \mathcal{I}\}$ can be written as $\mathbf{u}_x = \sum_{y\in\mathcal{I}} a_x(y)\widehat{\mathbf{w}}_y$. Then we have

$$b_x(y) := (\widehat{\mathbf{v}}_x, \widehat{\mathbf{w}}_y) = (\mathbf{u}_x, \widehat{\mathbf{w}}_y) = \sum_{y'\in\mathcal{I}} a_x(y') A_{y'y}. \tag{5.20}$$

Here the inner product is defined as $(\mathbf{v}, \mathbf{w}) := \sum_x \mathbf{v}(x)\overline{\mathbf{w}}(x)$, and the matrix A is defined by

$$A_{y'y} := (\widehat{\mathbf{w}}_{y'}, \widehat{\mathbf{w}}_y) = \frac{\sum_x \overline{G}_{xy'}(\widetilde{z}) G_{xy}(\widetilde{z})}{\|\mathbf{w}_y\|_2 \|\mathbf{w}_{y'}\|_2} = \frac{G_{y'y}(\widetilde{z}) - \overline{G_{yy'}(\widetilde{z})}}{2\mathrm{i}\widetilde{\eta}\|\mathbf{w}_y\|_2 \|\mathbf{w}_{y'}\|_2}, \tag{5.21}$$

where we used (5.15) in the third step. Notice that by definition, $A \equiv A(\mathcal{I})$ is a positive definite Hermitian matrix with indices in $\mathcal{I}$. We define two *row vectors* $\mathbf{a}_x := (a_x(y))_{y\in\mathcal{I}}$ and $\mathbf{b}_x := (b_x(y))_{y\in\mathcal{I}}$. Then (5.20) gives that $\mathbf{a}_x = \mathbf{b}_x A^{-1}$, with which we can get that

$$1 = \|\widehat{\mathbf{v}}_x\|_2^2 \geqslant \|\mathbf{u}_x\|_2^2 = \mathbf{a}_x A \mathbf{a}_x^* = \mathbf{b}_x A^{-1} \mathbf{b}_x^* \geqslant \|\mathbf{b}_x\|^2 \|A\|_{\ell^2\to\ell^2}^{-1}.$$

This inequality implies that on Ξ,

$$\|\mathbf{b}_x\|^2 \leqslant \|A\|_{\ell^2\to\ell^2} \lesssim \|\mathcal{A}\|_{\ell^2\to\ell^2}, \tag{5.22}$$

where the matrix $\mathcal{A}$ is defined by

$$\mathcal{A}_{y'y} = \frac{1}{2\mathrm{i}} \left[G_{y'y}(\widetilde{z}) - \overline{G_{yy'}(\widetilde{z})} \right].$$

On the other hand, we have the resolvent identity

$$\begin{aligned} G_{xy}(z) = G_{xy}(\widetilde{z}) - \mathrm{i}(\widetilde{\eta} - \eta) \sum_\alpha G_{x\alpha}(z) G_{\alpha y}(\widetilde{z}) &= \overline{w}_y(x) - \mathrm{i}(\widetilde{\eta} - \eta)(\widehat{\mathbf{v}}_x, \widehat{\mathbf{w}}_y)\|\mathbf{v}_x\|_2 \|\mathbf{w}_y\|_2 \\ &= \overline{w}_y(x) - \mathrm{i}(\widetilde{\eta} - \eta) b_x(y) \|\mathbf{v}_x\|_2 \|\mathbf{w}_y\|_2. \end{aligned} \tag{5.23}$$

With this identity, we obtain that on Ξ,

$$\sum_{y\in\mathcal{I}} |G_{xy}(z)|^2 \lesssim \sum_{y\in\mathcal{I}} |G_{xy}(\widetilde{z})|^2 + \left(\frac{\widetilde{\eta}}{\eta}\right)^2 \sum_{y\in\mathcal{I}} |b_x(y)|^2 \lesssim \sum_{y\in\mathcal{I}} |G_{xy}(\widetilde{z})|^2 + \left(\frac{\widetilde{\eta}}{\eta}\right)^2 \|\mathcal{A}\|_{\ell^2\to\ell^2}, \tag{5.24}$$

where we used (5.19) in the first step and (5.22) in the second step, Similarly, we can get that on Ξ,

$$\sum_{y\in\mathcal{I}} |G_{yx}(z)|^2 \lesssim \sum_{y\in\mathcal{I}} |G_{yx}(\widetilde{z})|^2 + \left(\frac{\widetilde{\eta}}{\eta}\right)^2 \|\mathcal{A}\|_{\ell^2\to\ell^2}. \tag{5.25}$$

For the specific index set $\mathcal{I} = \{y : |y - x_0| \leqslant K\}$, using (5.6) we can bound that

$$\sum_{y \in \mathcal{I}} \left(|G_{xy}(\widetilde{z})|^2 + |G_{yx}(\widetilde{z})|^2\right) \prec 1 + \sum_{y\in\mathcal{I}} B_{xy} \lesssim \frac{K^2}{W^2}. \tag{5.26}$$

It remains to bound $\|\mathcal{A}\|_{\ell^2\to\ell^2}$ in (5.24) and (5.25). A simple bound can be obtained by using the Hilbert-Schmidt norm:

$$\|\mathcal{A}\|^2_{\ell^2\to\ell^2} \leqslant \|\mathcal{A}\|^2_{HS} \lesssim \sum_{y,y'\in\mathcal{I}} \left(|G_{y'y}(\widetilde{z})|^2 + |G_{yy'}(\widetilde{z})|^2\right) \prec |\mathcal{I}| + \sum_{y,y'\in\mathcal{I}} B_{yy'} \lesssim \frac{K^{d+2}}{W^2}.$$

This estimate is not strong enough to give the bound (5.7). To obtain a better bound on $\|\mathcal{A}\|_{\ell^2\to\ell^2}$, we use the following lemma, which is based on the classical method of moments. The proof of this lemma will be given in [64].

Lemma 5.6 (Lemma 9.1 of [64])**.** *Suppose the assumptions of Lemma 5.3 hold. We choose the index set $\mathcal{I} = \{y : |y - x_0| \leqslant K\}$ for $K \in [W, L/2]$. Then for any fixed $p \in \mathbb{N}$ and small constant $\varepsilon > 0$, we have the estimate*

$$\mathbb{E}\,\mathrm{Tr}\left(\mathcal{A}^{2p}\right) \leqslant K^d \left(W^{\varepsilon}\frac{K^4}{W^4}\right)^{2p-1}. \tag{5.27}$$

With Lemma 5.6, we obtain that

$$\mathbb{E}\|\mathcal{A}\|^{2p}_{\ell^2\to\ell^2} \leqslant \mathbb{E}\,\mathrm{Tr}\left(\mathcal{A}^{2p}\right) \leqslant K^d \left(W^{\varepsilon}\frac{K^4}{W^4}\right)^{2p-1}.$$

Since p can be arbitrarily large, using Markov's inequality we get that

$$\|\mathcal{A}\|_{\ell^2\to\ell^2} \prec \frac{K^4}{W^4}. \tag{5.28}$$

Inserting (5.26) and (5.28) into (5.24) and (5.25), we obtain (5.7).

Proof of (5.8)**:** To prove (5.8), as in (5.12), we define the following interpolations between z and $\widetilde{z}$:

$$z_j = E + \mathrm{i}\widetilde{\eta} - \mathrm{i}(\widetilde{\eta} - \eta)\cdot jL^{-10d}, \quad j \in [\![0, L^{10d}]\!]. \tag{5.29}$$

By (5.14), we have the perturbation estimate

$$\|G(z_j) - G(z_{j-1})\| \leqslant L^{-d}. \tag{5.30}$$

Moreover, by (5.7) and the fact $\widetilde{\eta}/\eta \leqslant W^{\varepsilon_0}$, we know that (5.9) holds for $z = z_j$ for all $j \in [\![0, L^{10d}]\!]$. Taking a union bound, we get that for any small constant $\tau > 0$ and large constant $D > 0$, the event

$$\begin{aligned}\Xi_0 := &\left\{|G_{xy}(\widetilde{z}) - m(\widetilde{z})\delta_{xy}| \leqslant W^{\tau}B_{xy},\ \forall\, x,y \in \mathbb{Z}_L^d\right\}\\ &\cap\left\{\max_{0\leqslant j\leqslant L^{10d}} \sum_{K\in[W,L/2]} \max_{x,x_0\in\mathbb{Z}_L^d} \frac{W^4}{K^4} \sum_{|y-x_0|\leqslant K} \left(|G_{xy}(z_j)|^2 + |G_{yx}(z_j)|^2\right) \leqslant W^{2\varepsilon_0+\tau}\right\}\end{aligned} \tag{5.31}$$

holds with probability $\mathbb{P}(\Xi_0) \geqslant 1 - L^{-D}$. Now fix a small constant $\delta_0 \in (0, d/20)$, we define the events

$$A_j := \{\|G(z_j) - m(z_j)\|_{\max} \leqslant W^{-\delta_0}\}, \quad B_j := \{\|T(z_j)\|_{\max} \leqslant W^{-d+2\varepsilon_0+3\tau}\}, \quad j \in [\![0, L^{10d}]\!].$$

By Lemma 5.1, we have that

$$\mathbb{P}\left(\|G(z_j) - m(z_j)\|_{\max} \geqslant W^{-d/2+\varepsilon_0+2\tau}; A_j \cap B_j\right) \leqslant L^{-D}. \tag{5.32}$$

Now we use the above facts to prove that

$$\mathbb{P}\Big(\max_{0\leqslant j\leqslant L^{10d}} \|G(z_j) - m(z_j)\| \leqslant W^{-d/2+\varepsilon_0+2\tau}\Big) \geqslant 1 - 2L^{-D+10d}, \tag{5.33}$$

which concludes (5.8) since τ and D are arbitrary.

Using (5.30), we can obtain that $\|G(z_1)-m(z_1)\|_{\max}\leqslant W^{-d/4}$ on Ξ_0, which gives $\Xi_0\subset A_1$. Moreover, on Ξ_0, we have that for any $x,y\in\mathbb{Z}_L^d$,

$$T_{xy}(z_1)=|m(z_1)|^2\sum_{\alpha}s_{x\alpha}|G_{\alpha y}(z_1)|^2\lesssim s_{xy}+W^{-d}\sum_{|\alpha-y|\leqslant W^{1+\frac{\tau}{4}}}|G_{\alpha y}(z_1)|^2+\mathrm{O}(W^{-100d})\lesssim\frac{W^{2\varepsilon_0+2\tau}}{W^d},$$

where in the second step we used (1.8) and in the third step we used the averaged bound in the definition of Ξ_0. This estimate gives that $\Xi_0\subset A_1\cap B_1$. Then (5.32) implies that $\mathbb{P}(\Xi_1)\geqslant\mathbb{P}(\Xi_0)-L^{-D}\geqslant 1-2L^{-D}$, where the event Ξ_1 is defined by

$$\Xi_1:=\Xi_0\cap\left\{\|G_{xy}(z_1)-m(z_1)\|_{\max}\leqslant W^{-d/2+\varepsilon_0+2\tau}\right\}.$$

Repeating the above argument, for any $j\in[\![0,L^{10d}]\!]$, we can obtain that $\mathbb{P}(\Xi_j)\geqslant 1-(j+1)\cdot L^{-D}$ for

$$\Xi_j:=\Xi_0\cap\left\{\max_{0\leqslant k\leqslant j}\|G_{xy}(z_k)-m(z_k)\|_{\max}\leqslant W^{-d/2+\varepsilon_0+2\tau}\right\}.$$

Taking $j=L^{10d}$, we conclude (5.33). □

5.3 Step 3: n-th order T-equation

In this step, we construct the n-th order T-equation in Lemma 5.7, whose proof will be postponed to [64]. In general, it is difficult to define the T-equation explicitly (there are already hundreds of terms when $n=6$). Instead, we will give a prescription to generate the T-equation by applying local and global expansions. Section 9 contains some more explanations.

Lemma 5.7 (n-th order T-equation, Theorem 3.7 of [64])**.** *Fix any $n\in\mathbb{N}$. Suppose we have defined the $(n-1)$-th order T-expansion. Then we can construct an n-th order T-equation satisfying Definition 2.17.*

In Step 5, we will solve the n-th order T-equation (2.28) to get the n-th order T-expansion (2.19). Before doing that, we need to show that $\mathcal{E}_n$ satisfies the properties (1.25)–(1.27) and (2.13)–(2.16), i.e. $\mathcal{E}_n$ is indeed an n-th order self-energy. This is the purpose of Step 4, where the proof of the sum zero properties (1.27) and (2.16) will be the core argument.

5.4 Step 4: Proving the sum zero properties

In this subsection, we prove that $\mathcal{E}_n$ constructed in Lemma 5.7 is indeed a self-energy.

Lemma 5.8 (Properties of $\mathcal{E}_n$)**.** *Fix any $n\in\mathbb{N}$. Suppose we have defined the $(n-1)$-th order T-expansion. The deterministic matrix $\mathcal{E}_n$ constructed in the n-th order T-equation in Lemma 5.7 satisfies the properties* (1.25)–(1.27) *and* (2.13)–(2.16) *with $l=n$.*

The proof of Lemma 5.8 is based on three main ingredients, Lemmas 5.9, 5.10 and 5.11. In Lemma 5.9, we show that the estimates (1.26), (2.14) and (2.15) hold. Its proof depends on the doubly connected property of $\mathcal{E}_n$ (cf. Definition 6.5) and we postpone it to Section 7.

Lemma 5.9. *Under the assumptions of Theorem 2.1 and Lemma 5.8, $\mathcal{E}_n^\infty$ exists. Moreover, $\mathcal{E}_n$ and $\mathcal{E}_n^\infty$ satisfy* (1.26), (2.14), (2.15) *and*

$$\left|\sum_{\mathfrak{a}\in\mathbb{Z}_L^d}(\mathcal{E}_n)_{0\mathfrak{a}}(m(z),\psi,W,L)-\sum_{\mathfrak{a}\in\mathbb{Z}^d}(\mathcal{E}_n^\infty)_{0\mathfrak{a}}(m(E),\psi,W)\right|\leqslant\eta W^{-(n-2)d/2+\varepsilon},\quad\forall\,\eta\in[W^2/L^{2-\varepsilon},L^{-\varepsilon}],\quad(5.34)$$

for any small constant $\varepsilon>0$. Here we have abbreviated $m(E)\equiv m(E+\mathrm{i}0_+)$.

By taking $L\to\infty$, the infinite space limit $\sum_{\mathfrak{a}}(\mathcal{E}_n^\infty)_{0\mathfrak{a}}$ depends only on n, $m(E)$, the function ψ in Assumption 1.2, and the band width W. Now using a standard calculation with Fourier transforms, we show that the W dependence can be pulled out as a scaling factor if ψ is compactly supported.

Lemma 5.10. *Fix any $n \in \mathbb{N}$. Under the assumptions of Theorem 2.1 and Lemma 5.8, we have that*

$$\sum_{\mathfrak{a}\in\mathbb{Z}^d} (\mathcal{E}_n^{\infty})_{0\mathfrak{a}}(m(E), \psi, W) = W^{-(n-2)d/2}\mathfrak{S}_n(m(E), \psi), \tag{5.35}$$

where $\mathfrak{S}_n$ is a constant independent of W.

The full proof of Lemma 5.10 will be given in Section 7.2. Here we give a sketch of the proof.

Sketch of the proof of Lemma 5.10. The proof is straightforward if we replace the variance profile $f_{W,L}$ in (1.6) with an exact Fourier series on $\mathbb{Z}_L^d$:

$$\widetilde{f}_{W,L}(x) := \frac{1}{L^d}\sum_{p\in\mathbb{T}_L^d} \psi(Wp)e^{\mathrm{i}p\cdot x}, \quad \text{with} \quad \mathbb{T}_L^d = \left(\frac{2\pi}{L}\mathbb{Z}_L\right)^d. \tag{5.36}$$

We define the matrix $\widetilde{S} = (\widetilde{s}_{xy})$ with entries $\widetilde{s}_{xy} = \widetilde{f}_{W,L}([x-y]_L)$ and

$$\widetilde{S}^+(z) := \frac{m^2(z)\widetilde{S}}{1-m^2(z)\widetilde{S}}, \quad \widetilde{S}^-(z) := \overline{\widetilde{S}^+(z)}, \quad \widetilde{\Theta}(z) := \frac{|m(z)|^2\widetilde{S}}{1-|m(z)|^2\widetilde{S}}. \tag{5.37}$$

Then their entries can be expressed as

$$\widetilde{S}^+_{xy}(z) = \frac{1}{L^d}\sum_{p\in\mathbb{T}_L^d} \frac{m^2(z)\psi(Wp)}{1-m^2(z)\psi(Wp)}e^{\mathrm{i}p\cdot(x-y)}, \quad \widetilde{\Theta}_{xy} = \frac{1}{L^d}\sum_{p\in\mathbb{T}_L^d} \frac{|m(z)|^2\psi(Wp)}{1-|m(z)|^2\psi(Wp)}e^{\mathrm{i}p\cdot(x-y)}. \tag{5.38}$$

By replacing S, $S^{\pm}$ and Θ with $\widetilde{S}$, $\widetilde{S}^{\pm}$ and $\widetilde{\Theta}$ in $\mathcal{E}_n$, we get a new matrix $\widetilde{\mathcal{E}}_n$. In Section 7.2, we will show that $\sum_{\mathfrak{a}}(\mathcal{E}_n)_{0\mathfrak{a}}$ has the same infinite space limit as $\sum_{\mathfrak{a}}(\widetilde{\mathcal{E}}_n)_{0\mathfrak{a}}$. Let $\mathcal{G}$ denote the graphs in $\widetilde{\mathcal{E}}_n$, p_e denote the momentum associated with each edge e in $\mathcal{G}$, Ξ_L be a subset of $(\mathbb{T}_L^d)^{n_e}$ given by the constraint that the total momentum at each vertex is equal to 0, where $n_e \equiv n_e(\mathcal{G})$ is the total number of edges in $\mathcal{G}$. Then using the Fourier series (5.36) and (5.38), we can write that

$$\sum_{\mathfrak{a}}(\widetilde{\mathcal{E}}_n)_{0\mathfrak{a}}(m(z),\psi,W,L) = \frac{1}{L^{(n-2)d/2}}\sum_{\mathcal{G}}\sum_{\{p_e\}\in\Xi_L}\mathcal{F}_{\mathcal{G}}\left(\{Wp_e\},z\right),$$

where $\mathcal{F}_{\mathcal{G}}$ is a function expressed in terms of $\psi(Wp_e)$ (cf. Section 7.2 for more details). Taking $L\to\infty$ and $\eta\to 0$, we get that

$$\sum_{\mathfrak{a}}(\mathcal{E}_n^{\infty})_{0\mathfrak{a}}(m(E),\psi,W) = \frac{1}{(2\pi)^{(n-2)d/2}}\sum_{\mathcal{G}}\int_{\{p_e\}\in\Xi}\mathcal{F}_{\mathcal{G}}\left(\{Wp_e\},E\right)\prod_e \mathrm{d}p_e,$$

where Ξ is a union of hyperplanes in the torus $(-\pi,\pi]^{dn_e}$ with the constraint that the total momentum at each vertex is equal to 0. Then applying a change of variables $q_e = Wp_e$ and using that ψ is compactly supported, we obtain that

$$\sum_{\mathfrak{a}}(\mathcal{E}_n^{\infty})_{0\mathfrak{a}}(m(E),\psi,W) = \frac{1}{(2\pi)^{(n-2)d/2}W^{(n-2)d/2}}\sum_{\mathcal{G}}\int_{\widetilde{\Xi}}\mathcal{F}_{\mathcal{G}}\left(\{q_e\},E\right)\prod_e \mathrm{d}q_e,$$

where $\widetilde{\Xi}$ is a union of hyperplanes in $(\mathbb{R}^d)^{n_e}$ given by the constraints of Ξ. Renaming the right-hand side, we conclude (5.35). ☐

In Lemma 5.11, we show that the row sum $\sum_{\mathfrak{a}}(\mathcal{E}_n)_{0\mathfrak{a}}$ is much smaller than $W^{-(n-2)d/2}$ under some particular choices of L and z.

Lemma 5.11. *Fix any $n\in\mathbb{N}$. Under the assumptions of Theorem 2.1 and Lemma 5.8, suppose $L\equiv L_n$ satisfies that*

$$W^{(n-3)d/2+c_0} \leqslant L_n^2/W^2 \leqslant W^{(n-2)d/2-c_0} \tag{5.39}$$

for a constant $c_0>0$, and $\eta\equiv\eta_n = W^{2+\varepsilon}/L_n^2$ for a small enough constant $\varepsilon>0$. Then for $z_n = E+\mathrm{i}\eta_n$,

$$\Big|\sum_{\mathfrak{a}}(\mathcal{E}_n)_{0\mathfrak{a}}\left(m(z_n),\psi,W,L_n\right)\Big| \leqslant W^{-(n-2)d/2}\cdot W^{-c} \tag{5.40}$$

for a constant $c>0$ depending only on c_0 and d.

Comparing (5.40) with (5.35), we see that $\sum_{\mathfrak{a}}(\mathcal{E}_n)_{0\mathfrak{a}}$ is much smaller than its scaling size $W^{-(n-2)d/2}$ if $\mathfrak{S}_n \neq 0$. We will use this contradiction to show that $\mathfrak{S}_n = 0$, and hence conclude the sum zero property (2.16) for $\mathcal{E}_n^\infty$. To prove Lemma 5.11, we need to use the following lemma, whose proof will be postponed to Section 6.5. The proof is based on some additional properties of $\mathcal{R}_{IT}^{(n)}$, $\mathcal{A}_{IT}^{(>n)}$ and $\mathcal{E}rr'_{n,D}$ that will be introduced in Definition 6.7 below (more precisely, their doubly connected properties that will be defined in Definition 6.5).

Lemma 5.12. *Fix any $n \in \mathbb{N}$. Under the assumptions of Lemma 5.11, we have the following estimates:*

$$\sum_{\mathfrak{a},\mathfrak{b}} \left|(\mathcal{R}_{IT,k})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}(z_n,\psi,W,L_n)\right| \prec L_n^d \cdot \eta_n^{-1} W^{-(k-2)d/2}, \quad 3 \leqslant k \leqslant n; \tag{5.41}$$

$$\sum_{\mathfrak{a},\mathfrak{b}} \left|(\mathcal{A}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}(z_n,\psi,W,L_n)\right| \prec L_n^d \cdot \eta_n^{-2} W^{-(n-1)d/2}; \tag{5.42}$$

$$\sum_{\mathfrak{a},\mathfrak{b}} \left|(\mathcal{E}rr'_{n,D})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}(z_n,\psi,W,L_n)\right| \prec L_n^d \cdot \eta_n^{-2} W^{-(D-1)d/2}. \tag{5.43}$$

In the proof of Lemma 5.11, we will use these estimates to control

$$\Big|\sum_{\mathfrak{a},\mathfrak{b}}\sum_x (\Theta\mathcal{E}_n)_{\mathfrak{a}x} T_{x,\mathfrak{b}\mathfrak{b}}\Big| \sim \frac{L_n^d}{\eta_n^2}\Big|\sum_{\mathfrak{a}}(\mathcal{E}_n)_{0\mathfrak{a}}\Big|.$$

Compared to the scaling size $L_n^d\eta_n^{-2}W^{-(n-2)d/2}$ of the right-hand side given by (5.35), (5.41) gains a factor $W^{(n-3)d/2}\eta_n$ when $k=3$, (5.42) gains a factor $W^{-d/2}$, and (5.43) is negligible because D is arbitrarily large.

Proof of Lemma 5.11. In the setting of Lemma 5.11, we have the $(n-1)$-th order T-expansion. Hence by Theorem 2.16, the local law (2.27) holds for $G(z_n) \equiv G(z_n,\psi,W,L_n)$ if L_n satisfies $L_n^2/W^2 \leqslant W^{(n-2)d/2-c_0}$. This explains the upper bound in (5.39).

Now given an n-th order T-equation (2.28) with $\mathfrak{b}_1 = \mathfrak{b}_2 = \mathfrak{b}$, taking the expectation of both sides and summing over $\mathfrak{a},\mathfrak{b} \in \mathbb{Z}_L^d$, we obtain that for large enough $D > 0$,

$$\begin{aligned}\sum_{\mathfrak{a},\mathfrak{b}} \mathbb{E}T_{\mathfrak{a}\mathfrak{b}}(z_n,L_n) = m(z_n)\sum_{\mathfrak{a},\mathfrak{b}}\mathbb{E}\overline{G}_{\mathfrak{b}\mathfrak{b}}(z_n,L_n)\Theta_{\mathfrak{a}\mathfrak{b}}(z_n,L_n) + \Big[\sum_{\mathfrak{a},x}(\Theta\Sigma^{(n)})_{\mathfrak{a}x}\sum_{\mathfrak{b}}\mathbb{E}T_{x\mathfrak{b}}\Big](z_n,L_n)\\ + \sum_{\mathfrak{a},\mathfrak{b}}\mathbb{E}(\mathcal{R}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}(z_n,L_n) + \sum_{\mathfrak{a},\mathfrak{b}}\mathbb{E}(\mathcal{A}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}(z_n,L_n) + \mathrm{O}(W^{-D}),\end{aligned} \tag{5.44}$$

where we used $\mathbb{E}\mathcal{Q}_{IT}^{(n)} = 0$ and (5.43). For simplicity, we have omitted the arguments ψ and W from the above equation. To further simplify the notation, we will also omit the arguments z_n and L_n in the following proof.

For the left-hand side of (5.44), using Ward's identity (5.16) we get that

$$\sum_{\mathfrak{a},\mathfrak{b}}\mathbb{E}T_{\mathfrak{a}\mathfrak{b}} = |m|^2\sum_{\mathfrak{a},x}s_{\mathfrak{a}x}\cdot\mathbb{E}\sum_{\mathfrak{b}}|G_{x\mathfrak{b}}|^2 = |m|^2\frac{\sum_x \mathrm{Im}\,(\mathbb{E}G_{xx})}{\eta_n} = L_n^d|m|^2\frac{\mathrm{Im}\,(\mathbb{E}G_{00})}{\eta_n}, \tag{5.45}$$

where in the last step we used $\mathbb{E}G_{xx} = \mathbb{E}G_{00}$ for all $x \in \mathbb{Z}_{L_n}^d$ by translational invariance. For the first term on the right-hand side of (5.44), using the identity in (1.42) we obtain that

$$m\sum_{\mathfrak{a},\mathfrak{b}}\mathbb{E}\overline{G}_{\mathfrak{b}\mathfrak{b}}\Theta_{\mathfrak{a}\mathfrak{b}} = m\mathbb{E}\overline{G}_{00}\cdot\sum_{\mathfrak{a},\mathfrak{b}}\Theta_{\mathfrak{a}\mathfrak{b}} = L_n^d\frac{|m|^2 m\cdot\mathbb{E}\overline{G}_{00}}{1-|m|^2}. \tag{5.46}$$

For the third term on the right-hand side of (5.44), using (5.41) we obtain that

$$\sum_{\mathfrak{a},\mathfrak{b}}\left|\mathbb{E}(\mathcal{R}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}\right| \leqslant \sum_{k=3}^n\sum_{\mathfrak{a},\mathfrak{b}}\left|\mathbb{E}(\mathcal{R}_{IT,k})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}\right| \leqslant L_n^d\frac{W^{-d/2+\varepsilon}}{\eta_n}. \tag{5.47}$$

For the fourth term on the right-hand side of (5.44), using (5.42) we obtain that

$$\sum_{\mathfrak{a},\mathfrak{b}} \left|\mathbb{E}(\mathcal{A}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}\right| \leqslant L_n^d \frac{W^{-(n-1)d/2+\varepsilon}}{\eta_n^2}. \tag{5.48}$$

Finally, for the second term on the right-hand side of (5.44), we decompose $\Sigma^{(n)}$ as (2.31). For $4 \leqslant l \leqslant n$, we can calculate that

$$\sum_{\mathfrak{a},x} (\Theta\mathcal{E}_l)_{\mathfrak{a}x} \sum_{\mathfrak{b}} \mathbb{E}T_{x\mathfrak{b}} = \frac{|m|^2}{1-|m|^2} \sum_{\alpha,x} (\mathcal{E}_l)_{\alpha x} \sum_{\mathfrak{b}} \mathbb{E}T_{x\mathfrak{b}} = L_n^d \frac{|m|^4 \cdot \operatorname{Im}(\mathbb{E}G_{00})}{(1-|m|^2)\eta_n} \sum_{\alpha} (\mathcal{E}_l)_{0\alpha}, \tag{5.49}$$

where in the first step we used (1.42) and in the second step we used the translational invariance of $\mathcal{E}_{\alpha x}$ and (5.45). Applying (1.27) to $\mathcal{E}_l$, $4 \leqslant l \leqslant n-1$, we get that

$$\Big|\sum_{\alpha} (\mathcal{E}_l)_{0\alpha}\Big| \leqslant \eta_n W^{-(l-2)d/2+\varepsilon}, \quad 4 \leqslant l \leqslant n-1.$$

Inserting it into (5.49) and using $1-|m|^2 \sim \eta_n$, we obtain that

$$\Big|\sum_{\mathfrak{a},x} (\Theta\mathcal{E}_l)_{\mathfrak{a}x} \sum_{\mathfrak{b}} \mathbb{E}T_{x\mathfrak{b}}\Big| \lesssim L_n^d \frac{W^{-(l-2)d/2+\varepsilon} \operatorname{Im}(\mathbb{E}G_{00})}{\eta_n}, \quad 4 \leqslant l \leqslant n-1. \tag{5.50}$$

Now plugging (5.45)–(5.50) into (5.44) and cancelling the L_n^d factor on both sides, we obtain that

$$\begin{aligned} |m|^2 \frac{\operatorname{Im}(\mathbb{E}G_{00})}{\eta_n} &= \frac{|m|^2 m \cdot \mathbb{E}\overline{G}_{00}}{1-|m|^2} + \frac{|m|^4 \cdot \operatorname{Im}(\mathbb{E}G_{00})}{(1-|m|^2)\eta_n} \sum_{\alpha} (\mathcal{E}_n)_{0\alpha} \\ &\quad + \mathrm{O}\left(\frac{W^{-d/2+\varepsilon}}{\eta_n} + \frac{W^{-(n-1)d/2+\varepsilon}}{\eta_n^2} + \sum_{l=4}^{n-1} \frac{W^{-(l-2)d/2+\varepsilon} \operatorname{Im}(\mathbb{E}G_{00})}{\eta_n}\right). \end{aligned} \tag{5.51}$$

Since (2.27) holds for $G(z_n, L_n)$, we have that

$$\mathbb{E}G_{00}(z_n, L_n) = m(z_n) + \mathrm{O}(W^{-d/2+\varepsilon}). \tag{5.52}$$

Moreover, taking the imaginary part of the equation $z_n = -m(z_n) - m^{-1}(z_n)$, we obtain that

$$\frac{|m(z_n)|^2}{1-|m(z_n)|^2} = \frac{\operatorname{Im} m(z_n)}{\eta_n}. \tag{5.53}$$

Inserting (5.52) and (5.53) into (5.51) and using $1-|m|^2 \sim \eta_n$, we get that

$$\Big|\sum_{\alpha} (\mathcal{E}_n)_{0\alpha}(m(z_n), \psi, W, L_n)\Big| \lesssim W^{-d/2+\varepsilon}\eta_n + W^{-(n-1)d/2+\varepsilon} = W^{-d/2+\varepsilon}\frac{W^{2+\varepsilon}}{L_n^2} + W^{-(n-1)d/2+\varepsilon}.$$

Together with the lower bound in condition (5.39), we conclude (5.40) for $c = \min(d/2-\varepsilon, c_0 - 2\varepsilon)$. ☐

Finally, combining Lemmas 5.9–5.11, we can conclude Lemma 5.8.

Proof of Lemma 5.8. The estimates (1.26), (2.14) and (2.15) for $\mathcal{E}_n$ follow from Lemma 5.9, and the equations (1.25) and (2.13) follow from Lemma A.1. Now we pick $L_n = W^{1+(n/4-5/8)d}$, which satisfies the condition (5.39) with $c_0 = d/4$. Applying Lemma 5.9 and Lemma 5.10 with $L = L_n$ and $z = z_n$ defined in Lemma 5.11, we obtain that

$$\Big|\sum_{\mathfrak{a}} (\mathcal{E}_n)_{0\mathfrak{a}}(m(z_n), \psi, W, L_n)\Big| \geqslant W^{-(n-2)d/2} \left|\mathfrak{S}_n(m(E), \psi)\right| - \eta_n W^{-(n-2)d/2+\varepsilon}. \tag{5.54}$$

Combining (5.40) and (5.54), we obtain that

$$|\mathfrak{S}_n(m(E), \psi)| \leqslant \eta_n W^{\varepsilon} + W^{-c} = \mathrm{o}(1).$$

Since $\mathfrak{S}_n$ is a constant, we must have $\mathfrak{S}_n(m(E), \psi) = 0$, which by Lemma 5.10 implies (2.16) for $\mathcal{E}_n^\infty$. Combining (2.16) for $\mathcal{E}_n^\infty$ with Lemma 5.9, we obtain (1.27) for $\mathcal{E}_n$. ☐

5.5 Step 5: The n-th order T-expansion

After showing that $\mathcal{E}_n$ is a self-energy satisfying Definition 2.13, we can now solve the n-th order T-equation to obtain the n-th order T-expansion.

Lemma 5.13 (n-th order T-expansion)**.** *Given the n-th order T-equation constructed in Lemma 5.7, if $\mathcal{E}_n$ satisfies* (1.25)–(1.27) *and* (2.13)–(2.16)*, then we can construct an n-th order T-expansion satisfying Definition 2.15.*

Proof. By property (iii) of Definition 2.17, we can write that

$$(\mathcal{R}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = \sum_x \Theta_{\mathfrak{a}x}(\Gamma_R^{(n)})_{x,\mathfrak{b}_1\mathfrak{b}_2}, \quad (\mathcal{A}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = \sum_x \Theta_{\mathfrak{a}x}(\Gamma_A^{(>n)})_{x,\mathfrak{b}_1\mathfrak{b}_2},$$
$$(\mathcal{Q}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = \sum_x \Theta_{\mathfrak{a}x}(\Gamma_Q^{(n)})_{x,\mathfrak{b}_1\mathfrak{b}_2}, \quad (\mathcal{E}rr'_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = \sum_x \Theta_{\mathfrak{a}x}(\Gamma_{err}^{(n,D)})_{x,\mathfrak{b}_1\mathfrak{b}_2}, \tag{5.55}$$

for some sums of graphs $\Gamma_R^{(n)}$, $\Gamma_A^{(>n)}$, $\Gamma_Q^{(n)}$ and $\Gamma_{err}^{(n,D)}$. Then moving the second term on the right-hand side of (2.28) to the left-hand side and multiplying both sides by $(1-\Theta\Sigma^{(n)})^{-1}$, we get that

$$T_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2} = m\Theta^{(n)}_{\mathfrak{a}\mathfrak{b}_1}\overline{G}_{\mathfrak{b}_1\mathfrak{b}_2} + \sum_x \Theta^{(n)}_{\mathfrak{a}x}\left[(\Gamma_R^{(n)})_{x,\mathfrak{b}_1\mathfrak{b}_2} + (\Gamma_A^{(>n)})_{x,\mathfrak{b}_1\mathfrak{b}_2} + (\Gamma_Q^{(n)})_{x,\mathfrak{b}_1\mathfrak{b}_2} + (\Gamma_{err}^{(n,D)})_{x,\mathfrak{b}_1\mathfrak{b}_2}\right]. \tag{5.56}$$

where we used $\Theta^{(n)} = (1-\Theta\Sigma^{(n)})^{-1}\Theta$ by (1.23). We can expand $\Theta^{(n)}$ as

$$\Theta^{(n)} = \sum_{k=0}^{D}(\Theta\Sigma^{(n)})^k\Theta + \Theta^{(n)}_{err}, \quad \Theta^{(n)}_{err} := \sum_{k>D}(\Theta\Sigma^{(n)})^k\Theta. \tag{5.57}$$

This expansion is well-defined because $\|\Theta\Sigma^{(n)}\|_{\ell^\infty\to\ell^\infty} \leqslant W^{-d+\tau}$ for any constant $\tau>0$ by estimate (6.1) below. Every $(\Theta\Sigma^{(n)})^k\Theta$ can be expanded into a sum of *labelled diffusive edges* (cf. Definition 6.3), which are allowed in the T-expansion (cf. Definition 6.6). Moreover, we regard $(\Theta^{(n)}_{err})_{xy}$ as a diffusive edge of scaling order $>2D$. Then we plug (5.57) into (5.56) and rearrange the resulting graphs as follows: $m\Theta^{(n)}_{\mathfrak{a}\mathfrak{b}_1}\overline{G}_{\mathfrak{b}_1\mathfrak{b}_2}$ will give the first two terms in (2.19) and some graphs in $(\mathcal{A}_T^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$; $\sum_x \Theta^{(n)}_{\mathfrak{a}x}(\Gamma_R^{(n)})_{x,\mathfrak{b}_1\mathfrak{b}_2}$ will give $(\mathcal{R}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ and some graphs in $(\mathcal{A}_T^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$; $\sum_x \Theta^{(n)}_{\mathfrak{a}x}(\Gamma_A^{(>n)})_{x,\mathfrak{b}_1\mathfrak{b}_2}$ will give some graphs in $(\mathcal{A}_T^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$; $\sum_x \Theta^{(n)}_{\mathfrak{a}x}(\Gamma_Q^{(n)})_{x,\mathfrak{b}_1\mathfrak{b}_2}$ will give $(\mathcal{Q}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$; $\sum_x \Theta^{(n)}_{\mathfrak{a}x}(\Gamma_{err}^{(n,D)})_{x,\mathfrak{b}_1\mathfrak{b}_2}$ will give $(\mathcal{E}rr_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$. This concludes Lemma 5.13. □

Finally, we collect the results in Sections 5.2–5.5 to complete the proof of Theorem 2.1. The following proof is simply a recap of the strategy described in Section 5.1.

Proof of Theorem 2.1. We follow the flow chart in Figure 2.

Step 1: By Lemma 2.5, we have defined the second order T-expansion.

Step 2: Suppose that we have defined the k-th order T-expansion for all $2 \leqslant k \leqslant n-1$. Then applying Theorem 2.16, we get that the local law (2.27) holds as long as $L \leqslant L_n := W^{1+(n-2)d/4-c_0/2}$.

Step 3: We can construct an n-th order T-equation by Lemma 5.7.

Step 4: Using the local law in Step 2 and the n-th order T-equation in Step 3, we show properties (1.25)–(1.27) and (2.13)–(2.16) for $\mathcal{E}_n$ in Lemma 5.8.

Step 5: Applying Lemma 5.13 we obtain the n-th order T-expansion.

By induction, we can construct the n-th order T-expansion for all $2 \leqslant n \leqslant n_{W,L}$ with

$$n_{W,L} = \left\lceil \frac{4}{d}\left(\log_W L - 1 + \frac{c_0}{2}\right)\right\rceil + 2.$$

Then we apply Theorem 2.16 to conclude Theorem 2.1. □

For the reader's convenience, we summarize the lemmas in this section which are still to be proved.

- Lemmas 5.2, 5.4, 5.6 and 5.7 will be proved in [64].
- Lemma 5.9 and Lemma 5.10 will be proved in Section 7.
- Lemma 5.12 will be proved in Section 6.5.

In Section 9, we will describe some key ideas in [64] that are needed to prove Lemmas 5.4, 5.6 and 5.7.

6 Doubly connected property

In this section, we will introduce an important structural property—the doubly connected property—satisfied by the graphs in the T-expansion.

6.1 Labelled Θ edges

In this subsection, we show how (2.25) follows from the properties (1.25)–(1.27) of the self-energies. The following lemma is a simple consequence of the sum zero property and will be proved in Appendix B.

Lemma 6.1. *Fix any $z = E + i\eta$ with $E \in (-2+\kappa, 2-\kappa)$ and $\eta \geqslant W^2/L^{2-\varepsilon}$ for a small constant $\varepsilon > 0$. Let $g : \mathbb{Z}_L^d \to \mathbb{R}$ be a symmetric function (i.e. $g(x) = g(-x)$) supported on a box $\mathcal{B}_K := [\![-K, K]\!]^d$ of scale $K \geqslant W$. Assume that g satisfies the sum zero property $\sum_x g(x) = 0$. Then for any $x_0 \in \mathbb{Z}_L^d$ such that $|x_0| \geqslant K^{1+c}$ for a constant $c > 0$, we have that*

$$\Big|\sum_x \Theta_{0x}(z) g(x - x_0)\Big| \leqslant \sum_{x\in\mathcal{B}_K} \frac{x^2}{|x_0|^2}|g(x)| \cdot \left(|x_0|^\tau B_{0x_0}\mathbf{1}_{|x_0|\leqslant \eta^{-1/2}W^{1+\tau}} + |x_0|^{-D}\right),$$

for any constants $\tau, D > 0$.

With Lemma 6.1, we can readily obtain the following lemma. The long proof is due to extra arguments needed to handle the facts that (1.27) is only an approximate "sum zero property" and $(\mathcal{E}_l)_{0x}$ satisfies the "compactly supported property" of $g(x)$ only approximately.

Lemma 6.2. *Fix $d \geqslant 6$. Given a self-energy $\mathcal{E}_{2l}$ satisfying Definition 2.13, we have that*

$$\Big|\sum_\alpha \Theta_{x\alpha}(\mathcal{E}_{2l})_{\alpha y}\Big| \leqslant \frac{W^\tau}{W^{(l-1)d}\langle x-y\rangle^d}, \quad \forall\, x, y \in \mathbb{Z}_L^d, \tag{6.1}$$

for any small constant $\tau > 0$. Let $\mathcal{E}_{2k_1}, \mathcal{E}_{2k_2}, \cdots, \mathcal{E}_{2k_l}$ be a sequence of self-energies satisfying Definition 2.13. We have that for any small constant $\tau > 0$,

$$\left|(\Theta\mathcal{E}_{2k_1}\Theta\mathcal{E}_{2k_2}\Theta\cdots\Theta\mathcal{E}_{2k_l}\Theta)_{xy}\right| \leqslant W^{-(k-2)d/2+\tau}B_{xy}, \quad \forall\, x, y \in \mathbb{Z}_L^d, \tag{6.2}$$

where $k := \sum_{i=1}^l 2k_i - 2(l-1)$ is the scaling order of $(\Theta\mathcal{E}_{2k_1}\Theta\mathcal{E}_{2k_2}\Theta\cdots\Theta\mathcal{E}_{2k_l}\Theta)_{xy}$. The estimate (6.2) implies (2.25) for the $\Sigma_{T,k}$ defined in (2.21).

Proof. We abbreviate $r := \langle x - y\rangle$. To prove (6.1), we decompose the sum over α according to the dyadic scales $\mathcal{I}_n := \{\alpha \in \mathbb{Z}_L^d : K_{n-1} \leqslant |\alpha - y| \leqslant K_n\}$, where K_n are defined by

$$K_n := 2^n W \quad \text{for} \quad 1 \leqslant n \leqslant \log_2(L/W) - 1, \quad \text{and} \quad K_0 := 0. \tag{6.3}$$

If $K_n \geqslant W^{-\varepsilon} r$ for a small constant $\varepsilon > 0$, then we have that

$$\Big|\sum_{\alpha\in\mathcal{I}_n}\Theta_{x\alpha}(\mathcal{E}_{2l})_{\alpha y}\Big| \leqslant \sum_{\alpha\in\mathcal{I}_n}|\Theta_{x\alpha}| \cdot \max_{\alpha\in\mathcal{I}_n}|(\mathcal{E}_{2l})_{\alpha y}| \leqslant W^\varepsilon\frac{K_n^2}{W^2}\cdot W^\varepsilon\frac{W^2}{W^{(l-1)d}K_n^{d+2}} \leqslant \frac{W^{(d+2)\varepsilon}}{W^{(l-1)d}r^d}, \tag{6.4}$$

where in the second step we used (1.26) (together with $2d - 4 \geqslant d + 2$ when $d \geqslant 6$) and $\sum_{\alpha\in\mathcal{I}_n}\Theta_{x\alpha} \leqslant W^\varepsilon K_n^2/W^2$ by (1.17). It remains to bound the sum

$$\sum_{\alpha\in\mathcal{I}_{near}}\Theta_{x\alpha}(\mathcal{E}_{2l})_{\alpha y}, \quad \mathcal{I}_{near} := \bigcup_{n:K_n\leqslant W^{-\varepsilon}r}\mathcal{I}_n.$$

In order for $\mathcal{I}_{near}$ to be nonempty, it suffices to assume that $r \geqslant W^{1+\varepsilon}$.

Using (1.26) and (1.27), we can obtain that

$$\sum_{\alpha\in\mathcal{I}_{near}} (\mathcal{E}_{2l})_{\alpha y} = \sum_{\alpha}(\mathcal{E}_{2l})_{\alpha y} - \sum_{x\notin\mathcal{I}_{near}} (\mathcal{E}_{2l})_{\alpha y} \leqslant W^{\varepsilon}\left(\frac{\eta}{W^{(l-1)d}} + \frac{W^2}{W^{(l-1)d}(W^{-\varepsilon}r)^2}\right). \tag{6.5}$$

Then we write $(\mathcal{E}_{2l})_{\alpha y} = \overline{R} + \mathring{R}_{\alpha y}$ for $\alpha \in \mathcal{I}_{near}$, where $\overline{R} := \sum_{\alpha\in\mathcal{I}_{near}} (\mathcal{E}_{2l})_{\alpha y}/|\mathcal{I}_{near}|$ is the average of $(\mathcal{E}_{2l})_{\alpha y}$ over $\mathcal{I}_{near}$. By (1.26) and (6.5), we have that

$$|\overline{R}| \leqslant \frac{W^{(d+3)\varepsilon}W^2}{W^{(l-1)d}r^{d+2}} + \frac{\eta W^{(d+1)\varepsilon}}{W^{(l-1)d}r^{d}}, \quad |\mathring{R}_{\alpha y}| \leqslant \frac{W^{2+\varepsilon}}{W^{(l-1)d}\langle\alpha - y\rangle^{d+2}} + |\overline{R}|. \tag{6.6}$$

Thus we can bound that

$$\begin{aligned}\Big|\sum_{\alpha\in\mathcal{I}_{near}} \Theta_{x\alpha}\overline{R}\Big| &\leqslant \left(\frac{W^{(d+3)\varepsilon}W^2}{W^{(l-1)d}r^{d+2}} + \frac{\eta W^{(d+1)\varepsilon}}{W^{(l-1)d}r^{d}}\right)\sum_{\alpha\in\mathcal{I}_{near}}\Theta_{x\alpha} \\ &\lesssim \left(\frac{W^{(d+3)\varepsilon}W^2}{W^{(l-1)d}r^{d+2}} + \frac{\eta W^{(d+1)\varepsilon}}{W^{(l-1)d}r^{d}}\right)\min\left\{W^{\varepsilon}\frac{W^{-2\varepsilon}r^2}{W^2}, \eta^{-1}\right\} \lesssim \frac{W^{(d+2)\varepsilon}}{W^{(l-1)d}r^{d}},\end{aligned} \tag{6.7}$$

where in the second step we used (1.17) and (1.42) to bound $\sum_{\alpha\in\mathcal{I}_{near}}\Theta_{x\alpha}$. Finally, we use Lemma 6.1 to bound the sum over $\mathring{R}$ as

$$\begin{aligned}\Big|\sum_{\alpha\in\mathcal{I}_{near}} \Theta_{x\alpha}\mathring{R}_{\alpha y}\Big| &\leqslant \left(\sum_{\alpha\in\mathcal{I}_{near}}\frac{|\alpha-y|^2}{r^2}|\mathring{R}_{\alpha y}|\right)\left(\frac{W^{\varepsilon}}{W^2r^{d-2}}\mathbf{1}_{r\leqslant\eta^{-1/2}W^{1+\varepsilon}} + W^{-D}\right) \\ &\leqslant \left(\sum_{\alpha\in\mathcal{I}_{near}}\frac{W^{2+\varepsilon}}{r^2\langle\alpha-y\rangle^dW^{(l-1)d}} + W^{-(d+2)\varepsilon}r^d|\overline{R}|\right)\left(\frac{W^{\varepsilon}}{W^2r^{d-2}}\mathbf{1}_{r\leqslant\eta^{-1/2}W^{1+\varepsilon}} + W^{-D}\right) \\ &\leqslant \left(\frac{W^{2+2\varepsilon}}{W^{(l-1)d}r^2} + \frac{\eta W^{-\varepsilon}}{W^{(l-1)d}}\right)\left(\frac{W^{\varepsilon}\mathbf{1}_{r\leqslant\eta^{-1/2}W^{1+\varepsilon}}}{W^2r^{d-2}} + W^{-D}\right) \leqslant \frac{W^{4\varepsilon}}{W^{(l-1)d}r^d},\end{aligned} \tag{6.8}$$

where in the second and third steps we used (6.6), and in the last step we used

$$\frac{\eta W^{-\varepsilon}}{W^{(l-1)d}}\frac{W^{\varepsilon}}{W^2r^{d-2}}\mathbf{1}_{r\leqslant\eta^{-1/2}W^{1+\varepsilon}} \leqslant \frac{W^{2\varepsilon}}{W^{(l-1)d}r^d}.$$

Combining (6.4), (6.7) and (6.8), we conclude (6.1) since ε is arbitrary.

From (6.1), we can obtain (6.2) easily by using the following simple facts: if f_1, f_2 and g are functions on $\mathbb{Z}_L^d \times \mathbb{Z}_L^d$ satisfying that

$$|f_1(x,y)| \leqslant \langle x-y\rangle^{-d}, \quad |f_2(x,y)| \leqslant \langle x-y\rangle^{-d}, \quad |g(x,y)| \leqslant W^{-2}\langle x-y\rangle^{-d+2},$$

then we have

$$\sum_{\alpha}|f_1(x,\alpha)f_2(\alpha,y)| \lesssim \langle x-y\rangle^{-d}, \quad \sum_{\alpha}|f_1(x,\alpha)g(\alpha,y)| \lesssim W^{-2}\langle x-y\rangle^{-d+2}.$$

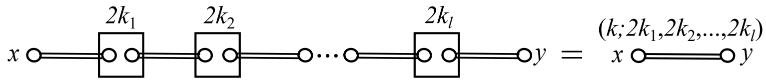
Finally, (2.25) follows from (6.2) directly by definition (2.21). □

The self-energies in Definition 2.13 will appear in the following *labelled diffusive edges*, which are formed by joining the self-energies with diffusive edges.

Definition 6.3 (Labelled diffusive edges)**.** *Given l self-energies $\mathcal{E}_{2k_i}$, $i = 1, 2, \cdots, l$, we represent the entry*

$$(\Theta\mathcal{E}_{2k_1}\Theta\mathcal{E}_{2k_2}\Theta\cdots\Theta\mathcal{E}_{2k_l}\Theta)_{xy} \tag{6.9}$$

by a labelled diffusive edge between atoms x and y with label $(k; 2k_1, \cdots, 2k_l)$, where $k := \sum_{i=1}^{l} 2k_i - 2(l-1)$ is the scaling order of this edge. In graphs, each labelled diffusive edge is drawn as one single double-line edge with a label but without any internal structure as in the following figure:



The scaling order of a labelled diffusive edge is calculated as follows. Taking (6.9) as an example, there are $l+1$ diffusive edges of total scaling order $2(l+1)$, self-energies $\mathcal{E}_{2k_1}, \cdots, \mathcal{E}_{2k_l}$ of total scaling order $\sum_{i=1}^{l} 2k_i$, and $2l$ internal atoms of total scaling order $-4l$. Hence the scaling order of (6.9) is $2(l+1)+\sum_{i=1}^{l} 2k_i - 4l = k$. By (6.2), (6.9) is bounded by $W^{-(k-2)d/2+\tau} B_{xy}$ for any constant $\tau > 0$, i.e. it has the same decay with respect to $|x-y|$ as Θ_{xy} except for an extra $W^{-(k-2)d/2}$ factor. As a convention, both diffusive and labelled diffusive edges will be called "diffusive edges".

The scaling order of a normal regular graph with labelled diffusive edges can be equivalently counted as

$$\begin{aligned}\text{ord}(\mathcal{G}) := &\ \#\{\text{off-diagonal } G \text{ edges}\} + \#\{\text{light weights}\} + 2\#\{\text{waved edges}\} + 2\#\{\text{diffusive edges}\} \\ &+ \sum_k k \cdot \#\{k\text{-th order labelled diffusive edges}\} - 2\left[\#\{\text{internal atoms}\} - \#\{\text{dotted edges}\}\right]. \quad (6.10)\end{aligned}$$

In other words, a k-th order labelled diffusive edges is simply counted as an edge of scaling order k, and there is no need to count its internal structures using Definition 2.12.

6.2 Doubly connected property

Recall the definition of molecules in Definition 3.4. We define the molecular graph as the quotient graph of the atomic graph with the equivalence relation that atoms belonging to the same molecule are equivalent.

Definition 6.4 (Molecular graphs). *Molecular graphs are graphs consisting of*

- *external molecules which represent the external atoms (such as the $\otimes$, $\oplus$ and $\ominus$ molecules);*
- *internal molecules;*
- *blue and red solid edges, which represent the plus and minus G edges between molecules;*
- *diffusive edges between molecules;*
- *dotted edges between external and internal molecules.*

Given any atomic graph $\mathcal{G}$, we define its molecular quotient graph $\mathcal{G}_{\mathcal{M}}$ in the following way:

- *each molecule of $\mathcal{G}$ is represented by a vertex in $\mathcal{G}_{\mathcal{M}}$;*
- *each blue or red solid edge of $\mathcal{G}$ between atoms in different molecules is represented by a blue or red solid edge between these two molecules in $\mathcal{G}_{\mathcal{M}}$;*
- *each diffusive edge of $\mathcal{G}$ between atoms in different molecules is represented by a diffusive edge between these two molecules in $\mathcal{G}_{\mathcal{M}}$;*
- *each dotted edge of $\mathcal{G}$ between an external atom and an internal atom is represented by a dotted edge between the corresponding external and internal molecules;*
- *we discard all the other components in $\mathcal{G}$ (including the weights, $\times$-dotted edges, and all edges inside any molecule).*

We emphasize that molecular graphs are used solely to analyze the graph structures; the expansions in Section 3 are only applied to atomic graphs. In the following proof, we assume that each atomic graph is automatically associated with a molecular graph. As discussed below Definition 3.4, we call the structure of the molecular graph as the *global structure* of the atomic graph.

The following *doubly connected* property is a key global property for our proof. It allows us to establish a direct connection between the scaling order of a graph and a bound on its value (cf. Lemma 6.10 below). In fact, all graphs in the T-expansion and T-equation will satisfy this property (cf. Definitions 6.6 and 6.7).

Definition 6.5 (Doubly connected property). *A subgraph $\mathcal{G}$ without external molecules is said to be* doubly connected *if its molecular graph $\mathcal{G}_{\mathcal{M}}$ satisfies the following property. There exist a collection, say $\mathcal{B}_{black}$, of diffusive edges and another collection, say $\mathcal{B}_{blue}$, of either blue solid or diffusive edges such that (a) $\mathcal{B}_{black} \cap \mathcal{B}_{blue} = \emptyset$, and (b) both $\mathcal{B}_{black}$ and $\mathcal{B}_{blue}$ contain a spanning tree that connects all molecules in the graph. For simplicity of notations, we call the diffusive edges in $\mathcal{B}_{black}$ as black edges, and the blue solid and diffusive edges in $\mathcal{B}_{blue}$ as blue edges. Correspondingly, $\mathcal{B}_{black}$ and $\mathcal{B}_{blue}$ are referred to as* black net *and* blue net*, respectively, where a "net" refers to a subset of edges that contains a spanning tree.*

A graph $\mathcal{G}$ with external molecules is said to be doubly connected if its subgraph with all external molecules removed is doubly connected, i.e. the spanning trees in the two nets are not required to contain the external molecules.

The doubly connected property is defined on molecular graphs, and thus is a *global property*. In the above definition, the diffusive edges also include labelled diffusive edges introduced in Definition 6.3. The red solid edges are not tracked in the doubly connected property, and the path connectivity of red solid edges can be broken in our expansion procedure in [64]. By symmetry, we can also define an expansion procedure so that graphs in the T-expansion satisfy the doubly connected property with a black net and a red net.

6.3 T-expansion with doubly connected structures

For graphs in the T-expansion, they are all doubly connected in the sense of Definition 6.5. By including this property and the labelled diffusive edges in Definition 6.3, we are now ready to state the rest of the details for the T-expansion in Definition 2.15 and the T-equation in Definition 2.17. We will design an expansion strategy in [64] so that all graphs in the T-expansion and T-equation are doubly connected.

Definition 6.6 (More properties of the n-th order T-expansion). *An n-th order T-expansion of $T_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ is an expression satisfying Definition 2.15 and the following additional properties.*

(i) A diffusive edge in the graphs on the right-hand side of (2.19) *is either a Θ edge or a labelled diffusive edge of the form* (6.9) *with $4 \leqslant 2k_i \leqslant n$.*

(ii) Each graph in $(\mathcal{R}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{A}_T^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{Q}_T^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ and $(\mathcal{E}rr_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ is doubly connected in the sense of Definition 6.5.

Definition 6.7 (More properties of the n-th order T-equation). *An n-th order T-equation of $T_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ is an expression satisfying Definition 2.17 and the following additional properties.*

(i) A diffusive edge in $\mathcal{E}_n$ and the graphs on the right-hand side of (2.28) *is either a Θ edge or a labelled diffusive edge of the form* (6.9) *with $4 \leqslant 2k_i \leqslant n-1$.*

(ii) Each graph in $\mathcal{E}_n$, $(\mathcal{R}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{A}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$, $(\mathcal{Q}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ and $(\mathcal{E}rr'_{n,D})_{\mathfrak{a},\mathfrak{b}_1\mathfrak{b}_2}$ is doubly connected in the sense of Definition 6.5.

6.4 Bounding doubly connected graphs

In this subsection, we give some important estimates on doubly connected graphs in Lemma 6.10. In particular, these estimates will be used crucially in the proofs of Lemma 5.9 and Lemma 5.12. Inspired by the maximum bound in (5.8) and the weak averaged bound in (5.9), we introduce the following weak and strong norms, which will be a convenient tool for the proof of Lemma 6.10.

Definition 6.8. *Given a $\mathbb{Z}_L^d \times \mathbb{Z}_L^d$ matrix $\mathcal{A}$ and some fixed $a, b > 0$, we define its weak-(a,b) norm as*

$$\|\mathcal{A}\|_{w;(a,b)} := W^{ad/2} \max_{x,y\in\mathbb{Z}_L^d} |\mathcal{A}_{xy}| + \sup_{K\in[W,L/2]} \left(\frac{W}{K}\right)^b K^{ad/2} \max_{x,x_0\in\mathbb{Z}_L^d} \frac{1}{K^d} \sum_{y:|y-x_0|\leqslant K} (|\mathcal{A}_{xy}| + |\mathcal{A}_{yx}|),$$

and its strong-(a,b) norm as

$$\|\mathcal{A}\|_{s;(a,b)} := \max_{x,y\in\mathbb{Z}_L^d} \left(\frac{W}{\langle x-y\rangle}\right)^b \langle x-y\rangle^{ad/2} |\mathcal{A}_{xy}|.$$

In this paper, we only use weak or strong-(a,b) norms with $a \leqslant 2$. In this case, it is easy to check that the strong-(a,b) norm is strictly stronger than the weak-(a,b) norm. By Definition 6.8, we immediately get the bounds

$$\max_{x,y\in\mathbb{Z}_L^d} |\mathcal{A}_{xy}| \leqslant W^{-ad/2}\|\mathcal{A}\|_{w;(a,b)}, \tag{6.11}$$

$$\max_{x,x_0\in\mathbb{Z}_L^d} \frac{1}{K^d}\sum_{y:|y-x_0|\leqslant K} (|\mathcal{A}_{xy}|+|\mathcal{A}_{yx}|) \leqslant \frac{1}{W^b K^{ad/2-b}}\|\mathcal{A}\|_{w;(a,b)}, \quad \text{for all } K\in[W,L/2], \tag{6.12}$$

$$|\mathcal{A}_{xy}| \leqslant \frac{1}{W^b\langle x-y\rangle^{ad/2-b}}\|\mathcal{A}\|_{s;(a,b)}. \tag{6.13}$$

Here we list the weak or strong norms of some key deterministic or random variables.

(i) $\|B\|_{s;(2,2)} \leqslant 1$ and $\|B^{(1/2)}\|_{s;(1,1)} \leqslant 1$, where $B^{(1/2)}$ is the matrix with entries $(B_{xy})^{1/2}$;

(ii) If (2.27) holds, then $\|G(z)-m(z)I_N\|_{s;(1,1)} \prec 1$.

(iii) If (5.8) and (5.9) hold, then $\|G(z)-m(z)I_N\|_{w;(1,2)} \prec W^{\varepsilon_0}$ and $W^{-2\varepsilon_0}\|T(z)\|_{w;(2,4)} \prec W^{2\varepsilon_0}$.

(iv) The following positive random variable Ψ_{xy} was defined in [65, Definition 3.4] for a small constant $\tau>0$ and a large constant $D>0$:

$$\Psi_{xy}^2 \equiv \Psi_{xy}^2(\tau,D) := W^{-D} + \max_{\substack{|x_1-x|\leqslant W^{1+\tau}\\ |y_1-y|\leqslant W^{1+\tau}}} s_{x_1y_1} + W^{-(2+2\tau)d}\sum_{|x_1-x|\leqslant W^{1+\tau}}\sum_{|y_1-y|\leqslant W^{1+\tau}} |G_{x_1y_1}|^2. \tag{6.14}$$

Note that $\|\Psi(z)\|_{w;(1,2)} \prec \|G(z)-m(z)I_N\|_{w;(1,2)}+1$ and $\|\Psi(z)\|_{s;(1,1)} \prec \|G(z)-m(z)I_N\|_{s;(1,1)}+1$ as long as D is large enough.

The motivation for introducing the Ψ matrix is as follows: given $x_1,x_2\in\mathbb{Z}_L^d$, suppose y_1 and y_2 satisfy that

$$|y_1-x_1| \leqslant W^{1+\tau/2}, \quad |y_2-x_2| \leqslant W^{1+\tau/2}. \tag{6.15}$$

If $y_1\ne y_2$ and we know that $\|G(z)\|_{\max}\prec 1$, then using Lemma 5.1 we can obtain the bound

$$\begin{aligned}
|G_{y_1y_2}(z)|^2 \prec T_{y_1y_2}(z) &= |m|^2 s_{y_1y_2}|G_{y_2y_2}(z)|^2 + |m|^2\sum_{\alpha\ne y_2} s_{y_1\alpha}|G_{\alpha y_2}(z)|^2\\
&= |m|^2 s_{y_1y_2}|G_{y_2y_2}(z)|^2 + |m|^2\sum_{\alpha\ne y_2} s_{y_1\alpha}|G_{y_2\alpha}(\overline{z})|^2\\
&\prec s_{y_1y_2} + \sum_{\alpha\ne y_2} s_{y_1\alpha}T_{y_2\alpha}(\overline{z}) \leqslant s_{y_1y_2} + \sum_{\alpha,\beta} s_{y_1\alpha}s_{y_2\beta}|G_{\alpha\beta}(z)|^2\\
&\leqslant W^{-D} + s_{y_1y_2} + W^{-2d}\sum_{|\alpha-y_1|\leqslant W^{1+\tau/2}}\sum_{|\beta-y_2|\leqslant W^{1+\tau/2}} |G_{\alpha\beta}(z)|^2 \leqslant W^{2d\tau}\Psi_{x_1x_2}^2(\tau,D),
\end{aligned} \tag{6.16}$$

where in the third and fifth steps we used the simple identity $G_{xy}(z)=\overline{G_{yx}(\overline{z})}$, and in the sixth step we used (1.8). In particular, if y_1 and y_2 are in the same molecules as x_1 and x_2, respectively, then we know that (6.15) holds, since otherwise the graph value will be smaller than W^{-D} for any fixed $D>0$ by (1.8) and (2.3). Then (6.16) shows that all the G edges between two molecules containing atoms x_1 and x_2 can be bounded with the same variable $\Psi_{x_1x_2}$. This fact will be convenient for our proof.

By (1.42), the row sums of Θ diverge when $L\to\infty$ (e.g. if $\eta = W^2/L^{2-\varepsilon}$). On the other hand, the following claim shows that the product of a Θ entry and a variable with bounded weak-$(1,2)$ norm is summable if $d\geqslant 8$. Although this claim will not be used in our proof directly, it explains why we require $d\geqslant 8$ in Theorem 1.4. Our proof of Lemma 6.10 is actually based on some more general versions of this claim in (6.32) and (6.33) below.

Claim 6.9. *Let $\mathcal{A}$ be a matrix satisfying $\|\mathcal{A}\|_{w;(a,b)}\prec 1$ for some fixed $a,b>0$. If*

$$ad/2-b-2\geqslant 0, \tag{6.17}$$

then we have that

$$\max_{x,y\in\mathbb{Z}_L^d}\sum_\alpha B_{x\alpha}\mathcal{A}_{y\alpha} \prec W^{-ad/2}. \tag{6.18}$$

Proof. We decompose the sum over α according to the dyadic scales:

$$\alpha \in \mathcal{I}_{n,m} := \{\alpha \in \mathbb{Z}_L^d : K_{n-1} \leqslant |x-\alpha| \leqslant K_n, K_{m-1} \leqslant |y-\alpha| \leqslant K_m\},$$

where K_n are defined in (6.3). Then using (6.12) and the fact that $\mathcal{I}_{n,m}$ is inside a box of scale $\mathrm{O}(K_n \wedge K_m)$, we can estimate that

$$\sum_{\alpha \in \mathcal{I}_{n,m}} B_{x\alpha}\mathcal{A}_{y\alpha} \prec \frac{1}{W^2 K_n^{d-2}} \sum_{\alpha \in \mathcal{I}_{n,m}} \mathcal{A}_{y\alpha} \prec \frac{1}{W^2 K_n^{d-2}} \cdot \frac{(K_n \wedge K_m)^d}{W^b (K_n \wedge K_m)^{ad/2-b}}$$
$$\leqslant \frac{1}{W^{b+2}(K_n \wedge K_m)^{ad/2-b-2}} \leqslant W^{-ad/2},$$

where in the last step we used (6.17). Summing over $\mathrm{O}((\log L)^2)$ many such sets $\mathcal{I}_{n,m}$, we get that

$$\sum_\alpha B_{x\alpha}\mathcal{A}_{y\alpha} \prec (\log L)^2 W^{-ad/2} \prec W^{-ad/2}.$$

This concludes the proof. □

If $\|G(z) - m(z)I_N\|_{w;(1,2)} \prec 1$, then by Claim 6.9 we have that

$$\max_{x,y \in \mathbb{Z}_L^d} \sum_\alpha B_{x\alpha}|G_{y\alpha}| \prec W^{-d/2}$$

if $d/2 - 4 \geqslant 0$, which gives $d \geqslant 8$. We now prove the following key estimates on doubly connected graphs.

Lemma 6.10. *Suppose $d \geqslant 8$ and $\|G(z) - m(z)I_N\|_{w;(1,2)} \prec 1$. Let $\mathcal{G}$ be a doubly connected normal regular graph without external atoms. Pick any two atoms of $\mathcal{G}$ and fix their values $x, y \in \mathbb{Z}_L^d$. Then the resulting graph $\mathcal{G}_{xy}$ satisfies that*

$$|\mathcal{G}_{xy}| \prec W^{-(n_{xy}-3)d/2} B_{xy}\mathcal{A}_{xy}, \tag{6.19}$$

where $n_{xy} := \mathrm{ord}(\mathcal{G}_{xy})$ is the scaling order of $\mathcal{G}_{xy}$ and $\mathcal{A}_{xy}$ is some positive variable satisfying $\|\mathcal{A}\|_{w;(1,2)} \prec 1$. Furthermore, if $\|G(z) - m(z)I_N\|_{s;(1,1)} \prec 1$, then we have that

$$|\mathcal{G}_{xy}| \prec W^{-(n_{xy}-3)d/2} B_{xy}^{3/2}. \tag{6.20}$$

If we fix an atom $x \in \mathcal{G}$, then the resulting graph $\mathcal{G}_x$ satisfies that

$$|\mathcal{G}_x| \prec W^{-\mathrm{ord}(\mathcal{G}_x)\cdot d/2}. \tag{6.21}$$

The above bounds hold also for the graph $\mathcal{G}^{\mathrm{abs}}$, which is obtained by replacing each component (including edges, weights and coefficients) in $\mathcal{G}$ with its absolute value and ignoring all the P or Q labels (if any). We emphasize that in defining $\mathcal{G}^{\mathrm{abs}}$, a labelled diffusive edge (6.9) will be regarded as one single edge and replaced by $|\left(\Theta\mathcal{E}_{2k_1}\Theta\mathcal{E}_{2k_2}\Theta\cdots\Theta\mathcal{E}_{2k_l}\Theta\right)_{xy}|$.

Note that a doubly connected graph $\mathcal{G}$ with at least two molecules must have $n_{xy} \geqslant 3$. If x and y are in the same molecule, then (6.19) gives the sharp bound $|\mathcal{G}_{xy}| \prec W^{-n_{xy}d/2}$.

Proof of Lemma 6.10. The estimate (6.21) is a special case of (6.19) with $x = y$. Hence we only need to prove (6.19) and (6.20). Moreover, due to the trivial bound $|\mathcal{G}_{xy}| \prec \mathcal{G}_{xy}^{\mathrm{abs}}$, it suffices to prove (6.19) and (6.20) for the graph $\mathcal{G}_{xy}^{\mathrm{abs}}$. As explained before, the η^{-1} factor in (1.42) is the main trouble for our proof. We will show that if we choose the order of summation in a proper way, then the following key property holds: for every summation over the global scale L, it involves a product of at least one diffusive edge and one variable whose weak-$(1,2)$ or strong-$(1,1)$ norm is bounded by $\mathrm{O}_\prec(1)$. In particular, every such summation does not provide a large η^{-1} factor as we have seen in (6.18).

By (1.8), (1.17), (2.3) and (6.2), we have the following maximum bounds on deterministic edges:

$$\max_{x,y} s_{xy} = \mathrm{O}(W^{-d}), \quad \max_{x,y} |S_{xy}^{\pm}| = \mathrm{O}(W^{-d}), \quad \max_{x,y} \Theta_{xy} \prec W^{-d},$$
$$\max_{x,y} |(\Theta\mathcal{E}_{2k_1}\Theta\mathcal{E}_{2k_2}\Theta\cdots\Theta\mathcal{E}_{2k_l}\Theta)_{xy}| \prec W^{-kd/2}, \tag{6.22}$$

where $k := \sum_{i=1}^{l} 2k_i - 2(l-1)$. For simplicity of notations, we will use $\alpha \sim_{\mathcal{M}} \beta$ to mean that "atoms α and β belong to the same molecule". Suppose there are ℓ internal molecules $\mathcal{M}_i$, $1 \leqslant i \leqslant \ell$, in $\mathcal{G}^{\rm abs}_{xy}$. We choose one atom in each $\mathcal{M}_i$, say x_i, as a representative. Moreover, let atoms x and y be the representatives of their respective molecules in $\mathcal{G}^{\rm abs}$. For definiteness, we assume that x and y belong to *different molecules.* The case where x and y belong to the same molecule can be dealt with in a similar way, and we omit the details. In the following proof, we fix a small constant $\tau > 0$ and a large constant $D > 0$. For any $y_i \sim_{\mathcal{M}} x_i$, it suffices to assume that

$$|y_i - x_i| \leqslant W^{1+\tau/2}, \tag{6.23}$$

because otherwise the graph is smaller than W^{-D}. Then under the assumption (6.23), for $y_i \sim_{\mathcal{M}} x_i$ and $y_j \sim_{\mathcal{M}} x_j$, by (1.17), (6.16) and (6.2) we have that

$$|G_{y_i y_j}| \prec W^{d\tau} \Psi_{x_i x_j}(\tau, D), \quad \Theta_{y_i y_j} \prec B_{y_i y_j} \lesssim W^{(d-2)\tau/2} B_{x_i x_j}, \tag{6.24}$$

$$\left|(\Theta \mathcal{E}_{2k_1} \Theta \mathcal{E}_{2k_2} \Theta \cdots \Theta \mathcal{E}_{2k_l} \Theta)_{y_i y_j}\right| \prec W^{-(k-2)d/2+(d-2)\tau/2} B_{x_i x_j}. \tag{6.25}$$

These estimates show that we can bound the edges between different molecules with Ψ or B entries that only contain the representative atoms x_i in their indices.

First, we bound the edges between different molecules. Due to the doubly connected property of $\mathcal{G}^{\rm abs}_{xy}$, we can pick two spanning trees of the black net and blue net, which we refer to as the *black tree* and *blue tree*, respectively. We bound the edges that do not belong to the two trees using the maximum bounds:

(i) each solid edge that is not in the blue tree is bounded by $\mathrm{O}_\prec(W^{-d/2})$ using (6.11) with $a = 1$;

(ii) each diffusive edge that is not in the black and blue trees is bounded by $\mathrm{O}_\prec(W^{-d})$;

(iii) each labelled diffusive edge that is not in the black and blue trees is bounded by $\mathrm{O}_\prec(W^{-kd/2})$, where k is the scaling order of this edge.

The edges in the two trees are bounded as follows:

(iv) the blue solid and diffusive edges in the two trees are bounded using (6.24) and (6.25).

In this way, we can bound that

$$\mathcal{G}^{\rm abs}_{xy} \prec W^{-n_1 d/2 + n_2 \tau} \sum_{x_1, \cdots, x_\ell} \Gamma_{global}(x_1, \cdots, x_\ell) \prod_{i=1}^{\ell} \mathcal{G}^{(i)}_{x_i}, \tag{6.26}$$

where $W^{-n_1 d/2 + n_2 \tau}$ is a factor coming from the above items (i)–(iv), Γ_{global} is a product of blue solid edges that represent Ψ entries and double-line edges that represent B entries, and every $\mathcal{G}^{(i)}_{x_i}$ is the subgraph inside the molecule $\mathcal{M}_i$, which has x_i as an external atom. We bound the local structure $\mathcal{G}^{(i)}_{x_i}$ inside $\mathcal{M}_i$ as follows:

- each waved or diffusive edge is bounded by $\mathrm{O}_\prec(W^{-d})$ using (6.22);
- each labelled diffusive edge is bounded by $\mathrm{O}_\prec(W^{-kd/2})$, where k is its scaling order;
- each off-diagonal G edge and light weight is bounded by $\mathrm{O}_\prec(W^{-d/2})$ using (6.11) with $a = 1$;
- each summation over an internal atom in $\mathcal{M}_i \setminus \{x_i\}$ provides a factor $\mathrm{O}(W^{(1+\tau/2)d})$ due to (6.23).

Thus with the definition of the scaling order in (6.10), we get that

$$|\mathcal{G}^{(i)}_{x_i}| \prec W^{-\mathrm{ord}(\mathcal{G}^{(i)}_{x_i}) \cdot d/2 + k_i \cdot \tau d/2}, \tag{6.27}$$

where k_i is the number of internal atoms in $\mathcal{G}^{(i)}_{x_i}$. Finally, for convenience of proof, we bound each diffusive edge in the *blue* (but not black) tree of $\Gamma_{global}(x_1, \cdots, x_\ell)$ as

$$B_{x_i x_j} \leqslant W^{-d/2} B^{1/2}_{x_i x_j}. \tag{6.28}$$

Then every edge in the blue tree represents a Ψ or $B^{(1/2)}$ entry, whose weak-$(1,2)$ or strong-$(1,1)$ norm is bounded by $\mathrm{O}_\prec(1)$ (depending on whether we want to prove (6.19) or (6.20)). Plugging (6.27) and (6.28) into (6.26), we obtain that

$$\mathcal{G}^{\rm abs}_{xy} \prec W^{-(n_{xy} - \ell - 3)d/2 + n_3 \tau} (\mathcal{G}_{xy})_{aux}, \tag{6.29}$$

where $n_3 := n_2 + \sum_{i=1}^{\ell} k_i d/2$ and the number $n_{xy} - \ell - 3$ in the exponent can be obtained by counting carefully the number of $W^{-d/2}$ factors from the above arguments. Here $(\mathcal{G}_{xy})_{aux}$ is an *auxiliary graph* defined as follows:

- it has two external atoms x and y, and some internal atoms x_i, $1 \leqslant i \leqslant \ell$, which are the representative atoms of the molecules in $\mathcal{G}_{xy}^{\text{abs}}$;
- each diffusive edge in the black tree of $\mathcal{G}_{xy}^{\text{abs}}$ is replaced by a double-line edge representing a B entry in $(\mathcal{G}_{xy})_{aux}$;
- each edge in the blue tree of $\mathcal{G}_{xy}^{\text{abs}}$ is replaced by a blue solid edge representing a Ψ or $B^{(1/2)}$ entry in $(\mathcal{G}_{xy})_{aux}$.

By the construction of $(\mathcal{G}_{xy})_{aux}$, it is doubly connected in the following sense: $(\mathcal{G}_{xy})_{aux}$ contains a black spanning tree consisting of black double-line edges and a blue spanning tree consisting of blue solid edges. Now with (6.29), to conclude the proof it suffices to show that after summing over all the internal atoms in $(\mathcal{G}_{xy})_{aux}$, the auxiliary graph can be bounded as

$$|(\mathcal{G}_{xy})_{aux}| \prec W^{-\ell d/2} B_{xy} \mathcal{A}_{xy}, \tag{6.30}$$

for a positive variable $\mathcal{A}_{xy}$ satisfying $\|\mathcal{A}\|_{w;(1,2)} \prec 1$ (resp. $\|\mathcal{A}\|_{s;(1,1)} \prec 1$) if $\|G(z) - m(z)I_N\|_{w;(1,2)} \prec 1$ (resp. $\|G(z) - m(z)I_N\|_{s;(1,1)} \prec 1$). The estimate (6.30) is an easy consequence of the following Claim 6.11. Our auxiliary graph $(\mathcal{G}_{xy})_{aux}$ satisfies its assumptions. We postpone its proof until we complete the proof of Lemma 6.10.

Claim 6.11. *Let $\widetilde{\mathcal{G}}_{xy}$ be a graph with two external atoms x and y, ℓ internal atoms $x_1, x_2, \cdots, x_\ell$, a black spanning tree consisting $\ell+1$ black double-line edges, and a blue spanning tree consisting $\ell+1$ blue solid edges. Suppose that each black edge between atoms, say α and β, represents a $B_{\alpha\beta}$ factor, and each blue edge represents a positive variable whose weak-(a,b) norm is bounded by $\mathrm{O}_\prec(1)$. If (6.17) holds, then*

$$|\widetilde{\mathcal{G}}_{xy}| \prec W^{-a\ell d/2} B_{xy} \mathcal{A}_{xy}, \tag{6.31}$$

for a positive variable $\mathcal{A}_{xy}$ satisfying $\|\mathcal{A}\|_{w;(a,b)} \prec 1$. Moreover, if the strong-(a,b) norm of each blue edge is bounded by $\mathrm{O}_\prec(1)$ and (6.17) holds, then (6.31) holds for a positive variable $\mathcal{A}_{xy}$ satisfying $\|\mathcal{A}\|_{s;(a,b)} \prec 1$.

Note that both $(a,b) = (1,2)$ and $(a,b) = (1,1)$ satisfy (6.17) for $d \geqslant 8$. Hence taking $a = 1$ in (6.31), we obtain (6.30). Combining (6.29) and (6.30), we conclude (6.19) and (6.20) for $\mathcal{G}_{xy}^{\text{abs}}$ since τ is arbitrary. □

Proof of Claim 6.11. Our proof is based on the following extensions of Claim 6.9. If $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$ are two matrices whose weak-(a,b) or strong-(a,b) norms are bounded by $\mathrm{O}_\prec(1)$, then we have that

$$\sum_{x_i} \mathcal{A}^{(2)}_{x_i\beta} \cdot \prod_{j=1}^{k} B_{x_i y_j} \prec W^{-ad/2} \Gamma(y_1, \cdots, y_k), \tag{6.32}$$

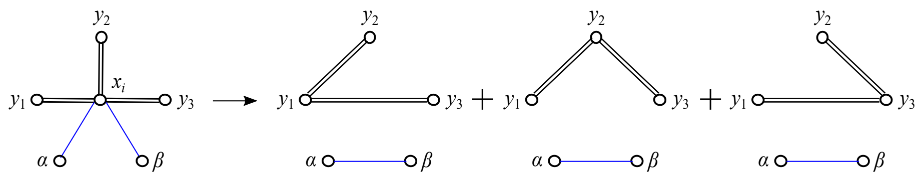
and

$$\sum_{x_i} \mathcal{A}^{(1)}_{x_i\alpha} \mathcal{A}^{(2)}_{x_i\beta} \cdot \prod_{j=1}^{k} B_{x_i y_j} \leqslant W^{-ad/2} \Gamma(y_1, \cdots, y_k) \mathcal{A}_{\alpha\beta}, \tag{6.33}$$

where $\mathcal{A}$ is a matrix with $\|\mathcal{A}\|_{w;(a,b)} \prec 1$ if $\|\mathcal{A}^{(1)}\|_{w;(a,b)} + \|\mathcal{A}^{(2)}\|_{w;(a,b)} \prec 1$ or $\|\mathcal{A}\|_{s;(a,b)} \prec 1$ if $\|\mathcal{A}^{(1)}\|_{s;(a,b)} + \|\mathcal{A}^{(2)}\|_{s;(a,b)} \prec 1$, and $\Gamma(y_1, \cdots, y_k)$ is defined as a sum of k different products of $(k-1)$ double-line edges:

$$\Gamma(y_1, \cdots, y_k) := \sum_{i=1}^{k} \prod_{j \neq i} B_{y_i y_j}. \tag{6.34}$$

Intuitively speaking, (6.33) means that after summing over a product of k double-line edges and two solid edges, we lose one double-line edge and one solid edge, which leads to the $W^{-ad/2}$ factor as in Claim 6.9. In each new graph, we have $(k-1)$ double-line edges connected with one of the neighbors of x_i on the black tree, and one solid edge between atoms α and β representing $\mathcal{A}_{\alpha\beta}$. In the following figure, we draw an example with $k = 3$, where there are three graphs corresponding to the three terms on the right-hand side of (6.34) and we have omitted the factor $W^{-ad/2}$ from them:



To prove (6.32), it suffices to assume the weaker condition $\|\mathcal{A}^{(2)}\|_{w;(a,b)} \prec 1$. We decompose the sum over x_i according to dyadic scales K_n defined in (6.3). Consider the case $x_i \in \mathcal{I}_{\vec{a}}$ for some $\vec{a} := (a_1, \cdots, a_k) \in (\mathbb{N} \setminus \{0\})^k$, where

$$\mathcal{I}_{\vec{a}} := \{x_i : K_{a_j - 1} \leqslant |x_i - y_j| \leqslant K_{a_j}, 1 \leqslant j \leqslant k\}.$$

For simplicity of notations, we abbreviate $L_j := K_{a_j}$ and $L_{\min} := \min_{1\leqslant j\leqslant k} L_j$. Then we have that

$$\sum_{x_i \in \mathcal{I}_{\vec{a}}} |\mathcal{A}^{(2)}_{x_i\beta}| \cdot \prod_{j=1}^k B_{x_i y_j} \leqslant \prod_{j=1}^k \frac{1}{W^2 L_j^{d-2}} \sum_{x \in \mathcal{I}_{\vec{a}}} |\mathcal{A}^{(2)}_{x_i\beta}| \prec \frac{L^d_{\min}}{W^b L_{\min}^{ad/2-b}} \frac{1}{\prod_{j=1}^k W^2 L_j^{d-2}}, \tag{6.35}$$

where in the second step we used (6.12) and the fact that $\mathcal{I}_{\vec{a}}$ is inside a box of scale $L_{\min}$. Let $s \in \{1, 2, \cdots, k\}$ be the value such that $L_s = L_{\min}$. Using (6.17), we obtain that

$$\frac{L^d_{\min}}{W^b L_{\min}^{ad/2-b}} \frac{1}{W^2 L_s^{d-2}} = \frac{1}{W^{b+2} L_{\min}^{ad/2-b-2}} \leqslant W^{-ad/2}.$$

Combining this bound with the fact that

$$\langle y_j - y_s \rangle \leqslant W + L_j + L_s \leqslant 3L_j, \quad j \neq s, \tag{6.36}$$

we can bound (6.35) as

$$\sum_{x_i \in \mathcal{I}_{\vec{a}}} |\mathcal{A}^{(2)}_{x_i\beta}| \cdot \prod_{j=1}^k B_{x_i y_j} \prec W^{-ad/2} \prod_{j\neq s} \frac{1}{W^2 L_j^{d-2}} \lesssim W^{-ad/2} \prod_{j \neq s} B_{y_j y_s} \leqslant W^{-ad/2} \Gamma(y_1, \cdots, y_k).$$

Summing over all possible scales $\mathcal{I}_{\vec{a}}$, we conclude (6.32).

Next we prove (6.33) when $\|\mathcal{A}^{(1)}\|_{w;(a,b)} + \|\mathcal{A}^{(2)}\|_{w;(a,b)} \prec 1$. Applying (6.11) to $\mathcal{A}^{(1)}_{x_i\alpha}$ and using (6.32), we get that

$$\sum_{x_i} |\mathcal{A}^{(1)}_{x_i\alpha}||\mathcal{A}^{(2)}_{x_i\beta}| \cdot \prod_{j=1}^k B_{x_i y_j} \prec W^{-ad/2} \sum_{x_i} |\mathcal{A}^{(2)}_{x_i\beta}| \cdot \prod_{j=1}^k B_{x_i y_j} \prec W^{-ad} \Gamma(y_1, \cdots, y_k).$$

Applying (6.12) to $\mathcal{A}^{(1)}_{x_i\alpha}$ and using (6.32), we get that for any $x_0 \in \mathbb{Z}_L^d$ and $K \in [W, L/2]$,

$$\frac{1}{K^d} \sum_{\alpha:|\alpha - x_0| \leqslant K} \sum_{x_i} |\mathcal{A}^{(1)}_{x_i\alpha}||\mathcal{A}^{(2)}_{x_i\beta}| \cdot \prod_{j=1}^k B_{x_i y_j} \prec \frac{\sum_{x_i} |\mathcal{A}^{(2)}_{x_i\beta}| \cdot \prod_{j=1}^k B_{x_i y_j}}{W^b K^{ad/2-b}} \prec W^{-ad/2} \frac{\Gamma(y_1, \cdots, y_k)}{W^b K^{ad/2-b}}.$$

We can obtain a similar estimate for the average over $\{\beta : |\beta - x_0| \leqslant K\}$. The above two estimates imply that $\|\mathcal{A}\|_{w;(a,b)} \prec 1$, where $\mathcal{A}$ is defined by

$$\mathcal{A}_{\alpha\beta} := \frac{W^{ad/2}}{\Gamma(y_1, \cdots, y_k)} \sum_{x_i} |\mathcal{A}^{(1)}_{x_i\alpha}||\mathcal{A}^{(2)}_{x_i\beta}| \cdot \prod_{j=1}^k B_{x_i y_j}. \tag{6.37}$$

This concludes (6.33) in one case. Then we prove (6.33) in the other case with $\|\mathcal{A}^{(1)}\|_{s;(a,b)} + \|\mathcal{A}^{(2)}\|_{s;(a,b)} \prec 1$. We decompose the sum over x_i according to dyadic scales as $x_i \in \mathcal{I}_{\vec{a}}$ for some $\vec{a} := (a_1, \cdots, a_{k+2}) \in (\mathbb{N} \setminus \{0\})^{k+2}$, where

$$\mathcal{I}_{\vec{a}} := \{x_i : K_{a_j-1} \leqslant |x_i - y_j| \leqslant K_{a_j}, 1 \leqslant j \leqslant k; K_{a_{k+1}-1} \leqslant |x_i - \alpha| \leqslant K_{a_{k+1}}, K_{a_{k+2}-1} \leqslant |x_i - \beta| \leqslant K_{a_{k+2}}\}.$$

For simplicity of notations, we abbreviate $L_j := K_{a_j}$, $1 \leqslant j \leqslant k+2$, and $L_{\min} := \min_{1\leqslant j\leqslant k} L_j$. Let $s \in \{1, 2, \cdots, k\}$ be the value such that $L_s = L_{\min}$. Then using (6.13) and the fact that $\mathcal{I}_{\vec{a}}$ is inside a box of scale $L_{\min} \wedge L_{k+1} \wedge L_{k+2}$, we obtain that

$$\begin{aligned}
&\sum_{x_i\in\mathcal{I}_{\vec{a}}} |\mathcal{A}^{(1)}_{x_i\alpha}||\mathcal{A}^{(2)}_{x_i\beta}| \cdot \prod_{j=1}^{k} B_{x_i y_j} \prec \prod_{j=1}^{k} \frac{1}{W^2 L_j^{d-2}} \cdot \frac{(L_{\min}\wedge L_{k+1}\wedge L_{k+2})^d}{W^{2b} L_{k+1}^{ad/2-b} L_{k+2}^{ad/2-b}} \\
&\leqslant \prod_{1\leqslant j\leqslant k, j\neq s} \frac{1}{W^2 L_j^{d-2}} \cdot \frac{1}{W^b (L_{k+1}\vee L_{k+2})^{ad/2-b}} \cdot \frac{1}{W^{b+2}(L_{k+1}\wedge L_{k+2})^{ad/2-b-2}} \\
&\lesssim W^{-ad/2} \frac{1}{W^b \langle \alpha-\beta\rangle^{ad/2-b}} \prod_{j\neq s} B_{y_j y_s}.
\end{aligned}$$

Here in the second step we used $L_{k+1}L_{k+2} = (L_{k+1}\vee L_{k+2})(L_{k+1}\wedge L_{k+2})$, and in the third step we used (6.17), (6.36) and $\langle\alpha-\beta\rangle \leqslant W + L_{k+1} + L_{k+2} \leqslant 3L_{k+1}\vee L_{k+2}$. Summing the above estimate over all possible scales $\mathcal{I}_{\vec{a}}$, we get that $\|\mathcal{A}\|_{s;(a,b)} \prec 1$, which concludes (6.33).

Now the proof of (6.31) involves repeated applications of (6.32) and (6.33) with a carefully chosen order of summations. Without loss of generality, we regard y as the root of the blue tree, and sum over the internal vertices from the leaves of the blue tree to the root. More precisely, we will sum over the vertices according to a partial order $x_{i_1} \preceq x_{i_2} \preceq \cdots \preceq x_{i_\ell} \preceq y$ that is compatible with the blue tree structure—if x_i is a child of x_j, then we have $x_i \preceq x_j$. By renaming the labels of vertices if necessary, we can assume that the partial order is $x_1 \preceq x_2 \preceq \cdots \preceq x_\ell$, so that we will perform the summations according to the order $\sum_{x_\ell}\cdots\sum_{x_2}\sum_{x_1}$. For simplicity of notations, we denote all the blue solid edges appearing in the proof by $\mathcal{A}$, including the old edges in $\widetilde{\mathcal{G}}_{xy}$ and the new edges coming from applications of (6.33). All these $\mathcal{A}$ variables have weak-(a,b) or strong-(a,b) norms bounded by $\mathrm{O}_\prec(1)$, and their exact expressions may change from one line to another.

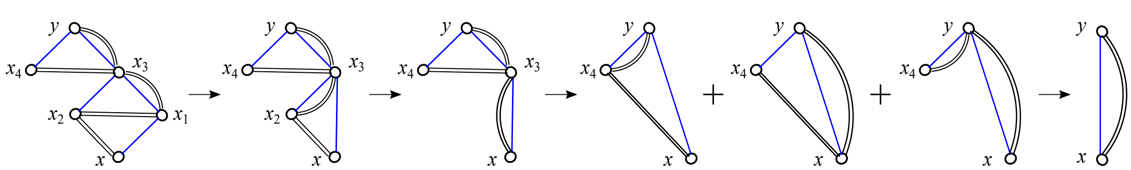
For the summation over x_1, using (6.32) (if x_1 is not connected with x in the blue tree) or (6.33) (if x_1 is connected with x in the blue tree), we can bound $\widetilde{\mathcal{G}}_{xy}$ as

$$\widetilde{\mathcal{G}}_{xy} \prec W^{-ad/2} \sum_{k=1}^{\ell_1} \mathcal{G}^{(1)}_{xy,k}, \tag{6.38}$$

where $\mathcal{G}^{(1)}_{xy,k}$ are new graphs obtained by replacing the edges connected to x_1 with the graphs on the right-hand side of (6.32) or (6.33), and ℓ_1 is the number of neighbors of x_1 on the black tree. More precisely, we perform the following operations to get these new graphs.

- We get rid of the blue solid and black double-line edges connected with x_1.
- If x and x_1 are connected through a blue solid edge in $\widetilde{\mathcal{G}}_{xy}$, then in each new graph x is connected to the parent of x_1 on the blue tree through a blue solid edge.
- Suppose $w_1, \cdots, w_{\ell_1}$ are the neighbors of x_1 on the black tree. Then corresponding to the k-th term in $\Gamma(w_1, \cdots, w_{\ell_1})$, $k = 1, \cdots, \ell_1$, the atoms $w_1, \cdots, w_{k-1}, w_{k+1}, \cdots, w_{\ell_1}$ are connected to w_k through double-line edges in the new graph.

Now it is crucial to observe that each new graph $\mathcal{G}^{(1)}_{xy,k}$ is still doubly connected. In (6.39), we show the reduction from the first graph to the second one through a summation over x_1, where we have omitted the factor $W^{-ad/2}$ from the graphs.



(6.39)

Similarly, we can bound the summations over atoms $x_2, \cdots, x_\ell$ one by one using (6.32) and (6.33). At each step we gain an extra factor $W^{-ad/2}$ and reduce the graphs into a sum of several new graphs, each of which has one fewer atom and a doubly connected structure. Finally, after summing over all internal atoms, we obtain a graph with atoms x and y only. In this case, the only doubly connected graph is the graph where x

and y are connected by a double-line B_{xy} edge and a blue solid edge whose weak-(a,b) or strong-(a,b) norm is bounded by $\mathrm{O}_{\prec}(1)$. This concludes (6.31). In (6.39), we give an example of the above graph reduction process by summing over the four internal atoms. □

If $\|G_{xy}(z) - m(z)I_n\|_{w;(a,b)} \prec W^{\varepsilon_0}$, then from (6.19) we immediately get that

$$|\mathcal{G}_{xy}| \prec W^{(n_{xy}-2)\varepsilon_0} \cdot W^{-(n_{xy}-3)d/2} B_{xy}\mathcal{A}_{xy}, \tag{6.40}$$

for a positive variable $\mathcal{A}_{xy}$ satisfying $\|\mathcal{A}\|_{w;(1,2)} \prec 1$. This follows from the fact that the number of light weights and off-diagonal G edges in $\mathcal{G}_{xy}$ is at most $n_{xy} - 2$, because by property (ii) of Definition 2.11, the number of internal atoms in $\mathcal{G}_{xy}$ is smaller than the number of waved and diffusive edges at least by 1.

Deterministic doubly connected graphs satisfy better bounds than Lemma 6.10, because all edges in the blue net are now (labelled) diffusive edges, whose strong-$(2,2)$ norms are bounded by $\mathrm{O}_{\prec}(1)$.

Corollary 6.12. *Suppose $d \geqslant 6$. Let $\mathcal{G}$ be a deterministic doubly connected normal regular graph without external atoms. Pick any two atoms of $\mathcal{G}$ and fix their values as $x, y \in \mathbb{Z}_L^d$. Then the resulting graph $\mathcal{G}_{xy}$ satisfies that*

$$|\mathcal{G}_{xy}| \prec W^{-(n_{xy}-4)d/2} B_{xy}^2, \quad \text{with} \quad n_{xy} := \mathrm{ord}(\mathcal{G}_{xy}). \tag{6.41}$$

This bound also holds for the graph $\mathcal{G}_{xy}^{\mathrm{abs}}$.

Proof. This corollary can be proved in the same way as Lemma 6.10, except that we need to apply Claim 6.11 to an auxiliary graph whose blue edges have strong-$(2,2)$ norms bounded by $\mathrm{O}_{\prec}(1)$. □

We also need another version of Corollary 6.12, which will be used in the proof of Lemma 5.9 in Section 7.1 below.

Corollary 6.13. *Under the assumptions of Corollary 6.12, suppose we replace every (labelled) diffusive edge in $\mathcal{G}$ between atoms, say α and β, with an edge bounded by $\mathrm{O}_{\prec}(B_{\alpha\beta})$. We treat these edges as double-line edges of scaling order 2 and call the resulting graph $\mathcal{G}'$. We pick any two atoms of $\mathcal{G}'$ and fix their values as $x, y \in \mathbb{Z}_L^d$. Then the resulting graph $\mathcal{G}'_{xy}$ satisfies the bound*

$$|\mathcal{G}'_{xy}| \leqslant \frac{W^{-\left(n'_{xy}-4\right)d/2}}{W^4\langle x-y\rangle^{2d-4-\tau}}, \quad \text{with} \quad n'_{xy} := \mathrm{ord}(\mathcal{G}'_{xy}), \tag{6.42}$$

for any constant $\tau > 0$. Furthermore, suppose we replace a double-line edge between atoms, say α_0 and β_0, in $\mathcal{G}'_{xy}$ with an edge bounded by $\mathrm{O}_{\prec}(\widetilde{B}_{\alpha_0\beta_0})$, where $\widetilde{B}_{\alpha_0\beta_0} := W^{-4}\langle \alpha_0 - \beta_0\rangle^{-(d-4)}$. We treat this edges as a double-line edge of scaling order 2 and denote the resulting graph by $\mathcal{G}''_{xy}$. Then it satisfies the bound

$$|\mathcal{G}''_{xy}| \leqslant \frac{W^{-\left(n''_{xy}-4\right)d/2}}{W^6\langle x-y\rangle^{2d-6-\tau}}, \quad \text{with} \quad n''_{xy} := \mathrm{ord}(\mathcal{G}''_{xy}), \tag{6.43}$$

for any constant $\tau > 0$.

Proof. The estimate (6.42) follows from (6.41). The estimate (6.43) can be proved in the same way as Lemma 6.10, except that we need to apply Claim 6.11 to an auxiliary graph whose blue edges have strong-$(2,4)$ norms bounded by $\mathrm{O}_{\prec}(1)$. □

6.5 Proof of Lemma 5.12

In this subsection, we complete the proof of Lemma 5.12 using Lemma 6.10. Recall that by Theorem 2.16, the local law (2.27) holds for $G(z_n, \psi, W, L_n)$, so $\|G(z_n) - m(z_n)I_N\|_{w;(1,1)} \prec 1$. For simplicity of notations, in the following proof we abbreviate $G \equiv G(z_n, f, W, L_n)$. Moreover, in the setting of Lemma 5.12, $\otimes$ represents the external atom $\mathfrak{a}$, while $\oplus$ represents the external atom $\mathfrak{b}$.

We first consider the $\oplus$-recollision graphs in $(\mathcal{R}_{IT,k})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}$. Take a graph from $(\mathcal{R}_{IT,k})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}$, say $\mathcal{G}_{\mathfrak{a}\mathfrak{b}}$. By Definitions 2.17 and 6.7, it has at least one dotted edge connected with $\oplus$, a diffusive edge connected with $\otimes$, is of scaling order $\geqslant 3$, and is doubly connected in the sense of Definition 6.5. Now we combine $\oplus$ with the internal atoms that connect to it through dotted edges. Then by property (iii) of Definition 2.17, we can write that

$$\mathcal{G}_{\mathfrak{a}\mathfrak{b}} = \sum_x \Theta_{\mathfrak{a}x}(\mathcal{G}_0)_{x\mathfrak{b}}, \quad \text{or} \quad \mathcal{G}_{\mathfrak{a}\mathfrak{b}} = \Theta_{\mathfrak{a}\mathfrak{b}}(\mathcal{G}_0)_{\mathfrak{b}}, \tag{6.44}$$

for a graph $(\mathcal{G}_0)_{x\mathfrak{b}}$ or $(\mathcal{G}_0)_{\mathfrak{b}}$ satisfying the assumptions of Lemma 6.10. Using (1.42) and (6.20), we can bound the first case of (6.44) as

$$\sum_{\mathfrak{a},\mathfrak{b}} |\mathcal{G}_{\mathfrak{a}\mathfrak{b}}| \leqslant \sum_{\mathfrak{a},\mathfrak{b},x} \Theta_{\mathfrak{a}x} |(\mathcal{G}_0)_{x\mathfrak{b}}| \prec \eta_n^{-1} W^{-(k-3)d/2} \sum_{x,\mathfrak{b}} B_{x\mathfrak{b}}^{3/2} \lesssim L_n^d \frac{W^{-(k-2)d/2}}{\eta_n}. \tag{6.45}$$

The second case of (6.44) is easier to bound and we omit the details.

The proof of (5.42) is similar. Recall that by Definitions 2.17 and 6.7, the graphs in $(\mathcal{A}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}$ are of scaling orders $\geqslant n+1$ and doubly connected in the sense of Definition 6.5 (i.e. the subgraphs induced on the internal atoms are doubly connected). Without loss of generality, we only consider the graphs in $(\mathcal{A}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}$ that are not $\oplus$-recollision graphs, because otherwise they can be bounded in the same way as the graphs in $(\mathcal{R}_{IT}^{(n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}$. Pick one such graph $\mathcal{G}_{\mathfrak{a}\mathfrak{b}}$ in $(\mathcal{A}_{IT}^{(>n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}$. It can be written into

$$\mathcal{G}_{\mathfrak{a}\mathfrak{b}} = \sum_{x,y,y'} \Theta_{\mathfrak{a}x} (\mathcal{G}_0)_{x,yy'} G_{y\mathfrak{b}} \overline{G}_{y'\mathfrak{b}}, \quad \text{or} \quad \mathcal{G}_{\mathfrak{a}\mathfrak{b}} = \sum_{x,y} \Theta_{\mathfrak{a}x} (\mathcal{G}_0)_{xy} \Theta_{y\mathfrak{b}}, \tag{6.46}$$

or some forms obtained by setting some indices of x, y, y' to be equal to each other. Without loss of generality, we only consider the two cases in (6.46), while all the other cases can be dealt with in similar ways. By the doubly connected property of $\mathcal{G}_{\mathfrak{a}\mathfrak{b}}$, we know that $\mathcal{G}_0$ is doubly connected. Using (1.42) and (6.20), we can bound the second term of (6.46) as

$$\sum_{\mathfrak{a},\mathfrak{b}} |\mathcal{G}_{\mathfrak{a}\mathfrak{b}}| \prec \eta_n^{-2} \sum_{x,y} |(\mathcal{G}_0)_{xy}| \prec \eta_n^{-2} W^{-(n-2)d/2} \sum_{x,y} B_{xy}^{3/2} \prec L_n^d \frac{W^{-(n-1)d/2}}{\eta_n^2},$$

where in the second step we used that $\text{ord}((\mathcal{G}_0)_{xy}) \geqslant n+1$. Using (1.42), (6.20) and Ward's identity (5.15), we can bound the first term in (6.46) as

$$\begin{aligned}\sum_{\mathfrak{a},\mathfrak{b}} |\mathcal{G}_{\mathfrak{a}\mathfrak{b}}| &\prec \sum_{\mathfrak{a},\mathfrak{b}} \sum_{x,y,y'} \Theta_{\mathfrak{a}x} \left|(\mathcal{G}_0)_{x,yy'} G_{y\mathfrak{b}} \overline{G}_{y'\mathfrak{b}}\right| \prec \eta_n^{-2} W^{-(n-2)d/2} \sum_{x,y,y'} \left(\mathcal{G}_0^{\text{abs}}\right)_{x,yy'} \\ &\prec \eta_n^{-2} W^{-(n-2)d/2} \sum_{y,y'} B_{yy'}^{3/2} \prec L_n^d \frac{W^{-(n-1)d/2}}{\eta_n^2}.\end{aligned}$$

Here in the third step we used that $\sum_x (\mathcal{G}_0^{\text{abs}})_{x,yy'}$ is a doubly connected graph satisfying the assumptions of Lemma 6.10 with two fixed atoms y and y', so that it satisfies (6.20). Combining the above estimates, we conclude (5.42).

Finally, (5.43) can be proved in the same way as (5.42) by using that the scaling orders of the graphs in $(\mathcal{E}rr'_{n,D})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}$ are at least $D+1$.

7 Infinite space limit

In this subsection, we study the infinite space limits of the self-energies $\mathcal{E}_{2l}^\infty$. In particular, we will complete the proofs of Lemma 5.9 and Lemma 5.10. We write the graphs $\mathcal{E}_{2l}$ as

$$\mathcal{E}_{2l} \equiv \mathcal{E}_{2l}\left(m(z), S, S^{\pm}(z), \Theta(z)\right), \tag{7.1}$$

where the matrices S, $S^{\pm}$ and Θ depend on W, L and η. We want to remove the L and η dependence by taking $L \to \infty$ and $\eta \to 0$. More precisely, we define the infinite space limit $\mathcal{E}_{2l}^\infty$ as follows.

Definition 7.1 (Infinite space limits)**.** *Given a deterministic regular graph $\mathcal{G} \equiv \mathcal{G}\left(m(z), S, S^{\pm}(z), \Theta(z)\right)$ with $z = E + \mathrm{i}\eta$, we define*

$$\mathcal{G}^\infty \equiv \mathcal{G}^\infty\left(m(E), S_\infty, S_\infty^{\pm}(E), \Theta_\infty(E)\right), \quad x \in \mathbb{Z}^d, \tag{7.2}$$

in the following way. Recall that we denote $m(E) := m(E + \mathrm{i}0_+)$.

(i) We replace the $s_{\alpha\beta}$ edges in $\mathcal{G}$ with $(S_\infty)_{\alpha\beta}$, where (recall (1.5))

$$(S_\infty)_{\alpha\beta} := \lim_{L\to\infty} f_{W,L}(\alpha-\beta). \tag{7.3}$$

(ii) We replace the $\mathcal{S}^{\pm}_{\alpha\beta}(z)$ edges in $\mathcal{G}$ with $(S^{\pm}_\infty)_{\alpha\beta}(E)$, where

$$S^+_\infty(E) := \frac{m^2(E)S_\infty}{1-m^2(E)S_\infty}, \quad S^-_\infty(E) := \overline{S}^+_\infty(E). \tag{7.4}$$

(iii) We replace the $\Theta_{\alpha\beta}$ edges in $\mathcal{G}$ with $(\Theta_\infty)_{\alpha\beta}$, where

$$(\Theta_\infty)_{\alpha\beta} := \lim_{L\to\infty} \Theta_{\alpha\beta}\left(E+\mathrm{i}\frac{W^2}{L^2}, L\right). \tag{7.5}$$

(iv) For all $m(z)$ in the coefficient (that is, $m(z)$'s that do not appear in $S^{\pm}(z)$ and $\Theta(z)$ entries), we replace them with $m(E)$.

(v) Finally, we let all the internal atoms take values over the whole $\mathbb{Z}^d$.

Note that $\mathcal{G}^\infty$ (if exists) only depends on E, W and ψ in Assumption 1.2, but does not depend on L and η.

We first show that $S^{\pm}_\infty(E)$ and $\Theta_\infty(E)$ are well-defined, and give some basic estimates on them. The proof of Lemma 7.2 will be given in Appendix B.

Lemma 7.2. *For any $x\in\mathbb{Z}^d$ and $E\in(-2+\kappa,2-\kappa)$, $(S^{\pm}_\infty(E))_{0x}$ and $(\Theta_\infty(E))_{0x}$ exist and we have that*

$$|(S^+_\infty)_{0x}(E)| \lesssim W^{-d}\mathbf{1}_{|x|\leqslant W^{1+\tau}} + (|x|+W)^{-D}, \tag{7.6}$$

and

$$|(\Theta_\infty)_{0x}(E)| \leqslant \frac{1}{W^2(|x|+W)^{d-2-\tau}}, \tag{7.7}$$

for any constants $\tau, D>0$. Moreover, for any $L\geqslant W$ and $z=E+\mathrm{i}\eta$ with $W^2/L^{2-\varepsilon}\leqslant\eta\leqslant 1$ for a small constant $\varepsilon>0$, we have that

$$|(S^+_\infty)_{0x}(E) - S^+_{0x}(z)| \lesssim \eta W^{-d}\mathbf{1}_{|x|\leqslant W^{1+\tau}} + (|x|+W)^{-D}, \quad \forall\, x\in\mathbb{Z}^d_L, \tag{7.8}$$

and

$$|(\Theta_\infty)_{0x}(E) - \Theta_{0x}(z)| \leqslant \frac{\eta}{W^4(|x|+W)^{d-4-\tau}} + (|x|+W)^{-D}, \quad \forall\, x\in\mathbb{Z}^d_L, \tag{7.9}$$

for any constants $\tau, D>0$,

We have the following counterpart of Lemma 6.1 with Θ replaced by Θ_∞. The proof of Lemma 7.3 will be given in Appendix B.

Lemma 7.3. *Fix any $E\in(-2+\kappa,2-\kappa)$. Let $g:\mathbb{Z}^d\to\mathbb{R}$ be a symmetric function supported on a box $\mathcal{B}_K := [\![-K,K]\!]^d$ of scale $K\geqslant W$. Assume that g satisfies the sum zero property $\sum_x g(x)=0$. Then for any $x_0\in\mathbb{Z}^d$ such that $|x_0|\geqslant K^{1+c}$ for a constant $c>0$, we have that*

$$\Big|\sum_x (\Theta_\infty)_{0x}(E)g(x-x_0)\Big| \leqslant \left(\sum_{x\in\mathcal{B}_K}\frac{x^2}{|x_0|^2}|g(x)|\right)\frac{1}{W^2|x_0|^{d-2-\tau}},$$

for any constant $\tau>0$.

With this lemma, we can obtain the following counterpart of Lemma 6.2 for the infinite space limits of the labelled diffusive edges.

Lemma 7.4. *Fix $d \geqslant 6$. For any $\mathcal{E}_{2l}^{\infty}$ satisfying (2.13), (2.14) and (2.16), we have that*

$$\Big|\sum_{\alpha}(\Theta_\infty)_{x\alpha}(\mathcal{E}_{2l}^{\infty})_{\alpha y}\Big| \leqslant \frac{W^{-(l-1)d}}{(|x-y|+W)^{d-\tau}}, \quad \forall\, x,y\in\mathbb{Z}^d, \tag{7.10}$$

for any constant $\tau>0$. If $\mathcal{E}_{2k_1}^{\infty},\cdots,\mathcal{E}_{2k_l}^{\infty}$ satisfy (2.13), (2.14) and (2.16), then we have that

$$\Big|\big(\Theta_\infty\mathcal{E}_{2k_1}^{\infty}\Theta_\infty\mathcal{E}_{2k_2}^{\infty}\Theta_\infty\cdots\Theta_\infty\mathcal{E}_{2k_l}^{\infty}\Theta_\infty\big)_{xy}\Big| \leqslant \frac{W^{-(k-2)d/2}}{W^2(|x-y|+W)^{d-2-\tau}}, \quad \forall\, x,y\in\mathbb{Z}^d, \tag{7.11}$$

for any constant $\tau>0$, where $k:=\sum_{i=1}^{l}2k_i-2(l-1)$.

Proof. As in (2.2), we abbreviate $\langle x-y\rangle := |x-y|+W$ for our current setting with $L=\infty$. With Lemma 7.3, the proofs of (7.10) and (7.11) are similar to the ones for (6.1) and (6.2). To prove (7.10), we decompose the sum over α according to $\mathcal{I}_{far}:=\{\alpha: |\alpha-y|\geqslant\langle x-y\rangle^{1-\tau}\}$ and $\mathcal{I}_{near}:=\{\alpha:|\alpha-y|<\langle x-y\rangle^{1-\tau}\}$. Using (7.7) and (2.14) (together with $2d-4\geqslant d+2$ when $d\geqslant 6$), we can bound that for any constant $\tau>0$,

$$\begin{aligned}\Big|\sum_{\alpha\in\mathcal{I}_{far}}(\Theta_\infty)_{x\alpha}(\mathcal{E}_{2l}^{\infty})_{\alpha y}\Big| &\leqslant \frac{1}{W^{(l-1)d}}\sum_{\alpha\in\mathcal{I}_{far}}\frac{1}{\langle x-\alpha\rangle^{d-2-\tau}}\frac{1}{\langle\alpha-y\rangle^{d+2-\tau}}\\ &\leqslant \frac{1}{W^{(l-1)d}\langle x-y\rangle^{(1-\tau)(d-3\tau)}}\sum_{\alpha\in\mathcal{I}_{far}}\frac{1}{\langle x-\alpha\rangle^{d-2-\tau}}\frac{1}{\langle\alpha-y\rangle^{2+2\tau}}\\ &\lesssim \frac{1}{W^{(l-1)d}\langle x-y\rangle^{(1-\tau)(d-3\tau)}}.\end{aligned} \tag{7.12}$$

For the sum over $\alpha\in\mathcal{I}_{near}$, we decompose it as $(\mathcal{E}_{2l}^{\infty})_{\alpha y}=\overline{R}+\mathring{R}_{\alpha y}$ with

$$\overline{R}:=\frac{\sum_{\alpha\in\mathcal{I}_{near}}(\mathcal{E}_{2l}^{\infty})_{\alpha y}}{|\mathcal{I}_{near}|}=-\frac{\sum_{\alpha\in\mathcal{I}_{far}}(\mathcal{E}_{2l}^{\infty})_{\alpha y}}{|\mathcal{I}_{near}|},$$

where we used (2.16) in the second step. Then using (2.14), we can obtain that

$$|\overline{R}|\leqslant\frac{\langle x-y\rangle^{(d+3)\tau}W^2}{W^{(l-1)d}\langle x-y\rangle^{d+2}},\quad |\mathring{R}_{\alpha y}|\leqslant\frac{W^2}{W^{(l-1)d}\langle\alpha-y\rangle^{d+2-\tau}}+|\overline{R}|. \tag{7.13}$$

We can bound the term with $\overline{R}$ as

$$\begin{aligned}\Big|\sum_{\alpha\in\mathcal{I}_{near}}(\Theta_\infty)_{x\alpha}\overline{R}\Big| &\leqslant \frac{\langle x-y\rangle^{(d+3)\tau}W^2}{W^{(l-1)d}\langle x-y\rangle^{d+2}}\sum_{\alpha\in\mathcal{I}_{near}}(\Theta_\infty)_{x\alpha}\leqslant\frac{\langle x-y\rangle^{(d+3)\tau}W^2}{W^{(l-1)d}\langle x-y\rangle^{d+2}}\frac{\langle x-y\rangle^{(1-\tau)\cdot(2+\tau)}}{W^2}\\ &\leqslant\frac{\langle x-y\rangle^{(d+3)\tau}}{W^{(l-1)d}\langle x-y\rangle^{d}},\end{aligned}$$

where in the second step we used (7.7) to bound $\sum_{\alpha\in\mathcal{I}_{near}}(\Theta_\infty)_{x\alpha}$. On the other hand, we use Lemma 7.3 and (7.13) to bound the term with $\mathring{R}$ as

$$\begin{aligned}\Big|\sum_{\alpha\in\mathcal{I}_{near}}\Theta_{x\alpha}\mathring{R}_{\alpha y}\Big| &\leqslant \sum_{\alpha\in\mathcal{I}_{near}}|\alpha-y|^2|\mathring{R}_{\alpha y}|\cdot\frac{\langle x-y\rangle^{\tau}}{W^2\langle x-y\rangle^{d}}\\ &\leqslant\Bigg(\sum_{\alpha\in\mathcal{I}_{near}}\frac{W^{-(l-1)d}}{\langle\alpha-y\rangle^{d-\tau}}+\frac{|\overline{R}|}{W^2}\langle x-y\rangle^{(1-\tau)(d+2)}\Bigg)\frac{\langle x-y\rangle^{\tau}}{\langle x-y\rangle^{d}}\lesssim\frac{\langle x-y\rangle^{2\tau}}{W^{(l-1)d}\langle x-y\rangle^{d}}.\end{aligned}$$

Combining the above two estimates with (7.12), we conclude (7.10). Finally, (7.11) follows from (7.10). □

We will refer to the infinite space limits of the diffusive and labelled diffusive edges as Θ_∞ and labelled Θ_∞ edges. The estimates (7.7) and (7.11) suggest that these two types of edges can be also used in the doubly connected property.

Definition 7.5 (Doubly connected property with Θ_∞ egdes)**.** *We extend the doubly connected property in Definition 6.5 by including Θ_∞ and labelled Θ_∞ edges, which can be used either in the black net $\mathcal{B}_{black}$ or the blue net $\mathcal{B}_{blue}$.*

From Corollary 6.13, we immediately obtain the following result, which explains why (2.14) holds.

Corollary 7.6. *Suppose $d \geqslant 6$. Let $\mathcal{G}$ be a deterministic doubly connected graph without external atoms. Denote its infinite space limit by $\mathcal{G}^\infty$. Pick any two atoms of $\mathcal{G}^\infty$ and fix their values as $x, y \in \mathbb{Z}^d$. Then the resulting graph $\mathcal{G}^\infty_{xy}$ satisfies that for any constant $\tau > 0$,*

$$\left|\mathcal{G}^\infty_{xy}\right| \leqslant W^{-n_{xy}d/2} \frac{W^{2d-4}}{(|x-y|+W)^{2d-4-\tau}}, \quad \text{with} \quad n_{xy} := \operatorname{ord}(\mathcal{G}_{xy}).$$

Proof. This result is a corollary of (6.42) by taking $L \to \infty$. ☐

7.1 Proof of Lemma 5.9

Now we prove the following lemma, which implies Lemma 5.9 as a special case.

Lemma 7.7. *Fix $d \geqslant 6$. Suppose we have a sequence of self-energies $\mathcal{E}_{2l}$, $4 \leqslant 2l \leqslant n-1$, satisfying Definition 2.13 and properties (2.13)–(2.16). Let $\mathcal{G}$ be a deterministic graph satisfying the assumptions of Corollary 6.12, and let $\mathcal{G}^\infty$ be its infinite space limit. Moreover, suppose the labelled diffusive edges in $\mathcal{G}_{xy}$ can only be of the form (6.9) with $4 \leqslant 2k_i \leqslant n-1$. Fix any $L \geqslant W$ and $z = E + \mathrm{i}\eta$ with $E \in (-2+\kappa, 2-\kappa)$ and $W^2/L^{2-\varepsilon} \leqslant \eta \leqslant L^{-\varepsilon}$ for a small constant $\varepsilon > 0$. Then for any $x \in \mathbb{Z}_L^d$, we have that*

$$\left|\mathcal{G}_{0x}(m(z), S, S^\pm(z), \Theta(z)) - \mathcal{G}^\infty_{0x}(m(E), S_\infty, S^\pm_\infty(E), \Theta_\infty(E))\right| \leqslant W^{-n_0 d/2} \frac{\eta W^{2d-6}}{\langle x\rangle^{2d-6-\tau}}, \tag{7.14}$$

for any constant $\tau > 0$, where $n_0 := \operatorname{ord}(\mathcal{G}_{0x})$. Moreover, (7.14) implies that for any constant $\tau > 0$,

$$\left|\sum_{x\in\mathbb{Z}_L^d} \mathcal{G}_{0x}(m(z), S, S^\pm(z), \Theta(z)) - \sum_{x\in\mathbb{Z}^d} \mathcal{G}^\infty_{0x}(m(E), S_\infty, S^\pm_\infty(E), \Theta_\infty(E))\right| \leqslant \frac{L^\tau \eta}{W^{(n_0-2)d/2}}. \tag{7.15}$$

Proof. Using (7.14) when $x \in \mathbb{Z}_L^d$ and applying Corollary 7.6 to $\mathcal{G}^\infty_{0x}$ when $x \notin \mathbb{Z}_L^d$, we obtain that

$$\begin{aligned}
&\left|\sum_{x\in\mathbb{Z}_L^d} \mathcal{G}_{0x}(m(z), S, S^\pm(z), \Theta(z)) - \sum_{x\in\mathbb{Z}^d} \mathcal{G}^\infty_{0x}(m(E), S_\infty, S^\pm_\infty(E), \Theta_\infty(E))\right| \\
&\leqslant \sum_{\|x\|_\infty \leqslant L/2} \frac{\eta W^{2d-6}}{W^{n_0 d/2}(|x|+W)^{2d-6-\tau}} + \sum_{\|x\|_\infty > L/2} \frac{W^{2d-4}}{W^{n_0 d/2}|x|^{2d-4-\tau}} \\
&\lesssim L^\tau \left(\eta + \frac{W^2}{L^2}\right) W^{-(n_0-2)d/2} \lesssim L^\tau \eta W^{-(n_0-2)d/2},
\end{aligned}$$

which concludes (7.15). It remains to prove (7.14).

First, using $|m(z) - m(E)| = \mathrm{O}(\eta)$, we observe that replacing $m(z)$ in the coefficient with $m(E)$ leads to an extra factor η:

$$\left|\mathcal{G}_{0x}(m(z), S, S^\pm(z), \Theta(z)) - \mathcal{G}_{0x}(m(E), S, S^\pm(z), \Theta(z))\right| \leqslant \frac{\eta W^{2d-4}}{W^{n_0 d/2}\langle x\rangle^{2d-4-\tau}},$$

for any small constant $\tau > 0$. It remains to prove that for $x \in \mathbb{Z}_L^d$,

$$\left|\mathcal{G}_{0x}(m(E), S, S^\pm(z), \Theta(z)) - \mathcal{G}^\infty_{0x}(m(E), S_\infty, S^\pm_\infty(E), \Theta_\infty(E))\right| \leqslant \frac{\eta W^{2d-6}}{W^{n_0 d/2}\langle x\rangle^{2d-6-\tau}}. \tag{7.16}$$

For this purpose, we define a new graph $\mathcal{G}^{[\eta]}_{0x}(m(E), S^{[\eta]}, S^{\pm,[\eta]}(z), \Theta^{[\eta]}(z))$ obtained by replacing the S, $S^\pm(z)$ and $\Theta(z)$ edges defined on $\mathbb{Z}_L^d$ with $S^{[\eta]}$, $S^{\pm,[\eta]}(z)$ and $\Theta^{[\eta]}(z)$ edges defined on $\mathbb{Z}^d$, where

$$S^{[\eta]}_{\alpha\beta} := S_{\alpha\beta}\mathbf{1}_{|\alpha-\beta|\leqslant L^\tau W\eta^{-1/2}}, \quad S^{\pm,[\eta]}_{\alpha\beta}(z) := S^\pm_{\alpha\beta}(z)\mathbf{1}_{|\alpha-\beta|\leqslant L^\tau W\eta^{-1/2}},$$

and

$$\Theta^{[\eta]}_{\alpha\beta}(z) := \Theta_{\alpha\beta}(z)\mathbf{1}_{|\alpha-\beta|\leqslant L^\tau W\eta^{-1/2}},$$

for $\alpha,\beta\in\mathbb{Z}^d$. Note that for a sufficiently small $\tau\in(0,\varepsilon/4)$, we have $L^\tau W\eta^{-1/2}\leqslant L^{1-\varepsilon/4}$. Hence in order for $\mathcal{G}^{[\eta]}_{0x}$ to be nonzero, any atom α in it must satisfy $|\alpha|\leqslant C_0L^\tau W\eta^{-1/2}\ll L$ for a constant $C_0>0$. By (1.8), (1.17) and (2.3), we have that for any constant $D>0$,

$$\max_{x\in\mathbb{Z}_L^d}\left|\mathcal{G}_{0x}(m(E),S,S^{\pm}(z),\Theta(z))-\mathcal{G}^{[\eta]}_{0x}(m(E),S^{[\eta]},S^{\pm,[\eta]}(z),\Theta^{[\eta]}(z))\right|\leqslant L^{-D}.$$

Hence to prove (7.16), it remains to show that for $x\in\mathbb{Z}_L^d$,

$$\left|\mathcal{G}^{[\eta]}_{0x}(m(E),S^{[\eta]},S^{\pm,[\eta]}(z),\Theta^{[\eta]}(z))-\mathcal{G}^{\infty}_{0x}(m(E),S_\infty,S^{\pm}_\infty(E),\Theta_\infty(E))\right|\leqslant\frac{\eta W^{2d-6}}{W^{n_0d/2}\langle x\rangle^{2d-6-\tau}}. \tag{7.17}$$

By Corollary 7.6, we have that for $|x|>C_0L^\tau W\eta^{-1/2}$,

$$\begin{aligned}&\left|\mathcal{G}^{[\eta]}_{0x}(m(E),S^{[\eta]},S^{\pm,[\eta]}(z),\Theta^{[\eta]}(z))-\mathcal{G}^{\infty}_{0x}(m(E),S_\infty,S^{\pm}_\infty(E),\Theta_\infty(E))\right|\\&=\left|\mathcal{G}^{\infty}_{0x}(m(E),S_\infty,S^{\pm}_\infty(E),\Theta_\infty(E))\right|\leqslant\frac{W^{2d-4}}{W^{n_0d/2}\langle x\rangle^{2d-4-\tau}}\leqslant\frac{\eta W^{2d-6}}{W^{n_0d/2}\langle x\rangle^{2d-6-\tau}}.\end{aligned}$$

It remains to prove (7.17) for $|x|\leqslant C_0W^{1+\tau}\eta^{-1/2}$. We will replace the $S^{[\eta]}$, $S^{\pm,[\eta]}$, $\Theta^{[\eta]}$ and labelled $\Theta^{[\eta]}$ edges in $\mathcal{G}^{[\eta]}_{0x}$ with the S_∞, $S^{\pm}_\infty$, Θ_∞ and labelled Θ_∞ edges one by one, and control the error of each replacement using the estimates (1.8), (7.8) and (7.9). We remark that when dealing with a labelled $\Theta^{[\eta]}$ edge, we will replace a self-energy $\mathcal{E}^{[\eta]}_{2l}$, $4\leqslant 2l\leqslant n-1$, with $\mathcal{E}^{\infty}_{2l}$ as a whole, and the estimate (2.15) will be used to bound the difference. For simplicity, in the following proof we use the notations

$$\mathcal{G}^{[\eta]}_{0x}(S^{[\eta]},S^{\pm,[\eta]},\Theta^{[\eta]}),\quad \mathcal{G}^{\infty}_{0x}(S_\infty,S^{\pm}_\infty(E),\Theta_\infty(E)),$$

with the understanding that the arguments $\Theta^{[\eta]}$ and Θ_∞ represent both diffusive and labelled diffusive edges. First, using (1.8), it is easy to see that replacing any $S^{[\eta]}$ edge with a S_∞ edge gives an error of order $\mathrm{O}(L^{-D})$. Second, using (7.8), it is easy to show that replacing any $S^{\pm,[\eta]}(z)$ edge with a $S^{\pm}_\infty(E)$ edge leads to an extra factor η. Hence after replacing all $S^{[\eta]}$ and $S^{\pm,[\eta]}(z)$ edges with S_∞ and $S^{\pm}_\infty$ edges, we get that

$$\left|\mathcal{G}^{[\eta]}_{0x}(S^{[\eta]},S^{\pm,[\eta]}(z),\Theta^{[\eta]}(z))-\mathcal{G}^{[\eta]}_{0x}(S_\infty,S^{\pm}_\infty(E),\Theta^{[\eta]}(z))\right|\leqslant\frac{\eta W^{2d-4}}{W^{n_0d/2}\langle x\rangle^{2d-4-\tau}}.$$

Here as a convention, we still add the superscript $[\eta]$ to the graph after the replacements, but its arguments are different from the original graph. It remains to show that replacing the $\Theta^{[\eta]}$ and labelled $\Theta^{[\eta]}$ edges with Θ_∞ and labelled Θ_∞ edges leads to a small enough error:

$$\left|\mathcal{G}^{[\eta]}_{0x}(S_\infty,S^{\pm}_\infty(E),\Theta^{[\eta]}(z))-\mathcal{G}^{\infty}_{0x}(S_\infty,S^{\pm}_\infty(E),\Theta_\infty(E))\right|\leqslant\frac{\eta W^{2d-6}}{W^{n_0d/2}\langle x\rangle^{2d-6-\tau}}. \tag{7.18}$$

Combining the above two estimates, we conclude (7.17).

It remains to prove (7.18). Notice that $\mathcal{G}^{[\eta]}_{0x}(S_\infty,S^{\pm}_\infty(E),\Theta^{[\eta]}(z))-\mathcal{G}^{\infty}_{0x}(S_\infty,S^{\pm}_\infty(E),\Theta_\infty(E))$ can be written into a sum of $\mathrm{O}(1)$ many graphs, each of which is of scaling order n_0 and has a doubly connected structure consisting of $\Theta^{[\eta]}$ and Θ_∞ edges, labelled $\Theta^{[\eta]}$ and Θ_∞ edges, and one edge of the form $(\Theta^{[\eta]}-\Theta_\infty)_{\alpha\beta}$ or

$$\left[\Theta^{[\eta]}\mathcal{E}^{[\eta]}_{2k_1}\Theta^{[\eta]}\mathcal{E}^{[\eta]}_{2k_2}\Theta^{[\eta]}\cdots\Theta^{[\eta]}\mathcal{E}^{[\eta]}_{2k_l}\Theta^{[\eta]}-\Theta_\infty\mathcal{E}^{\infty}_{2k_1}\Theta_\infty\mathcal{E}^{\infty}_{2k_2}\Theta_\infty\cdots\Theta_\infty\mathcal{E}^{\infty}_{2k_l}\Theta_\infty\right]_{\alpha\beta}, \tag{7.19}$$

with $4\leqslant 2k_i\leqslant n-1$, $1\leqslant i\leqslant l$, and scaling order $2s:=\sum_{i=1}^l 2k_i-2(l-1)$. Let $(\mathcal{G}_\omega)_{0x}$ be one of these graphs. We claim that

$$|(\mathcal{G}_\omega)_{0x}|\leqslant\frac{\eta W^{2d-6}}{W^{n_0d/2}\langle x\rangle^{2d-6-\tau}}\quad\text{for}\quad |x|\leqslant C_0W^{1+\tau}\eta^{-1/2}. \tag{7.20}$$

With (7.20), we immediately conclude (7.18).

Finally we prove (7.20). If $(\mathcal{G}_\omega)_{0x}$ contains a $(\Theta^{[\eta]}-\Theta_\infty)_{\alpha\beta}$ edge, then by (7.7) and (7.9) we obtain that for any constant $\tau>0$,

$$\left|(\Theta^{[\eta]}-\Theta_\infty)_{\alpha\beta}\right| \leqslant \frac{\eta \mathbf{1}_{|\alpha-\beta|\leqslant L^\tau W\eta^{-1/2}}}{W^4(|\alpha-\beta|+W)^{d-4-\tau}} + \frac{\mathbf{1}_{|\alpha-\beta|>L^\tau W\eta^{-1/2}}}{W^2(|\alpha-\beta|+W)^{d-2-\tau}} \leqslant \frac{\eta}{W^4(|\alpha-\beta|+W)^{d-4-\tau}}, \quad (7.21)$$

where in the second step we used that $W^2/|\alpha-\beta|^2\leqslant\eta$ for $|\alpha-\beta|>L^\tau W\eta^{-1/2}$. Thus we can write that $\mathcal{G}_\omega=\eta\widetilde{\mathcal{G}}_\omega$ for a graph $\widetilde{\mathcal{G}}_\omega$ which has a doubly connected structure consisting of $\Theta^{[\eta]}$ and Θ_∞ edges, labelled $\Theta^{[\eta]}$ and Θ_∞ edges, and one special edge between α and β bounded by $\mathrm{O}_\prec(\widetilde{B}_{\alpha\beta})$. Then applying (6.43) (in the $L\to\infty$ case), we obtain that

$$\left|(\widetilde{\mathcal{G}}_\omega)_{0x}\right| \leqslant \frac{W^{-(n_0-4)d/2}}{W^6(|x|+W)^{2d-6-\tau}},$$

which implies (7.20). On the other hand, suppose $\mathcal{G}_\omega$ contains an edge of the form (7.19). Following the same argument as above, in order to show (7.20), it suffices to prove that for any constant $\tau>0$,

$$|(7.19)| \leqslant \frac{\eta W^{-(s-1)d}}{W^4(|\alpha-\beta|+W)^{d-4-\tau}}. \quad (7.22)$$

We prove this estimate by replacing the $\Theta^{[\eta]}$ and $\mathcal{E}^{[\eta]}_{2k_i}$ entries one by one, and bounding the error of each replacement using (7.9) and (2.15). First, with (7.9) and (6.1), we get that

$$\begin{aligned}
&\left|\left[(\Theta_\infty-\Theta^{[\eta]})\mathcal{E}^{[\eta]}_{2k_1}\Theta^{[\eta]}\mathcal{E}^{[\eta]}_{2k_2}\Theta^{[\eta]}\cdots\Theta^{[\eta]}\mathcal{E}^{[\eta]}_{2k_l}\Theta^{[\eta]}\right]_{\alpha\beta}\right| \\
&\leqslant \sum_{\alpha_1,\cdots,\alpha_l}\frac{\eta}{W^4(|\alpha-\alpha_1|+W)^{d-4-\tau}}\prod_{i=1}^{l-1}\frac{W^{-(k_i-1)d}}{(|\alpha_i-\alpha_{i+1}|+W)^{d-\tau}}\frac{W^{-(k_l-1)d}}{(|\alpha_l-\beta|+W)^{d-\tau}} \\
&\lesssim \frac{\eta W^{-(s-1)d}}{W^4(|\alpha-\beta|+W)^{d-4-(l+1)\tau}}.
\end{aligned}$$

Second, using (7.7), (6.1) and (2.15) for $\mathcal{E}_{2k_1}$, we get that

$$\begin{aligned}
&\left|\left[\Theta_\infty\left(\mathcal{E}^{[\eta]}_{2k_1}-\mathcal{E}^{\infty}_{2k_1}\right)\Theta^{[\eta]}\mathcal{E}^{[\eta]}_{2k_2}\Theta^{[\eta]}\cdots\Theta^{[\eta]}\mathcal{E}^{[\eta]}_{2k_l}\Theta^{[\eta]}\right]_{\alpha\beta}\right| \\
&\leqslant \sum_{\alpha_1,\cdots,\alpha_l,\beta_1}\frac{1}{W^2(|\alpha-\alpha_1|+W)^{d-2-\tau}}\frac{\eta W^{2d-6}}{W^{k_1 d}(|\alpha_1-\beta_1|+W)^{2d-6-\tau}}\frac{1}{W^2(|\beta_1-\alpha_2|+W)^{d-2-\tau}} \\
&\quad\times\prod_{i=2}^{l-1}\frac{W^{-(k_i-1)d}}{(|\alpha_i-\alpha_{i+1}|+W)^{d-\tau}}\frac{W^{-(k_l-1)d}}{(|\alpha_l-\beta|+W)^{d-\tau}} \\
&\lesssim \frac{\eta W^{-(s-1)d}}{W^4(|\alpha-\beta|+W)^{d-4-(l+2)\tau}}.
\end{aligned}$$

Continuing the above process, we can replace $\Theta^{[\eta]}$ with Θ_∞ and $\mathcal{E}^{[\eta]}_{2k_i}$ with $\mathcal{E}^{\infty}_{2k_i}$ one by one. Moreover, using (2.15), (6.1), (7.7), (7.9) and (7.10) at each step, we can show that each replacement gives an error at most

$$\frac{\eta W^{-(s-1)d}}{W^4(|\alpha-\beta|+W)^{d-4-(l+2)\tau}}.$$

This implies (7.22) since τ is arbitrarily small, and hence concludes (7.20). □

Now we can complete the proof of Lemma 5.9.

Proof of Lemma 5.9. As given by Definition 6.7, $\mathcal{E}_n$ is a sum of O(1) many deterministic doubly connected graphs satisfying the assumptions of Lemma 7.7. Hence we immediately conclude Lemma 5.9 using Corollary 6.12, Corollary 7.6 and Lemma 7.7. □

7.2 Proof of Lemma 5.10

Finally, in this subsection we give the full proof of Lemma 5.10. Recall the matrices $\widetilde S$, $\widetilde S^{\pm}$ and $\widetilde\Theta$ defined in (5.37). The following claim shows that $\widetilde S$, $\widetilde S^{\pm}$ and $\widetilde\Theta$ are close to S, $S^{\pm}$ and Θ. Its proof will be given in Section B.

Claim 7.8. *Under the assumptions of Lemma 5.10, fix any $L \geqslant W$ and $z = E + \mathrm{i}\eta$ with $E \in (-2+\kappa, 2-\kappa)$ and $W^2/L^{2-\varepsilon} \leqslant \eta \leqslant L^{-\varepsilon}$ for a small constant $\varepsilon > 0$. For any $x \in \mathbb{Z}_L^d$, we have that*

$$|\widetilde S_{0x} - S_{0x}| \lesssim \frac{W^2}{L^2}\frac{1}{W^d}\mathbf{1}_{|x|\leqslant W^{1+\tau}} + \langle x\rangle^{-D}, \tag{7.23}$$

$$|\widetilde S^+_{0x}(z) - \mathcal{S}^+_{0x}(z)| \lesssim \frac{W^2}{L^2}\frac{1}{W^d}\mathbf{1}_{|x|\leqslant W^{1+\tau}} + \langle x\rangle^{-D}, \tag{7.24}$$

$$|\widetilde\Theta_{0x}(z) - \Theta_{0x}(z)| \leqslant \frac{W^2}{L^2}\frac{1}{W^4(|x|+W)^{d-4-\tau}} + \langle x\rangle^{-D}, \tag{7.25}$$

for any constants $\tau, D > 0$.

Corresponding to the self-energies in Definition 2.13, we define $\widetilde{\mathcal{E}}_{2l}$, $4 \leqslant 2l \leqslant n$, as the sum of graphs obtained by replacing the S, $S^{\pm}$ and Θ edges in $\mathcal{E}_{2l}$ with the $\widetilde S$, $\widetilde S^{\pm}$ and $\widetilde\Theta$ edges. With Claim 7.8, we can show that $\widetilde{\mathcal{E}}_{2l}$ is sufficiently close to $\mathcal{E}_{2l}$.

Claim 7.9. *Under the assumptions of Lemma 5.10, fix any $L \geqslant W$ and $z = E + \mathrm{i}\eta$ with $E \in (-2+\kappa, 2-\kappa)$ and $W^2/L^{2-\varepsilon} \leqslant \eta \leqslant L^{-\varepsilon}$ for a small constant $\varepsilon > 0$. Then for any $4 \leqslant 2l \leqslant n$, we have that*

$$\left|(\widetilde{\mathcal{E}}_{2l})_{0x}(z) - (\mathcal{E}_{2l})_{0x}(z)\right| \leqslant W^{-ld}\frac{W^2}{L^2}\frac{W^{2d-6}}{\langle x\rangle^{2d-6-\tau}}, \quad \forall\, x \in \mathbb{Z}_L^d, \tag{7.26}$$

and

$$\Big|\sum_{x\in\mathbb{Z}_L^d}(\widetilde{\mathcal{E}}_{2l})_{0x}(z)\Big| \leqslant L^{\tau}W^{-(l-1)d}\frac{W^2}{L^2}, \tag{7.27}$$

for any constant $\tau > 0$.

Proof. We prove (7.26) and (7.27) by induction on l. First, we trivially have $\widetilde{\mathcal{E}}_2 = \mathcal{E}_2 = 0$. Now suppose we have shown that (7.26) and (7.27) hold for $\mathcal{E}_{2l}$ with $l \leqslant k-1$. Then with this induction hypothesis and the estimates (7.23)–(7.25), using the same argument as in the proof of Lemma 7.7, we can prove that (7.26) and (7.27) hold for $\mathcal{E}_{2k}$. □

Claim 7.9 shows that $\sum_{\mathfrak a}(\mathcal{E}_n)_{0\mathfrak a}$ has the same infinite space limit as $\sum_{\mathfrak a}(\widetilde{\mathcal{E}}_n)_{0\mathfrak a}$. Hence to prove Lemma 5.10, we first calculate the sum $\sum_{\mathfrak a}(\widetilde{\mathcal{E}}_n)_{0\mathfrak a}$ for a finite L, and then take $L \to \infty$. In the following proof, we choose $z \equiv z(L) = E + \mathrm{i}W^2/L^{2-\varepsilon}$ for a small enough constant $\varepsilon > 0$. Now we express $\sum_{\mathfrak a}(\widetilde{\mathcal{E}}_n)_{0\mathfrak a}$ using the Fourier series (5.36) and (5.38). For simplicity of notations, we denote the $\widetilde S$, $\widetilde S^{\pm}$, $\widetilde\Theta$ and labelled $\widetilde\Theta$ edge in a unified way as

$$\widetilde S^{(a)}_{xy}(z) = \frac{1}{L^d}\sum_{p\in\mathbb{T}_L^d}\psi_a(Wp,z)e^{\mathrm{i}p\cdot(x-y)},$$

for

$$a \in \{\emptyset, \pm, \Theta\} \cup \Big\{(k; 2k_1, \cdots, 2k_l) : l \geqslant 1, \max_i(2k_i) \leqslant n-1, k = \sum_{i=1}^{l} 2k_i - 2(l-1)\Big\},$$

where $\widetilde S^{(\emptyset)} := \widetilde S$, $\widetilde S^{(\pm)}(z) := \widetilde S^{\pm}(z)$, $\widetilde S^{(\Theta)}(z) := \widetilde\Theta(z)$ and $\widetilde S^{(k;2k_1,\cdots,2k_l)}$ corresponds to a labelled $\widetilde\Theta$ edge as in (6.9) (with Θ and $\mathcal{E}_{2k_i}$ replaced by $\widetilde\Theta$ and $\widetilde{\mathcal{E}}_{2k_i}$). The functions $\psi_a(Wp,z)$ are given by (5.36) and (5.38) for $a \in \{\emptyset, \pm, \Theta\}$, and we have

$$\psi_{(k;2k_1,\cdots,2k_l)}(Wp,z) = \psi_\Theta(Wp,z)^{l+1}\prod_{i=1}^{l}\psi_{\mathcal{E}_{2k_i}}(Wp,z), \tag{7.28}$$

where $\psi_{\widetilde{\mathcal{E}}_{2k_i}}(Wp,z)$ is the Fourier transform of $(\widetilde{\mathcal{E}}_{2k_i})_{0x}$ (which can be calculated inductively with respect to $2k_i$). For each edge e in $(\widetilde{\mathcal{E}}_n)_{0\mathfrak{a}}$, we assign a label a_e and a momentum p_e to it.

For a vertex x in the graph, suppose that it is connected with k edges with labels a_i, $1 \leqslant i \leqslant k$. Then summing over $x \in \mathbb{Z}_L^d$, we get that

$$\begin{aligned}\sum_{x\in\mathbb{Z}_L^d}\prod_{i=1}^k \widetilde{S}^{(a_i)}_{xy_i} &= \sum_{x\in\mathbb{Z}_L^d}\prod_{i=1}^k \frac{1}{L^d}\sum_{p_i\in\mathbb{T}_L^d}\psi_{a_i}(Wp_i,z)e^{\mathrm{i}p_i\cdot(x-y_i)} \\ &= \frac{1}{L^{(k-1)d}}\sum_{\substack{p_1,\cdots,p_k\in\mathbb{T}_L^d:\\ p_1+p_2+\cdots+p_k=0 \mod 2\pi}}\prod_{i=1}^k \psi_{a_i}(Wp_i,z)e^{-\mathrm{i}p_i\cdot y_i},\end{aligned}$$

where for a vector $v\in\mathbb{R}^d$, we use $v=0 \mod 2\pi$ to mean that $v_i = 0 \mod 2\pi$ for all $1\leqslant i\leqslant d$. Note that $p_1+\cdots+p_k = 0 \mod 2\pi$ is a momentum conservation condition. The momentum p_i associated with y_i will be used later in the summation over y_i, and so on. Let $\mathcal{G}$ denote the graphs in $\widetilde{\mathcal{E}}_n$, $c(\mathcal{G},z)$ be the coefficient of $\mathcal{G}$, p_e denote the momentum associated with each edge e in $\mathcal{G}$, Ξ_L be a subset of $(\mathbb{T}_L^d)^{n_e}$ given by the constraint that the total momentum at each vertex is equal to 0 modulo 2π, where $n_e\equiv n_e(\mathcal{G})$ is the total number of edges in $\mathcal{G}$. Then after summing over all indices in $(\widetilde{\mathcal{E}}_n)_{0\mathfrak{a}}$, we obtain that

$$\sum_{\mathfrak{a}}(\widetilde{\mathcal{E}}_n)_{0\mathfrak{a}}(m(z),\psi,W,L) = \frac{1}{L^{(n-2)d/2}}\sum_{\mathcal{G}}c(\mathcal{G},z)\sum_{\{p_e\}\in\Xi_L}\prod_e \psi_{a_e}(Wp_e,z). \tag{7.29}$$

Taking $L\to\infty$, (7.29) gives that

$$\sum_{\mathfrak{a}}(\mathcal{E}_n^\infty)_{0\mathfrak{a}} = \frac{1}{(2\pi)^{(n-2)d/2}}\sum_{\mathcal{G}}c(\mathcal{G},E)\int_{\{p_e\}\in\Xi}\prod_e\psi_{a_e}(Wp_e,E)\mathrm{d}p_e, \tag{7.30}$$

where Ξ is a union of hyperplanes in the torus $(-\pi,\pi]^{dn_e}$ given by the constraint that the total momentum at each vertex is equal to zero modulo 2π. To give a more rigorous proof of (7.30), we need to deal with the singularities of $\psi_{a_e}(Wp_e,E)$ at $p_e=0$ for (labelled) diffusive edges. We introduce an infrared cutoff on these edges, i.e. $\psi_{a_e,\varepsilon}(Wp_e,z) := \psi_{a_e}(Wp_e,z)\mathbf{1}_{|Wp_e|\leqslant\varepsilon}$. Then we define $\sum_{\mathfrak{a}}(\widetilde{\mathcal{E}}_{n,\varepsilon})_{0\mathfrak{a}}$ by replacing $\psi_{a_e}(Wp_e,z)$ with $\psi_{a_e,\varepsilon}(Wp_e,z)$ on the right-hand side of (7.29). Since $\psi_{a_e,\varepsilon}(Wp_e,z)$'s are nonsingular, taking $L\to\infty$ we readily get that

$$\sum_{\mathfrak{a}}(\mathcal{E}_{n,\varepsilon}^\infty)_{0\mathfrak{a}} = \frac{1}{(2\pi)^{(n-2)d/2}}\sum_{\mathcal{G}}c(\mathcal{G},E)\int_{\{p_e\}\in\Xi}\prod_e\psi_{a_e,\varepsilon}(Wp_e,E)\mathrm{d}p_e.$$

Then taking $\varepsilon\to 0$, we can show that this equation converges to (7.30), which again follows from the doubly connected property of the graphs $\mathcal{G}$ by using a similar argument as in the proof of Lemma 7.7. We omit the details.

Now applying a change of variables $q_e = Wp_e$ to (7.30), we get that

$$\sum_{\mathfrak{a}}(\mathcal{E}_n^\infty)_{0\mathfrak{a}} = \frac{1}{(2\pi)^{(n-2)d/2}W^{(n-2)d/2}}\sum_{\mathcal{G}}c(\mathcal{G},E)\int_{W\Xi}\prod_e\psi_{a_e}(q_e,E)\mathrm{d}q_e.$$

Since ψ is compactly supported in the assumption of Theorem 2.1, we have that

$$\frac{1}{(2\pi)^{(n-2)d/2}W^{(n-2)d/2}}\int_{W\Xi}\prod_e\psi_{a_e}(q_e,E)\mathrm{d}q_e = \frac{1}{(2\pi)^{(n-2)d/2}W^{(n-2)d/2}}\int_{\widetilde{\Xi}}\prod_e\psi_{a_e}(q_e,E)\mathrm{d}q_e, \tag{7.31}$$

where $\widetilde{\Xi}$ is a union of hyperplanes in $(\mathbb{R}^d)^{n_e}$ given by the constraint that the total momentum at each vertex is equal to 0 (without modulo 2π). Combining the above two equations, we obtain (5.35) by renaming

$$\mathfrak{S}_n(m(E),\psi) := \frac{1}{(2\pi)^{(n-2)d/2}}\sum_{\mathcal{G}}c(\mathcal{G},E)\int_{\widetilde{\Xi}}\prod_e\psi_{a_e}(q_e,E)\mathrm{d}q_e.$$

Remark 7.10. The equation (7.31) is the only place where the compactly supported condition of ψ is used. If we only assume that ψ is a Schwartz function, then equation (7.31) does not hold exactly, but with an additional error of order $\mathrm{O}(W^{-D})$ for any large constant $D>0$. Such a small error does not affect our proofs, and we refer the reader to Section 8 below for the necessary modifications of the proof in the setting of Theorem 1.4.

8 Proof of the main results

In this section, we complete the proofs of the main results—Theorem 1.3, Theorem 1.4, Theorem 1.5 and Corollary 1.6. First, we prove Theorem 1.3 using the local law (1.19). In fact, we will prove a slightly stronger result in Lemma 8.1. For any constants $M, K > 1$, we define the following random subset of indices that contains $B_{\gamma,K,\ell}$ as a subset:

$$\widetilde{\mathcal{B}}_{M,K,\ell} := \left\{ \alpha : \lambda_\alpha \in (-2+\kappa, 2-\kappa) \text{ so that } \min_{x_0} \sum_x |u_\alpha(x)|^2 \left(\frac{\|x-x_0\|_L}{\ell} + 1 \right)^M \leqslant K \right\}.$$

Note that this subset contains all indices associated with bulk eigenvectors that are localized *super-polynomially* in balls of radius $\mathrm{O}(\ell)$.

Lemma 8.1. *Suppose the assumptions of Theorem 1.3 hold. Fix any constants $c_0 > 0$ and $M, K > 1$. For any $W \leqslant \ell \leqslant L^{1-c_0}$, we have that*

$$|\widetilde{\mathcal{B}}_{M,K,\ell}|/N \prec \left(\ell^{\frac{M}{M+4}}/L^{\frac{M-d}{M+4}}\right)^2 + W^{-d/2}. \tag{8.1}$$

Using Theorem 1.4 and Lemma 8.1, we can easily conclude Theorem 1.3.

Proof of Theorem 1.3. Since $\mathcal{B}_{\gamma,K,\ell} \subset \widetilde{\mathcal{B}}_{M,C_M K,\ell}$ for arbitrarily large M and a constant $C_M > 0$, we obtain from (8.1) that

$$|\mathcal{B}_{\gamma,K,\ell}|/N \prec (\ell/L)^2 + W^{-d/2}.$$

This implies (1.12) by Definition 2.3. To prove (1.13), using the spectral decomposition of $G(z)$, we obtain from (1.19) that

$$|u_\alpha(x)|^2 \leqslant \eta \operatorname{Im} G_{xx}(\lambda_\alpha + \mathrm{i}\eta) \prec W^2/L^{2-\varepsilon},$$

for $\eta = W^2/L^{2-\varepsilon}$. Since $\varepsilon > 0$ is arbitrarily small, we conclude (1.13). □

Now we give the proof of Lemma 8.1 based on Theorem 1.4.

Proof of Lemma 8.1. We define the following characteristic function $P_{x,l}$ projecting onto the complement of the ℓ-neighborhood of x: $P_{x,\ell}(y) := \mathbf{1}(|y-x| \geqslant \ell)$. Define the following random subset of indices

$$\mathcal{A}_{\delta,\ell} := \Big\{ \alpha : \lambda_\alpha \in (-2+\kappa, 2-\kappa), \sum_x |u_\alpha(x)| \|P_{x,\ell}\mathbf{u}_\alpha\| \leqslant \delta \Big\},$$

where $\delta \equiv \delta(L)$ may depend on L and is not necessarily a constant. Using Theorem 1.4, we get that

$$\max_{x,y} |G_{xy}(z) - m(z)\delta_{xy}| \prec W^{-d/2}, \qquad \max_{x \in \mathbb{Z}_L^d} \eta \sum_{y:|y-x| \leqslant \ell} |G_{xy}|^2 \prec \eta\ell^2/W^2 \leqslant \ell^2/L^{2-\varepsilon},$$

if we take $\eta = W^2/L^{2-\varepsilon}$. With these estimates, following the proof of Proposition 7.1 of [27], we can obtain that

$$|\mathcal{A}_{\delta,\ell}|/N \leqslant C\sqrt{\delta} + \mathrm{O}_\prec\left(\ell^2/L^{2-\varepsilon} + W^{-d/2}\right). \tag{8.2}$$

Next we use a similar argument as in the proof of [25, Corollary 3.4] to derive the estimate (8.1) from (8.2). Let $\widetilde{\ell} := \ell L^{c_1}$ for a constant $c_1 \in (0, c_0)$. If $\alpha \in \widetilde{\mathcal{B}}_{M,K,\ell}$, then for any $x_0 \in \mathbb{Z}_L^d$ we have that

$$\begin{aligned}
\left(\sum_x |u_\alpha(x)| \|P_{x,\widetilde{\ell}}\mathbf{u}_\alpha\| \right)^2 &\leqslant \left[\sum_x |u_\alpha(x)|^2 \left(\frac{\|x-x_0\|_L}{\ell} + 1 \right)^M \right] \left[\sum_x \|P_{x,\widetilde{\ell}}\mathbf{u}_\alpha\|^2 \left(\frac{\|x-x_0\|_L}{\ell} + 1 \right)^{-M} \right] \\
&\leqslant K \left(\frac{\ell}{\widetilde{\ell}} \right)^M \sum_{x,y:\|x-y\|_L \geqslant \widetilde{\ell}} \|u_\alpha(y)\|^2 \left(\frac{\|x-x_0\|_L}{\ell} + 1 \right)^{-M} \left(\frac{\|x-y\|_L}{\ell} + 1 \right)^M \\
&\leqslant K \left(\frac{\ell}{\widetilde{\ell}} \right)^M \sum_{x,y} \|u_\alpha(y)\|^2 \left(\frac{\|y-x_0\|_L}{\ell} + 1 \right)^M \leqslant K^2 \left(\frac{\ell}{\widetilde{\ell}} \right)^M L^d =: \delta_L^2.
\end{aligned}$$

Thus we have proved that $\widetilde{\mathcal{B}}_{M,K,\ell} \subset \mathcal{A}_{\delta_L,\widetilde{\ell}}$. Then we get from (8.2) that

$$\left|\widetilde{\mathcal{B}}_{M,K,\ell}\right|/N \leqslant CK\left(\ell/\widetilde{\ell}\right)^{M/2} L^{d/2} + \mathrm{O}_{\prec}\left(\widetilde{\ell}^2/L^{2-\varepsilon} + W^{-d/2}\right) \prec L^{\varepsilon}\left(\ell^{\frac{M}{M+4}}/L^{\frac{M-d}{M+4}}\right)^2 + W^{-d/2},$$

where in the second step we minimized the sum over $\widetilde{\ell}$. Since ε is arbitrarily small, we conclude (8.1). $\square$

The proof of Theorem 1.4 is almost the same as the one for Theorem 2.1, except for some minor differences regarding the infinite space limits of the self-energies. Here we only describe the necessary modifications to the arguments in Section 5, without writing down all the details of the proof of Theorem 1.4. First, fix any $n \in \mathbb{N}$, we define the renormalized self-energies $\mathcal{E}_l^{(r)}$, $4 \leqslant l \leqslant n$, as follows.

Definition 8.2 (Renormalized self-energies)**.** *Let $z = E + \mathrm{i}\eta$ with $E \in (-2+\kappa, 2-\kappa)$ and $W^2/L^{2-\varepsilon} \leqslant \eta \leqslant L^{-\varepsilon}$ for a small constant $\varepsilon > 0$. Let $\mathcal{E}_l$, $4 \leqslant l \leqslant n-1$, be a sequence of self-energies satisfying Definition 2.13, and $\mathcal{E}_n$ be a sum of scaling order n deterministic graphs constructed in the n-th order T-equation in Lemma 5.7. Then we define the renormalized self-energies inductively as follows. First, we define $\mathcal{E}_4^{(r)} = \mathcal{E}_4$ and $\mathcal{E}_{2l+1}^{(r)} = 0$. Suppose we have defined $\mathcal{E}_l^{(r)}$ for all $l \leqslant k-1$. Then we define $\mathcal{E}_k^{(r)}$ as the sum of deterministic graphs obtained from $\mathcal{E}_k$ by replacing all the lower order self-energies $\mathcal{E}_l$, $4 \leqslant l \leqslant k-1$, in it with*

$$\mathring{\mathcal{E}}_l := \mathcal{E}_l^{(r)} - \Big[\sum_x (\mathcal{E}_l^{(r)})_{0x}\Big] \cdot S. \tag{8.3}$$

By definition, we trivially have that $\mathring{\mathcal{E}}_l$ and its infinite space limit $\mathring{\mathcal{E}}_l^{\infty}$ satisfy the sum zero properties. In particular, the sum zero properties of $\mathring{\mathcal{E}}_l^{\infty}$, $4 \leqslant l \leqslant n-1$, are necessary for $\mathcal{E}_n^{(r),\infty}$, the infinite space limit of $\mathcal{E}_n^{(r)}$, to be well-defined. For example, we consider the function in (7.28). In the proof below, we will see that $\sum_{x\in\mathbb{Z}^d} (\mathcal{E}_l^{\infty})_{0x} = \mathrm{O}(W^{-D})$ under the assumptions of Theorem 1.4, which gives that $\psi_{\mathcal{E}_{2k_i}}(Wp, E) = \mathrm{O}(W^2|p|^2 + W^{-D})$ for $p \ll W^{-1}$. If we do not perform the renormalization in Definition 8.2, then we have

$$\left|\psi_{(k;2k_1,\cdots,2k_l)}(Wp,E)\right| \lesssim W^{-2(l+1)}|p|^{-2(l+1)}(W^2|p|^2 + W^{-D})^{2l} \quad \text{for } |p| \ll W^{-1}.$$

Thus it may give a non-integrable singularity around $p = 0$ in the infinite space limit. (However, notice that W^{-D} is negligible in finite space with $W \geqslant L^{\varepsilon}$ for a constant $\varepsilon > 0$.) In (8.3), the matrix S can be replaced by any doubly stochastic matrix whose Fourier transform is a Schwartz function (e.g. S^k for any fixed $k \in \mathbb{N}$).

Using properties (1.25)–(1.27) for $\mathcal{E}_l$, $4 \leqslant l \leqslant n-1$, we can obtain the following result.

Lemma 8.3. *Under the assumptions of Theorem 1.4 and in the setting of Definition 8.2, for $4 \leqslant l \leqslant n$, we have that*

$$\left|(\mathcal{E}_l^{(r)})_{0x}(z) - (\mathcal{E}_l)_{0x}(z)\right| \leqslant W^{-ld/2}\frac{\eta W^{2d-6}}{\langle x\rangle^{2d-6-\tau}}, \quad \forall\, x \in \mathbb{Z}_L^d, \tag{8.4}$$

and

$$\Big|\sum_x (\mathcal{E}_l^{(r)})_{0x}(z) - \sum_x (\mathcal{E}_l)_{0x}(z)\Big| \leqslant L^{\tau}\eta W^{-(l-2)d/2}, \tag{8.5}$$

for any constant $\tau > 0$.

Proof. We prove (8.4) and (8.5) by induction on l. First, we trivially have $\mathcal{E}_4^{(r)} = \mathcal{E}_4$. Then suppose we have shown that (8.4) and (8.5) hold for $\mathcal{E}_l^{(r)}$ for all $l \leqslant k-1$. Combining this induction hypothesis with (1.27), we get that

$$\Big|\sum_x (\mathcal{E}_l^{(r)})_{0x}(z)\Big| \leqslant L^{\tau}\eta W^{-(l-2)d/2}.$$

Then using (8.3) and (8.4), it is trivial to see that

$$\left|(\mathring{\mathcal{E}}_l)_{0x}(z) - (\mathcal{E}_l)_{0x}(z)\right| \leqslant W^{-ld/2}\frac{\eta W^{2d-6}}{\langle x\rangle^{2d-6-\tau}}, \quad \forall\, x \in \mathbb{Z}_L^d, \quad l \leqslant k-1.$$

Now using the same argument as in the proof of Lemma 7.7, we can get that (8.4) and (8.5) hold for $\mathcal{E}_k$. $\square$

We can obtain the following result on the infinite space limits $\mathcal{E}_l^{(r),\infty}$ of $\mathcal{E}_l^{(r)}$, $4 \leqslant l \leqslant n$.

Lemma 8.4. *Under the assumptions of Theorem 1.4 and in the setting of Definition 8.2, for* $4 \leqslant l \leqslant n$*, we have that*

$$\mathcal{E}_l^{(r),\infty}(x, x+a) = \mathcal{E}_l^{(r),\infty}(0, a), \quad \mathcal{E}_l^{(r),\infty}(0, a) = \mathcal{E}_l^{(r),\infty}(0, -a), \quad \forall\, x, a \in \mathbb{Z}^d, \tag{8.6}$$

and

$$\left|(\mathcal{E}_l^{(r),\infty})_{0x}\right| \leqslant W^{-ld/2} \frac{W^{2d-4}}{\langle x\rangle^{2d-4-\tau}}, \quad \forall\, x \in \mathbb{Z}^d, \tag{8.7}$$

for any constant $\tau > 0$*. Furthermore, we have that for any fixed* $D > 0$*,*

$$\sum_{x\in\mathbb{Z}^d} (\mathcal{E}_n^{(r),\infty})_{0x}(m(E), \psi, W) = W^{-(n-2)d/2}\mathfrak{S}_n(m(E), \psi) + \mathrm{O}(W^{-D}), \tag{8.8}$$

where $\mathfrak{S}_n$ *is a constant that does not depend on* W*.*

Proof. The property (8.6) follows from Lemma A.1, and the estimate (8.7) follows from the doubly connected property by Corollary 7.6. Equation (8.8) can be proved using the same argument as in Section 7.2. There is only one difference that has been discussed in Remark 7.10—the equation (7.31) does not hold exactly if ψ is not compactly supported. However, using the fact that ψ is a Schwartz function, we get that (7.31) holds up to a small error $\mathrm{O}(W^{-D})$, which leads to the extra $\mathrm{O}(W^{-D})$ in (8.8). We omit the details. ☐

As in Lemma 7.7, we can bound the difference between $\mathcal{E}_l^{(r)}$ and $\mathcal{E}_l^{(r),\infty}$.

Lemma 8.5. *Under the assumptions of Theorem 1.4 and in the setting of Definition 8.2, for* $4 \leqslant l \leqslant n$*, we have that for any constant* $\tau > 0$*,*

$$\left|(\mathcal{E}_l^{(r)})_{0x}(m(z), \psi, W, L) - (\mathcal{E}_l^{(r),\infty})_{0x}(m(E), \psi, W)\right| \leqslant W^{-ld/2} \frac{\eta W^{2d-6}}{\langle x\rangle^{2d-6-\tau}}, \quad \forall\, x \in \mathbb{Z}_L^d, \tag{8.9}$$

and

$$\Big|\sum_{x\in\mathbb{Z}_L^d} (\mathcal{E}_l^{(r)})_{0x}(m(z), \psi, W, L) - \sum_{x\in\mathbb{Z}^d} (\mathcal{E}_l^{(r),\infty})_{0x}(m(E), \psi, W)\Big| \leqslant L^\tau \eta W^{-(l-2)d/2}. \tag{8.10}$$

Proof. We prove this lemma by induction on l. First, (8.9) and (8.10) trivially hold for $\mathcal{E}_4^{(r)} = \mathcal{E}_4$. Now suppose we have shown that (8.9) and (8.10) hold for $\mathcal{E}_l^{(r)}$ for all $l \leqslant k-1$. Then with this induction hypothesis and the same argument as in the proof of Lemma 7.7, we can show that(8.9) and (8.10) hold for $\mathcal{E}_k^{(r)}$. We omit the details. ☐

Now we are ready to complete the proof of Theorem 1.4.

Proof of Theorem 1.4. We repeat the five-step strategy in Section 5.1, where the steps 1, 2, 3 and 5 stay the same as in the proof of Theorem 2.1, because these steps only involve the self-energies $\mathcal{E}_l$ but do not use the infinite space limits $\mathcal{E}_l^\infty(z)$ at any place. Regarding Step 4, we need to prove a counterpart of Lemma 5.8 in the setting of Theorem 1.4. The properties (1.25) and (1.26) follow from Lemma A.1 and Corollary 6.12. The properties (2.13), (2.14) and (2.15) will be replaced by (8.6), (8.7) and (8.9). It remains to prove the following sum zero properties:

$$\Big|\sum_{x\in\mathbb{Z}_L^d} (\mathcal{E}_n)_{0x}(m(z), \psi, W, L)\Big| \leqslant L^\tau \eta W^{-(n-2)d/2}, \quad \forall\, \eta \in [W^2/L^{2-\varepsilon}, L^{-\varepsilon}], \tag{8.11}$$

and

$$\Big|\sum_{x\in\mathbb{Z}^d} (\mathcal{E}_n^{(r),\infty})_{0x}(m(E), \psi, W)\Big| \leqslant W^{-D}, \tag{8.12}$$

for any constants $\tau, D > 0$. By (8.5) and (8.10), we have that

$$\Big|\sum_{x\in\mathbb{Z}_L^d} (\mathcal{E}_n)_{0x}(m(z), \psi, W, L) - \sum_{x\in\mathbb{Z}^d} (\mathcal{E}_n^{(r),\infty})_{0x}(m(E), \psi, W)\Big| \leqslant L^\tau \eta W^{-(n-2)d/2}. \tag{8.13}$$

Hence the estimate (8.11) is a consequence of (8.12).

The proof of (8.12) is similar to the one for Lemma 5.8. Since the proof of Lemma 5.11 does not involve infinite space limits, the estimate (5.40) also holds in the current setting for L_n satisfying (5.39). Combining this estimate with (8.13) and (8.8), we can obtain that $\mathfrak{S}_n(m(E),\psi) = \mathrm{o}(1)$, which gives $\mathfrak{S}_n(m(E),\psi) = 0$. Together with (8.8), it implies (8.12), and hence completes Step 4 of the five-step strategy in Section 5.1. Finally, applying the argument in Figure 2, we complete the proof of Theorem 1.4. □

Finally, we give the proof of Theorem 1.5 and Corollary 1.6.

Proof of Theorem 1.5. By Theorem 1.4, we know that $G(z)$ satisfies the local law (1.19). Moreover, in the proof of Theorem 1.4, we have constructed the M-th order T-equation (2.28) with $n = M$. Setting $\mathfrak{b}_1 = \mathfrak{b}_2 := \mathfrak{b}$ in (2.28), solving $T_{\mathfrak{a}\mathfrak{b}}$ and taking expectation, we get that

$$\begin{aligned}\mathbb{E}T_{\mathfrak{a}\mathfrak{b}} = |m|^2\left(\frac{1}{1-\Theta\Sigma^{(M)}}\Theta\right)_{\mathfrak{a}\mathfrak{b}} + \sum_x \left(\frac{1}{1-\Theta\Sigma^{(M)}}\right)_{\mathfrak{a}x}\mathbb{E}\left[m\Theta_{x\mathfrak{b}}\left(\overline{G}_{\mathfrak{b}\mathfrak{b}}-\overline{m}\right) + (\mathcal{R}_{IT}^{(M)})_{x,\mathfrak{b}\mathfrak{b}}\right] \\ + \sum_x \left(\frac{1}{1-\Theta\Sigma^{(n)}}\right)_{\mathfrak{a}x}\mathbb{E}\left[(\mathcal{A}_{IT}^{(>M)})_{x,\mathfrak{b}\mathfrak{b}} + (\mathcal{E}rr'_{M,D})_{x,\mathfrak{b}\mathfrak{b}}\right]. \qquad (8.14)\end{aligned}$$

Recall that $(\mathcal{R}_{IT}^{(M)})_{x,\mathfrak{b}\mathfrak{b}}$, $(\mathcal{A}_{IT}^{(>M)})_{x,\mathfrak{b}\mathfrak{b}}$ and $(\mathcal{E}rr'_{M,D})_{x,\mathfrak{b}\mathfrak{b}}$ can be written into the forms in (5.55). Plugging them into (8.14) and using the identity $\Theta^{(M)} = (1-\Theta\Sigma^{(M)})^{-1}\Theta$, we obtain that

$$\mathbb{E}T_{\mathfrak{a}\mathfrak{b}} = |m|^2\Theta^{(M)}_{\mathfrak{a}\mathfrak{b}} + \left(\Theta^{(M)}\mathcal{G}^{(M)}\right)_{\mathfrak{a}\mathfrak{b}} + \sum_x \Theta^{(M)}_{\mathfrak{a}x}\sum_\omega \mathbb{E}\left(\mathcal{G}^{err}_\omega\right)_{x\mathfrak{b}}, \qquad (8.15)$$

where $\mathcal{G}^{(M)}$ is defined as

$$\mathcal{G}^{(M)}_{x\mathfrak{b}} := m\delta_{x\mathfrak{b}}\mathbb{E}\left(\overline{G}_{\mathfrak{b}\mathfrak{b}}-\overline{m}\right) + \mathbb{E}(\Gamma_R^{(n)})_{x,\mathfrak{b}\mathfrak{b}}, \qquad (8.16)$$

and $\mathcal{G}^{err}_\omega$ are the graphs in $(\Gamma_A^{(>n)})_{x,\mathfrak{b}\mathfrak{b}}$ and $(\Gamma_{err}^{(n,D)})_{x,\mathfrak{b}\mathfrak{b}}$, i.e. $(\Gamma_A^{(>n)})_{x,\mathfrak{b}\mathfrak{b}} + (\Gamma_{err}^{(n,D)})_{x,\mathfrak{b}\mathfrak{b}} = \sum_\omega \left(\mathcal{G}^{err}_\omega\right)_{x\mathfrak{b}}$. To conclude the proof, it remains to prove (1.24), (1.28) and that

$$\sum_x \left|\Theta^{(M)}_{\mathfrak{a}x}\mathbb{E}\left(\mathcal{G}^{err}_\omega\right)_{x\mathfrak{b}}\right| \leqslant W^{-dM/2}. \qquad (8.17)$$

First, we can expand $\Theta^{(M)}$ as

$$\Theta^{(M)} = \left[1-(\Theta\Sigma^{(M)})^{K+1}\right]^{-1}\sum_{k=0}^K(\Theta\Sigma^{(M)})^k\Theta, \qquad (8.18)$$

for a large constant $K\in\mathbb{N}$. Using (6.1) and (6.2), we can obtain that

$$\left\|\Theta\Sigma^{(M)}\right\|_{\ell^\infty\to\ell^\infty} \prec W^{-d}, \quad \text{and} \quad \left[(\Theta\Sigma^{(M)})^k\Theta\right]_{xy} \prec B_{xy}. \qquad (8.19)$$

Combining (8.18) with (8.19), we get that

$$\left|\Theta^{(M)}_{xy}\right| \prec B_{xy} + W^{-(K+1)d} \leqslant 2B_{xy},$$

as long as K is chosen to be sufficiently large. This concludes (1.24). Second, we notice that every graph in $\mathcal{G}^{(M)}$ is doubly-connected. Then using (6.20), we immediately conclude (1.28). Finally, we prove (8.17). Each graph $(\mathcal{G}^{err}_\omega)_{x\mathfrak{b}}$ can be written into

$$\left(\mathcal{G}^{err}_\omega\right)_{x\mathfrak{b}} = \sum_{y,y'}(\mathcal{G}_0)_{x,yy'}G_{y\mathfrak{b}}\overline{G}_{y'\mathfrak{b}}, \quad \text{or} \quad \left(\mathcal{G}^{err}_\omega\right)_{x\mathfrak{b}} = \sum_{x,y}(\mathcal{G}_0)_{xy}\Theta_{y\mathfrak{b}}, \qquad (8.20)$$

or some forms obtained by setting some indices of x,y,y' to be equal to each other. Without loss of generality, we only consider the first form in (8.20), while all the other forms are easier to bound. By Definition 6.7, the graph $\mathcal{G}_0$ is doubly connected. Then we can get the bound

$$\sum_x \left|\Theta^{(M)}_{\mathfrak{a}x}\mathbb{E}\left(\mathcal{G}^{err}_\omega\right)_{x\mathfrak{b}}\right| \prec \sum_{x,y,y'} B_{\mathfrak{a}x}\left|\mathbb{E}(\mathcal{G}_0)_{x,yy'}G_{y\mathfrak{b}}\overline{G}_{y'\mathfrak{b}}\right| \prec W^{-d/2}\sum_{x,y}B_{\mathfrak{a}x}\mathbb{E}\sum_{y'}\left(\mathcal{G}_0^{\mathrm{abs}}\right)_{x,yy'}B^{1/2}_{y\mathfrak{b}}$$

$$\prec W^{-d/2}W^{-(M-2)d/2}\sum_{x,y}B_{\mathfrak{a}x}B_{xy}^{3/2}B_{y\mathfrak{b}}^{1/2}\prec W^{-(M+1)d/2},$$

where in the second step we used (1.19) and (1.24), and in the third step we used that $\sum_{y'}(\mathcal{G}_0^{\rm abs})_{x,yy'}$ is a graph satisfying the assumptions of Lemma 6.10 with two fixed atoms x and y, so that it satisfies (6.20). This concludes (8.17). $\square$

Proof of Corollary 1.6. If S is invertible, then multiplying both sides of (8.15) by $|m|^{-2}S^{-1}$, we obtain that

$$\mathbb{E}|G_{\mathfrak{a}\mathfrak{b}}|^2=|m|^2\left[\frac{1}{1-\left(1+\Sigma^{(M)}\right)|m|^2S}\right]_{\mathfrak{a}\mathfrak{b}}+\left[\frac{1}{1-\left(1+\Sigma^{(M)}\right)|m|^2S}\mathcal{G}^{(M)}\right]_{\mathfrak{a}\mathfrak{b}}$$
$$+\sum_x\left[\frac{1}{1-\left(1+\Sigma^{(M)}\right)|m|^2S}\right]_{\mathfrak{a}x}\sum_\omega\mathbb{E}\left(\mathcal{G}^{err}_\omega\right)_{x\mathfrak{b}}.$$

The last term can be bounded by $\mathrm{O}(W^{-Md/2})$ using the same argument as the one below (8.20), and we omit the details.

On the other hand, if S is singular, we can choose $0<\varepsilon_N<L^{-100Md}$ so that $S+\varepsilon_N I$ is nonsingular. Then we define another random band matrix $\widetilde{H}$ with variance profile $\widetilde{S}:=(S+\varepsilon_N I)/(1+\varepsilon_N)$ and denote its resolvent by $\widetilde{G}(z):=(\widetilde{H}-z)^{-1}$. The above argument shows that (1.29) holds for $\mathbb{E}|\widetilde{G}_{xy}|^2$. Moreover, it is easy to show that $\mathbb{E}|G_{xy}|^2=\mathbb{E}|\widetilde{G}_{xy}|^2+\mathrm{O}(W^{-Md/2})$ with a simple perturbation argument. $\square$

9 Main ideas for Lemmas 5.4, 5.6 and 5.7

In this section, we discuss some key ideas that will be used in the proofs of three key lemmas, Lemmas 5.4, 5.6 and 5.7, in [64].

Main idea for Lemma 5.4. To prove Lemma 5.4, it suffices to prove the following self-improving estimate on T-variables. If $T_{xy}\prec B_{xy}+\widetilde{\Phi}^2$ for a deterministic parameter $\widetilde{\Phi}$, then

$$T_{xy}(z)\prec B_{xy}+W^{-c_1}\widetilde{\Phi}^2 \tag{9.1}$$

for a constant $c_1>0$ depending only on d and c_0 in (5.4). Iterating this estimate for D/c_1 many times, we will get that $T_{xy}(z)\prec B_{xy}+W^{-D}$. This concludes (5.10) as long as D is large enough.

The estimate (9.1) follows from the high moment bound for any fixed $p\in\mathbb{N}$:

$$\mathbb{E}T_{\mathfrak{a}\mathfrak{b}}(z)^p\prec(B_{\mathfrak{a}\mathfrak{b}}+W^{-c}\widetilde{\Phi}^2)^p,\quad\forall\ \mathfrak{a},\mathfrak{b}\in\mathbb{Z}_L^d. \tag{9.2}$$

We regard $T_{\mathfrak{a}\mathfrak{b}}(z)^p$ as a graph with p copies of $T_{\mathfrak{a}\mathfrak{b}}(z)$. Now we replace one of them with the n-th order T-expansion. If we replace $T_{\mathfrak{a}\mathfrak{b}}$ with the first two terms on the right-hand side of (2.19), then using (1.17) and (2.25) we can bound that

$$\left|\mathbb{E}T_{\mathfrak{a}\mathfrak{b}}^{p-1}m\overline{G}_{\mathfrak{b}\mathfrak{b}}\left[\Theta_{\mathfrak{a}\mathfrak{b}}+\left(\Theta\Sigma_T^{(n)}\Theta\right)_{\mathfrak{a}\mathfrak{b}}\right]\right|\prec B_{\mathfrak{a}\mathfrak{b}}\mathbb{E}T_{\mathfrak{a}\mathfrak{b}}^{p-1}. \tag{9.3}$$

Next we replace $T_{\mathfrak{a}\mathfrak{b}}$ with a graph $\mathcal{G}_{\mathfrak{a}\mathfrak{b}}$ in $(\mathcal{R}_{T,k})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}$, $k\geqslant 3$. It can be written into the forms in (6.44) or some variants of them with Θ replaced by a labelled diffusive edge. As an example, if $\mathcal{G}_{\mathfrak{a}\mathfrak{b}}=\sum_x\Theta_{\mathfrak{a}x}(\mathcal{G}_0)_{x\mathfrak{b}}$, then using (1.17) and (6.40), we can bound that

$$|\mathcal{G}_{\mathfrak{a}\mathfrak{b}}|\prec W^{(k-3)(-d/2+\varepsilon_0)+\varepsilon_0}\sum_x B_{\mathfrak{a}x}B_{x\mathfrak{b}}\mathcal{A}_{x\mathfrak{b}}\prec W^{(k-2)(-d/2+\varepsilon_0)}B_{\mathfrak{a}\mathfrak{b}}, \tag{9.4}$$

where $\mathcal{A}_{x\mathfrak{b}}$ is a variable satisfying $\|\mathcal{A}\|_{w;(1,2)}\prec 1$ and in the last step we used (6.32). So with (9.4) and the fact $k\geqslant 3$, we can bound that

$$\left|\mathbb{E}T_{\mathfrak{a}\mathfrak{b}}^{p-1}(\mathcal{R}_T^{(n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}\right|\prec W^{-d/2+\varepsilon_0}B_{\mathfrak{a}\mathfrak{b}}\mathbb{E}T_{\mathfrak{a}\mathfrak{b}}^{p-1}. \tag{9.5}$$

Then we replace $T_{\mathfrak{a}\mathfrak{b}}$ with a graph $\mathcal{G}_{\mathfrak{a}\mathfrak{b}}$ in $(\mathcal{A}_T^{(>n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}$. It can be written into the forms in (6.46) or some variants of them. As an example, if $\mathcal{G}_{\mathfrak{a}\mathfrak{b}} = \sum_x \Theta_{\mathfrak{a}x}(\mathcal{G}_0)_{xy}|G_{y\mathfrak{b}}|^2$, then using (1.17), (6.40) and $\text{ord}((\mathcal{G}_0)_{xy}) > n$, we can bound that

$$|\mathcal{G}_{\mathfrak{a}\mathfrak{b}}| \prec W^{(n-2)(-d/2+\varepsilon_0)+\varepsilon_0} \sum_{x,y} B_{\mathfrak{a}x} B_{xy} \mathcal{A}_{xy} \left(B_{y\mathfrak{b}} + \widetilde{\Phi}^2\right) \lesssim W^{(n-1)(-d/2+\varepsilon_0)} \sum_y B_{\mathfrak{a}y} \left(B_{y\mathfrak{b}} + \widetilde{\Phi}^2\right)$$
$$\lesssim W^{(n-1)(-d/2+\varepsilon_0)} \left(\frac{1}{W^4 \langle \mathfrak{a} - \mathfrak{b} \rangle^{d-4}} + \frac{L^2}{W^2} \widetilde{\Phi}^2\right) \lesssim W^{-c_0+(n-1)\varepsilon_0} \left(B_{\mathfrak{a}\mathfrak{b}} + \widetilde{\Phi}^2\right), \tag{9.6}$$

where we used (6.32) in the second step, $\sum_y B_{\mathfrak{a}y} B_{y\mathfrak{b}} \lesssim W^{-4} \langle \mathfrak{a} - \mathfrak{b} \rangle^{-(d-4)}$ and $\sum_y B_{\mathfrak{a}y} \lesssim L^2/W^2$ in the third step, and $\langle \mathfrak{a} - \mathfrak{b} \rangle \leqslant L$ and (2.26) in the fourth step. With (9.6), we obtain that

$$\left| \mathbb{E} T_{\mathfrak{a}\mathfrak{b}}^{p-1} (\mathcal{A}_T^{(>n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}} \right| \prec W^{-c_0+(n-1)\varepsilon_0} \left(B_{\mathfrak{a}\mathfrak{b}} + \widetilde{\Phi}^2\right) \mathbb{E} T_{\mathfrak{a}\mathfrak{b}}^{p-1}. \tag{9.7}$$

Finally, if we replace $T_{\mathfrak{a}\mathfrak{b}}$ with $(\mathcal{Q}_T^{(n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}}$, then we apply the Q-expansions mentioned in Section 3.5 to $T_{\mathfrak{a}\mathfrak{b}}^{p-1} (\mathcal{Q}_T^{(n)})_{\mathfrak{a}\mathfrak{b}}$, and show that

$$\left| \mathbb{E} T_{\mathfrak{a}\mathfrak{b}}^{p-1} (\mathcal{Q}_T^{(n)})_{\mathfrak{a},\mathfrak{b}\mathfrak{b}} \right| \prec \sum_{k=2}^{p} \left[W^{-d/4+\varepsilon_0/2} \left(B_{\mathfrak{a}\mathfrak{b}} + \widetilde{\Phi}^2\right) \right]^k \mathbb{E} T_{\mathfrak{a}\mathfrak{b}}^{p-k}. \tag{9.8}$$

The details will be given in [64]. Combining (9.3), (9.5), (9.7) and (9.8), and applying Hölder's inequality and Young's inequality to each term, we obtain that

$$\mathbb{E} T_{\mathfrak{a}\mathfrak{b}}^p \prec W^{\varepsilon} \left(B_{\mathfrak{a}\mathfrak{b}} + W^{-c_1} \widetilde{\Phi}^2\right)^p + W^{-\varepsilon} \mathbb{E} T_{\mathfrak{a}\mathfrak{b}}^p \quad \Rightarrow \quad \mathbb{E} T_{\mathfrak{a}\mathfrak{b}}^p \prec W^{\varepsilon} \left(B_{\mathfrak{a}\mathfrak{b}} + W^{-c_1} \widetilde{\Phi}^2\right), \tag{9.9}$$

for $c_1 := \min(c_0 - (n-1)\varepsilon_0, d/4 - \varepsilon_0/2)$ and any constant $\varepsilon > 0$. This concludes (9.2) as long as ε_0 is sufficiently small.

Main idea for Lemma 5.6. We expand $\mathbb{E}\,\text{Tr}\left(\mathcal{A}^{2p}\right)$ as

$$\mathbb{E}\,\text{Tr}\left(\mathcal{A}^{2p}\right) = \sum_{x_1,\cdots,x_{2p} \in \mathcal{I}} \sum_{a_1,\cdots,a_{2p}} c(a_1,\cdots,a_{2p}) \prod_{i=1}^{2p} G_{x_i x_{i+1}}^{a_i}, \quad a_i \in \{+,-\}, \tag{9.10}$$

where we adopted the conventions $x_{2p+1} \equiv x_1$, $G_{xy}^+ \equiv G_{xy}$ and $G_{xy}^- \equiv \overline{G}_{xy}$, and each $c(a_1,\cdots,a_{2p})$ is a deterministic coefficient of order $\mathrm{O}(1)$. To conclude (5.27), it suffices to show that

$$\sum_{x_1,\cdots,x_{2p} \in \mathcal{I}} \mathbb{E} \prod_{i=1}^{2p} G_{x_i x_{i+1}}^{a_i} \leqslant K^d \left(W^{\varepsilon} \frac{K^4}{W^4}\right)^{2p-1}, \quad \forall\, (a_1,\cdots,a_{2p}) \in \{+,-\}^{2p}. \tag{9.11}$$

We regard the above graph as a $2p$-gon graph with $2p$ external vertices $x_1,\cdots,x_{2p}$. We will expand it using the operations defined in Section 3, and a similar expansion strategy as the one for Lemma 5.7 that will be introduced below. In the expansions, we will get internal molecules. Our goal is to expand every $2p$-gon graph into a linear combination of connected deterministic graphs that satisfy a weaker doubly connected property: there exist two disjoint nets $\mathcal{B}_{black}$ and $\mathcal{B}_{blue}$ of black and blue diffusive edges, so that each internal molecule connects to external molecules through a path of edges in $\mathcal{B}_{black}$ and a path of edges in $\mathcal{B}_{blue}$. (Note that if we remove the external molecules from these graphs, the remaining internal molecules do not form doubly connected graphs, so this new property is weaker than the one in Definition 6.5.) Such deterministic graphs will satisfy the bound in (9.11).

Main idea for Lemma 5.7. The proof of Lemma 5.7 is based on a carefully designed global expansion strategy. This strategy is also used in the proof of Lemma 5.6 as discussed above.

The main difficulty with our expansions is how to maintain the doubly connected structures of the graphs. It is not hard to check that local expansions will not affect the doubly connected property. However, this is not the case with global expansions introduced in Section 3.5, because new molecules created in a global expansion may break the doubly connected property. In fact, a global expansion preserves the doubly connected structure only when we expand a T-variable containing a *redundant* blue solid edge. Here we call a blue solid edge redundant if and only if after removing it, the resulting graph is still doubly connected.

In [64], we will show that the graphs in our expansions actually satisfy a stronger *pre-deterministic property*. Roughly speaking, a doubly connected graph $\mathcal{G}$ is said to be pre-deterministic if the following property holds: there exists an order of all the internal blue solid edges in $\mathcal{G}$, denoted by $b_1 \preceq b_2 \preceq \cdots$, such that for any k, after changing the edges $b_1, \cdots, b_{k-1}$ into diffusive edges, the blue solid edge b_k becomes a redundant edge. We call this order of blue solid edges a *pre-deterministic order*.

Now the highlight of our expansion strategy is that if we expand the T-variable containing the *first redundant edge* in a pre-deterministic order using the global expansion in Section 3.5, then the resulting graphs are still pre-deterministic. Then in every new graph, we find the first redundant edge in a pre-deterministic order and expand it further. Continuing in this way, we finally obtain a linear combination of graphs that can be written into the form (2.28). Here we remark that after one step of global expansion, we need to apply local expansions to the resulting graphs to turn them into locally standard graphs before we apply the next step of global expansion. In [64], we will show that local expansions also do not affect the pre-deterministic property.

Finally, we remark that the above discussion is only for heuristic purpose, and they are not completely rigorous regarding some technical details. In fact, we will use a slightly weaker property, called the *sequentially pre-deterministic property*, instead of the pre-deterministic property. The interested reader can refer to [64] for more details.

A Symmetry and translational invariance

In this section, we record the following simple fact: any deterministic graph with two external atoms satisfies the properties in (1.25).

Lemma A.1. *Let $\mathcal{M}$ be a deterministic matrix in terms of S, $S^{\pm}(z)$ and $\Theta(z)$. Then we have*

$$\mathcal{M}(x, x+a) = \mathcal{M}(0, a), \quad \mathcal{M}(0, a) = \mathcal{M}(0, -a), \quad \forall x, a \in \mathbb{Z}_L^d. \tag{A.1}$$

Proof. This lemma is a simple consequence of the fact that all the matrices S, $S^{\pm}(z)$ and $\Theta(z)$ satisfy the two properties in (A.1). Suppose $\mathcal{M}(x, y)$ can be written into the general form

$$\mathcal{M}(x, y) = \sum_{x_1, \cdots, x_\ell} \prod_{(x, x_i) \in \mathbf{E}} f_i(x, x_i) \cdot \prod_{(x_i, x_j) \in \mathbf{E}} f_{ij}(x_i, x_j) \cdot \prod_{(x_j, y) \in \mathbf{E}} g_j(x_j, y),$$

where $\mathbf{E}$ is the set of all the edges in the graph $\mathcal{M}(x, y)$, and f_i, f_{ij} and g_j satisfy the two properties in (A.1). Then we have that

$$\begin{aligned}
\mathcal{M}(x, x+a) &= \sum_{x_1, \cdots, x_\ell} \prod_{(x, x_i) \in \mathbf{E}} f_i(x, x_i) \cdot \prod_{(x_i, x_j) \in \mathbf{E}} f_{ij}(x_i, x_j) \cdot \prod_{(x_j, x+a) \in \mathbf{E}} g_j(x_j, x+a) \\
&= \sum_{x_1, \cdots, x_\ell} \prod_{(0, x_i - x) \in \mathbf{E}} f_i(0, x_i - x) \cdot \prod_{(x_i - x, x_j - x) \in \mathbf{E}} f_{ij}(x_i - x, x_j - x) \cdot \prod_{(x_j - x, a) \in \mathbf{E}} g_j(x_j - x, a) \\
&= \sum_{x_1, \cdots, x_\ell} \prod_{(0, x_i) \in \mathbf{E}} f_i(0, x_i) \cdot \prod_{(x_i, x_j) \in \mathbf{E}} f_{ij}(x_i, x_j) \cdot \prod_{(x_j, a) \in \mathbf{E}} g_j(x_j, a) = \mathcal{M}(0, a),
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{M}(0, a) &= \sum_{x_1, \cdots, x_\ell} \prod_{(0, x_i) \in \mathbf{E}} f_i(0, x_i) \cdot \prod_{(x_i, x_j) \in \mathbf{E}} f_{ij}(x_i, x_j) \cdot \prod_{(x_j, a) \in \mathbf{E}} g_j(x_j, a) \\
&= \sum_{x_1, \cdots, x_\ell} \prod_{(0, -x_i) \in \mathbf{E}} f_i(0, -x_i) \cdot \prod_{(-x_i, -x_j) \in \mathbf{E}} f_{ij}(-x_i, -x_j) \cdot \prod_{(-x_j, -a) \in \mathbf{E}} g_j(-x_j, -a) \\
&= \sum_{x_1, \cdots, x_\ell} \prod_{(0, x_i) \in \mathbf{E}} f_i(0, x_i) \cdot \prod_{(x_i, x_j) \in \mathbf{E}} f_{ij}(x_i, x_j) \cdot \prod_{(x_j, -a) \in \mathbf{E}} g_j(x_j, -a) = \mathcal{M}(0, -a),
\end{aligned}$$

where we used that for f satisfying (A.1), $f(-x, -y) = f(0, x-y) = f(0, y-x) = f(x, y)$. $\square$

B Proofs of some deterministic estimates

In this section, we collect the proofs of some deterministic estimates, including Lemma 6.1, Lemma 7.2, Lemma 7.3 and Claim 7.8. We start with the following Taylor expansion:

$$\Theta = (1-|m|^2 S)^{-1}|m|^2 S = \left(1-|m|^{2K}S^K\right)^{-1}\sum_{k=1}^{K}|m|^{2k}S^k. \tag{B.1}$$

Since $\|S\|_{\ell^\infty\to\ell^\infty}=1$ and $|m|\leqslant 1-c\eta$ for some constant $c>0$, taking $K=\eta^{-1}W^\tau$ or $K=\eta^{-1}\langle x-y\rangle^\tau$ in (B.1) for a small constant $\tau>0$, we get that

$$\Theta_{xy} = \sum_{k=1}^{\eta^{-1}W^\tau}|m|^{2k}(S^k)_{xy} + \mathrm{O}(e^{-cW^\tau/2}) = \sum_{k=1}^{\eta^{-1}\langle x-y\rangle^\tau}|m|^{2k}(S^k)_{xy} + \mathrm{O}(e^{-c\langle x-y\rangle^\tau/2}). \tag{B.2}$$

Since S is a doubly stochastic matrix, $(S^k)_{xy}$ can be understood through a k-step random walk on the torus $\mathbb{Z}_L^d$. We first prove the following lemma for the random walk on $\mathbb{Z}^d$.

Lemma B.1. *Let $B_n=\sum_{i=1}^n X_i$ be a random walk on $\mathbb{Z}^d$ with i.i.d. steps X_i such that*

$$\mathbb{P}(|X_1|=x)=f_{W,L}(x),$$

for a function $f_{W,L}$ satisfying Assumption 1.2. Let $\mathcal{C}$ be the covariance matrix of X_1 with $\mathcal{C}_{ij}=\mathbb{E}[(X_1)_i(X_1)_j]$. Assume that $n\geqslant W^{c_0}$ for a constant $c_0>0$. Then for any large constant $D>0$, we have that

$$\mathbb{P}\left(B_n=x\right)=\frac{1+\mathrm{o}(1)}{(2\pi n)^{d/2}\sqrt{\det(\mathcal{C})}}e^{-\frac12 x^\top(n\mathcal{C})^{-1}x}+\mathrm{O}(n^{-D}). \tag{B.3}$$

Moreover, suppose $x,a,b\in\mathbb{Z}^d$ satisfy that $|x|\geqslant W^{1+2\varepsilon_0}$ and $|a|\leqslant|b|\leqslant|x|^{1-\varepsilon_0}$ for a small constant $\varepsilon_0>0$. Then if $n\geqslant|x|^{2-2\varepsilon_0+\varepsilon_1}/W^2$ for a constant $\varepsilon_1>0$, we have that

$$\begin{aligned}&|\mathbb{P}\left(B_n=x+a\right)+\mathbb{P}\left(B_n=x-a\right)-\mathbb{P}\left(B_n=x+b\right)-\mathbb{P}\left(B_n=x-b\right)|\\&\leqslant\frac{|b|^2}{W^2}\frac{n^\tau}{n^{d/2+1}W^d}e^{-\frac12 x^\top(n\mathcal{C})^{-1}x}+\mathrm{O}(|x|^{-D}),\end{aligned} \tag{B.4}$$

for any constants $\tau,D>0$.

Proof. The estimate (B.3) has been proved in Lemma 30 of [65]. We only need to prove (B.4). By (1.8), we have that for any fixed $\tau,D>0$,

$$\mathbb{P}(|X_1|\leqslant W^{1+\tau})\geqslant 1-W^{-D}.$$

Then using a simple Chernoff bound, we can get that for any fixed $\tau,D>0$,

$$\mathbb{P}\left(B_n=x\right)=\mathrm{O}(n^{-D}),\quad\text{for}\quad |x|\geqslant Wn^{1/2+\tau}. \tag{B.5}$$

Thus to prove (B.3) and (B.4), we only need to focus on the case

$$|x|=\mathrm{O}(Wn^{1/2+\tau_0}) \tag{B.6}$$

for a small constant $\tau_0>0$. In the following proof, we always make this assumption. Using characteristic functions and cumulants, the following estimate has been shown in the proof of Lemma 30 in [65]:

$$\begin{aligned}\mathbb{P}\left(B_n=x\right)=&\frac{1}{(2\pi)^d}\int_{|p|\leqslant W^{-1}n^{-1/2}|x|^{\tau_0}}\mathrm{d}p\,e^{-\mathrm{i}p\cdot x}e^{-\frac12 np^\top\mathcal{C}p}\left[1+\sum_{3\leqslant k\leqslant K_D}\alpha_k(\hat p)(Wn^{1/2}|p|)^k\right]\\&+\mathrm{O}(|x|^{-D}),\end{aligned} \tag{B.7}$$

where K_D is a fixed integer depending only on D, and $\alpha_k(\hat p)\in\mathbb{C}$ are complex coefficients defined as

$$\alpha_k(\hat p):=\frac{\kappa_k(\hat p)}{k!\cdot W^k}\mathrm{i}^k n^{1-k/2}.$$

Here $\hat{p} := p/|p|$ and $\kappa_k(\hat{p})$ denotes the k-th cumulant of $\hat{p}\cdot X_1$. Using (1.8), we can check that

$$|\kappa_k(\hat{p})| \leqslant C^k k! \cdot W^k, \quad \forall\, \hat{p} \in \mathbb{S}^d,$$

for a large enough constant $C>0$. Thus we have $\alpha_k(\hat{p}) = \mathrm{O}(n^{1-k/2})$. Using (B.7), we obtain that

$$\begin{aligned} &|\mathbb{P}(B_n = x+a) + \mathbb{P}(B_n = x-a) - \mathbb{P}(B_n = x+b) - \mathbb{P}(B_n = x-b)| \qquad \text{(B.8)} \\ &\lesssim \left| \int_{|p|\leqslant W^{-1}n^{-1/2}|x|^{\tau_0}} \mathrm{d}p\, [\cos(p\cdot a) - \cos(p\cdot b)]\, e^{-\mathrm{i}p\cdot x} e^{-\frac{1}{2}np^\top \mathcal{C} p}\left[1 + \sum_{3\leqslant k\leqslant K_D} \alpha_k(\hat{p})(Wn^{1/2}|p|)^k\right]\right| + \mathrm{O}(|x|^{-D}). \end{aligned}$$

For $n \geqslant |x|^{2-2\varepsilon_0+\varepsilon_1}/W^2$, $|a|\leqslant |b| \leqslant |x|^{1-\varepsilon_0}$ and $|p| \leqslant W^{-1}n^{-1/2}|x|^{\tau_0}$, we have

$$|p\cdot a| + |p\cdot b| \lesssim |x|^{-\varepsilon_1/2}|x|^{\tau_0} \leqslant |x|^{-\varepsilon_1/4},$$

as long as $\tau_0 < \varepsilon_1/4$. Thus using the Taylor expansions of $\cos(p\cdot a)$ and $\cos(p\cdot b)$, we can write that

$$\cos(p\cdot a) - \cos(p\cdot b) = \sum_{k=1}^{K_D'} (-1)^k \frac{(p\cdot a)^{2k} - (p\cdot b)^{2k}}{(2k)!} + \mathrm{O}(|x|^{-D}), \tag{B.9}$$

where K_D' is a fixed integer depending only on D and ε_1. Inserting it into (B.8) and bounding each term in the resulting expression, we can obtain (B.4). For example, the leading term is

$$\begin{aligned} &\int_{|p|\leqslant W^{-1}n^{-1/2}|x|^{\tau_0}} \mathrm{d}p \frac{(p\cdot a)^2 - (p\cdot b)^2}{2} e^{-\mathrm{i}p\cdot x} e^{-\frac{1}{2}np^\top \mathcal{C}p} \\ &= \frac{1}{\sqrt{n^d \det(\mathcal{C})}} \int_{|W\mathcal{C}^{-1/2}q|\leqslant |x|^{\tau_0}} \mathrm{d}q \frac{\left[q\cdot (n\mathcal{C})^{-1/2}a\right]^2 - \left[q\cdot(n\mathcal{C})^{-1/2}b\right]^2}{2} e^{-\mathrm{i}q\cdot y} e^{-q^2/2} + \mathrm{O}(|x|^{-D}), \qquad \text{(B.10)} \end{aligned}$$

where we used change of variables $y := (n\mathcal{C})^{-1/2}x$ and $q := (n\mathcal{C})^{1/2}p$. Using the conditions in Assumption 1.2, we can check that

$$C^{-1}W^2 \leqslant \lambda_{\min}(\mathcal{C}) \leqslant \lambda_{\max}(\mathcal{C}) \leqslant CW^2 \tag{B.11}$$

for some large constant $C>0$, where $\lambda_{\min}$ and $\lambda_{\max}$ respectively denote the maximum and minimum eigenvalues of $\mathcal{C}$. By (B.11), we have that $C^{-1/2}|q| \leqslant |W\mathcal{C}^{-1/2}q| \leqslant C^{1/2}|q|$. Then in (B.10) we can replace the domain of the integral by $\int_{q\in\mathbb{R}^d}$, because the integral over the domain $\{q : |W\mathcal{C}^{-1/2}q| > |x|^{\tau_0}\}$ can be bounded by $\mathrm{O}(|x|^{-D})$ for any fixed $D>0$ due to the term $e^{-q^2/2}$. Hence we can estimate (B.10) as

$$\begin{aligned} |\text{(B.10)}| &\lesssim \frac{1}{n^{d/2}W^d}\left| \int_{q\in\mathbb{R}^d} \mathrm{d}q \frac{[q\cdot(n\mathcal{C})^{-1/2}a]^2 - [q\cdot(n\mathcal{C})^{-1/2}b]^2}{2} e^{-\mathrm{i}q\cdot y}e^{-q^2/2}\right| + \mathrm{O}(|x|^{-D}) \\ &\lesssim \frac{n^{2\tau_0}}{n^{d/2+1}W^d}\frac{|b|^2}{W^2} e^{-y^2/2} + \mathrm{O}(|x|^{-D}) \end{aligned}$$

where we have bounded the integral using the stationary phase approximation and the fact that $|y| = \mathrm{O}(n^{\tau_0})$ for x satisfying (B.6). All the other integrals coming from the $k\geqslant 2$ terms in (B.9) can be bounded in a similar way, and they all give sub-leading terms. This proves the bound in (B.4). □

Now we prove Lemma 6.1, Lemma 7.2, Lemma 7.3 and Claim 7.8 one by one using Lemma B.1.

Proof of Lemma 6.1. Fix a small constant $\tau > 0$. We need to estimate the sum in (B.2). Let $B_n = \sum_{i=1}^n X_i$ be a random walk on $\mathbb{Z}_L^d$ with *i.i.d.* steps X_i, such that $\mathbb{P}(X_i = y - x) = s_{xy}$. Then we have

$$(S^k)_{xy} = \mathbb{P}(B_k = y-x). \tag{B.12}$$

If $W^\tau \leqslant k \leqslant L^{2-\tau}/W^2$, using a simple Chernoff bound we can get the large deviation estimate

$$\mathbb{P}\left(|B_k| \geqslant k^{1/2+\delta}W\right) \leqslant k^{-D} \tag{B.13}$$

for any constants $\delta, D > 0$. In particular, it shows that with high probability, B_k can be regarded as a random walk on the full lattice $\mathbb{Z}^d$ if $k \leqslant L^{2-\tau}/W^2$, so that both (B.3) and (B.4) can be applied.

Since $g(\cdot)$ is a symmetric function and $\sum_x g(x) = 0$, we can write that

$$\sum_x \Theta_{0x}(z) g(x - x_0) = \sum_{a \in \mathfrak{A}} g(a) \left[\Theta(0, x_0 + a) + \Theta(0, x_0 - a) - \Theta(0, x_0 + y_a) - \Theta(0, x_0 - y_a)\right], \tag{B.14}$$

where $\mathfrak{A}$ is a subset of $\mathcal{B}_K$, and $y_a \in \mathcal{B}_K$ depends on a and satisfies $|y_a| \leqslant |a|$. By (B.2) and (B.12), we have

$$|\Theta(0, x_0 + a) + \Theta(0, x_0 - a) - \Theta(0, x_0 + y_a) - \Theta(0, x_0 - y_a)| \leqslant \sum_{k=1}^{|x_0|^\tau \eta^{-1}} |m|^{2k} P_k(a, y_a) + |x_0|^{-D}, \tag{B.15}$$

for any constant $D > 0$, where

$$P_k(a, y_a) := \left|\mathbb{P}(B_k = x_0 + a) + \mathbb{P}(B_k = x_0 - a) - \mathbb{P}(B_k = x_0 + y_a) - \mathbb{P}(B_k = x_0 - y_a)\right|.$$

Suppose $\tau < \varepsilon/2$ so that $\eta \geqslant W^2/L^{2-2\tau}$. Using the large deviation estimate (B.13), we can bound that

$$P_k(a, y_a) = \mathrm{O}(W^{-D}) \quad \text{for} \quad 1 \leqslant k \leqslant |x_0|^{2-\tau}/W^2. \tag{B.16}$$

For $|x_0|^{2-\tau}/W^2 < k \leqslant |x_0|^\tau \eta^{-1}$, $P_k(a, y_a)$ can be bounded using (B.4) as

$$P_k(a, y_a) \leqslant \frac{|a|^2}{W^2} \frac{n^\tau}{k^{d/2+1} W^d} e^{-\frac{1}{2} x_0^\top (k\mathcal{C})^{-1} x_0}. \tag{B.17}$$

Plugging (B.16) and (B.17) into (B.15), we obtain that for any constant $D > 0$,

$$\begin{aligned} |(B.15)| &\leqslant \sum_{|x_0|^{2-\tau}/W^2 \leqslant k \leqslant |x_0|^\tau \eta^{-1}} \frac{|a|^2}{W^2} \frac{n^\tau}{k^{d/2+1} W^d} e^{-\frac{1}{2} x_0^\top (k\mathcal{C})^{-1} x_0} + |x_0|^{-D} \\ &\lesssim \frac{|a|^2}{W^2} \frac{n^\tau}{\langle x_0 \rangle^{d - d\tau/2}} \mathbf{1}_{|x_0| \leqslant \eta^{-1/2} W^{1+\tau}} + |x_0|^{-D}. \end{aligned}$$

Plugging it into (B.14), we conclude Lemma 6.1 since τ can be arbitrarily small. □

Proof of Lemma 7.2. Using (1.8), we get that for any fixed $D > 0$,

$$|(S_\infty)_{0x} - S_{0x}(L)| \leqslant (|x| + W)^{-D}, \quad \forall\, x \in \mathbb{Z}^d, \tag{B.18}$$

where $S(L)$ refers to the variance matrix defined on $\mathbb{Z}_L^d$, and we adopted the convention that $S_{0x}(L) = 0$ for $x \notin (-L/2, L/2]^d$. Now we prove the estimates (7.6)–(7.9) one by one. First, by the arguments in the proof of [12, Lemma 4.2], there exist constants $c_0, c_1 > 0$ such that

$$\left\| \left(\frac{m^2(E) S + c_0}{1 + c_0} \right)^2 \right\|_{\ell^\infty \to \ell^\infty} \leqslant 1 - c_1, \quad \left\| \left(\frac{m^2(E) S_\infty + c_0}{1 + c_0} \right)^2 \right\|_{\ell^\infty \to \ell^\infty} \leqslant 1 - c_1. \tag{B.19}$$

Then we use (B.19) to estimate the Taylor expansion

$$S_\infty^+ = \frac{m^2(E) S_\infty}{1 + c_0} \sum_{k=0}^\infty \left(\frac{m^2(E) S_\infty + c_0}{1 + c_0} \right)^k. \tag{B.20}$$

Using (B.19), we immediately obtain that $(S_\infty^+)_{0x}$ exists for any $x \in \mathbb{Z}^d$, $\max_x (S_\infty^+)_{0x} = \mathrm{O}(W^{-d})$ and

$$\sum_{k \geqslant |x|/W} \left[S_\infty \left(\frac{m^2(E) S_\infty + c_0}{1 + c_0} \right)^k \right]_{0x} = \mathrm{O}\left(W^{-d} (1 - c_1)^{|x|/(2W)} \right).$$

On the other hand, when $|x| \geqslant W^{1+\tau}$, we have that for any fixed $D > 0$,

$$\left| \sum_{k < |x|/W} \left[S_\infty \left(\frac{m^2(E) S_\infty + c_0}{1 + c_0} \right)^k \right]_{0x} \right| \leqslant |x|^{-D}.$$

Here we used a similar large deviation estimate as in (B.13) to derive this estimate. Combining the above two estimates, we conclude (7.6). Subtracting the Taylor expansion of $S^+(z)$ from the expansion (B.20), and using $|m(z)-m(E)| = \mathrm{O}(\eta)$ and (B.18), we can readily conclude (7.8).

It remains to study Θ_∞. Suppose we have shown that $(\Theta_\infty)_{0x}$ exists for any $x\in\mathbb{Z}^d$. Then the estimate (7.7) follows from (1.17). Now we prove (7.9). For any $\widetilde L \geqslant L > 2|x|$, we abbreviate $\widetilde z := E + \mathrm{i}W^2/\widetilde L^2$, $\widetilde m \equiv m(\widetilde z)$, $\widetilde S \equiv S(\widetilde L)$ and $\widetilde\Theta \equiv \Theta(\widetilde z, \widetilde L)$. Then using (B.2), we obtain that

$$\left|\widetilde\Theta_{0x} - \Theta_{0x}\right| \leqslant \sum_{k=1}^{\langle x\rangle^{2-\tau}/W^2} \left| |m|^{2k} S^k_{0x} - |\widetilde m|^{2k}\widetilde S^k_{0x}\right| + \sum_{k=\langle x\rangle^{2-\tau}/W^2}^{\langle x\rangle^{-\tau}\eta^{-1}} \left| |m|^{2k} S^k_{0x} - |\widetilde m|^{2k}\widetilde S^k_{0x}\right|$$
$$+ \sum_{k=\langle x\rangle^{-\tau}\eta^{-1}}^{\langle x\rangle^{\tau}\eta^{-1}} |m|^{2k} S^k_{0x} + \sum_{k=\langle x\rangle^{\tau}\eta^{-1}}^{\langle x\rangle^{-\tau}\widetilde L^2/W^2} |\widetilde m|^{2k}\widetilde S^k_{0x} + \sum_{k=\langle x\rangle^{-\tau}\widetilde L^2/W^2}^{\langle x\rangle^{\tau}\widetilde L^2/W^2} |\widetilde m|^{2k}\widetilde S^k_{0x} + \langle x\rangle^{-D}, \qquad \text{(B.21)}$$

for any constants $\tau, D>0$. Using (B.18), $\left||m|^{2k} - |\widetilde m|^{2k}\right| \lesssim k\eta$ and (B.3), we can bound the five terms on the right-hand side of (B.21) one by one as follows:

$$\sum_{k=1}^{\langle x\rangle^{2-\tau}/W^2} \left| |m|^{2k} S^k_{0x} - |\widetilde m|^{2k}\widetilde S^k_{0x}\right| \leqslant \langle x\rangle^{-D},$$

$$\sum_{k=\langle x\rangle^{2-\tau}/W^2}^{\langle x\rangle^{-\tau}\eta^{-1}} \left| |m|^{2k} S^k_{0x} - |\widetilde m|^{2k}\widetilde S^k_{0x}\right| \lesssim \sum_{k=\langle x\rangle^{2-\tau}/W^2}^{\langle x\rangle^{-\tau}\eta^{-1}} \frac{k\eta}{W^d k^{d/2}} \leqslant \frac{\eta}{W^4\langle x\rangle^{d-4-d\tau}},$$

$$\sum_{k=\langle x\rangle^{-\tau}\eta^{-1}}^{\langle x\rangle^{\tau}\eta^{-1}} |m|^{2k} S^k_{0x} \lesssim \frac{\langle x\rangle^{\tau}\eta^{-1}\mathbf{1}_{|x|\leqslant \langle x\rangle^{\tau}W\eta^{-1/2}}}{(\langle x\rangle^{-\tau}\eta^{-1})^{d/2}W^d} + \langle x\rangle^{-D} \leqslant \frac{\eta}{W^4\langle x\rangle^{d-4-2d\tau}} + \langle x\rangle^{-D},$$

$$\sum_{k=\langle x\rangle^{\tau}\eta^{-1}}^{\langle x\rangle^{-\tau}\widetilde L^2/W^2} |\widetilde m|^{2k}\widetilde S^k_{0x} \lesssim \sum_{k=\max(\langle x\rangle^{\tau}\eta^{-1},\langle x\rangle^{2-\tau}/W^2)}^{\langle x\rangle^{-\tau}\widetilde L^2/W^2} \frac{1}{W^d k^{d/2}} + \langle x\rangle^{-D} \leqslant \frac{\eta}{W^4\langle x\rangle^{d-4-d\tau}} + \langle x\rangle^{-D},$$

$$\sum_{k=\langle x\rangle^{-\tau}\widetilde L^2/W^2}^{\langle x\rangle^{\tau}\widetilde L^2/W^2} |\widetilde m|^{2k}\widetilde S^k_{0x} \lesssim \frac{\langle x\rangle^{\tau}\widetilde L^2}{W^2}\frac{1}{(\langle x\rangle^{-\tau}\widetilde L^2/W^2)^{d/2}W^d} + \langle x\rangle^{-D} \leqslant \frac{\langle x\rangle^{d\tau}}{W^2\widetilde L^{d-2}} + \langle x\rangle^{-D}.$$

Combining the above estimates and taking $\widetilde L\to\infty$, we can conclude (7.9) since τ is arbitrary.

Finally, it remains to show that the limit in (7.5) exists. For any $x\in\mathbb{Z}^d$, we choose $\widetilde L \geqslant L \geqslant L^c\langle x\rangle$ for a constant $c>0$, $\eta = W^2/L^2$, $\widetilde\eta = W^2/\widetilde L^2$, $m\equiv m(E+\mathrm{i}\eta)$ and $\widetilde m \equiv m(E+\mathrm{i}\widetilde\eta)$. Again using the Taylor expansion (B.2), we obtain that for any constants $\tau, D>0$,

$$\left|\widetilde\Theta_{0x}\left(E+\mathrm{i}\widetilde\eta,\widetilde L\right) - \Theta_{xy}\left(E+\mathrm{i}\eta, L\right)\right| \leqslant \sum_{k=1}^{L^{-\tau}\eta^{-1}} \left| |m|^{2k} S^k_{0x} - |\widetilde m|^{2k}\widetilde S^k_{0x}\right| + \sum_{k=L^{-\tau}\eta^{-1}}^{L^{\tau}\eta^{-1}} |m|^{2k} S^k_{0x}$$
$$+ \sum_{k=L^{\tau}\eta^{-1}}^{L^{-\tau}\widetilde\eta^{-1}} |\widetilde m|^{2k}\widetilde S^k_{0x} + \sum_{k=L^{-\tau}\widetilde\eta^{-1}}^{L^{\tau}\widetilde\eta^{-1}} |\widetilde m|^{2k}\widetilde S^k_{0x} + L^{-D}.$$

Applying similar arguments as above to each term on the right-hand side, we can obtain that

$$\left|\widetilde\Theta_{0x}\left(E+\mathrm{i}\widetilde\eta,\widetilde L\right) - \Theta_{0x}\left(E+\mathrm{i}\eta, L\right)\right| \leqslant \frac{\eta L^{2d\tau}}{W^4\langle x\rangle^{d-4}} + L^{-D}.$$

This shows that for any fixed $x\in\mathbb{Z}^d$, $\widetilde\Theta_{0x}(E+\mathrm{i}W^2/\widetilde L, \widetilde L)$ is a Cauchy sequence in $\widetilde L$. Hence the limit in (7.5) exists. □

Proof of Lemma 7.3. We choose L such that $L \leqslant |x_0|^2 \leqslant 2L$ and $\eta = W^2/L^{2-\tau}$ for a small constant $\tau > 0$. By (7.9), we have that for any constant $D > 0$,

$$|(\Theta_\infty)_{0x}(E) - \Theta_{0x}(E + \mathrm{i}\eta)| \leqslant \frac{\eta}{W^4 \langle x\rangle^{d-4-\tau}} + \langle x\rangle^{-D}.$$

With this estimate, we get that

$$\left|\sum_x \left[(\Theta_\infty)_{0x}(E) - \Theta_{0x}(E + \mathrm{i}\eta)\right] g(x - x_0)\right| \leqslant \sum_{x\in\mathcal{B}_K} \left[\frac{|x_0|^\tau}{|x_0|^d} + |x_0|^{-D}\right] |g(x)|.$$

On the other hand, by Lemma 6.1 we have that

$$\left|\sum_x \Theta_{0x}(E + \mathrm{i}\eta) g(x - x_0)\right| \leqslant \left(\sum_{x\in\mathcal{B}_K} \frac{x^2}{|x_0|^2}|g(x)|\right) \left[|x_0|^\tau B_{0x_0} + |x_0|^{-D}\right].$$

Combining the above two estimates, we conclude the proof. □

Proof of Claim 7.8. Fix any constants $\tau, D > 0$, by (1.8) we have that

$$\left|\widetilde{S}_{0x} - S_{0x}\right| \leqslant |x|^{-D} \quad \text{for} \quad |x| \geqslant W^{1+\tau}. \tag{B.22}$$

We now bound the difference $\widetilde{S}_{0x} - S_{0x}$ for $|x| < W^{1+\tau}$:

$$\widetilde{S}_{0x} - S_{0x} = \frac{\langle x\rangle^2}{L^2}\frac{1}{L^d} \sum_{p_0\in\mathbb{T}_L^d} \phi(p_0, x) e^{\mathrm{i}p_0\cdot x} + \mathrm{O}(W^{-D}), \tag{B.23}$$

where we used (1.10) and that

$$\int_{|p|>\pi} \psi(Wp) e^{\mathrm{i}p\cdot x} \mathrm{d}p = \mathrm{O}(W^{-D}),$$

because ψ is a Schwartz function. Moreover, the function $\phi(p_0, x)$ is defined as

$$\phi(p_0, x) := \frac{L^2}{\langle x\rangle^2}\frac{L^d}{(2\pi)^d} \int_{p\in B(p_0)} \left[\psi(Wp_0) - \psi(Wp)e^{\mathrm{i}(p-p_0)\cdot x}\right] \mathrm{d}p,$$

with $B(p_0)$ being the box centered at p_0 with side length $2\pi/L$. It is easy to check that

$$|\phi(p_0, x)| \lesssim \sup_{p\in B(p_0)} \left(|\psi(Wp)| + |\psi'(Wp)| + |\psi''(Wp)|\right).$$

Plugging this estimate into (B.23), we can get that

$$\left|\widetilde{S}_{0x} - S_{0x}\right| \lesssim \frac{\langle x\rangle^2}{L^2}\frac{1}{W^d} + \langle x\rangle^{-D}, \quad \text{for} \quad |x| \leqslant W^{1+\tau}. \tag{B.24}$$

Combining (B.22) and (B.24), we obtain (7.23). Then using (B.19), (7.23) and the Taylor expansions of S^+ and $\widetilde{S}^+$ as in (B.20), we can readily get (7.24). We omit the details.

It remains to prove (7.25). By (7.23), we have that for any constants $\tau, D > 0$,

$$\left|\widetilde{S}^k_{0x} - S^k_{0x}\right| \lesssim \frac{W^2}{L^2}\frac{1}{W^d} \sum_{k_1=1}^{k-1} \sum_{|x_1-x_2|\leqslant\langle x\rangle^\tau W} S^{k_1}_{0x_1}\widetilde{S}^{k-1-k_1}_{x_2x} + \langle x\rangle^{-D}. \tag{B.25}$$

Moreover, we have a similar inequality as (B.21):

$$\begin{aligned}|\widetilde{\Theta}_{0x} - \Theta_{0x}| \leqslant &\sum_{k=1}^{\langle x\rangle^{2-\tau}/W^2} |m|^{2k}\left|S^k_{0x} - \widetilde{S}^k_{0x}\right| + \sum_{k=\langle x\rangle^{2-\tau}/W^2}^{\langle x\rangle^{-\tau}\eta^{-1}} |m|^{2k}\left|S^k_{0x} - \widetilde{S}^k_{0x}\right| \\ &+ \sum_{k=\langle x\rangle^{-\tau}\eta^{-1}}^{\langle x\rangle^{\tau}\eta^{-1}} |m|^{2k}\left(\left|\widetilde{S}^k_{0x}\right| + \left|S^k_{0x}\right|\right) + \langle x\rangle^{-D}.\end{aligned} \tag{B.26}$$

Using (B.3) and (B.25), we can bound the three terms on the right-hand side of (B.26) one by one as in the estimates below (B.21), which concludes (7.25). We omit the details. □

C Proofs for local expansions

In this section, we provide the proofs for the lemmas in Section 3.

Proof of Lemma 3.9. We first prove that $\mathcal{O}^{(x)}_{weight}[\mathcal{G}]$ is a canonical local expansion by verifying the properties (i)–(iv) of Definition 3.8 one by one. To prove property (i), it suffices to show that (3.5) is an identity in the sense of graph values. This follows from Lemma 3.5 together with the facts that $b_{x\alpha} = \delta_{x\alpha} + S^{+}_{x\alpha}$ and $P_\alpha = 1 - Q_\alpha$. The properties (iii) and (iv) are trivial by definition. It remains to prove property (ii). First, it is easy to see that $\mathcal{O}^{(x),1}_{weight}$ acting on a regular (resp. normal regular) graph gives a linear combination of regular (resp. norma regular) graphs. Second, the $\mathcal{O}_{dot}$ in (3.7) will expand a regular graph into a sum of normal regular graphs by Lemma 3.3. Hence to prove property (ii) of Definition 3.8, it suffices to show that $\mathcal{O}^{(x),2}_{weight}$ acting on a regular graph also gives a linear combination of regular graphs. Now given any regular graph $\mathcal{G}$, we need to check the properties (i)–(iii) of Definition 2.11 for the graphs in $\mathcal{O}^{(x),2}_{weight}[\mathcal{G}]$. The properties (i) and (iii) of Definition 2.11 are trivially true, while the property (ii) follows from the fact that the new atoms are connected to x through paths of waved edges.

In sum, we have shown that $\mathcal{O}^{(x)}_{weight}[\mathcal{G}]$ is a canonical local expansion. Now we prove statements (a) and (b) of Lemma 3.9. If $\mathcal{G}$ contains some regular weights G_{xx} or $\overline{G}_{xx}$ on x, then there is a graph in $\mathcal{O}^{(x),1}_{weight}[\mathcal{G}]$ obtained by replacing all these weights by m or $\overline{m}$, and this graph satisfies (b). All the other graphs in $\mathcal{O}^{(x),1}_{weight}[\mathcal{G}]$ satisfy (a). To conclude the proof, it remains to prove that if $\mathcal{G}$ only contains light weights on the atom x, then every graph without Q-labels in $\mathcal{O}^{(x)}_{weight}[\mathcal{G}]$ satisfies either (a) or (b). For this purpose, we study the graphs on the right-hand side of (3.5) one by one.

(1) The first two graphs on the right-hand side of (3.5) both have strictly higher scaling orders than $\mathcal{G}$ because they contain one more light weight than $\mathcal{G}$.

(2) We consider any graph, say $\mathcal{G}_1$, in $\mathcal{O}^{(x),1}_{weight} \circ \mathcal{O}_{dot}\,[m \sum_\alpha s_{x\alpha} G_{\alpha x} \partial_{h_{\alpha x}} f(G)]$. We have the following cases.

   – If α is identified with x or some other atoms in $f(G)$, then we have that

$$\mathrm{ord}\,[\mathcal{G}_1] \geqslant \mathrm{ord}\,[\mathcal{G}] + 1, \tag{C.1}$$

   because the scaling order of $s_{x\alpha}$ is larger than $\mathrm{ord}(G_{xx} - m) = 1$ and the scaling orders of the graphs in $\partial_{h_{\alpha x}} f(G)$ are $\geqslant \mathrm{ord}[f(G)]$. Hence $\mathcal{G}_1$ satisfies (a).

   – If α is not identified with any other atom, then $G_{\alpha x}$ is of scaling order 1. Moreover, suppose $\mathcal{G}_1$ contains a subgraph in $\partial_{h_{\alpha x}} f(G)$ obtained through the partial derivatives in (3.8). If $b \neq x$, then the scaling order of $G_{a\alpha} G_{xb}$ or $\overline{G}_{bx} \overline{G}_{\alpha a}$ is strictly larger than the scaling order of the original component G_{ab}, $\overline{G}_{ba}$, $G_{aa} - m$ or $\overline{G}_{aa} - \overline{m}$ (where the last two cases happen if the partial derivative acts on a light weight with $a = b$). Hence $\mathcal{G}_1$ satisfies (a).

   – Suppose α is not identified with any other atom, and $\mathcal{G}_1$ contains a subgraph in $\partial_{h_{\alpha x}} f(G)$ obtained through the partial derivatives in (3.8). If $b = x$, then $\mathcal{O}^{(x),1}_{weight}$ will expand (3.8) into

$$G_{a\alpha} G_{xx} = m G_{a\alpha} + G_{a\alpha}(G_{xx} - m), \quad \text{or} \quad \overline{G}_{xx} \overline{G}_{\alpha a} = \overline{m} \overline{G}_{\alpha a} + (\overline{G}_{xx} - \overline{m}) \overline{G}_{\alpha a}.$$

   If $\mathcal{G}_1$ contains $G_{a\alpha}(G_{xx} - m)$ or $(\overline{G}_{xx} - \overline{m})\overline{G}_{\alpha a}$, then it satisfies (a); otherwise, if $\mathcal{G}_1$ contains $m G_{a\alpha}$ or $\overline{m}\overline{G}_{\alpha a}$, then it satisfies (b).

(3) The graphs in $\mathcal{O}^{(x),1}_{weight} \circ \mathcal{O}_{dot} \left[m \sum_{\alpha,\beta} S^{+}_{x\alpha} s_{\alpha\beta} G_{\beta\alpha} \partial_{h_{\beta\alpha}} f(G)\right]$ can be dealt with in the same way as (2).

Combining the above cases (1)-(3), we conclude the proof of Lemma 3.9. □

Proof of Lemma 3.13. First, using Lemma 3.10 and a similar argument as in the above proof of Lemma 3.9, we can prove that $\mathcal{O}^{(x)}_{multi-e}[\mathcal{G}]$ is a canonical local expansion. To prove the property (a), we study the graphs on the right-hand side of (3.13) one by one.

(1) Notice that $\mathcal{O}_{dot} \left[\sum_\alpha s_{x\alpha} G_{\alpha y_1} \overline{G}_{\alpha y'_i}\right]$ is a sum of subgraphs of scaling orders $\geqslant \mathrm{ord}(G_{xy_1} \overline{G}_{xy'_i}) = 2$. In addition, there is an extra light weight $\overline{G}_{xx} - \overline{m}$ in the third graph on the right-hand side of (3.13), so it gives graphs of strictly higher scaling orders than $\mathcal{G}$ after the dotted edge partition $\mathcal{O}_{dot}$. The fourth graph on the right-hand side of (3.13) can be handled in the same way.

(2) The fifth graph on the right-hand side of (3.13) obviously has strictly higher scaling order than $\mathcal{G}$, because it contains one more light weight $G_{\alpha\alpha} - m$.

(3) Notice that $\mathcal{O}_{dot}\left[\sum_\alpha s_{x\alpha} G_{x\alpha} G_{\alpha y_1}\right]$ gives subgraphs of scaling orders $\geqslant 2 > \mathrm{ord}(G_{xy_1}) = 1$. Hence the dotted edge partition of the sixth graph on the right-hand side of (3.13) gives graphs of strictly higher scaling orders than $\mathcal{G}$. The seventh graph on the right-hand side of (3.13) can be handled in the same way.

(4) Regarding the eighth graph on the right-hand side of (3.13), we consider any graph, say $\mathcal{G}_1$, in

$$\mathcal{O}_{dot}\left[\sum_\alpha s_{x\alpha} \frac{\mathcal{G}}{G_{xy_1} f(G)} G_{\alpha y_1} \partial_{h_{\alpha x}} f(G)\right].$$

We have the following two cases.

- If α is identified with x or some other atoms in $f(G)$, then we have (C.1), because the scaling order of $s_{x\alpha}$ is larger than G_{xy_1} and the scaling orders of the graphs in $\partial_{h_{\alpha x}} f(G)$ are $\geqslant \mathrm{ord}[f(G)]$.
- If α is not identified with any other atom, then $G_{\alpha y_1}$ is of scaling order 1. Moreover, suppose $\mathcal{G}_1$ contains a graph in $\partial_{h_{\alpha x}} f(G)$ obtained through the partial derivatives in (3.8). By our assumption on $f(G)$, we must have $a \neq x$ and $b \neq x$. Then the scaling order of $G_{a\alpha} G_{xb}$ or $\overline{G}_{bx}\overline{G}_{\alpha a}$ is strictly larger than the scaling order of the original component G_{ab}, $\overline{G}_{ba}$, $G_{aa} - m$ or $\overline{G}_{aa} - \overline{m}$. Hence $\mathcal{G}_1$ satisfies (C.1).

(5) Regarding the first graph on the right-hand side of (3.13). we consider any graph, say $\mathcal{G}_1$, in

$$\mathcal{O}_{dot}\left[\sum_\alpha s_{x\alpha} G_{\alpha y_1} \overline{G}_{\alpha y_i'} \cdot \frac{\mathcal{G}_x}{G_{xy_1}\overline{G}_{xy_i'}}\right].$$

We have the following three cases.

- If α is identified with y_1 or y_i' and $y_1 \neq y_i'$, then the subgraph $s_{xy_1} G_{y_1 y_1} \overline{G}_{y_1 y_i'}$ or $s_{xy_i'} G_{y_i' y_1} \overline{G}_{y_i' y_i'}$ is of strictly higher scaling order than $G_{xy_1}\overline{G}_{xy_i'}$. Hence $\mathcal{G}_1$ satisfies (C.1).
- If α is not identified with either y_1 or y_i', then $\mathcal{G}_1$ satisfies (a.1).
- if $\alpha = y_1 = y_i'$, then $\mathcal{G}_1$ satisfies (a.2).

The second graph on the right-hand side of (3.13) can be handled in the same way.

Combining the cases (1)–(5), we conclude property (a).

Finally, if $\deg(x) = 1$ or x is connected with exactly two mismatched solid edges in $\mathcal{G}$, then one can check that the first two leading terms on the right-hand side of (3.13) vanish, and the above case (5) cannot happen. Hence we get that property (b) holds. □

Proof of Lemma 3.16. First, using Lemma 3.14 and a similar argument as in the proof of Lemma 3.9, we can prove that $\mathcal{O}_{GG}^{(x)}[\mathcal{G}]$ is a canonical local expansion. To prove the properties (a) and (b), we consider the graphs on the right-hand side of (3.17) one by one.

(1) The second to fifth graphs on the right-hand side of (3.17) obviously have strictly higher scaling orders than $\mathcal{G}$, because they contain one more light weight than $\mathcal{G}$.

(2) With a similar argument as in item (4) of the above proof of Lemma 3.13, we can show that the sixth and seventh graphs on the right-hand side of (3.17) will give graphs satisfying (C.1) after the dotted edge partition $\mathcal{O}_{dot}$.

(3) The first graph on the right-hand side of (3.17) satisfies (C.1) if $y \neq y'$. Otherwise, if $y = y'$, then the graph is obtained by replacing $G_{xy}G_{yx}$ with $mS^+_{xy}G_{yy}$.

Combining the cases (1)–(3), we conclude the properties (a) and (b). □

Proof of Lemma 3.19. All the statements are corollaries of Lemma 3.13, except for the properties (a) and (b) of $\mathcal{G}_1$. We prove these properties by studying the graphs on the right-hand side of (3.19) one by one.

(1) The second and third graphs on the right-hand side of (3.19) obviously satisfy (C.1) because they contain one more light weight than $\mathcal{G}$.

(2) With a similar argument as in item (4) of the above proof of Lemma 3.13, we can show that the fourth graph on the right-hand side of (3.19) will give graphs satisfying (C.1) after the dotted edge partition.

(3) Any graph in $\mathcal{O}_{dot}[\sum_\alpha s_{x\alpha} G_{\alpha y}\overline{G}_{\alpha y'} f(G)]$ satisfies (C.1) if $\alpha \in \{y, y'\}$ and $y \neq y'$, satisfies (a) if $\alpha \notin \{y, y'\}$, and satisfies (b) if $\alpha = y = y'$.

Combining the cases (1)-(3), we conclude the proof. □

Finally, we give the proof of Lemma 3.22.

Proof of Lemma 3.22. To conclude the proof, we need to show that the expansion process in Figure 1 will finally stop after O(1) many iterations.

The $\mathcal{O}_{dot}$–$\mathcal{O}_{weight}$ loop. First, we prove that the $\mathcal{O}_{dot}$–$\mathcal{O}_{weight}$ loop stops after O(1) many iterations. By Lemma 3.9, after a weight expansion, every resulting graph satisfies at least one of the following conditions: (1) it already satisfies the stopping rules; (2) it has strictly higher scaling order than the input graph; (3) it has strictly fewer weights than the input graph. Thus there exists a fixed $k \in \mathbb{N}$ depending on n and the number of weights such that after k iterations of the $\mathcal{O}_{dot}$–$\mathcal{O}_{weight}$ loop, every new graph either satisfies the stopping rules already or has no weights in it. A graph in the former case will be sent to the output directly. For a graph in the latter case, $\mathcal{O}_{weight}$ will be a null operation in the next iteration and this graph exits the $\mathcal{O}_{dot}$–$\mathcal{O}_{weight}$ loop successfully.

The $\mathcal{O}_{dot}$–$\mathcal{O}_{weight}$–$\mathcal{O}_{multi-e}$ loop. Suppose we apply the $\mathcal{O}_{dot}$-$\mathcal{O}_{weight}$-$\mathcal{O}_{multi-e}$ iteration once to an input graph, say $\mathcal{G}_0$, and get a collection of new graphs, say $\mathscr{G}_1$. For any new graph in $\mathscr{G}_1$, if it already satisfies the stopping rules, then we send it to the output directly; otherwise we send it back to the first step $\mathcal{O}_{dot}$. For a graph in the latter case, if it contains no weights and every atoms in it either has degree 0 or is connected with two matched solid edges, then $\mathcal{O}_{weight}$ and $\mathcal{O}_{multi-e}$ are both null operations and this graph exits the $\mathcal{O}_{dot}$–$\mathcal{O}_{weight}$–$\mathcal{O}_{multi-e}$ loop successfully. On the other hand, if either $\mathcal{O}_{weight}$ or $\mathcal{O}_{multi-e}$ is a non-trivial operation for a graph $\mathcal{G}_1 \in \mathscr{G}_1$, then we will apply the $\mathcal{O}_{dot}$–$\mathcal{O}_{weight}$–$\mathcal{O}_{multi-e}$ iteration to it and get a collection of new graphs, say $\mathscr{G}_2$. By Lemmas 3.9 and 3.13, every graph in $\mathscr{G}_2$ either satisfies the stopping rules already, or falls into at least one of the following categories:

(1) it has strictly higher scaling order than $\mathcal{G}_1$;

(2) there is one new atom of degree 2 and one old atom whose degree decreases by 2, while the degree of any other atom stays the same;

(3) there is no new atom and one old atom whose degree decreases by 2, while the degree of any other atom either stays the same or decreases by 2.

If a graph in $\mathscr{G}_2$ satisfies the stopping rules, then we send it to the output. Otherwise, we send it back to $\mathcal{O}_{dot}$ and apply another $\mathcal{O}_{dot}$–$\mathcal{O}_{weight}$–$\mathcal{O}_{multi-e}$ iteration to it.

We repeat the above iterations, and construct correspondingly a tree diagram $\mathcal{T}$ of graphs as follows. Let $\mathcal{G}_0$ be the root, which represents the input graph. Given a graph $\mathcal{G}$ represented by a vertex of the tree, its children are the graphs obtained from an $\mathcal{O}_{dot}$–$\mathcal{O}_{weight}$–$\mathcal{O}_{multi-e}$ iteration acting on $\mathcal{G}$. If a graph $\mathcal{G}$ satisfies the stopping rules or if $\mathcal{O}_{weight}$ and $\mathcal{O}_{multi-e}$ are null operations for $\mathcal{G}$, then $\mathcal{G}$ is a leaf of the tree, and it exits the $\mathcal{O}_{dot}$–$\mathcal{O}_{weight}$–$\mathcal{O}_{multi-e}$ successfully. Let the height of $\mathcal{T}$ be the maximum distance between a leaf of $\mathcal{T}$ and the root. To show that the $\mathcal{O}_{dot}$–$\mathcal{O}_{weight}$–$\mathcal{O}_{multi-e}$ loop stops after O(1) many iterations, it is equivalent to show that $\mathcal{T}$ is a finite tree with height of order O(1).

Let $\mathcal{G}_0 \to \mathcal{G}_1 \to \mathcal{G}_2 \to \cdots \to \mathcal{G}_h$ be a self-avoiding path on $\mathcal{T}$ from the root to a leaf. We let $k_0 = 0$. After having defined k_i, let $k_{i+1} := \min\{j > k_i : \mathrm{ord}(\mathcal{G}_j) > \mathrm{ord}(\mathcal{G}_{k_i})\}$. Then the sequence $\{k_0, k_1, k_2, \cdots\}$ has length $\leqslant n$, since a graph of scaling order $\geqslant n+1$ already satisfies the stopping rule (S3). Moreover, we claim that $|k_{i+1} - k_i| < n$. In fact, from the above discussion, we see that in order for the scaling order of a child to be the same as the scaling order of its parent, it has to be in category (2) or (3). Note that the total degree of the atoms in $\mathcal{G}_{k_i}$ is at most $2n$ (because there are at most n off-diagonal solid edges in it), and each iteration decreases at least one atom's degree by 2. Hence we immediately get that $|k_{i+1} - k_i| \leqslant n$.

The above argument shows that we must have $h \leqslant n^2$, i.e. the height of $\mathcal{T}$ is at most n^2. This means that the $\mathcal{O}_{dot}$–$\mathcal{O}_{weight}$–$\mathcal{O}_{multi-e}$ loop will stop after at most n^2 many iterations.

Next with Lemmas 3.16 and 3.19, using a similar tree diagram argument as above, we can show that both the $\mathcal{O}_{dot}$–$\cdots$–$\mathcal{O}_{GG}$ and $\mathcal{O}_{dot}$–$\cdots$–$\mathcal{O}_{G\overline{G}}$ loops will stop after O(1) many iterations. In particular, exiting the $\mathcal{O}_{dot}$–$\cdots$–$\mathcal{O}_{G\overline{G}}$ loop means that the expansion process in Figure 1 is completed successfully, which concludes the proof of Lemma 3.22. □

References

[1] E. Abrahams, P. W. Anderson, D. C. Licciardello, and T. V. Ramakrishnan, *Scaling theory of localization: Absence of quantum diffusion in two dimensions*, Phys. Rev. Lett. **42** (1979), 673–676.

[2] M. Aizenman and S. Molchanov, *Localization at large disorder and at extreme energies: an elementary derivation*, Communications in Mathematical Physics **157** (1993), no. 2, 245–278.

[3] Michael Aizenman and Simone Warzel, *Extended states in a Lifshitz tail regime for random Schrödinger operators on trees*, Phys. Rev. Lett. **106** (2011), 136804.

[4] ______, *Resonant delocalization for random Schrödinger operators on tree graphs*, J. Eur. Math. Soc. **15** (2013), no. 4, 1167–1222.

[5] P. W. Anderson, *Absence of diffusion in certain random lattices*, Phys. Rev. **109** (1958), 1492–1505.

[6] Z. Bao and L. Erdős, *Delocalization for a class of random block band matrices*, Probab. Theory Related Fields **167** (2017), no. 3, 673–776.

[7] R. Bauerschmidt, J. Huang, and H.-T. Yau, *Local Kesten–Mckay law for random regular graphs*, Communications in Mathematical Physics **369** (2019), no. 2, 523–636.

[8] R. Bauerschmidt, A. Knowles, and H.-T. Yau, *Local semicircle law for random regular graphs*, Communications on Pure and Applied Mathematics **70** (2017), no. 10, 1898–1960.

[9] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Isotropic local laws for sample covariance and generalized Wigner matrices*, Electron. J. Probab. **19** (2014), no. 33, 1–53.

[10] P. Bourgade, *Random band matrices*, Proceedings of the International Congress of Mathematicians, 2018, pp. 2759–2783.

[11] P. Bourgade, L. Erdős, H.-T. Yau, and J. Yin, *Universality for a class of random band matrices*, Advances in Theoretical and Mathematical Physics **21** (2017), no. 3, 739–800.

[12] P. Bourgade, F. Yang, H.-T. Yau, and J. Yin, *Random band matrices in the delocalized phase, II: Generalized resolvent estimates*, Journal of Statistical Physics **174** (2019), no. 6, 1189–1221.

[13] P. Bourgade, H.-T. Yau, and J. Yin, *Random band matrices in the delocalized phase, I: Quantum unique ergodicity and universality*, Communications on Pure and Applied Mathematics **73** (2020), no. 7, 1526–1596.

[14] J. Bourgain and C. Kenig, *On localization in the continuous Anderson-Bernoulli model in higher dimension*, Inventiones mathematicae **161** (2005), no. 2, 389–426.

[15] R. Carmona, A. Klein, and F. Martinelli, *Anderson localization for Bernoulli and other singular potentials*, Communications in Mathematical Physics **108** (1987), no. 1, 41–66.

[16] R. Carmona and J. Lacroix, *Spectral theory of random Schrödinger operators*, Birkhäuser, Boston, 1990.

[17] G. Casati, I. Guarneri, F. Izrailev, and R. Scharf, *Scaling behavior of localization in quantum chaos*, Phys. Rev. Lett. **64** (1990), 5–8.

[18] G. Casati, L. Molinari, and F. Izrailev, *Scaling properties of band random matrices*, Phys. Rev. Lett. **64** (1990), 1851–1854.

[19] D. Damanik, R. Sims, and G. Stolz, *Localization for one-dimensional, continuum, Bernoulli-Anderson models*, Duke Mathematical Journal **114** (2002), no. 1, 59–100.

[20] J. Ding and C. Smart, *Localization near the edge for the Anderson Bernoulli model on the two dimensional lattice*, Inventiones mathematicae **219** (2020), no. 2, 467–506.

[21] M. Disertori, L. Pinson, and T. Spencer, *Density of states for random band matrices*, Comm. Math. Phys. **232** (2002), 83–124.

[22] J. T. Edwards and D. J. Thouless, *Numerical studies of localization in disordered systems*, Journal of Physics C: Solid State Physics **5** (1972), no. 8, 807–820.

[23] K. Efetov, *Supersymmetry in disorder and chaos*, Cambridge University Press (1997).

[24] L. Erdős and A. Knowles, *Quantum diffusion and delocalization for band matrices with general distribution*, Ann. Henri Poincaré **12** (2011), no. 7, 1227–1319.

[25] ______, *Quantum diffusion and eigenfunction delocalization in a random band matrix model*, Communications in Mathematical Physics **303** (2011), no. 2, 509–554.

[26] L. Erdős, A. Knowles, and H.-T. Yau, *Averaging fluctuations in resolvents of random band matrices*, Ann. Henri Poincaré **14** (2013), 1837–1926.

[27] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Delocalization and diffusion profile for random band matrices*, Comm. Math. Phys. **323** (2013), no. 1, 367–416.

[28] ———, *The local semicircle law for a general class of random matrices*, Elect. J. Prob. **18** (2013), no. 59, 1–58.

[29] ———, *Spectral statistics of Erdős–Rényi graphs I: Local semicircle law*, The Annals of Probability **41** (2013), no. 3B, 2279 –2375.

[30] L. Erdős, M. Salmhofer, and H.-T. Yau, *Quantum diffusion of the random Schrödinger evolution in the scaling limit*, Acta Mathematica **200** (2008), no. 2, 211 –277.

[31] L. Erdős, B. Schlein, and H.-T. Yau, *Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices*, Annals of Probability **37** (2009), 815–852.

[32] ———, *Local semicircle law and complete delocalization for Wigner random matrices*, Commun. Math. Phys. **287** (2008), no. 2, 641–655.

[33] L. Erdős, H.-T. Yau, and J. Yin, *Bulk universality for generalized Wigner matrices*, Probab. Theory Related Fields **154** (2012), no. 1-2, 341–407.

[34] ———, *Rigidity of eigenvalues of generalized Wigner matrices*, Adv. Math. **229** (2012), no. 3, 1435–1515.

[35] ———, *Universality for generalized Wigner matrices with Bernoulli distribution*, Journal of Combinatorics **2** (2011), no. 1, 15–81.

[36] M. Feingold, D. M. Leitner, and M. Wilkinson, *Spectral statistics in semiclassical random-matrix ensembles*, Phys. Rev. Lett. **66** (1991Feb), 986–989.

[37] J. Fröhlich, F. Martinelli, E. Scoppola, and T. Spencer, *Constructive proof of localization in the Anderson tight binding model*, Communications in Mathematical Physics **101** (1985), no. 1, 21–46.

[38] J. Fröhlich and T. Spencer, *Absence of diffusion in the Anderson tight binding model for large disorder or low energy*, Communications in Mathematical Physics **88** (1983), no. 2, 151–184.

[39] Y. V. Fyodorov and A. D. Mirlin, *Scaling properties of localization in random band matrices: A σ-model approach.*, Phys. Rev. Lett. **67** (1991), 2405–2409.

[40] F. Germinet and A. Klein, *A comprehensive proof of localization for continuous Anderson models with singular random potentials*, J. Eur. Math. Soc. **15** (2013), no. 1, 53–143.

[41] Y. He, A. Knowles, and M. Marcozzi, *Local law and complete eigenvector delocalization for supercritical Erdős–Rényi graphs*, The Annals of Probability **47** (2019), no. 5, 3278 –3302.

[42] Y. He and M. Marcozzi, *Diffusion profile for random band matrices: A short proof*, Journal of Statistical Physics **177** (2019), no. 4, 666–716.

[43] J. Huang and H.-T. Yau, *Spectrum of random d-regular graphs up to the edge*, arXiv:2102.00963.

[44] A. Khorunzhy, B. Khoruzhenko, and L. Pastur, *Asymptotic properties of large random matrices with independent entries*, Journal of Mathematical Physics **37** (1996), no. 10, 5033–5060.

[45] W. Kirsch, *An invitation to random Schroedinger operators*, arXiv:0709.3707 (2007).

[46] L. Li and L. Zhang, *Anderson-Bernoulli localization on the 3D lattice and discrete unique continuation principle*, arXiv:1906.04350 (2019).

[47] A. Lytova and L. Pastur, *Central limit theorem for linear eigenvalue statistics of random matrices with independent entries*, Ann. Probab. **37** (2009), no. 5, 1778–1840.

[48] R. Peled, J. Schenker, M. Shamis, and S. Sodin, *On the Wegner Orbital Model*, International Mathematics Research Notices **2019** (2017), no. 4, 1030–1058.

[49] J. Schenker, *Eigenvector localization for random band matrices with power law band width*, Comm. Math. Phys. **290** (2009), 1065–1097.

[50] M. Shcherbina and T. Shcherbina, *Characteristic polynomials for 1d random band matrices from the localization side*, Communications in Mathematical Physics **351** (2017), no. 3, 1009–1044.

[51] ———, *Universality for 1d random band matrices: Sigma-model approximation*, Journal of Statistical Physics **172** (2018), no. 2, 627–664.

[52] ———, *Universality for 1d random band matrices*, arXiv:1910.02999 (2019).

[53] T. Shcherbina, *On the second mixed moment of the characteristic polynomials of 1d band matrices*, Comm. Math. Phys. **328** (2014), 45–82.

[54] ———, *Universality of the local regime for the block band matrices with a finite number of blocks*, J. Stat. Phys. **155** (2014), 466–499.

[55] ———, *Universality of the second mixed moment of the characteristic polynomials of the 1d band matrices: Real symmetric case*, J. Math. Phys. **56** (2015).

[56] S. Sodin, *The spectral edge of some random band matrices*, Ann. of Math. **173** (2010), no. 3, 2223–2251.

[57] T. Spencer, *Duality, statistical mechanics and random matrices*, Current Developments in Mathematics **2012**, 229–260.

[58] ———, *Mathematical aspects of Anderson localization*, International Journal of Modern Physics B **24** (2010), 1621–1639.

[59] ———, *Random banded and sparse matrices*, Oxford handbook of random matrix theory, 2011.

[60] ———, *SUSY statistical mechanics and random band matrices*, Quantum many body systems, 2012.

[61] D. J. Thouless, *Maximum metallic resistance in thin wires*, Phys. Rev. Lett. **39** (1977), 1167–1169.

[62] E. P. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, Annals of Mathematics **62** (1955), no. 3, 548–564.

[63] M. Wilkinson, M. Feingold, and D. M. Leitner, *Localization and spectral statistics in a banded random matrix ensemble*, Journal of Physics A: Mathematical and General **24** (1991), no. 1, 175.

[64] F. Yang, H.-T. Yau, and J. Yin, *Delocalization and quantum diffusion of random band matrices in high dimensions II: T-expansion*, arXiv:2107.05795 (2021).

[65] F. Yang and J. Yin, *Random band matrices in the delocalized phase, III: averaging fluctuations*, Probability Theory and Related Fields **179** (2021), 451–540.