Random band matrices in the delocalized phase, II: Generalized resolvent estimates

P. Bourgade
Courant Institute
bourgade@cims.nyu.edu

F. Yang
U. of California, Los Angeles
fyang75@math.ucla.edu

H.-T. Yau
Harvard University
htyau@math.harvard.edu

J. Yin
U. of California, Los Angeles
jyin@math.ucla.edu

This is the second part of a three part series of papers. In this paper, we consider a general class of $N \times N$ random band matrices $H = (H_{ij})$ whose entries are centered random variables, independent up to a symmetry constraint. We assume that the variances $\mathbb{E}|H_{ij}|^2$ form a band matrix with typical band width $1 \ll W \ll N$. We consider the generalized resolvent of H defined as $G(Z) := (H - Z)^{-1}$, where Z is a deterministic diagonal matrix such that $Z_{ij} = (z\mathbb{1}_{1 \leqslant i \leqslant W} + \widetilde{z}\mathbb{1}_{i>W})\,\delta_{ij}$, with two distinct spectral parameters $z \in \mathbb{C}_+ := \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$ and $\widetilde{z} \in \mathbb{C}_+ \cup \mathbb{R}$. In this paper, we prove a sharp bound for the local law of the generalized resolvent G for $W \gg N^{3/4}$. This bound is a key input for the proof of delocalization and bulk universality of random band matrices in [2]. Our proof depends on a fluctuations averaging bound on certain averages of polynomials in the resolvent entries, which will be proved in [10].

1 The model and the results.

1.1 The model. Our goal in this paper is to establish estimates on Green's functions which were used in the proof of delocalization conjecture and bulk universality for random band matrices. All results in this paper apply to both real and complex band matrices. For simplicity of notations, we consider only the real symmetric case. Random band matrices are characterized by the property that the matrix element H_{ij} becomes negligible if $\operatorname{dist}(i, j)$ exceeds the band width W. We shall restrict ourselves to the convention that $i, j \in \mathbb{Z}_N = \mathbb{Z} \cap (-N/2, N/2]$, and $i - j$ is defined modular N. More precisely, we consider the following matrix ensembles.

Definition 1.1 (Band matrix H_N with bandwidth W_N)**.** *Let H_N be an $N \times N$ matrix with real centered entries $(H_{ij}: i, j \in \mathbb{Z}_N)$ which are independent up to the condition $H_{ij} = H_{ji}$. We say that H_N is a random band matrix with (typical) bandwidth $W = W_N$ if*

$$s_{ij} := \mathbb{E}|H_{ij}|^2 = f(i - j) \tag{1.1}$$

The work of P.B. is partially supported by the NSF grant DMS#1513587. The work of H.-T. Y. is partially supported by NSF Grant DMS-1606305 and a Simons Investigator award. The work of J.Y. is partially supported by the NSF grant DMS#1552192.

for some non-negative symmetric function $f : \mathbb{Z}_N \to \mathbb{R}_+$ *satisfying*

$$\sum_{x\in\mathbb{Z}_N} f(x) = 1, \tag{1.2}$$

and there exist some (small) positive constant c_s *and (large) positive constant* C_s *such that*

$$c_s\, W^{-1} \cdot \mathbb{1}_{|x|\leqslant W} \leqslant f(x) \leqslant C_s\, W^{-1} \cdot \mathbb{1}_{|x|\leqslant C_s W}, \quad i, j \in \mathbb{Z}_N. \tag{1.3}$$

The method in this paper also allows to treat cases with exponentially small mass away from the band width (e.g. $f(x) \leqslant C_s W^{-1} e^{-c_s|x|^2/W^2}$). We work under the hypothesis (1.3) mainly for simplicity.

We assume that the random variables H_{ij} have arbitrarily high moments, in the sense that for any fixed $p \in \mathbb{N}$, there is a constant $\mu_p > 0$ such that

$$\max_{i,j} \left(\mathbb{E}|H_{ij}|^p\right)^{1/p} \leqslant \mu_p \operatorname{Var}\left(H_{ij}\right)^{1/2} \tag{1.4}$$

uniformly in N.

In this paper, we will *not* need the following moment condition assumed in Part I of this series [2]: there is fixed $\varepsilon_m > 0$ such that for $|i-j| \leqslant W$, $\min_{|i-j|\leqslant W} \left(\mathbb{E}\,\xi_{ij}^4 - (\mathbb{E}\,\xi_{ij}^3)^2 - 1\right) \geqslant N^{-\varepsilon_m}$, where $\xi_{ij} := H_{ij}(s_{ij})^{-1/2}$ is the normalized random variable with mean zero and variance one.

All the results in this paper will depend on the parameters c_s, C_s in (1.3) and μ_p in (1.4). But we will not track the dependence on c_s, C_s and μ_p in the proof.

Denote the eigenvalues of H_N by $\lambda_1 \leqslant \ldots \leqslant \lambda_N$. It is well-known that the empirical spectral measure $\frac{1}{N}\sum_{k=1}^N \delta_{\lambda_k}$ converges almost surely to the Wigner semicircle law with density

$$\rho_{\mathrm{sc}}(x) = \frac{1}{2\pi}\sqrt{(4-x^2)_+}.$$

The aim of this paper is to estimate "the generalized resolvent" $G(z, \widetilde{z}\,)$ of H_N defined by

$$G(z, \widetilde{z}\,) := \left(H_N - \begin{pmatrix} zI_{W\times W} & 0 \\ 0 & \widetilde{z}I_{(N-W)\times(N-W)} \end{pmatrix}\right)^{-1}, \quad z,\ \widetilde{z} \in \mathbb{C}^+ \cup \mathbb{R}, \tag{1.5}$$

where $\mathbb{C}^+$ denotes the upper half complex plane $\mathbb{C}^+ := \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$. The generalized resolvent is an important quantity used in Part I of this series [2]. The key point of this generalization, compared with the usual resolvent, is the freedom to choose different z and $\widetilde{z}$. To the best of our knowledge, the local law for this type of generalized resolvent has only been studied in the preceding paper [1], where it was assumed that $W \geqslant cN$ for some constant $c > 0$.

To understand the role of the generalized resolvent, we block-decompose the band matrix H_N and its eigenvectors as

$$H_N = \begin{pmatrix} A & B^* \\ B & D \end{pmatrix}, \quad \boldsymbol{\psi}_j := \begin{pmatrix} \mathbf{w}_j \\ \mathbf{p}_j \end{pmatrix},$$

where A is a $W \times W$ Wigner matrix. From the eigenvector equation $H\boldsymbol{\psi}_j = \lambda_j \boldsymbol{\psi}_j$, we get

$$Q_{\lambda_j}\mathbf{w}_j = \lambda_j \mathbf{w}_j, \quad Q_e := A - B^* \frac{1}{D-e} B.$$

Thus $\mathbf{w}_j$ is an eigenvector of $Q_e := A - B^*(D-e)^{-1}B$ with eigenvalue λ_j when $e = \lambda_j$. A key input to the proof of universality and QUE for random band matrices is an estimate on the Green's function of Q_e. Since some eigenvalues of D can be very close to e, the matrix $(D-e)^{-1}$ can be very singular. It is thus very difficult (if possible) to estimate the Green's function of Q_e directly. On the other hand, the Green's function of Q_e is just the $W \times W$ minor of the generalized resolvent $G(z, e)$ of H_N, which we find to be relatively more doable.

Due to the need in Part I, we will consider generalized resolvent for a general class of band matrices. More precisely, we introduce the following Definition 1.2. Here and throughout the rest of this paper, we will use the notation that for any $a, b \in \mathbb{Z}$,

$$[\![a, b]\!] := [a, b] \cap \mathbb{Z}.$$

Definition 1.2 (Definition of $H_\zeta^{\mathbf{g}}$). *For any sufficiently small $\zeta > 0$ and any $\mathbf{g} = (g_1, g_2, \cdots, g_N) \in \mathbb{R}^N$, H_ζ and $H_\zeta^{\mathbf{g}}$ will denote $N \times N$ real symmetric matrices satisfying the following properties. The entries $(H_\zeta)_{ij}$ are centered and independent up to the symmetry condition, satisfy* (1.4)*, and have variances*

$$\mathbb{E}|(H_\zeta)_{ij}|^2 = (s_\zeta)_{ij} := s_{ij} - \frac{\zeta(1+\delta_{ij})}{W}\mathbf{1}_{i,j\in[\![1,W]\!]},$$

where s_{ij}, $i, j \in \mathbb{Z}_N$, satisfy the conditions in Definition 1.1. Then the matrix $H_\zeta^{\mathbf{g}}$ is defined by

$$(H_\zeta^{\mathbf{g}})_{ij} := (H_\zeta)_{ij} - g_i\delta_{ij}.$$

We denote by S_0 and Σ the matrices with entries $(S_0)_{ij} = s_{ij}$ and $\Sigma_{ij} = \frac{(1+\delta_{ij})}{W}\mathbf{1}_{i,j\in[\![1,W]\!]}$, respectively. Then the matrix of variances is

$$S_\zeta := S_0 - \zeta\Sigma, \quad (S_\zeta)_{ij} = (s_\zeta)_{ij}.$$

1.2 The results. The generalized resolvent $G_\zeta^{\mathbf{g}}(z, \widetilde{z})$ of $H_\zeta^{\mathbf{g}}$ is defined similarly as in (1.5) by

$$G_\zeta^{\mathbf{g}}(z,\widetilde{z}) := \left(H_\zeta^{\mathbf{g}} - \begin{pmatrix} zI_{W\times W} & 0 \\ 0 & \widetilde{z}I_{(N-W)\times(N-W)}\end{pmatrix}\right)^{-1}.$$

Define $\big((M_\zeta^{\mathbf{g}})_i(z,\widetilde{z})\big)_{i=1}^N$ as the solution vector to the system of self-consistent equations

$$\Big((M_\zeta^{\mathbf{g}})_i(z,\widetilde{z})\Big)^{-1} = -z\mathbb{1}_{i\in[\![1,W]\!]} - \widetilde{z}\mathbb{1}_{i\notin[\![1,W]\!]} - g_i - \sum_j (s_\zeta)_{ij}(M_\zeta^{\mathbf{g}})_j(z,\widetilde{z}), \tag{1.6}$$

for $z, \widetilde{z} \in \mathbb{C}^+ \cup \mathbb{R}$ and $i \in \mathbb{Z}_N$, with the constraint that

$$(M_0^{\mathbf{0}})_i(\widetilde{z},\widetilde{z}) = m_{\mathrm{sc}}(\widetilde{z} + \mathrm{i}0^+),$$

where m_{sc} denotes the Stieltjes transform of the semicircle law

$$m_{\mathrm{sc}}(z) := \frac{-z + \sqrt{z^2-4}}{2}, \quad z \in \mathbb{C}^+. \tag{1.7}$$

(The existence, uniqueness and continuity of the solution is given by Lemma 1.3 below.) For simplicity of notations, we denote by $M_\zeta^{\mathbf{g}}(z,\widetilde{z})$ the diagonal matrix with entries

$$(M_\zeta^{\mathbf{g}})_{ij} := (M_\zeta^{\mathbf{g}})_i\delta_{ij}.$$

We will show that $M_\zeta^{\mathbf{g}}(z,\widetilde{z})$ is the asymptotic limit of the generalized resolvent $G_\zeta^{\mathbf{g}}(z,\widetilde{z})$. We now list some properties of $M_\zeta^{\mathbf{g}}$ needed for the proof of local law stated in Theorem 1.4. Its proof is delayed to Section 4.

Lemma 1.3. *Assume $|\operatorname{Re}\widetilde{z}| \leqslant 2-\kappa$ and $|\widetilde{z}| \leqslant \kappa^{-1}$ for some (small) constant $\kappa > 0$. Then there exist constants $c, C > 0$ such that the following statements hold.*

- *(Existence and Lipschitz continuity) If*

$$\zeta + \|\mathbf{g}\|_\infty + |z - \widetilde{z}| \leqslant c, \tag{1.8}$$

then there exist $(M_\zeta^{\mathbf{g}})_i(z,\widetilde{z})$, $i \in \mathbb{Z}_N$, which satisfy (1.6) *and*

$$\max_i \left|(M_\zeta^{\mathbf{g}})_i(z,\widetilde{z}) - m_{\mathrm{sc}}(\widetilde{z}+\mathrm{i}0^+)\right| \leqslant C\left(\zeta + \|\mathbf{g}\|_\infty + |z-\widetilde{z}|\right). \tag{1.9}$$

If, in addition, we have $\zeta' + \|\mathbf{g}'\|_\infty + |z' - \widetilde{z}'| \leqslant c$, then

$$\max_i \left|(M_{\zeta'}^{\mathbf{g}'})_i(z',\widetilde{z}') - (M_\zeta^{\mathbf{g}})_i(z,\widetilde{z})\right| \leqslant C\left(\|\mathbf{g}-\mathbf{g}'\|_\infty + |z'-z| + |\widetilde{z}'-\widetilde{z}| + |\zeta'-\zeta|\right). \tag{1.10}$$

- *(Uniqueness) The solution vector* $\left((M_\zeta^{\mathbf{g}})_i(z,\widetilde{z})\right)_{i=1}^N$ *to* (1.6) *is unique under* (1.8) *and the constraint*

$$\max_i \left|(M_\zeta^{\mathbf{g}})_i(z,\widetilde{z}) - m_{\mathrm{sc}}(\widetilde{z}+\mathrm{i}0^+)\right| \leqslant c.$$

We now state our results on the generalized resolvent of $H_\zeta^{\mathbf{g}}$. In this paper, we will always use τ to denote an arbitrarily small positive constant independent of N, and D to denote an arbitrarily large positive constant independent of N. Define for any matrix X the max norm

$$\|X\|_{\max} := \max_{i,j}|X_{ij}|.$$

The notations η_*, η^* and r in next theorem were used in Assumptions 2.3 and 2.4 of Part I of this series [2]. Their meanings are not important for this paper and the reader can simply view them as some parameters. In this paper, all the statements hold for sufficiently large N and we will not repeat it everywhere.

Theorem 1.4 (Local law)**.** *Define a set of parameters with some constants* $\varepsilon_*, \varepsilon^* > 0$*:*

$$\eta_* := N^{-\varepsilon_*}, \quad \eta^* := N^{-\varepsilon^*} \quad r := N^{-\varepsilon_*+3\varepsilon^*}, \quad T := N^{-\varepsilon_*+\varepsilon^*}, \quad 0 < \varepsilon^* \leqslant \varepsilon_*/20. \tag{1.11}$$

Fix any $|e| < 2-\kappa$ *for some constant* $\kappa > 0$*. Then for any deterministic* z*,* ζ*,* $\mathbf{g}$ *satisfying*

$$|\operatorname{Re} z - e| \leqslant r, \quad \eta_* \leqslant \operatorname{Im} z \leqslant \eta^*, \quad 0 \leqslant \zeta \leqslant T, \quad \|\mathbf{g}\|_\infty \leqslant W^{-3/4}, \tag{1.12}$$

and W*,* ε_**,* ε^* *satisfying*

$$\log_N W \geqslant \max\left\{\frac{6}{7}+\varepsilon^*,\ \frac{3}{4}+\frac{3}{4}\varepsilon_*+\varepsilon^*\right\}, \tag{1.13}$$

we have that for any fixed $\tau > 0$ *and* $D > 0$*,*

$$\mathbb{P}\left(\|G_\zeta^{\mathbf{g}}(z,e) - M_\zeta^{\mathbf{g}}(z,e)\|_{\max} \geqslant N^\tau\left(\frac{1}{\sqrt{W\operatorname{Im} z}} + \frac{N^{1/2}}{W}\right)\right) \leqslant N^{-D}. \tag{1.14}$$

In fact, the last estimate holds under the weaker assumption

$$\log_N W \geqslant \max\left\{\frac{3}{4}+\varepsilon^*,\ \frac{1}{2}+\varepsilon_*+\varepsilon^*\right\}. \tag{1.15}$$

We will refer to the first statement, i.e., (1.14) under the assumption (1.13), as the weak form of this theorem, and the statement (1.14) under assumption (1.15) as the strong form. This paper gives a full and self-contained proof for the weak form, which helps the reader understand the basic strategy of our proof. On the other hand, the proof for the strong form is much more involved, and we include a substantial part into a separate paper [10]. Only the strong form of Theorem 1.4 was used in part I of this series [2], where we took $\log_N W > 3/4$, $\varepsilon_* < 1/4$ and ε^* to be a sufficiently small constant.

The main purpose of this part and part III [10] of this series is to prove the above Theorem 1.4. In fact, the bound (1.14) is almost optimal under our setting in the sense that it (at least) gives the correct size of $\mathbb{E}|(G_\zeta^{\mathbf{g}})_{ij}|^2$ for $i \neq j$ up to an N^τ factor. This sharp bound is very important for the proof of the complete delocalization of eigenvectors and the bulk universality of random band matrices in part I [2]. As explained there, the bound must be of order $\mathrm{o}(W/N)$ to allow the application of the so-called *mean field reduction* method, which was introduced in [1] and is the starting point of this series. Compared with the local law for regular resolvents, the main difficulty in proving the local law for the generalized resolvents is due to the small and even vanishing imaginary part of $\widetilde{z}$. As a result, some key inputs, such as Ward's identity (see (3.2)) for the regular resolvents estimates are missing. In fact, as discussed before, the case $\|G(z,\widetilde{z})\|_{\max} = \infty$ could occur when $\widetilde{z} = e$ is real. This difficulty has already appeared in the case $W \geqslant cN$ in [1], where some "uncertainty principle" was introduced to solve this problem. Unfortunately, this method seems difficult to apply in the $W \ll N$ case. Instead, in this paper, we shall use a totally different strategy, i.e, the T-equation method, which was introduced in [4]. Moreover, we have to improve the induction (bootstrap) argument used in [4], as explained below. We remark that the proofs of the weak form and strong form of Theorem 1.4 are completely parallel, except that we will apply a stronger T-equation estimate (Lemma 2.14) than the one (Lemma 2.8) used in the proof of the weak form. We shall give a simple proof of the weak T-equation estimate using the standard fluctuation averaging mechanism as in the previous proof of local semicircle law [5, 8]. The proof of the strong T-equation estimate is based on an improved (and substantially more involved) fluctuation averaging result, whose proof is delayed to part III of this series [10].

1.3 Sketch of proof. In the following discussion, for two random variables X and Y, we shall use the notation $X \prec Y$ if for any fixed $\tau > 0$, $|X| \leqslant N^{\tau}|Y|$ with high probability for large enough N.

We define the T matrix with entries

$$T_{ij} := \sum_k S_{ik}|G_{kj}|^2, \quad G \equiv G_\zeta^{\mathbf{g}}, \quad S_{ik} \equiv (S_\zeta)_{ik}, \tag{1.16}$$

With a standard self-consistent equation estimate (see Lemma 2.1), one can show that

$$\|G - M\|_{\max}^2 \prec \|T\|_{\max}, \quad M \equiv M_\zeta^{\mathbf{g}}. \tag{1.17}$$

Our proof of Theorem 1.4 is based on an induction argument combined with a self-consistent T-equation estimate as explained below. We introduce the following notation:

$$|||G|||^2(z, \widetilde{z}) := \max_j \sum_{1 \leqslant i \leqslant N} |G_{ij}(z, \widetilde{z})|^2, \quad \Lambda(z, \tilde{z}) := \|G - M\|_{\max}(z, \widetilde{z}). \tag{1.18}$$

Fix z and $\operatorname{Re}\widetilde{z} = e$. We perform the induction with respect to the imaginary part of $\widetilde{z}$. Define a sequence of $\widetilde{z}_n$ such that

$$\operatorname{Im}\widetilde{z}_n = N^{-n\varepsilon}\operatorname{Im} z, \quad \operatorname{Re}\widetilde{z}_n = e,$$

for small enough constant $\varepsilon > 0$. In the $n = 0$ case with $\operatorname{Im}\widetilde{z}_0 = \operatorname{Im} z$, using the methods in [5, 8], we can obtain the local law (1.14) for $G(z, \widetilde{z}_0)$. Suppose we has proved the local law for $G(z, \widetilde{z}_{n-1})$:

$$\Lambda(z, \widetilde{z}_{n-1}) \prec \Phi_{\text{goal}}, \quad \Phi_{\text{goal}} := \frac{1}{\sqrt{W \operatorname{Im} z}} + \frac{N^{1/2}}{W}. \tag{1.19}$$

Then with $\operatorname{Im}\widetilde{z}_n = N^{-\varepsilon}\operatorname{Im} z_{n-1}$ and a simple (but quite sharp up to an $N^{2\varepsilon}$ factor) L^2-estimate, we get a bound on the n-th level:

$$|||G|||^2(z, \widetilde{z}_n) \prec N\widetilde{\Phi}^2, \quad \widetilde{\Phi}^2 := N^{2\varepsilon}\Phi_{\text{goal}}^2, \tag{1.20}$$

which gives a rough bound $\Phi^{(0)}$ by the self-consistent equation estimate (1.17):

$$\|T\|_{\max}(z, \widetilde{z}_n) \leqslant \frac{C_s}{W}|||G|||^2(z, \widetilde{z}_n) \prec (\Phi^{(0)})^2 \Rightarrow \Lambda(z, \widetilde{z}_n) \prec \Phi^{(0)}, \quad \Phi^{(0)} := \sqrt{\frac{N}{W}}\widetilde{\Phi}, \tag{1.21}$$

where C_s is the constant from (1.3). Note that $\widetilde{\Phi}$ is very close to Φ_{goal}, while $\Phi^{(0)}$ is not. Now with the strong T-equation estimate (see Lemma 2.14), one can get an improved bound $(\Phi^{(1)})^2$ on T as follows:

$$\|T\|_{\max}(z, \widetilde{z}_n) \prec (\Phi^{(1)})^2 \Rightarrow \Lambda(z, \widetilde{z}_n) \prec \Phi^{(1)}, \quad \Phi^{(1)} := \Phi_{\text{goal}}^2 + \left(\frac{N}{W \operatorname{Im} z} + \frac{N^2}{W^2}\right)\left(\widetilde{\Phi}^2 + N^{-1/2}\right)(\Phi^{(0)})^2, \tag{1.22}$$

where we used (1.17) to get a better bound $\Lambda(z, \widetilde{z}_n) \prec \Phi^{(1)}$. With (1.15), one can verify that $\Phi^{(1)} \leqslant \Phi_{\text{goal}} + N^{-\varepsilon'}\Phi^{(0)}$ for some constant $\varepsilon' > 0$. After at most $l := 1/\varepsilon'$ many iterations with (1.22) and (1.17), i.e. $\Phi^{(0)} \to \Phi^{(1)} \to \cdots \to \Phi^{(l)}$, we can obtain the local law (1.19) for $G(z, \widetilde{z}_n)$, which is used as the input for the next induction. The key point of this induction argument is that one has a good L^2-bound (1.20) inherited from the local law on the upper level, and this L^2-bound can be used in the T-equation estimate (1.22) to give an improved bound for $\Lambda(z, \widetilde{z}_n)$ on this level. Finally, after finitely many inductions in n, we can obtain the local law (1.14) for, say, $G(z, e + \mathrm{i}N^{-10})$. Then with a continuity argument, we can prove the local law (1.14) for $G(z, e)$. In Fig. 1, we illustrate the flow of the induction argument with a diagram.

We remark that the above induction argument is not a continuity argument, as used e.g. in the works [5, 7, 8] on local semicircle law of regular resolvents. The multiplicative steps $\operatorname{Im}\widetilde{z}_n \to N^{-\varepsilon}\operatorname{Im}\widetilde{z}_n$ that we made are far too large for a continuity argument to work. The main reason for choosing this multiplicative step is that the T-equation estimate can only be applied for O(1) number of times due to the degrade of the probability set (see Remark 2.9).

The main difficulty of our proof lies in establishing the T-equation estimate (1.22). The starting point is a self-consistent equation for the T matrix, i.e. the T-equation, see (2.14) below. In this paper, we focus on proving the stability of the T-equation, i.e. bounding $\| \left(1 - S|M|^2\right)^{-1} S\|_{\max}$ in (2.14), where we abbreviate $S \equiv S_\zeta$. For regular resolvent of generalized Wigner matrices (i.e. $\tilde{z} = z$, $\zeta = 0$ and $\mathbf{g} = \mathbf{0}$), we have

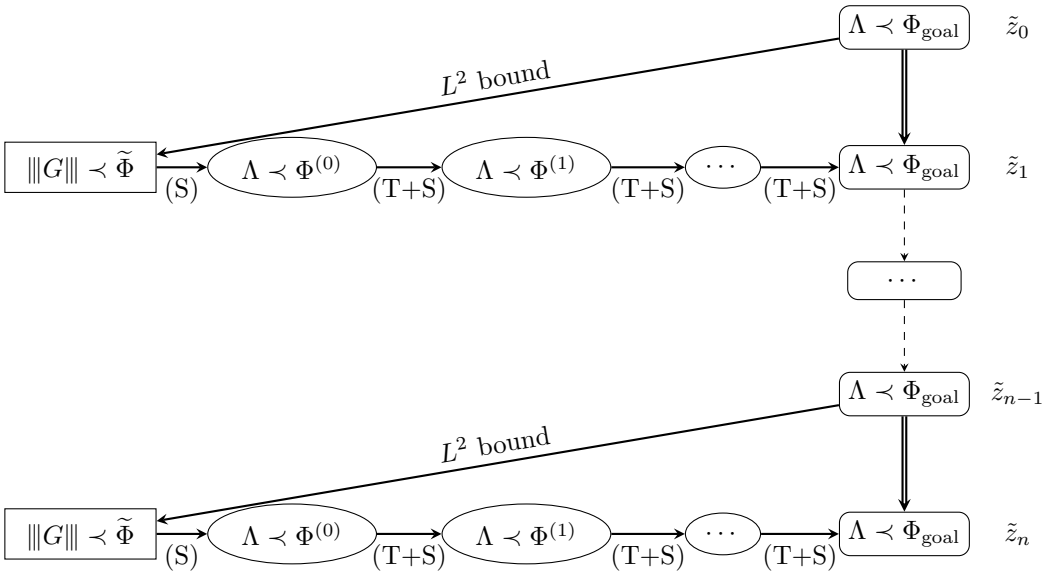

Figure 1: The diagram for the induction argument with respect to n. At each level $n-1$, we obtain the local law (1.19), which gives the rough bound $\Phi^{(0)}$ on level n through (1.20) and (1.21). Applying (1.22) and (1.17) iteratively, one can improve the initial bound $\Phi^{(0)}$ to the sharp bound $\Phi_{\rm goal}$. In the diagram, (S) stands for an application of the self-consistent equation estimate (1.17), and (T+S) stands for an application of the T-equation estimate (1.22) followed by a self-consistent equation estimate (1.17).

$|M| \leqslant 1 - c\,\mathrm{Im}\, z$ for some constant $c > 0$. However, in our general setting and in particular when $\mathrm{Im}\,\tilde{z}$ is small, we actually have $\|M\|_\infty > 1$ and $\|S|M|^2\|_{l^\infty \to l^\infty} > 1$. Therefore, the usual Taylor expansion approach cannot be used (in fact, it is not even easy to see that 1 is outside the spectrum of $|M|^2 S$). In this paper, we will establish the following bound

$$\left\|(1 - S|M|^2)^{-1} S\right\|_{\max} = \mathrm{O}\left(\frac{1}{W\,\mathrm{Im}\, z} + \frac{N}{W^2}\right).$$

One important component for the proof is the estimate $\sum_i (|M_i|^2 - 1) \leqslant -cW\,\mathrm{Im}\, z$ for some constant $c > 0$. To see this bound is useful, we can intuitively view $(|M|^2 S)^n$ as an n-step inhomogeneous random walk on $\mathbb{Z}_N$ with annihilation, where the average annihilation rate is $-W\,\mathrm{Im}\, z/N$ by the above bound. This shows that we can explore some decay properties of $(|M|^2 S)^n$ as n increase, which may give some useful bounds on the Taylor expansion of $(1 - S|M|^2)^{-1}$. However, our proof actually will not follow this heuristic argument, see Section 4.

Finally, to finish the proof of the strong version of the T-equation estimate (Lemma 2.14), we need a fluctuation averaging results for a quantity of the form $N^{-1}\sum_k \mathcal{E}_k$, where $\mathcal{E}_k$'s are some polynomials of the generalized resolvent entries. The proof involves a new graphical method and we include it in part III of this series [10].

2 Tools for the proof of Theorem 1.4

The basic strategy to prove Theorem 1.4 is to apply the self-consistent equation estimate: Lemma 2.1, and the T-equation estimate: Lemma 2.8 or 2.14, in turns. We collect these results in this section, and use them to prove Theorem 1.4 in next section.

For simplicity, we will often drop the superscripts ζ and $\mathbf{g}$ from our notations. In particular, G and M are always understood as $G_\zeta^{\mathbf{g}}$ and $M_\zeta^{\mathbf{g}}$, while H and S are understood as H_ζ and S_ζ in the rest of this paper.

In the proof, for quantities A_N and B_N, we will use the notations $A_N = \mathrm{O}(B_N)$ and $A_N \asymp B_N$ to mean that $|A_N| \leqslant C|B_N|$ and $C^{-1}|B_N| \leqslant |A_N| \leqslant C|B_N|$, respectively, for some constant $C > 0$.

2.1 The self-consistent equation estimate. The self-consistent equation estimate is the starting point of almost every proof of the local law of the (generalized) resolvents of random matrices. We now state the self-consistent equation estimate for our model.

Lemma 2.1 (Self-consistent equation estimate)**.** *Suppose that* $|\operatorname{Re}\widetilde{z}| \leqslant 2-\kappa$ *for some constant* $\kappa > 0$*. Then there exists constant* $c_0 > 0$ *such that if*

$$\zeta + \|\mathbf{g}\|_\infty + |z-\widetilde{z}| \leqslant c_0,$$

then the following statement holds. If there exist some fixed $\delta > 0$ *and some deterministic parameter* $\Phi \geqslant W^{-1/2}$ *such that*

$$\|G(z,\widetilde{z}) - M(z,\widetilde{z})\|_{\max} \leqslant N^{-\delta}, \quad \|T\|_{\max} \leqslant \Phi^2, \tag{2.1}$$

in a subset Ω *of the sample space of the random matrices, then for any fixed* $\tau > 0$ *and* $D > 0$*,*

$$\mathbb{P}\left(\mathbf{1}_\Omega \|G(z,\widetilde{z}) - M(z,\widetilde{z})\|_{\max} \geqslant N^\tau \Phi\right) \leqslant N^{-D}. \tag{2.2}$$

Note that by the definition of T-matrix in (1.16), we have

$$\|T\|_{\max} \leqslant \|G(z,\widetilde{z}) - M(z,\widetilde{z})\|_{\max}^2 + \mathrm{O}(W^{-1}).$$

Hence we can always choose $\Phi = \mathrm{O}(N^{-\delta})$ in (2.1). The proof of Lemma 2.1 follows the standard idea of using a vector-level self consistent equation method [5,8]. In preparation for the proof, we recall the following definition of minors.

Definition 2.2 (Minors)**.** *For any* $N \times N$ *matrix* A *and* $\mathbb{T} \subset \{1,\dots,N\}$*, we define the minor of the first kind* $A^{[\mathbb{T}]}$ *as the* $(N-|\mathbb{T}|)\times(N-|\mathbb{T}|)$ *matrix with*

$$(A^{[\mathbb{T}]})_{ij} \;:=\; A_{ij}, \quad i,j \notin \mathbb{T}.$$

For any $N \times N$ *invertible matrix* B*, we define the minor of the second kind* $B^{(\mathbb{T})}$ *as the* $(N-|\mathbb{T}|)\times(N-|\mathbb{T}|)$ *matrix with*

$$(B^{(\mathbb{T})})_{ij} = \left((B^{-1})^{[\mathbb{T}]}\right)^{-1}_{ij}, \quad i,j \notin \mathbb{T},$$

whenever $(B^{-1})^{[\mathbb{T}]}$ *is invertible. Note that we keep the names of indices when defining the minors. By definition, for any sets* $\mathbb{U},\mathbb{T} \subset \{1,\dots,N\}$*, we have*

$$(A^{[\mathbb{T}]})^{[\mathbb{U}]} = A^{[\mathbb{T}\cup\mathbb{U}]}, \quad (B^{(\mathbb{T})})^{(\mathbb{U})} = B^{(\mathbb{T}\cup\mathbb{U})}.$$

For convenience, we shall also adopt the convention that for $i \in \mathbb{T}$ *or* $j \in \mathbb{T}$*,*

$$(A^{[\mathbb{T}]})_{ij} = 0, \quad (B^{(\mathbb{T})})_{ij} = 0.$$

For $\mathbb{T} = \{a\}$ *or* $\mathbb{T} = \{a,b\}$*, we shall abbreviate* $(\{a\}) \equiv (a)$ *and* $(\{a,b\}) \equiv (ab)$*.*

Remark 2.3. *In previous works, e.g. [6, 8], we have used the notation* $(\cdot)$ *for both the minor of the first kind and the minor of the second kind. Here we try to distinguish between* $(\cdot)$ *and* $[\cdot]$ *in order to be more rigorous.*

The following identities were proved in Lemma 4.2 of [8] and Lemma 6.10 of [6].

Lemma 2.4 (Resolvent identities)**.** *For an invertible matrix* $B \in \mathbb{C}^{N\times N}$ *and* $k \notin \{i,j\}$*, we have*

$$B_{ij} \;=\; B^{(k)}_{ij} + \frac{B_{ik}B_{kj}}{B_{kk}}\,, \quad \frac{1}{B_{ii}} \;=\; \frac{1}{B^{(k)}_{ii}} - \frac{B_{ik}B_{ki}}{B^{(k)}_{ii}B_{ii}B_{kk}}\,, \tag{2.3}$$

and

$$\frac{1}{B_{ii}} \;=\; (B^{-1})_{ii} - \sum_{k,l}^{(i)} (B^{-1})_{ik} B^{(i)}_{kl} (B^{-1})_{li}. \tag{2.4}$$

Moreover, for $i \neq j$ *we have*

$$B_{ij} \;=\; -B_{ii}\sum_{k}^{(i)} (B^{-1})_{ik} B^{(i)}_{kj} \;=\; -B_{jj}\sum_{k}^{(j)} B^{(j)}_{ik} (B^{-1})_{kj}\,. \tag{2.5}$$

The above equalities are understood to hold whenever the expressions in them make sense.

Since the N^τ factor and the N^{-D} bound for small probability event appear very often in our proof, we introduce the following notations.

Definition 2.5. *For any non-negative A, we denote*

$$\mathrm{O}_\tau(A) := \mathrm{O}(N^{\mathrm{O}(\tau)}A).$$

We shall say an event $\mathcal{E}_N$ holds with high probability (w.h.p.) if for any fixed $D > 0$,

$$\mathbb{P}(\mathcal{E}_N) \geqslant 1 - N^{-D}$$

for sufficiently large N. Moreover, we say $\mathcal{E}_N$ holds with high probability in Ω if for any fixed $D > 0$,

$$\mathbb{P}(\Omega \setminus \mathcal{E}_N) \leqslant N^{-D}$$

for sufficiently large N.

The following lemma gives standard large deviation bounds that will be used in the proof of Lemma 2.1.

Lemma 2.6 (Lemma 3.5 of [9])**.** *Let (X_i) be a family of independent random variables and (b_i), (B_{ij}) be deterministic families of complex numbers, where $i, j = 1, \ldots, N$. Suppose the entries X_i satisfy $\mathbb{E}X_i = 0$, $\mathbb{E}|X_i|^2 = 1$ and the bound(1.4). Then for any fixed $\tau > 0$, we have*

$$\left|\sum_i b_i X_i\right| \leqslant N^\tau \left(\sum_i |b_i|^2\right)^{1/2}, \quad \left|\sum_{i,j} \bar{X}_i B_{ij} X_j\right| \leqslant N^\tau \left(\sum_{i,j} |B_{ij}|^2\right)^{1/2},$$

with high probability.

The following lemma provides estimates on the entries of $(1 - M^2 S)^{-1}$ and $\left(1 - S|M|^2\right)^{-1} S$. It will be used in the proof of Lemma 2.1 and Theorem 1.4, and its proof is delayed until Section 4.

Lemma 2.7. *Suppose that the assumptions for the strong form of Theorem 1.4, i.e.,* (1.11), (1.12) *and* (1.15)*, hold. If $\widetilde{z}$ satisfies*

$$\operatorname{Re}\widetilde{z} = e, \quad 0 \leqslant \operatorname{Im}\widetilde{z} \leqslant \operatorname{Im} z,$$

then we have for $M \equiv M_\zeta^{\mathbf{g}}(z, \widetilde{z})$ and $S \equiv S_\zeta$,

$$\left[(1 - M^2 S)^{-1}\right]_{ij} = \begin{cases} \delta_{ij} + \mathrm{O}(W^{-1}), & \text{if} \quad |i-j| \leqslant (\log N)^2 W \\ \mathrm{O}(N^{-c \log N}), & \text{if} \quad |i-j| > (\log N)^2 W \end{cases}, \tag{2.6}$$

and

$$\left\| \left(1 - S|M|^2\right)^{-1} S \right\|_{\max} = \mathrm{O}\left(\frac{1}{W \operatorname{Im} z} + \frac{N}{W^2}\right). \tag{2.7}$$

Now we can give the proof of Lemma 2.1.

Proof of Lemma 2.1. The following proof is fairly standard in random matrix theory and we will omit some details. For simplicity, we drop ζ and $\mathbf{g}$ in superscripts. Using (2.5), we have $G_{ij} = -G_{ii} \sum_k^{(i)} H_{ik} G_{kj}^{(i)}$ for $i \neq j$. Since the elements in $\{H_{ik}\}_{k=1}^N$ are independent of $G^{(i)}$, by the standard large deviations estimates in Lemma 2.6, we have that for any fixed $\tau > 0$ and $D > 0$,

$$\mathbb{P}\left(|G_{ij}|^2 \leqslant N^\tau |G_{ii}|^2 \sum_k S_{ik} |G_{kj}^{(i)}|^2\right) \geqslant 1 - N^{-D}, \quad i \neq j. \tag{2.8}$$

Since $G_{ii} \asymp 1$ in Ω, (2.8) implies that

$$\mathbb{P}\left(\mathbf{1}_\Omega |G_{ij}|^2 = \mathrm{O}_\tau\left(\sum_k S_{ik} |G_{kj}^{(i)}|^2\right)\right) \geqslant 1 - N^{-D}, \quad i \neq j.$$

By (2.3), the definition of T in (1.16), and the bound for T in (2.1), we have

$$\sum_k S_{ik}|G^{(i)}_{kj}|^2 \leqslant 2\sum_k S_{ik}|G_{kj}|^2 + 2\sum_k S_{ik}\frac{|G_{ki}G_{ij}|^2}{|G_{ii}|^2} = \mathrm{O}(\Phi^2) \quad \text{in } \Omega.$$

Therefore, we obtain (2.2) for the $i \neq j$ case.

For the diagonal case, we define

$$\mathcal{Z}_i := Q_i\left(\sum_{kl}^{(i)} H_{ik}H_{il}G^{(i)}_{kl}\right) - H_{ii}.$$

Using (2.4), (2.3), the off-diagonal case for (2.2) we just proved, and the standard large deviations estimates in Lemma 2.6, we can get that for any fixed $\tau > 0$,

$$\frac{1}{G_{ii}} \;=\; -z\mathbb{1}_{i\in[\![1,W]\!]} - \widetilde{z}\mathbb{1}_{i\notin[\![1,W]\!]} - g_i - \sum_j S_{ij}G_{jj} - \mathcal{Z}_i + \mathrm{O}_\tau(\Phi^2), \quad \text{with } \mathcal{Z}_i = \mathrm{O}_\tau\left(\Phi\right),$$

holds with high probability in Ω. With the definition of M_i in (1.6), we have

$$G^{-1}_{ii} - M^{-1}_i = -\sum_j S_{ij}\left(G_{jj} - M_j\right) + \mathrm{O}_\tau(\Phi), \quad \textit{w.h.p. in } \Omega,$$

which implies

$$M_i - G_{ii} = -\sum_j M_i^2 S_{ij}\left(G_{jj} - M_j\right) + \mathrm{O}_\tau(\Phi) + \mathrm{O}\left(\max_i |G_{ii} - M_i|^2\right), \quad \textit{w.h.p. in } \Omega.$$

We rewrite the above estimate as

$$\sum_j (1 - M^2S)_{ij}\left(G_{jj} - M_j\right) = \mathrm{O}_\tau(\Phi) + \mathrm{O}\left(\max_i |G_{ii} - M_i|^2\right).$$

Then with (2.6) and the first bound in (2.1), we can get (2.2) for the diagonal entries and complete the proof of Lemma 2.1 □

2.2 The T-equation estimate. A key component for the proof of Theorem 1.4 is the self-consistent equation for the T variables. It leads to a self-improved bound on $\|G - M\|_{\max}$. This kind of approach was also used in [4] to prove a weak type delocalization result for random band matrices. To help the reader understand the proof, we first prove a weak T-equation estimate, i.e. Lemma 2.8, which will give the weak form of Theorem 1.4. The stronger T-equation estimate will be stated in Lemma 2.14, and its proof is put in the companion paper [10].

Lemma 2.8 (Weak T-equation estimate)**.** *Under the assumptions of Theorem 1.4 (i.e.,* (1.11), (1.12), (1.15) *and the assumption on e), the following statements hold provided $\varepsilon_* > 0$ is a sufficiently small constant. Let $\widetilde{z}$ satisfy*

$$\mathrm{Re}\,\widetilde{z} = e, \quad N^{-10} \leqslant \mathrm{Im}\,\widetilde{z} \leqslant \mathrm{Im}\,z, \tag{2.9}$$

and Φ be any deterministic parameter satisfying

$$W^{-1} \leqslant \Phi^2 \leqslant N^{-\delta}$$

for some fixed $\delta > 0$. Fix some z and $\widetilde{z}$ (which can depend on N). If for any constants $\tau' > 0$ and $D' > 0$,

$$\mathbb{P}\left(\|G(z,\widetilde{z}) - M(z,\widetilde{z})\|_{\max} \geqslant N^{\tau'}\Phi\right) \leqslant N^{-D'}, \tag{2.10}$$

then for any fixed (small) $\tau > 0$ and (large) $D > 0$, we have

$$\mathbb{P}\left(\|T(z,\widetilde{z})\|_{\max} \geqslant N^{\tau}(\Phi^w_{\#})^2\right) \leqslant N^{-D}, \quad (\Phi^w_{\#})^2 := \left(\frac{N}{W\,\mathrm{Im}\,z} + \frac{N^2}{W^2}\right)(\Phi^3 + N^{-1}). \tag{2.11}$$

Furthermore, if the parameter Φ satisfies

$$\Phi \leqslant \min\left\{\frac{W}{N^{1+\varepsilon_*+\varepsilon^*}}, \frac{W^2}{N^{2+\varepsilon^*}}\right\}, \tag{2.12}$$

then for any fixed $\tau > 0$ and $D > 0$ we have

$$\|G(z,\widetilde{z}) - M(z,\widetilde{z})\|_{\max} \leqslant \Phi N^{-\frac{1}{3}\varepsilon^*} + N^{\tau}\left(\frac{1}{\sqrt{W \operatorname{Im} z}} + \frac{N^{1/2}}{W}\right) \tag{2.13}$$

with probability at least $1 - N^{-D}$.

Remark 2.9. *The above statements should be understood as follows. For any small constant $\tau > 0$ and large constant $D > 0$, (2.11) and (2.13) hold if (2.10) holds for some constants τ', D' that depend on τ and D. In general, we need to take $\tau' < \tau$ to be sufficiently small and $D' > D$ to be sufficiently large. Compared with Lemma 2.1, we lose a much "larger" portion of the probability set. Hence Lemma 2.8 can only be iterated for $\mathrm{O}(1)$ number of times, while Lemma 2.1 can be applied for $\mathrm{O}(N^C)$ times for any fixed $C > 0$.*

Proof of Lemma 2.8. From the defining equation (1.16) of T, we add and subtract $\sum_k S_{ik}|M_k|^2 T_{kj}$ so that

$$T_{ij} = \sum_k S_{ik}|M_k|^2 T_{kj} + \sum_k S_{ik}\left(|G_{kj}|^2 - |M_k|^2 T_{kj}\right).$$

Therefore, we have

$$T_{ij} = \sum_k \left[\left(1 - S|M|^2\right)^{-1} S\right]_{ik}\left(|G_{kj}|^2 - |M_k|^2 T_{kj}\right). \tag{2.14}$$

Isolating the diagonal terms, we can write the T-equation as

$$T_{ij} = T_{ij}^0 + \sum_{k\neq j}\left[(1 - S|M|^2)^{-1}S\right]_{ik}\left(|G_{kj}|^2 - |M_k|^2 T_{kj}\right), \quad T_{ij}^0 := \left[(1 - S|M|^2)^{-1}S\right]_{ij}\left(|G_{jj}|^2 - |M_j|^2 T_{jj}\right). \tag{2.15}$$

By the definition of T, the assumption (2.10) and the estimate (1.9) on M_i, we can get the simple bounds $G_{jj} = \mathrm{O}(1)$ and $T_{jj} = \mathrm{O}_\tau(\Phi^2)$. Applying these bounds to the definition of T_{ij}^0, we get

$$T_{ij}^0 = \mathrm{O}\left(\left[(1 - S|M|^2)^{-1}S\right]_{ij}\right), \tag{2.16}$$

which will be shown to be the main term of T_{ij} up to an N^τ factor. By (2.7) and the condition (1.12) on $\operatorname{Im} z$, we have

$$\left[(1 - S|M|^2)^{-1}S\right]_{ij} = \mathrm{O}\left(\frac{1}{W \operatorname{Im} z} + \frac{N}{W^2}\right). \tag{2.17}$$

Definition 2.10 ($\mathbb{E}_k$, P_k and Q_k)**.** *We define $\mathbb{E}_k$ as the partial expectation with respect to the k-th row and column of H, i.e. $\mathbb{E}_k(\cdot) := \mathbb{E}(\cdot|H^{[k]})$. For simplicity, we will also use the notations*

$$P_k := \mathbb{E}_k, \quad Q_k := 1 - \mathbb{E}_k. \tag{2.18}$$

Using this definition and the bound (2.17), we rewrite the off-diagonal terms in (2.15) into two parts:

$$\begin{aligned}&\sum_{k\neq j}\left[(1 - S|M|^2)^{-1}S\right]_{ik}\left(|G_{kj}|^2 - |M_k|^2 T_{kj}\right)\\ &= \left(\frac{N}{W \operatorname{Im} z} + \frac{N^2}{W^2}\right)\left(\sum_{k\neq j} c_k\left(\mathbb{E}_k|G_{kj}|^2 - |M_k|^2 T_{kj}\right) + \sum_{k\neq j} c_k Q_k|G_{kj}|^2\right),\end{aligned} \tag{2.19}$$

where c_k is a sequence of deterministic numbers satisfying

$$c_k := \left[(1 - S|M|^2)^{-1}S\right]_{ik}\left(\frac{N}{W \operatorname{Im} z} + \frac{N^2}{W^2}\right)^{-1} = \mathrm{O}(N^{-1}).$$

The following two lemmas provide estimates for the two parts in (2.19), where Lemma 2.12 is a standard fluctuation averaging lemma.

Lemma 2.11. *Suppose that b_k, $k \in \mathbb{Z}_N$, are deterministic coefficients satisfying $\max_k |b_k| = \mathrm{O}(N^{-1})$. Then under the assumptions of Lemma 2.8, we have that for any fixed (small) $\tau > 0$,*

$$\sum_{k \neq j} b_k \left(\mathbb{E}_k |G_{kj}|^2 - |M_k|^2 T_{kj} \right) = \mathrm{O}_\tau \left(\Phi^3 \right), \quad j \in \mathbb{Z}_N, \tag{2.20}$$

with high probability.

Proof. By (2.5) and (2.10), we have $-\sum_l^{(k)} H_{kl} G_{lj}^{(k)} = G_{kj}/G_{kk} = \mathrm{O}_\tau(\Phi)$ and $G_{kk} - M_k = \mathrm{O}_\tau(\Phi)$ (*w.h.p.*). Then we can obtain that for $k \neq j$,

$$\mathbb{E}_k |G_{kj}|^2 = \mathbb{E}_k |M_k|^2 \left| \sum_l^{(k)} H_{kl} G_{lj}^{(k)} \right|^2 + \mathrm{O}_\tau(\Phi^3) = |M_k|^2 \sum_l^{(k)} s_{kl} \left| G_{lj}^{(k)} \right|^2 + \mathrm{O}_\tau(\Phi^3) \tag{2.21}$$

with high probability. Using (2.3), we have

$$G_{lj}^{(k)} = G_{lj} + \mathrm{O}_\tau(|G_{lk}||G_{kj}|) = G_{lj} + \mathrm{O}_\tau(\Phi^2), \quad l, j \neq k,$$

with high probability. Inserting it into (2.21) and using the definition (1.16), we can obtain (2.20). □

Lemma 2.12. *Suppose that b_k, $k \in \mathbb{Z}_N$ are deterministic coefficients satisfying $\max_k |b_k| = \mathrm{O}(N^{-1})$. Then under the assumptions of Lemma 2.8, we have for any fixed (large) $p \in 2\mathbb{N}$ and (small) $\tau > 0$,*

$$\mathbb{E} \left| \sum_{k \neq j} b_k Q_k |G_{kj}|^2 \right|^p \leqslant \left(N^\tau \Phi^3 \right)^p, \quad j \in \mathbb{Z}_N. \tag{2.22}$$

Proof. Our proof follows the arguments in [5, Appendix B]. We consider the decomposition of the space of random variables using P_k and Q_k defined in (2.18). It is evident that P_k and Q_k are projections, $P_k + Q_k = 1$, $P_k Q_k = 0$, and all of these projections commute with each other. For a set $A \subset \mathbb{Z}_N$, we denote $P_A := \prod_{k \in A} P_k$ and $Q_A := \prod_{k \in A} Q_k$. Now fix any $j \in \mathbb{Z}_N$, we set $X_k := Q_k |G_{kj}|^2$. Then for $p \in 2\mathbb{N}$, we can write

$$\begin{aligned} \mathbb{E} \left| \sum_{k \neq j} b_k X_k \right|^p &= \sum_{k_1, k_2, \ldots, k_p}^{*} c_{\mathbf{k}} \mathbb{E} \prod_{s=1}^{p} X_{k_s} = \sum_{\mathbf{k}}^{*} c_{\mathbf{k}} \mathbb{E} \prod_{s=1}^{p} \left(\prod_{r=1}^{p} (P_{k_r} + Q_{k_r}) X_{k_s} \right) \\ &= \sum_{\mathbf{k}}^{*} c_{\mathbf{k}} \sum_{A_1, \ldots, A_p \subset [\mathbf{k}]} \mathbb{E} \prod_{s=1}^{p} \left(P_{A_s^c} Q_{A_s} X_{k_s} \right), \end{aligned}$$

where $\mathbf{k} := (k_1, k_2, \ldots, k_p)$, $[\mathbf{k}] := \{k_1, k_2, \ldots, k_p\}$, $\sum^*$ means summation with indices not equal to j, and $c_{\mathbf{k}}$ are deterministic coefficients satisfying $c_{\mathbf{k}} = \mathrm{O}(N^{-p})$. Then with the same arguments as in [5] (more specifically, the ones between (B.21)-(B.24)), we see that to conclude (2.22), it suffices to prove that for $k \in A \subset \mathbb{Z}_N \setminus \{j\}$ and any fixed $\tau > 0$,

$$|Q_A X_k| = \mathrm{O}_\tau \left(\Phi^{|A|+1} \right) \quad w.h.p. \tag{2.23}$$

We first recall the following simple bound for partial expectations, which is proved in Lemma B.1 of [5]. Given a nonnegative random variable X and a deterministic control parameter Ψ such that $X \leqslant \Psi$ with high probability. Suppose $\Psi \geqslant N^{-C}$ and $X \leqslant N^C$ almost surely for some constant $C > 0$. Then for any fixed $\tau > 0$, we have

$$\max_i P_i X = \mathrm{O}_\tau(\Psi) \quad w.h.p. \tag{2.24}$$

In fact, (2.24) follows from Markov's inequality, using high-moments estimates combined with the definition of high probability events in Definition 2.5 and Jensen's inequality for partial expectations. In the application to resolvent entries, the deterministic bound follows from $\|G\| \leqslant (\operatorname{Im} \widetilde{z})^{-1} \leqslant N^{10}$ by (2.9).

Now the bound (2.23) in the case $|A| = 1$ follows from (2.24) directly. For the case $|A| = n \geqslant 2$, we assume without loss of generality that $j = 1$, $k = 2$ and $A = \{2, \dots, n+1\}$. It suffices to prove that

$$Q_{n+1} \cdots Q_3 |G_{21}|^2 = \mathrm{O}_\tau \left(\Phi^{n+1} \right). \tag{2.25}$$

Using the identity (2.3), we can write

$$Q_3 |G_{21}|^2 = Q_3 \left(G_{21}^{(3)} + \frac{G_{23} G_{31}}{G_{33}} \right) \overline{\left(G_{21}^{(3)} + \frac{G_{23} G_{31}}{G_{33}} \right)} = Q_3 \left(\overline{G_{21}^{(3)}} \frac{G_{23} G_{31}}{G_{33}} + G_{21}^{(3)} \overline{\frac{G_{23} G_{31}}{G_{33}}} + \left| \frac{G_{23} G_{31}}{G_{33}} \right|^2 \right).$$

Note that the leading term $Q_3 \left| G_{21}^{(3)} \right|^2$ vanishes since $G_{21}^{(3)}$ is independent of the 3rd row and column of H, and the rest of the three terms have at least three off-diagonal resolvent entries. We now act Q_4 on these terms, apply (2.3) with $k = 4$ to each resolvent entry, and multiply everything out. This gives a sum of fractions, where all the entries in the numerator are off-diagonal and all the entries in the denominator are diagonal. Moreover, the leading order terms vanish,

$$Q_4 Q_3 \left(\overline{G_{21}^{(34)}} \frac{G_{23}^{(4)} G_{31}^{(4)}}{G_{33}^{(4)}} + G_{21}^{(34)} \overline{\frac{G_{23}^{(4)} G_{31}^{(4)}}{G_{33}^{(4)}}} \right) = 0,$$

and each of the surviving term has at least four off-diagonal resolvent entries. We then continue in this manner, and at each step the number of off-diagonal resolvent entries in the numerator increases at least by one. Finally, $Q_{n+1} \cdots Q_3 |G_{kj}|^2$ is a sum of fractions where each of them contains at least $n+1$ off-diagonal entries in the numerator. Together with (2.24), this gives the estimate (2.25), which further proves (2.23). □

Remark 2.13. *Lemma 2.12 asserts that the Q_k operation yields an improvement by a factor Φ. In fact, for the regular resolvents of band matrices, a stronger version of averaging fluctuation results was proved in [3]. We believe that following the methods there, the bounds in Lemma 2.11 and Lemma 2.12 can be improved to*

$$\mathrm{O}_\tau \left(\Phi^4 + W^{-1/2} \Phi^2 \right). \tag{2.26}$$

In this paper, however, we will skip the discussion on the strategy in [3], since its proof is rather involved, and more importantly, we will prove an even stronger bound, i.e., (2.30) *below, in Part III of this series [10]. With* (2.26)*, the $\Phi_{\#}^w$ in* (2.11) *can be improved to*

$$(\Phi_{\#}^w)^2 = \left(\frac{N}{W \operatorname{Im} z} + \frac{N^2}{W^2} \right) (\Phi^4 + W^{-1/2} \Phi^2 + N^{-1}),$$

and the condition (2.12) *becomes*

$$\Phi^2 \leqslant \min \left\{ \frac{W}{N^{1+\varepsilon_* + \varepsilon^*}}, \frac{W^2}{N^{2+\varepsilon^*}} \right\}. \tag{2.27}$$

Using this estimate, the conditions (1.13) *can be weaken to*

$$\log_N W \geqslant \max \left\{ \frac{4}{5} + \varepsilon^*, \ \frac{2}{3} + \frac{2}{3} \varepsilon_* + \varepsilon^* \right\}. \tag{2.28}$$

Now we finish the proof of Lemma 2.8. Using (2.19), Lemma 2.11, Lemma 2.12 and Markov's inequality, we can get that

$$\sum_{k \neq j} \left[(1 - S|M|^2)^{-1} S \right]_{ik} \left(|G_{kj}|^2 - |M_k|^2 T_{kj} \right) = \mathrm{O}_\tau \left(\left(\frac{N}{W \operatorname{Im} z} + \frac{N^2}{W^2} \right) \Phi^3 \right)$$

with high probability. Note that it only includes the off-diagonal terms, i.e. $k \neq j$ terms. Now plugging it into the T-equation (2.15) and using (2.16), we obtain (2.11).

Finally, we need to prove (2.13). Clearly, if (2.12) holds, then $\Phi \leqslant N^{-\delta}$ and $(\Phi_{\#}^w)^2 \leqslant N^{-2\delta}$ for some constant $\delta > 0$. Thus (2.1) is satisfied, and then (2.13) follows from an application of (2.11) and Lemma 2.1. This completes the proof of Lemma 2.8. □

The following lemma gives a stronger form of Lemma 2.8. It will be proved in the companion paper [10]. Here we recall the notation in (1.18).

Lemma 2.14 (Strong T-equation estimate)**.** *Suppose the assumptions of Theorem 1.4 (i.e.,* (1.11)*,* (1.12)*,* (1.15) *and the assumption on e) and* (2.9) *hold. Let* Φ *and* $\widetilde{\Phi}$ *be deterministic parameters satisfying*

$$W^{-1} \leqslant \widetilde{\Phi}^2 \leqslant \Phi^2 \leqslant \widetilde{\Phi} \leqslant N^{-\delta} \tag{2.29}$$

for some constant $\delta > 0$*. Fix some* z *and* $\widetilde{z}$ *(which can depend on* N*). If for any constants* $\tau' > 0$ *and* $D' > 0$*,*

$$\mathbb{P}\Big(\|G(z,\widetilde{z}) - M(z,\widetilde{z})\|_{\max} \geqslant N^{\tau'}\Phi\Big) + \mathbb{P}\left(|||G|||^2(z,\widetilde{z}) \geqslant N^{1+\tau'}\widetilde{\Phi}^2\right) \leqslant N^{-D'},$$

then for any fixed (small) $\tau > 0$ *and (large)* $D > 0$*, we have*

$$\mathbb{P}\left(\|T(z,\widetilde{z})\|_{\max} \geqslant N^{\tau}\Phi_{\#}^2\right) \leqslant N^{-D}, \quad \Phi_{\#}^2 := \left(\frac{N}{W\operatorname{Im} z} + \frac{N^2}{W^2}\right)\left(\Phi^2\widetilde{\Phi}^2 + \Phi^2 N^{-1/2} + N^{-1}\right). \tag{2.30}$$

Furthermore, if the parameter $\widetilde{\Phi}$ *satisfies*

$$\widetilde{\Phi}^2 \leqslant \min\left\{\frac{W}{N^{1+\varepsilon_*+\varepsilon^*}}, \frac{W^2}{N^{2+\varepsilon^*}}\right\}, \tag{2.31}$$

then for any fixed $\tau > 0$ *and* $D > 0$ *we have*

$$\|G(z,\widetilde{z}) - M(z,\widetilde{z})\|_{\max} \leqslant \Phi N^{-\frac{1}{3}\varepsilon^*} + N^{\tau}\left(\frac{1}{\sqrt{W\operatorname{Im} z}} + \frac{N^{1/2}}{W}\right) \tag{2.32}$$

with probability at least $1 - N^{-D}$*.*

The Remark 2.9 also applies to this lemma. Note that (2.13) or (2.32) gives a self-improved bound on $\|G - M\|_{\max}$, which explains how we can improve the estimate on G (from Φ to $\Phi_{\#}$) via T equations. As long as we have an initial estimate such that (2.12) or (2.31) holds, we can then iterate the proof and improve the estimate on G to $\Phi_{\text{goal}} = \left(\frac{1}{\sqrt{W\operatorname{Im} z}} + \frac{N^{1/2}}{W}\right)$ in (1.14).

Proof of Lemma 2.14. See the proof of Theorem 2.7 in part III of this series [10]. □

3 Proof of Theorem 1.4

Fix a parameter $0 < \varepsilon_0 < \varepsilon_*/5$. We define

$$\widetilde{z}_n := \operatorname{Re}\widetilde{z} + \mathrm{i}N^{-n\varepsilon_0}\operatorname{Im} z,$$

so that $\operatorname{Im}\widetilde{z}_{n+1} = N^{-\varepsilon_0}\operatorname{Im}\widetilde{z}_n$. The basic idea in proving Theorem 1.4 is to use mathematical induction on $n \in \mathbb{N}$.

The proofs of the weak form and strong form of Theorem 1.4 are completely parallel. In the following proof, we will only remark on the minor differences between them.

Step 0: The special case with $\widetilde{z} = z$ and $\zeta = 0$, $\mathbf{g} = \mathbf{0}$ (i.e. $G(H,z)$ is the ordinary resolvent of a generalized Wigner matrix) was proved in [5]. The proof given there can be carried over to our case without changes under the assumptions of Theorem 1.4 when $\widetilde{z} = z$ and $\operatorname{Im} z \geqslant W^{-1+\delta}$ for some fixed $\delta > 0$.

This gives that

$$\mathbb{P}\left(\|G(z,z) - M(z,z)\|_{\max} \geqslant \frac{N^{\tau}}{\sqrt{W\operatorname{Im} z}}\right) \leqslant N^{-D},$$

for any fixed $\tau > 0$. This bound is clearly stronger than the one in (1.14).

Step 1: Consider the case $n = 0$, i.e., $G(z, \widetilde{z}_0)$, where we have

$$\operatorname{Re}\widetilde{z}_0 = \operatorname{Re}\widetilde{z}, \quad \operatorname{Im}\widetilde{z}_0 = \operatorname{Im} z.$$

We claim that for any $w, \widetilde{w} \in \mathbb{C}_+$,

$$\|G(w, \widetilde{w})\|_{L^2 \to L^2} \leqslant \frac{1}{\min(\operatorname{Im} w, \operatorname{Im}\widetilde{w})}. \tag{3.1}$$

To prove it, we first assume that $\operatorname{Im} w = a + \operatorname{Im}\tilde{w}$ with $a \geqslant 0$. We write

$$G(w, \widetilde{w}) = (A - \mathrm{i}aJ - \mathrm{i}\operatorname{Im}\tilde{w})^{-1}, \quad J_{kl} = \mathbf{1}_{k \in [\![1, W]\!]}\delta_{kl},$$

where A is a symmetric matrix. Then

$$(A - \mathrm{i}aJ - \mathrm{i}\operatorname{Im}\tilde{w})^*(A - \mathrm{i}aJ - \mathrm{i}\operatorname{Im}\tilde{w}) = (A - \mathrm{i}aJ)^*(A - \mathrm{i}aJ) + 2a(\operatorname{Im}\tilde{w})J + (\operatorname{Im}\tilde{w})^2 \geqslant (\operatorname{Im}\tilde{w})^2.$$

Obviously, we have a similar estimate with $\operatorname{Im}\tilde{w}$ replaced by $\operatorname{Im} w$ when $\operatorname{Im} w \leqslant \operatorname{Im}\tilde{w}$. This proves the claim (3.1).

Now by the definition of T and (1.3), we know

$$|T_{ij}(z, \widetilde{z}_0)| \leqslant \frac{C_s}{W}\sum_k |G_{kj}(z, \widetilde{z}_0)|^2 = \frac{C_s \operatorname{Im} G_{jj}(z, \widetilde{z}_0)}{W \operatorname{Im} z},$$

where in the second step we used the so-called Ward identity that for any symmetric matrix A and $\eta > 0$,

$$\sum_k |R_{kj}(A, \mathrm{i}\eta)|^2 = \frac{\operatorname{Im} R_{jj}(A, \mathrm{i}\eta)}{\eta}, \quad R(A, \mathrm{i}\eta) := (A - \mathrm{i}\eta)^{-1}. \tag{3.2}$$

Obviously, the same argument gives that

$$\|T(z, \widetilde{z}_0(t))\|_{\max} \leqslant \frac{C_s \max_j \operatorname{Im} G_{jj}(z, \widetilde{z}_0(t))}{W \operatorname{Im} z}, \quad \widetilde{z}_0(t) := (1-t)z + t\widetilde{z}_0, \quad t \in [0, 1]. \tag{3.3}$$

Now we claim that for any small enough $\tau > 0$,

$$\sup_{s \in [0,1]} \mathbb{P}\left(\|G(z, \widetilde{z}_0(t)) - M(z, \widetilde{z}_0(t))\|_{\max} \geqslant \frac{N^\tau}{\sqrt{W \operatorname{Im} z}}\right) \leqslant N^{-D}. \tag{3.4}$$

To prove (3.4), we first note that for any $w, w' \in \mathbb{C}$,

$$G(z, w) = G(z, w') + G(z, w)(w - w')\widetilde{J}G(z, w'), \quad \widetilde{J}_{kl} = \mathbf{1}_{k \notin [\![1, W]\!]}\delta_{kl}. \tag{3.5}$$

This implies that

$$\|\partial_{\widetilde{z}} G(z, \widetilde{z})\|_{\max} \leqslant \sqrt{N}\|G(z, \widetilde{z})\|_{L^2 \to L^2}\|G(z, \widetilde{z})\|_{\max} \leqslant \frac{\sqrt{N}}{\min(\operatorname{Im} z, \operatorname{Im}\widetilde{z})}\|G\|_{\max}.$$

In particular, in this step we have

$$\|\partial_s G(z, \widetilde{z}_0(t))\|_{\max} \leqslant CN^{1/2+\varepsilon_*}|z - \widetilde{z}_0|\|G(z, \widetilde{z}_0(t))\|_{\max}. \tag{3.6}$$

This provides some continuity estimate on $G(z, \widetilde{z}_0(t))$, which shows that (3.4) can be obtained from the following estimate:

$$\max_{k \in [\![0, N^5]\!]} \mathbb{P}\left(\left\|G(z, \widetilde{z}_0(kN^{-5})) - M(z, \widetilde{z}_0(kN^{-5}))\right\|_{\max} \geqslant \frac{N^\tau}{\sqrt{W \operatorname{Im} z}}\right) \leqslant N^{-D}. \tag{3.7}$$

From Step 0, this estimate holds for $k = 0$. By induction, we assume that (3.7) holds for $k = k_0$. Then using (3.6) and (1.10), we know that the first estimate of (2.1) holds for $G(z, \widetilde{z}_0(t))$ with $t = (k_0 + 1)N^{-5}$.

Then by (3.3) and applying Lemma 2.1, we obtain (3.7) for $k = k_0 + 1$. This completes the proof of (3.7) and (3.4). Note that the estimate (3.4) applied to $G(z, \widetilde{z}_0(1))$ is the result we want for this step

Step 2: Suppose that for some $n \in \mathbb{N}$ with $\operatorname{Im} \widetilde{z}_n \geqslant N^{-10}$, (1.14) holds for $G(z, \widetilde{z}_n)$ and $M(z, \widetilde{z}_n)$ for any large $D > 0$. We first prove the following estimate for $G(z, \widetilde{z}_{n+1}) - M(z, \widetilde{z}_{n+1})$, which is weaker than (1.14):

$$\mathbb{P}\left(\|G(z, \widetilde{z}_{n+1}) - M(z, \widetilde{z}_{n+1})\|_{\max} \geqslant N^{\tau} \left(\frac{N^{1/2+\varepsilon_0}}{W\sqrt{\operatorname{Im} z}} + \frac{N^{1+\varepsilon_0}}{W^{3/2}} \right) \right) \leqslant N^{-D} \tag{3.8}$$

for any fixed $\tau > 0$.

For any $w, w' \in \mathbb{C}^+$ satisfying

$$\operatorname{Re} w = \operatorname{Re} w', \quad N^{-\varepsilon_0} \operatorname{Im} w' \leqslant \operatorname{Im} w \leqslant \operatorname{Im} w', \tag{3.9}$$

using (3.9) and (3.5), we have

$$\begin{aligned} \sum_i |G_{ij}(z, w)|^2 &\leqslant 2\Big(1 + |w - w'|^2 \, \|G(z, w)\|_{L^2 \to L^2}^2 \Big) \sum_i |G_{ij}(z, w')|^2 \\ &\leqslant 2\Big(1 + \frac{(\operatorname{Im} w')^2}{(\operatorname{Im} w)^2}\Big) \sum_i |G_{ij}(z, w')|^2 \leqslant 3N^{2\varepsilon_0} |||G(z, w')|||^2, \end{aligned}$$

where we have used (3.1) to bound $\|G(z, w)\|_{L^2 \to L^2}^2$. We apply this inequality with $w' = \widetilde{z}_n$ and w satisfying (3.9). Using (1.14) and the definition (1.18), we can bound $|||G(z, \widetilde{z}_n)|||^2$ as

$$\sup_{\substack{\operatorname{Re} w = \operatorname{Re} \widetilde{z}_n, \\ \operatorname{Im} \widetilde{z}_{n+1} \leqslant \operatorname{Im} w \leqslant \operatorname{Im} \widetilde{z}_n}} \|T(z, w)\|_{\max} \leqslant \sup_{\substack{\operatorname{Re} w = \operatorname{Re} \widetilde{z}_n, \\ \operatorname{Im} \widetilde{z}_{n+1} \leqslant \operatorname{Im} w \leqslant \operatorname{Im} \widetilde{z}_n}} \frac{C}{W} |||G(z, w)|||^2 = \mathrm{O}_{\tau}\left(\frac{N^{1+2\varepsilon_0}}{W^2 \operatorname{Im} z} + \frac{N^{2+2\varepsilon_0}}{W^3} \right) \tag{3.10}$$

with high probability for any fixed $\tau > 0$.

We now consider interpolation between $\widetilde{z}_n$ and $\widetilde{z}_{n+1}$:

$$\widetilde{z}_{n,m} = \widetilde{z}_n - \mathrm{i}(\operatorname{Im} \widetilde{z}_n - \operatorname{Im} \widetilde{z}_{n+1}) m N^{-50}, \quad m \in [\![0, N^{50}]\!].$$

We would like to use Lemma 2.1 and induction to prove that (3.8) holds for $G(z, \widetilde{z}_{n,m}) - M(z, \widetilde{z}_{n,m})$ for all m. First, we know (3.8) holds for $G(z, \widetilde{z}_n)$. Then suppose (3.8) holds for $G(z, \widetilde{z}_{n,j})$ for all $j \leqslant m - 1$. We now verify that (2.1) holds for $G(z, \widetilde{z}_{n,m})$ with $\Phi^2 = N^{\tau} \Phi_0^2$ for any fixed $\tau > 0$, where

$$\Phi_0^2 := \frac{N^{1+2\varepsilon_0}}{W^2 \operatorname{Im} z} + \frac{N^{2+2\varepsilon_0}}{W^3}.$$

To this end, we note that (3.10) already verifies the bound on $\|T(z, \widetilde{z}_{n,m})\|_{\max}$ in (2.1) for all $m \in [\![0, N^{50}]\!]$. By using $\|\partial_{\widetilde{z}} G\|_{\max} \leqslant N \|G\|_{\max}^2$ (which follows from (3.5)), (1.10), $|\widetilde{z}_{n,m-1} - \widetilde{z}_{n,m}| \leqslant N^{-50}$, and (3.10) (to bound $\|G\|_{\max}^2$ by $|||G|||^2$), we note that for sufficiently small constant $\delta > 0$,

$$\|G(z, \widetilde{z}_{n,m-1}) - M(z, \widetilde{z}_{n,m-1})\|_{\max} \leqslant N^{-2\delta} \implies \|G(z, \widetilde{z}_{n,m}) - M(z, \widetilde{z}_{n,m})\|_{\max} \leqslant N^{-\delta}.$$

This proves the first bound in (2.1) for $G(z, \widetilde{z}_{n,m})$. Then Lemma 2.1 asserts that (2.2) holds for $G(z, \widetilde{z}_{n,m})$ with $N^{\tau} \Phi_0$ for any fixed $\tau > 0$. This proves (3.8) (i.e. the $m = N^{50}$ case) by induction.

Step 3: Suppose that for some $n \in \mathbb{N}$ with $\operatorname{Im} \widetilde{z}_n \geqslant N^{-10}$, (1.14) holds for $G(z, \widetilde{z}_n)$ and $M(z, \widetilde{z}_n)$ for any large $D > 0$. We have proved that (3.8) and (3.10) hold for $G(z, \widetilde{z}_{n+1})$. We now apply Lemma 2.8 to prove the weak form of Theorem 1.4. First, the condition (2.10) holds with $\Phi = \frac{N^{1/2+\varepsilon_0}}{W\sqrt{\operatorname{Im} z}} + \frac{N^{1+\varepsilon_0}}{W^{3/2}}$. In order for the condition (2.12) to hold, we need

$$\frac{N^{1/2+\varepsilon_0}}{W\sqrt{\operatorname{Im} z}} + \frac{N^{1+\varepsilon_0}}{W^{3/2}} \leqslant \min\left\{ \frac{W}{N^{1+\varepsilon_* + \varepsilon^*}}, \frac{W^2}{N^{2+\varepsilon^*}} \right\}, \tag{3.11}$$

which is satisfied if

$$W \geqslant 2 \max\left(N^{\frac{6}{7} + \frac{2}{7}\varepsilon_0 + \frac{2}{7}\varepsilon^*}, N^{\frac{3}{4} + \frac{3}{4}\varepsilon_* + \frac{1}{2}\varepsilon_0 + \frac{1}{2}\varepsilon^*} \right).$$

If we take $\varepsilon_0 < \varepsilon^*$, (2.10) implies (2.13) under the condition (1.13). We then apply Lemma 2.8 again, and after at most $3/\varepsilon^*$ iterations we obtain that

$$\|G(z,\widetilde{z}_{n+1}) - M(z,\widetilde{z}_{n+1})\|_{\max} \leqslant N^{\tau}\left(\frac{1}{\sqrt{W\operatorname{Im} z}} + \frac{N^{1/2}}{W}\right). \tag{3.12}$$

By induction on n (with the number of inductions $\leqslant 10/\varepsilon_0$), the main estimate (3.12) for $G(z,\widetilde{z}_n)$ holds for all n as long as $\operatorname{Im}\widetilde{z}_n \geqslant N^{-10}$.

Similarly, we can apply Lemma 2.14 to prove the strong form of Theorem 1.4. As in the previous argument, (3.8) and (3.10) hold for $G(z,\widetilde{z}_{n+1})$ assuming (1.14) for $G(z,\widetilde{z}_n)$ and $\operatorname{Im}\widetilde{z}_n \geqslant N^{-10}$. Therefore, we can choose Φ and $\widetilde{\Phi}$ as

$$\Phi = \frac{N^{1/2+\varepsilon_0}}{W\sqrt{\operatorname{Im} z}} + \frac{N^{1+\varepsilon_0}}{W^{3/2}}, \quad \widetilde{\Phi} = \frac{N^{\varepsilon_0}}{\sqrt{W\operatorname{Im} z}} + \frac{N^{1/2+\varepsilon_0}}{W},$$

where the choice of $\widetilde{\Phi}$ follows from using (3.10). It is easy to see that (2.29) holds. In order to apply Lemma 2.14, we need (2.31), i.e.,

$$\left(\frac{N^{\varepsilon_0}}{\sqrt{W\operatorname{Im} z}} + \frac{N^{1/2+\varepsilon_0}}{W}\right)^2 \leqslant \min\left\{\frac{W}{N^{1+\varepsilon_*+\varepsilon^*}}, \frac{W^2}{N^{2+\varepsilon^*}}\right\},$$

which is satisfied if

$$W \geqslant 2\max\left(N^{\frac{3}{4}+\frac{1}{2}\varepsilon_0+\frac{1}{4}\varepsilon^*}, N^{\frac{1}{2}+\varepsilon_*+\varepsilon_0+\frac{1}{2}\varepsilon^*}\right).$$

Clearly, the assumption (1.15) guarantees this condition if we choose $\varepsilon_0 < \varepsilon^*/2$. Again, we can apply Lemma 2.14 iteratively until we get (3.12) for $G(z,\widetilde{z}_{n+1})$. The rest of the proof for the strong form of Theorem 1.4 is the same as the proof for the weak form.

Step 4: We now prove (1.14) for $G(z,\widetilde{z})$ with $\operatorname{Im}\widetilde{z} = 0$ by using continuity from the estimate for $G(z,\widetilde{z})$ with $\operatorname{Im}\widetilde{z} = N^{-10}$ established in **Step 3**. It is easy to see that

$$\partial_{\widetilde{z}}\|G(z,\widetilde{z})\|_{\max} \leqslant \|\partial_{\widetilde{z}}G(z,\widetilde{z})\|_{\max} \leqslant N\|G(z,\widetilde{z})\|_{\max}^2. \tag{3.13}$$

With (3.13) and using (3.12) for $G(z,\operatorname{Re}\widetilde{z} + \mathrm{i}N^{-10})$, we can obtain that

$$\sup_{0\leqslant\eta\leqslant N^{-10}}\|G(z,\operatorname{Re}\widetilde{z}+\mathrm{i}\eta)\|_{\max} = \mathrm{O}(1), \quad w.h.p.$$

Then using (1.10), (3.5) and (3.12) for $G(z,\operatorname{Re}\widetilde{z} + \mathrm{i}N^{-10})$, we obtain that (1.14) holds for $G(z,\operatorname{Re}\widetilde{z})$.

Remark 3.1. *If we use the bound in Remark 2.13 and the condition* (2.27) *instead of* (2.12)*, then the restriction* (3.11) *becomes*

$$\left(\frac{N^{1/2+\varepsilon_0}}{W\sqrt{\operatorname{Im} z}} + \frac{N^{1+\varepsilon_0}}{W^{3/2}}\right)^2 \leqslant \min\left\{\frac{W}{N^{1+\varepsilon_*+\varepsilon^*}}, \frac{W^2}{N^{2+\varepsilon^*}}\right\}$$

which gives restriction in (2.28)*. So we get a result in between the weak and strong forms of Theorem 1.4.*

4 Properties of M

The main goal of this section is to derive some deterministic estimates related to $(M_\zeta^{\mathbf{g}})_i$, $i \in \mathbb{Z}_N$. In particular, we will finish the proof of Lemma 1.3 and Lemma 2.7.

4.1 The stability. The system of self-consistent equations (1.6) is a perturbation of the standard self-consistent equation

$$m_{\mathrm{sc}}^{-1} = -\widetilde{z} - m_{\mathrm{sc}}$$

for $m_{\mathrm{sc}}(\widetilde{z})$. Thus our basic strategy is to use the standard perturbation theory (see (4.13) below) combined with a stability estimate for the self-consistent equation (i.e. the operator bound (4.4)). We first recall the following elementary properties of m_{sc}, which can be proved directly using (1.7).

Lemma 4.1. *We have for all $z = E + \mathrm{i}\eta$ with $\eta > 0$ that*

$$|m_{\mathrm{sc}}(z)| = |m_{\mathrm{sc}}(z) + z|^{-1} \leqslant 1.$$

Furthermore, there is a constant $c > 0$ such that for $E \in [-10, 10]$ and $\eta \in (0, 10]$ we have

$$c \;\leqslant\; |m_{\mathrm{sc}}(z)| \;\leqslant\; 1 - c\eta\,, \tag{4.1}$$

$$|\partial_z m_{\mathrm{sc}}(z)| \;\leqslant\; c^{-1}(\kappa+\eta)^{-1/2}, \tag{4.2}$$

$$|1 - m_{\mathrm{sc}}^2(z)| \;\asymp\; \sqrt{\kappa+\eta}\,,$$

as well as

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

where $\kappa := \big||E| - 2\big|$ denotes the distance of E to the spectral edges.

The following lemma will be used in the proof of Lemma 1.3 and Lemma 2.7. Recall that S_0 is the matrix with entries s_{ij}, which is defined in Definition 1.2.

Lemma 4.2. *Assume $|\operatorname{Re}\widetilde{z}| \leqslant 2 - \kappa$ for some constant $\kappa > 0$ and denote $m = m_{\mathrm{sc}}(\widetilde{z} + \mathrm{i}0^+)$. Then for any fixed $\tau > 0$, there exist constants $c_1, C_1 > 0$ such that*

$$\left\| \left(\frac{m^2 S_0 + \tau}{1+\tau} \right)^2 \right\|_{L^\infty \to L^\infty} < 1 - c_1. \tag{4.3}$$

Furthermore,

$$\left\| (1 - m^2 S_0)^{-1} \right\|_{L^\infty \to L^\infty} \leqslant C_1. \tag{4.4}$$

Proof. For some small constant $\tau > 0$ we write

$$(1 - m^2 S_0)^{-1} = \frac{1}{1+\tau} \sum_{k=0}^{\infty} \left(\frac{m^2 S_0 + \tau}{1+\tau} \right)^k. \tag{4.5}$$

Assuming (4.3), we get that

$$\|(1 - m^2 S_0)^{-1}\|_{L^\infty \to L^\infty} \leqslant \frac{1}{1+\tau} \left(1 + \left\| \frac{m^2 S_0 + \tau}{1+\tau} \right\|_{L^\infty \to L^\infty} \right) \sum_{j=0}^{\infty} \left\| \frac{m^2 S_0 + \tau}{1+\tau} \right\|_{L^\infty \to L^\infty}^{2j} \leqslant C_1,$$

which proves (4.4).

We now prove (4.3). Suppose that there is a vector $\mathbf{v} \in \mathbb{C}^N$ so that $\|\mathbf{v}\|_\infty = 1$ and

$$\left| \frac{[(m^2 S_0 + \tau)^2 \mathbf{v}]_i}{(1+\tau)^2} \right| = 1 - \varepsilon$$

for some $i \in \mathbb{Z}_N$ and $\varepsilon \equiv \varepsilon_N \to 0^+$. Hence

$$(1 + 2\tau + \tau^2)(1 - \varepsilon) = \left| m^4 b + 2\tau m^2 a + \tau^2 v_i \right| \leqslant |b| + 2\tau|a| + \tau^2 |v_i| \leqslant 1 + 2\tau + \tau^2, \tag{4.6}$$

where $a := (S_0 \mathbf{v})_i$, $b := (S_0^2 \mathbf{v})_i$ and we have used the bounds $|m| \leqslant 1$, $|a| \leqslant 1$ and $|b| \leqslant 1$ (since $\|S_0\|_{L^\infty \to L^\infty} = 1$). It will be clear that the $|m| = 1$ case is most difficult and we will assume this condition in the following proof. Moreover, we assume with loss of generality that $v_i > 0$ (by changing the global phase of $\mathbf{v}$). Now m,

a and b are complex numbers, and the inequality (4.6) implies that $m^4 b$, $m^2 a$ and v_i have almost the same phases. Since $|v_i| \leqslant 1$, $|b| \leqslant 1$ and $|a| \leqslant 1$, (4.6) implies that for some constant $C > 0$ independent of ε,

$$v_i \geqslant 1 - C\varepsilon, \quad |b - m^{-4}| \leqslant C\varepsilon, \quad |a - m^{-2}| \leqslant C\varepsilon. \tag{4.7}$$

Since m is a unit modulus complex number with imaginary part of order 1, we have that $\delta := |m^{-2} - m^{-4}|$ is a number of order 1 and

$$|a - b| > \delta/2.$$

Fix the index i and denote $c_j := (S_0)_{ij}$, $d_j := (S_0^2)_{ij}$. Then $\sum_j c_j = 1 = \sum_j d_j$. Hence (4.7) implies

$$1 - \mathrm{O}(\varepsilon) = \mathrm{Re}(a\bar{a}) = \sum_j c_j \,\mathrm{Re}(v_j \bar{a}), \quad 1 - \mathrm{O}(\varepsilon) = \mathrm{Re}(b\bar{b}) = \sum_j d_j \,\mathrm{Re}(v_j \bar{b}),$$

where $\mathrm{O}(\varepsilon)$ denote a positive number bounded by $C\varepsilon$ for some constant $C > 0$ independent of ε. For any $0 < r < 1$, denote by $A_r := \{j : \mathrm{Re}(v_j \bar{a}) \geqslant 1 - r\}$ and let $\alpha_r := \sum_{j \in A_r} c_j$. Then we have

$$\alpha_r \geqslant \sum_{j \in A_r} c_j \,\mathrm{Re}(v_j \bar{a}) = 1 - \mathrm{O}(\varepsilon) - \sum_{j \notin A_r} c_j \,\mathrm{Re}(v_j \bar{a}) \geqslant 1 - \mathrm{O}(\varepsilon) - (1 - \alpha_r)(1 - r) = \alpha_r + r - \alpha_r r - \mathrm{O}(\varepsilon),$$

which implies that

$$\sum_{j \in A_r} c_j = \alpha_r \geqslant 1 - \mathrm{O}(\varepsilon) r^{-1}, \quad \sum_{j \notin A_r} c_j = \mathrm{O}(\varepsilon) r^{-1}. \tag{4.8}$$

Similarly, if we define $B_r := \{j : \mathrm{Re}(v_j \bar{b}) \geqslant 1 - r\}$, then

$$\sum_{j \notin B_r} d_j = \mathrm{O}(\varepsilon) r^{-1}. \tag{4.9}$$

We claim that if $r \geqslant C\varepsilon$ for some large enough constant $C > 0$, then $A_r \cap B_r \neq \emptyset$. To see this, we define $U := \{j : |i - j| \leqslant W\}$. By (1.3) and the definition of c_j, we have $c_j \geqslant c_s W^{-1}$ for $j \in U$. Clearly, we also have $d_j \geqslant \frac{1}{2} c_s W^{-1}$ for $j \in U$. Then with (4.8) and (4.9), we have

$$\#\{j \in U \setminus A_r\} = \mathrm{O}(\varepsilon) r^{-1} c_s^{-1} W, \quad \#\{j \in U \setminus B_r\} = \mathrm{O}(\varepsilon) r^{-1} c_s^{-1} W.$$

If we choose $r = C\varepsilon$ for some large enough constant $C > 0$, then the above two inequalities imply $A_r \cap B_r \neq \emptyset$, since $|U| = W$. Thus there is an index j such that

$$\mathrm{Re}(v_j \bar{a}) \geqslant 1 - r, \quad \mathrm{Re}(v_j \bar{b}) \geqslant 1 - r. \tag{4.10}$$

Since $|a| \leqslant 1$, $|b| \leqslant 1$, $|v_j| \leqslant 1$ and $|a - b| > \delta/2$, (4.10) is possible only if $r \gtrsim \delta$, which contradicts the fact that $r \to 0$ when $\varepsilon \to 0$. This proves (4.3). □

4.2 Proof of Lemma 1.3. With Lemma 4.2, we can now give the proof of Lemma 1.3.

Proof of Lemma 1.3. We first prove the existence and continuity of the solutions to (1.6). The proof is a standard application of the contraction principle. Denote by $\mathbf{z} := (z_1, \ldots, z_N)$, $\mathbf{x} := (x_1, \ldots, x_N)$ and $\mathbf{M} := ((M_\zeta^{\mathbf{g}})_1, \ldots, (M_\zeta^{\mathbf{g}})_N)$ with

$$z_i = z \mathbf{1}_{i \in [\![1, W]\!]} + \widetilde{z} \mathbf{1}_{i \notin [\![1, W]\!]},$$

and

$$x_i \equiv (x_\zeta^{\mathbf{g}})_i(z, \widetilde{z}) := (M_\zeta^{\mathbf{g}})_i(z, \widetilde{z}) - m, \quad \mathbf{M} = \mathbf{x} + m\mathbf{e}_1, \quad m := m_{\mathrm{sc}}(\widetilde{z} + \mathrm{i}0^+), \quad \mathbf{e}_1 = (1, 1, \cdots, 1). \tag{4.11}$$

Using the above notations and recalling Definition 1.2, we can rewrite (1.6) into the following form

$$(m + x_i)^{-1} = M_i^{-1} = -z_i - g_i - (S_0 M)_i + \zeta(\Sigma M)_i = -z_i - g_i - (S_0 \mathbf{x})_i - m(S_0 \mathbf{e}_1)_i + \zeta(\Sigma \mathbf{x})_i + \zeta m(\Sigma \mathbf{e}_1)_i. \tag{4.12}$$

Subtracting $m^{-1} = -\widetilde{z} - m$ from the last equation and using $S_0 \mathbf{e}_1 = \mathbf{e}_1$, we get that

$$m^{-1} - (m + x_i)^{-1} = g_i + (z_i - \widetilde{z}) + (S_0 \mathbf{x}) - \zeta m(\Sigma \mathbf{e}_1)_i - \zeta(\Sigma \mathbf{x})_i.$$

Then (4.12) is equivalent to

$$[(1-m^2S_0)\mathbf{x}]_i = m^2(g_i+(z_i-\widetilde{z}))+m^2\left(\frac{1}{m+x_i}-\frac{1}{m}+\frac{x_i}{m^2}\right)-\zeta m^3(\Sigma\mathbf{e}_1)_i-\zeta m^2(\Sigma\mathbf{x})_i. \tag{4.13}$$

Define iteratively a sequence of vectors $\mathbf{x}^k\in\mathbb{C}^N$ such that $\mathbf{x}^0=\mathbf{0}\in\mathbb{C}^N$ and

$$\left[(1-m^2S_0)\mathbf{x}^{k+1}\right]_i := m^2(g_i+(z_i-\widetilde{z}))+m^2\left(\frac{1}{m+(\mathbf{x}^k)_i}-\frac{1}{m}+\frac{(\mathbf{x}^k)_i}{m^2}\right)-\zeta m^3(\Sigma\mathbf{e}_1)_i-\zeta m^2(\Sigma\mathbf{x}^k)_i. \tag{4.14}$$

In other words, (4.14) defines a mapping $h: l^\infty(\mathbb{Z}_N)\to l^\infty(\mathbb{Z}_N)$:

$$\mathbf{x}^{k+1}=h(\mathbf{x}^k),\quad h_i(\mathbf{x}):=\sum_j(1-m^2S_0)^{-1}_{ij}\left[m^2(g_j+(z_j-\widetilde{z}))+q(x_j)-\zeta m^3(\Sigma\mathbf{e}_1)_j-\zeta m^2(\Sigma\mathbf{x}^k)_j\right], \tag{4.15}$$

where

$$q(x):=m^2\left(\frac{1}{m+x}+\frac{x}{m^2}-\frac{1}{m}\right)=\frac{x^2}{m+x}.$$

Note by the assumptions of Lemma 1.3, $c_\kappa\leqslant m\leqslant 1$ for some constant $c_\kappa>0$ depending only on κ. Then with (4.4), it is easy to see that there exists a sufficiently small constant $0<\alpha<c_\kappa/2$, such that h is a self-mapping

$$h: B_r\left(l^\infty(\mathbb{Z}_N)\right)\to B_r\left(l^\infty(\mathbb{Z}_N)\right),\quad B_r\left(l^\infty(\mathbb{Z}_N)\right):=\{\mathbf{x}\in l^\infty(\mathbb{Z}_N):\|\mathbf{x}\|_\infty\leqslant r\},$$

as long as $r\leqslant\alpha$ and

$$\zeta+\|\mathbf{g}\|_\infty+|z-\widetilde{z}|\leqslant c_r \tag{4.16}$$

for some constant $c_r>0$ depending on r. Now it suffices to prove that h restricted to $B_r\left(l^\infty(\mathbb{Z}_N)\right)$ is a contraction, which then implies that $\mathbf{x}:=\lim_{k\to\infty}\mathbf{x}^k$ exists and is a unique solution to (4.13) subject to the condition $\|\mathbf{x}\|_\infty\leqslant r$.

From the iteration relation (4.15), we obtain that

$$\mathbf{x}^{k+1}-\mathbf{x}^k=\frac{1}{1-m^2S_0}\left[q(\mathbf{x}^k)-q(\mathbf{x}^{k-1})\right]-\frac{\zeta m^2}{1-m^2S_0}\Sigma(\mathbf{x}^k-\mathbf{x}^{k-1}), \tag{4.17}$$

where $q(\mathbf{x})$ denotes a vector with components $q(x_i)$. Using $|q'(0)|=0$ and (4.4), we get from (4.17) that

$$\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_\infty\leqslant C_\kappa\left(\zeta+\|\mathbf{x}^k\|_\infty+\|\mathbf{x}^{k-1}\|_\infty\right)\cdot\|\mathbf{x}^k-\mathbf{x}^{k-1}\|_\infty$$

for some constant $C_\kappa>0$ depending only on κ. Thus we can first choose a sufficiently small constant $0<r<\alpha$ and then the constant $c_r>0$ such that $C_\kappa(c_r+2r)<1$, and h is a self-mapping on $B_r\left(l^\infty(\mathbb{Z}_N)\right)$ under the condition (4.16). In other words, h is indeed a contraction, which proves the existence and uniqueness of the solution.

Note that with (4.4) and $\mathbf{x}^0=\mathbf{0}$, we get from (4.15) that

$$\|\mathbf{x}^1\|_\infty=\mathrm{O}\left(|z-\widetilde{z}|+\zeta+\|\mathbf{g}\|_\infty\right).$$

With the contraction mapping, we have the bound

$$\|\mathbf{x}\|_\infty\leqslant\sum_{k=0}^\infty\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_\infty\leqslant\frac{\|\mathbf{x}^1\|_\infty}{1-C_\kappa(\zeta+2r)}=\mathrm{O}\left(|z-\widetilde{z}|+\zeta+\|\mathbf{g}\|_\infty\right).$$

This gives the bound (1.9).

We now prove (1.10). We have proved above that both $(M^{\mathbf{g}}_\zeta)_i(z,\widetilde{z})$ and $(M^{\mathbf{g}'}_{\zeta'})_i(z',\widetilde{z}\,')$ exist and satisfy (1.9). Denote by $m':=m_{\mathrm{sc}}(\widetilde{z}\,'+\mathrm{i}0^+)$ and $x'_i:=(M^{\mathbf{g}'}_{\zeta'})_i(z',\widetilde{z}\,')-m'$. By (4.2), we have

$$|m'-m|=\mathrm{O}(|\widetilde{z}-\widetilde{z}\,'|). \tag{4.18}$$

Then using (4.13) we can obtain that

$$\begin{aligned}\|\mathbf{x}'-\mathbf{x}\|_\infty \leqslant & C\|(1-m^2S_0)^{-1}\|_{L^\infty\to L^\infty}\cdot\big\{|\widetilde{z}-\widetilde{z}'|\cdot\big[\|\mathbf{x}'\|_\infty+\|\mathbf{g}'\|_\infty+|z'-\widetilde{z}'|+\|\mathbf{x}'\|_\infty^2+\zeta'(1+\|\mathbf{x}'\|_\infty)\big]\\ &+\big[\|\mathbf{g}-\mathbf{g}'\|_\infty+|z-z'|+|\widetilde{z}-\widetilde{z}'|+|\zeta-\zeta'|(1+\|\mathbf{x}'\|_\infty)+(\zeta+\|\mathbf{x}\|_\infty+\|\mathbf{x}'\|_\infty)\cdot\|\mathbf{x}'-\mathbf{x}\|_\infty\big]\big\}\\ \leqslant & C\left(\zeta+\|\mathbf{x}\|_\infty+\|\mathbf{x}'\|_\infty\right)\cdot\|\mathbf{x}'-\mathbf{x}\|_\infty+C\left(\|\mathbf{g}-\mathbf{g}'\|_\infty+|z-z'|+|\widetilde{z}-\widetilde{z}'|+|\zeta-\zeta'|\right).\end{aligned}$$

Applying (1.9) to both $(M_\zeta^{\mathbf{g}})_i(z,\widetilde{z})$ and $(M_{\zeta'}^{\mathbf{g}'})_i(z',\widetilde{z}')$, we see that for small enough c,

$$\|\mathbf{x}'-\mathbf{x}\|_\infty\leqslant C\left(\|\mathbf{g}-\mathbf{g}'\|_\infty+|z-z'|+|\widetilde{z}-\widetilde{z}'|+|\zeta-\zeta'|\right).$$

Together with (4.18), we obtain (1.10) as desired. □

4.3 Proof of Lemma 2.7. To prove Lemma 2.7, it suffices to prove the result for the case $\mathbf{g}=0$, and we will describe how to relax to the condition $\mathbf{g}=\mathrm{O}(W^{-3/4})$ by using the Lipschitz continuity estimate (1.10) at the end of the proof. In preparation for the proof, we first prove the following lemma.

Lemma 4.3. *Suppose that* $\mathbf{g}=0$ *and the assumptions* (1.11), (1.12) *and* (1.15) *hold. Then there exist constants* $c>0$ *and* $C>0$ *such that*

$$\left||(M_\zeta^{\mathbf{0}})_n|^2-|m|^2\right|\leqslant C\left(|z-\widetilde{z}|+\zeta\right)e^{-c\frac{|n|}{W}},\quad n\in\mathbb{Z}_N,\tag{4.19}$$

and

$$\frac{1}{W}\sum_{n\in\mathbb{Z}_N}(|m|^2|(M_\zeta^{\mathbf{0}})_n|^{-2}-1)\geqslant c(\operatorname{Im}z-\operatorname{Im}\widetilde{z})-\zeta+\mathrm{O}\left(N^{-\frac{3}{2}\varepsilon_*}+N^{-\varepsilon^*}\operatorname{Im}\widetilde{z}\right),\tag{4.20}$$

where $m:=m_{\mathrm{sc}}(\widetilde{z}+\mathrm{i}0^+)$.

Proof of Lemma 4.3. First with (4.5) and the fact that $(S_0)_{ij}=0$ if $|i-j|\geqslant C_sW$, we get that

$$[(1-m^2S_0)^{-1}]_{ij}-\delta_{ij}=[m^2(1-m^2S_0)^{-1}S_0]_{ij}=\mathrm{O}(W^{-1})\sum_{k\geqslant\frac{|i-j|}{C_sW}}\left\|\frac{m^2S_0+\tau}{1+\tau}\right\|_{L^\infty\to L^\infty}^k$$

Therefore with (4.3), we obtain immediately that

$$\left|[(1-m^2S_0)^{-1}]_{ij}-\delta_{ij}\right|\leqslant CW^{-1}e^{-c\frac{|i-j|}{W}}\tag{4.21}$$

for some constants $c,C>0$. As in the proof of Lemma 1.3, with $\mathbf{x}^k$ defined in (4.14), we know that

$$x_n=M_n-m=x_n^1+\sum_{k\geqslant1}(x_n^{k+1}-x_n^k),\quad M_n:=(M_\zeta^{\mathbf{0}})_n.\tag{4.22}$$

(Recall that we have proved that $x_n=\lim_{k\to\infty}x_n^k$ in the proof of Lemma 1.3 above.) In particular, according to (4.14), $\mathbf{x}^1$ is given by

$$[(1-m^2S_0)\mathbf{x}^1]_i=m^2(z_i-\widetilde{z})-\zeta m^3(\Sigma\mathbf{e}_1)_i.\tag{4.23}$$

Then with (4.21) and (4.23), one can show that

$$|x_n^1|\leqslant Ce^{-c\frac{|n|}{W}}\left(|z-\widetilde{z}|+\zeta\right),\quad n\in\mathbb{Z}_N.\tag{4.24}$$

By (4.17) and (4.21), we have

$$|x_i^{k+1}-x_i^k|\leqslant C\sum_j\left(W^{-1}e^{-c\frac{|i-j|}{W}}+\delta_{ij}\right)\left[\left(|x_j^k|+|x_j^{k-1}|\right)|x_j^k-x_j^{k-1}|+\zeta\mathbb{1}_{j\in[\![1,W]\!]}\max_{j'\in[\![1,W]\!]}|x_{j'}^k-x_{j'}^{k-1}|\right].$$

By induction, it is easy to prove that there are constants $c,C>0$ such that

$$|x_n^{k+1}-x_n^k|\leqslant Ce^{-c\frac{|n|}{W}}\left(|z-\widetilde{z}|+\zeta\right)^{k+1}.\tag{4.25}$$

Together with (4.24) and (4.22), this implies

$$|x_n| = |M_n - m| \leqslant C\left(|z-\widetilde{z}\,| + \zeta\right) e^{-c\frac{|n|}{W}}, \quad n \in \mathbb{Z}_N. \tag{4.26}$$

This proves (4.19) since $\big||M_n|^2 - |m|^2\big| \leqslant |M_n^2 - m^2|$.

We now prove (4.20). Using (4.19), we have

$$\frac{1}{W}\sum_{n\in\mathbb{Z}_N}\left(|m|^2|M_n|^{-2}-1\right) = \frac{1}{W|m|^2}\sum_{n\in\mathbb{Z}_N}\left(|m|^2-|M_n|^2\right) + \mathrm{O}\left(|z-\widetilde{z}|^2+\zeta^2\right). \tag{4.27}$$

By definition (4.11),

$$|M_n|^2 = |m|^2 + 2\,\mathrm{Re}(\bar{m}x_n) + |x_n|^2.$$

Then with (4.26) we get that

$$\frac{1}{W}\sum_n\left(|M_n|^2-|m|^2\right) = \frac{1}{W}\sum_n\left[2\,\mathrm{Re}(\bar{m}x_n)+|x_n|^2\right] = \frac{2}{W}\sum_n \mathrm{Re}(\bar{m}x_n) + \mathrm{O}(|z-\widetilde{z}|^2+\zeta^2).$$

By (1.11) and (1.12), we have

$$\zeta^2 + |\,\mathrm{Re}(z-\widetilde{z})|^2 \leqslant T^2 + r^2 \leqslant N^{-3\varepsilon_*/2}, \quad 0 \leqslant \mathrm{Im}\,\widetilde{z} \leqslant \mathrm{Im}\,z \leqslant N^{-\varepsilon^*},$$

which implies that

$$\zeta^2 + |z-\widetilde{z}\,|^2 \leqslant \zeta^2 + |\,\mathrm{Re}(z-\widetilde{z})|^2 + \mathrm{Im}(z-\widetilde{z})^2 \leqslant N^{-3\varepsilon_*/2} + N^{-\varepsilon^*}\,\mathrm{Im}(z-\widetilde{z}).$$

Then using (4.22) and (4.25), we obtain that

$$\begin{aligned}\frac{1}{W}\sum_n\left(|M_n|^2-|m|^2\right) &= \frac{2}{W}\sum_n \mathrm{Re}(\bar{m}x_n) + \mathrm{O}\left(N^{-\frac{3}{2}\varepsilon_*} + N^{-\varepsilon^*}\,\mathrm{Im}(z-\widetilde{z})\right)\\ &= \frac{2}{W}\sum_n \mathrm{Re}\left(\bar{m}x_n^1\right) + \mathrm{O}\left(N^{-\frac{3}{2}\varepsilon_*} + N^{-\varepsilon^*}\,\mathrm{Im}(z-\widetilde{z})\right).\end{aligned} \tag{4.28}$$

Summing (4.23) over i, we get that (recall that we take $\mathbf{g}=0$)

$$(1-m^2)\sum_i x_i^1 := m^2\sum_i(z_i-\widetilde{z}\,) - \zeta m^3\,(W+1) = m^2W(z-\widetilde{z}) - \zeta m^3 W + \mathrm{O}(1),$$

where we used that $\sum_i (S_0)_{ij} = 1$ and $(\Sigma\mathbf{e}_1)_i = 1 + W^{-1}$ for $i \in [\![1, W]\!]$. Thus for the second term in the second line of (4.28), we have

$$\begin{aligned}\sum_n \mathrm{Re}(\bar{m}x_n^1) &= |m|^2 W\,\mathrm{Re}\left(\frac{(z-\widetilde{z}\,)m - \zeta m^2}{1-m^2}\right) + \mathrm{O}(1)\\ &= |m^2|W\left(\frac{\zeta}{2} - \frac{\mathrm{Im}\,z - \mathrm{Im}\,\widetilde{z}}{\sqrt{4-|\,\mathrm{Re}\,\widetilde{z}|^{\,2}}} + \mathrm{O}\left(N^{-\varepsilon^*}\,\mathrm{Im}\,\widetilde{z}\right)\right) + \mathrm{O}(1),\end{aligned} \tag{4.29}$$

where we have used the following special properties of $m(\widetilde{z}+\mathrm{i}0^+)$ when $\widetilde{z}$ is a real number, in which case $m(\widetilde{z}+\mathrm{i}0^+)$ has unit modulus:

$$\mathrm{Re}\,\frac{m(a^+)}{1-m^2(a^+)} = 0, \quad \mathrm{Im}\,\frac{m(a^+)}{1-m^2(a^+)} = \frac{1}{\sqrt{4-a^2}}, \quad \mathrm{Re}\,\frac{m^2(a^+)}{1-m^2(a^+)} = -\frac{1}{2}, \quad |a|<2, \quad a^+ := a + \mathrm{i}0^+. \tag{4.30}$$

Here the error $\mathrm{O}\left(N^{-\varepsilon^*}\,\mathrm{Im}\,\widetilde{z}\right)$ in (4.29) is due to $|m(\widetilde{z}) - m(\mathrm{Re}\,\widetilde{z}+\mathrm{i}0^+)| \leqslant C\,\mathrm{Im}\,\widetilde{z}$. Inserting (4.29) into (4.28), we obtain that for some constant $c>0$,

$$\frac{1}{W}\sum_n\left(|M_n|^2-|m|^2\right) \leqslant -c(\mathrm{Im}\,z - \mathrm{Im}\,\widetilde{z}) + \zeta|m|^2 + \mathrm{O}\left(N^{-\frac{3}{2}\varepsilon_*} + N^{-\varepsilon^*}\,\mathrm{Im}\,\widetilde{z}\right),$$

which, together with (4.27), proves (4.20). □

With Lemma 4.3, we now finish the proof of Lemma 2.7.

Proof of Lemma 2.7. We first assume that $\mathbf{g} = 0$. With (4.3) and a perturbation argument, we can show that

$$\left\| \left(\frac{M^2 S + \tau}{1+\tau} \right)^2 \right\|_{L^\infty \to L^\infty} < 1 - c$$

for some constant $c > 0$. Then (2.6) can be proved as in (4.21). Our main task is to prove (2.7). Assume that

$$(1 - |M|^2 S)\mathbf{u}^0 = \mathbf{v}^0 \tag{4.31}$$

for some vectors $\mathbf{u}^0, \mathbf{v}^0 \in \mathbb{R}^N$. Multiplying (4.31) with $\mathbf{u}^0 |M|^{-2}$ from the left and using the definition of S, we obtain that

$$\sum_i (|M_i|^{-2} - 1)|\mathbf{u}_i^0|^2 + \sum_{1 \leqslant i \leqslant W} \zeta(1 + W^{-1})|\mathbf{u}_i^0|^2 + \frac{1}{2}\sum_{i,j} S_{ij}\left(\mathbf{u}_i^0 - \mathbf{u}_j^0\right)^2 = (\mathbf{u}^0, |M|^{-2}\mathbf{v}^0). \tag{4.32}$$

We define a symmetric operator $H : L^2(\mathbb{T}) \mapsto L^2(\mathbb{T})$, where $\mathbb{T} := [\![-(\log N)^4 W, (\log N)^4 W]\!]$ and

$$H := H_0 + H_1,$$

with

$$H_0 : \quad (\mathbf{u}, H_0 \mathbf{v}) = \frac{1}{4}\sum_{i,j \in \mathbb{T}} S_{ij}\left(\mathbf{u}_i - \mathbf{u}_j\right)\left(\mathbf{v}_i - \mathbf{v}_j\right), \quad \mathbf{u}, \mathbf{v} \in L^2(\mathbb{T}),$$

and

$$H_1 : \quad (H_1)_{ij} := \delta_{ij}\left[(|M_i|^{-2} - 1) + \zeta\, \mathbf{1}_{1 \leqslant i \leqslant W}(1 + W^{-1})\right].$$

For any vector $\mathbf{u}$, we denote by $\mathbf{u}|_{\mathbb{T}}$ the restriction of $\mathbf{u}$ to $L^2(\mathbb{T})$. Then with (4.19) and the fact that $|m| \leqslant 1$, we can rewrite (4.32) as

$$(\mathbf{u}^0|_{\mathbb{T}}, H\mathbf{u}^0|_{\mathbb{T}}) + \frac{1}{4}\sum_{i,j} S_{ij}\left(\mathbf{u}_i^0 - \mathbf{u}_j^0\right)^2 \leqslant (\mathbf{u}^0, |M|^{-2}\mathbf{v}^0) + \mathrm{O}(N^{-10})\|\mathbf{u}^0\|_2^2. \tag{4.33}$$

First we claim that

$$H \geqslant c \operatorname{Im} z (\log N)^{-4} \tag{4.34}$$

for some constant $c > 0$. With Temple's inequality, we have the following estimate on the ground state energy of H:

$$H \geqslant E_0(H) \;\geqslant\; \langle H \rangle_\phi - \frac{\langle (H)^2 \rangle_\phi - \langle H \rangle_\phi^2}{E_1(H) - \langle H \rangle_\phi}, \tag{4.35}$$

for any $\phi \in L^2(\mathbb{T})$ such that $\|\phi\|_2 = 1$ and $\langle H \rangle_\phi < E_1(H)$, where $E_0(H)$ and $E_1(H)$ are the lowest two eigenvalues of H. Applying min-max principle to $H \geqslant H_0 - \|H_1\|_{L^2 \to L^2}$, we obtain that

$$E_1(H) \;\geqslant\; E_1(H_0) - \|H_1\|_{L^2 \to L^2}. \tag{4.36}$$

By (4.19), we have $\|H_1\|_{L^2 \to L^2} = \mathrm{O}\left(|z - \widetilde{z}\,| + \zeta + \operatorname{Im} \widetilde{z}\right)$. We then claim that

$$E_1(H_0) \geqslant c(\log N)^{-13} \tag{4.37}$$

for some constant $c > 0$. Recall that $S \equiv S^\zeta = S_0 - \zeta\Sigma$ with

$$\frac{1}{4}\sum_{i,j \in \mathbb{T}} \Sigma_{ij}\left(\mathbf{u}_i - \mathbf{u}_j\right)^2 \leqslant 1, \quad \forall \mathbf{u} \in L^2(\mathbb{T}), \quad \|\mathbf{u}\|_2 = 1.$$

Then again by min-max principle, it suffices to prove the following lemma.

Lemma 4.4. *For s_{ij} satisfying* (1.1)-(1.3)*, there exists a constant $c > 0$ such that*

$$\frac{1}{4}\sum_{i,j \in \mathbb{T}} s_{ij}\left(\mathbf{u}_i - \mathbf{u}_j\right)^2 \geqslant c(\log N)^{-13}, \quad \forall \mathbf{u} \in L^2(\mathbb{T}), \quad \|\mathbf{u}\|_2 = 1, \quad \mathbf{u} \perp (1, 1, \cdots, 1).$$

We postpone its proof until we finish the proof of Lemma 2.7. We now choose the trial state ϕ as a constant vector in (4.35), i.e.,

$$\phi_0 = \frac{1}{\sqrt{|\mathbb{T}|}}(1, 1, \cdots, 1).$$

Then by definition, $H_0\phi_0 = \mathbf{0}$ and $\langle H\rangle_{\phi_0} \leqslant \|H_1\|_{L^2\to L^2} \ll E_1(H)$ by (4.36) and (4.37). Then by (4.35) and (4.36), we have

$$H \geqslant \langle H_1\rangle_{\phi_0} - \frac{\|H_1\|^2_{L^2\to L^2}}{E_1(H) - \langle H_1\rangle_{\phi_0}} \geqslant \langle H_1\rangle_{\phi_0} - \frac{\|H_1\|^2_{L^2\to L^2}}{E_1(H_0) - 2\|H_1\|_{L^2\to L^2}}. \tag{4.38}$$

By the definition of H_1, we have

$$\begin{aligned}
\langle H_1\rangle_{\phi_0} &= \frac{1}{|\mathbb{T}|}\sum_{n\in\mathbb{T}}\left[(|M_n|^{-2} - 1) + \zeta\,\mathbf{1}_{n\in[\![1,W]\!]}(1 + W^{-1})\right] \\
&= \frac{1}{|\mathbb{T}|}\sum_{n\in\mathbb{T}}(1 - |m|^2)|M_n|^{-2} + \frac{1}{|\mathbb{T}|}\sum_{n\in\mathbb{T}}(|m|^2|M_n|^{-2} - 1) + \frac{\zeta(W+1)}{|\mathbb{T}|} \\
&\geqslant c\,\mathrm{Im}\,\widetilde{z} + \mathrm{O}(N^{-10}) + \frac{1}{|\mathbb{T}|}\sum_{n\in\mathbb{Z}_N}(|m|^2|M_n|^{-2} - 1) + \frac{\zeta(W+1)}{|\mathbb{T}|} \\
&\geqslant c\,\mathrm{Im}\,z(\log N)^{-4} + \mathrm{O}\left(N^{-\frac{3}{2}\varepsilon_*} + N^{-\varepsilon^*}\,\mathrm{Im}\,z\right),
\end{aligned}$$

where we used (4.19) and $|m|^2 \leqslant 1 - c\,\mathrm{Im}\,\widetilde{z}$ (by (4.1)) in the third step, and (4.20) in the last step. Together with (4.38), $\|H_1\|^2_{L^2\to L^2} = \mathrm{O}(N^{-\frac{3}{2}\varepsilon_*} + N^{-\varepsilon^*}\,\mathrm{Im}\,z)$ and (4.37), this proves (4.34).

With (4.34), (4.33) gives that for some $c > 0$,

$$c\,\mathrm{Im}\,z(\log N)^{-4}\sum_{i\in\mathbb{T}}|\mathbf{u}^0_i|^2 + \frac{1}{4}\sum_{i,j}S_{ij}\left(\mathbf{u}^0_i - \mathbf{u}^0_j\right)^2 \leqslant (\mathbf{u}^0, |M|^{-2}\mathbf{v}^0) + \mathrm{O}(N^{-10})\|\mathbf{u}^0\|_2^2.$$

Now for some fixed $i_0 \in \mathbb{Z}_N$, we choose $\mathbf{v}^0 = S\mathbf{e}_{i_0}$. Then the above inequality becomes

$$c\,\mathrm{Im}\,z(\log N)^{-4}\sum_{i\in\mathbb{T}}|\mathbf{u}^0_i|^2 + \frac{1}{4}\sum_{i,j}S_{ij}\left(\mathbf{u}^0_i - \mathbf{u}^0_j\right)^2 \leqslant (S|M|^{-2}\mathbf{u}^0)_{i_0} + \mathrm{O}(N^{-10})\|\mathbf{u}^0\|_2^2. \tag{4.39}$$

In the following, we suppose $\|\mathbf{u}^0\|_\infty \gg W^{-1}$, otherwise the proof is done. Since for any $i \in \mathbb{Z}_N$,

$$(\mathbf{u}^0 - |M|^2S\mathbf{u}^0)_i = (S\mathbf{e}_{i_0})_i = \mathrm{O}(W^{-1}), \tag{4.40}$$

we must have

$$\|\mathbf{u}^0\|_\infty \asymp \|S\mathbf{u}^0\|_\infty.$$

Now we decompose $\mathbf{u}^0$ as follows:

$$\mathbf{u}^0_i = u + \widetilde{\mathbf{u}}_i, \quad \text{with } u = \frac{1}{N}\sum_{i\in\mathbb{Z}_N}\mathbf{u}^0_i, \quad \sum_i \widetilde{\mathbf{u}}_i = 0.$$

Suppose $|u| \geqslant 10\|\widetilde{\mathbf{u}}\|_\infty$, then we have

$$\max_i|\mathbf{u}^0_i| \leqslant 2\min_i|\mathbf{u}^0_i|.$$

Together with (4.39), it implies that if $|u| \geqslant 10\|\widetilde{\mathbf{u}}\|_\infty$, then

$$\|\mathbf{u}^0\|_\infty \leqslant 2|u| \leqslant C(W\mathrm{Im}\,z)^{-1}. \tag{4.41}$$

On the other hand, if $|u| \leqslant 10\|\widetilde{\mathbf{u}}\|_\infty$, with (4.31), (4.19) and the definition of S in Definition 1.2, we get that

$$\widetilde{\mathbf{u}} - |M|^2S\widetilde{\mathbf{u}} = \mathrm{O}\left(W^{-1} + (\zeta + |z - \widetilde{z}|)|u|\right). \tag{4.42}$$

Then in this case, with (4.40) and (4.42) it is easy to see that

$$\|\mathbf{u}^0\|_\infty \asymp \|S\mathbf{u}^0\|_\infty \asymp \|\widetilde{\mathbf{u}}\|_\infty \asymp \|S\widetilde{\mathbf{u}}\|_\infty \asymp \|S_0\widetilde{\mathbf{u}}\|_\infty. \tag{4.43}$$

By (1.2), we have

$$\sum_j (S_0\widetilde{\mathbf{u}})_j = 0,$$

which implies

$$\|S_0\widetilde{\mathbf{u}}\|_\infty \leqslant \max_{i,j} |(S_0\widetilde{\mathbf{u}})_j - (S_0\widetilde{\mathbf{u}})_i|\,. \tag{4.44}$$

Using (1.2), for fixed $i \leqslant j \in \mathbb{Z}_N$ we have

$$|(S_0\widetilde{\mathbf{u}})_j - (S_0\widetilde{\mathbf{u}})_i|^2 = \left|\sum_{x,y}(S_0)_{ix}(S_0)_{jy}(\widetilde{\mathbf{u}}_x - \widetilde{\mathbf{u}}_y)\right|^2 \leqslant \sum_{x,y}(S_0)_{ix}(S_0)_{jy}|\widetilde{\mathbf{u}}_x - \widetilde{\mathbf{u}}_y|^2. \tag{4.45}$$

The lower bound in (1.3) shows that S_0 has a core, i.e., there is a constant $c_s > 0$ such that $(S_0)_{xy} \geqslant c_s W^{-1}$ if $|x-y| \leqslant W$. Then for any fixed $i \leqslant j \in \mathbb{Z}_N$, we choose $x_0, x_1, x_2, \cdots, x_n$ for some $n = \mathrm{O}(N/W)$ such that

$$i = x_0 \leqslant x_1 \leqslant x_2 \leqslant \cdots \leqslant x_{n-1} \leqslant x_n = j, \quad \text{with} \quad W/3 \leqslant |x_k - x_{k+1}| \leqslant W/2,\ \forall k.$$

Furthermore, set $x'_0 = x$ and $x'_n = y$. Clearly for any choices of x'_k, $1 \leqslant k \leqslant n-1$, we have

$$\widetilde{\mathbf{u}}_y - \widetilde{\mathbf{u}}_x = \sum_{k=1}^{n}\left(\widetilde{\mathbf{u}}_{x'_k} - \widetilde{\mathbf{u}}_{x'_{k-1}}\right) \Rightarrow |\widetilde{\mathbf{u}}_y - \widetilde{\mathbf{u}}_x|^2 \leqslant \frac{CN}{W}\sum_{k=1}^{n}\left|\widetilde{\mathbf{u}}_{x'_k} - \widetilde{\mathbf{u}}_{x'_{k-1}}\right|^2.$$

For our goal, we will choose x'_k's such that

$$x'_k \in [\![x_k - W/4, x_k + W/4]\!], \quad 1 \leqslant k \leqslant n-1.$$

Taking averaging over all x'_k, $1 \leqslant k \leqslant n-1$, in the above regions, we get that

$$|\widetilde{\mathbf{u}}_y - \widetilde{\mathbf{u}}_x|^2 \leqslant \frac{N}{W}\left(\mathrm{Average}_{x'_1, x'_2, \cdots, x'_{n-1}}\right)\sum_{k=1}^{n}\left|\widetilde{\mathbf{u}}_{x'_k} - \widetilde{\mathbf{u}}_{x'_{k-1}}\right|^2.$$

Note that by our choices, we always have $|x'_k - x'_{k-1}| \leqslant W$ and $S_{x'_k x'_{k-1}} \geqslant \frac{1}{2}c_s W^{-1}$ for $2 \leqslant k \leqslant n-1$, which gives that

$$\begin{aligned}\mathrm{Average}_{x'_{k-1}, x'_k}\left|\widetilde{\mathbf{u}}_{x'_k} - \widetilde{\mathbf{u}}_{x'_{k-1}}\right|^2 &\leqslant \frac{4}{W^2}\sum_{x'_k, x'_{k-1} \in [\![x_{k-1} - W/4, x_k + W/4]\!]}\left|\widetilde{\mathbf{u}}_{x'_k} - \widetilde{\mathbf{u}}_{x'_{k-1}}\right|^2 \\ &\leqslant \frac{8c_s^{-1}}{W}\sum_{x'_k, x'_{k-1} \in [\![x_{k-1} - W/4, x_k + W/4]\!]} S_{x'_k x'_{k-1}}\left|\widetilde{\mathbf{u}}_{x'_k} - \widetilde{\mathbf{u}}_{x'_{k-1}}\right|^2.\end{aligned}$$

Together with (4.45), we get that for some constant $C > 0$,

$$\begin{aligned}|(S_0\widetilde{\mathbf{u}})_j - (S_0\widetilde{\mathbf{u}})_i|^2 \leqslant &\sum_{x,y}(S_0)_{ix}(S_0)_{jy}\left[\frac{CN}{W^2}\sum_{k=2}^{n-1}\sum_{x'_k, x'_{k-1} \in [\![x_{k-1} - W/4, x_k + W/4]\!]} S_{x'_k x'_{k-1}}\left|\widetilde{\mathbf{u}}_{x'_k} - \widetilde{\mathbf{u}}_{x'_{k-1}}\right|^2\right] \\ &+ \sum_{x,y}(S_0)_{ix}(S_0)_{jy}\frac{CN}{W}\left[\frac{2}{W}\sum_{x' : |x' - x_1| \leqslant W/4}|\widetilde{\mathbf{u}}_{x'} - \widetilde{\mathbf{u}}_x|^2 + \frac{2}{W}\sum_{y' : |y' - x_{n-1}| \leqslant W/4}|\widetilde{\mathbf{u}}_y - \widetilde{\mathbf{u}}_{y'}|^2\right].\end{aligned}$$

For the first term on the right-hand side, we have

$$\sum_{k=2}^{n-1}\sum_{x'_k, x'_{k-1} \in [\![x_{k-1} - W/4, x_k + W/4]\!]} S_{x'_k x'_{k-1}}\left|\widetilde{\mathbf{u}}_{x'_k} - \widetilde{\mathbf{u}}_{x'_{k-1}}\right|^2 \leqslant C\sum_{k,l \in \mathbb{Z}_N} S_{kl}\left(\widetilde{\mathbf{u}}_k - \widetilde{\mathbf{u}}_l\right)^2.$$

For the terms in the second line, we notice that

$$|x' - x| \leqslant |x' - x_1| + |x_1 - i| + |i - x| \leqslant C_s W + W$$

for all x' such that $|x'-x_1|\leqslant W/4$, where C_s is the constant appeared in (1.3). Then we can subdivide the interval $[\![x,x']\!]$ or $[\![x',x]\!]$ into subintervals with lengths $\leqslant W/2$, and proceed as above to get

$$\sum_x (S_0)_{ix}\frac{2}{W}\sum_{|x'-x_1|\leqslant W/4}|\widetilde{\mathbf{u}}_{x'}-\widetilde{\mathbf{u}}_x|^2\leqslant\frac{C}{W}\sum_{1\leqslant k,l\leqslant N}S_{kl}\left(\widetilde{\mathbf{u}}_k-\widetilde{\mathbf{u}}_l\right)^2$$

for some constant $C>0$ that is independent of the choice of x'. In sum, we have obtained that

$$|(S_0\widetilde{\mathbf{u}})_j-(S_0\widetilde{\mathbf{u}})_i|^2\leqslant\frac{CN}{W^2}\sum_{1\leqslant k,l\leqslant N}S_{kl}\left(\widetilde{\mathbf{u}}_k-\widetilde{\mathbf{u}}_l\right)^2=\frac{CN}{W^2}\sum_{1\leqslant k,l\leqslant N}S_{kl}\left(\mathbf{u}_k^0-\mathbf{u}_l^0\right)^2.$$

Then from (4.43) and (4.44), we obtain that

$$\|\mathbf{u}^0\|_\infty^2\leqslant\frac{CN}{W^2}\sum_{1\leqslant k,l\leqslant N}S_{kl}\left(\mathbf{u}_k^0-\mathbf{u}_l^0\right)^2.$$

Plugging it into (4.39), we get that if $|u|\leqslant 10\|\widetilde{\mathbf{u}}\|_\infty$, then

$$\frac{W^2}{N}\|\mathbf{u}^0\|_\infty^2\leqslant C\sum_{1\leqslant k,l\leqslant N}S_{kl}\left(\mathbf{u}_k^0-\mathbf{u}_l^0\right)^2\leqslant C\|\mathbf{u}^0\|_\infty+\mathrm{O}(N^{-10})\|\mathbf{u}^0\|_2^2\Rightarrow\|\mathbf{u}^0\|_\infty\leqslant\frac{CN}{W^2}.\tag{4.46}$$

In sum, by our choice of $\mathbf{v}^0=S\mathbf{e}_{i_0}$ and (4.31), we obtain from (4.41) and (4.46) that

$$\left\|\left(1-S|M|^2\right)^{-1}S\right\|_{\max}\leqslant C\left(\frac{1}{W\operatorname{Im}z}+\frac{N}{W^2}\right),$$

which completes the proof of (2.7) in the case with $\mathbf{g}=0$.

Given any $\mathbf{g}\in\mathbb{R}^N$ such that $\|\mathbf{g}\|_\infty\leqslant W^{-3/4}$, we can write

$$M_\zeta^{\mathbf{g}}=M_\zeta^{\mathbf{0}}+\mathcal{E},$$

where $\mathcal{E}$ is a diagonal matrix with $\max_i|\mathcal{E}_{ii}|=\mathrm{O}(\|\mathbf{g}\|_\infty)=\mathrm{O}(W^{-3/4})$ by the Lipschitz continuity estimate (1.10). Then (2.6) can be obtained by combing (2.6) in the case $\mathbf{g}=0$ with a standard perturbation argument. For (2.7), we write

$$\left(1-S|M_\zeta^{\mathbf{g}}|^2\right)^{-1}S=\left(1-S|M_\zeta^{\mathbf{0}}|^2\right)^{-1}S+\left(1-S|M_\zeta^{\mathbf{0}}|^2\right)^{-1}S(|M_\zeta^{\mathbf{g}}|^2-|M_\zeta^{\mathbf{0}}|^2)\left(1-S|M_\zeta^{\mathbf{g}}|^2\right)^{-1}S.\tag{4.47}$$

Using (2.7) in the case $\mathbf{g}=0$ and the bound

$$\left\|\left(1-S|M_\zeta^{\mathbf{0}}|^2\right)^{-1}S\right\|_{L^\infty\to L^\infty}\leqslant N\left\|\left(1-S|M_\zeta^{\mathbf{0}}|^2\right)^{-1}S\right\|_{\max},$$

we get from (4.47) that

$$\left\|\left(1-S|M_\zeta^{\mathbf{g}}|^2\right)^{-1}S\right\|_{\max}\leqslant\left\|\left(1-S|M_\zeta^{\mathbf{0}}|^2\right)^{-1}S\right\|_{\max}+\mathrm{O}\left(\left(\frac{N}{W\operatorname{Im}z}+\frac{N^2}{W^2}\right)W^{-3/4}\right).\left\|\left(1-S|M_\zeta^{\mathbf{g}}|^2\right)^{-1}S\right\|_{\max}.$$

Together with (1.15), this implies (2.7) for any $\mathbf{g}$ such that $\|\mathbf{g}\|_\infty\leqslant W^{-3/4}$. □

Proof of Lemma 4.4. Since the matrix $S_0=(s_{ij})$ has a core by (1.3), it suffices to prove that

$$\sum_{i,j\in\mathbb{T}}\hat{s}_{ij}\left(\mathbf{u}_i-\mathbf{u}_j\right)^2\geqslant c(\log N)^{-13},\quad\forall\mathbf{u}\in L^2(\mathbb{T}),\ \ \|\mathbf{u}\|_2=1,\ \ \mathbf{u}\perp(1,1,\cdots,1),\tag{4.48}$$

where

$$\hat{s}_{ij}:=\frac{1}{W}\mathbb{1}_{|i-j|\leqslant W}.$$

Then we define the following two symmetric operators $F_{0,1} : L^2(\mathbb{T}) \mapsto L^2(\mathbb{T})$ such that for any $\mathbf{u}, \mathbf{v} \in L^2(\mathbb{T})$,

$$(\mathbf{u}, F_0\mathbf{v}) = \frac{1}{W(\log N)^5} \sum_{i,j\in\mathbb{T}, |i-j|_{\mathbb{T}}\leqslant W} (\mathbf{u}_i - \mathbf{u}_j)(\mathbf{v}_i - \mathbf{v}_j),$$

where $|\cdot|_{\mathbb{T}}$ denotes the periodic distance on $\mathbb{T}$, and

$$(\mathbf{u}, F_1\mathbf{v}) = \sum_{i,j\in\mathbb{T}} \tilde{s}_{ij}(\mathbf{u}_i - \mathbf{u}_j)(\mathbf{v}_i - \mathbf{v}_j), \quad \tilde{s}_{ij} := \hat{s}_{ij} - \frac{1}{W(\log N)^5}\mathbb{1}_{|i-j|_{\mathbb{T}}\leqslant W}.$$

We first show that for some constant $c > 0$,

$$E_1(F_0) \geqslant c(\log N)^{-13}, \tag{4.49}$$

where $E_1(F_0)$ denotes the second lowest eigenvalue of F_0. Without loss of generality, we can regard F_0 as an operator on $L^2(\mathbb{T}, \mathbb{C})$ consisting of *complex* L^2 vectors. Since F_0 is a periodic operator on $L^2(\mathbb{T}, \mathbb{C})$, its eigenvectors are the unit complex vectors with Fourier components:

$$\mathbf{w}_p : \quad (\mathbf{w}_p)_k := \frac{1}{\sqrt{|\mathbb{T}|}} e^{\mathrm{i}pk}, \quad k \in \mathbb{T}, \quad \text{with} \quad p = \frac{2\pi n}{|\mathbb{T}|}, \quad n \in \mathbb{T}.$$

Then for any $p \neq 0$, we have

$$\begin{aligned}(\mathbf{w}_p, F_0\mathbf{w}_p) &= \frac{1}{W(\log N)^5} \sum_{|k-l|_{\mathbb{T}}\leqslant W} |(\mathbf{w}_p)_k - (\mathbf{w}_p)_l|^2 = \frac{1}{|\mathbb{T}|W(\log N)^5} \sum_{|k-l|_{\mathbb{T}}\leqslant W} [2 - 2\cos(p(k-l))] \\ &= \frac{1}{W(\log N)^5} \sum_{|n|\leqslant W} [2 - 2\cos(pn)] \geqslant \frac{c}{W(\log N)^5} \frac{W^3}{|\mathbb{T}|^2} \geqslant c(\log N)^{-13}.\end{aligned}$$

This proves (4.49).

We now show that F_1 defines a positive operator. For simplicity of notations, we let $L = |\mathbb{T}|$ and shift $\mathbb{T}$ to $\mathbb{T} := [\![1, L]\!]$. Then $\tilde{s}_{ij}$ can be written as

$$\tilde{s}_{ij} = \left(1 - (\log N)^{-5}\right)\hat{s}_{ij} - \frac{1}{W(\log N)^5}\left(\mathbb{1}_{1\leqslant i\leqslant W, L-W+i\leqslant j\leqslant L} + \mathbb{1}_{1\leqslant j\leqslant W, L-W+j\leqslant i\leqslant L}\right). \tag{4.50}$$

Fix any $\mathbf{u} \in L^2(\mathbb{T})$. The following proof is very similar to the one below (4.45), so we shall omit some details. For any fixed $1 \leqslant i \leqslant W$ and $L - W \leqslant j \leqslant L$, we choose $x_0, x_1, \cdots, x_n$ for some $n = \mathrm{O}((\log N)^4)$ such that

$$i = x_0 \leqslant x_1 \leqslant x_2 \leqslant \cdots \leqslant x_{n-1} \leqslant x_n = j, \quad \text{with} \quad W/3 \leqslant |x_k - x_{k+1}| \leqslant W/2, \ \forall k.$$

Moreover, we set $x'_0 = i$ and $x'_n = j$. Then we can get as before that

$$|\mathbf{u}_i - \mathbf{u}_j|^2 \leqslant C(\log N)^4 \left(\mathrm{Average}_{x'_1, x'_2, \cdots, x'_{n-1}}\right) \sum_{k=1}^{n} \left|\mathbf{u}_{x'_k} - \mathbf{u}_{x'_{k-1}}\right|^2,$$

where we took average over all $x'_k \in [\![x_k - W/4, x_k + W/4]\!]$, $1 \leqslant k \leqslant n-1$. Note that by our choices, we always have $|x'_k - x'_{k-1}| \leqslant W$ and $\hat{s}_{x'_k x'_{k-1}} = W^{-1}$ for $1 \leqslant k \leqslant n$, which gives that

$$\begin{aligned}&\frac{1}{W(\log N)^5} \sum_{1\leqslant i\leqslant W, L-W\leqslant j\leqslant L} |\mathbf{u}_i - \mathbf{u}_j|^2 \\ &\leqslant \frac{1}{W(\log N)^5} \sum_{1\leqslant i\leqslant W, L-W\leqslant j\leqslant L} \left[\frac{C(\log N)^4}{W} \sum_{k=2}^{n-1} \sum_{x'_k, x'_{k-1} \in [\![x_{k-1} - W/4, x_k + W/4]\!]} \hat{s}_{x'_k x'_{k-1}} \left|\mathbf{u}_{x'_k} - \mathbf{u}_{x'_{k-1}}\right|^2\right] \\ &+ \frac{1}{W(\log N)^5} \sum_{1\leqslant i\leqslant W, L-W\leqslant j\leqslant L} C(\log N)^4 \left[\sum_{x:|x-x_1|\leqslant W/4} \hat{s}_{xi} |\mathbf{u}_x - \mathbf{u}_i|^2 + \sum_{y:|y-x_{n-1}|\leqslant W/4} \hat{s}_{jy} |\mathbf{u}_y - \mathbf{u}_j|^2\right] \\ &\leqslant C(\log N)^{-1} \sum_{k,l\in\mathbb{T}} \hat{s}_{kl} (\mathbf{u}_k - \mathbf{u}_l)^2.\end{aligned}$$

Then by (4.50), it is easy to see that F_1 is a positive operator. Thus by min-max principle we have

$$E_1(F_0 + F_1) \geqslant E_1(F_0),$$

which proves (4.48) together with (4.49). □

References

[1] 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.

[2] P. Bourgade, H.-T. Yau, and J. Yin, *Random band matrices in the delocalized phase, I: Quantum unique ergodicity and universality*, in preparation (2018).

[3] L. Erdős, A. Knowles, and H.-T. Yau, *Averaging fluctuations in resolvents of random band matrices*, Ann. Henri Poincaré **14** (2013), 1837–1926.

[4] 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.

[5] ———, *The local semicircle law for a general class of random matrices*, Elect. J. Prob. **18** (2013), no. 59, 1–58.

[6] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, *Spectral statistics of Erdős-Rényi graphs II: Eigenvalue spacing and the extreme eigenvalues*, Comm. Math. Phys. **314** (2012), 587–640.

[7] L. Erdős, B. Schlein, and H.-T. Yau, *Local semicircle law and complete delocalization for Wigner random matrices*, Commun. Math. Phys. **287** (2008), no. 2, 641–655.

[8] 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.

[9] A. Knowles and J. Yin, *The isotropic semicircle law and deformation of Wigner matrices*, Comm. Pure Appl. Math. **66** (2013), 1663–1749.

[10] F. Yang and J. Yin, *Random band matrices in the delocalized phase, III: Averaging fluctuations*, in preparation (2018).